Forecasting With Big Data: A Review

[1]Adline Rajasenah Merryton and [2]Dr. Gethsiyal Augasta. M,
[1]Assistant Professor of Computer Science, Sarah Tucker College, Tirunelveli, Tamilnadu, India.
[2]Assistant Professor of Computer Science, Kamaraj College, Thoothukudi, Tamilnadu, India.

Abstract: Nowadays the forecasting approach shapes up on the improvements in economic measurements. Recent studies record that Big Data is a new variety of strategic resource in the digital era and a key factor to drive innovation which is changing the ways of mankind. The recent demonetization has seen an increase in the fertile importance of risk management in organizations and firms. At this point, huge organizations find that Big Data is useful in Risk Management whilst Big Data forecasting has the power to improve their performance which in turn enables better Risk Management. The capturing, cleaning, integrating, storing, processing, indexing, searching, sharing, transferring, mining and visualization of large volumes of fast moving highly complex data are identified as the major challenges which obstruct the process of forecasting using Big Data. As per the reviews Big Data has deeply rooted in the fields of Business, Economics, Population Dynamics, Weather, Crime and so on. The standard tools adopted for forecasting with Big Data are Time Series Regression Model, Factor Models, Neural Networks, Simulation Models and Bayesian Models. This paper deals with a study on the use of Big Data in forecasting through Reviews, Techniques, Tools, Fields, challenges and applications.

Keywords: *Forecasting, Techniques, Regression, Decomposition Forecasting, Challenges.*

I. INTRODUCTION

A. *Big data*

Big data is one of the most frequently discussed topics in this digital era. With the evolution of technology and the increased multitudes of data flowing in and out of organizations daily, there has become a need for faster and more efficient ways of analyzing such data. Having piles of data on hand is no longer enough to make efficient decisions at the right time. Such data sets are no longer be easily analyzed with traditional data management and analysis techniques and infrastructures. There arises a need for new tools and methods specialized for big data analytics, as well as the required architecture for storing and managing such data. Accordingly, the emergence of big data has an effect on everything from the data itself and its collection, to the processing, to the final extracted decisions. For instance, demonetization in India in the year 2016. The demonetization drive mandated the Smartphone-wielding Indians to e-wallets. The drive ushered in digital governance with massive use of data collected from various sources to strategize the demonetization scheme [1].

"Big data analysts and data scientists will link this emerging data with existing information already with the tax and other departments or relevant ministries including customs and excise, property registration bodies, courts, police and phone records, district authorities, utilities, net banking transactions (linked to digital wallets and credit cards) etc., unearthing even more data of relevance to highlight unorthodox or revealing financial insights" [1].

Implementing such gigantic scheme is impossible without the use of new age technologies like big data, Internet of Things (IOT), e-wallets, robotics and all that algorithm build-up and writing. Digitization and Big Data also help the tax revenues to increase and plug existing loopholes in the system. Though there are many upsides to the revolutionary demonetization drive, using Big Data the downside is data storage capabilities. The push for a cashless economy comes with a warning about the privacy of data generated via e-wallets and other similar apps.

Using Big Data techniques the data can be taken from any source and can be analyze it to find answers that enable Cost Reduction, Time Reduction, New Product development, Smart Decision Making. Big Data is a term for Data Set that is so large or complex, that traditional data processing application software like RDBMS can't handle to give a best result. Data must be analyzed and the value is extracted to a particular size of Data set. It ranges from a few dozen of Terabytes to Petabytes [2]. Latest technologies and cloud based applications are the prerequisites to overcome the limits of conventional RDBMS. Big Data challenges include capturing data, Data Storage, Data Analysis, Search, Storing, Transfer, Visualization, Querying, Updating and Information Privacy. There are five main features characterize Big Data: Volume, Variety, Velocity, Variability and Veracity or the five V's. The Volume of the data is its size and enormous it is. Velocity refers to the rate with which data is changing, or how often it is created. Variety includes the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data. Variability refers to the inconsistency of the data set can store processes to handle and manage it. Veracity refers to the quality of captured data that can be varying greatly, affecting the accurate analysis [3].

Large sites like Twitter, Face book, Linked in, Google, Amazon and so on need most up-to-date database management technologies to work with active and complex datasets.

The distinction between Big Data and the Databases which were used by those older organizations is its size, complication and its rapid growth. This growth needs innovative tools to handle the challenges. It also needs competent and effective technology to process massive volume of data in an organized approach.

B. *Forecasting*

The fundamental activity of any organization is planning. The basis of planning is forecasting. It is a scientifically calculated presumption and is basic to all planning activity long or short range—whether it is national, regional, organizational, or functional planning. The scientific basis of forecasting rests with studying the past, present and future trends, present and future actions and their effects. The happenings of the past are relevant to what is happening now and what will happen in the future. Thus, forecasting takes into account all the three dimensions of time – past, present and future. In spite of all the calculations, forecasting remains a calculated presumption. Even though there are errors it remains the foundation for management planning. Forecasting is essentially the study of internal and external forces that shape demand and supply [4].

Prediction is a similar, but more general term. Both might refer to formal statistical methods employing time series and cross sectional or longitudinal data, or alternatively to less formal judgmental methods. Usage can differ between areas of application: for example, in hydrology the terms "forecast" and "forecasting" are sometimes reserved for estimates of values at certain specific future times, while the term "prediction" is used for more general estimates, such as the number of times floods will occur over a long period [5].

Risk and uncertainty are central to forecasting and prediction. In any case, the data must be up to date in order for the forecast to be as accurate as possible. Forecasting starts with certain presuppositions based on the management's experience, knowledge, and judgment. These estimates are projected into the coming months or years using one or more techniques such as Box-Jenkins models, Delphi method, Exponential Smoothing, Moving Averages, Regression Analysis, and Trend Projection[6]. The paper is organized as follows: Section 2 explains the Literature Review of Forecasting with Big Data, Section 3 describes the Techniques and Methods of Forecasting, Section 4 describes the Tools for Forecasting with Big Data, Section 5 describes the Fields in which Forecasting with Big Data is used, Section 6 describes the Issues and Challenges of Forecasting with Big Data.

II. LITERATURE REVIEW

Lirby [7] compared three different time-series methods viz., moving averages, exponential smoothing, and regression. He found that in terms of month-to-month forecasting, horizon was increased to six months. The regression models included was found to be the best method for longer-term forecasts of one year or more. Sleckler[8] found that econometric models were not entirely successful in improving the accuracy in forecasting.

Richards and King [9] consent that predictions can be improved through data driven decision making. Tucker [10] believes Big Data will soon be predicting our every move, and according to Einav and Levin [11], Big Data is most commonly sought after for building predictive models in a world where forecasting continues to remain a vital statistical problem [12].

Rey and Wells [13] believe Data Mining techniques can be exploited to help forecasting with Big Data. Cukier [14] finds fault with Big Data for the recent financial crisis as he believes the financial models adopted were unable to handle the huge amounts of data that was being inputted into the systems, and thereby resulted in inaccurate forecasts. Dazhi Chong and Hui Shi [15] states that big data analytics has become a key factor for companies to reveal hidden information and achieve competitive advantages in the market.

III. FORECASTING TECHNIQUES

A. *Moving Averages*

Moving average method uses the average of the actual past and projects it for future. This method can be used for products with quite little change like fixed demand, no seasonality, limited trends or cycles, and no significant demand shifts. Most of the companies apply this method because of its simplicity and user friendliness.

Intuitively, the simplest way to smooth a time series is to calculate a simple, or unweighted, moving average. This is known as using a rectangular or "boxcar" window function. The smoothed statistic s_t is then just the mean of the last k observations and it is calculated using equation (1)

$$s_t = \frac{1}{k}\sum_{n=0}^{k-1} x_{t-n} = \frac{x_t + x_{t-1} + x_{t-2} + \cdots + x_{t-k+1}}{k} = s_{t-1} + \frac{x_t - x_{t-k}}{k},$$

--(1)

Where the choice of an integer $k > 1$ is arbitrary. A small value of k will have less of a smoothing effect and be more responsive to recent changes in the data, while a larger k will have a greater smoothing effect, and produce a more pronounced lag in the smoothed sequence. [16]

B. *Exponential Smoothing*

This is an advanced form of time series forecasting. Unlike moving averages, this method can capture trends and recurring patterns. They achieve this by emphasizing the more recent data and smoothing out the noise, which are often caused by pure randomness in the data. The simplest form of exponential smoothing is given by the equation (2):

$$s_t = \alpha \cdot x_t + (1 - \alpha) \cdot s_{t-1}.$$

- (2)

where α is the *smoothing factor*, and $0 < \alpha < 1$. In other words, the smoothed statistic s_t is a simple weighted average of the current observation x_t and the previous smoothed statistic s_{t-1}. [16]

C. *Regression Analysis Models*

Regression [17] comes in all shapes and colors. This model is used to determine the association between demand and demand drivers. They are especially useful for knowing trends and seasonality. Regression analysis is used to find equations that fit data. Once we have the regression equation, we can use the model to make predictions. One type of regression analysis is linear analysis. The equation for a line is **y = mx + b**. Linear regression is a way to model the relationship between two variables. We might also recognize the equation as the slope formula. The equation has the form Y=a+bX, where Y is the dependent variable (that's the variable that goes on the Y axis), X is the independent variable (i.e. it is plotted on the X axis), b is the slope of the line and a is the y-intercept, calculated by the equation (3).

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} \qquad b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

- (3)

D. *Hybrid Forecasting Methods*

Hybrid Forecasting Methods combine regression, data smoothing, and other techniques to forecast. For instance, the forecasting methods which are great at short-term forecasting cannot capture seasonality.

E. *Decomposition Forecasting Methods*

Decomposition Forecasting Methods are highly effective at finding multiple and delicate patterns. They are good at daily forecasts. These methods decompose historical data into various patterns that can concurrently contain annual seasonality, day-of-the-week patterns, fixed-date and floating-holiday impacts, and others**.**

Classical time series decomposition[18] separates a time series into five components: mean, long-range trend, seasonality, cycle, and randomness. The decomposition model is,

Value = (Mean) x (Trend) x (Seasonality) x (Cycle) x (Random)

The basic decomposition method consists of estimating the five components of the model using the equation (4)

$$X_t = UT_tC_t\ S_t\ R_t \qquad - (4)$$

Where,

X_t denotes the series or, optionally, log of series.

U denotes the mean of the series.

T_t denotes the linear trend.

C_t denotes cycle.

S_t denotes season.

R_t denotes random error.

t denotes the time period.

F. Spectral Analysis

Spectral-analysis forecasting models are effective at filtering noise out of cyclical data. It can draw out seasonal patterns from data which also have a sturdy monthly pattern. Otherwise, the seasonality would be complicated to perceive. Fig. 1 shows the spectral analysis of the sounding note c' (261.6 Hz) played *mf* on bass which shows the differing volumes of harmonics 1-8 which make up the instruments' characteristic sounds.

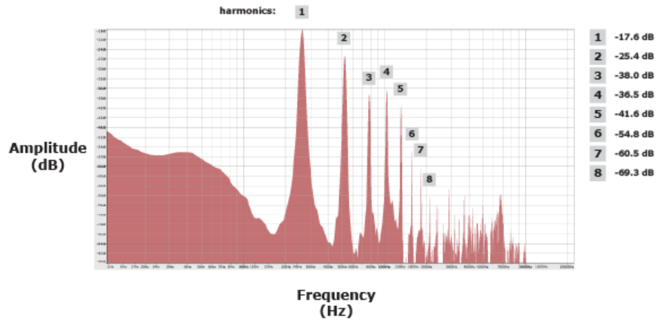


Fig. 1 Spectral Analysis of a Bass Flute

IV. TOOLS FOR FORECASTING WITH BIGDATA

Almost all the Big Data tools which are available in the market are open source. The following tools are commonly used nowadays.

Hadoop and MapReduce[19] is one of the commonly used Big Data tool. It is a programming model used for scripting applications to process enormous amount of data. Here, the master takes care of scheduling, monitoring and re-execution of the failed tasks whereas the slave perform the tasks as per the command of the master. Gridgain which is an alternative to MapReduce is used for fast analysis of real time data usage in memory processing. High Performance Computing Cluster (HPCC) and STORM are also considered as tools of Big Data.

Apart from tools Big Data is widely connected with some Databases. Some of them are Apache Cassandra, Apache HBase, Mongo DB Neo4J, Apache CouchDB, Terrastore, FlockDB, RIAK, Hypertable, Hive and so on[19].

Apache Cassandra is a distributed database management system which is an open source developed by Facebook. Apache HBase is designed to run on Hadoop Distributed File system. It gives real time access to Hadoop. MongoDB is used by MapReduce for batch processing. Neo4j is a graph database model which is thousand times faster than conventional DBMS. Apache CouchDB performs MapReduce queries through JavaScript. It brings together the Smart Objects. Terrastore is highly scalable and reliable. FlockDB is a graph oriented database like Neo4j. RIAK is a disseminated key-value data store. Hypertable is designed after Bigtable. It uses HQL (Hypertable Querying Language) as its own querying language. Hive is a Hadoop based data warehouse which uses its own querying language called HiveQL .

Like those commonly used Big Data tools there is a set of tools which is used for business intelligence. They are Talend, Jaspersoft, Jedox, Pentaho, SpagoBI, Knime, BIRT etc.

V. FIELDS OF FORECASTING WITH BIGDATA

In the present scenario of data processing, the concept Big Data is everywhere. The drastic increase in the development of data paved way for Big Data. It has its roots in many fields. The following areas will have the great opportunities.

A. Manufacturing Industries

Many manufacturing companies are having automated machines in their production process which generates more data. Big Data tools help the manufacturing industries to store, retrieve and analyze the data in forecasting the future demand of their products. Fig. 2 shows the demand forecast for a product for 12 months.

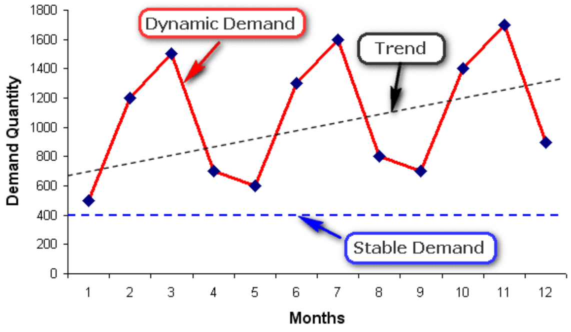


Fig. 2. Demand Forecast for a product

B. Defense

Information is an important treasure in the defense sector. Data received from satellites and messages from various devices are important in identifying the enemies and other terrorist movements. Fig. 3 shows the forecast for Indian Other Defence Spending for the years 2015-2025.

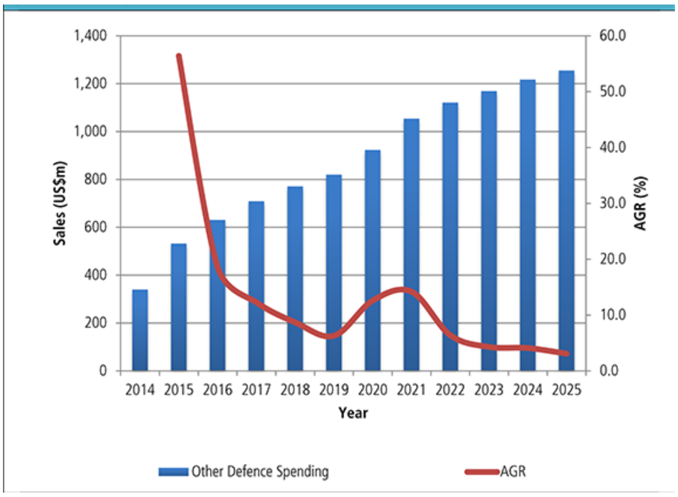
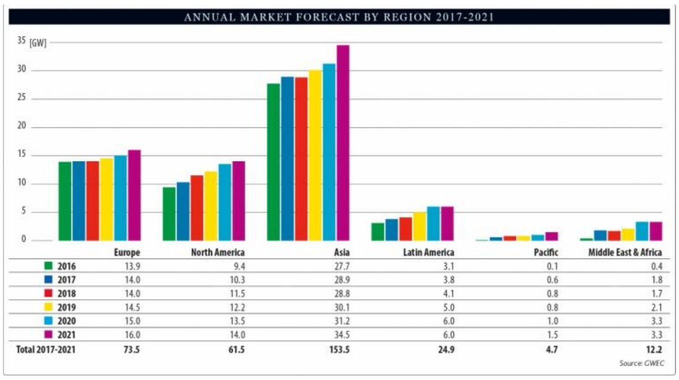


Fig. 3. Indian Other Defence Spending Forecast 2015-2025

C. *Marketing*

Big Data is not a magic box which directly leads to better marketing of products. Careful and detailed analysis using the tools of Big Data gives success. Proper analytics helps marketing companies to predict future requirements using the present day purchase. Fig. 4 shows the Annual Forecast of wind energy marketing in various continents for the years 2017-2021.



	Europe	North America	Asia	Latin America	Pacific	Middle East & Africa
🟩 2016	13.9	9.4	27.7	3.1	0.1	0.4
🟦 2017	14.0	10.3	28.9	3.8	0.6	1.8
🟥 2018	14.0	11.5	28.8	4.1	0.8	1.7
🟨 2019	14.5	12.2	30.1	5.0	0.8	2.1
🟦 2020	15.0	13.5	31.2	6.0	1.0	3.3
🟪 2021	16.0	14.0	34.5	6.0	1.5	3.3
Total 2017-2021	**73.5**	**61.5**	**153.5**	**24.9**	**4.7**	**12.2**

Source: GWEC

Fig. 4. Annual Forecast of Wind Energy Marketing for the years 2017-2021

D. *Healthcare*

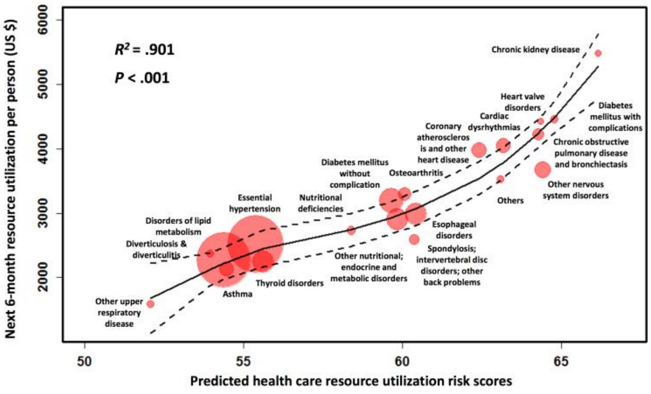


Fig. 5. Close examination of the prospective analysis of next 6-month resource utilizations stratified by the top 20 most common chronic diseases.

A big revolution is happening in the genetic field. New research direction arrived by Genome project. Big Data plays significant role in data storage, retrieval, sequence analysis and visualization. The goal of modern healthcare systems is to provide optimal health care through the meaningful use of health information technology in order to improve healthcare quality and coordination, to reduce healthcare costs; reduce avoidable overuse and to provide support for reformed payment structures. Fig. 5 shows the forecast of resource utilizations[20] by top 20 most common chronic diseases.

E. *Economics*

The data revolution of the past decade is likely to have a further and profound effect on economic research. Increasingly, economists make use of newly available large-scale administrative data or private sector data that often are obtained through collaborations with private firms, giving rise to new opportunities and challenges [21].

VI. ISSUES ON FORECASTING WITH BIG DATA

This paper focuses mainly on the challenges which need to overcome when forecasting with Big Data. A good example is the existence of a vast amount of data on earthquakes, but the lack of reliable model that can accurately predict earthquakes [22]. Some existing challenges are related to hypothesis, testing and models utilized for Big Data forecasting.

Big Data is a challenge for forecasting as a result of its inherent characteristics. Firstly Big Data evolves and changes in real time, and as such it is important that the techniques used to forecast Big Data are able to transform unstructured data into structured data [23], accurately capture these dynamic changes and detect change points in advance. Secondly, there are challenges stemming from Big Data's highly complex structure and as [24] point out, it is challenge to build forecasting model that do not result in poor out-of-sample forecasts owing to the over use of potential predictors. There are a wide range of problems that Big Data will tackle over the next few years, but there are several factors that will slow down its development. These include:

- Limited supply of data scientists and people capable of working with Big Data.
- The growth in noise is corrupting the signal that businesses hope to find in the data.
- There is a lack of models, making data less useful than it might be.

VII. FUTURE CHALLENGES

The success of Forecasting with Big Data in the enterprises requires biggest cultural and technological change. Enterprise wise strategy required to derive the business value by integrating the available traditional data. Biggest challenges in Forecasting analysis are Heterogeneity and Incompleteness, Scalability, Timeliness and security of the data [25]. Privacy is one of the major concerns for the outsourced data. Policies should be deployed and rule violators should be identified to avoid the misuse of data. Data integrity is a challenge for the data available in cloud platform. Organizational leaders should take the initiative to understand and move towards Forecasting with Big Data. Skilled people required for the shift to Big Data. It requires people in the area of system analysis, domain knowledge, data analytics, database management and software developers. Huge numbers of open source technologies are available in the market for Forecasting. Few are discussed in the previous section. Selection of right tool is also a challenge.

CONCLUSION

Various models and tools regarding forecasting using Big Data has been discussed in this paper. The outcome of this study brings out the importance of Big Data in Forecasting and the requirements of change and adoption to latest technologies. Big Data databases ensure better performance than traditional RBDMS in various use cases. There are many open source software available in the market. But the choice of selecting best Big Data tool is a challenge for the programmers for developing efficient scalable application. Clear analysis required before selecting the tools from developer and users point of view.

Most of the Big Data tools for Forecasting available in the market are open source. The biggest challenges in front of all the enterprises are the requirement of cultural and technological change to adopt the new technology. Valuable insights will be derived from available traditional data also. Organizational leaders should take the initiative to understand and move towards the Big Data. Based on past researchers it is evident that Factor models are the most common and popular tool currently used for Big Data forecasting. Future research problems in forecasting will promise the benefits of Big Data.

References

[1] Deoras, Srishti. "Demonetization and the role of big data analytics in preventing financial malpractices" [Accessed 18 August 2017]

[2] Big Data Analytics: A Literature Review – Nada Elgendy and Ahmed Elragal

[3] EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012)

[4] Chary, S N (2004), Production and Operations Management, Tata McGraw Hill: New Delhi. PP 8.1-8.18.

[5] Garg, P.K. *Forecasting Management: Futurism on Management*, Global India Publications, 2009.

[6] Angwin, Duncan, Stephen Cummings, Chris Smith, *The Strategy Pathfinder: Core Concepts and Live Cases*, John Wiley & Sons, 2011, pp. 234.

[7] R.M.Lirby, " A comparison of Short and Medium Range Statistical Forecasting Methods", Management Science, 4: 8202-210, 1966.

[8] Olatunji ,Oladejo Michael, Abdullahi Bello. "A Suitable Model for the Forecast of Exchange Rate in Nigeria (Nigerian Naira versus US Dollar)", International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064. Pp 2670. (Accessed on 23.08.2017)

[9] Richards, N.M., King, J.H. (2013). "Three Paradoxes of Big Data", Stanford Law Review Online, 66(41), PP. 41- 46.

[10] Tucker, P, "The Future is Not A Destination", 2013.

[11] Einav, L., and Levin, J.D. (2013). "The Data Revolution and Economic Analysis", Working aer no. 19035, National Bureau of Economic Research.

[12] Hand, D. J. (2009). Mining the Past to Determine the Future: Problems and Possibilities. International Journal of Forecasting, 25(3), pp. 441-451.

[13] Rey, T., and Wells, C. (2013). Integrating Data Mining and Forecasting. OR/MS Today, 39(6).

[14] Cukier, K. Data, data everywhere. The Economist, 2010, PP. 14.

[15] Chong, Dazhi & Hui Shi. "Big data analytics: a literature review", Journal of Management Analytics, Volume 2, 2015 - Issue 3. Pp 175-201.

[16] Armenico, Bertiz R, Vinluan Albert A , et. al. "Forecasting Model for Criminality in Barangay Commonwealth, Quezon City, Philippines using Data Mining Techniques", *International Journal of Conceptions on Computing and Information Technology*, Vol. 3, Issue. 3,2015, PP 28-33.

[17] Stephanie,"Find a Linear Regression Equation by Hand or in Excel", (2017), [accessed August 23,2017]

[18] Alam, Maskurul, Matiur Rahman, Sharmin Akter Sumy & Yasin Ali Parh, "Time Series Decomposition and Seasonal Adjustment", *Global Journal of Science Frontier Research: F Mathematics and Decision Sciences*, Volume 15, Issue 9, 2015.

[19] Caldarola, Enrico Giacinto, Antonio Maria Rinaldi (2015), "Big Data: A Survey The New Paradigms, Methodologies and Tools". [Accessed on August 20,2017]

[20] Hu , Zhongkai, Shiying Hao, et. al., "Online Prediction of Health Care Utilization in the Next Six Months Based on Electronic Health Record Information: A Cohort and Validation Study", *Journal of Medical Internet Research*, Vol 17, No 9 (2015).

[21] Einav, Liran, Jonathan Levin, "Economics in the Age of Big Data", 07 NOV 2014, [accessed on 17.08.2017]

[22] Silver N (2012) The signal and the Noise: The Art and Science of Prediction. Penguin Books, Westmins

[23] Shi Y (2014) Big Data: history, current status, and challenges going forward. Bridge 44(4):6–11

[24] Einav L, Levin JD (2013) The data revolution and economic analysis. Working Paper No. 19035, National Bureau of Economic Research.

[25] Padmavalli. M, "Big Data: Emerging Challenges of Big Data and Techniques for Handling", IOSR Journal of Computer Engineering, PP 13-18. (Accessed on August 23, 2017)