Journal of International Technology and Information Management

Volume 26 | Issue 2 Article 1

6-1-2017

Comparative Analysis of Big Data Analytics Software in Assessing Sample Data

Soly Mathew Biju
UOWD, dr.solymathew@gmail.com

Alex Mathew
alex.k.mathew@gmail.com

Follow this and additional works at: https://scholarworks.lib.csusb.edu/jitim

 Part of the Business Intelligence Commons, Communication Technology and New Media Commons, Computer and Systems Architecture Commons, Data Storage Systems Commons, Digital Communications and Networking Commons, E-Commerce Commons, Information Literacy Commons, Management Information Systems Commons, Management Sciences and Quantitative Methods Commons, Operational Research Commons, Science and Technology Studies Commons, Social Media Commons, and the Technology and Innovation Commons

Recommended Citation

Biju, Soly Mathew and Mathew, Alex (2017) "Comparative Analysis of Big Data Analytics Software in Assessing Sample Data," *Journal of International Technology and Information Management*: Vol. 26 : Iss. 2 , Article 1.
Available at: https://scholarworks.lib.csusb.edu/jitim/vol26/iss2/1

This Article is brought to you for free and open access by CSUSB ScholarWorks. It has been accepted for inclusion in Journal of International Technology and Information Management by an authorized editor of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.

COMPARATIVE ANALYSIS OF SELECTED BIG DATA ANALYTICS TOOLS

Soly Mathew Biju,
dr.solymathew@gmail.com

Alex Mathew,
alex.k.mathew@gmail.com

ABSTRACT

Over the last few years, big data has emerged as an important topic of discussion in most firms owing to its ability of creation, storage and processing of content at a reasonable price. Big data consists of advanced tools and techniques to process large volumes of data in organisations. Investment in big data analytics has almost become a necessity in large-sized firms, particularly multinational companies, for its unique benefits, particularly in prediction and identification of various trends. Some of the most popular big data analytics software used today are MapReduce, Hive, Tableau and Hive, while the framework Hadoop enables easy processing of such extremely large data sets. The current research attempts to create a comparative assessment of five such applications namely IBM SPSS, IBM Watson Analytics, R, Minitab and SAS. The case taken into effect for the test was that of the factors affecting housing affordability in the US. Based on the statistics obtained from the American Housing Survey (AHS) database, the researcher has identified different factors impacting the affordability in the states. The technique of reducing variables though Principal Component Analysis (PCA) and a model based on partial least square regression/polynomial regression was fitted to check the impact on the affordability. The primary findings suggest that majorly age of the head of the household, income earned were the two most important factors affecting the pricing in the region. Also, a comparison is drawn at the end of study with interpretation of the most and least effective applications.
KEY WORDS: Big data, data analytics, prediction analysis, analytic tools.

©International Information Management Association, Inc. 2017

ISSN: 1941-6679-On-line Copy

INTRODUCTION

Since the advent of the internet, the amount of data being generated every day is increasing manifold. Today, over 2.5 quintillion bytes of data is being generated in the form of videos, text, pictures, transaction records, and GPS signals, among many others (VCloudNews, 2015). Companies possess massive amounts of data due to widespread internet usage, and they are now realizing the benefits of using analytics to analyse and extract information from the data. Business intelligence and analytics (BI&A) has become an indispensable part of business operations. The opportunities associated with big data analysis have helped in generating significant interest in techniques, analytical knowledge and methodologies practiced by firms in the field of extended data processing (Chen, Chiang and Storey, 2012). The rush to analyse critical data has resulted in the advent of advanced analytical intelligence for different data segments such as warehousing, suppliers, aviation industry and many others to forecast future predictions. Progressive technological landscape has given a new vision to accumulate a data of large size of population availing the services either through social media or cell phones.

When communicating information through traditional forms of technology such as telephone or telefax, the type of data transmitted is relatively simple and limited in size as compared to data transmitted through emails (Murray, 2014). Web analytics, social media and ecommerce websites often have a large database as per changes in scale and demand. However the problem is that the structure and design of conventional data analytics software are not capable of handling data of enormous size and complexity (Marz and Warren, 2015). Many such software are unable to generate potentially significant results for the data amassed and provide an idea about extensive details of the model owing to this complexity.

Big data is complex because of its diversity and rapidly evolving environment, creating a challenge for companies, particularly those engaged in science and engineering domains like biological sciences (Xindong Wu *et al.*, 2014). The factors creating the complexity have been explored diversely in literature; for instance, (Jin *et al.*, 2015) identified the 5 Vs responsible for making big data complex: huge Volume, high Velocity, high Variety, low Veracity, and high Value. Out of these, 3Vs (Variety, Volume and Velocity) have been recognized by more number of researchers like Kwon et al. (2014) and Chen et al. (2012). Further, it is observed that advanced trends in big data in recent years have improved access to different parts of the world and enabled framing of more effective business strategies (Lohr, 2012). Data analytics generate a systematic framework to form new system architecture for data acquisition, sharing, storing

©International Information Management Association, Inc. 2017

ISSN: 1941-6679-On-line Copy

and large scale processing (Hu *et al.*, 2014). The accurate economic forecasting contributed through big data analytics helps in developing algorithms to achieve more efficient results.

The Big Data Applications are a new type of software applications that leverage large data sets for processing. Statistical Analysis Systems (SAS), Sisene, High Performance Computing Cluster (HPCC) Systems, Talend, Pentaho, Tableau, Informatica and many others are used to understand information and process such a big volume of data (Oracle, 2013). Large sized companies have shown progress in generating these analytical systems to handle critical issues and challenges of big data due to the financial and technical expertise available at their disposal (Khan *et al.*, 2014). The inherent developmental approach interconnects multiple phenomena and identifies patterns on several factors and problems to which not many solutions existed earlier (Belle *et al.*, 2015). With the emerging trends in big data, the new software applications vary vastly in features, scalability and usability. The aim of this study is to evaluate some such existing popular big data analytics software used and compare them for their limitations in identification of patterns/ trends in the data.

LITERATURE REVIEW

Big data provides a scientific paradigm to solve different problems in the fields of economic and business activities. It draws huge attention because of its highly useful insights to produce an idea about trends forecasting in the field (Chen & Zhang, 2014). Big data systems expand into all dimensions generating information about the different sources and factors effecting trends and patterns of business trends (George, Haas and Pentland, 2014). According to Forbes (2016), the total value of business analytics software will be valued at over $187 billion by 2019. It is affirmed that this amount of data will enhance the productivity of organizations, plus documenting and collecting the data has unveiled new potential for publishing revenue statements in more sophisticated manner(Cao, Chychyla and Stewart, 2015). Varian (2014) opined that a considerable number of economic transactions made today are processed electronically and the traditional methods of computing regression and econometric techniques do not work well these big data sets. Compression and management of such a huge volume of data is done using machine learning techniques such as decision trees, neural network and others(Mishra and Badhe, 2016). Cloud computing, another path-breaking technology in this regard, helps alleviate technology costs and patterns by creating new ways for individuals to consume goods and services and develop efficient business models (Eddy, 2015). The flexible nature of cloud computing makes it

ideal for handling big data projects efficiently. Catlett (2013) also corroborated the same in his research on big data and cloud computing.

Predictive modeling is effective in any commercial industry where standardization of data is the norm, regardless of geographical boundaries, technology or other standards (Gantz and Reinsel, 2012). The major role of big data technology is observed in organizations having periodic requirement for processing data related to different functions such as merchandising, supply chain management, promotions, fraud detection, pricing, loss prevention and customer feedback. Different applications specializing in each of these functions are available, apart from a universal software. Chen, Alspaugh, & Katz (2012) cited in their research that MapReduce based systems are used for processing large dataset and interact with Cloudera customers dealing in retail, media, telecommunications and other sectors. In the backdrop of diversified customer requirements, deployment of information technology in retail sector has developed business models that deliver seamless customer experience (Piotrowicz and Cuthbertson, 2014). Big data companies aggregate supply chain data in many of their functions catering to manufacturing customers and sell software tools to improve their performance (Brown, Chui and Manyika, 2011). Tan, Zhan, Ji, Ye, & Chang (2015) stated in their research that the deduction graph technique (techniques with structure of acyclic directed graphs with boxes) provides an analytical infrastructure to incorporate the competency of the firms in gaining supply chain innovation capabilities. Moreover, the concept of big data analysis has helped in fraud detection by detecting the outliers and non-standard data types (Russom, 2011). Tan et al. (2015) found in their research that fraud detection using Security Information and Event Management (SIEM) tools can help manage unstructured data in heterogeneous and noisy formats efficiently.

Analysis of big data used is both advantageous and disadvantageous for a company. A major research into massive amount of data helps in revealing patterns and predicting the model accurately (Sagiroglu and Sinanc, 2013). The uses of real time information along with IT capabilities help in transforming decision and reap better results than the conventional business (Davenport, Barth and Bean, no date). However there exist some disadvantages with respect to big data analytics software which cause some of them to perform better than the other in analysis and prediction of trends. Companies have to maintain a competitive edge in order to survive in a cut-throat competitive environment, absence of which results in failure and elimination from the market (Villars and Olofson, 2014).

©International Information Management Association, Inc. 2017

ISSN: 1941-6679-On-line Copy

RESEARCH METHODOLOGY

The current research undertakes a case of the US housing market in order to identify the trends and factors affecting the ***housing affordability*** in the US. The housing market in the US demonstrates the economic strength of potential buyers in the market, particularly after the global recession of 2008, which warrants an independent study on the factors affecting the demand and prices of housing. For this purpose, the data of 64,535people , measuring their affordability of housing units along with the measurement of economic variables such as cost burdens of households, income and market rent was extracted from the US Housing Database (Department of Housing and Urban Developmen, 2013). The affordability relates cost of a housing unit to safe housing, median income earned and basic amenities with respect to the location. Categorical variables for two types of structural units available for rent/sale divided into four census regions were provided. Other types of fees included condominium fees, or any added expenses. The software used for analysis are R programming, Minitab, SAS, IBM Watson Analytics and IBM SPSS. Among all the available software available for the analysis these software were selected because they are easily available and easy to use and do not require extensive knowledge of coding. Also these softwares are more compatible with Windows operating system as compared to other software such as Hive and MapReduce.

The data sets for IBM Watson also include the data from Twitter, which was used for the sentimental analysis. IBM Watson allows the user to use the hash tags(#) related to the topic and extract data from twitter whenever the hash tag has been used in tweets for the given time period. Depending upon whether the hash tag has been used to make positive statements or negative statements IBM Watson perform the sentimental analysis. It is also able to extract the data from other social media platforms such as Facebook and Youtube.

ANALYSIS PROCESS:

The main of the current research to compare different software for big data analytics. For the current research the comparison has been done on the basis of different parameters such as size of data handle, types of data, loading of data, predictive capabilities, users interface , add ons, presentation of results, review of data etc. Since all of the mentioned parameters can be performed in each software by the user themselves these parameters has been selected. Different software can be better from another in some parameters and some software can have advantage over other in different conditions. For example some software can handle huge set of data but it may not have good graphical representation. So on the basis of these

parameters this research compares the selected software and suggests the most efficient software for big data analysis**.**

ANALYSIS

The analysis presents the factors specifically affecting the housing affordability in the US divided as per four regions. The dimensions on the basis of which the data was analysed and compared are: size of data, types of data, user interface, additional tools, results presentation, and data loading. These dimensions were identified from four important studies related to assessment and comparison of big data applications: Wimmer & Powell (2015); Mujawar & Joshi (2015); and IBM (2014). Based on the given data, the researcher has tried to use 5 popular statistics software in order to find scalability in buying or renting a house in the US and the analysis is presented as follows:

R PROGRAMMING

The R software is highly useful in importing and analyzing big data sets and is accessible for both structured and unstructured data. Initially, the data was given for 99 variables and 64,535 observations, so the first step in research study involved reducing the observations primarily to obtain the principal components affecting the housing affordability in the region. The researcher has reduced the variables using Principal Component Analysis (PCA) for identifying a linear combination of a set of variables that has maximum variance and removing its effect. Further decision tree model was developed to understand the factors enhancing the cost burden of customers. Also, regression is conducted and a polynomial regression model is developed to study the coefficient of determination R square. To streamline the procedure, caret package was used to simplify classification problem. Further, glmnet and tree were also installed to form the regression model and decision tree respectively.

The data was first reduced to 45 variables and on further reduction, 5 independent factors identified were Total Salary (TOTALSAL), Current Market Value (VALUE), Number of Units in the building (NUNITS),), Relative Cost w.r.t Market Rent (COSTMedRELAMIPCT) because the eigen values showed significant results for these variables. The dependent variable in the research is Median Cost (CostMed).

©International Information Management Association, Inc. 2017

ISSN: 1941-6679-On-line Copy

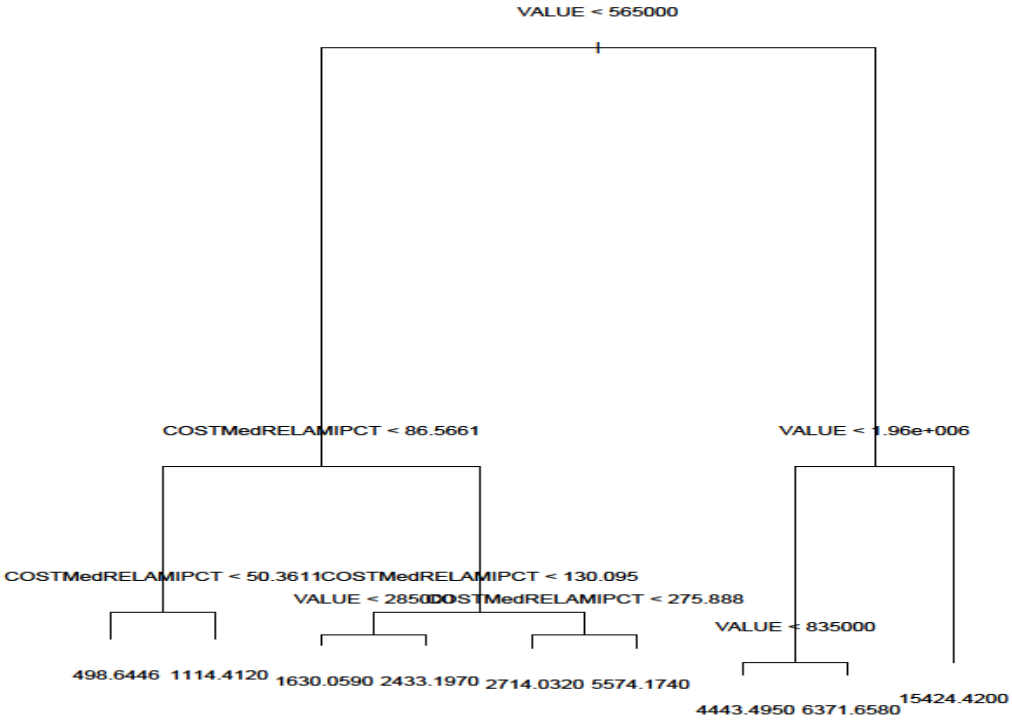

Figure 1: Decision Tree showing different nodes and factors affecting housing affordability

The decision tree is plotted on and the nodes illustrate a significant relationship between the current market value and other costs affecting housing units. The node explains that the price to rent ratio, a ratio which helps a consumer decide whether to buy or rent a property, is related to the current market price and other factors do not display highly significant results. The Multiple Linear Regression model plotted is:

COSTMED = -73.77 + 0.000568 TOTALSAL + 0.005122 VALUE
+ 1.8449 NUNITS + 300.66 COSTMedRELFMRCAT

The variable VALUE, current market value of the uni when tested at 5% significance level showed insignificant results with a value of 0.534 and also, COSTMedRELFMRCAT also had a significant value of 0.171, displaying insignificant results. All the remaining factors NUNITS and TOTALSAL also displayed insignificant results at 5% significance level..

Further, R square shows a value of 0.85 indicating that model is a good fit and also Q-Q and plotting of residuals against leverage and fitted values give idea about normality of the data.

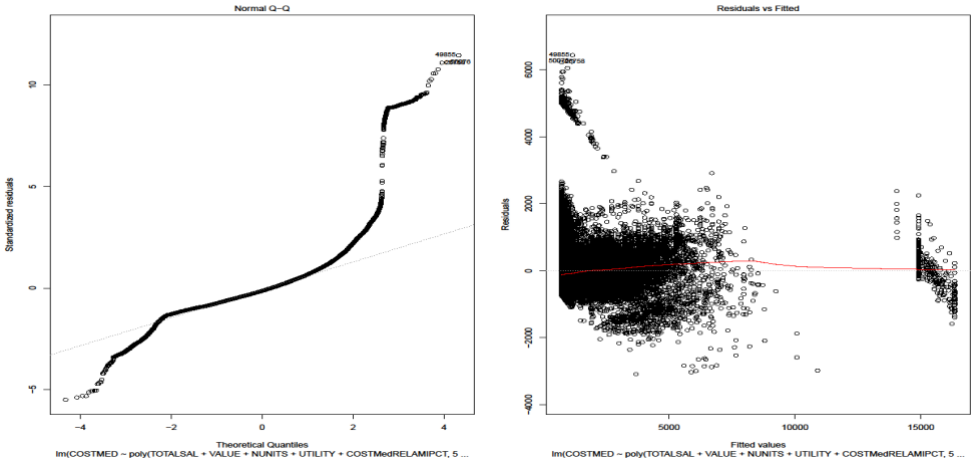

Figure 2: Q-Q Plot displaying normality Residual VS Fitted Plot

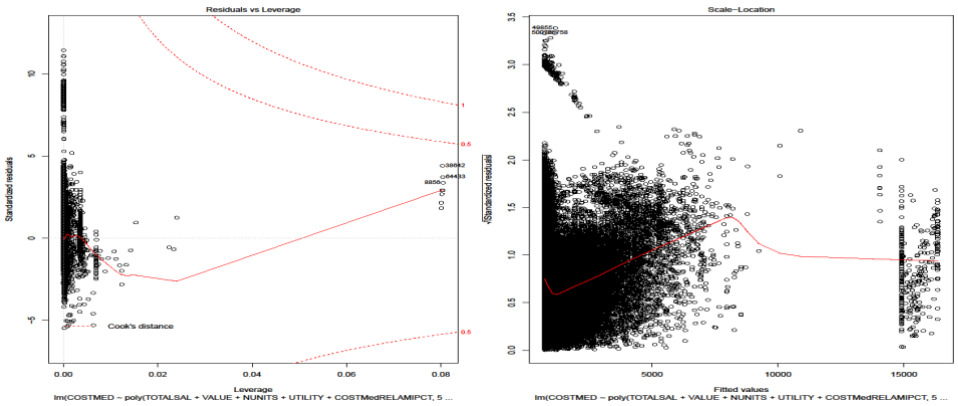

Figure 3: Residuals VS Leverage Plot Scale-Location Plot

Similarly figure 3 shows the residuals vs leverage plot scale or the location plot. Residuals show the gap between the observed value (of dependent variable) and the predicted value. The smaller the residuals better the regression model. One the other hand the leverage is the point which is far from the mean value of the variable. The residuals vs leverage plots scale helps to analyze the impact of outliers if any on the regression model. In some case the outliers do not influence the regression results so including and excluding those values do not make any difference in the regression model. However in some cases where the outliers are very far from the mean value they can influence the regression model.

MINITAB SOFTWARE

Majorly, Minitab is a tool for statistical use and is highly used in six sigma and quality improvement of rugged dataset. The dimensionality reduction technique used in Minitab software was Principal Component Analysis (PCA) and a Multiple Linear Regression model was fitted to see the dependency of the dependent variables on the independent variables.

The PCA Technique using eigen values ascertained 9 independent component impacting housing cost at median interest (COSTMED) and the factors that were identified were: Age of the head of Household (AGE1), Region Specific location (METRO3), Census Region (REGION), Area Median Income (LMED), Fair Market Rent (FMR), Low Income Limit (L80). Poverty Income (IPOV), Number of Units in the bedroom (BEDRMS) and Year Unit was constructed (BUILT). The regression model fitted is:

COSTMED = -10061 + 4.861 AGE1 + 48.58 METRO3 + 8.60 REGION
- 0.001327 LMED + 1.1800 FMR + 0.02868 L80 - 0.04111 IPOV
+ 222.00 BEDRMS + 4.355 BUILT

At 5% LOS, all the factors are significant with p-value 0.000 except LMED, which has a significant value of 0.044. The remaining factors have an impact on the dependent variable COSTMED and displayed a significant value at 5% LOS.

Further the adjusted R square showed a value of 0.24 indicating that the regression model fitted is not a good fit and the independent factors selected do not have a significant impact on the dependent factor.

©International Information Management Association, Inc. 2017

ISSN: 1941-6679-On-line Copy

IBM WATSON ANALYTICS

IBM Watson analytics software combines artificial intelligence and sophisticated human knowledge to give out results visually for large dataset. The software has helped in producing a visualization of the analytical factors displaying impact on the dependent factor. Moreover, it has also accounted for sentiments data from Twitter and other news agency about the housing affordability in US. Through social media analysis, it was observed that 68% males accounted in buying houses or renting. Further, 50% showed positive sentiments who own a house whereas 20% people who did not have a house of their own showed negative sentiments.

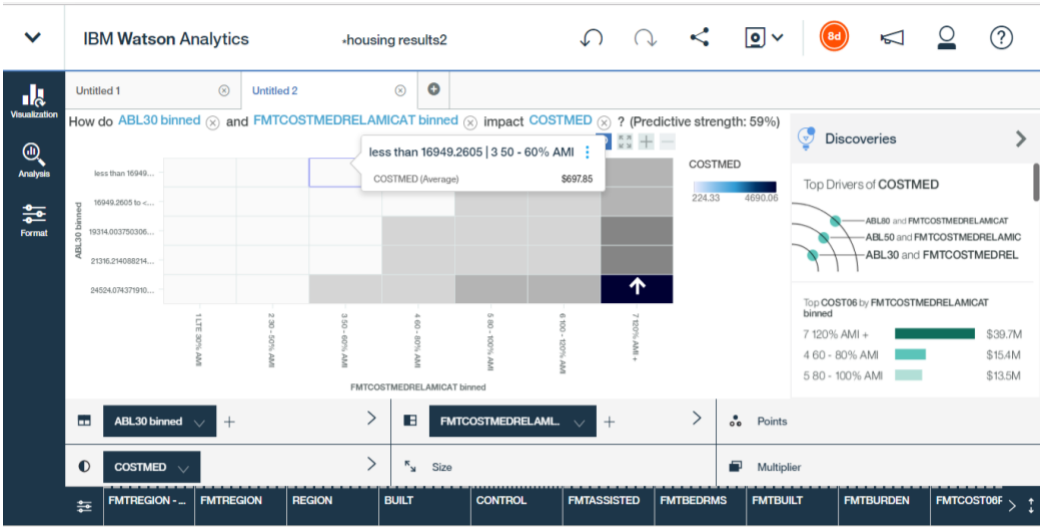

Figure 4: Social Media Analysis using IBM Watson

Now, after running the analysis under decision tree model, the researcher found out that COSTMED was majorly dependent on the factors: Very Low Income Adjusted for Number of Bedrooms (ABL50), Cost Relative to Median Income (FMTCOSTMEDRELAMICAT), Housing Cost at 6% Interest (COST06), Cost12 Relative to Median Income (FMTCOST12RELFMRCAT) and Monthly housing Costs (ZSMHC) which together showed a value of 0.99. Other important factors such as Extremely Low Income Adjusted for Number of Bedrooms (ABL30), Low Income Adjusted for number of bedrooms (ABL80), and Median Income Adjusted for Number of bedrooms (ABLMED) together showed strength of 59% in impacting the dependent variable.

©International Information Management Association, Inc. 2017

ISSN: 1941-6679-On-line Copy

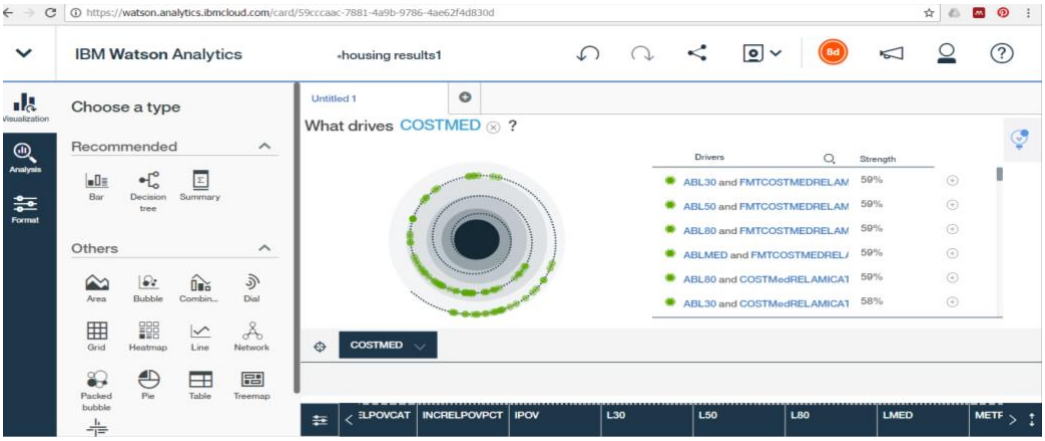

Figure 5: Independent factors affecting the dependent variable

SAS

SAS holds a dominant market position in advanced analytics and business intelligence for data management in different sectors (Kadre and Konasani, 2015). SAS software has also been used to determine the effect of several independent variables on the market share value of housing. Principal component analysis has been followed by Partial least square regression analysis for analysis.

Initially principal component analysis has been taken to reduce the number of variables up to a small variable set where the variables which explain variation in the value of housing affordability have been identified based on their Eigen values. Major variables which have Eigen value>1 and are considered to proceed to partial least square regression analysis are - Age1 (Age of the headperson of household), LMED (Area Median Income), FMR (Fair Market Rent), L30 (Extremely low income), L50 (very low income), L80 (Low Income Limit), IPOV (Poverty Income), BEDRMS (Number of Units in the bedroom), and BUILT (year of construction). Variation explained by all the variables in housing affordability can be seen in figure 6.

©International Information Management Association, Inc. 2017

ISSN: 1941-6679-On-line Copy

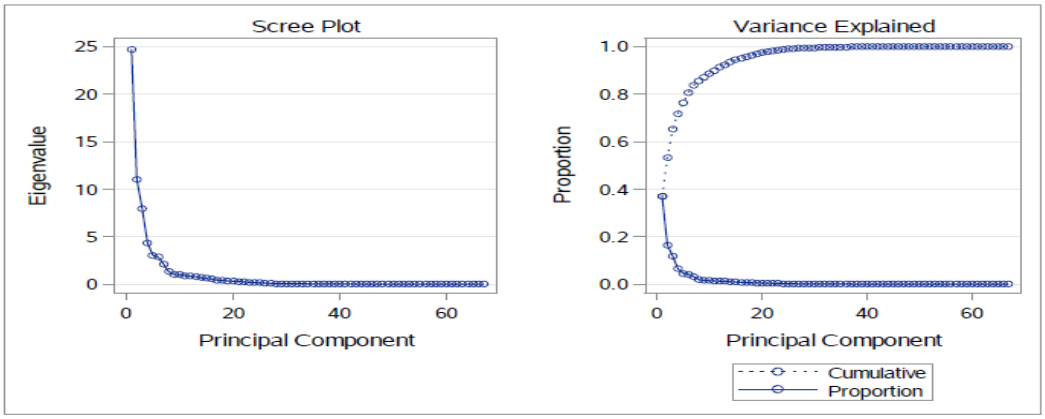

Figure 6: Variation explained by dependent variables in the model

Further, Partial least square analysis has been used, which has generalized and combined PCA factors and regression, to analyze dependent variable market share from 9 major derived independent variables of PCA. It has been found that only 24% of the response variation is explained by Age1 and Area Median Income, but only these two variables have explained 62% variation in the overall model. After running PLS, two factors have been identified based on their variation in housing affordability, where first factor contains Age1 and Area Median Income 2 variables out of 9 and the second factor contain other 7 remaining variables. Figure 7 has depicted the variance explained by both factors.

©International Information Management Association, Inc. 2017

ISSN: 1941-6679-On-line Copy

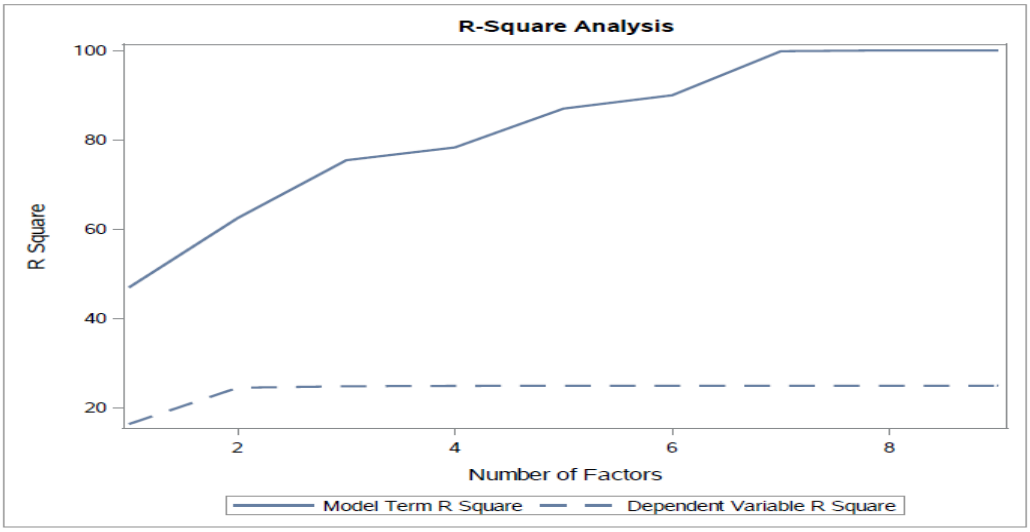

Figure7: Variance explained by both factors of the model

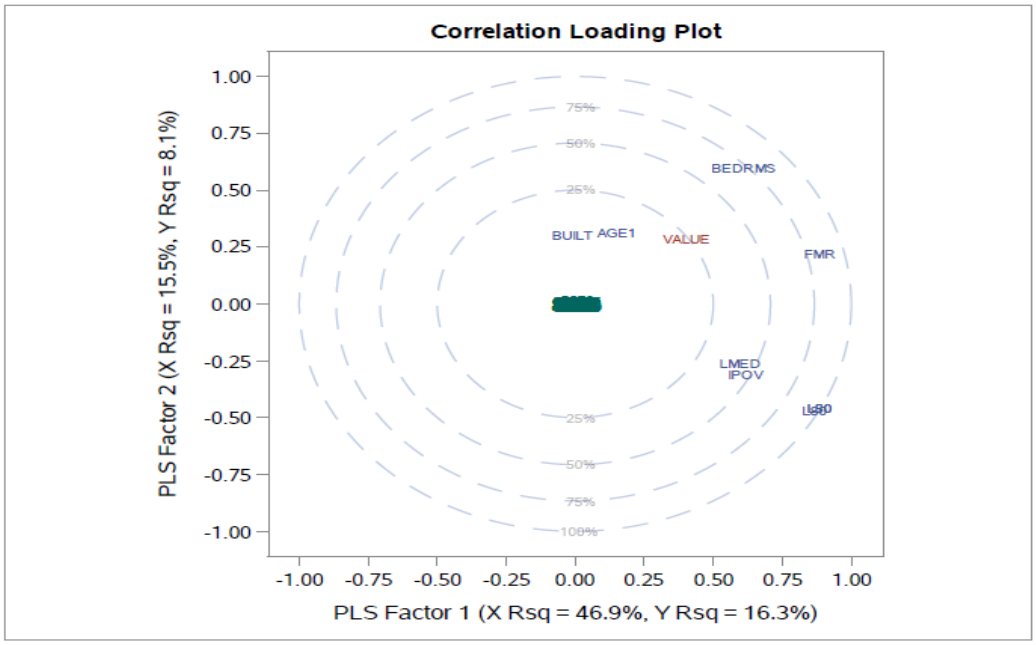

Figure 8: Correlation loading plot of Age1 and income with other 7 dependent variables

From figure 8, it has been seen that factor 1 (Age) explains more variation in the model than factor 2 (Area media income). This also signifies Age of the householder and Area median income majorly impact housing affordability system more than any other variables. On the basis of the results it can be said that the change in the age of owner of the house and the median income will significantly change the housing demand.

The Housing Affordability Data considered in the current paper explains the burden of buying/renting a house amidst different factors, there are existing studies on the topic (Housing Alliance of Pennsylvania, 2012; Nystrom and Zaidi, 2013) which allowed the researcher to understand the topic deeply and analyze it using five types of analytics software.

SPSS

The last software which was taken into consideration for the current research was SPSS, statistical software developed by IBM. However SPSS was not able to handle the data used for this analysis as it can accept only a limited number of rows unless external memory is inserted for further analysis. One of the major reasons behind IBM introducing new software such as IBM Watson analytics (discussed above) is because its previous software such as SPSS is not able to handle big data with given computer specification (IBM, n.d.). To analyze huge amount of data the extra memory needs to be installed in the computer. For this research such huge memory was not available so it was not possible to conduct the analysis in SPSS.

©International Information Management Association, Inc. 2017

ISSN: 1941-6679-On-line Copy

COMPARISON OF SOFTWARE

Parameters	R	Minitab	IBM Watson	SAS	SPSS
Size of Data	Any large dataset	10,000,000 rows	Any large datset	Any large datset	Not large dataset (except if external memory inserted)
Type of Data	Structured, Unstructured and complex dataset	Qualitative, Quantitative and Negative dataset	Structured and Unstructured dataset	Structured and Unstructured dataset	Structured numerical dataset
Data Loading	Easy Access	Easy Access	Easy Access	Easy Access	N/A
Prediction Capabilities	NA	NA	NA	NA	NA
User Interface	Coding Knowledge	Easy to use	Easy to use	Coding Knowledge	Easy to use
Additional Tools	Yes	Yes	Yes	Yes	Yes
Results Presentation	Yes	Yes	No	Yes	No
Independent installation	Yes	yes	no	No	Not easily
Exporting output	Yes	Yes	Not easy	Yes	yes
Review data	Yes	No	Yes	Yes	No

All the five software used for the current research has been compared on the basis of certain parameters. The selected parameters cover all the aspects of the software starting from the size of the data, user interface to independent

installation of other packages in the software. Another parameter used for the comparison is ***presentation of results*** which is one of the most important parameters. Even if the analysis performed is complex but the presentation of results should be easy for the users to understand and explain it to other. The current study shows that both the IBM software (IBM Watson and SPSS) lack proper results presentation as compared to other software. In case of ***independent installation***, for example in R there are various new packages which are published and updated regularly and any user can easily update those packages and use them. However on the other hand such updated packages are not available for SPSS. Similarly it is easy to ***export output*** from all the selected software except IBM Watson. For example one can easily export the results from R to the word document or other required place which is not the case with IBM Watson. The last parameter used for comparison is ***review of data***, which is done to ensure that there are no measure errors in the collected data. It also includes the process of verifying the qualification of the data.

CONCLUSION

The quantitative research conducted on housing affordability respondents using different types of software estimated and predicted the factors impacting buying of the houses. Wimmer & Powell (2015) stated in their research that open source tools have effectively increased the demand of big data analytics and provides a competitive advantage in the times of rapid data generation and technological advancement. The predictable factors have shown that polynomial regression model used for different software explained their variation (impact of each variable on the dependent variable) adequately and models were eliminated accordingly. The number of variables in the model was more than 90. However for the research to find the best fit regression model only the variable which affects the demand of the housing in US has to be included. So using the polynomial regression and principal factor analysis the number of variable has been reduced so that only the variables which has significant contribution to change in the dependent variable has been included.SAS holds the maximum market leadership in commercial space, but the advanced version is paid, hence becomes an expensive option to operate statistical functions. Majorly, the prediction of regression models on the basis of factors identified in SAS has helped the researcher provide a scope for further detailed analysis in the future. SAS also posseses good data handling capabilities with its stable Graphical user interface. It is also better than other similar software in terms of deploying visual analytic and providing data warehousing service. However SAS is not the only software which can be used for big data analysis. There are many other software

©International Information Management Association, Inc. 2017

ISSN: 1941-6679-On-line Copy

which has been introduced such as Hive, Tableau, Spark etc. Moreover, customization of software has become one of the most popular techniques for data analysis which allows the user to modify the software as per the requirements. A major improvement required is that different model in the study would have given a better result in comparison to the existing results. For the current study only the predictive analysis technique has been used to compare the software. However the comparison results could have been different if some other techniques were used as a basis of comparison.

SCOPE FOR FURTHER RESEARCH

The current research has only included five different software for big data analysis. However further studies can be conducted including more software can be included for the comparison purpose. Furthermore apart from the predictive analysis the software can be compared using other techniques (using unsupervised learning tests) which are used in big data analysis. Also different sets of data can be used for further research.

©International Information Management Association, Inc. 2017

ISSN: 1941-6679-On-line Copy

REFERENCES

Belle, A., Thiagarajan, R., Soroushmehr, S. M. R., Navidi, F., Beard, D. A. and Najarian, K. (2015) 'Big Data Analytics in Healthcare.', *BioMed research international*. Hindawi Publishing Corporation, 2015, p. 370194. doi: 10.1155/2015/370194.

Brown, B., Chui, M. and Manyika, J. (2011) 'Are you ready for the era of "big data"?', *McKinsey Quarterly*, 4(October), pp. 24–35. doi: 00475394.

Cao, M., Chychyla, R. and Stewart, T. (2015) 'Big Data Analytics in Financial Statement Audits', *Accounting Horizons*, 29(2), pp. 423–49.

Catlett, C. (2013) *Cloud computing and big data*. IOS Press.

Chen, H., Chiang, R. and Storey, V. (2012) 'Business intelligence and analytics: From big data to big impact', *MIS quarterly*, 36(4), pp. 1165–1188.

Chen, P. C. L. and Zhang, C.-Y. (2014) 'Data-intensive applications, challenges, techniques and technologies: A survey on Big Data', *Information Sciences*, 275, pp. 314–347. doi: 10.1016/j.ins.2014.01.015.

Chen, Y., Alspaugh, S. and Katz, R. (2012) 'Interactive analytical processing in big data systems', *Proceedings of the VLDB Endowment*. VLDB Endowment, 5(12), pp. 1802–1813. doi: 10.14778/2367502.2367519.

Davenport, T. H., Barth, P. and Bean, R. (no date) 'How "Big Data" Is Different'.

Department of Housing and Urban Developmen (2013) *AMERICAN HOUSING SURVEY: HOUSING AFFORDABILITY DATA SYSTEM, Office of Policy Development and Research (PD&R)*.

Eddy, N. (2015) *Cloud Computing Reducing Costs, Improving Productivity, EWeek*. Available at: http://www.eweek.com/small-business/cloud-computing-reducing-costs-improving-productivity.

Forbes (2016) 'Roundup Of Analytics, Big Data & BI Forecasts And Market Estimates, 2016'.

Gantz, J. and Reinsel, D. (2012) 'The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east', *IDC Analyze the future*.

©International Information Management Association, Inc. 2017 ISSN: 1941-6679-On-line Copy

George, G., Haas, M. R. and Pentland, A. (2014) 'Big Data and Management', *Academy of Management Journal*. Academy of Management, 57(2), pp. 321–326. doi: 10.5465/amj.2014.4002.

Housing Alliance of Pennsylvania (2012) *A NEW VISION for Housing Market Recovery: What the Data Tells Us About What Works*. Glenside. Available at: https://housingalliancepa.org/wp-content/uploads/2012/03/HAPAAlleghCoBookFINAL.pdf.

Hu, H., Wen, Y., Chua, T.-S. and Li, X. (2014) 'Toward Scalable Systems for Big Data Analytics: A Technology Tutorial', *IEEE Access*, 2, pp. 652–687. doi: 10.1109/ACCESS.2014.2332453.

IBM (2014) *Performance and Capacity Implications for Big Data*.

IBM (no date) *Maximum number of cases in an SPSS dataset, IBM*. Available at: http://www-01.ibm.com/support/docview.wss?uid=swg21476061.

Jin, X., Wah, B. W., Cheng, X. and Wang, Y. (2015) 'Significance and Challenges of Big Data Research', *Big Data Research*, 2(2), pp. 59–64. doi: 10.1016/j.bdr.2015.01.006.

Kadre, S. and Konasani, V. R. (2015) *Practical Business Analytics Using SAS: A Hands-on Guide*. APress.

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, W. K. M., Alam, M., Shiraz, M. and Gani, A. (2014) 'Big data: survey, technologies, opportunities, and challenges.', *TheScientificWorldJournal*. Hindawi Publishing Corporation, 2014, p. 712826. doi: 10.1155/2014/712826.

Kwon, O., Lee, N. and Shin, B. (2014) 'Data quality management, data usage experience and acquisition intention of big data analytics', *International Journal of Information Management*, 34(3), pp. 387–294.

Lohr, S. (2012) 'The Age of Big Data', *The New York Times*, pp. 1–5. doi: 10.1126/science.1243089.

Marz, N. and Warren, J. (2015) *Big Data: Principles and best practices of scalable realtime data systems*.

Mishra, S. and Badhe, V. (2016) 'A Survey of approaches for mining large data sets', *International Journal Of Engineering And Computer Science*, 5(5), pp. 16354–16358.

Mujawar, S. and Joshi, A. (2015) 'Data Analytics Types, Tools and their Comparison', *International Journal of Advanced Research in Computer and Communication Engineering*, 4(2).

Murray, D. (2014) *Knowledge Machines: Language and Information in a Technological Society*. Oxon: Routledge.

Nystrom, S. and Zaidi, Al. (2013) *THE ECONOMIC, DEMOGRAPHIC, AND CLIMATE IMPACT OF ENVIRONMENTAL TAX REFORM IN WASHINGTON AND KING COUNTY*. Washington. Available at: http://www.cbuilding.org/sites/default/files/etr-wa-remi-dec-13-2013.pdf.

Oracle (2013) 'Big Data Analytics', (March).

Piotrowicz, W. and Cuthbertson, R. (2014) 'Introduction to the Special Issue Information Technology in Retail: Toward Omnichannel Retailing', *International Journal of Electronic Commerce*, 18(4), pp. 5–16. doi: 10.2753/JEC1086-4415180400.

Russom, P. (2011) *BIG DATA ANALYTICS*.

Sagiroglu, S. and Sinanc, D. (2013) 'Big data: A review', in *2013 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, pp. 42–47. doi: 10.1109/CTS.2013.6567202.

Tan, K. H., Zhan, Y., Ji, G., Ye, F. and Chang, C. (2015) 'Harvesting big data to enhance supply chain innovation capabilities: An analytic infrastructure based on deduction graph', *International Journal of Production Economics*, 165, pp. 223–233. doi: 10.1016/j.ijpe.2014.12.034.

Varian, H. R. (2014) 'Big Data: New Tricks for Econometrics', *Journal of Economic Perspectives*, 28(2), pp. 3–28. doi: 10.1257/jep.28.2.3.

VCloudNews (2015) *EVERY DAY BIG DATA STATISTICS – 2.5 QUINTILLION BYTES OF DATA CREATED DAILY, V Cloud News*. Available at: http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/.

©International Information Management Association, Inc. 2017 ISSN: 1941-6679-On-line Copy

Villars, R. L. and Olofson, C. W. (2014) *WHITE P APER Big Data : What It Is and Why You Should Care INFORMATION EVERYWHE RE , BUT WHERE'S THE KNOWLEDGE ?*

Wimmer, H. and Powell, L. (2015) 'A comparison of open source tools for data science', in *Proceedings of the Conference on Information Systems Applied Research*. Wilmington: (Information Systems & Computing Academic Professionals.

Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding (2014) 'Data mining with big data', *IEEE Transactions on Knowledge and Data Engineering*, 26(1), pp. 97–107. doi: 10.1109/TKDE.2013.109.