PREDICTIVE ANALYTICS AS APPLIED TO BIG DATA: CONSIDERATIONS FOR ANALYTICS TEAMS

J. OGUNLEYE, PhD
Professor at Middlesex University and the Editor of the International Journal of Developments in Big Data and Analytic

Middlesex University, United Kingdom
E-mail: oj08@hotmail.com

Abstract. The phenomenon of big data has brought home the importance of predictive analytics as a technology and statistical technique critical to taking the sting out of the big data mayhem. Although predictive analytics has been around for some time, the benefits of predictive analytics have only recently been appreciated due largely to the phenomenon of big data. This new-found appreciation of predictive analytics is coupled with a desire by many corporate organisations not only to inform strategic business decisions with evidence, but also to predict future trends with a high level of confidence. While many organisations are able to use predictive analytics technology with greater success, the outcome for some organisations has been less than successful. This paper argues that predictive analytics can only achieve so much when organisations and their analytics teams consider the main limiting factors in predictive analytics projects. The paper suggests a number of measures that analytics teams can take to minimise the main limitations of predictive analytics. The paper concludes that although some machine learning algorithms based on artificial intelligence are increasingly being used to minimise aspects of limitations of predictive analytics but with ever present danger of lurking variables or unknown factors, there is no sustainable alternative to good data quality assurance.

Keywords: Predictive analytics, data quality, limitations, model and modelling, return on investment, legal and ethics.

Introduction-predictive analytics as applied to Big Data. In the recent years, predictive analytics has been transformed by the phenomenon of big data, the innovation that surrounds the use of digital information. These datasets — 85% of which are non metric data or unstructured (SAS Institute, 2012) — are huge and complex in volume, velocity, variety, veracity and variability they are significantly beyond the capability of standard data processing and analytic tools, and even threatens traditional computing architectures (Ogunleye, 2014).

Predictive analytics can be conceptualised as both an analytical process and technology. OPCC (2012, p.3) conceptualise predictive analytics as a ‘general purpose analytical process that enables organisations to identify patterns in data that can be used to make predictions of various outcomes, not all of which have an impact on

Библиотека БГУИР

individuals.' But predictive analytics is more than an analytical process: it combines human skills and capability with technology such as machine learning of patterns in current and historical data and the application of algorithms not only to identify patterns in the data but also to forecast future probabilities of the outcome of those patterns (Ogunleye, 2014). According to Miller (2014, p.2):

> Predictive analytics, like much of statistics, involves searching for meaningful relationships among variables and representing those relationships in models. There are response variables – things we are trying to predict. There are explanatory variables or predictors – things we observe, manipulate, or control that could relate to the response.

It is these generally accepted conceptions of predictive analytics that have given organisations confidence to deploy and, for some organisations, embed predictive analytics in functional and operational decision making (Accenture, 2013) to ensure that decisions are based on hard evidence and to achieve high level of confidence in prediction. In other words, these widely accepted conceptions of predictive analytics have enabled data-driven organisations to take the sting out of the big data mayhem as well as strengthen their ability to 'generate better decisions, greater consistency, and lower costs' (CGI, 2013, p. 2).

Predictive analytics technology is therefore critical to sense- and meaning - making of Big Data, as predictive analytics not only 'makes it possible to harness the power of big data' (Heitmueller, et al. 2014, p.1523) thereby leveraging organisation data assets, but also critical to translating Big Data into 'meaningful, useable business information' (Abbott, 2014).

What is clear from the foregoing discourse is that people, tools and algorithms are critical in any predictive analytics project as shown in figure 1.

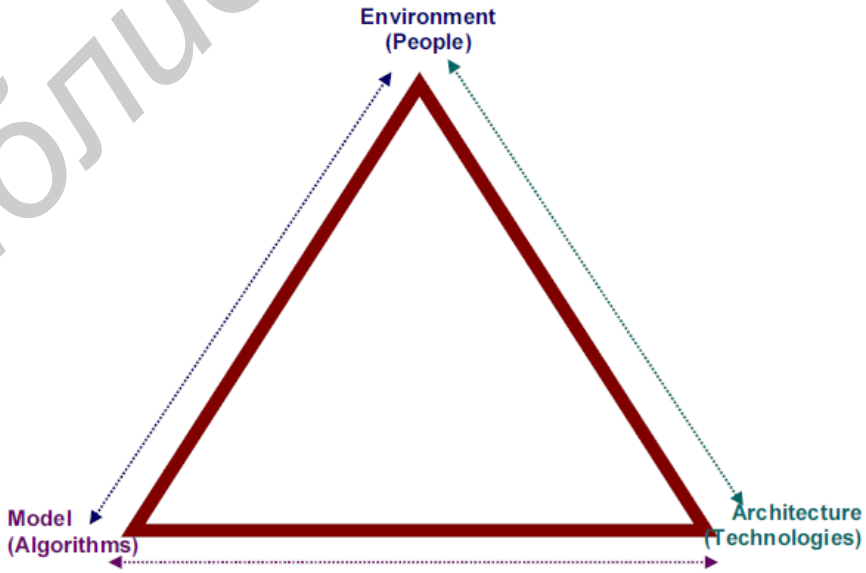

Fig.1. Triangular connection of people, model and tools in predictive analytics

Библиотека БГУИР

Considerations for Analytics Teams. Like some technologies or statistical techniques, predictive analytics has its limiting factors. However, predictive analytics as conceptualised above can only be limited by errors of omission or commission on the part of businesses, users of predictive analytics and their analytics teams. Since so much depends on human input (figure 2) in making any predictive analytics project successful, the likelihood of human errors cannot be over stated. Thus, a major consideration in the deployment of any predictive analytics project is subject knowledge or the extent to which an organisation is familiar with the concepts of predictive analytics. In other words, before the project is launched or embarked upon, businesses and their analytics team need to have a deeper understanding of the concepts of predictive analytics noting in particular what is involved in operationalising predictive analytics and how people, tools and algorithms are connected, issues in predictive analytics, and the application of predictive analytics to big data (see Ogunleye, 2014). The following paragraphs highlight some of the practical considerations for analytics teams.

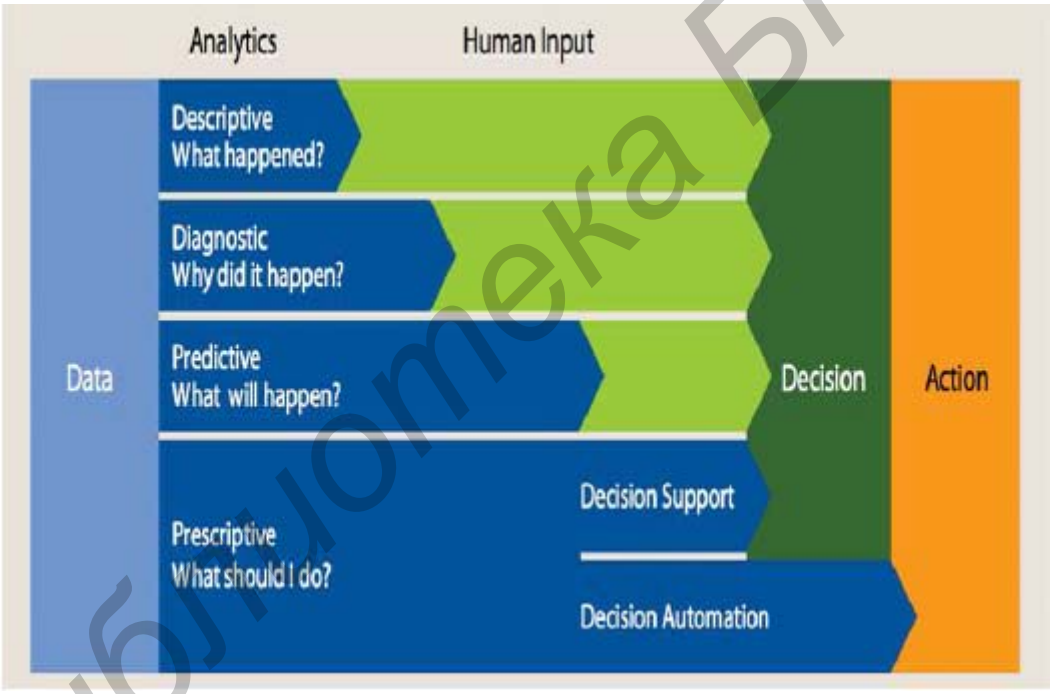

Fig.2. Analytics and human input (Source: Gartner, NG00254653, September 2013 in FICO, 2013, p. 3)

Quality of data. A significant issue in predictive analytics relates to the quality of the data sets. As a statistical technique, data is central to predictive analytics and it is important that organisations their analytics teams have a deeper knowledge of the quality of the available data before investing in predictive analytics technology (see also McCue, 2007).

A recent study by Andreescu, et al. (2014, p.15) has underlined the significant of good quality of data in analytics projects. Andreescu, et al found that poor data could 'cause serious consequences for the efficiency of organizations' and that poor attention

Библиотека БГУИР

to 'quality issue could potentially lead to erroneous data mining and analysis results which in turn could lead to severe consequences, financial or otherwise.' To prevent the likelihood of poor data being used in predictive analytics, Ogunleye (2014) points out that it is important that analytics teams secure the integrity of data and super secure in their judgement that the data is free of bias or that bias has been corrected and everything else that goes with operationalising predictive analytics – including (predictive) model development and implementation, monitoring, calibration and re-calibration. This process of assuring and measuring data quality involves data preparation, cleaning and formatting—all of which are essential for data mining, 'the process of discovering interesting patterns and knowledge from large amounts of data' (Han, et al. 2011, p6).

Two issues are also worth mentioning. First, lack of historic data as a limitation can be minimised by recreating the data 'back in time' as Jain (2015) has argued. Second, it is possible to use artificial intelligence based on machine learning algorithms to minimise the impact of data quality on the outcome of predictive analytics - for example, to reduce noise, correct errors and address biases especially when data are mixed and contain thousands of independent variables. The only proviso is that, with the ever present danger of the so-called lurking variables or unknown factors, good algorithms might not be a sustainable alternative to good data quality assurance.

Model and modelling. Model refers to a 'representation of the world, a rendering or description of reality, an attempt to relate one set of variables to another' (Miller, 2014, p. 2). Modelling, therefore, is a mathematical representation of an entity and very important in any predictive analytics project. According to Dickey (2012):

> Predictive modelling is a name given to a collection of mathematical techniques having in common the goal of finding a mathematical relationship between a target, response, or "dependent" variable and various predictor or "independent" variables with the goal in mind of measuring future values of those predictors and inserting them into the mathematical relationship to predict future values of the target variable.

As modelling is important in predictive analytics, it can also be a major limitation if the modelling process is not well understood. As Taylor (2012) explains, the modelling process has to be 'repeatable, industrial-scale' to ensure effective development of 'dozens or even thousands' of required predictive analytic models – in order to search for 'meaningful relationship among models and representing those relationships in models' (Miller, 2014, p.2). So, whatever the type of predictive models is deployed – be it regression or classification – an important issue for consideration is the level of user discretion that is considered acceptable. User discretion, judgement and experience (or lack of them) will have impact on the outcome of predictive analytics. Thus, without an effective model life-cycle management build into the production system or environment, predictive analytics project will not archive desire results (see also Chu, et el., 2007). Similarly, a predictive modelling based on an out-

Библиотека БГУИР

of-date or incorrect data might 'wrongfully skew' analytics results (Viswanathan, 2013) or leads to wrong conclusions (Heger, 2014).

In deploying modelling in predictive analytics, there is, sometimes, a lack of a clear understanding of the difference between 'prediction' and 'projection' and how the two terms compare. Although both 'projection' and 'prediction' have the term 'forecast' in their meaning, but an understanding of how the two terms compare is very important. Prediction is about the predicting future trends and projection is about projecting (forecasting) events. Analytics teams might run into trouble when projections are used to predict future trends when the underlying assumptions, implicit and explicit assumptions, of their models are not constant or, at best, susceptible to seasonality trap. In other words, when analytics teams are not clear about how the two terms compare, they might not be able to guarantee the stability in the phenomenon to be predicted (see also Elkan, 2013).

Additionally, there is a general tendency to equate correlation with causation. As Viswanathan (2013) argues there is evidence to 'mistake' correlation for causation in predictive analytics especially if the latter is conceptualised as in Miller (2014, p.2), involving 'searching for meaningful relationships among variables and representing those relationships in models'. Coefficient of Correlation highlights the linearity of relationships between variables in the model or data item, but implies little about the nature of those relationships. Thus, in making judgement about how the two terms—correlation and causation—compare in predictive analytics results, the nature of the organisation and the phenomenon to which the statistical technique applies are a significant consideration. As Huang (2013, p. 1) argues:

> Genetics and molecular biology have historically been blessed with simple cases of unidirectional, linear causality, which have taught us a great deal about gene function, but also stifled the intellectual embrace of mutual causation. .. Thus, unintentionally, equating correlation with causation is warranted in complex, networked systems where positive feedback loops are a characteristic feature and entail mutual causation. Understanding such relationships will help optimize approaches to disrupting the cycle – for instance, by treating certain symptoms. But it will also open the door to the deeper cause: what kicked off the causation cycle in the first place?

Return on investment. The return of investment (ROI) is an often-overlooked issue in predictive analytics projects. There is evidence that return of investment is as higher as 250% in predictive analytics projects, compare to the 89% return of investment of projects that focused solely on accessing information and seeking internal gains in productivity, according to a survey by the International Data Corporation (Vesset and Harries, 2011). There is also evidence that many organisations deploy predictive analytics projects with little or cognisance of the return on investment and those organisations that did have 'struggled to see a meaningful' ROI (Accenture, 2013).

Historically, the return of investment is conceptualised in financial term, used as a critical performance evaluation tool for financial returns and costs. However, in the

Библиотека БГУИР

recent decades, the concept of return of investment has been extended to and applied in a range of contexts including information system especially when organisations are making decisions about acquisitions of intellectual property or software (Botchkarev and Andru, 2011).

Operationalising predictive analytics involves a sizeable amount of investments in people, tools and technology. An earlier survey by the IDC (Harries, 2003) found that predictive analytics projects required significant higher levels of investment. Even though the IDC survey took place over a decade ago, the levels of investment required for predictive analytics project remain high (Accenture, 2013). Take human resources, for instance, investment begins well before the initial model-build and the continuous maintenance phases. This investment in human resources is particularly justifiable because of the rate at which models deteriorate over time (Jain, 2015).

With regards to the methodological approach to or the process of calculating return of investment, there is no universally agreed formula or method for calculating. A lot depends on the individual organisation and the context in which an organisation deploys predictive analytics. However, it is generally accepted that calculating ROI will compare technology and labour costs with before- business results and after-business results. Additionally, as McCann (2014) explains:

…Many companies run controlled studies where, for example, a new marketing tactic suggested by the analytics is directed at a portion of an audience, with the benefit quantified by comparing it with results for a control group.

What is clear from the foregoing is that irrespective of the methodological approach, an organisation should strike a balance between the available resources and the ROI before deploying predictive analytics technology.

Legal and ethical. There are legal and ethical considerations in the deployment of predictive analytics, especially where a company operates in different jurisdictions or cultures. The way information about customers are kept and mined and the ‘extent to which data mining’s outcomes are themselves ethical’ with respect to individual customers in corporate and non organisations (Johnson, 2014; EDUCAUSEreview, 2013; Kay, et al., 2012) should conform to the highest ethical standards. According to Schwartz (2010, p.3), it is critical that an organisation ‘assess whether its decision-making with analytics reflects legal, cultural, and social norms about acceptable activities and take steps, when needed to comply with these norms.’ (See also Johnson, 2013; OPCC, 2012).

*Business case for predictive analytics.*Another consideration for organisations and their analytics teams is a perception that the case for a predictive project has to be about technologies or has to be predicated on information technology infrastructure. Any decision to deploy predictive analytics must reflect the business proposition. This is what Heger (2014, p. 47) says about any organisation considering a big data analytics project, an argument that also applies to any predictive analytics project:

> Any company considering a Big Data project has to first evaluate the business cases, specify the goals and objectives, and stipulate the outcomes of the proposed Big

Библиотека БГУИР

> Data initiatives. After the people and business impact and outcome is clearly understood, the IT capabilities and requirements can be evaluated. Developing a roadmap of how to achieve the desired business outcomes provides the organization with the understanding of what is required from a financial and organizational perspective.

It is therefore important to understand the business and people aspects of any predictive analytics project. It is important that the case for predictive analytics starts with a business problem/preposition that requires a multidisciplinary team of data scientists, statisticians, data analysts as well as individuals with risk management expertise. It is important that members of the multidisciplinary team come together and agree on how predictive analytics technology can be used to address critical business problems. A central part of the role of the analytics team is therefore to demonstrate how businesses can seamlessly and effectively align predictive analytics with IT decisions (see Boris, 2014).

Human factor. An often-overlooked issue in predictive analytics is the challenge pose by human factor. The introduction of predictive analytics technology will require attitudinal change and people in the organisation who have used to making decision based on intuition or gut feeling and who considered themselves an essential part of the existing decision making process might feel that their toes are being stepped on. These individuals have an interest to protect – which is to make sure that no machine takes over their jobs! They need to be listened to, won over and assured that the predictive analytics technology is required solely as a decision making-supporting tool for the organisation. It is therefore important that champions of predictive analytics anticipate the human relations challenges that will arise as a result of the introduction of predictive analytics technology. In other words, these champions should be mindful of concerns by those in the organisations who might have reservations about the project.

Conclusion. The phenomenon of big data occasioned by the recent explosion in digital data has underlined the significance of predictive analytics as both a technology and statistical technique critical to taking the sting of the big data mayhem. Although predictive analytics has been around for some time and has been used successfully by large companies operating in a small number of industrial sectors, it was only in the recent years that the benefits and potential of predictive analytics have been appreciated. Predictive analytics offers data-driven organisations in particular and users alike tremendous benefits— the main being the ability to base operational and functional decisions on hard facts and the ability to embed analytics across enterprise operations and functions. Thus, predictive analytics is new approach to decision making as the technology enables organisations to make real time predictions with a high degree of confidence. However, predictive analytics has its limitations. The main limitation is quality of data. Out-of-date or incorrect data can skew analytics results or produce wrong or incorrect conclusions. Although, it is possible to use artificial intelligence based on machine learning algorithms to minimise the impact of data

Библиотека БГУИР

quality on the outcome of predictive analytics, but with the ever present danger of lurking variables or unknown factors, good algorithms cannot possibly be a sustainable alternative to good data quality assurance.

References

[1]. Abbott, D. (2014) Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst, New Jersey: John Wiley & Sons.

[2]. Accenture (2013) Analytics in Action: Breakthroughs and Barriers on the Journey to ROI, online: http://www.accenture.com/sitecollectiondocuments/ pdf/accenture-analytics-in-action-survey.pdf; accessed: 28.5.15

[3]. Andreescu, I. A., Anda Belciu, A., Alexandra Florea, A. and Diaconita, V. (2014) Measuring Data Quality in Analytical Projects, Database Systems Journal, 5 (1), pp. 15-25.

[4]. Boris, Z. (2014) 'Application of Predictive Analytics for Better Alignment of Business and IT', International Conference on Knowledge, Innovation and Enterprise, Big Data Summit, 25 July, Riga, Latvia, available at: http:// www.kiecon.org/pdf%20%20Application%20of%20Predictive% 20Analytics%20for%20Better%20Alignment%20of%20Business%20and% 20IT.pdf.

[5]. Cao, C. (2012) Sports data mining technology used in basketball outcome prediction. Masters Dissertation. Dublin Institute of Technology, Ireland. http://arrow.dit.ie/cgi/viewcontent.cgi?article=1040&context=scschcomdis; accessed: 13 April 2013

[6]. Chu, R. , Duling, D. and Thompson, W. (2007) Best Practices for Managing Predictive Models in a Production Environment, SAS Global Forum 2007, Paper 076-2007, SAS Institute Inc. Available: http://www2.sas.com/ proceedings/forum2007/076-2007.pdf; accessed: 20 September 2012.

[7]. CGI (2013) Predictive analytics. The rise and value of predictive analytics in enterprise decision making, White paper, [Online], www.cgi.com; accessed: 20.4.2015

[8]. Dickey, A. D. (2012) 'Introduction to Predictive Modeling with Examples', *SAS Global Forum 2012:* Statistics and data Analysis, Paper 337-2012, SAS Institute Inc. http://support.sas.com/resources/papers/proceedings 12/3 37- 2012.pdf; accessed: 16.5.15.

[9]. EDUCAUSEreview (2013) 'Ethics, Big Data, and Analytics: A Model for Application', online: http://www.educause.edu/ero/article/ethics-big-data- and-analytics-model-application; accessed: 10.2.2014.

[10]. Elkan, C. (2013) 'Predictive analytics and data mining', Online: http:// cseweb.ucsd.edu/~elkan/255/dm.pdf; accessed: 10/3/15.

[11]. Harries, D. H. (2003) Predictive Analytics and ROI: Lessons from IDC's Financial Impact Study, International Data Corporation (IDC), Online: http:// www.analyticalinsights.com/web_images/IDC Predictiveanalyticsan- dROI.pdf; accessed: 3.5.15.

[12]. Vesset, D. and Harries, D. H. (2011) 'The Business Value of Predictive Analytics', White Paper, International Data Corporation (IDC), Online: http:// www.nexdimension.net/resources/products/ibm/spss/ibm-spss-predictive- analytics-business-value-whitepaper.pdf; accessed: 3.5.15.

[13]. Han, J., Kamber, M., and Pei, J. (2011) *Data Mining Concepts and Techniques* (Third ed). Elsevier Inc.: p.6 and 8.

[14]. Heger, D. (2014) Big Data Analytics—'Where to go from Here', *International Journal of Developments in Big Data and Analytics,* 1, 1, pp.42-58.

[15]. Heitmueller, A., Henderson, S., Warburton, W., Elmagarmid, A. Pentland , A. and Darzi, A. (2014) Developing Public Policy To Advance The Use Of Big Data In Health Care, *Health Aff (Millwood), 33, 9, pp.* 1523-1530.

[16]. Huang, S. (2013) 'When correlation and causation coincide', *BioEssays,* 36, 1, pp1-2, Online: http://onlinelibrary.wiley.com/doi/10.1002/ bies.201370003/pdf; accessed: 22.5.15.

[17]. Jain, P. (2015) 'Predictive Analytics: 4 Key Constraints', Online: http://

Библиотека БГУИР

www.aryng.com/blog/predictive-analytics-4-key-constraints/; accessed: 15.5.15.

[18]. Johnson, J. A. (2013) 'Ethics of Data Mining and Predictive Analytics in Higher Education', Association for *Institutional Research Annual Forum,* Long Beach, California, May 19-22, 2013. Available at http://ssrn.com/ abstract=2156058 or http://dx.doi.org/10.2139/ssrn.2156058; accessed: 19 June 2014.

[19]. Johnson, J. A. (2014) 'The Ethics of Big Data in Higher Education', *International Review of Information Ethics,* http://www.i-r-i-e.net/inhalt/021/IRIE- 021-Johnson.pdf; accessed: 10 September 2014.

[20]. Kay, D., Korn, N. and Oppenheim, C. (2012) Legal, Risk and Ethical Aspects of Analytics in Higher Education, JISC CETIS Analytics Series, Vol.1, No. 6 Online: http://publications.cetis.org.uk/wp-content/uploads/2012/11/ Legal-Risk-and-Ethical-Aspects-of-Analytics-in-Higher-Education-Vol1- No6.pdf; accessed: 26.5.15.

[21]. McCann, D. (2014) Predictive Analytics: How Clear Is the ROI?, CFO Magazine, Online: http://ww2.cfo.com/technology/2014/07/predictive- analytics-clear-roi/; accessed: 26.5.15.

[22]. McCue, C. (2007) Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis, Butterworth-Heinemann: Oxford, UK.

[23]. Miller, M. T. (2014) Modeling Techniques in Predictive Analytics: Business Problems and Solutions with R, Pearson Education, Inc

[24]. Ogunleye, J. (2014) The Concepts of Predictive Analytics, *International Journal of Developments in Big Data and Analytics,* 1, 1, pp. 86-94.

[25]. OPCC (2012) The Age of Predictive Analytics: From Patterns to Predictions. Report prepared by the Research Group of the Office of the Privacy Commissioner of Canada, available: https:// www.priv.gc.ca/information/research -recherche/2012/pa_201208_e.pdf; accessed: 23 July2014.

[26]. SAS Institute (2012) Big Data Meets Big Data Analytics, White Paper. SAS Institute Inc. Available:http://www.sas.com/content/dam/SAS/en_us/doc/ whitepaper1/big-data-meets-bigdata-analytics-105777.pdf; accessed: 2 January 2014.

[27]. Schwartz, M. P. (2010) Data Protection Law and the Ethical Use of Analytics, The Centre for Information Policy Leadership LLP [online] http://www.huntonfiles.com/files/webupload/CIPL_Ethical_Undperinnings_of_ Analytics_Paper.pdf); accessed: 26 January 2014.

[28]. Taylor, J. (2014) 'Three steps to put Predictive Analytics to Work, Decision Management Solutions, SaS [online] https://www.sas.com/content/dam/SAS/ en_us/doc/whitepaper2/threesteps-put-predictive-analytics-work-105837.pdf; accessed: 20 August 2014.

[29]. Viswanathan, V. (2013) 'Avoiding Pitfalls in Predictive Analytics Models', Online: http://data-informed.com/avoiding-pitfalls-in-predictive-analytics-models/; accessed: 20.11.14.

Библиотека БГУИР