AJIT-e: Online Academic Journal of Information Technology
2016 Yaz/Summer – Cilt/Vol: 7 - Sayı/Num: 24
DOI: 10.5824/1309-1581.2016.3.007.x
http://www.ajit-e.org/?menu=pages&p=details_of_article&id=216

Received : 16.03.2016 *Editorial Process Begin: 15.06.2016* *Published: 28.07.2016*

Bemo: A Parsimonious Big Data Mining Methodology

Joseph M. Woodside, *Stetson University, Department of Decision and Information Sciences, DeLand, FL, jmwoodsi@stetson.edu*

ABSTRACT *The Problem: Standardized processes are often followed to systematically conduct data mining projects. However while current models provide good descriptions, they are in need of updates given current Big Data challenges. Current data mining methods do not meet all requirements of businesses, in addition current methods are difficult to remember and do not cover all requisite steps. Given these limitations, usage of the traditional data mining process methods are fading in favor of independent data mining processes.*

What Was Done: BEMO (Business Opportunity, Exploration, Modeling, and Operationalization) is a standard parsimonious process developed for conducting data mining projects in a reusable and repeatable fashion in a Big Data environment. This model is vendor, technology, and industry agnostic. The process model is applied to a practical project example.

Why this Work is Important: This manuscript allows a reusable and simplified model for data mining that can be applied to a variety of applications given a formalized and detailed process template. Given new technologies, Big Data and other developments a new data mining methodology is required to adequately meet these needs. The contribution of a parsimonious Big Data mining model also permits utilizing simpler models over complex models that can more efficiently generalize new problems.

Keywords *:* Data Mining, Big Data, BEMO, CRISP-DM, SEMMA, EMSMA, Parsimonious

INTRODUCTION

Data Mining

Data mining is the process of exploring, analyzing, and uncovering meaningful patterns and trends by reviewing data through various mechanisms. The first International Conference on Knowledge Discovery and Data Mining (KDD) was started in 1995. Data mining is still considered a relatively new and evolving field of study and draws from statistics, mathematics, machine learning, and artificial intelligence. Data mining importance has been accelerated in recent years through the exponential growth of data as well as the decreasing cost to capture, store, and process data (Shmueli et al., 2010).

Standardized processes are often followed to systematically conduct data mining projects. In a KDNuggets survey CRISP-DM remained the top methodology when compared between 2007 and 2014 with 43% of respondents, followed by custom methodology 27.5%, and followed by SEMMA with 8.5%. The authors indicate that while CRISP-DM provide a good description, the details are in need of update since these methodologies have not been adapted

DOI: 10.5824/1309-1581.2016.3.007.x

to Big Data challenges and current data science. As further evidence the original website for the CRISP-DM methodology (crisp-dm.org) is no longer available and IBM SPSS is the main tool using the methodology. While CRISP-DM usage remained relatively flat, the other well-known methodology SEMMA has declined, while custom methodologies have increased significantly which in part is attributed to the current methodologies shortcomings (Piatetsky, 2014).

CRISP-DM

The CRoss Industry Standard Process for Data Mining (CRISP-DM) model is intended to allow data mining projects to be more reliable, cost less, be repeatable, and improve management. The market benefits of a common model include customer satisfaction and establishing a specific data mining process (Hipp and Wirth, 2001). The CRISP-DM model was a non-proprietary methodology for data mining developed in the mid-1990's through a European group of companies. While steps are intended to be sequential, often backtracking is required (Sharda et al., 2014).

The CRISP-DM phases include business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Business understanding concerns outlining project objectives, business requirements, problem definition, and initial project plan. Data understanding begins data collection to familiarize oneself with the data, locate data quality issues, and form hypotheses. The two sections of business understanding and data understanding are closely linked. Data preparation covers table creation, record selection, cleansing, deriving attributes and the transfer of data into modeling tools. In the modeling phase different techniques are chosen and applied, data preparation and modeling are closely linked. In the evaluation phase, models have been constructed and verify the model meets business requirements. Deployment, while the final phase does not indicate the end of a project, and may require additional steps by the data analytics or customer, and care should be taken to clearly define the setup of the working model (Hipp and Wirth, 2001).

SEMMA

SEMMA is an acronym for Sample, Explore, Modify, Model and Assess and follows a data mining process developed by the SAS Institute and is included in SAS Enterprise Miner software. In the Sample step data tables are created that contain enough information to be significant but small enough to process efficiently. In the Explore step the analyst looks for relationships, trends, and anomalies to improve overall understanding. In the Modify step data is transformed and chosen to prepare for the next step. In modeling analytical tools and methods are used to predict outcomes. In the Assess step findings are reviewed among different predictive models (Truxillo and Wells, 2014).

EMSMA

A slight variation of the SEMMA approach is developed for Big Data and in-memory analytics called EMSMA. Big Data is a term to describe the exponential growth of data in structured and unstructured forms. Big Data is beyond just large amounts of data and characterized by

"V's": volume, variety, velocity, veracity, and value. This follows similarly to SEMMA though the steps include Explore, Modify, Segment, Model, and Assess. This approach effectively bypasses the sampling step since the high-performance systems are capable of processing large amounts of data eliminating the requirement to sample first (Sharda et al., 2014; Ravenna et al., 2015).

Model Development

Parsimony

The principle of parsimony requires abandoning complex models and utilizing simpler models that can generalize new problems. A parsimonious model is one that meets the necessary requirements while limiting factors (Vandekerckhove and Matzke, 2014). Occam's principle states that all things being equal the simpler solution is preferred. For example with data mining decision trees are often pruned to develop simpler solutions and prevent over fitting (Bensusan, 2014).

Background

BEMO was introduced during an analytics and data mining course for application. This course is offered as a required major course and general elective course throughout the University. The initial conceptualization stemmed from the complexity for learners from varying degree backgrounds to recall the exact steps of CRISP-DM as the naming is not a direct equivalent of the process steps, and constraint of SEMMA omitting process steps contained in CRISP-DM along with ability to clearly segment steps of SEMMA. For example, the exploration and modification steps were often combined or reversed since many of these activities occur in parallel and in successive iterations. In addition, the data mining steps were tied to specific vendor tools and thereby limited a generalized learning model to be applied in business application scenarios.

To address these limitations a simpler, easily remembered and non-proprietary methodology was introduced for applications. BEMO is a parsimonious data mining process with the steps of Business Opportunity, Exploration, Modeling, and Operationalization. During business opportunity the goals or objectives are defined along with the problem being solved. During the exploration step data quality is reviewed and updated including variable selection, outlier correction, duplication correction, etc. During the Modeling steps the predictive models are utilized based on the combination of inputs and outputs. In the operationalization step the model is implemented and practical considerations of operations are incorporated. A comparison between general data mining steps, CRISP-DM, SEMMA, EMSMA and BEMO are shown below for illustrative purposes for methodology coverage, process steps, and parsimony of process design.

DOI: 10.5824/1309-1581.2016.3.007.x

Table 1: Comparison of Data Mining Methodologies

#	General Steps	CRISP-DM	SEMMA	EMSMA	BEMO
1	Define/understand purpose	Business Understanding			Business Opportunity
2	Obtain data	Data Understanding	Sample		Exploration
3	Explore, clean, pre-process data	Data Preparation	Explore	Explore	
4	Reduce, partition the data		Modify	Modify	
5	Specify task	Modeling	Model	Segment	Model
6	Choose the data mining techniques			Model	
7	Iterative implementation and model tuning				
8	Assess results	Evaluation	Assess	Assess	
9	Deploy model	Deployment			Operationalize

BEMO Application

Netflix

To aid in application of the BEMO process steps, a business case study of Netflix is used to help conceptualize and apply a real-world example demonstration. Netflix was co-founded in 1997 by Reed Hastings and Marc Randolph to offer online movie rentals. Today Netflix is the global leader in Internet television with 81+ million customers in 190+ countries consuming 125+ million hours of programming per day (Netflix, 2016). This case demonstration example of the BEMO data mining process follows the Netflix prize (Netflix, 2009).

Business Opportunity

Netflix's goal is to connect people to the movies they love and as a result developed a movie recommendation software called Cinematch, which helps determine if someone will. The Netflix Prize of $1 Million was aimed at the opportunity of improving the prediction accuracy of whether a customer would enjoy a movie based on their preferences. The Prize of $1 Million would be awarded for a prediction accuracy improvement of 10% on the same training data set, when comparing predictions vs. actual ratings.

Exploration

Netflix developed the training data and test data. The training data contained 100 million ratings on 18,000 movie titles, from 480,000 randomized and anonymized customers collected between October 1998 and December 2005 with ratings from 1 to 5 stars and the date and

rating of each title. The test data contained 2.8 million customer movie pairs with rating dates but the actual rating to be predicted. To prevent identification, some rating data was changed in the following ways: deletions, insertion of alternate rating and date, rating dates change.

Model

The primary data mining model algorithms that showed the best results was Singular Value Decomposition (SVD) and Restricted Boltzmann Machines (RBM). While alone SVD slightly outperformed RBM, a combination of the two reduced the error further. The test data was assessed by computing the square root of the averaged squared difference between the predicted and actual rating or the root mean squared error RMSE. The Cinematch RMSE was 0.9514 on the test data and 0.9525 on the training data. To meet the prize the RMSE must be 90% of 0.9525 or 0.8572. While in the prize description, the uncertainty of whether the improvement would take months or years was noted, on September 21, 2009 Netflix awarded the $1 Million Grand Prize to team "BellKor's Pragmatic Chaos" with a RMSE of 0.8567.

Operationalization

Often times the focus is spent on predictive modeling, with little time spent on project management and operationalization of the project. For example, Netflix awarded the $1 million prize but unfortunately the underlying model that won the $1 Million was never put into operation due to limitations. The models were built to accommodate 100 million ratings from the training set and only a few million ratings on the testing set. To put this in perspective Netflix contains a Big Data repository on the order of 5 Billion+ ratings and growing which could not adapt to the algorithms. Also the additional accuracy gains of the new models did not justify the engineering effort required to move to production. Additional factors in the model also were impacted during the timeframe, for example in 2007 Netflix launched a streaming service which changed the way customers were using the Netflix services versus traditional mail rentals and also changed the data that was being collected (Masnick, 2012).

DISCUSSION

Following application, a formalized BEMO process is defined with additional details and components within each step for generalizability. Finally a process template is provided for repeatability and reusability of the process.

BEMO – Business Understanding

During business opportunity the goals or objectives are defined along with the problem being solved. This is a critical, though sometimes overlooked component of a data mining project that ensures successful definition and outcomes in support of the business objectives. During this step an executive summary is provided to communicate to key stakeholders from a customer, financial, employee, or operational perspective. An organizational and industry

DOI: 10.5824/1309-1581.2016.3.007.x

background is provided along with key objectives or goals that are expected as a result of the data mining project. This section is kept intentionally brief and at a summary level which is intended to capture the interest and motivation of stakeholders in conducting the data mining project.

BEMO – Exploration

During the exploration step data quality is reviewed and updated including variable selection, outlier correction, duplication correction, data reduction, data visualization, data standardization or normalization. During exploration variables must be assessed for type for future model use. There are two general variable types of continuous and categorical or put another way quantitative and qualitative. Continuous variables may be further broken down to include interval, ratio, and numeric data types. Categorical or qualitative variables may be further broken down to include nominal, ordinal, binary, or class variables (Laerd Statistics, 2013; **Shmueli et al., 2010;** Georges and Anderson, 2014; Truxillo and Wells, 2014).

With Big Data, traditional data processes, data management and data quality components must be reassessed and improved (Woodside, 2014). A McKinsey Global Institute report identifies metadata, data classification, data acquisition, and data fusion as key Big Data quality and management components. Metadata increases in importance with Big Data based on complexity and varying sources of data. For classification Big Data taxonomies must be developed to allow use of the information across the organization such as demographic, financial, geospatial, and protected health identifiers. Organization is important within data acquisition for Big Data to allow easier exchange and accessibility of information. Finally data fusion is the combination of structured and unstructured Big Data sources allows multiple areas information to be used for decision making and uncovering new insights as most business problems or opportunities are not defined in advance with Big Data processes**.** One of the greatest impacts on Big Data is veracity. Traditionally this has been measured as data quality based on the output or usage. For Big Data projects veracity must be measured beyond the output or usage to include validity, accuracy, timeliness, reasonableness, and completeness and transparent to users. Unstructured data also required special attention and process definitions (Prevosto and Marotta, 2014).

BEMO – Model

During the Modeling steps the predictive models are utilized based on the combination of inputs and outputs. Depending on the software application and reference source several varying naming conventions may be utilized (Laerd Statistics, 2013; Georges and Anderson, 2014; Brannick, 2016). Table 2 is provided as a terminology lookup for potential combinations.

Table 2: Model Input and Output Terms

Input	->	Output
Experimental	->	Response
Predictor	->	Target

X	->	Y
Independent	->	Dependent

Data is partitioned into training, validation, and testing sets as appropriate for the model parameters. Various models may be applied, interpreted, and tuned iteratively. A final model is chosen based on performance factors and business constraints. To assist in model selection, a table is provided with common factors and evaluation components for the data mining models of linear regression, logistic regression, k nearest neighbor, decision tree, neural network and support vector machine. For example logistic regression is utilized for categorical output, typically a binary output such as 0/1. To evaluate logistic regression commonly a classification matrix, error rate, lift, and receiver operating curve (ROC) is assessed (**Shmueli et al., 2010;** Georges and Anderson, 2014; Truxillo and Wells, 2014).

Table 3: Model Selection

	Output		Evaluation		
	Continuous	Categorical	R^2	Error/Class	Lift/ROC
Linear Regression	X	-	X	X	X
Logistic Regression	-	X	-	X	X
K Nearest Neighbor	X	X	-	X	X
Decision Tree	X	X	-	X	X
Neural Network	X	X	-	X	X
Support Vector Machine	-	X	-	X	X

BEMO – Operationalize

In the operationalization step the model is implemented and practical considerations of operations are incorporated. Model advantages and disadvantages are discussed in the context of a business setting application. For example model performance in a production setting, model development cost, among others. An overall project plan is developed to ensure on time, on budget, and on scope delivery. This plan includes practical considerations of resources, timeline, and iterative phases. A formal assessment of the project is also made including items such as return on investment (ROI), payback, net present value (NPV), and breakeven analysis.

Process Template	
Title:	

DOI: 10.5824/1309-1581.2016.3.007.x

Business Understanding	*Provide an executive summary* - *Organizational/Industry business background* - *Business problem/opportunity* - *Key business objective(s)/goal(s) from data mining* - *Business success criteria* - *Business requirements, assumptions, and constraints* - *Available resources* - *Key business terms*
Exploration	*Provide a results overview following data exploration* - *Data collection* - *Data dictionary* - *Data sampling* - *Graphical data exploration* - *Summary statistics (average, standard deviation, min, max, etc.)* - *Data quality exploration* - *Missing data* - *Outliers* - *Data cleansing* - *Data formatting* - *Data derivation* - *Data integration* - *Data normalization* - *Data relationships, correlations* - *Dimension reduction* - *Principal component analysis*

Model	*Partition*
	- *Train, # records*
	- *Validation/Test, # records*
	Apply a data mining model
	- *Model description*
	- *Model assumptions*
	- *Model parameters*
	- *Model steps*
	- *Inputs*
	- *Outputs*
	- *Variable descriptions*
	- *Model results*
	- *Model/variable tuning*
	- *Model results*
	- *Model solution*
	  o *Final model chosen*
	  ▪ *Variables in final model*
	  o *Model performance evaluation results*
	  ▪ *R^2*
	  ▪ *Adjusted R^2*
	  ▪ *Variable Significance*
	  ▪ *Lift Chart*

	▪ *Decile Chart* ▪ *Receiver Operating Curve* ▪ *Sum of Squared Errors* ▪ *RMSE* ▪ *Classification Matrix* ▪ *Error Rate*
Operationalize	*Overall Model Recommendations* - *Model advantages / disadvantages* - *Performance evaluation* - *Selection recommendation* *Implementation Recommendations* - *Project plan* - *Project timeline, resources, phases* - *Risks and contingencies* - *Costs and benefits* - *ROI, NPV, breakeven, implementation cost*

Figure 1: BEMO Process Template

Conclusion

Management Implications

Business Understanding is the critical first step of the BEMO process, and can address a variety of business opportunities such as process efficiency, cost reduction, revenue maximization, customer improvements, financial improvements, employee learning, and strategic advantages. Big Data, data mining, and analytics are invaluable without clear business objectives and value creation. In addition, committed sponsors such as senior executives, board, and key stakeholders are required for successful project initiation and development. Exploration is a fundamental second step of the BEMO process and often an area of significant time investment. Big Data veracity is the accuracy, quality, truthfulness, or trustworthiness of the data. Due to data limitations, data quality, and data capture many organizations spend considerable time and resources preparing data. Leading organizations have invested the resources to have high levels of data quality, capture and availability which facilitates improved project completion time and successful modeling results. Modeling is commonly thought of area within data mining, though is strongly dependent on preceding steps. In the modeling step efficient analytical systems and tools are required to easily interpret and apply data mining models. New high performance technologies such as in-memory analytics and distributed grid computing permit Big Data to processed in near-real-time for improved problem solving speed using a much large non-sampled dataset. Operationalization is the final step and involves putting the project into practice. During this step the project plan, resources, and costs are developed. The solution is maintained and tuned over time in response to new business challenges and opportunities (Sharda et al., 2014).

CONTRIBUTIONS

This paper provides a number of key contributions including developing a parsimonious model that can be more easily followed for a standard business process and developing a standard reusable and repeatable BEMO process template. Current data mining methods do not meet all requirements of business challenges, in addition current methods are difficult to remember and do not cover all requisite steps. The contribution of a parsimonious model also requires utilizing simpler models over complex models that can more efficiently generalize new problems.

REFERENCES

Bensusan, H. (2014). *God doesn't always shave with Occam's razor.* School of Cognitive and Computing Sciences.

Brannick, M. (2016). *Regression Basics*. University of South Florida. Retrieved from: http://faculty.cas.usf.edu/mbrannick/regression/regbas.html.

Georges, J., Anderson, C. (2014). *Advanced Predictive Modeling Using SAS Enterprise Miner 13.1*. SAS Institute.

Laerd Statistics. (2013). *Linear Regression Analysis Using Stata*. Lund Research. Retrieved From: https://statistics.laerd.com/stata-tutorials/linear-regression-using-stata.php.

Laerd Statistics. (2013). *Types of Variable*. Lund Research. Retrieved from: https://statistics.laerd.com/statistical-guides/types-of-variable.php.

Mike Masnick, M. (2012). Why Netflix Never Implemented The Algorithm That Won The Netflix $1 Million Challenge. Tech Dirt. Retrieved from: https://www.techdirt.com/blog/innovation/articles/20120409/03412518422/why-netflix-never-implemented-algorithm-that-won-netflix-1-million-challenge.shtml

Netflix. (2009). *Netflix Prize*. Retrieved from: http://www.netflixprize.com/

Netflix. (2016). *Netflix Media Center*. Retrieved from: https://media.netflix.com/en/about-netflix

Piatetsky, G. (2014). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. KDnuggets. Retrieved from: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Prevosto, Virginia, Marotta, Peter. Does Big Data Need Bigger Data Quality and Data Management? Verisk Review, 2014.

Ravenna, A., Truxillo, C., Wells, C. (2015). *SAS Visual Statistics*. SAS Institute.

Shmueli, G., Patel, N., Bruce, P. (2010). *Data Mining for Business Intelligence*. Wiley, Hoboken, NJ.

Sharda, R., Delen, D., Turban, E. (2014). *Business Intelligence A Managerial Perspective on Analytics*. Pearson, Upper Saddle River, NJ.

Truxillo, C., Wells, C. (2014). *Advanced Business Analytics*. SAS Institute.

Vandekerckhove, J., Matzke, D. (2014). *Model Comparison and the Principle of Parsimony*. CIDLab. Retrieved from http://www.cidlab.com/prints/vandekerckhove2014model.pdf

Wirth, R., Hipp, J. (2001). *CRISP-DM: Towards a Standard Process Model for Data*

Mining. DaimlerChrysler Research & Technology, University of Tübingen.

Woodside, J.M. (2014). Big Data Veracity in Healthcare. The 2014 International Conference on Advances in Big Data Analytics.