Six Sigma Methodology with Recency, Frequency and Monetary Analysis Using Data Mining

Andrej Trnka

***Abstract*—Data Mining methods provide a lot of opportunities in the market sector. This paper deals with Data Mining algorithms and methods (especially RFM analysis) and their use in Six Sigma methodology, especially in DMAIC phases. DMAIC stands for Define, Measure, Analyze, Improve and Control. Our research is focused on improvement of Six Sigma phases (DMAIC phases). With implementation of RFM analysis (as a part of Data Mining) to Six Sigma (to one of its phase), we can improve the results and change the Sigma performance level of the process. We used C5.0, QUES, CHAID and Neural Network algorithms. The results are in proposal of selected Data Mining methods into DMAIC phases.**

***Index Terms*—Data mining; DMAIC, RFM, six sigma.**

I. Six Sigma Research

In our research we tried to implement several Data Mining algorithms to the Six Sigma methodology and improve performance level of this methodology. We focused on existing process, so we implemented Data Mining algorithms to DMAIC phases. DMAIC stands for Define, Measure, Analyze, Improve and Control. These phases represent "life cycle of Six Sigma".

Six Sigma is a rigorous, focused, and highly effective implementation of proven quality principles and techniques. Incorporating elements from the work of many quality pioneers, Six Sigma aims for virtually error – free business performance. Sigma is a letter in the Greek alphabet used by statisticians to measure the variability in any process. A company's performance is measured by the sigma level of their business processes. Traditionally companies accepted three or four sigma performance levels as the norm, despite the fact that these processes created between 6200 and 67000 problems per million opportunities! The Six Sigma standard of 3.4 problems-per-million opportunities is a response to the increasing expectations of customers and the increased complexity of modern products and processes. [1]

About 95 percent of all Six Sigma projects follow the DMAIC phases. Fig. 1 shows the DMAIC phases in the Six Sigma methodology.

The Define phase involves preparing a business charter (rationale for the project); understanding the relationships between Suppliers–Inputs–Processes–Outputs–Customers (called SIPOC analysis); analyzing Voice of the Customer data to identify the critical-to-quality (CTQs) characteristics important to customers; and developing a project objective.

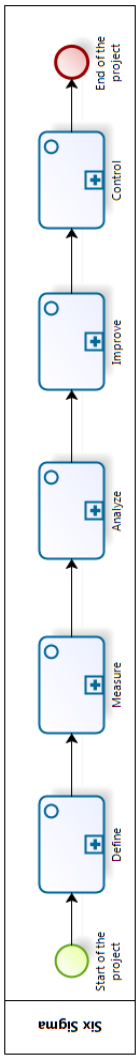

Fig. 1. DMAIC phases.

The Measure phase involves developing operational definitions for each CTQ variable; determining the validity of the measurement system for the CTQs; and establishing baseline capabilities for each CTQ.

The Analyze phase involves identifying the upstream variables (Xs) for each CTQ using a flowchart. Upstream variables are the factors (Xs) that affect the performance of a CTQ. Additionally, the Analyze Phase involves using Failure Modes and Effects Analysis (FMEA) to eliminate Xs that are not likely to impact the CTQ; operationally define each X; collect baseline data for each X; perform studies to determine the ability of the measurement system for each X to adequately reflect the behavior of each X; establish baseline capabilities for each X; and understand the effect of each X on each CTQ.

Manuscript received May 12, 2012 revised June 25, 2012. This work supports the VEGA project No. 1/0214/11.

Andrej Trnka is with the University of SS Cyril and Methodius in Trnava, Faculty of Mass Media Communication, Nam. J. Herdu 2, 917 01 Trnava, Slovak Republic (e-mail: andrej.trnka@ucm.sk).

The Improve phase involves designing experiments to understand the relationships between the CTQs and the Xs; determining the levels of the critical Xs that optimize the CTQs; developing action plans to formalize the level of the Xs that optimize the CTQs; and conducting a pilot test of the revised process using the levels of the critical Xs that will hypothetically optimize the CTQs.

The Control phase involves avoiding potential problems with the Xs with risk management and mistake proofing (discussed in the next paragraph); standardizing successful process revisions; controlling the critical Xs; creating a set of instructions for turning the improved process over to the process owner, called a control plan; turning the revised process over to the process owner; and disbanding the team and celebrating its success. [2]

II. Six Sigma With RFM Analysis (Data Mining)

Data Mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques [3].

A lot of algorithms are used in Data Mining process. Figure 2 shows a widely used approach in Data Mining called CRISP-DM. The steps of CRISP-DM should be followed in each Data Mining project.

As a Data Mining tool we used PASW Modeler 14. RFM stands for Recency, Frequency, Monetary. RFM analysis helps to identify high spending customers. We decided to implement this analysis, because Six Sigma is customer oriented methodology. With RFM analysis we should improve performance level of Six Sigma methodology and minimize process costs.

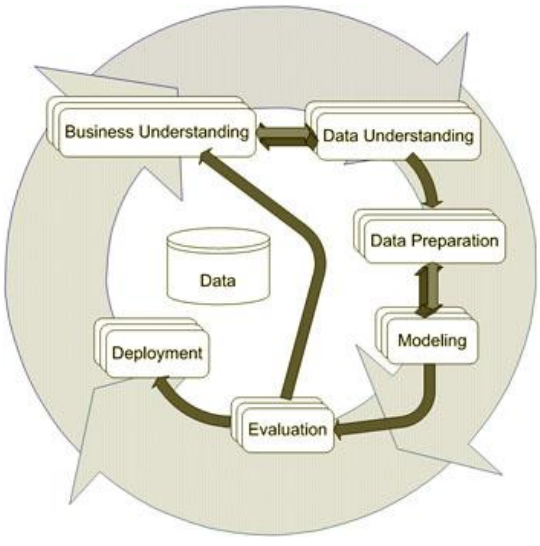

Fig. 2. CRISP-CD approach.

III. RFM Research

In PASW Modeler are two RFM nodes: RFM Aggregate and RFM Analysis. RFM Aggregate enables to take a customer's historical transactional data and combine that data into a single row that lists when they last deal with vendor, how many transactions they have made, and the total monetary value of those transactions. RFM Analysis then enables to further analyze this prepared data.

In our research we have to respect our local laws. Some variable values represented in this paper were changed. We wanted to identify the highest spending customers in order to offer loyalty rewards.

For RFM Score we used the historical data showing each purchase made by a customer, incorporating both the amount spent and the date of the transaction. Our dataset consists of 59470 records. Fig. 3 shows built model. Node called RFM Score is so-called Super Node. This node consists of several nodes, including RFM Score and RFM Analysis. The results were descending sorted.

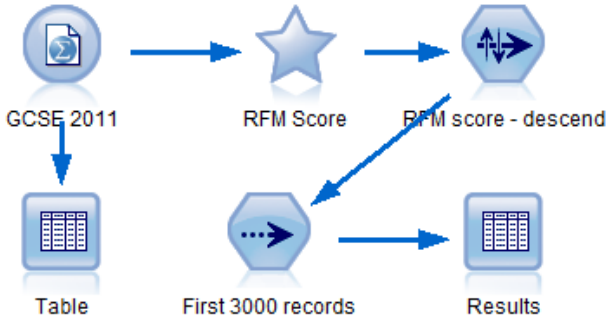

Fig. 3. Model with RFM score.

IV. Predicting Promotional Target

The next step in our research with RFM took the aggregated RFM data and combined it with customer and historic promotional campaign data to predict responses to future marketing campaigns. We wanted to identify recent customers who have positively responded to previous marketing campaigns and then test different models to see which gives the best predictions of target customers for future promotions.

We used the historical data showing the each purchase made by a customer, incorporating both the amount spent and the date of the transaction. In separate two dataset we stored the details about each customer, such as income, marital status and information on whether they have responded to previous promotional campaigns (used in Merge Super Node). The proposed model was created from the part of the previous model. Fig. 4 shows built model.

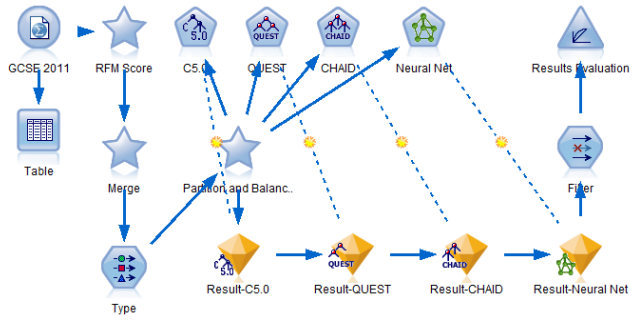

Fig. 4. Predictive model.

To assess how well the model works with data in the real world, we used a Partition node to hold out a subset of

records for purposes of testing and validation. This node splits the data into two equal halves: one for training the model against and one for testing against.

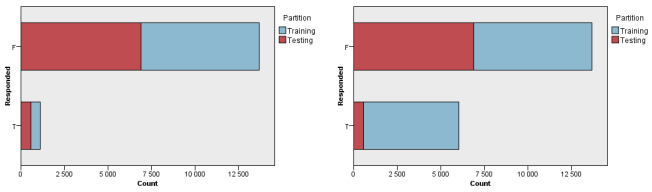

Fig. 5. (a) Distribution of merged data ; (b) Balanced distribution of data.

Fig. 5a shows the differences of distribution true and false responses. Some modeling techniques have trouble with such biased data, so we balanced the data with approximately equal numbers of true and false outcomes. This enabled the model to have a better chance of finding patterns. Fig. 5b shows the balanced data.

After balancing the response data, we tested a couple of model types (C5.0, CHAID, QUEST and Neural Net) against it to see which is better at predicting future responses.

V. Results

The results of RFM Analysis (Fig. 3) are the scores for customers who were both recent and frequent purchasers and who spent a reasonable amount on each transaction.

Fig. 6 shows the results of RFM Analysis. The most important is column RFM score, which shows the current value of RFM Score. These results should be saved and combined with another data set that contains personal details about each customer.

Results (5 fields, 3 000 records) #1

File Edit Generate

Table Annotations

	Card_ID	Recency	Frequency	Monetary	RFM Score
1	CID010...	18	22	679.730	150.000
2	CID010...	13	17	687.870	150.000
3	CID010...	26	22	1064.310	150.000
4	CID010...	11	23	715.810	150.000
5	CID010...	13	23	662.260	150.000
6	CID010...	28	30	1118.660	150.000
7	CID010...	12	26	1162.160	150.000
8	CID010...	16	27	833.700	150.000
9	CID010...	23	16	1399.890	150.000
10	CID010...	12	41	2089.060	150.000
11	CID010...	11	7	914.980	150.000
12	CID010...	12	27	1266.730	150.000
13	CID010...	15	18	832.360	150.000
14	CID010...	19	17	688.850	150.000
15	CID010...	12	17	1921.880	150.000
16	CID010...	14	7	655.960	150.000
17	CID010...	20	25	882.740	150.000
18	CID010...	27	12	1924.940	150.000
19	CID010...	25	8	1306.480	150.000
20	CID010...	12	28	1377.680	150.000

OK

Fig. 6. Results (a part) of RFM analysis.

The results of predicting promotional target (Fig. 4) shows Fig. 7. To compare the models predicted accuracy, we built a gains chart. In addition, we plot the RFM Score to see if this is a good predictor.

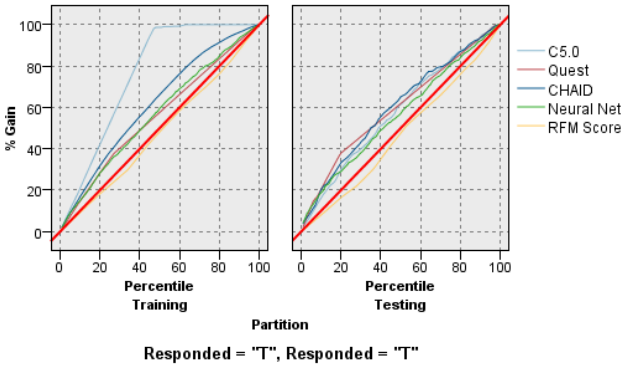

Fig. 7. Evaluation graph of used models.

Using the testing data, a comparison between the model types shows that the QUEST or CHAID model produces a more accurate prediction model and is therefore the model to use.

For storage the data we suggest creating the Data Warehouse, because integrity of the data from process is variable [4-6].

We implemented these techniques and Data Mining algorithms to the Define phase of DMAIC. Fig. 8 shows position of RFM analysis in this phase.

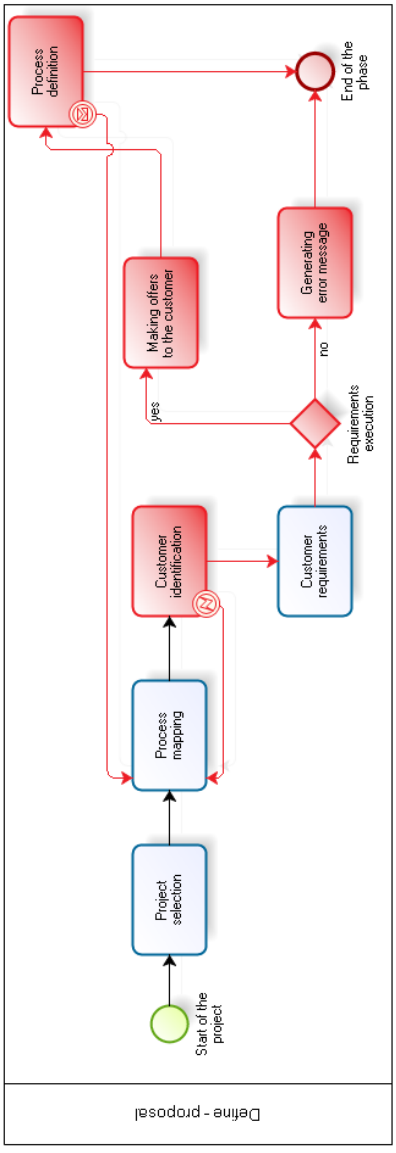

Fig. 8. Proposed place of RFM Analysis in the Define phase.

Proposed techniques are represented by Customer identification step.

ACKNOWLEDGMENT

Grateful acknowledgement for translating the English edition goes to Juraj Mistina.

REFERENCES

[1] T. Pyzdek and P. Keller, "The Six Sigma HandBook: A Complete Guide for Green Belts, Black Belts, and Managers at All Levels," The McGraw-Hill, 2010.
[2] H. S. Gitlow, *A Guide to Lean Six Sigma Management Skills*, 2009.
[3] D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley, 2005.
[4] PASW Modeler 14: Application Guide. Integral Solution Limited, 2010.
[5] P. Tanuska and T. Skripcak, "Data Driven Scenario Test Generation for Information Systems," *International Journal of Computer Theory and Engineering*, vol. 3, No. 4, pp. 565-570.
[6] R. Halenar, "Loading data into data warehouse and their testing," *Journal of Information Technologies*, vol. 2, pp. 7-14, 1999.

Andrej Trnka is with Faculty of Mass Media Communication, University of SS Cyril and Methodius in Trnava, Slovak Republic. He is a member of IAENG, IEEE, ACM and IACSIT. He received Ph.D. degree (2010) in area of informatics and automation from Slovak University of Technology, Faculty of Material Science and Technology Trnava, Slovak Republic. His research includes the field of Information Systems, Data Mining, Six Sigma methodlogy and Statistical Process Control. He published papers in national and international conference proceedings and journals.