Name: Mutunga Samuel Mutua

Roll no: S22/0145

Year, Semester: Year 2, Semester 4

Courser Unit: BCS2222 – Data Science Algorithms and Tools

Qn 1. Explain the concept of CRISP-DM in the Big Data Cycle and propose your contributions towards advancing the theory, this may be your new way of the cycle, citing at least 5 different research papers. Your work must be published in an online journal and the link must be shared. Please note that your work will not be considered without the link

CRISP-DM is the de facto standard and an industry-independent process model for data mining. It is an acronym for Cross Industry Standard Process for Data Mining.

The objectives and benefits of CRISP-DM are.

1. Ensures quality of knowledge discovery project results

2. Reduce skills required for knowledge discovery

3. Reduce cost and time

4. It is general purpose, i.e., stable across varying applications

5. Robust, i.e., insensitive to changes in env.

6. It is tool and technique independent

7. It is tool supportable

9. Supports documentation of projects

10. Captures experience for reuse

11. Supports knowledge transfer and training.

CRISP-DM can be described in terms of a hierarchical process model, having 4 levels of abstraction. i.e., Phases, Generic tasks, Specialized tasks, and process instances. These four levels of abstraction can be briefly explained as follows.

1. Phases. This is at the top level, the data mining process is organized into small no of phases and each phase comprises of second-level generic tasks.

2. Generic tasks. This level is intended to be general enough to cover all possible data mining situations. These tasks are designed to be as complete and as stable as possible. Complete means to cover the whole process of data mining and possible data mining applications. Stable means to be valid for yet unforeseen developments like new modeling techniques.

3. Specialized tasks. This level describes how tasks in the generic tasks should be carried out in specific situations. e.g., from build model to build response model having specific activities to the problem and to the data mining tool chosen. CRISP-DM does not attempt to capture all possible routes through data mining process because it would require complex process model and the expected benefits would be very low.

4. Process instances. This is a record of actions, decisions, and results of an actual data mining engagement. A process instance is organized according to tasks defined at the higher level but describes what happened in a particular engagement rather than what happens in general.

Further, CRISP-DM distinguishes between reference model and user guide. Whereas reference model represents a quick overview of phases, tasks, and outputs and describes what to do in a data mining project, the User guide gives more detailed tips and hints for each phase and each task within a phase and depicts how to do a data mining project.

CRISP-DM User Guide comprises the following items.

1. check lists.

2. questionnaires.

3. tools and techniques.

4. sequence of steps.

5. decision points.

6. pitfalls.

Generic CRISP-DM reference model

The reference model has six phases namely, business understanding, data understanding, data preparation, modelling, evaluation, and deployment. These phases are briefly explained below.

1. Business understanding

Focuses on understanding project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition, a preliminary project plan should be created.

2. Data understanding

This starts with initial data collection from data sources and proceeds with other activities to get familiar with data to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information. There is a close link between Business Understanding and Data Understanding. The formulation of the

data mining problem and the project plan require at least some understanding of the available data.

3. Data preparation

In this phase data selection should be conducted by defining inclusion and exclusion criteria. Bad quality data should be cleaned, derived attributes should be constructed, data transformation for modeling tools. Further, different methods are possible and model dependent. The final dataset is fed into modeling tools.

4. Modelling

Here various modeling techniques are selected, applied to building test case and the model. The choice of modeling technique depends on business problems and the data. But how do you explain the choice? Specific parameters are required and calibrated to optimal values to build the model. In addition, to assess the model it is appropriate to evaluate the model against evaluation criteria and select the best ones. Often one realizes data problems while modeling or one gets ideas for constructing new data.

5. Evaluation

At this point you have built one or more models that appear to be of high quality, from a data analysis perspective. Before proceeding thoroughly evaluate the model, and review the steps executed to construct the model. This is to ensure it achieves business objectives.

Key objective: Are there important business issues that have not been sufficiently considered? The results are checked against defined business objectives.

A decision on the use of the data mining results should be reached.

6. Deployment

This phase is described generally in the user guide (It desc that the deployment phase consists of planning the deployment, monitoring, and maintenance). It could be the final report or a software component. Further, this can be as simple as generating a report or as complex as implementing a repeatable data mining process. This phase is carried out by the user.

The CRISP-DM is a popular process in practice and research, however, I believe the following areas should be considered as possible additions to the current process.

1. Enrich specialized process model with examples gained from case studies.

2. Project management framework to ensure firm deadlines, timely completion of tasks and proper usage of resources. This also enable one to plan for iterations explicitly.

3. Quality assurance to take care of the any quality issues that may arise in the specialized process model.

References:

1. CRISP-DM: Towards a Standard Process Model for Data Mining by *Rüdiger Wirth* DaimlerChrysler Research & Technology FT3/KL PO BOX 2360 89013 Ulm, Germany, ruediger.wirth@daimlerchrysler.com, and *Jochen Hipp* Wilhelm-Schickard-Institute, University of Tübingen Sand 13, 72076 Tübingen, Germany jochen.hipp@informatik.uni-tuebingen.de.
2. A systematic Literature Review on Applying CRISP-DM Process Model by Christoph Schroer, Felix Kruse and Jorge Marx Gomez available online on www.sciencedirect.com.
3. KDD, SEMMA AND CRISP-DM: A parallel overview by Ana Azevedo and M. F. Santos.
4. Data Engineering in CRISP-DM Process Production Data – Case Study by *Jolanta BRZOZOWSKA* [0000-0002-4355-2847]*, *Jakub PIZOŃ* [0000-0002-0806-6771]** *Gulzhan BAYTIKENOVA****, *Arkadiusz GOLA* [0000-0002-2935-5003]*****, *Alfiya ZAKIMOVA****, *Katarzyna PIOTROWSKA* [0000-0003-0899-7610]*****
5. CENTER FOR APPLIED RESEARCH TECHNOLOGY - Data Mining in MRO – Amsterdam University of Applied Sciences by Maurice Pelt, Asteris Apostolidis, Robert J. de Boer, Maaik Borst, Jonno Broodbakker, Ruud Jansen, Lorance Helwani, Roberto Felix Patron, Konstantinos Stamoulis