Towards a Data Mining Methodology for the Banking Domain

Veronika Plotnikova

Supervisors: Marlon Dumas, Fredrik P. Milani, Robert Kitt

University of Tartu, Institute of Computer Science, J. Liivi 2, 50409 Tartu, Estonia
veronika.plotnikova@ut.ee

Abstract. Telecoms and financial service industries are leaders in adopting data analytics technologies, practices, and heavily invest into 'Big Data' tools and related competence development. However, many of them fail to realize benefits of data-driven decision making and maximize 'Big Data' business value due to lack of knowledge on how to frame, approach and tackle complex data analytics projects. Existing data mining methodologies are domain-independent, general, abstract and partially outdated. Several refinements of data mining methodologies have been proposed, but they address specific aspects or tasks and remain fragmented. The goal of this doctoral project is to develop a domain-specific data mining methodology for the financial sector, which (1) represents consolidation of existing body of knowledge, and (2) is validated on the sample of real life data-mining projects. The proposed illustrative case studies approach is based on broad, typical data mining use cases portfolio executed across different geographical regions and business areas of the financial institution.

Keywords: Big data, Data mining, CRISP-DM, Banking, Financial services.

1 Introduction

The 'Big Data' phenomenon, technological advances in data processing and development of algorithmic techniques have fostered widespread adoption of data analytics across different industries. According to the most recent market studies [1-2] adoption rate of 'Big Data' analytics tripled for all companies reaching 53% in 2017, up from 17% in 2015. Study based on global in-depth survey of 583 business and IT professionals [3] revealed that 40% of organizations are already using data analytics across key business functions, and it forecasted to double: the rate should exceed 70% in 2018 and reach 90% in 2020. Telecommunications and financial services are the leading industry adopters with 87% and 76% of the respective sector companies already reporting the data analytics usage [1-2] – well above average figures.

Telecoms and financial sectors as early adopters have developed specific datasets, varieties of data and execute broad set of data mining tasks to solve industry-specific business problems. Therefore, both industries are naturally the most suitable sectors for in-depth exploration of data analytics[1] phenomena and its impact on organizations and business practices. Also, both telecoms and financial services explicitly demonstrate the trend of heavy investments into data analytics technologies and competences seeking to realize benefits from data-driven decision-making and maximize 'Big Data' business value. However, many of them consequently fail due to lack of knowledge on how to approach and tackle complex data analytics projects. Well-developed, comprehensive, domain-specific methodologies and guidelines to govern data analytics deliveries is key pre-requisite to ensure their success. Business value is realized by reusability, repeatability, scaling and actionability of resulting data analytics products, solutions and insights across organization and is dependent on domain-specific factors.

Academic literature to date have studied [4, 10] data mining use cases catering to broad variety of business problems along with application-specific issues [5]. In contrast, existing standard data-mining methodologies have not been extensively and explicitly discussed; they are domain-independent, rather generic, abstract and partially outdated. There are attempts to introduce refinements, but they are also fragmented and concentrated at two opposite ends of the spectrum - either proposing additional elements into a data mining process, or focusing on organizational aspects (general data mining processes and tools integration into business, enterprise and IT architectures); domain-specific factors are not considered.

Comprehensive, domain-specific methodologies for data analytics projects are critical for business value realization, but they do not exist. The purpose of this PhD project is to bridge the gap and develop such data mining methodology. As telecoms and financial services are identified as one of the most suitable sectors for in-depth exploration of data analytics business practices, the new methodology will be designed for one of them - banking domain[2]. The project's research proposal is structured as follows. *Section 2* introduces necessary basic concepts and terminology, and reflects on their current usage by practitioners. *Section 3* offers literature review followed by identification of existing research gaps and formulation of research questions, *Section 4* proposes research methodology while *Section 5* concludes.

2 Basic Concepts and Related Terminology

Data Mining is defined as set of rules, processes, algorithms that are designed to find valuable 'knowledge', extract patterns, identify relationships, etc. from large date warehouses or datasets [10]. This involves automated data extraction, processing, modeling with the help of vast range of methods and techniques of statistics, machine learning, artificial intelligence, etc. There are three major standard methodologies

[1] In this paper, data analysis and data mining are used as synonyms, even though it is acknowledged that data analytics is broader field, as it encompasses statistical analysis methods that are traditionally not associated with data mining.

[2] In this paper, banking domain refers to universal banking business model with extensive products and services portfolio offered to all types of clientele, and with variety of support functions (risk, operations, etc.).

developed and widely used in academic research and in business practices, CRISP-DM, SEMMA, ASUM-DM. Short overview of each and current usage practices are presented in the following subsections.

2.1 Overview of Existing Standard Data Mining Methodologies

CRISP-DM (Cross-Industry Standard Process for Data Mining) is industry–driven guidelines to perform data mining on large datasets [9-11]. It originated from KDD (Knowledge Discovery in Databases) field which also had KDD process developed in 1996 [8]. Essentially, CRISP-DM was built on KDD process fundamentals[3], however, with several abstraction layers it has achieved much higher level of complexity and details (eg. generic tasks level consists of 24 tasks and outputs), thereby, representing refinement of KDD process. CRIPS-DM development was led by industrial consortium with the final version published in 2000; attempts to update initiated in 2006 were unsuccessful. CRISP-DM divides data mining process into six not strictly sequential, but iterative phases – business understanding, data understanding and data preparation, modeling, evaluation, and deployment.

SEMMA (Sample, Explore, Modify, Model and Assess) is list of sequential steps guiding implementation of data mining process developed by SAS Institute [10-11].

ASUM-DM (Analytics Solutions Unified Method for Data Mining) was released in 2015 by IBM with the purposes to refine and extend CRISP-DM.

2.2 Data Mining Methodologies Usage Patterns

According to KDNuggets[4] polls results presented in the Table 1, the leading methodology for data mining process is CRISP-DM, followed by SEMMA and KDD [6].

Table 1. KDNuggets Poll on Data Mining Methodology results, [6]

Poll Years	**2002**	**2004**	**2007**	**2014**
CRISP-DM	51%	42%	42%	43%
SEMMA	12%	10%	13%	8.5%
KDD process			7%	7.5%
My organization's	7%	6%	5%	3.5%
My own	23%	28%	19%	27.5%
Other (incl. domain specific)	4%	6%	9% (5%)	10% (2%)
None	4%	7%	5%	0%

However, the usage of CRIPS-DM has reached plateau while others are steadily declining. Importantly, data scientists own methodologies usage stays above 25% rate

[3] KDD process consists of 9 steps: learning application domain, dataset creation, data cleaning & processing, data reduction & projection, choosing the function of data mining, choosing data mining algorithm, interpretation, using discovered knowledge.

[4] One of the leading websites on Business Analytics, Data Mining, and Data Science (edited by Gregory I. Piatetsky-Shapiro, one of the major contributors to Knowledge Discovery and Data Mining concepts).

and coupled with other ones (domain and non-domain specific) is steadily increasing reaching usage rate of over 30% [6]. This indicates decline in adoption rates of CRISP-DM and potential need for revision and modification. Indeed, this methodology though widely used was not updated since 2000 while data mining usage, methods and tools have developed exponentially.

3 Literature Review

The literature review was conducted using key principles of Systematic Literature Review approach [7]. The corpus of scientific research articles, publications and books was retrieved and the following steps conducted.

Step 1 - Scopus and Web of Science databases have been searched with the search string of the three standard major methodologies described in *Section 2*, i.e. “CRISP-DM”, “SEMMA”, “ASUM-DM” jointly with domain keyword “banking”[5]. All texts referred from databases were retrieved and included into literature corpus.

Step 2 - Identical procedure as in Step 1 was performed for Google Scholar database, but with the delimitation - the texts corpus was retrieved for the first 100 hits. The threshold was determined empirically based on evaluation of relevancy of texts spanning beyond first 100 search results. The relevancy of the retrieved texts after the given threshold declined significantly and did not contribute to additional insights.

In both steps, there were no time restrictions set, all texts were retrieved as many years back as database contained, oldest publication dated back to 1998, newest to 2018. 1/3 of studies have been published over last 3 years while approximately half of the scientific texts are concentrated over last 5 years period. Overall text corpus was reviewed and evaluated on iterative basis with respect to the relevancy of studies. Summary statistics of the literature reviewed is presented in the Table 2 below.

Table 2. Summary statistics on retrieved publications

Database	Scopus and Web of Science	Google Scholar	Total	Class 1 texts	Class 2 texts
No. of texts (string Crisp-DM)	57	91	**148**		
No. of texts (string SEMMA)	9	94	**103**		
No. of texts (string ASUM-DM)	1	3	**4**		
Total (excl. duplications)	**61**	**163**	**224**		
Total (excl. irrelevant)	**55**	**132**	**187**	**83**	**104**

Scientific publications from databases were supplemented by additional set of general materials (over 20 various texts). They were primarily retrieved from industry web-

[5] As CRISP-DM methodology is elaborated derivation, refinement of KDD process (as described in Section 2.1), KDD was omitted from the direct search.

sites via general search and provide descriptive information on data mining methodologies and processes in industry context.

Analysis of the selected publications corpus enables to perform next research steps:

1. construct high-level typification of research performed in the field over the last 10 years,
2. identify and categorize the existing research gaps, and
3. formulate research questions.

Based on analysis of scientific publications, existing research can be broadly typified into two major classes.

The first research class (hereinafter, *Class 1*) relates to application of various data mining methodologies for specific case studies. Importantly, the typical purpose of case studies is to solve various business problems of the financial institutions by the means of modeling tasks. The case studies can be further categorized as follows:

1. customer behavior modeling with the purpose to identify customer likely to churn or loyal customer [13],
2. profiling customers either according to the usage patterns of various digital channels while interacting with the bank, patterns of electronic transactions, eg., [13-14] or based on other features,
3. overall customer relationship management including customer segmentation tasks, customer targeting [15],
4. modeling tasks to support variety of risk management processes:
    a. credit risk identification and management – credit scoring, modeling and identifications of defaults [16],
    b. identification and prevention of fraud behavior and/or ALM risks,
    c. risk control activities including auditing (internal/external in bank domain) [17],
5. efficiency studies, eg. optimization of branch network [18].

In *Class 1* publications, the relevant data mining methodologies are used to structure the data mining process and achieve data mining goals. Critical discussions are not common, and if present, are structured around the method application at best, typically considering data.

Also, *Class 1* research concentrates on the application of the particular scientific technique processing aspects, types of modeling techniques with associated selection of the best one based on evaluation results, model validation aspects, feature selection and the final set of the best predictors. At the same time, there is lack of critical evaluation of methodology aspects, discussions on the methodology steps, substeps that need to be modified, added, or are redundant is largely omitted. Knowledge discovery in relation to executing the data mining task methodologically remains ‘hidden’, ‘tacit’ and confined within individual experience of the data mining experts. This might be evidenced by own methodologies usage growth as identified in subsection *2.2*.

The second class of publications (hereinafter, *Class 2*) concentrates on data mining methodologies or processes on a higher abstraction levels. A subset of these studies also contains case studies similarly to *Class 1* publications, but in contrast, these experiments are conducted on a broader scope with larger number of organizations and/or data mining tasks. Also, *Class 2* publications typically present critical evaluation of existing standard data mining methodologies. Such approach supports identification of deficiencies and suggests improvements. Importantly, *Class 2* research takes various domain and industry perspectives. However, most of the studies focus on the analysis of specific step of the methodologies. Very rare exception is [12] which proposes novel direction - design of fuzzy expert system to evaluate overall success of data mining projects by evaluating each step of the process methodology.

Critical evaluation results and proposed suggestions can be structured based on the following methodology phases, steps or areas.

Deployment phase and business process. CRISP-DM methodology is identified as lacking deployment phase details which can support integration of data mining results into business process [19]. Pivk, et al identify relationship between data and data mining sophistication levels, and propose improvements by use of ontologies (domain, business process and data mining) including extension elements to CRISP-DM, and Service-Oriented Architecture for data mining. [20] proposes new deployment framework (DEEPER). Associated concepts of ontologies and broader business architecture for establishing data mining systems in organization are also discussed [21].

Data preparation phase and data requirements. Number of studies proposes additional substeps and techniques for data preparations stage starting from adjustments to KDD initiated in [22] or alternatively, specific methodologies on gathering and structuring data requirements in the broader context of data-intensive projects and data governance [23]. These studies are performed in the context of IT system architecture, discussing enterprise data warehouses, 'data lakes' and associated data and information modeling and management concepts (eg. Business Information Modeling). Given the fact that ~80% time in data mining process is taken by data preprocessing and preparation steps, this part of research is of utmost importance.

Model evaluation and selection phase. This research direction focuses on relevant methodology enhancements to model evaluation and selection steps based on decision-support framework, eg. [24] proposes hybrid methodology and procedure for generating and selecting the most appropriate casual explanatory model.

Novel methodology enhancements and adjustments. Limited, but valuable number of studies has emerged as a response to legislation and regulatory requirements, eg. [25] developed DADM (Discrimination-aware data mining) framework. Other valuable direction of research is represented by authors proposing extension of methodological frameworks from other business areas or processes. Adaptive Software Development (ASD) methodology is adopted and introduced as ASD-DM for predictive data mining in [26]. Other research is associated with Sex Sigma Lean methodologies modifications and application in data mining process context, eg. DMAIC[6] application discussed in [27].

BI technologies, tools and IT architectures perspectives. Part of the studies acknowledge importance of data mining processes and associated methodologies when designing and implementing respective BI, Data Science technologies and tools

[6] Acronym for Define, Measure, Analyze, Improve and Control, refers to a data-driven improvement cycle used for improving, optimizing and stabilizing business processes and designs.

in the organizations. Such studies lack enhancement prospective, however, they discuss relevant aspects for successful integration of data mining process into overall IT architecture [28].

Organizational prospective. Finally, there is set of *Class 2* publications progressing to higher levels of generalization [29]. These studies do not focus on application of data mining methodologies, but rather concentrate on broader investigation on adoption of data mining as such. These studies, though not addressing concrete methodological aspects are rather important as they discuss relevant motivational and organizational aspects. These aspects are disregarded in existing standard data mining methodologies, however, they do represent an inseparable part of practical context and implementation environment in which data mining methodology is used.

The literature review showed a few well-developed frameworks for data mining, and they have been created for wide industry application. Existing data mining methodologies do not cater to specific industry needs such as banking domain. Thus, existing research gap can be formulated as follows:

***Research Gap** – Lack of comprehensive data-mining methodology applicable, adapted for banking industry.*

The following research questions address it:

***RQ1:** What are the existing data mining frameworks and what components they include?*

***RQ2:** What within the existing frameworks could be re-used, removed or needed to be added in order to develop the data mining methodology for banking domain?*

The research methodology to address research questions is presented in *Section 4*.

4 Research Methodology

The research methodology consists of two phases summarized in the Table 3.

Table 3. Research methodology overview for ***Phase 1*** and ***2***

Phase	RQs	Activities	Expected Outcome
1 - Comprehensive review of existing frameworks	RQ1	Systematic Literature Review and analysis of its results	Comprehensive overview of existing DM frameworks
2.1 - Refinements generation	RQ2	Identification, consolidation of refinements towards existing DM methodologies from existing literature	Structured list of refinements to DM methodologies phases, steps and deliverables
2.2 - Validation	RQ2	Validating refinements proposed in phase 2.1 with sample of real-life data mining projects. Removal of conflicting, irrelevant refinements	Common, validated refinements set

Expected outcome of the research is conceptualized, refined data mining methodology with adaptations to financial services domain, which (1) represents consolidation of existing body of knowledge, and (2) is validated on the sample of real life data-mining projects. The proposed illustrative case studies approach is based on broad, typical data mining use cases portfolio executed across different geographical regions and business areas of the financial institution.

5 Conclusion

The Systematic Literature Review for the research project (documented in *Section 3)* has demonstrated a few well-developed frameworks for data mining created for wide industry application, which do not cater to specific industry needs such as banking domain. Also, scarce research concerned with this topic in specific financial services domain provides opportunities for new insights and novel findings relevant for both practitioners and academia. *Section 4* proposed project research methodology to: (1) elicit and consolidate domain-specific refinements towards existing data mining methodologies from existing body of knowledge, and (2) to validate against portfolio of real-life data mining projects executed in banking domain. The result of the study will be conceptualized, enhanced data-mining methodology specifically designed to frame and tackle complex data analytics projects in financial services industry.

References

1. Nasdaq Globe Newsire, https://globenewswire.com/news-release/2017/12/20/1267022/0/en/Dresner-Advisory-Services-Publishes-2017-Big-Data-Analytics-Market-Study.html, news feed Dresdner Advisory Services Publishes 2017 Big Data Analytics Market Study, last accessed 2018/04/06
2. Forbes homepage, https://www.forbes.com/sites/louiscolumbus/2017/12/24/53-of-companies-are-adopting-big-data-analytics/#4cf12a2139a1, last accessed 2018/04/06
3. Forrester Consulting: The Future Belongs To Those Who Monetize And Maximize Their Data, Industry report, January 2017, last accessed 2018/04/06
4. Jayasree, V., Balan, R.V.S.: A review on data mining in banking sector. American Journal of Applied Sciences, 10 (10), 1160-1165 (2013)
5. David L. Olson: Data mining in business services. Service Business, 1 (3), pp 181–193 (2007)
6. KDNuggets Homepage, https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html, last accessed 2018/04/07
7. Soltani, Z., Navimipour, N.J.: Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future research. Computers in Human Behavior 61, 667–688 (2016)
8. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD process for extracting useful knowledge from volumes of data. Comminications of the ACM, 39 (11), 27-34 (1996)
9. Chapman, S., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R: CRISP-DM 1.0, step-by-step data mining guide, SPSS Inc. (2000)
10. Morabito, V.: The future of digital business innovation: Trends and practices. 1st edition. Springer International Publishing Switzerland (2016)
11. Rohanizadeha, S., Moghadama, M.: A Proposed Data Mining Methodology and its Application to Industrial Procedures. Journal of Industrial Engineering, 4, 37-50 (2009)

12. Nadali, A.; Kakhky N.E.; Nosratabadi, E. H.: Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system. 3rd International Conference on Electronics Computer Technology (ICECT), 161-165 (2011)
13. García, D.L., Nebot, À. Vellido, A.: Intelligent data analysis approaches to churn as a business problem: a survey. Knowledge and Information Systems 51 (3), 719-774 (2017)
14. Mansingh, G., Osei-Bryson, K.-M., Rao, L., Mills, A.: Application of a data mining process model: A case study- profiling internet banking users in Jamaica. In: AMCIS 2010 Proceedings, Paper 439 (2010)
15. Daihani, D.U., Feblian, D.: Implementation of CRISP-DM model in order to define the sales pipelines of PT X. In: Proceeding of 9th International Seminar on Industrial Engineering and Management, 1-10 (2016)
16. Geng, R., Bose, I., Chen, X.: Prediction of financial distress: An empirical study of listed Chinese companies using data mining. European Journal of Operational Research, 241 (1), 236-247 (2015)
17. Shaikh, J.M.: E-commerce impact: Emerging technology - Electronic auditing. Managerial Auditing Journal, 20 (4), 408-421 (2005)
18. Met, I., Tunali, G., Erkoç, A., Tanrikulu, S.: Branch Efficiency and Location Forecasting Application of Ziraat Bank. Journal of Applied Finance & Banking, 7 (4), 1-13 (2017)
19. Pivk, A., Vasilecas, O., Kalibatiene, D., Rupnik, R.: On approach for the implementation of data mining to business process optimisation in commercial companies. Technological and Economic Development of Economy, 19 (2), 237-256 (2013)
20. Balkan, S., Goul, M.: A portfolio theoretic approach to administering advanced analytics: The case of multi-stage campaign management. In: Proceedings of the 44th Annual Hawaii International Conference on System Sciences, 1-10 (2011)
21. Xin, G., Enjie, D., Hongxia, X.: Promoting data mining methodologies by architecture-level optimizations. In: Proceedings 2009 2nd International Workshop on Knowledge Discovery and Data Mining, WKKD 2009, 179-182 (2009)
22. Tianrui Li, Da Ruan: An extended process model of knowledge discovery in database, Journal of Enterprise Information Management, Vol. 20 Issue: 2, pp. 169-177 (2007)
23. Priebe, T., Markus, S.: Business information modeling: A methodology for data-intensive projects, data science and big data governance. In: Proceedings 2015 IEEE International Conference on Big Data (IEEE Big Data 2015), pp. 2056-2065 (2015)
24. Osei-Bryson, K.-M.: A hybrid decision support framework for generating and selecting causal explanatory regression splines models for information systems research. Information System Frontiers, 17 (4), 845 - 856 (2015)
25. Berendt, B., Preibusch, S.: Better decision support through exploratory discrimination-aware data mining: Foundations and empirical evidence. Artificial Intelligence and Law, 22 (2), 175-209 (2014)
26. Alnoukari, M., Alzoabi, Z., Hanna, S.: Applying Adaptive Software Development (ASD) agile modeling on predictive data mining applications: ASD-DM methodology. In: Proceedings International Symposium on Information Technology, ITSim 2008, 2 (2008)
27. Zwetsloot, M.I, Kuiper, A., Akkerhuisc, S., T., de Koningd, H.: Lean Six Sigma meets data science: Integrating two approaches based on three case studies. Quality Engineering (online journal), DOI: 10.1080/08982112.2018.1434892 (2018)
28. Narayanan, L.V.: Data warehousing and analytics in banking: Implementation. Editor Vadlamani Ravi, Advances in Banking Technology and Management: Impacts of ICT and CRM, 217-231, publisher Information Science Reference, Hershey, New York (2008)
29. Debuse, J.C.W.: Extending data mining methodologies to encompass organizational factors. Systems Research and Behavioral Science, 24 (2), 183-190 (2007)