Adapting the CRISP-DM Data Mining Process: A Case Study in the Financial Services Domain

Veronika Plotnikova, Marlon Dumas, Fredrik Milani

Institute of Computer Science, University of Tartu,
Narva mnt 18, 51009 Tartu, Estonia
firsname.lastname@ut.ee

Abstract. Data mining techniques have gained widespread adoption over the past decades, particularly in the financial services domain. To achieve sustained benefits from these techniques, organizations have adopted standardized processes for managing data mining projects, most notably CRISP-DM. Research has shown that these standardized processes are often not used as prescribed, but instead, they are extended and adapted to address a variety of requirements. To improve the understanding of how standardized data mining processes are extended and adapted in practice, this paper reports on a case study in a financial services organization, aimed at identifying perceived gaps in the CRISP-DM process and characterizing how CRISP-DM is adapted to address these gaps. The case study was conducted based on documentation from a portfolio of data mining projects, complemented by semi-structured interviews with project participants. The results reveal 18 perceived gaps in CRISP-DM alongside their perceived impact and mechanisms employed to address these gaps. The identified gaps are grouped into six categories. The study provides practitioners with a structured set of gaps to be considered when applying CRISP-DM or similar processes in financial services. Also, number of the identified gaps are generic and applicable to other sectors with similar concerns (eg. privacy), such as telecom, e-commerce.

Keywords: data mining · CRISP-DM · case study.

1 Introduction

The use of data mining to support decision making has grown considerably in the past decades. This growth is especially notable in the service industries, such as the financial sector, where the use of data mining has generally become an enterprise-wide practice [1]. In order to ensure that data mining projects consistently deliver their intended outcomes, organisations use standardised processes, such as KDD, SEMMA, and CRISP-DM[1], for managing data mining projects.

[1] KDD - Knowledge Discovery in Databases; SEMMA - Sample, Explore, Modify, Model, and Assess; CRISP-DM - Cross-Industry Process for Data Mining.

These processes are industry agnostic and, thus, do not necessarily fulfil all requirements of specific industry sectors. Therefore, efforts have been made to adapt standard data mining processes for domain-specific requirements [2,3,4]. Although the financial services industry was early to employ data mining techniques, no approach has been proposed to address the specific requirements for data mining processes of this sector [5]. Yet, business actors in this sector adapt and extend existing data mining processes to fit requirements of their data mining projects [5]. This observation suggests that practitioners in the financial sector encounter needs that standardized data mining processes do not satisfy.

In this setting, this research aims to identify perceived gaps in CRISP-DM within the financial sector, to characterize the perceived impact of these gaps, and the mechanisms practitioners deploy to address such gaps. To this end, we conduct a case study in a financial services company where CRISP-DM is recommended and widely used, but not mandated. We studied a collection of data mining projects based on documentation and semi-structured interviews with project stakeholders. We discovered and documented 18 perceived gaps within and across all phases of the CRISP-DM lifecycle, their perceived impact and how practitioners addressed them. Our findings could support experts in applying CRISP-DM or similar standardized processes for data mining projects.

The rest of this paper is structured as follows. Sect. 2 introduces CRISP-DM and related work. This is followed by the presentation of the case study design (Sect. 3) and results (Sect. 4). Sect. 5 discusses the findings and threats to validity. Finally, Section 6 draws conclusions and future work directions.

2 Theoretical Background

Several data mining processes have been proposed by researchers and practitioners, including KDD and CRISP-DM, with the latter regarded as a 'de facto' industry standard [6]. CRISP-DM consists of six phases executed in iterations [6]. The first phase is business understanding including problem definition, scoping, and planning. Phase 2 (data understanding) involves initial data collection, data quality assurance, data exploration, and potential detection and formulation of hypotheses. It is followed by Phase 3 (data preparation), where the final dataset from the raw data is constructed. In Phase 4 (modelling phase) model building techniques are selected and applied. Next, in Phase 5 (model evaluation), findings are assessed and decisions are taken on the basis of these findings. Finally, in the deployment phase, the models are put into use.

CRISP-DM is often adapted to accommodate domain-specific requirements [7]. For example, Niaksu [2] extended CRISP-DM to accommodate requirements of the healthcare domain, such as non-standard datasets, data interoperability, and privacy constraints. Solarte [3] adapted CRISP-DM to address aspects specific for data mining in the industrial engineering domain. These adaptations concern, for instance, defining project roles and stakeholders, analysis of additional data requirements, and selection of data mining techniques according to organisational goals and data requirements. Meanwhile, Marban

et al. [4] propose adaptations specifically targeting the industrial engineering domain by introducing new tasks, steps, and deliverables.

In [5], we conducted a systematic review of adaptations of CRISP-DM in the financial domain. This review identified three types of adaptations: modification, extension, and integration. Modification refers to the situation where adjustments are made at the level of sub-phases, tasks or deliverables. Extension refers to significant changes, including new elements, which affect multiple phases of the process. Lastly, integration refers to the combination of a standardized process (e.g. CRISP-DM) with approaches originating from other domains.

3 Case study design

A case study is an empirical research method aimed at investigating a specific reality within its real-life context [9]. This method is suitable when the defining boundaries between what is studied and its context are unclear [10], which is the case in our research. The case study was conducted according to a detailed protocol[2]. The protocol provides details of the case study design and associated artifacts, including interview questions, steps taken to validate these questions, the procedure used to code the interview responses, etc.

The first step in the case study is to define its objective and research questions. We decomposed our research objectives into three components: perceived gaps, their respective impact, and the adopted workarounds. Accordingly, we defined three research questions: (1) What gaps in CRISP-DM practitioners perceive in the financial services industry? (RQ1); (2) Why do practitioners perceive these gaps, i.e. what is the perceived impact of the identified gaps? (RQ2); (3) How is CRISP-DM adapted to address these gaps (RQ3).

The second step was to define the organisational context and scope of the case study. We sought an organisation that: (1) operates within the financial service industry, (2) has systematically engaged with data mining over the last 3 years, (3) uses CRISP-DM for their data mining projects, and (4) grants access to domain experts and documentation. In line with these requirements, we conducted the case study in the data mining department of a bank operating in Northern Europe. This department acts as a centralised data mining function (Centre of Excellence), responsible for execution of data mining projects across the organisation. The department's portfolio of projects spans over several years and covers several regions and business lines.

We selected a representative subset of projects (Table 1), covering four project types. The first is Business Delivery, i.e., the development of models for different banking products or complex algorithms for analysis of a bank's customers, such as private customers micro-segmentation. The second type is Model Rebuild. These projects share the commonality of rebuilding, retraining, and re-deploying existing models and algorithms. The third is "Proof of Concept" (POC), which explores the use of new analytics techniques, namely

[2] The protocol is available at: `https://figshare.com/s/33c42eda3b19784e8b21`

process mining, for discovering improvement opportunities in lending processes. The fourth, last category is Capability Development, i.e., projects aimed at the development of competencies and tools for repeatable usage in other data mining projects. The selected project in this category concerns exploration of advanced graph analytics methods and development of a visualisation algorithm library. All projects adhered to key phases of CRISP-DM.

Table 1. Projects Characteristics.

No.	*Project Definition*	*Geography*	*Project Type*	*Time span*	*No.of interviews*	*Participants*
1	Product propensity model	1	Business-driven	2018	2	Data Scientist, Project Manager
2	Retail customers micro-segmentation	2,3,4	Business-driven	2017-2019	2	Data Scientist, Project Manager
3	Product propensity model	2,3,4	Business-driven	2018	1	Data Scientist
4	Lending process mining	2,3,4	POC	2019	1	Data Scientist
5	Payments categorization model	2,3,4	Model rebuild	2019	1	Data Scientist
6	Graph analytics library	1	Capabilities development	2019	1	Data Scientist

The third step of the case study was data collection. We approached this step in a two-pronged manner. First, we collected documentation about each project. Second, we conducted semi-structured interviews with data scientists and project managers involved in their execution. The interview questions were derived from the research questions and literature review. The interviews were transcribed (total of 115 pages) and encoded following the method proposed by [11]. The first level coding scheme was derived and refined in iterations. It resulted in combining a set of initial codes (based on reviews and research questions) and codes elicited during coding process. Second level coding, also obtained by an iterative approach, was based on themes that emerged from the analysis. The final coding scheme is available in the case study protocol referenced above.

4 Case study results

In this section, we present the results of our case study. We have structured the results according to the main components of ITIL framework (Information Technology Infrastructure Library). ITIL is industry-agnostic and an accepted approach for management of IT services widely adopted across different business domains [12]. It consists of three main elements: process inputs and outputs, process controls, and process enablers. We view data mining projects as instances of IT delivery and, thereby, encompassed by the scope of ITIL. Therefore, the

results for each of the five phases of CRISP-DM correspond to the main process according to ITIL, which we present first. Next, aspects concerning process controls and enablers related to CRISP-DM lifecycle are described.

4.1 Phase 1: Business Understanding (BU)

The business understanding (BU) phase focuses on identifying business objectives and requirements of the project. Our study shows a significant interdependency between BU and the other phases. All interviewees noted "numerous" iterations and reversals back to the BU phase during the project. One participant expressed that BU *"...had a lot of back and forth with business. It is basically spread over the whole duration of the project*". Another participant highlighted that although such iterations are time-consuming, they enable adequate elicitation and management of business requirements.

The number and degree of iterations vary across projects. Projects with multiple stakeholders reported higher degree of iterations. As noted, *"...the CRISP-DM process, when it is applied to use cases which are unsupervised, especially when there is some kind of segmentation exercise with a lot of different interested business counterparties, it is little bit more difficult to apply [...] because there's [sic] lots of going back to the business discussion, and scoping and Business Understanding part"*. More complex data mining solutions, such as project 2 that required layers of multidimensional calculations, reported more extensive iterations. Exploratory projects (e.g. project 4), required iterations when the obtained results were first applied by end-users. The introduction of new data types, and the discovery of previously unknown data limitations, necessitated reverting to the BU phase for continuous updating and understanding of the requirements, making it essentially intertwined with the data understanding.

Projects that deliver a model as a product (projects 1,3) reported less iterations, but the BU phase was both demanding and crucial for delivery of the right product. One participant underscored BU's significance when defining it as *"one of the most important [...] just a little mistake on the focus and not understanding well what you are targeting [...] you have to start all over again"*. Another interviewee emphasized the necessity of the BU phase and its iterations as *"... you don't really exactly know the scope [...] you might have an idea and you need to present that, but then it can go back and forth a couple of times before you even know the actual population and what kind of products are we looking at..."*. Unexpected iterations are also necessitated by the introduction of new regulations and compliance requirements (projects 2,3).

To summarize, CRISP-DM does not fully reflect the interdependence between BU and the other phases. The main gap (RQ1) of BU is the lack of specific tasks and activities to capture, validate, and refine business requirements. This can cause a (RQ2) mismatch between a business' needs and the outputs of data mining projects. Furthermore, it can lead to missed insights and incorrect inferences. Practitioners commonly address this gap (RQ3) by iterating back to the BU phase in order to align the project outputs with the business needs, regularly eliciting new requirements, and validating existing ones.

4.2 Phases 2-3: Data Understanding (DU) and Preparation (DP)

The Data Understanding (DU) and Data Preparation (DP) phases concentrate on data collection, dataset construction, and data exploration. Here, the interviewees highlighted a recurrent need to iterate between DU and DP, as well as between DU, DP and BU. This need was more emphasized in complex project (project 2) and in both POC projects (4,6). In one of these projects (project 4), the three phases DU, DP and BU, had been merged altogether. The participants indicated that the reason for iterating between DU, DP, and BU is that business requirements (identified in BU) often give rise to new data requirements or refinement of existing ones, and reciprocally, insights derived during DU give rise to observations that are relevant from a business perspective and thus affect the BU phase. Also, data limitations identified during DP may require stakeholders to refine the questions raised during BU.

Data quality issues were continuously detected when working with new data types, methods, techniques, and tools. Such issues required referring back to the DU phase. Furthermore, modelling, analysis, and interpretation of results prompted replacing certain data points or enhancing the initial dataset with new ones. Such changes required an iterative process between DU and other phases. In project 5, though it aimed at rebuilding and releasing updated version of already deployed model, data scientists had to redo the entire process, as interviewee expressed, *"I would say that from one side, we have this Data Understanding from first version, but due to different data preparation tools planned to use, it kind of required pretty much to start from scratch.... it's kind of requires completely different data sources... "*.

We also observed an important adjustment in regard to data privacy. CRISP-DM includes privacy as a sub-activity to the "Assess Situation" task in the context of project requirements elicitation. GDPR[3] strictly regulates personal data processing. Institutions can use privacy preserving technologies to reduce efforts and secure compliance. However, if such solutions are lacking, the data mining projects have to include assessment of data falling under GDPR and consider how to act (anonymize or remove). The interviewees underscored the impact of GDPR requirements throughout entire data mining lifecycle (discussed in *4.6 Lifecycle Gaps, Data Mining Process Enablers*).

Our findings indicate that DU and DP phases have inter-dependencies, data requirement elicitation, and privacy compliance gaps (RQ1). In CRISP-DM, the inter-dependencies between DU/DP and other phases are not addressed, and it does not provide specific tasks for capturing, validating, and refining data requirements throughout data mining lifecycle. Tasks to ensure compliant data processing are also lacking. Such gaps prolong the projects execution (RQ2), and practitioners mitigate them with extensive iterations between the phases. In some cases, DU/DP and BU phases are practically merged into one. The iterations between these phases are also used to address new data requirements, in particular, in regard to data privacy, and to validate existing ones (RQ3).

[3] A recently introduced EU legislation to safeguard customer data.

4.3 Phase 3: Modelling

The Modelling phase focuses on constructing the model after selecting suitable method and technique. The case study showed that this phase was not limited to prototyping only, as stipulated by CRISP-DM. Rather, models (especially, in projects 1,2,4) were developed in iterations, mostly between the modelling and the other phases, such as DU/DP, BU, and Deployment. These iterations were born of the need to improve the models. For instance, the requirements discovered during the BU and deployment phases influenced both which technique to use and model design. One interviewee expressed that *"...there is one quite new dependency or requirement for our side, this is actually latency, because we need to classify or scoring part should happen very fast. ...even here in the Modelling phase, we kinda consider [...] at least kept in mind this latency thing..."*.

For models to be accepted, their outcomes have to satisfy pre-defined performance criteria as measured with evaluation metrics. In contrast to standard CRISP-DM, we observed that model performance metrics and requirements have been adjusted and adapted to business stakeholders' requirements, such as acceptable level of false positives, accuracy, and other criteria, to make model fit real business settings and needs. Projects with complex modelling tasks (projects 1,3,5) adopted a distinct step-up modelling approach. These projects were characterized by first creating a baseline model (benchmark) followed by a set of experiments to identify the best approach to improve the models, i.e., satisfy specific performance metrics, *"... I think we just started off the model, any model just to get start, to get some sort of results to incorporate that in a pipeline [...] once we got one model up and running, then we started to incorporate several other models just to make any comparisons. [...] So, that's what we tried to, a lot of different models and, and we, we wanted [...] the model to be suitable for amount of data that we had, the skewed data, the number of rows and the number of the attributes. And since the data was very skewed and we didn't have that many targets, so to say...then we didn't want that many features and that's, that limited our dataset and in turn that limited which model we would use [...] So we compare these models also by different measures, and the one we ended up with stood out quite significantly."* Also, the Modelling phase explicitly incorporated elements of software development approaches resembling agile processes (project 1), specifically a Test-Driven Development approach (project 6), *"[...] we tried to then develop the actual function, and it could only pass that test if the criteria was met. So, it was a test-based or test driven implementation what we did [...] So even in the code we have all the test cases available...."*.

Practitioners commented on the restrictive notion that the outcome of the Modelling phase should be a model. They discussed situations where the results of the modelling phase were various interpretations of the model and different analytical metrics (projects 2,4,6). To this end, interviewees reported on both applying actual modelling techniques and executing algorithm-based data processing (e.g. using Natural Language Processing techniques) or experimenting with various process representations (project 4). For one of the POC projects, it was noted, *"[...] it can be quite questionable what we consider as the modelling*

here...process map, or the more formal process model in the process model language but as next steps in more advanced process mining projects, there could also be, additional models, for prediction and detection and so on. So, the process mining project can end up as a quite big project where many different types of modelling are involved." Thus, the Modelling phase can be defined as 'multi-modelling' with the set of unsupervised and supervised modelling outcomes.

In summary, the Modelling phase of CRISP-DM does not cater to needs of developing, improving, and refining models in data mining lifecycle. Furthermore, explicit guidelines how to iterate between phases, in particular, the BU, DU/DP, and Deployment, are lacking. Refinement of existing requirements and capturing new requirements, which originate from the Modelling phase and other phases, is not supported. Finally, CRISP-DM is restrictive with respect to modelling outcomes, not catering to 'multi-modelling', unsupervised modelling, and specialized modelling techniques (RQ1). These gaps can prolong data mining projects execution and increase the risk of mismatch between business need and outcome (RQ2). Commonly, practitioners address these gaps by employing an iterative and metric-driven modelling process, frequent iterations with other phases, and calibration with requirements from other phases. Also, tasks and activities are introduced to deliver various analytical outcomes ('multi-modelling') and to accommodate use of various techniques (RQ3).

4.4 Phase 4: Evaluation

The Evaluation phase is concerned with quality assessment and confirming that the business objectives of the projects are met. Majority of interviewed practitioners underscored the importance of validating and testing the models in a real usage scenario setting. While CRISP-DM prescribes assessing if the models meets business objectives, the 'how' is not discussed. As noted, *"... Crisp-DM should be updated specifically on the step of Evaluation to include how to test the model in business industry. I mean taking into account real scenarios, and there should be a list of steps in there. Which actually we have figured, figured out these steps [...] in an empirical way."* CRISP-DM prescribes a two-step validation. The first is a technical model validation which is conducted in the Modelling phase and considers metrics such as accuracy. The second step assesses if the models meet the business objectives which is conducted in the Evaluation phase. However, practitioners conducted these validations concurrently (projects 1-4). Stakeholders evaluated the models by considering the technical aspects, such as accuracy, and if the models are meaningful in a business setting, as noted, *"...important thing is that we like to think the evaluation through and really measure the thing that we want to measure and, and also not rely on only one measure, but can see the results from different angles."*

For unsupervised models (project 2,4), we noted that the evaluation was primarily subjective. The consideration was given to how meaningful the results were for the business, how they could be interpreted, and to what extent actions could be taken based on the results. Thus, suitability and model usage, i.e., business sensibility, were the basis for model evaluation. As one participant noted

that *"... it is difficult to define some sort of quality measure for this kind of unsupervised result other than, well, actionability and future usage because we could have a quality measures for the clustering itself that just means that the clustering, cluster is distinct, but they don't mean that clusters are actionable for business and there can be non-distinct groups which on the other hand are interesting for business. So, there was, in this case it's kind of [...] technical quality measures are not necessarily suitable for a practical, practical quality."*

Our findings show that the Evaluation phase of CRISP-DM does not specify how models can be assessed to determine if they meet the business objectives. In particular, there are gaps related to assessing and interpreting the models in their business context. Also, the separation of technical and business evaluation, as outlined by CRISP-DM, can be problematic (RQ1). These gaps can lead to poor model performance in real settings and reduce actionability (RQ2). Practitioners address these deficiencies by piloting models in actual business settings (RQ3).

4.5 Phase 5: Deployment

The Deployment phase is concerned with implementing data mining project outcomes to ensure they are available and serve business needs of end-users. In CRISP-DM, deployment tasks and activities are first considered in this phase. However, we observed the necessity to address deployment strategy and elicitation of deployment requirements earlier (project 1,3,5). As one participant noted, *"... when we develop a model, we think about what's important to us..and the business side, it could be interested in to see the results in a different way, or to include different columns or some things... So I understood after that process that one should have the Deployment phase already on your mind when [making] up the model, also, more or less from the very start,.....and to see the actual data that the business will pick up, and in the way they will pick it up...."*.

In addition, CRISP-DM does not address specification of deployment requirements well. Therefore, practitioners adapt reference process, especially to elicit requirements towards the format of the deployed solution and its end-usage in business contexts (projects 1,2,4). As noted, *"... the results were meant to be used on a daily basis by frontline people ... so, in this sense there are different levels of results that are needed ...there have to be some very simple KPIs and some very simple visualizations that don't need this more advanced process knowledge and understandable for everyone. So, that was something that we didn't know at the beginning that actually we need to report it not only to the Business Development department and process managers, but also to really frontline people ..."*. Thus, the deployment phase can involve calibrating requirements to adapt models for their ongoing end-usage.

We also observed that the practitioners adopted a different deployment process compared to CRISP-DM, which focuses on the deployment plan rather than implementation. Also, participants reported using a wider range of deployment formats. For instance, projects based on unsupervised models might not require deployment at all as their purpose is discovery of features and interpreting said

features within the context of a specific business problem (project 4). In contrast, algorithm library was reported as deployed solution in project 6.

Our main finding from the Deployment phase is that CRISP-DM stipulates the elicitation of deployment requirements too late. The often-needed calibration of deployment requirements elicited in earlier phases of CRISP-DM, is not covered. Also, this phase, as stipulated by CRISP-DM, assumes a restrictive stance and is not open to different deployment strategies. Lastly, CRISP-DM focuses on producing a deployment plan, but does not address implementation itself (RQ1). These gaps can prolong project execution, and increase the risk of a mismatch between the project outcome and the intended end-usage, i.e., the business need (RQ2). Practitioners address these gaps by considering deployment scenarios and eliciting deployment requirements early on, as well as extending the Deployment phase to include implementation tasks (RQ3).

4.6 Lifecycle Gaps

We also identified gaps that concern the whole CRISP-DM lifecycle rather than a specific phase thereof. Below, we present these gaps, organized according to two key pillars of the ITIL framework: process controls and enablers.

Data Mining Process Controls ITIL identifies five process controls: process documentation, process owners, policy, objectives, and feedback. Furthermore, in the context of IT delivery projects, it specifies process owners, process quality measurement, and reporting as key controls [14]. In our case study, practitioners highlighted three main aspects of data mining project controls – governance, quality, and compliance – which are in line with ITIL.

Our analysis shows that the practitioners have adopted elements of agile practices into their data mining life cycles. This is explicit in recent projects where 2-week sprints have been used, requirements are captured in epics, teams have daily stand-ups, sprint planning, and retrospectives. Practitioners also noted that CRISP-DM does not support the agility required for some data mining projects. As one practitioner stated, there is a *"...flaw in this methodology. It [...] tells you that it's dynamic, but it does not tell you what to do, [...] when do you have to go back to step 1, and how to do it faster"*.

Another aspect mentioned in relation to governance, is roles and responsibilities of both internal and external stakeholders. The importance of stakeholder management was emphasized. When stakeholders were not identified early on, *" [...] there was a lot of additional tasks...and secondly, each part [...] could have different stakeholders."* The stakeholders' understanding of the business problem and what the team delivering data mining projects can achieve, matter. For instance, in project 4, it was noted that *"...it can be two ways like either we present to the stakeholder a solution for a potential problem they could have, meaning we have to, to sell an idea. Or the other way around, they already have a problem that they have identified very clearly, and then they come looking for a solution, that we will provide. So I think as stakeholders understand more what we do it's more than the second one because they know we can help as they data scientists."* External stakeholders (customers) are not included in the validation

of the data mining solutions. Nevertheless, they can, potentially, contribute to improving the quality of the results, as was expressed,*"... the end customer, the client [...] The only thing we can measure currently is if they accept or not accept a product or service, but there will be a very interesting [to] involve,...to have them ask why they took a consumer loan or why didn't they, and in what way? [...] get a lot more information about the end user. [...] just to understand what, what is the actual driver. We can only read black and white on data, but we don't know if something else motivates them to do certain choices."*

Another aspect observed is that of quality. In particular, we noted the evolution of adopting quality assurance mechanisms in the data mining process. These measures are expressed as the implementation of a formal peer-review process that is integrated in the project execution. Such quality assurances are visible as checkpoints – both in the daily work routines, and via review-based checklists. These quality assurance measures serve to validate five key aspects of data mining projects: (1) privacy-compliant data processing, (2) project scope, business goals, and data mining target, (3) input dataset quality, (4) modelling method application, and (5) code quality (software development controls).

Lastly, participants highlighted compliance, in particular in regard to GDPR, as an example of external requirements that impact data mining projects. GDPR has introduced a set of privacy-compliance requirements , limited the usage of certain data types and how final results can be used. For instance, in project 5, the company required customers to express GDPR consent, resulting in a limited number of customers using the data mining-based solution.

Data Mining Process Enablers The data mining process enablers, in this context, refer to capabilities required for the organisation to be able to execute data mining projects. These enablers concern aspects that support projects that follow, to different extent, the CRISP-DM process. The capabilities discussed are related to data, data mining code, tools, infrastructure, technology, and organisational factors.

Data quality, understood as reliability, persistence, and stability, was reported as crucial for all projects. Practitioners expressed that more important than tools is *"... it's about the data because you have great tools, but if the quality of the data is not good enough, then, it doesn't matter, so to me this is like the most important thing"*. Another practitioner stated that *"[...] the thing people often are referring to is if they have like a lot of nulls maybe, or like missing fields in the data. So that would be one side of the quality, but that's just according to me lack of data, that wouldn't be [...] really a concern in my mind. Quality of data would be that it's reliable and that the sources are stable and not changing. So that would be quite important, and I guess you just have to incorporate a lot of sanity checks in order to trust the sources."*

We also observed a consensus that data should be made readily available for (self-service) usage and as underscored by one interviewee, "good databases are the key". Data consistency and completeness across various data sources is another critical aspect reported in, for instance, project 4 (specialized process mining project). In this case, setting up correct workflow registration in source

systems was a prerequisite to obtain acceptable data. In addition, it was emphasized that self-reported data was subject to biases and interpretations, thus it may be less reliable and, therefore, should be used with caution. The practitioners also referred to the quality of data mining project code. Its importance was chiefly noted for projects 5 and 6, in the context of scalability and optimization.

Available tools and infrastructure that ensure adequate prototyping, scaling, and deployment are regarded as pre-requisite capabilities for all projects. Limitations to operate with large datasets and difficulties in applying methods and algorithms were cited as consequences of tools and infrastructure limitations in projects 2 and 4. *"[...] we had a lot of impediments on the technology side, in the sense that we were using quite big amount of data, and we were doing the analysis on local computer so there were some restrictions or issues with data size, sometimes the data size actually didn't allow to compute clustering quality measures.. this data size also put restrictions on the algorithms that You can use, for example, it was not possible to use many different clustering methods because they just do not scale so we were somewhat limited in choice of algorithms...."*.

Furthermore, a critical requirement mentioned in regard to tools and infrastructure, was the ability to support automated, repeatable, and reproducible data mining deployments, *"I would say it's important that we have, that entire pipeline, the infrastructure for the entire pipeline would be prioritized. So, so like going from start to end, maybe in a very thin or narrow manner in the sense that we might not have that many different systems or programs to use, but that we can deploy going also fast to, to market."* Also, interviewees reported that available tools, platforms, and infrastructure had an impact on the choice of model design and language used. For instance, for project 6 *"....the main concern was to create a library that is usable inside our team [...] And we even considered a different programming language like Scala, as it could be more efficient. But since most of the end users, which are basically our team members were Python programmers, we decided to go for Python library [...]... of course we could just do Python, but we wanted the solution to be also scalable for, for large graphs. And that's why we chose Spark...that depended on the...business requirements, and business requirements indicated that we need to work on large datasets... "*.

Finally, organisational factors, such as data-driven decision-making culture and maturity have been referred to as crucial elements enabling adoption of data mining solutions in business practice (project 2). Interviewees referred to 'push-pull' paradigm whereby stakeholders actively 'pushed' for solution initially, and with active participation have converted to 'pull'. Further, education of stakeholders to support data-driven decision-making culture thus transforming organisation towards 'pull' paradigm has been emphasized and reported.

To summarize, we found that the CRISP-DM life cycle has gaps related to governance, quality assurance, and external compliance management. Also, CRISP-DM has deficiencies associated with data quality management and stakeholder management (RQ1). These gaps prolong project execution, increase a risk of mismatch between project outputs and business needs, and negatively impact business value realization (RQ2). We found that these gaps are filled by adopt-

ing agile software development practices, specifically Test-Driven Development (TDD), Scrum ceremonies and Scrum boards, via regular interaction with business stakeholders across all CRISP-DM phases, and via integration of regulatory compliance requirements into the data mining process (RQ3).

5 Discussion and Threats to Validity

In this section, we discuss the gaps identified both in the phases and entire CRISP-DM data mining lifecycle (Figure 1 below). We group them, based on their characteristics, into six distinct categories and, for each category, discuss the gaps (RQ1), perceived impact (RQ2), and how practitioners adapt CRISP-DM to mitigate the gaps (RQ3).We conclude this section with threats to validity.

The first category of gaps, *Inter-dependency Gaps (1,3,6)*, concerns the lack of iterations between different CRISP-DM phases. As practitioners noted, these gaps lead to missed insights, skewed interpretations, and an increased risk of incorrect inferences. If the *Inter-dependency gaps* are not addressed, an increased effort in the form of re-work and repeated activity cycles is required, resulting in prolonged project execution. Practitioners address this gap by making numerous iterations between the CRISP-DM phases or merging them.

The *Requirement Gaps (2,4,7,8,13)* relate to the lack of tasks for validation and modification of existing requirements and elicitation of new ones, and they are present in all CRISP-DM phases except for the Evaluation phase. Such deficiencies increase the risk of a mismatch between the outputs of the data mining project and the business needs. Practitioners reduce their impact by adding validation and calibration steps and by iteratively eliciting new requirements. Practitioners also adopt software development support tools, methods and incorporate elements from agile practices in their data mining projects.

The *Inter-dependency* and *Requirements Gaps* constitute the lion share of the gaps. These gaps stem from the largely sequential structure of CRISP-DM. Although iterations between the phases are possible, the procedural structure of CRISP-DM prescribes a linear approach where each phase is dependent on deliverables from the previous phase.

The third category, *Universality Gaps (9,10,14)* concerns a lack of support for various analytical outcomes, unsupervised and specialized techniques, as well as deployment formats. This category has been discovered for the Modelling and Deployment phases. Our results indicate that the standard CRISP-DM is, at times, overly specialized. In the case of the Modelling phase, it is restrictive in supporting standard, supervised, modelling techniques and associated data mining outcomes. For deployment, CRISP-DM does not provide tasks for implementation and associated technical requirements. These gaps lead to an increased risk of mismatch between data mining outcomes and business needs. Practitioners address these gaps by adding tasks to support unsupervised, specialized models' development and the delivery of various non-modelling analytical outcomes ('multi-modelling') as well as different deployment formats.

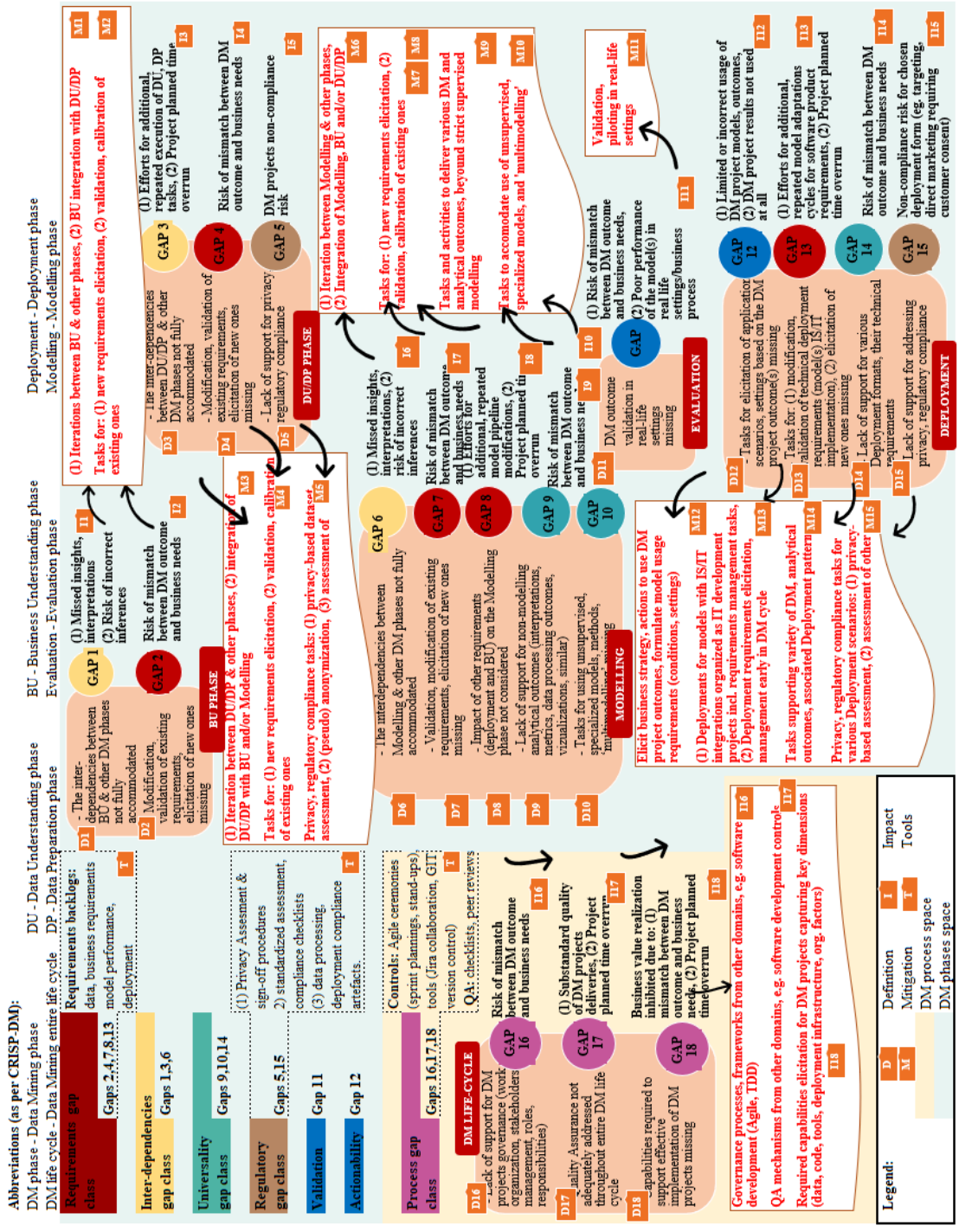

Fig. 1. Identified Gaps Mindmap

Further, we discovered *Validation Gap (11)* and *Actionability Gap (12)*, which concern the Evaluation and Deployment phases respectively. These gaps refer to a lack of support for piloting models in real-life settings. Thus, if models are not validated in practical settings, they are likely to exhibit poor performance when deployed. Also, CRISP-DM does not address elicitation of scenarios for model application. The lack of a model usage strategy, and an insufficient understanding of the models' application settings, leads to limited or incorrect usage of the models, or result in models not being used at all ('producing models to the shelf' scenario). These gaps were filled by extensively using pilots in real-life settings, as well as addressing the actionability of the created analytical and model assets. The gaps of Universality, Validation, and Actionability stem from CRISP-DM's over-emphasis on classical data mining and supervised machine learning modelling. Data mining itself is regarded as mostly a modelling exercise, rather than addressing business problems or opportunities using data.

The sixth category of gaps, *Privacy and Regulatory Compliance Gaps (5,15)* , deals with externally imposed restrictions. These gaps are related to the DU, DP, and Deployment phases. CRISP-DM does not, generally, cater for privacy and compliance and, in particular, lacks tasks to address the processing of customer data. The impact of these gaps can result in non-compliance. Thus, practitioners have established standardized privacy risk assessments, adopted compliance procedures, and checklists. These gaps stem from the fact that CRISP-DM was developed over two decades ago, when a different regulatory environment existed.

We also identified *Process Gaps (16,17,18)* which do not concern a specific phase but, rather, the entire data mining life cycle. These gaps encompass data mining process controls, quality assurance, and critical process enablers required for the effective execution of data mining projects. We note that CRISP-DM does not address projects governance aspects such as work organisation, stakeholders, roles, and responsibilities. Further, procedures for quality assurance are not provided for, and required key capabilities, i.e., for data, code, tools, infrastructure and organisational factors, are not taken into consideration. These gaps can reduce data mining project effectiveness and inhibit their business value realization. Practitioners mitigate them by incorporating quality assurance peer-reviews into the execution of data mining projects. Process gaps appear as CRISP-DM only partially incorporates project management activities, and does not take broader organisational and technical aspects for project management into consideration. Thus, process controls and enablers needed to support multiple data mining projects on organizational level continuously are not addressed.

When conducting case study research, there are threats to validity that should be considered, particularly, construct validity, external validity, and reliability [9]. Construct validity refers to the extent to which what is studied corresponds to what is defined and intended and defined to be studied. In our study, the interview method can be a source of construct validity risk. We mitigated this threat by including internal validity checkpoints (reconfirming questions, answers summaries with interviewee) to verify interviewee's understanding of the questions. We also confirmed the contents (interview transcripts) with the

participants. External validity concerns the extent by which the findings can be generalized. Case study approach has inherent limitation of generalizability, and further studies will be required to assert the generalizability of our findings. Finally, reliability concerns the level of dependency between the researcher and study results. We have tackled this risk by adopting iterative research process with regular validations within our research group. We have also reduced reliability threats by using triangulation of projects documentation and interviews. We also maintained appropriate chain of evidence keeping track of the research materials and process and in that way ensuring replicability of the research steps and results.

6 Conclusion

This paper presented a case study in a financial services organization aimed at identifying perceived gaps in the CRISP-DM lifecycle, their perceived impact, and workarounds to mitigate these gaps. The case study involved a representative subset of 6 projects within this company. Data was collected from project documentation and via semi-structured interviews with project participants. By combining these data sources, we identified 18 gaps in the CRISP-DM data mining process, as perceived by projects stakeholders. For each gap, the study elicited its potential impact and the adaptations that the interviewed project participants have made to the CRISP-DM process in order to address them.

The identified gaps are spread across all phases of the CRISP-DM lifecycle. About half of the gaps relate to *Requirements* management or insufficient recognition of *Inter-dependencies* between CRISP-DM phases. These findings confirm those discussed in [13], which highlighted that, in practice, there are many pathways for navigating across the tasks and phases of the CRISP-DM lifecycle. Our study also highlighted that CRISP-DM does not explicitly address *Privacy and Regulatory Compliance* issues and that it does not explicitly tackle *Validation and Actionability* concerns. Finally, we found a category of gaps (*Process Gaps*) arising from the fact that CRISP-DM does not fully consider the wider organisational and technical context of a data mining project.

The study also identified five adaptations: (1) inclusion of explicit iterations between phases or merging of phases, (2) addition of tasks to address requirements elicitation and management concerns, (3) addition of 'piloting' tasks for validation, (4) combination of CRISP-DM with IT development project management practices, and (5) addition of quality assurance mechanisms. A direction for future work is to define an extension of CRISP-DM that addresses the identified gaps, taking as a basis the adaptations identified in this study as well as insights from adaptations of CRISP-DM in other domains. We also foresee that the gaps in the *Process* class could be addressed by combining CRISP-DM with the ITIL framework. Another direction for future work is to conduct similar case studies in other organisations, both within the financial sector and in other services sectors, such as telecom, where similar concerns (privacy, compliance, risk management) arise.

References

1. Forbes Homepage, https://www.forbes.com/sites/louiscolumbus/2017/12/24/53-of-companies-are-adopting-big-data-analytics (2017), accessed Jan 30, 2021
2. Niaksu, O.: CRISP data mining methodology extension for medical domain. Baltic Journal of Modern Computing **3**(2), 92 (2015)
3. Solarte, J.: A Proposed Data Mining Methodology and its Application to Industrial Engineering. PhD Thesis, University of Tennessee (2002)
4. Marban, O., Mariscal, G., Menasalvas, E., Segovia, J.: An engineering approach to data mining projects. In: International Conference on Intelligent Data Engineering and Automated Learning, Springer, 578–588 (2007)
5. Plotnikova, V., Dumas M., Milani, F.P.: Data Mining Methodologies in the Banking Domain: A Systematic Literature Review. In: International Conference on Business Informatics Research, 104-118, Springer, Cham (2019).
6. Marban, O., Mariscal, G., Segovia, J.: A data mining and knowledge discovery process model. Data Mining and Knowledge Discovery in Real Life Applications, edited by P. Julio and K. Adem, pp. 438-453, Paris, I-Tech, Vienna, Austria (2009)
7. Plotnikova, V., Dumas M., Milani, F.P.: Adaptations of data mining methodologies: a systematic literature review. PeerJ Computer Science **6** (2020): e267
8. Chapman, P., 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)
9. Runeson, P., Host, M., Rainer A., Regnell, B.: Case study research in software engineering: Guidelines and examples. John Wiley & Sons, (2012)
10. Yin, R. K.: Case study research and applications: Design and methods. Sage publications, (2017)
11. Saldana, J.: The coding manual for qualitative researchers. Sage publications, (2015)
12. McNaughton, B., Ray, P., Lewis, L: Designing an evaluation framework for IT service management. Information & Management, **47**:4, 219–225 (2010)
13. Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J.H., Kull, M., Lachiche, N., Quintana, M. J. R., Flach, P. A.: CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, (2019)
14. AXELOS Limited. ITIL® Foundation, ITIL 4 Edition. TSO (The Stationery Office), (2019).