AGILE MINING
- A NOVEL DATA MINING PROCESS FOR INDUSTRY PRACTICE BASED ON AGILE METHODS AND VISUALIZATION

Xiao Zhu

Advanced Analytics Institute

School of Software

Faculty of Engineering and Information Technology

University of Technology Sydney

This dissertation is submitted for the degree of Master by Research

November 2017

Publication List during Master Degree Study:

X. Zhu, G. Xu. Applying Visual Analytics on Traditional Data Mining Process: Quick Prototype, Simple Expertise Transformation, and Better Interpretation. *International Conference on Enterprise Systems: Advances in Enterprise Systems* (ES), 2016.

S. Liu, G. Xu, X. Zhu. Towards Simplified Insurance Application via Sparse Questionnaire Optimization. *International Conference on Behavioral, Economic, and Socio-Cultural Computing* (BESC), 2017.

DECLARATION

This dissertation is the result of my work and includes nothing, which is the outcome of work done in collaboration except where specifically indicated in the text. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Production Note:
Signed:______ Signature removed prior to publication.

Date:______ 15/11/2017

Xiao Zhu

ABSTRACT

Current standard data mining processes like CRoss-Industry Standard Process for Data mining (CRISP-DM) are vulnerable to frequent change of customer requirement. Meanwhile, Stakeholders might not acquire sufficient understanding to generate business value from analytic results due to a lack of intelligible explanatory stage. These two cases repeatedly happen on those companies which are inexperienced in data mining practice. Towards this issue, Agile Mining, a refined CRISP-DM based data mining (DM) process, is proposed to address these two friction points between current data mining processes and inexperienced industry practitioners. By merging agile methods into CRISP-DM, Agile Mining process achieves a requirement changing friendly data mining environment for inexperienced companies. Moreover, this Agile Mining transforms traditional analytic-oriented evaluation to business-oriented visualization-based evaluation. In the case study, two industrial data mining projects are used to illustrate the application of this new data mining process and its advantages.

Keyword: Data Mining, Data Mining Standard Process, CRISP-DM, Visualization.

CONTENTS

List of Abbreviations and Acronyms

CRISP-DM: Cross Industry Standard Process for Data Mining

AM: Agile Mining

RR: Requirement Reconfirming

PR: Performance Refining

RF: Random Forest

LR: Logistic Regression

SVM: Support Vector Machine

NB: Naïve Bayes

1 INTRODUCTION

1.1 Research Background

Current prevailing data mining processes are vulnerable to handle requirement changes in cooperative projects with industry[1].

Requirement change in industry data mining project is frequently deemed as a failure of early work including stakeholder engagement or business understanding during industry data mining practice [2], though same issues have been accepted as a natural and recurring process in software developing domain [3]. The motivation of requirement changing comes along with stakeholder's growing awareness towards data mining practice and should be considered in the design of standard data mining process for industry project. In contrast, current prevailing data mining processes attach deficient priority on handling requirement change actively, which is regarded as incompatible with the project purpose of industry stakeholders. As a consequence, industry data mining projects, especially those from inexperienced practitioners, frequently encounter a dilemma: either project deliverable does not match stakeholder's expectation, which will damage customer satisfaction; or proposed project schedule suffers severe delay, which extremely expands the project cost.

Moreover, analyst always fails on transferring meaningful interpretation and inside the logic of the data mining project deliverables to stakeholders, which as a result, might limit the productivity and restrain the application of valuable data mining findings. The main reason for that is the bias of project success criterion. Analysts always put major efforts on the accuracy of calculation or performance of the model, while stakeholders only trust the result which they can understand [4]. After deployment, this doubt significantly restrains the productivity that the analytic outcome should provide to business. This misunderstanding also violates industry data mining project's commercial success.

1.2 Related Work

1.2.1 Data Mining Process

Data mining can be categorized in different perspective: 1. from data angle, it can be supervised, semi-supervised or unsupervised; 2. from the purpose of data mining practice, it can be predictive analysis or descriptive analysis; 3. from the angle of the applied algorithm, it can be classification, regression, clustering, pattern mining, etc. Given all practice mentioned above should be undertaken in a certain guided way to be manageable and traceable, a collection of data mining standard process[5] has been proposed such as CRISP-DM and SEMMA.

It is important to understand the whole approach of the how data mining can be conducted before one starts running algorithms to discover interesting patterns from data. Blindly applying data mining model on input data without well-organized application also called "data dredging" [6], could always produce meaningless or unintelligible patterns, which may eventually fail a project. In comparison, a formalized, well-defined data mining application could frequently discover valid, understandable and novel patterns. Moreover, a common data mining framework could also help as a measurable roadmap for people to follow during project planning and implementing.

1.2.2 Agile Methods

Agile Method is a set of novel software development method that arose in the 1990s. It attracts widespread attention for its capability to handle rapid requirement changing. Agile Method includes varied approaches with their terms such as eXtreme Programming (XP) and Scrum.[7]

Compared with non-Agile software development method, Agile Method emphasizes a close interaction between the development team and business experts, frequently shippable result, face-to-face communication, and changing-adaptive coding skill. Agile Method is occasionally considered as undisciplined and planless. But in fact, it attaches more adaptability rather than foreseeability onto its development process.

Characters of Agile Method are transferable on to data mining process, while so far no literature indicates any data mining process applying Agile Method to solve the requirement changing issue.

1.2.3 Visualization in Data Mining

Visualization tools are well applied in data mining domain, especially in data understanding, white-box model explanation and result evaluating stage. Benefit by its intuitive expression, visualization takes a major work for communication between business and analysts. There are also different usages of visualizations. Given analytic purpose, there is explanatory and exploratory visualization. Given its objective types, visualization can also be procedure visualization and result visualization.

Despite visualization’s convenience, in traditional visualization tools, analyst-oriented presenting formats and statistical indicators makes visualization result hard to absorb for business. Given the complexity of model itself, the visualization result might be the only way for business to understand how the model is functioning. As a result of this lack of trust, deployed deliverable may not be as productive as it is proposed, no matter how accurate the result is. To solve the insufficient business-analyst interaction during data mining process, a collection of business-oriented visualization tools is presented in this thesis.

1.3 Research Issue

Current prevailing data mining standard processes has two main shortages:

1. Current prevailing data mining standard processes are not genuinely iterative. Neither conditions to start or finish iteration nor fixed iteration length is denoted. In most cases, the non-recurring process is still preferred to achieve the best result if possible, in which iteration will occur only if the current job cannot proceed without a refined pre-construction work. No shippable outcomes will be presented as transition results after one repeat during iteration. As a result, business-accessible-deliverables can only be generated at a late stage with margin space for alteration, which disables business to update requirement in time. This type of process is shown in figure 1.1.
2. Current prevailing data mining standard processes attach insufficient emphasis on interaction with business. After business understanding phase, most of the work is supposed to be finished by analyst only, with no specific step for interpreting the work to business. This may restrain business understanding about the model functionality and character. Another side effect of the lack of interaction is that business cannot manage the progress of performance refining work to achieve a balance between accuracy and cost from stakeholder's interest.

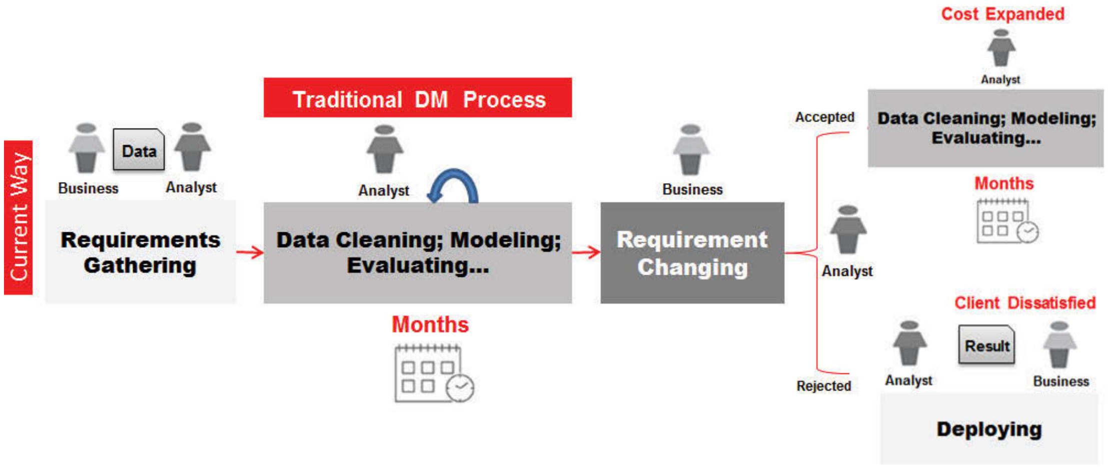

Figure 1.1 Common Scenario in Traditional CRISP-DM Projects

1.4 Research Novelty

Aiming to make current CRISP-DM process more adaptable for requirement changing and more manageable for business, the Agile Mining (AM) process is proposed. This AM process is based on CRISP-DM, integrated with advantages taken from agile method and illustrated in business-friendly visual style.

This novel data mining process firstly acknowledges requirement change as an acceptable factor during industry project. It takes advantages from Agile Method in software development to minimize the waste of requirement changing by adding a Requirement Reconfirming stage before major analysis starts.

Also, AM defines fixed-length iteration in its Performance Refining stage to provide a frequent transitional deliverable, which makes data mining project more manageable for business. This character enables business to do a trade-off between outcome’s accuracy and cost, specifically for those projects with a secondary accuracy demand. Moreover, the frequent transitional deliverables help business understand the model’s function and character, which could add their trust and confidence when the model is applied in real business cases and therefore make the most of model’s productivity. These two novelties constitute AM’s major advantage.

1.5 Structure of This Paper

Chapter 1 introduces the nature of requirement changing during data mining industry project and a lack of adaptability of current data mining processes to solve this issue.

Chapter 2 reviews current prevailing data mining processes, agile methods and novel visualization tools.

Chapter 3 elaborates the AM process in stage and step level.

Chapter 4 describes how AM can solve the requirement change during data mining project and enhance business manageability.

Chapter 5 describes how AM can improve insufficient interaction with business with a collection of novel visualization of data mining.

Chapter 6 summarizes this thesis.

2 Literature Review

2.1 Data Mining Process

Before creation of standard data mining process, data mining industry was at the chasm [8]between early market and main stream market [9]. No one could assure about its commercial success. If the early adopters fail with their data mining projects, they will usually assert that data mining does not work instead of blaming their incompetence in using data mining properly [10].

The establishment and application of data mining have been regarded as common good among all domains in recent years. Data mining technology has significantly evolved concerning both the advancement of the analytic algorithm and the range of result application. In contrast, attention paid to methods to systematically implement a data mining project is apparently much less than other aspects. The job of data analyst contains varied tacit knowledge, which is majorly empirical and difficult to be articulated as reducible steps [11]. A variety of data mining areas have revealed this informality, including standard of data mining process, creation of training and testing data, data mining model building approaches to model reporting [12]. In this review, the focus is limited to standard of data mining process.

A standard model of data mining process can benefit in multiple fields. It can bridge a platform for all parties to discuss data mining and add common understanding on essential data mining issues, especially for business people. Moreover, an engineering structure will be integrated into the data mining process. By that, business customers could understand a similar story by different tool or service providers more easily. Practically, business expectations become more reasonable regarding project proceeding and outcomes. A standard model also makes mining tools selection easier, as the model offers a step-wise frame to compare the compatibility and performance among different vendors.

The standard data mining process can benefit analysts as well. For those inexperienced practitioners, this process could be used as a roadmap with detailed guidance for each step. For masters in data mining, it can be used as a checklist to ensure all necessary work is finished before starting a new phase. Most of all, a standard process is a foundation for discussing and documenting the result during data mining. By setting measurable milestones, different tools and different people with diverse skills and backgrounds are connected to form an efficient and effective project [10].

Despite that data mining models are considered as very specific towards its application domain, they can be differentiated by the feature that whether the industrial issue is taken into account or not [13]. Meanwhile, academic models can be made easily applicable on industrial issue by minor transformation and so as industrial models. In this literature review, processes put on comparison are limited to those only have been renowned in both published papers and real industrial practice.

2.1.1 CRISP-DM

The CRISP-DM process was initially conceived in late 1996 by DaimlerChrysler, SPSS and NCR, three “veterans” of the young and immature data mining market [14]. The name CRISP-DM is coined as CRoss-Industry Standard Process for Data mining. CRISP-DM consists in a cycle that contains six phases (figure 2.1):

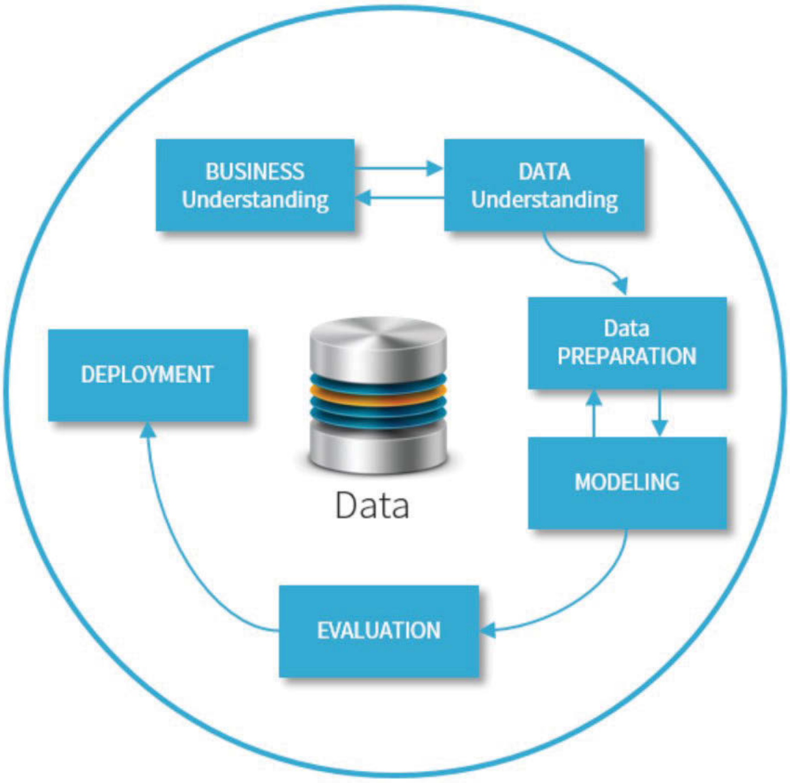

Figure 2.1 CRISP-DM Life Cycle [10]

Sequence in this cycle is not fixed. Based on the performance of the outcome of one phase, it always requires going backward to improve the performance in specific tasks to meeting the demand of moving to next step. Recurring dependencies between phases are demonstrated by the arrows. The outer circle symbolizes the nature of the data mining practice [10].

The six phases are outlined as below. [15]

1. Business Understanding: this starting phase aims at understanding the project goal from business view, transforming the goal into identified data mining problems, and forming a preliminary project plan to solve these problems.

This phase comprises specific tasks including determining business objectives, assessing situation, determining data mining goals and produce project plan.

2. Data Understanding: analysts get familiar with the data through a series operation on business original database including initial data collection and combination, data quality assessment, initial insight generation and initial hypothesis creation.

This phase comprises specific tasks including collecting initial data, describing data, exploring data and verifying data quality.

3. Data Preparation: this phase covers all kinds of operations to convert raw data to suitable input for modeling in next phase.

This phase comprises all common data preprocessing operations including selecting data, cleaning data, constructing data, integrating data and formatting data.

4. Modeling: applicable modeling algorithms will be performed on input data. Parameter calibration will be conducted with specific requirement on input data form which needs moving back to data preparation phase.

This phase comprises specific tasks including selecting modeling technique, generating test design, building model and assessing model from analytic view.

5. Evaluation: conduct a thorough review about the construction of model and result' capability to resolve proposed project objectives from business view. The decision about next move (whether deploy the result or not) is based on the evaluation result.

This phase comprises specific tasks including evaluating results, reviewing process and determining next steps.

6. Deployment: integrate project outcome with business operating system. Organize and present the generated insights properly to suit the business decision-making process.

This phase comprises specific tasks including planning deployment, planning monitoring and maintenance, producing final report and reviewing project.

CRISP-DM is demonstrated as guidance in various research projects: evaluating heating, ventilation, and air-conditioning systems' performance [16]; medical data predictive analysis [17]; vending machine data analysis [18]; centralized model evaluation of collaborative data mining [19]and text analysis [20].

CRISP-DM is illustrated as data mining approach in following industry practice as well: warranty claims analysis about vehicles [21]; marketing strategy generating automation [22]; analysis of population-intensive building construction [23]; data-driven air pollution conservation analysis in Asia [24]; quality diagnostic system for package manufacturing [25].

2.1.2 SEMMA

SAS Institute defines SEMMA (sampling, exploration, manipulation, modeling and assessment) as “a logical organization of the functional toolset of SAS enterprise miner for carrying out the core tasks of data mining” [26]. Compared with CRISP-DM, which provides a comprehensive step-wise guidance, SEMMA attaches more concentration on exploratory statistical analysis and visual-driven data mining solutions [27].

SEMMA’s five phases are explained as figure 2.2.

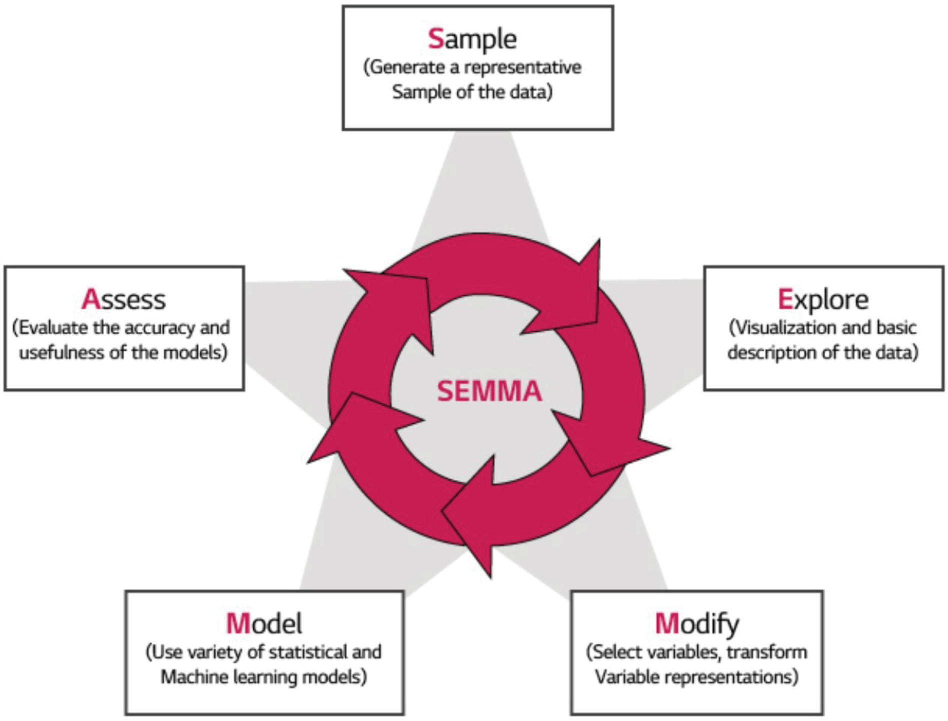
Sample
(Generate a representative
Sample of the data)
Explore
(Visualization and basic
description of the data)
Modify
(Select variables, transform
Variable representations)
Model
(Use variety of statistical and
Machine learning models)
Assess
(Evaluate the accuracy and
usefulness of the models)
SEMMA

Figure 2.2 SEMMA Cycle [28]

1. Sample: A subset containing all necessary information will be sampled from large raw data for quickly preliminary manipulation. This phase is considered as optional [28].

2. Explore: conduct exploratory analysis on sampled subset for getting familiar with the whole data. Initial hypotheses and interesting trends will also be pointed out.

3. Modify: contains varied kinds of data manipulation. By creating, selecting, and transforming the feature, the raw data is qualified to be fed into modeling tools.

4. Model: perform different modeling techniques on the modified input data to automatically search the matched model for the desired outcome.

5. Assess: assess the performance of modeling outcome and evaluate its reliability to solve the analytic problem.

SEMMA offers users freely choices on the data mining tools. However SAS is linked to the guidance in each phase to help users proceeding data mining practice. This limitation restrains its dissemination to certain extent [28].

2.1.3 KDD

The KDD process, as presented in figure 2.3, is “the process of using data mining methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, subsampling, and transformation of the database”[28]. This process is viewed as five stages, symbolized in figure 2.3:

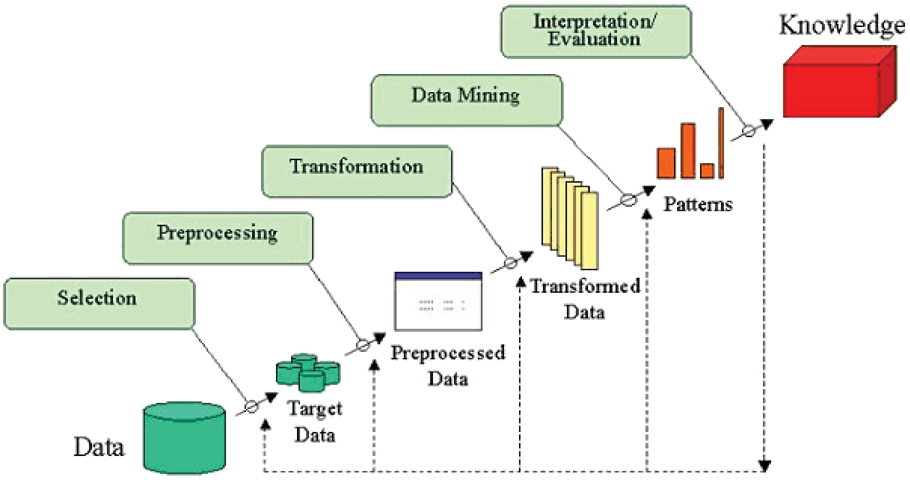

Figure 2.3 The Five Stages of KDD [29]

1. Selection: combine all necessary parts of raw dataset or sample a subset of the raw data to perform the knowledge discovery activity.

2. Preprocessing: clean and preprocess the selected dataset to make it in a structured format.

3. Transformation: transform the data to meet the knowledge discovering tool's demand. Usually the operation contains reducing the number of features and features deriving.

4. Data Mining: discover interesting patterns and represent the result in a reasonable form regarding specific objectives, e.g. prediction.

5. Interpretation/Evaluation: properly interpret and evaluate the mined result.

2.1.4 Comparison of KDD, SEMMA and CRISP-DM

Table 2-1 demonstrates the procedures conducting same function in the three data mining processes. Both KDD and CRISP-DM contain all procedure in SEMMA. Meanwhile, business understanding in CRISP-DM has an identical purpose with Pre-KDD procedures in KDD process, while Deployment in CRISP-DM summarizes the Post-KDD procedures. In the integrality view, CRISP-DM and KDD process is more comprehensive than SEMMA.

Table 2-1 Procedure Comparison of KDD, SEMMA and CRISP-DM

CRISP-DM	SEMMA	KDD
Business Understanding	N/A	N/A
Data Understanding	Sample	Selection
	Explore	Pre Processing
Data Preparation	Modify	Transformation
Modeling	Model	Data Mining
Evaluation	Assessment	Interpretation/Evaluation
Deployment	N/A	N/A

KDnuggets, a leading website of data mining, has conducted two polls about the usage of different data mining processes on 2007 and 2014 respectively. Among the 200 answers from poll 2007, 42 percent used CRISP-DM, 19 percent used their own model, 13 percent used SEMMA, 7.3 percent used KDD Process, 5.3 percent used their organization's model and 4 percent used some other model or no model. In the poll 2014, 43 percent of the 200 respondents used CRISP-DM, 27.5 percent used their model, 8.5 percent used SEMMA, 8 percent used other models, 7.5 percent used KDD Process and 3.5 percent used their organization's specific model. The results demonstrate CRISP-DM model's advantage on its overwhelmingly wide-spread usage.

Despite the outstanding popularity of CRISP-DM, it is recognized by data scientists that there are still friction points for the various phases include:

The business understanding phase is based on the hypothesis that business people know exactly what they are expecting and describe it to analysts. In practice, business expectations are always as vague as 'make smarter decisions by using data'[30] for lack of awareness about what data mining can provide. Only by update the knowledge background of business can the data mining practice proceed smoothly.

Another noteworthy issue is that frequently project objectives and project requirements are not generated from same group of people. Usually project objectives are raised by the management, instead of the end-user who can better tailor the requirements. Consequently, end-users always have to post-rationalize the model function. This mismatch also restrains model's productivity.

A common issue with the evaluation phase is the deficiency on business interpretation about the analytic model. According to the nine laws of data mining, it states that 'Data mining amplifies perception in the business domain'[31]. To bridge the background gap of business and analytics, interactive reporting methods such as visualization and dashboards reporting should be included as part of model evaluation output. These methods can benefit both the expression of analytic reporting such as what the model does and the operational reporting as how the model impacts the business [32].

2.2 Agile Method for Software Development

Because of the continuous change of requirement during software development, the developing period is frequently extended. As a consequence, project budget soars week by week. This common case makes delivering customer satisfied software in time more and more difficult.

Traditional methods consider that as long as developers put enough efforts on early stage, customer requirement is able to be defined well at the start. Cost can also be reduced for later on requirement change [33]. Meanwhile, Traditional methods view requirement change as mistakes from early work stage. Therefore, traditional processes always try to avoid taking a requirement change into the developing workflow.

Change of external environment can always cause crucial requirement change, which is inevitable and cannot be eliminated by developers. This kind of change can be from a lack of awareness of customer at early stage. Developers will do several attempts to ensure customer expectation for the final system. The change can also come from the exchange between different tasks' priority. Refusing these requirements could be viewed as irresponsible for customer. If the developing team is not experienced for a new domain project, the process will also fluctuate and unpredictable change will frequently happen.

To address the issue mentioned above, a proper solution should focus on reducing the cost of each change, instead of avoiding changes. In light of that, a more flexible kind of software developing methods has been raised, including Scrum, Extreme Programming (XP), Crystal method, Lean Development, etc.. These methods are collectively named Agile Methods. Compared to traditional methods, Agile Methods are designed to be adaptive to these changes. Traditional methods value process formalization and completeness very much, while Agile Methods try maintain balance between process flexibility and complexity. This enables Agile Methods gain relatively satisfied result with fewer steps of process.

2.2.1 Character of Agile Method

Seventeen software developers came to the Snowbird resort in Utah in February 2001 for a lightweight development methods seminar. Then the Manifesto for Agile Software Development was published by the seventeen. Through their collective experience of software developing and for the purpose to help other developers, the common value for Agile Methods was announced:[34]

- **Individuals and Interactions** over processes and tools
- **Working Software** over comprehensive documentation
- **Customer Collaboration** over contract negotiation
- **Responding to Change** over following a plan

In this list, though process, tools, documentation, contract and plan play important roles during development, it is the flexibility towards inevitable changes that should be regarded as top priority. Frequent internal communication would be beneficial for sharing information and quick response to change. Time-consuming document reading and writing work could also be simplified; prioritizing current working software could evaluate the working progress more specifically and easily; paying attention to customer means to absorb customer represent into developing team and make full use of customer’s professional experience and domain knowledge to guide the team to meet real business needs, which is better than just negotiating the contract and sticking to it;

Responding to change is valued because along with development proceeding, initial plan will be inappropriate from time to time. Among above four aspects, traditional methods value more on the formers, while Agile Methods value more on the latter.

Apart from that, traditional methods tend to provide complete and trivial guidance, while Agile Methods offer only core discipline that can be globally applied. By that Agile Methods expect developers could creatively solve problems in dynamic practical situations. Agile Methods believe that individual creativity of development team is the only approach to deal with increasingly complicated developing environment.

Individuals and interactions emphasize the importance of Self-organization and motivation, which is represented as interactions including co-location and pair programming. Working software is more preferable and actionable than progress reporting for Customer. Customer collaboration acknowledges that development team should not expect all requirements are fully collected at the beginning. A consistent engagement with customer is essential for keeping development on the right track. Responding to change focus on rapid responsiveness for dynamic requirement change to keep development in steady progress. [35].

Agile Methods adopt short iteration circle ranging from two to six weeks. If a developing team takes six months to get customer feedback, it certainly has not applied Agile Methods.

The content of iteration will be modified based on customer feedback from last iteration. Priority of functions can also be reordered dynamically. Compared with traditional methods using technical terms in function description, Agile Methods prefer describing functions in customer language, which is beneficial for productive communication. The customer can adjust priorities, or even delete or add function after iteration as well.

2.2.2 Extreme Programming

Extreme programming (XP) is a software development methodology. It aims to refine software quality and responding speed to better handle customer requirements change. As a type of agile methods [33, 36], it suggests using frequent releases in short development cycles to enhance developers' productivity. Meanwhile, refer checkpoints among big milestones to better adapt software to customer requirement change.

It is regarded to match small developing team in a requirement-frequently-changing background. XP originated from SmallTalk domain. At the early 1990s, after a series industry projects, Kent Beck, Ron Jeffries and Ward Cunningham [37] changed their minds about traditional development methods. They believed that the process of software development should be dynamic, flexible and individual-oriented.

To address requirement change emerged during developing process, Agile Methods manage to minimize its cost in global view.[33] Consider XP, for example, which requires developing team:

1. Issue first demo within 1-3 weeks for acquiring quick feedback.
2. Adopt programming code as easy as possible for the convenience of future alternation.
3. Keep refining system design to achieve better-developing structure.
4. Run tests continuously to reduce cost for later tests and modification.
5. Customer participates whole process and plays important role.

XP has four essential values: **Communication**, **feedback**, **simplicity** and **courage**. Communication stands for interaction within developing team and with customer. Feedback requires results from quick unit test and function test. Simplicity asks for focusing on finding the solution with minimum cost for certain scenario. Courage indicates to actively seek easier solution and bravely improve design and code quality. Furthermore, over ten practical rules have formed based on these four values, some of which are already existed but frequently ignored. XP brought them up once again but attached more emphasis on them.

Paying extreme attention to test is XP’s most important feature and the foundation during development. XP believes requirement and test are closely correlated. By adding test at early stage, code that inconsistent to customer’s requirement can be removed from the function module. In light of that, XP requires developers writing test case before writing the functional code. Meanwhile, XP requires automatic testing tools to conduct test. When a code module passes the test, both testing code and functional code are integrated into holistic system. Newly added code should be able to pass all the existing systemic tests and its test. Moreover, customer will be asked to take an acceptance test after iteration to see whether current system is competent enough or tell developers what else should be added.

Other components of extreme programming contains: pair programming and intensive code review, global test for unit, avoiding prior feature programming until necessary, a flat management structure, code simplicity and clarity, expecting requirement changes emerging along with project progress and customer awareness growth, close collaboration among development team and intimate communication with customer[34]. XP is named from the notion that the advantage of traditional software engineering practices is taken to "extreme" levels. Code reviews are considered, for example, as a beneficial practice. Once it is taken to the extreme, the code can be reviewed continuously, such as the practice of pair programming. Figure 2.4 shows a typical XP loop.

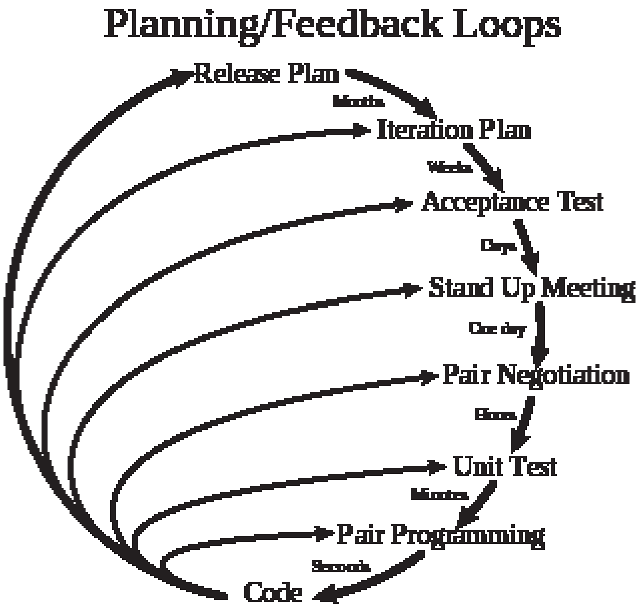

Figure 2.4 XP's Planning and Feedback Loops

XP takes a gradually iterative process. For example, arrange around 12 iterations during the development period, with iterative length taking 1-3 weeks. Every time after iteration, an executable transition version will be released to customer. XP requires developing team to focus on current iteration. All design is made to meet current function without consideration for others. As a consequence, XP becomes one of the most popular methods among Agile Methods for its strict discipline and high adaptability.

2.2.3 Scrum Method

The term "Scrum" is a metaphor from rugby. During a scrum, forwards from two teams will cross their arms together and move towards the same direction. Meanwhile, dynamic information including strategy, position, weather condition, etc. are all shared among the whole team. Decisions are also made quickly from short discussions.

As a software development method, Scrum is viewed as “an iterative and incremental agile software development framework for managing product development” [38, 39]. The aim of Scrum is regarded as "a flexible, holistic product development strategy where a development team works as a unit to reach a common goal" [40]. It proposes opposite assumptions of the "traditional, sequential approach" [40] for software development, which empowers self-organization for developers by advocating offline co-location or intensive online collaboration among the whole team. Face-to-face meeting in regular pace with specific rules are also suggested.

Scrum’s core rationale is the mutual acknowledgement that what customer want or need will always be changed (often called requirements volatility [41]) and this challenge is often unpredictable. Consequently, an overall thorough plan is not suitable. In light of that, Scrum applies an evidence-based empirical approach, which accepts that the problem cannot be fully understood or identified at early stage. Instead of trying to make a big plan, it should focus on exploiting development team's capability to deliver quickly, to rapidly respond to dynamic requirement, and be better adaptive towards evolving techniques and changes in business domain [42].

Sprint defines the basic unit of development in Scrum. The sprint stands for a timed effort, which is restricted to a specific duration [42]. The length is fixed prior for each sprint and is normally within one month, averagely fortnight[43].

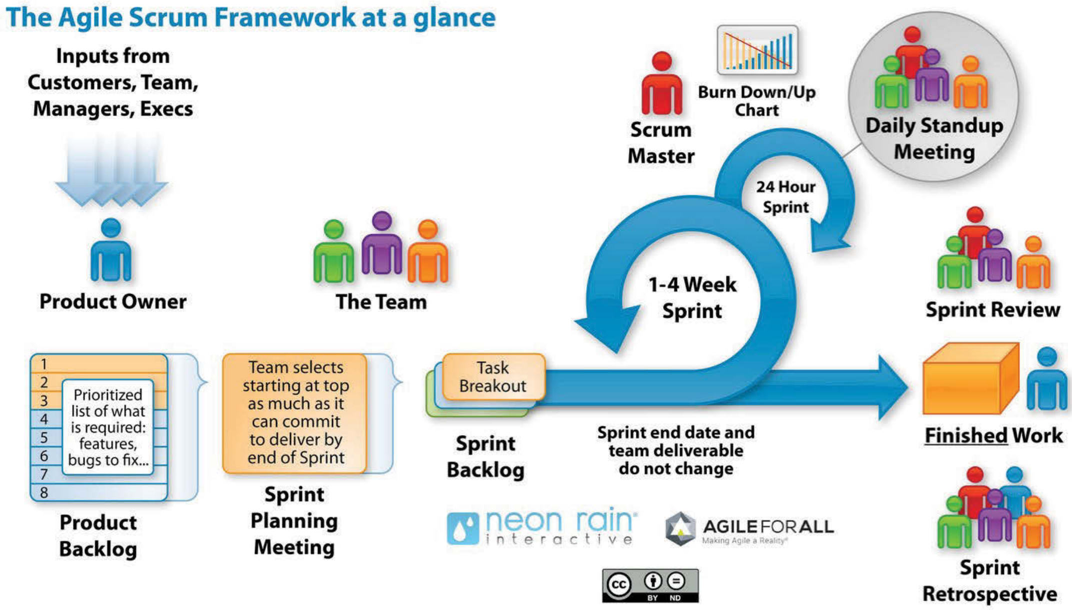

Figure 2.5 Agile Scrum Framework

As is shown in figure2.5, each sprint starts with a sprint planning event that aims to define a sprint backlog, identify the work for the sprint, and make an estimated forecast for the sprint goal. Each sprint ends with a sprint review and sprint retrospective [44], that reviews progress to show to stakeholders and identify lessons and improvements for the next sprints.

Scrum emphasizes working product at the end of the sprint that is done. In the case of software, this likely includes that the software has been fully integrated, tested and documented, and is potentially shippable [45].

Apart from the sprint and backlog, Scrum has another system to measure customer requirement. This requirement engineering system is called backlog grooming, with its structure shown in figure 2.6.

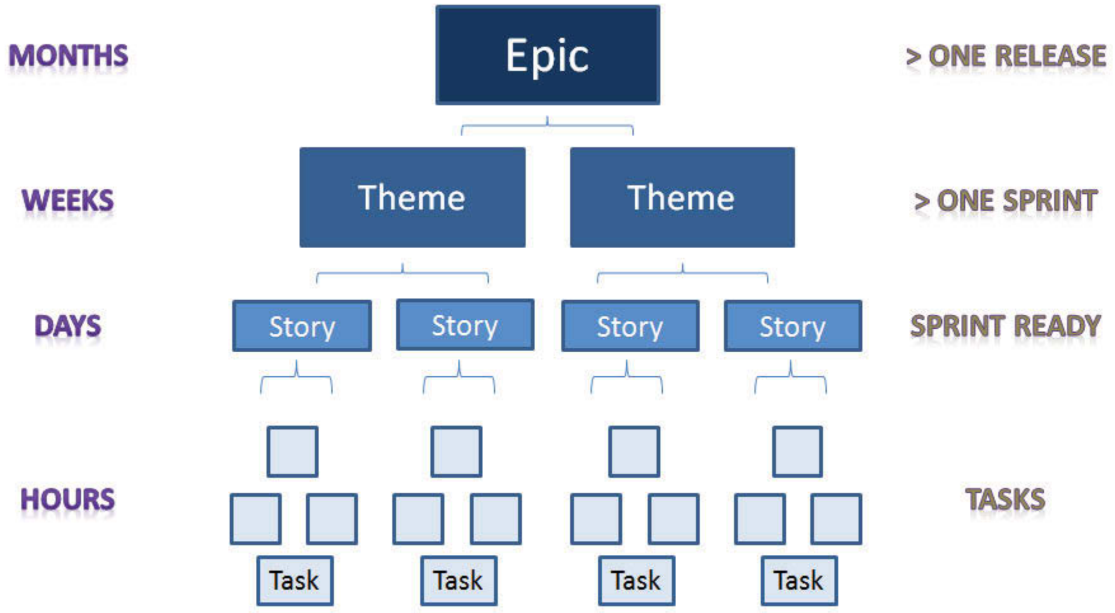

Figure 2.6 Backlog Grooming in Scrum

User stories are a simple format that can be used to state business values and is suitable for a variety of product backlog item, especially features. User stories are designed to help both business and developers understand the needs.

Traditional waterfall model considers requirements as necessary, not negotiable, independent and refined already. In contrast, Scrum method considers that the details of customer requirements are discussed in the ongoing dialogue during the development process. Requirement refining should be provided at the same time when developing team start building the function. The format of user story could be: “as a certain role, I want to achieve certain goals so that I can enjoy certain benefit.” The requirement’s abstract level could be Epic, Theme and User story. A noticeable character of story is that a story is not a task which needs elaboration for technical details. When writing a story, all the task details should be avoided to add into requirement description. Figure 2.7 shows a hierarchical view of customer requirement engineering in Scrum.

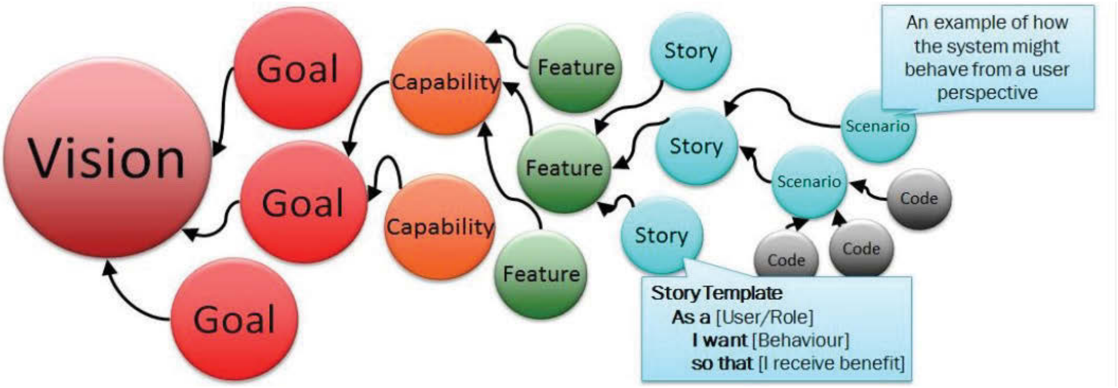

Figure 2.7 Hierarchy of Customer Requirement

A good story should follow the “invest” rules [46, 47]:

Independent: The user story should be independent, at least should be mutually loosely coupled; a high degree of dependence on the story will complicate the estimating, ordering and planning work; this is not to suggest that dependencies should be eliminated, however in writing stories, dependencies should be avoided.

Negotiable: The details of the story are negotiable instead of unalterable contract; negotiated help to avoid misleading and accusations. If the story can be negotiated, the developer will not say, "If you want it, write it to the document," the business people will not say: "You obviously did not understand the requirements of the document, because you develop something wrong ." A noticeable issue for this is it should not be the product owner to tell the developing team about how to realize the story, which is a common case to violate the negotiable rule. If the realization method becomes not negotiable, developing team will waste their creativity.

Valuable: Story should be valuable for both business and user the story is. Otherwise, it will be a waste.

Estimable: Normally there are two reasons that developing team cannot estimate the size of the story: the story is too large or too vague to be estimated; team lacks accumulated experience. For the former, the team has to work with business to break it into a more easily estimated story; if it is the latter, it is necessary to obtain information through some form of exploratory activity.

Small: Size of the story should be reasonable and alike.

Testable: A story either passes the test or fails. "Testable" means that the story has a corresponding set of quality acceptance criteria. If there is no testable standard at the end of each sprint, developing team would not able to know whether they have finished the story or not. Also, because these tests tend to provide important story details, so the team is likely to need these tests to estimate the size of the story.

2.2.4 Comparison of Scrum and XP

In Sum, XP provides detailed guidance for practical development process. In comparison, scrum tends to offer a general framework for project management tools. With regards to the application field, XP focuses on the developing and testing stage, while Scrum puts more emphases on requirement engineering. Below are comparisons between XP and Scrum:

1. Scrum and XP teams work iteratively, but Scrum's cycle is generally from two weeks to a month, while XP cycle is one or two weeks.
2. Scrum team in a sprint is not subject to any task changes. Once the sprint planning fulfilled the task, until the sprint end, the developing team will not accept any changes. In comparison, in XP iteration, if the new feature with the original-feature-alike size and size difference is not too much, new feature can still be used to replace the original feature under the premise that the original feature has not yet started.
3. XP team will work in strict accordance with the priority of the task. All tasks are prioritized by the customer, and the team is required to work at that priority. In contrast, product owner in the Scrum team also prioritizes the backlog; the scrum team members can also decide in which order they will do to complete all the tasks. It is rare to see a scrum team does not choose to start with highest priority entry. Occasionally, scrum team will choose to do a slightly lower

priority as a starter, such as some of the task is not just the beginning of the implementation of the sprint, although its priority is high. Or who is suitable for a task, is doing other work. In this situation, the priority will be adjusted.

4. Scrum does not define any specific practical methods. It just provides a practical framework for developing team. But XP does provide practical details such as test-driven development, automated testing, pair programming, simple design, refactoring, etc.

2.3 Visualization in Data Mining

Visualization is the technology of computer graphics and image processing. By transform the data into graphics or images displayed on the screen, and interactive processing theory, methods and techniques, visualization can support various fields such as computer graphics, image processing, computer vision, computer-aided design and other fields with its ability of data representation, data processing, decision support and a series of integrated technology.

Data mining is the process of checking and analyzing data to obtain potentially useful information that is implicit in the data. That is through complex statistical analysis and modeling techniques applied on existing data to reveal patterns and relationships that are hidden in the organization's dataset. These patterns and relationships are difficult to find by ordinary methods.

Getting the right information at the right time is critical for making right decisions. Visualization can not only reflect the current situation in real time but also predict future trends. To obtain effective information from massive data, efficient data mining techniques must be used. It is regarded as the most effective solution by using a variety of visual methods to express multi-dimensional data, and then use the human cognitive ability to large-scale multi-dimensional data set for data mining and knowledge discovery. It can be said that visualization integrated data mining is a new trend in data mining technology. This chapter attempts to combine data mining technology with data visualization technology to explore the meaning, content, application status and development trend of data mining visualization.

2.3.1 Definition of Data Mining Visualization

The so-called "visualization" refers to the process of visualizing and forming images of objective things in human mind, which is a mental process. Visualization improves people's ability to observe things and the formation of the overall concept. Visualization results facilitate the memory and understanding of people, while its advantages on information processing and expression cannot be replaced by other methods.

Visualization technology is a form and a process in which people are accustomed to accepting graphics, images, and supplemented by information processing techniques that will be perceived, imagined, reasoned, integrated and abstracted. Visualization is not only an image re-projection of reality but also an integration of trends, knowledge and information. Visualization technology is not only used to express static knowledge but also can be used to dynamically describe the development of objective objects and evolutionary knowledge acquisition.

Visualization of data mining now does not yet have a well-defined definition, as defined in data mining visualization refers to the use of visualization techniques in some invisible or abstract things that represent visible graphics or images. Visualization here refers to the use of computers to create visual images, to understand the large number of complex data to help.

Visualization can be used in many ways. It can help user visually understand the complex patterns of multidimensional data. By observing the existence of data in multiple dimensions and multiple graphic forms, it is possible to visually and quickly reveal data trends. Data mining can also help to examine the data before modeling and verify the results of its data mining tools. Also, visualization also plays an important role in data pattern discovery.

2.3.2 Objective of Data Mining Visualization

According to the visualized content, data mining visualization is divided into the following three categories:

1. Data visualization: The data in the database and the data warehouse can be viewed as having different granularity or different levels of abstraction, or as a combination of different attributes and dimensions. Data can be described in a variety of visual ways, such as box-like, three-dimensional cube, data distribution charts, curves, surfaces, connection diagrams, or any combination of the above methods to complete the visualization of data structure. Traditional geometric methods such as point graphs, line graphs, histograms, pie charts and so on. Selection of visualization method should be varied along with different purposes of data analysis.

   Discrete Plot may be the most widely used visualization tool in data mining. It helps people to analyze data clustering, observe the distribution of data, with or without singularity. For data with only two or three attributes, either a planar or stereoscopic representation can be used. And for a dataset with multiple attributes, a discrete dot matrix is used.

   Discrete Point Matrix is an extension of the function of discrete points, which can represent multidimensional data distribution. Each unit of the matrix is based on a two-dimensional representation of the data. Due to symmetry, the discrete dot matrix can be drawn half (usually the lower left corner).

   Parallel Coordinate: It is another way of representing multidimensional data. If the data has M dimensions, there are M parallel lines with each line representing one dimension. Each point is represented by a broken line, and the intersection of the polyline and the coordinate axis indicates the value of the point in that dimension.

2. Data mining process visualization: The visualization of the data mining process is illustrated in a visual way in which the user can visually see each procedure of data mining, such as which data warehouse or database current data is extracted

from; how the data is extracted; how the selected data is preprocessed and mined, what method is selected in the modeling process; how the results are stored and displayed.

The visualization of the knowledge discovery process makes the knowledge discovery process easy to understand and contributes to the application of knowledge. At present, visualization based on visual knowledge refers to the use of visual knowledge discovery tool through the visualization of the operation process to complete the knowledge of spatial data discovery. This mainly reflected in two aspects: visualization of the knowledge discovery system interface and visualization of the knowledge discovery process of navigation; as well as visualized queries and descriptions.

3. Model visualization: Not every user is an expert in data mining. Users are not expected to know what kind of information can be found before data mining. Some models are difficult to understand, such as neural network. Therefore, the data mining model must be transformed into the most natural representation. Only in this way can we understand the model more effectively and then take action. Also, some models get very large results, such as association rules. It is possible that a data mining gets a lot of rules, and how to find information from these rules is a tricky problem.

   Model visualization can be considered from two aspects: make the model output visualized, which make model expressed in a meaningful way; or allow interaction to enable the user to manipulate the model.

At present, the integration of knowledge discovery and data visualization technology has aroused the attention of researchers, more and more people began this research. It is accepted that the key to successfully combining knowledge discovery with information visualization is to create a shared data model that able to intuitively guide users to select toolsets (including mining tools and visualization). On this basis, a task-driven interactive intelligent knowledge discovery system could be built. This allows the user to monitor and guide the discovery process at any time based on the intermediate results obtained by the computer so that the computer's mining process is combined into the user's decision-making process until a satisfactory conclusion is reached.

2.4 Summary of Literature Review

This chapter reviews the definition, content and application of standard data mining process, agile method and visualization.

In standard data mining process part, CRISP-DM, SEMMA and KDD process are introduced. And CRISP-DM is regarded as more adaptable for various industrial projects.

In agile method part, shared common values for agile method and prevailing methods including Scrum and XP are introduced.

In visualization in data mining part, its definition, content, application and future scope are introduced.

3 A Novel Data Mining Process Based on Agile Methods and Visualization

3.1 Introduction to Agile Mining Process

In light of CRISP-DM and Agile methods, this Agile Mining process for industry analytic project contains two major iterative stages as figure 3.1, which are：

1. Requirement Reconfirming (RR)stage, and;
2. Performance Refining (PR) stage.

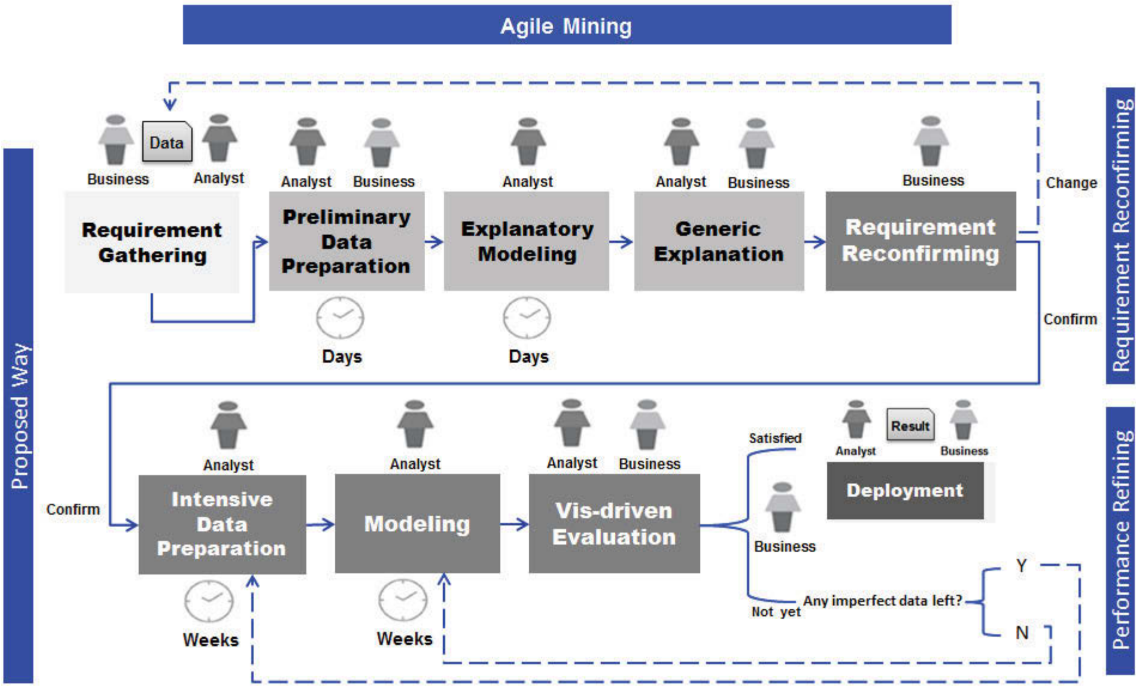

Figure 3.1 Agile Mining Process

RR and PR stage are two iterative processes with clear requirements on iteration's start/end. Iteration is measured by sprint, which represents a completed, fixed-length iteration as in Scrum.

By the end of each sprint, an explanatory (RR stage) or measurable (PR stage) result will be provided to business. This regular progress generation could enlighten business increasing understanding about the purpose of current analysis and then adjust their expectation and further plan.

The iterative attempt on seeking the best suitable requirement for project objective matches business people's growing awareness of the capability of data mining. By providing room for business to optimize the requirement at an acceptable cost, a clearer project goal will be shared in further analysis. The fixed-length iteration also provides analyst team a framework for shipping software on a regular cadence. Consider milestone, the end of a sprint, for example, which frequently comes, bringing with business and also analyst a feeling of tangible progress with each cycle that focuses and energizes everyone, as the "continuous inspiration" for the win [48]. Short iterations also reinforce the importance of good estimation and fast feedback from tests, both of which are believed as recurring struggles in traditional process.

3.2 Requirement Reconfirming

In RR stage, a sprint contains following five steps:

1. **Requirement Gathering**: Collect stakeholder's description about project purpose. The description could be stories from business view but should be convertible to individual analytic objectives. After that, these objectives will be prioritized by project manager and a business representative from cooperating company to form an initial project backlog. This business representative should be the end-user of this project deliverable, so that essential concerns from practical situation will be taken into consideration and attract higher priority. The analyst then will take project backlog items (PBIs) with top priority and divide it into several logic steps as backlog grooming in Scrum.

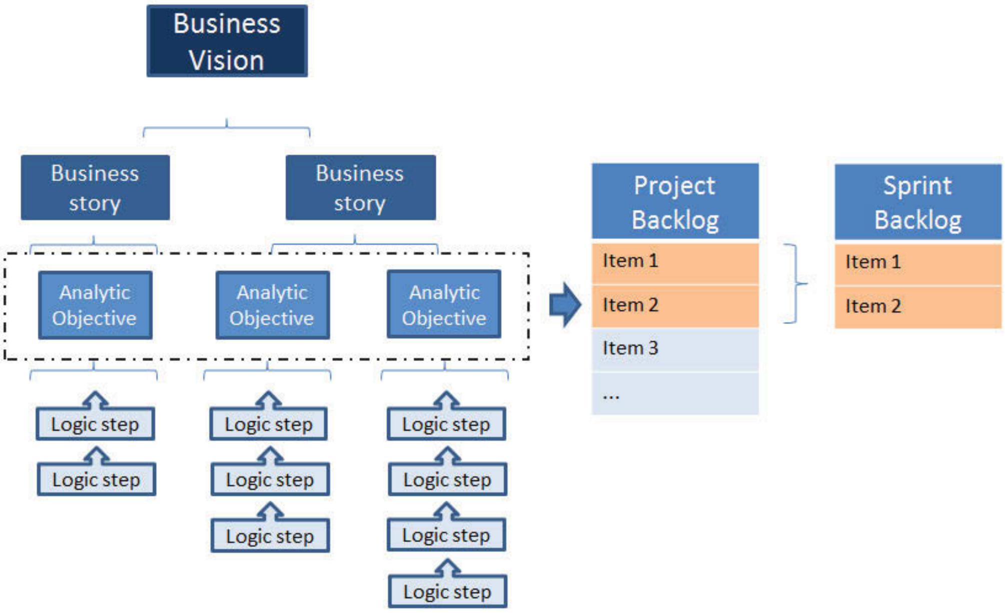

Figure 3.2 Requirement Gathering Process

After that, depending on the amount and complexity of PBI's logic steps, one or several PBIs will be selected into sprint backlog as the maximum workload that analyst team can handle within one sprint. A logic step could be an operation to achieve analytic objectives, such as alternation on existing features (join, transform, etc.), derivation of new features (group, segment, etc.), or result calculation.

2. **Preliminary Data Preparation**: connect relevant dataset from different sources. Based on each logic step of current sprint backlog item, all necessary dataset will be linked to form a structured data set as needed in corresponding logic step, which later on will become input of the model. This data preparation step should focus on testing the connectivity and sufficiency of required data. Rather than going deep through and trying to get most of the data prepared at once traditionally, this step should collect a relatively small portion with best data quality only. So that it can avoid complicated data cleaning work and efficiently validate the accessibility of current analytic objective. Meanwhile, Analyst should get business representative fully involved in this step to take advantage of their experience to locate target dataset and differentiate similar features.
3. **Explanatory Modeling**: perform explainable model on the pilot dataset. After getting all necessary features structured as input dataset, this step use white-box model such as Linear Regression, Logistic Regression or Decision Tree to reveal the insight generation process. This Explanatory Modeling step emphasizes the pattern of how business insight could be generated rather than business insight itself. In light of that, the intelligibility of model attracts top priority rather than model accuracy. The analyst should consider the best-understood model from business view in this step.
4. **Generic Explanation**: interpret each logic step and result from model to stakeholders. The analyst should brief the logic steps in Preliminary Data Preparation step to stakeholders to get them familiar with the input data. Then give generic explanation of the insight generation pattern in Explanatory

Modeling step to stakeholder. Business representative is expected to contribute to model explanation using business language and domain expertise.

5. **Requirement Reconfirming**: Stakeholders and business representative decide whether provided objective is desired for project purpose. Based on explanation from previous step, stakeholders could gain clearer understanding about how their business requirement will be achieved, or which kind of insights can be generated from current analytic objective. Based on this evolved awareness, business can better judge whether initial proposed requirement is competent enough to suit the project purpose.

The purpose of RR stage is to test the viability of current requirement rather than to generate insights. By skipping traditional data preparation and modeling, this stage is expected to validate two preconditions which are fundamental to subsequent PR stage:

1. Whether existing data storage is competent to support all logic steps in each analytic objective.
2. Whether the form and potential of generated insight from this stage suit the purpose that business desires in this project.

After RR Stage, if initial requirement of an analytic objective is confirmed, this objective can enter PR stage for further refinement. If initial requirement is altered, the process will roll back to Requirement Gathering step for updated requirement description.

3.3 Performance Refining

In PR stage, a sprint contains following four steps:

1. **Intensive Data Preparation**: Structure the input dataset for modeling step. In this step analyst should manage to get available as much as possible as traditional CRISP-DM data preparation. Operations such as combination, segmentation and format modification will be conducted to prepare input dataset ready for modeling tools. With the target dataset for each logic step located in

previous Preliminary Data Preparation step, this intensive data preparation is expected to take much less time than in CRISP-DM.

2. **Modeling**: Select relevant models that suit current objective and perform on input dataset with fine-tuned parameters. Some models may have specific requirement on input dataset. Therefore, it always needs to go back to intensive data preparation step. However, analyst should stick to the sprint rule, which is to deliver a testable result after a fixed period. In this way, business representative could gain a tangible awareness of how accurate current model could achieve and later on help during the performance judging step.

   At initial PR stage, sprints might have their major time spent on data preparation rather than modeling since no advanced models will perform well without high-quality input data. After that, sprint's major effort should be given to model calibrating.
3. **Visualization-driven Evaluating**: Compare performance of different models using visualization tools. Regarding the background gap between industry domain and analytics, visualization is suggested as the communicating tool which could bridge mutual understanding with graphic language. Business representative is expected to be familiar with indicators used in visualization after generic explanation step in RR stage. By interactively demonstrating model performance and comparing model's character, Analyst could offer business representative thorough interpretation about each model's capacity and how it can cope with analytic objective. As an end-user, business representative could also provide feedback from practical perspective as crucial criteria for model selection and performance evaluation. This step ends with a shippable data mining solution for current analytic objective.
4. **Performance Judging**: Stakeholder and business representative decides whether to further this performance refining stage with another sprint or to accept and deploy current solution. If further improvement is preferred, next sprint will start with bring more available data source which could be skipped by previous sprints.

> Usually the improvement on model performance will present a descending trend. After several initial sprints, business representative would have an empirical estimation of the potential of current model after next sprint. Based on this estimation and the practical requirement of the objective, business representative can decide that whether current analytic objective still needs a refined solution or not.

The purpose of PR stage is to complete the analytic objective in a business-manageable process. Through interpretation and interaction in Visual-driven Evaluating stage, business is enabled to participate in the analysis and provide domain knowledge as an important criterion for model selection. Meanwhile, through an increasing awareness towards the model ability, business is competent to make following judgments:

1. Compare model performance and character from an end-user perspective.
2. Accept solution from certain sprint to stop PR stage to achieve a trade-off between accuracy and cost from a business angle.

With these two considerations included, PR stage can improve business manageability on data mining project and also improve stakeholder’s confidence to adopt data mining result after deployment.

3.4 Chapter Summary

This chapter gives a thorough elaboration about the AM process in both stage and step level.

In requirement reconfirming stage, AM asks analysts to use frequent initial deliverable from partial dataset to clarify stakeholder’s real concern. In performance refining stage, with the definition of scrum being applied in the modeling procedure, stakeholders are able to do trade-off on their preferences.

4 AM Refined Requirement Confirmation and Accuracy-Cost Balance

4.1 Gap Research

Traditional data mining models prefer consistent customer requirements throughout whole data mining project. However, this situation rarely happens in increasingly complicated modern business environment. Moreover, a large amount of enterprise from traditional domain has heard about the benefits that data mining can produce, but lack the basic background to get ready for a data mining project. This insufficient preparation could be reflected in three aspects.

4.1.1 Data Insufficiency

This situation arises when a company lacks fundamental data to conduct relevant data mining practice.

Data sufficiency is the basic of a successful data mining project. As its name indicated, Data Mining is a method based on data. Only by possessing necessary data from business workflow can analysts conduct analysis and abstract trustful patterns.

However, in real case, not all enterprises are fully aware of their data readiness. By knowing the advantages which data mining can bring to business, some of them tend to blindly believe that data mining is a mantra that can be applied to any company and bring same benefits. This problem could impact the effectiveness of analytic result such as a predictive model with low accuracy. A worse situation case could be that after understanding business and data, analysts considered that current data quality couldnot support the proposed analysis, and by that time, considerably cost has been wasted. This dilemma can frequently cause argument about liability on the wasted cost and vitally damage the cooperating relationship between enterprise and analyst.

4.1.2 Evolving Analytic Objectives

This situation happens when stakeholders change their requirement after getting clearer of how data mining can help with their business.

Traditional data mining processes suggest analysts pay plenty of time and efforts to understand customer’s data, business workflow at early stage before start getting hands-on analytic algorithms. In CRISP-DM, it will always cost 50 percent to 70 percent of planned project schedule to finish the preparation job before modeling, and this preparation job is expected as non-repeatable once finished if early work has been done well. After that, priority will be given to the refinement of model, which could represent analysts’ professional skill. After getting a relatively good result from evaluation, the outcome will be explained to customer and then comes to deployment.

From the view of customer, the first executive version of project delivery is provided at last quarter of project schedule. This situation leaves no opportunities for customer to adjust within original plan once they acquire fresh view from this executive deliverable. Conventionally, customer could either suffer the imperfect result or negotiate a new schedule with analysts.

4.1.3 Unmanageable Accuracy-Cost Tradeoff

This situation occurs when customer has a budget plan and negotiable requirement on result accuracy.

During traditional modeling and evaluation stage, before reporting to business, analysts prefer getting optimal result at once, which reveals their academic competency. Meanwhile, this process may take quite much time on calibrating parameters and comparing output between different models, or even fixing data error to enlarge input dataset. Despite all these efforts, the improvement of model's accuracy usually follows a semi-parabolic trend. After getting a high level of accuracy, no matter how long analysts have worked on a model, its accuracy will not be improved significantly. But from the view of analysts, it is still worth trying, given that model accuracy is the most important criterion for academic success. In this situation, time spent on later modeling stage might be not that productive as earlier.

In contrast, business customer may value the inspiration by the generated patterns more than a precise pattern or prediction itself. If they had a choice to make a stop during the modeling stage, it would be much beneficial for business to maintain a balance between model accuracy and project cost. However, this is never an option for business in traditional data mining process, since the first version of presented deliverable is almost the best that analysts can achieve.

4.2 Agile Based Data Preprocessing and Modeling

According to the above mentioned three concerns in chapter 4.1, Agile Mining process provides a set of niche targeting solution for them:

1. By adding a quick global reconfirming stage before drilling down to get more accurate model, AM achieves a preliminary test on data sufficiency.
2. By generating an explanatory deliverable from initial reconfirming stage, AM helps stakeholders clarify their requirement and also provides analysts a stable direction for subsequent polishing work.
3. By iteratively improving model's accuracy and interpret model performance to business partner with visualization-driven evaluating, customer could be fully aware of the model's function and stop the refining stage after certain iteration to achieve preferable accuracy-cost balance.

4.2.1 Data Sufficiency Test

To know whether a company's current data can support a data mining practice, analysts usually have to go through every detailed step of the company's workflow. In real cases, these steps can be extremely complicated. Given this situation, traditional processes suggest a remarkable percentage of time on this business understanding. As a result, there is no proper test of data sufficiency before project officially starts. During the data understanding process, once this shortcoming is revealed, there will be no shippable deliverables to business and all previous investment is wasted.

In light of that, AM proposed a quick global data sufficiency test in requirement reconfirming stage to ensure currently available is enough to support the solution of analytic objective. Referring to Scrum's backlog grooming method, first the project deliverable should be elaborated into several logic steps, with each step describing which data features will be needed. After the grooming stage, analysts will select the best part of all the data sets containing features mentioned above and get these data sets combined to see whether all necessary features are accessible.

In this step, plenty of discordances will emerge due to data quality. Normally analyst should only take the data set that perfectly connected. However, some simple alternations are still suggested to be done if more data can be taken in after doing so.

This type of discordance frequently emerges when combining data sets from different systems. This kind of modifications including:

Feature type: float and integer, factor and character, etc.

Feature format: date (dd/mm/yy or mm/dd/yyyy), name (first name + last name or first name + last name), etc.

By skipping traditional complex data preparation, this data sufficiency test could detect the missing part in the logic chain of proposed deliverable, and stop the further waste at first time with refinement suggestions. If all necessary data sets can combine, AM process can move on with ensured data sufficiency.

4.2.2 Deliverable Reconfirmation

To suit the evolving requirement of deliverable, the iterative developing methods from Scrum is referred in AM. By targeting the current iteration at the top priority user story, analyst could generate a preliminary but executable result. This result might be inaccurate because it is based on partial data preprocessing, while it carries all logic steps to realize current story. By presentation this result to business partner, the initial business requirement can be either confirmed or modified.

4.2.3 Business-Oriented Accuracy-Cost balancing

To make the most of project budget, AM suggests a Performance Refining stage. It is built by fixed length iterations to achieve a gradually increasing model performance. Through Visual-Driven Evaluating step, business representative would understand model’s capability after each sprint and help stakeholders decide whether current solution is competent enough for corresponding analytic objective.

4.3 Case Study of AM application

Two industry projects demonstrated here used to exemplify the application of AM process and its advantage in RR and PR stage respectively.

4.3.1 Requirement Reconfirming in Workforce Modeling

Data mining was applied during a project with one of Australian leading age care service vendors. Project purpose was to modify current workforce model to cut cost from employee salary while maintaining competency for service demand. Both the organization and its stakeholders were inexperienced for an actionable requirement description about analytic objective. Given that, AM process was applied to confirm the requirement first.

There are two kinds of employment contract in this company: casual care worker (CCA) and permanent part-time worker (PPT). CCA is on a higher hourly rate, roughly 30 percent higher than PPT, while PPT enjoys a minimum-working-hour guarantee, which ensures PPT at least 15 working hour wage per week, in case there is few work assigned to a PPT. At that time, over 80 percent care workers were CCA. In this case, the age care company planned to transfer a portion of selected CCA to PPT. By doing so, company could cut the cost of work which used to be done by CCA. Meanwhile, the gap between minimum-working-hour-guarantee and PPT’s real working hour should be minimized.

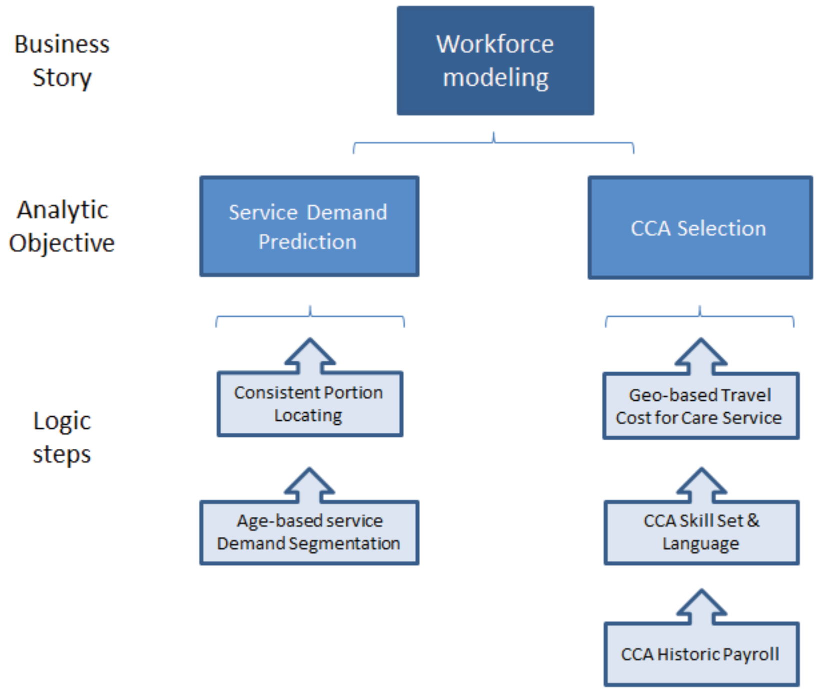
Business Story
Workforce modeling
Analytic Objective
Service Demand Prediction
CCA Selection
Logic steps
Consistent Portion Locating
Age-based service Demand Segmentation
Geo-based Travel Cost for Care Service
CCA Skill Set & Language
CCA Historic Payroll

Figure 4.1 Initial Requirement for Workforce Modeling

In light of above description, this workforce modeling requirement was separated into two analytic objectives, which were Service Demand Prediction and CCA Selection. The former story calculated the potential capacity for future market. The latter story filtered those CCA with better productivity and preferable skill set. Also an initial plan for first sprint in RR stage was designed with logic steps shown in blue blocks in figure 4.1.

Service Demand Prediction was to estimate the scale of the aged care company's service in demand.[49] To achieve that, during first RR sprint, only the company's growth in different age groups was considered as indicator for future market. After dataset combination, linear regression was applied to past four years data in Explanatory Modeling step to get initial result. However, after seeing how regression model works in Generic Modeling step, business representative indicated that the total population growth in the area should also be taken into consideration. In light of that, ABS population pyramid data was added in next RR sprint as a logical step to improve regression's accuracy (dash block in figure 4.2).

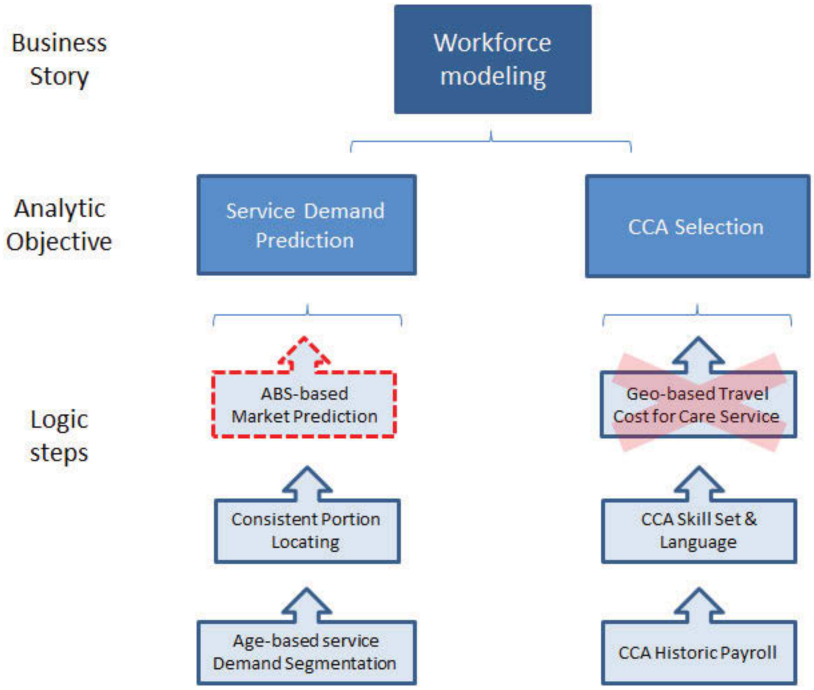
Business Story
Workforce modeling
Analytic Objective
Service Demand Prediction
CCA Selection
Logic steps
ABS-based Market Prediction
Geo-based Travel Cost for Care Service
Consistent Portion Locating
CCA Skill Set & Language
Age-based service Demand Segmentation
CCA Historic Payroll

Figure 4.2 Altered Requirement for Workforce Modeling

CCA selection was to find those CCAs with relatively higher productivity among peers. To achieve that, CCA with a demonstrated higher average weekly wage in historic service record was selected. Next, selected CCAs were segmented again by their language and skill set. Those CCA with more popular language and broader skill set were preferred. Then CCAs' geo-locations were examined since the travel distance between CCA and customer address was also paid as service cost. These three logic steps formed the CCA selection in first RR sprint.

During first RR sprint of CCA selection, Google Map API was used for the estimation of service travel distance. This logic step took major computational power and time of the sprint. However, the estimated result was found unmatched the historic payroll in later result explanation. During discussion with business representative, it was found that those CCA with loaded service demand were usually assigned with nearby orders in a row, which means distance between CCA and customers were not necessarily influential among CCA selection. This logic then was removed in next RR sprint as shown in figure 4.2.

By making these adjustments to initial PBI, company gained direct understanding about what will be implemented in this project, how the data is integrated and what function could they expected after deliverables are deployed. This enhanced understanding helped build trust in upcoming PR stage and avoided potential waste of revision from late stage.

4.3.2 Performance Refining in Customer Churn Prediction

One of Australian leading superannuation management and investment companies expected to apply data mining on its existing data to enhance their marketing productivity. Customer churn was regarded as one of their most concerned issues in its marketing business. An early warning system to detect potential customer churn was desired to prevent the financial loss.[50] Meanwhile, this decision-support system should be well understood and convincing to business so that correspondingly effective marketing strategy can be customized based on the output of this system. To achieve this purpose, AM process was used to guide this analysis.

A ten-year dataset including millions customer account was provided with label indicating customer's current status (churned or not) with 3 percent churned customer among the whole customer base. Customer's demographic information was provided except identity information. Customer's behavioral information including balance changes and phone record for service was also available in the dataset.

After confirming the objective requirement with the company in RR stage, a regression model was proposed, and a ranking list regarding every customer's predicted churn possibility was used to illustrate the churn risk. Five algorithms were selected to calculate this churn possibility, which was Random Forest [51] (RF), Support Vector Machine [52] (SVM), Naïve Bayes [53] (NB) and Logistical Regression [54] (LR).

Sprint length was set as fortnight. To better interpret the result, a gains chart was designed to determine the model's ability to discriminate churners and non-churners as shown in figure 4.3.

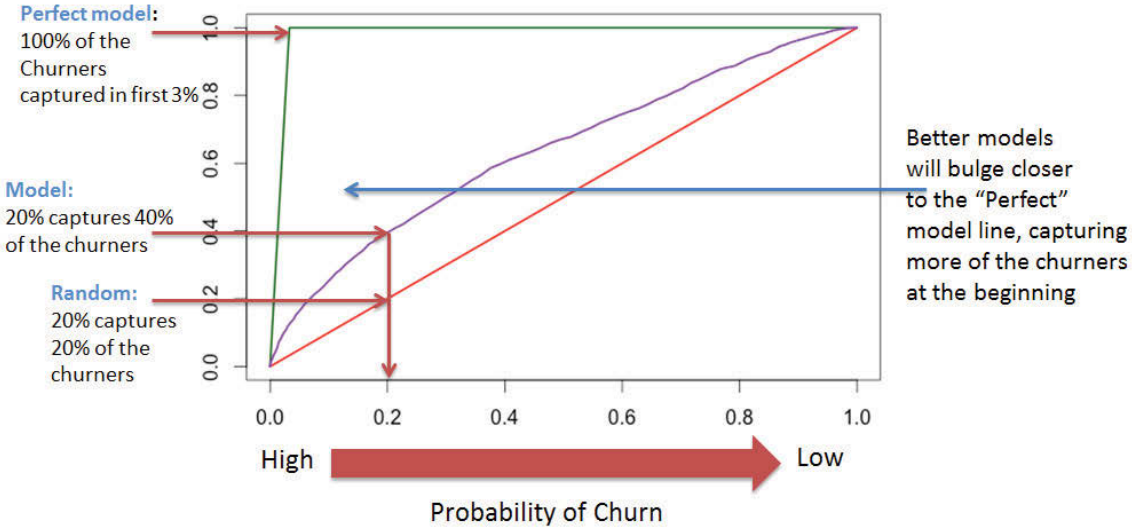

Figure 4.3 Gains Charts for Model Evaluation

In this gains chart, ideally, if an algorithm can predict all real churned customers with higher rank than others, its curve in this gains chart will be shown as the perfect line. The better the algorithm is, the steeper the line rises, which means it can precisely capture churned customer only with higher predicted possibility. For better intuitive understanding, another reference line is also drawn to compare the model result with purely random guess.

In first sprint, more tasks were concerned about intensive data preparation. After combining available data including all necessary features, RF, LR, SVM, NB models were performed on this combined dataset with their default parameters. Results of these four algorithms were shown using gains chart in figure 4.4.

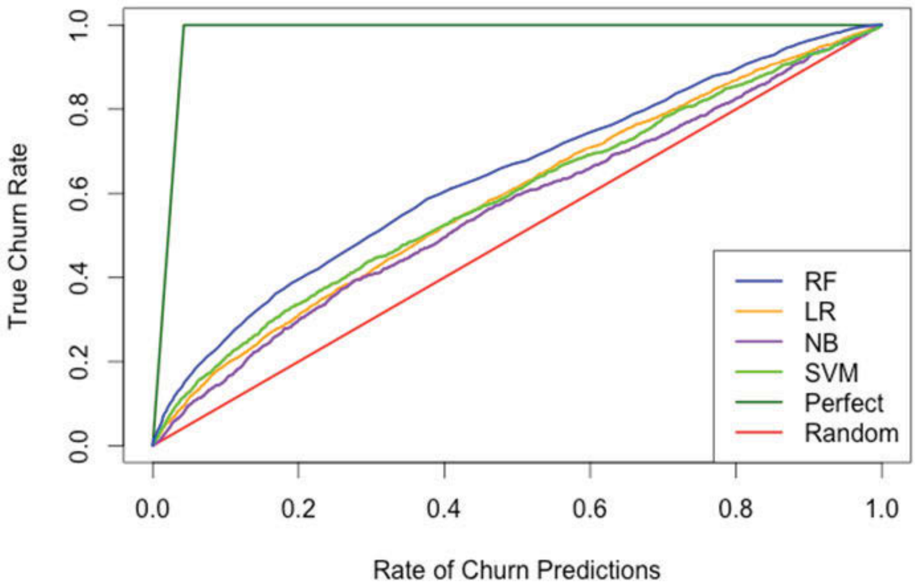

Figure 4.4 Result of First Sprint

In first sprint result evaluation, RF presented the best result with its top 20 percent prediction capturing 40 percent of real churner, which matches its character of less parameter calibrating demand. Next was SVM, which showed a better performance at its top 40 percent prediction than LR. NB presented the lowest accuracy with a remarkable disadvantage on its top 15 percent prediction. After presenting this result to business representative and stakeholders, second PR sprint was agreed to proceed.

Then second sprint started with original features in sprint one and attaches more concern on parameter tuning. However, as shown in figure 4.5, except LR gained a considerable enhancement on its top 40 predictions, results on other three algorithms remain stable compared with first sprint. In Performance Judging step, idea of creating derived feature was proposed for enhancing the prediction. Agreed by business, the third sprint was conducted.

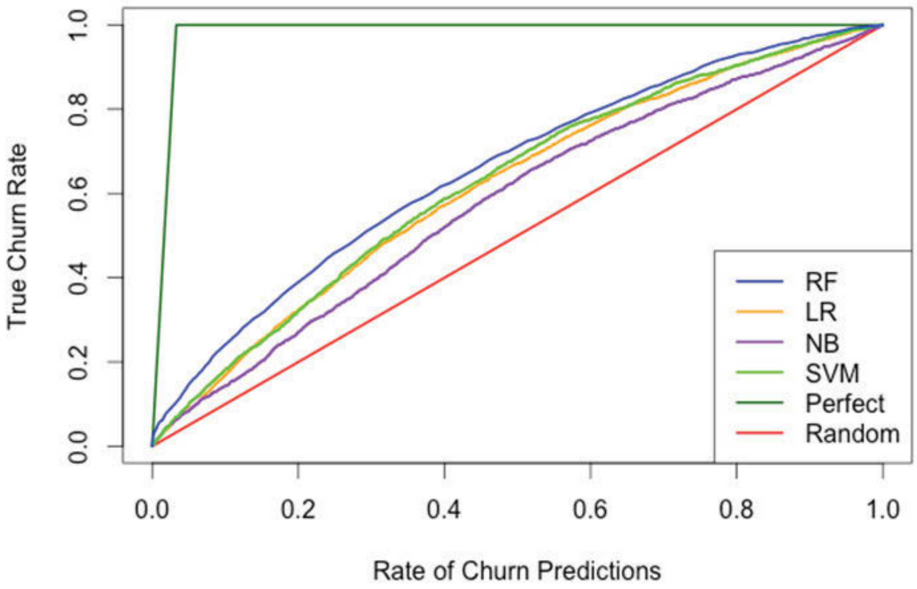

Figure 4.5 Result of Second Sprint

To refine the model performance, derived features were created based on initial dataset. For example, one derived feature showed whether customer withdrew more than 50 percent of account balance within ten days after making a phone call query. These features were believed as able to indicate potential churn risk. After adding these derived features into the training dataset, RF's predictive accuracy was significantly improved as the blue line in figure 4.6. With its top 5 percent prediction, 35 percent of churners were captured. Except for NB's result, LR and SVM's results are also amazingly refined.

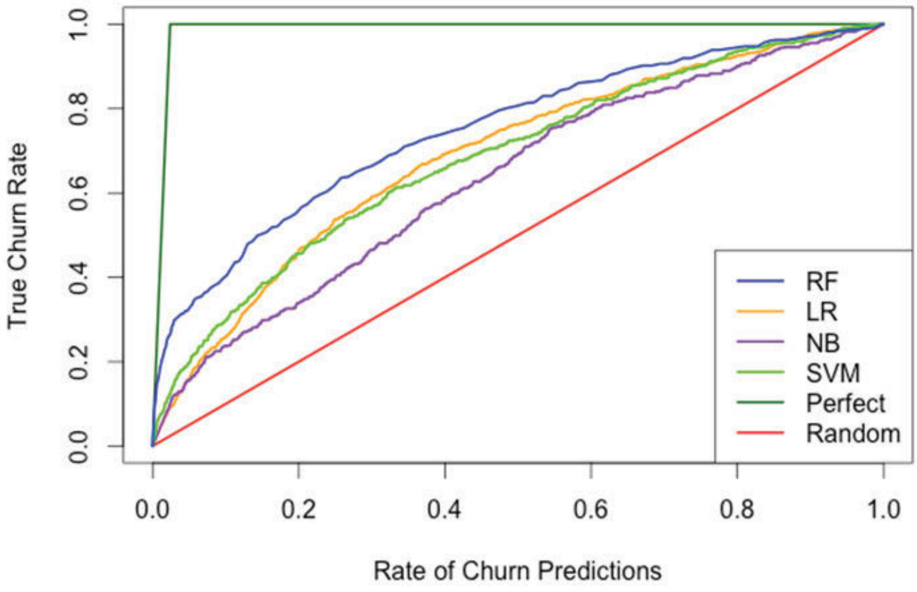

Figure 4.6 Result of Third Sprint

After presenting this third sprint result to business, the performance was accepted as competent for further marketing strategy design. Meanwhile business representative and stakeholders were confident about applying this churn detecting system into their real business and highly satisfied about the project deliverable.

4.4 Chapter Summary

This chapter elaborates how AM can better resolve the three issues in traditional data mining process, which are data insufficiency, evolving analytic objective and unmanageable accuracy-cost trade-off. Two industry projects are used to illustrate its improvement and application.

5 AM Enhanced Interpretation and Interaction

5.1 Gap Description: Bias on Understanding Project’s Success

It is not rare that data mining project fails on transferring meaningful interpretation and inside logic of the data mining project deliverables to stakeholders. The main reason for that is the bias of project success criterion. Analysts always put major efforts on accuracy of calculation or performance of model, which will then become statistical indicators or analytic terminologies. In comparison, stakeholders would only trust the result which can be understood or validated within their knowledge, which requires the analytic outcome being expressed more intuitively[4].

As a result, the bias of unmatched expression style limits the productivity and restrains the application of valuable data mining findings. After deployment, stakeholders’ doubt would significantly restrain the productivity that the analytic outcome should contribute. This recurring scenario remarkably violates industry data mining project’s commercial success.

Though the CRISP-DM is regarded as ‘very industry-oriented’ [55], in practice, there is still a considerable limitation when analyst introducing the process to their industry partner or acquiring business understanding from their domain expertise, owing to the lack of common understanding communication platform. Occasionally, industry partner, small and medium-size businesses specifically, remain unsure about whether their expectations are achievable or actionable with existing data when signing the project contract. In addition, industry partner’s expertise usually suffers remarkable reduction during the transformation to data analysts as business understanding, which increases risks of project.

5.2 Visual-Driven Interpretation and Interaction

To make the most of the analytic outcome, AM suggests a visual-driven interpretation and interactive way of expression. By using a series of novel data mining visualization method, AM manages to combine statistical indicators and analytic terms with business-friendly figures. One of the Visual-Driven Evaluation main characters is that it focuses on analytical reasoning facilitated by interactive visual interfaces [56] Interactive interface is applied to help interpreting the meaning of each dimension in the visual graphs, which enhances stakeholders’ confidence significantly by enabling business people interact with their data. In addition, the feedback from business people can also be added into consideration for model’s further refinement.

With the support of data mining visualization tools, AM process straddles top-down and bottom-up approaches, suitable for both industry partner and data analyst, and are typically deployed more quickly and at lower cost than business-centric tools [57]. Moreover, these visual tools offer many of the same capabilities as traditional business intelligence platforms but are typically much easier to deploy and manage.

In practice, not all the project industry partners are capable of conducting the initial prototyping work due to various limitations. However, by visualizing the data into interactive dashboard, analyst can achieve a jumpstart of a project to obtain quick insights from raw data and interpret both project objectives and project process in an intelligible way to industry partner. This will help industry partner building confidence and acknowledgment to the project and therefore gain better client satisfaction. Similarly, this interactive way can also benefit data analysts on bringing industry partner's domain knowledge into the project to facilitate project progress, especially when the problem requires domain knowledge more than analytics techniques.[57]

As a consequence, this visual-driven interpretation and interaction fill the gap between outcome performance and business comprehension. With the elimination of communication barrier, business people could benefit from intelligible analytic outcomes; analyst could also gain better business understanding from feedback of the interaction.

5.3 Case Study

One of Australia's leading providers of insurance has become our industry partner, seeking solution from data mining to improve their current insurance flow and customer experience. The company has a comprehensive range of insurance products available through financial advisers or direct to customers.

This is the first data mining practice that the company launched on their historic database. One insurance underwriter is assigned as business representative to provide support and receive feedback from analyst team. To effectively interact with business representative, a series of visual-based analyses on its insured customer behavior was conducted to find solutions that best suit the individual customer needs.

5.3.1 Information Gain Based Heatmap

Applying life insurance often requires in-person meetings with underwriters, tedious paperwork, and an average waiting time of six weeks to get policy activated.[58] This outdated process has become a barrier to broader consumer adoption, resulting large coverage gap. One of the major issues within this process is that applicants have to complete an overcomplicated questionnaire to get assessed by underwriters.

The questionnaire, which contains both useful and useless questions, can be as lengthy as 100 pages with more than 2.5K possible questions. Filling such questionnaire is tedious even though. Some nested questions can be skipped. Therefore, optimizing the questionnaire has become a very first step towards simplified insurance application process.

Advances in machine learning have made it possible to evaluate the importance of each question in a data-driven manner. Specifically, the questionnaire optimization can be considered as a feature selection problem in machine learning. Essentially, each question is considered as a feature and the claims are considered as the labels. The feature selection task is then to select the subset of questions that can efficiently determine how likely an applicant will have a claim in the future.

Life insurance application takes five typically steps [59]:

Compare Quotes: The first step is to compare quotes from different companies, even though the final price will vary depending on applicant’s situation.

The Application: The second step is then to lodge an application by filling some basic personal information, which normally takes half an hour to complete.

The Medical Exam: The third step is to take a quick medical exam for blood pressure, weight, etc. This normally takes half an hour to complete.

Underwriting: The longest part of the whole application is this underwriting process, in which applicants and underwriters need to complete a lengthy application form. This process takes six months on average.

Decision: Once the underwriting is complete, a decision will be made and an offer is given to the applicant. However, if the offer does not meet applicant’s expectation, it might be rejected, and the whole process is wasted. This six-month process may lead to a rejection of offer, therefore it is crucial to shorten the application process, particularly, the underwriting process.

Given the questionnaire with more than 2.5K questions, the underwriter will try to estimate the risk of an applicant. However, the questionnaire is extremely sparse as some questions are only enabled if certain answers are provided to some parent questions. To identify which subset of questions have strong impact on claims (risks), we adopted the Minimum Redundancy-Maximum-Relevance (mRMR) [60] feature selection method.

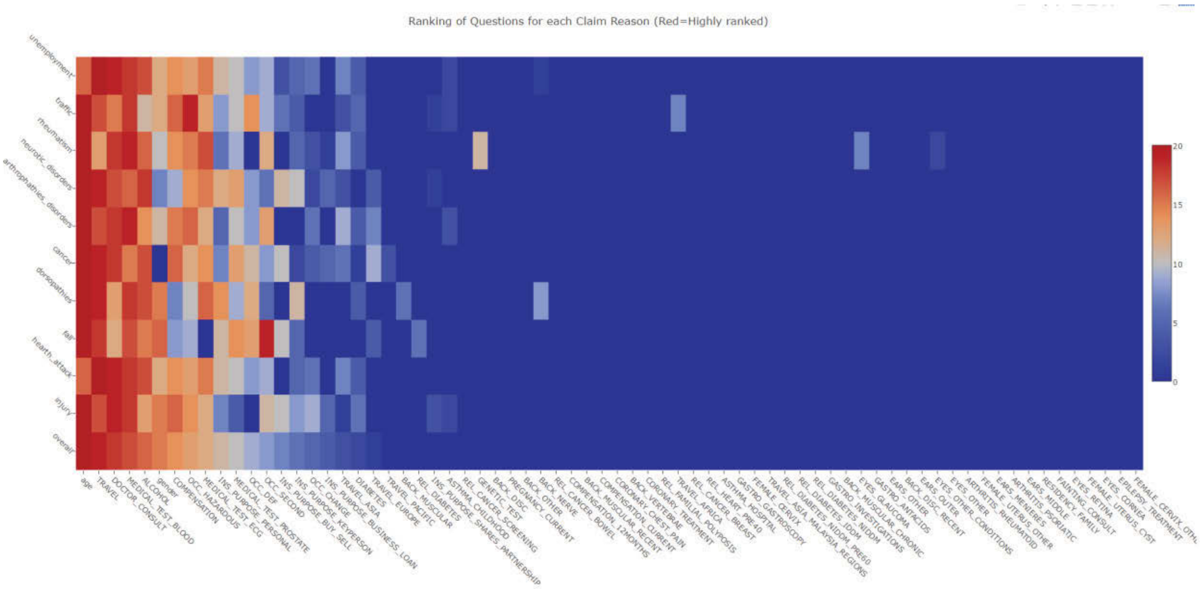

Figure 5.1 Information Gain based Heatmap

The goal of mRMR is to measure the relevance of each question to claim reasons. Meanwhile, the redundancy of each question should also be minimized, as some questions may be highly correlated. Specifically, the relevance of a question set Q for the claim reason c is defined by the mean of mutual information between each question qi and the claimed reason c:

$$D(Q,c) = \frac{1}{|Q|} \sum_{q_i \in Q} I(q_i; c)$$

Equation 5-1

On the other hand, the redundancy of question set Q is calculated as the mean of mutual information between each pair of questions qi and qj within question set Q:

$$R(Q) = \frac{1}{|Q|^2} \sum_{q_i, q_j \in Q} I(q_i; q_j)$$

Equation 5-2

Put relevance and redundancy together gives the mRMR measure:

$$\text{mRMR} = \max_{Q} \left[\frac{1}{|Q|} \sum_{q_i \in Q} I(q_i; c) - \frac{1}{|Q|^2} \sum_{q_i, q_j \in Q} I(q_i; q_j) \right]$$

Equation 5-3

The mRMR algorithm approximates the optimal maximum dependency feature selection algorithm. The algorithm considers the pairwise interactions of two questions in the question set.

5.3.2 Association Rule-Based Directed Graph

Association rule has been a popular technique for discovering interesting relationships between multiple attributes in a dataset. The relationship is precisely defined in the form of an association rule A towards B, where A and B are possible values of attributes. The strength of rules is measured by confidence or lift.

We applied association rule technique to discover how do exclusions related to the claim reasons. For example, some claimers may have been excluded for what they planned claim. Thus they claim for something else. We applied this technique to the exclusion vs. claiming dataset and a large number of rules are extracted and need manual selection. A subset of the rules is shown as in table 5-1:

Table 5-1 Rule-Based Association expression

Horse Riding 7	Accidental Falls 6	conf:(0.86)
Cancer 119	Cancer(Claim) 87	conf:(0.73)
Tinnitus 16	Accidental Falls 10	conf:(0.63)
Oesophagus 16	Accidental Falls 9	conf:(0.56)
Endometriosis 22	Cancer(Claim) 12	conf:(0.55)
Thoracic spine 11	Accidental Falls 6	conf:(0.55)
Noninfectious Enteritis and Colitis 12	Lumbar Spine 6	conf:(0.5)
Bowel 12	Motor Vehicle Traffic Accidents 6	conf:(0.5)
Social Instability 27	Accidental Falls 13	conf:(0.48)
Suicide and Self-Inflicted Injury 23	Mental Illness 11	conf:(0.48)
Elbow 42	Accidental Falls 20	conf:(0.48)
Leg 26	Accidental Falls 12	conf:(0.46)
Scuba Diving 13	Unemployment 6	conf:(0.46)
Newly Self Employed Exclusion 165	Accidental Falls 74	conf:(0.45)
Shoulder 27	Accidental Falls 11	conf:(0.41)
Multiple sclerosis 30	Cancer(Claim) 12	conf:(0.4)
Sky Diving 15	Accidental Falls 6	conf:(0.4)
Hands 53	Accidental Falls 21	conf:(0.4)
Skin cancer 33	Cancer(Claim) 13	conf:(0.39)
Territorial 41	Accidental Falls 16	conf:(0.39)
Offset Clause - Other Income Cover 18	Accidental Falls 7	conf:(0.39)
Breast Cancer 37	Cancer(Claim) 14	conf:(0.38)
Malignant 40	Cancer(Claim) 15	conf:(0.38)
Ca in situ cervix uteri 30	Cancer(Claim) 11	conf:(0.37)
Diabetes 25	Cancer(Claim) 9	conf:(0.36)

The above table shows a fraction of a large number of rules discovered from the dataset. Identifying the patterns and interesting rules can be difficult. Therefore, we proposed to use Force-Direct Graph to visualize the rules thus the selection process can be simplified. The force-directed graph of the above rules is shown as follows:

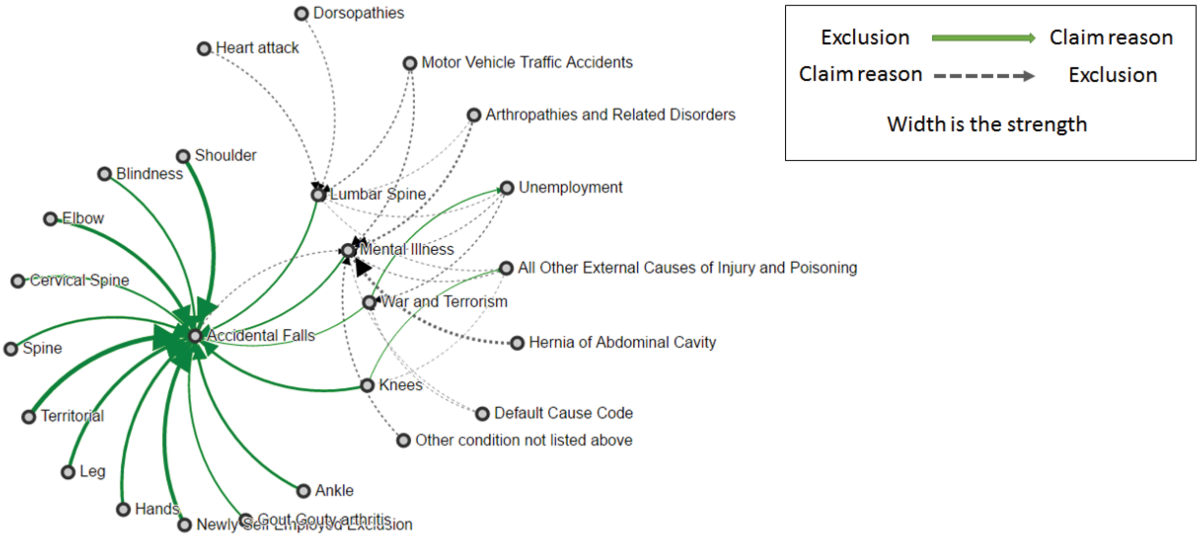

Figure 5.2 Association Rule based Force-Directed Graph

It can be observed that the rules are grouped by some nodes, such as “Accidental Falls”. By using this method, we managed to identify interesting rules as well as discovering the previously unknown relations among rules.

5.3.3 Customer Segmentation Based Parallel Coordinate

High dimensional data visualization is a challenge for many classic visualization methods, such as histogram and scatter plot, which has been proved useful in many visualization applications [61]. Although some works tried to use classic visualization methods for high dimensional data, such as using color and other attributes, the number of displayable dimensions is still limited. However, parallel coordinates is an efficient visualization method for high-dimensional multivariate data [62].

In this case, the project goal was to do customer segmentation on income protection (IP) insurance customer base. Business team consisted of underwriting and actuary staff from the insurance company. These people were familiar with both insurance workflow and meaning of each feature in calculated result but inexperienced to handle statistical parameters. Meanwhile, their domain expertise was expected to contribute to the business value identification.

In light of above circumstances, we decided to use parallel coordinates to present the customer segmentation result with each line representing a unique customer pricing category and each node on each vertical axis representing corresponding value (either numeric or categorical) of a feature (e.g. age and occupation). By comparing the rate of cost vs. revenue (cost rate) among all customer groups (defined by age, gender, etc.) with IP average cost rate, customer groups were labeled with A and B (above and below IP avg. cost rate). By showing all axes with features' original business name, this parallel coordinate visualization could be used to interact with domain experts from collaborating team without time-consuming introduction of analytic background. And by following their experiential guidance of zoom-in and highlighting, target customer groups with abnormal performance were discovered and existing intuitive business hypotheses were validated.

Insurance has the top level of data security requirement among all the domains. In our cooperation with an Australian leading insurance company, we were aiming to testify whether the company's current applied price scheme for each customer category is in accordance with the corresponding claim cost.

To achieve that, we structured customer demographic information (including age, gender, postcode, occupation, smoking status, etc.) for each globally unique identifier (GUID) and calculated the cost vs. revenue rate for each customer category. For showing the findings in a direct way, we applied Parallel Coordinate to embed each customer category into one line with each node standing for an untraceable feature. At the right end of the chart, we used A and B to reveal whether a certain customer category is above or below the average cost-revenue rate.

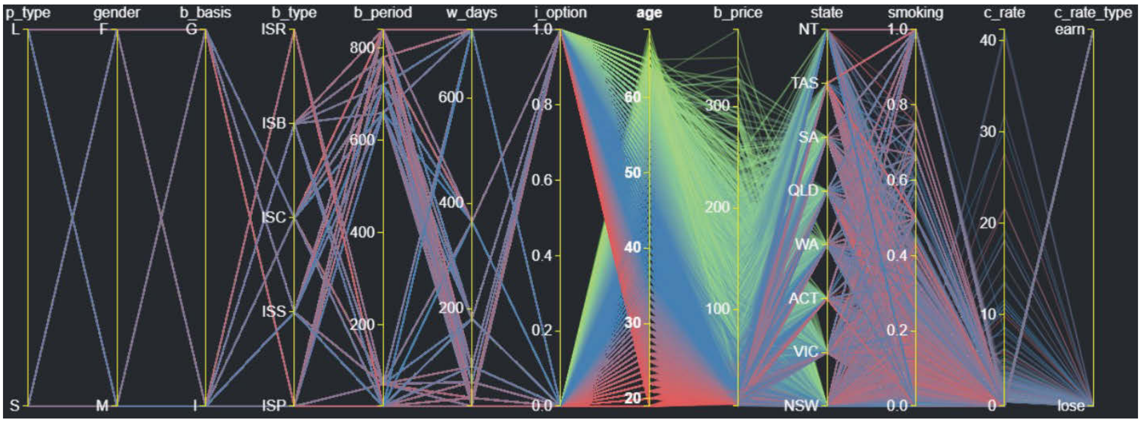

Figure 5.3 Parallel Coordinates without Clustering

In above mentioned presentation, customer demographic features are included such as gender, age, occupation category, smoking status and premium amount (which is relevant to individual annual income). These features are considered as unidentifiable information. A collection of unidentifiable features, however, can still be used to locate unique qualified individuals, given that there large amount of customer groups containing one customer only.

Privacy risk was considered as acceptable when segmentation result was shared only among analysts and insurance team. Then circumstance was changed that insurance company required sharing the findings with their collaborative reinsurance company. Meanwhile, same type of parallel coordinates chart with original features' name was also preferred.

To achieve customer privacy-preserving, we proposed to apply clustering on original result. The benefits of conducting clustering include:

1. Hide exact customer information behind aggregated value to protect customer privacy in minor categories.

2. Scale down the amount of total customer category with acceptable accuracy loss to make segmentation result more generic.

3. Release resolution competency and computational power by reducing the amount of line. This case risked the customer privacy issue.

Table 5-2

	Max	Min	1% Perc.	5% Perc.	10% Perc.	Speed
Canopy	679	6	9	25	48	0.47s
K-Means	658	13	20	35	51	71.65s

The performance of two clustering models is shown in table 5-2. For each value in 3rd to 5th columns means there are 1%, 5% or 10% clusters whose size is lower than this value. Since tiny size cluster may still lead to the risk of privacy, higher percentile value model is preferred in our project. The KMeans model has higher percentile, while canopy model has faster speed. The choice of model should be based on detail requirements.

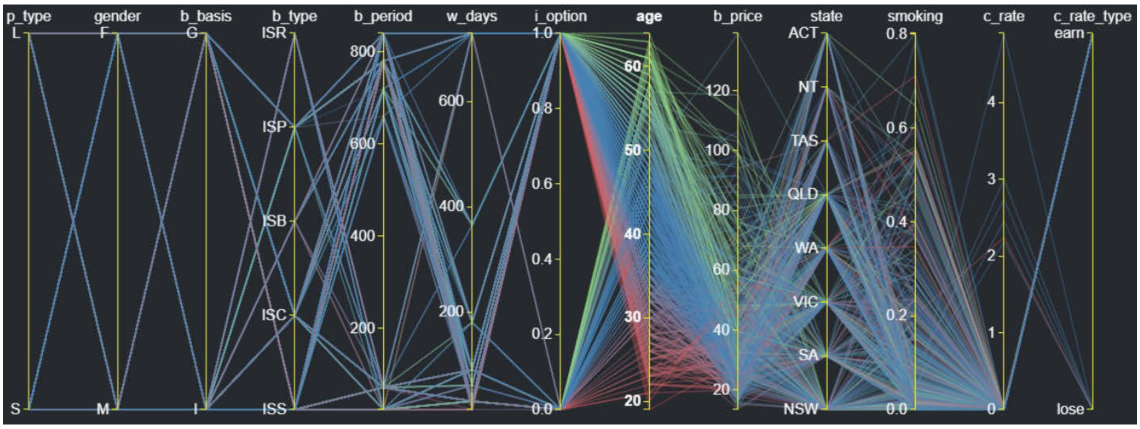

Figure 5.4 Parallel Coordinates with Clustering

By comparison of two clustering algorithms, K-means clustering was selected and the clustered result was put into the parallel coordinates as a privacy-preserved presentation. The parallel coordinates with clustering are shown in figure 5.4, compared with the one without clustering shown in figure 5.3.

5.4 Chapter Summary

This chapter introduces the interactive expression of AM process and its benefit for both business and analysts. Case study introduces three novel data mining visualization methods to effectively interact with business and achieve customer satisfaction in data mining project.

6 Conclusion

In sum, this agile-method-integrated novel data mining process, Agile Miming, can minimize the cost of business requirement change to deliver the business desired project outcome. By prioritizing visual interaction in evaluation step, AM also achieves an effective and efficient way for mutual knowledge exchange. With these advantages, AM improves customer satisfaction in industry projects and provides business-intelligible deliverables.

Due to the innovation of the Agile Mining, specific guidance within each step is still insufficient. In future work, it is suggested that generic solutions for typical problem scenarios should be collected such as criteria of requirement reconfirming for different analytic purpose. Also, along with the development of Agile methods, the requirement engineering technology can also be better merged with data quality checking work.

7 REFERENCE

[1] E. W. Ngai, L. Xiu, and D. C. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," *Expert systems with applications,* vol. 36, no. 2, pp. 2592-2602, 2009.

[2] R. N. Charette, "Why software fails [software failure]," *IEEE Spectrum,* vol. 42, no. 9, pp. 42-49, 2005.

[3] K. Beck, *Extreme programming explained: embrace change*. addison-wesley professional, 2000.

[4] H. Ian, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[5] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[6] G. D. Smith and S. Ebrahim, "Data dredging, bias, or confounding," *Bmj,* vol. 325, no. 7378, pp. 1437-8, 2002.

[7] P. Abrahamsson, O. Salo, J. Ronkainen, and J. Warsta, "Agile software development methods: Review and analysis," *arXiv preprint arXiv:1709.08439,* 2017.

[8] G. Moore, "Marketing and selling high-tech products to mainstream customers. Crossing the chasm," *HarperBusiness, NewYork,* 1991.

[9] R. Agrawal, "Data mining: Crossing the chasm," 1999.

[10] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 2000, pp. 29-39.

[11] R. Schutt and C. O'Neil, *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.", 2013.

[12] O. Laudy, "New Standard Methodology for Analytical Models," IBM Analytics, Asia-Pacific, KDnuggets2015.

[13] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. Kurgan, *Data mining: a knowledge discovery approach*. Springer Science & Business Media, 2007.

[14] P. Chapman *et al.*, "CRISP-DM 1.0 Step-by-step data mining guide," 2000.

[15] C. Shearer, "The CRISP-DM model: the new blueprint for data mining," *Journal of data warehousing,* vol. 5, no. 4, pp. 13-22, 2000.

[16] H. Mari Oriyad, F. Zare Derisi, M. Jahangiri, M. Rismanchian, and A. Karimi, "Evaluation of Heating, Ventilation, and Air conditioning (HVAC) System Performance in an Administrative Building in Tehran (Iran)," *Health and Safety at Work,* vol. 4, no. 3, pp. 59-66, 2014.

[17] S. Jensen and U. SPSS, "Mining medical data for predictive and sequential patterns: PKDD 2001," in *Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2001.

[18] G. Butler, "System and method of extracting data from vending machines," ed: Google Patents, 2002.

[19] H. Blockeel and S. Moyle, "Collaborative data mining needs centralised model evaluation," in *Proceedings of the ICML-2002 Workshop on Data Mining Lessons Learned*, 2002, pp. 21-28.

[20] E. Silva, H. Do Prado, and E. Ferneda, "Text mining: crossing the chasm between the academy and the industry," *WIT Transactions on Information and Communication Technologies,* vol. 28, 2002.

[21] J. Hipp and G. Lindner, "Analysing warranty claims of automobiles," *Internet Applications,* pp. 31-40, 1999.

[22] W. Gersten, R. Wirth, and D. Arndt, "Predictive modeling in automotive direct marketing: tools, experiences and open issues," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 398-406: ACM.

[23] W. F. Morris, P. L. Bloch, B. R. Hudgens, L. C. Moyle, and J. R. Stinchcombe, "Population viability analysis in endangered species recovery plans: past use and future improvements," *Ecological Applications,* vol. 12, no. 3, pp. 708-712, 2002.

[24] S.-T. Li and L.-Y. Shue, "Data mining to aid policy making in air pollution management," *Expert Systems with Applications,* vol. 27, no. 3, pp. 331-340, 2004.

[25] N. de Abajo, A. B. Diez, V. Lobato, and S. R. Cuesta, "ANN quality diagnostic models for packaging manufacturing: an industrial data mining case study," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 799-804: ACM.

[26] G. Mariscal, O. Marban, and C. Fernandez, "A survey of data mining and knowledge discovery process models and methodologies," *The Knowledge Engineering Review,* vol. 25, no. 2, pp. 137-166, 2010.

[27] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines," *International journal of medical informatics,* vol. 77, no. 2, pp. 81-97, 2008.

[28] A. I. R. L. Azevedo and M. F. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview," *IADS-DM,* 2008.

[29] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine,* vol. 17, no. 3, p. 37, 1996.

[30] A. McAfee and E. Brynjolfsson, "Big data: the management revolution," *Harvard business review,* vol. 90, no. 10, pp. 60-68, 2012.

[31] T. Khabaza. (2010). *Nine Laws of Data Mining*. Available: http://khabaza.codimension.net/index_files/9laws.htm

[32] O. Laudy. (2015). *New Standard Methodology for Analytical Models*. Available: http://www.kdnuggets.com/2015/08/new-standard-methodology-analytical-models.html

[33] J. Highsmith and A. Cockburn, "Agile software development: The business of innovation," *Computer,* vol. 34, no. 9, pp. 120-127, 2001.

[34] K. Beck *et al.*, "Manifesto for agile software development," 2001.

[35] S. W. Ambler, "Examining the agile manifesto," http://www. *ambysoft. com/essays/agileManifesto. html>. Acesso em,* vol. 8, no. 04, p. 2010, 2006.

[36] K. Conboy and B. Fitzgerald, "Method and developer characteristics for effective agile method tailoring: A study of XP expert opinion," *ACM Transactions on Software Engineering and Methodology (TOSEM),* vol. 20, no. 1, p. 2, 2010.

[37] L. Williams and R. Upchurch, "Extreme programming for software engineering education?," in *Frontiers in Education Conference, 2001. 31st Annual*, 2001, vol. 1, pp. T2D-12: IEEE.

[38] S. Alliance, "What is Scrum? An Agile Framework for Completing Complex Projects-Scrum Alliance," *Scrum Alliance. Available at:* https://www. *scrumalliance. org,* 2016.

[39] K. Schwaber, "Scrum development process," in *Business object design and implementation*: Springer, 1997, pp. 117-134.

[40] H. Takeuchi and I. Nonaka, "16 The new new product development game," *Japanese Business: Part 1, Classics Part 2, Japanese management Vol. 2: Part 1, Manufacturing and production Part 2, Automotive industry Vol. 3: Part 1, Banking and finance Part 2, Corporate strategy and inter-organizational relationships Vol. 4: Part 1, Japanese management overseas Part 2, Innovation and learning,* vol. 64, no. 1, p. 321, 1998.

[41] J. Henry and S. Henry, "Quantitative assessment of the software maintenance process and requirements volatility," in *Proceedings of the 1993 ACM conference on Computer science*, 1993, pp. 346-351: ACM.

[42] N. R. Interactive, "Agile Scrum framework," ed, 2017.

[43] K. Schwaber, *Agile project management with Scrum*. Microsoft press, 2004.

[44] J. Sutherland, "Agile development: Lessons learned from the first scrum," *Cutter Agile Project Management Advisory Service: Executive Update,* vol. 5, no. 20, pp. 1-4, 2004.

[45] P. Deemer, G. Benefield, C. Larman, and B. Vodde, "A lightweight guide to the theory and practice of scrum," *Ver,* vol. 2, p. 2012, 2012.

[46] M. Cohn, *User stories applied: For agile software development*. Addison-Wesley Professional, 2004.

[47] L. Buglione and A. Abran, "Improving the user story agile technique using the invest criteria," in *Software Measurement and the 2013 Eighth International Conference on Software Process and Product Measurement (IWSM-MENSURA), 2013 Joint Conference of the 23rd International Workshop on*, 2013, pp. 49-53: IEEE.

[48] E. Ibrahim, *A case study of Texas regional education service center multicultural/diversity trainers' perception of teacher resistance and structural barriers to multicultural education*. Texas A&M University, 2007.

[49] M. Verma *et al.*, "Dynamic resource demand prediction and allocation in multi‐tenant service clouds," *Concurrency and Computation: Practice and Experience,* vol. 28, no. 17, pp. 4429-4442, 2016.

[50] K. Coussement, S. Lessmann, and G. Verstraeten, "A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry," *Decision Support Systems,* vol. 95, pp. 27-36, 2017.

[51] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news,* vol. 2, no. 3, pp. 18-22, 2002.

[52] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning,* vol. 20, no. 3, pp. 273-297, 1995.

[53] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in neural information processing systems*, 2002, pp. 841-848.

[54] D. G. Kleinbaum and M. Klein, "Analysis of matched data using logistic regression," in *Logistic regression*: Springer, 2010, pp. 389-428.

[55] L. A. Kurgan and P. Musilek, "A survey of Knowledge Discovery and Data Mining process models," *The Knowledge Engineering Review,* vol. 21, no. 1, pp. 1-24, 2006.

[56] D. Keim, G. Andrienko, J.-D. Fekete, C. Gorg, J. Kohlhammer, and G. Melançon, "Visual analytics: Definition, process, and challenges," *Lecture notes in computer science,* vol. 4950, pp. 154-176, 2008.

[57] X. Zhu and G. Xu, "Applying Visual Analytics on Traditional Data Mining Process: Quick Prototype, Simple Expertise Transformation, and Better Interpretation," in *Enterprise Systems (ES), 2016 4th International Conference on*, 2016, pp. 208-213: IEEE.

[58] M. G. Cruz, G. W. Peters, and P. V. Shevchenko, *Fundamental aspects of operational risk and insurance analytics: A handbook of operational risk*. John Wiley & Sons, 2015.

[59] K. A. Dodge and R. Haskins, "Children and government," *Handbook of Child Psychology and Developmental Science,* 2015.

[60] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 27, no. 8, pp. 1226-1238, 2005.

[61] S. Dean and B. Illowsky, "Descriptive Statistics: Histogram," *Retrieved from the Connexions Web site:* http://cnx*. org/content/m16298/1.11,* 2009.

[62] A. Inselberg, "Multidimensional detective," in *Information Visualization, 1997. Proceedings., IEEE Symposium on*, 1997, pp. 100-107: IEEE.