Enhancing the Human Sensemaking Process with the Use of Social Network Analysis and Machine Learning Techniques

A thesis submitted for the degree of Doctor of Philosophy

By

Alaa Marshan

Department of Computer Science,

Brunel University

January 2018

ABSTRACT

Sensemaking is often associated with processing large or complex amount of data obtained from diverse and distributed sources. Sensemaking enables leaders to have a better grasp of what the data represents and what insights they can get from it. Thus, sensemaking is considered extremely important in mature markets where the competition is fierce. To-date, the research base on sensemaking has not moved far from the conceptual realm, however. In response, this research provides a conceptual framework that explains the core processes of sensemaking – noticing, interpretation and action – and examines how emerging technologies such as Social Network Analysis (SNA) and Machine Learning (ML) techniques help to enhance the human sensemaking process in generating valuable insights during data analysis. Design Science Research (DSR) is adopted as a research methodology in the context of financial transactional data analysis, aiming to make sense of the data while exploring conceptions of customer value for a mainstream commercial bank alongside the perceived need for banking products. Three analytical models are introduced, examining Connected Customer Lifetime Value (CCLV), Network Relationship Equity (NRE) and product purchasing frequency based on customer 'personas'. The former models employ SNA techniques in providing novelty, the latter combines the outcomes of SNA with ML clustering algorithms to provide a base on which product holdings and purchase frequency analysis are overlaid – providing a novel form of recommendation. Ongoing evaluation of the developed models is used to explore the nuances of the sensemaking process and the ability of such models to support that process (in the given domain).

DEDICATION

Dear Mom and Dad,

No words in the words will be enough for me to express my feelings while I am writing this letter. No words in the words will be enough for me to express my gratitude for the unconditional love you showered me with, everything you have taught me, and for the values and lessons I learnt from you.

It has been several years since I last saw you, spoke to you, touched your hands, hugged you, felt your heart beats or sat in your presence. I really miss you so much. I miss your loving hugs. I miss your sweet voices. I miss your smiles and laughter. I miss you so much, and I will miss you very much during the graduation ceremony, but I know that you will be there watching me, just like I know that you were with me throughout this journey.

Dear Mom and Dad, I dedicate this work to you for all the love you have given me and for everything you taught me to be who I am now. May your souls rest in peace and heaven.

Thank you, Mom and Dad,

ACKNOWLEDGEMENTS

First and foremost, to my supervisor, Prof Mark Lycett, which without him, this PhD would have never happened. For the constant support and advice he provided me with at all times. I really cannot measure how much he helped and guided me with his knowledge and wisdom. I have become a better learner because of him. I will remain forever grateful to him.

To my brother who supported me tremendously during this journey and I know that I will never be able to pay him back, and his family, wife, beautiful daughters. May God protect and bless them all, I remain grateful for everything.

To my sister who supported me throughout this adventure, and her son and daughter, who were there when I was down and sad to bring and cheer me up, may God protect and bless them all, I remain grateful for everything.

To Dr David Bell, my second supervisor, and Dr Alan Serrano, my mentor since the Masters' degree. Deep appreciation to their advice, help and support. I remain always grateful to them.

To Dr George Foy, Dr Elisa Tosetti and Prof Sergio De Cesare who provided me with their invaluable help and support to help me finish my PhD study. I remain always grateful to them.

To Dr Athina Ioannou, who constantly helped, supported, encouraged and provided me with great advice to help me tackle the obstacles I faced throughout this journey, I remain always and forever grateful for everything.

To all my friends who supported me throughout this journey especially Dr Asmat Monaghan and my friends, the doctors at the PhD office, Roja, Zainab, Nida, Nurul, Diana, Huda and Mohammad. I remain grateful for everything.

To the university, academics, staff members and anyone I learnt from throughout the PhD journey. Special thanks to the Connect team at the university, Raj, Simon, Peter, Chris, Wayne and Simon who were like a family to me for five years. Also, special thanks to the IT support team at the Computer Science Department, Jeremy, Asif and Vinod for all their help and support. I remain always grateful.

Thank you all very much.

DECLARATION

The following papers have been published as a direct result of the research discussed in this thesis:

Lycett, M., and Marshan, A., (2016) ‘Capturing Sensemaking Pattern during Data Analysis: A Conceptual Framework.’ *Proceedings of the 25th International Conference on Information Systems Development (ISD).*, pp. 106–116.

Lycett, M., and Marshan, A., (2017) ‘Modelling Connected Customer Lifetime Value (CCLV) in the Banking Domain.’ *Proceedings of the 23rd Americas Conference on Information Systems (AMCIS).*

ABBREVIATIONS

AF	Asset Finance
ANN	Artificial Neural Networks
CCLV	Connected Customer Lifetime Value
CE	Customer Equity
CLV	Customer Lifetime Value
CM&P	Cash management and payments
CRM	Customer Relationship Management
CRV	Customer Referral Value
CSMV	Customer Social Media Value
CTA	Cognitive Task Analysis
DSR	Design Science Research
FoR	Frame of Reference
FX	Foreign Exchange
HCC	Human Centred Computing
IF	Invoice Finance
IS	Information Systems
ISDT	Information Systems Design Theory
ML	Machine Learning
NCLV	Network Customer Lifetime Value
NDM	Naturalist Decision Making
NRE	Network Relationship Equity
PAM	Partitioning Around Medoids
RE	Relationship Equity
SME	Small and Medium Enterprise
SNA	Social Network Analysis
SVM	Support Vector Machine
WSS	Within-cluster Sum of Square

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

1.1 Overview

This chapter provides an overview of this research and the project associated with it, which both investigate how scientific advances in automated technologies can support human sensemaking process to generate valuable insights during data analysis. This chapter sets the basis for the research and describes how it was conducted.

The structure of the chapter is as follows. Section 1.2 provides background information about the research problem and explains the motivation behind this research. Section 1.3 defines the research aim and objectives that address the needs of this project. Section 1.4 describes the research process adopted in the study. Finally, Section 1.5 presents the thesis structure and shows a diagram that summarises the contents of the thesis.

1.2 Background and Motivation

Sensemaking is a relatively new concept that mainly emerged from the work of Weick (1995), who focused on the domain of organisational behaviour. The term sensemaking refers to decision making in ambiguous or unclear situations as it requires the processing of complex and large amounts of data from diverse, unrelated and independent resources. Sensemaking is concerned with studying how people comprehend and analyse events and data that are characterised by ambiguity and equivocality. At its simplest, sensemaking can be defined as the interrelated recurring processes of noticing, interpreting and action (Jeong and Brower, 2008). Noticing is a process in which individual actors single out some problematic stimuli as cues for further conscious processing out of their streams of experience in the situations that they face. Interpretation is a combining process in which the cue and data are connected to a frame of reference, through which meaning of the cue is constructed and hypotheses are generated. Lastly, action represents the 'work' of testing the generated hypotheses, and is motivated by the goal of the sensemaking process and at the same time makes it real. Action often triggers subsequent sensemaking processes (noticing and interpretation) (Jeong and Brower, 2008).

Like other essential factors of production such as hard assets and human capital, it is increasingly the case that much of modern economic activity, innovation, and growth simply could not take place without making sense of the available data. Uncertain decision problems within organisational and social realms, however, are often difficult to explicitly model, to (completely) formalize or make sense of – often due to the constraints within organisations, people, data, technology, functionality, time, budget and resources (Singh and Singh, 2012). The volume of information is increasingly measured in hundreds of gigabytes or terabytes and provides a potentially valuable resource for insight into the factors affecting people's preferences (Chen, Chiang and Storey, 2012). Additionally, the diversity of such resources and the diverse natures of the data show how 'Big Data' is different from regular and structured data (Davenport, Barth and Bean, 2012). Decision criteria and alternatives, thus, are burgeoning due to the continuing increase in the Variety, Volume, Veracity, Velocity and Value (5Vs) of (big) data (Davenport, Barth and Bean, 2012). Hence, a challenge in relation to making sense of data is coming to the fore (Chen, Chiang and Storey, 2012): 'Big data' arguably requires 'big insight'.

The latter aspect is important as there is a suggestion that a paradigm shift is underway, fusing methods associated with social science such as network analysis with those of data mining and machine learning in order for humans to be able to make sense of big data (Klein, Moon and Hoffman, 2006a). This shift is oriented at developing systems that have the capability to:

- Create meaningful inferences by fusing massive volumes of data
- Give people the ability to achieve insights and to access others' intuitions

- Present information in relevant ways that help infer the hypotheses that the human is considering applicable to the situation under investigation

Sensemaking is the term that combines these features under one umbrella. Intelligent systems equipped with the features listed above, arguably, can enhance human sensemaking ability to gain valuable insights from data analysis (Klein, Moon and Hoffman, 2006a). To date, however, the discussion and the research base on sensemaking have not moved far from the conceptual realm.

More specifically, from an operational perspective, there are few studies that attempt to operationalise the sensemaking process to make it fit for real-life (big) data analysis situations. Exceptions include a data/frame model presented by Orasanu and Connolly (1993), theory-inspired design principles for sensemaking support systems provided by Seidel *et al.* (2017), and a method for reconstructing analysts' reasoning process from their interaction logs with data analysis and visualisation software during complex sensemaking tasks proposed by Kodagoda *et al.* (2017). What these models lack, however, is a demonstration of how emerging technologies enhance the human sensemaking process in business analysis contexts. As the state-of-the-art stands, there is an increasing focus from a research perspective (Seidel *et al.*, 2017; Muhren and Van de Walle, 2010; Klein, Moon and Hoffman, 2006b) on how automated techniques such as data fusion techniques (Klein, Moon and Hoffman, 2006a) and machine learning algorithms (e.g., classification and clustering) (Chen, Chiang and Storey, 2012) can be utilised to enhance sensemaking and to get useful insights during data analysis.

1.3 Research Aim and Objectives

Given the overview above, the aim of this research is:

> *To explore how human sensemaking and (automated) analytical techniques can work together to improve the generation of insights from the use of business analytics*

The objectives of the work are to:

Objective 1: Review the literature in order to understand the state-of-the-art in sensemaking.

Objective 2: Develop a framework that captures the core sense making processes that support data exploration and analysis.

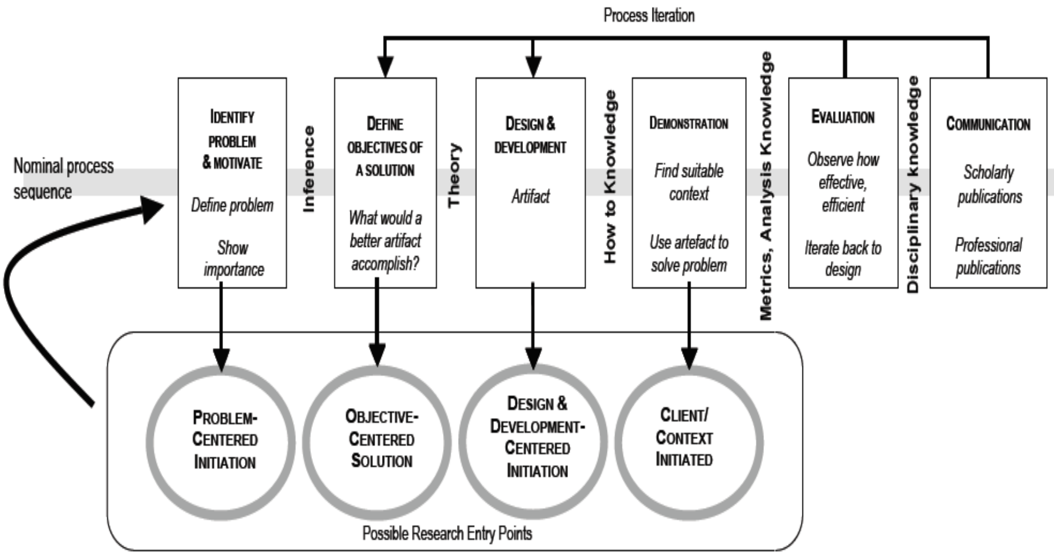
Objective 3: To empirically examine that framework to explore the relationship between the human sensemaking and (automated) analytical techniques (in a financial services context).

Objective 4: To evaluate the insights gained from empirical application.

1.4 Research Approach

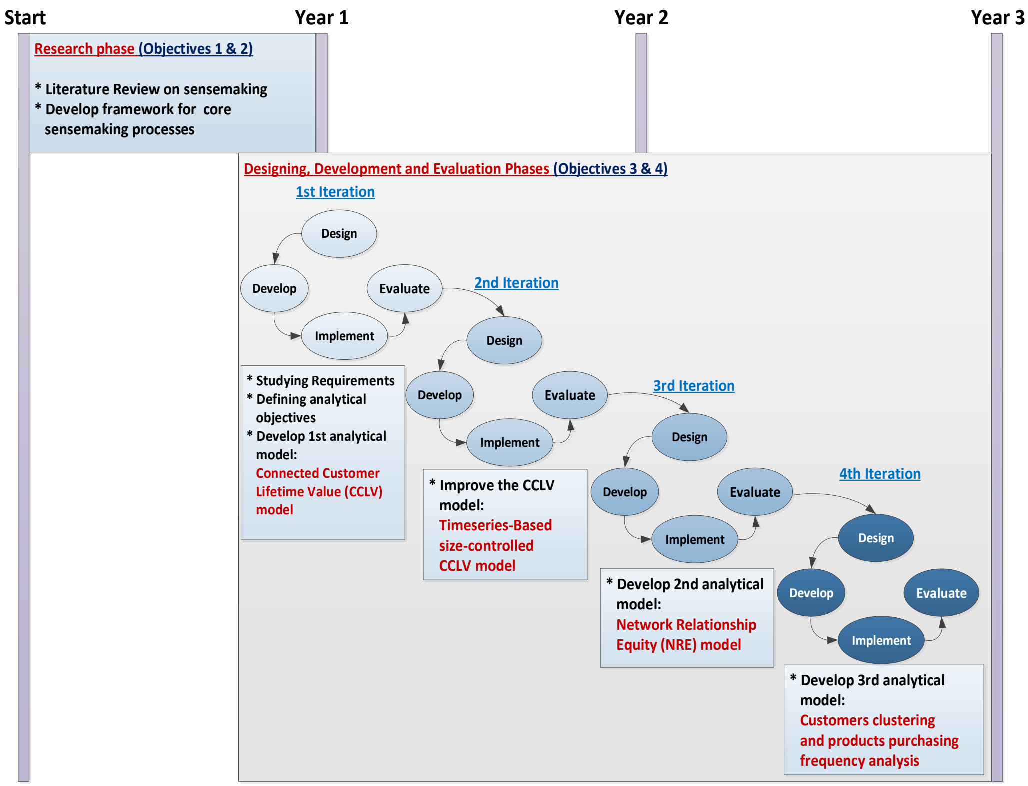
Design Science Research (DSR) is adopted as the methodological approach for the work here (Vaishnavi and Kuechler, 2013; Zahedi and Sinha, 2010; Mingers, 2001; Guba and Lincoln, 1994). A DSR approach suits the aim of this project because the emphasis is on gradually developing improved knowledge and business analytic models. The DSR process is structured into phases and involves incrementally designing, developing and evaluating the knowledge acquired during the course of the DSR process (Peffers et al., 2007; Vaishnavi and Kuechler, 2007). Figure (1.1) illustrates the general process model for Design Science Research Methodology (DSRM) proposed by Peffers *et al.* (2007).

Figure 1.1: DSRM Process Model (adapted from (Peffers *et al.*, 2007))



The DSR methodology is utilised to create artefacts that are founded on (semi) automated analytical techniques, helping to provide a path to successfully understand the problem or the analytical requirements (Peffers *et al.*, 2007). Consequently, successive design, develop, implement, and evaluate phases are iterated until the aim and objectives of this research are deemed have to beeen met. The design phase focuses on investigating the problem in more details in order to provide a better understanding of the required outcomes and to define clear objectives of the intended artefacts. The development phase builds on the objectives set in the design phase to develop the artefact that can address the problem. Finally, the implement and evaluate phases involve instantiating the artefact and evaluating the outcome in preparation for the next DSR iteration. Figure (1.2) illustrates the four iterations involved in this research.

Figure 1.2: Research Iterations



The first iteration in Figure (1.2) is about understanding the problem in more detail and defining the objectives of the required solution. Then, the first analytical model, the Connected Customer Lifetime Value (CCLV) model is designed, developed, implemented and evaluated with the concerned stakeholders from a mainstream banking organisation (referred to as BankCo from this point). The second iteration considers the evaluation outcome of the developed model and involves another cycle of design, development, implementation and evaluation in order to improve the first model by using the size of the firms as a controlling variable and making it a timeseries-based CCLV model. Following the same pattern, the third iteration builds on the evaluation outcome from the second iteration in order to develop a second model, the Network Relationship Equity (NRE) model, which meets another analytical requirement. Finally, within the fourth iteration, the final cycle of design, development, implementation and evaluation utilises emerging automated technologies to develop a third analytical model, called the persona-products purchasing behaviour model. The aim of the third model is to understand how customers acquire banking products and recommend new products to them. The four iterations aim at producing ‘satisficing’ artefacts (Simon, 1996), which can address the analytical requirements requested by BankCo.

1.5 Thesis Structure

Figure (1.3) illustrates the full structure of this thesis. The rest of this thesis is structured as follows:

Chapter 2 explains the similarities and differences between decision making and sensemaking in the organisational realm. The chapter provides an overview of the decision-making process within organisations, as well as the cognitive nature of this mental process and its relation to sensemaking. It is argued that naturalistic decision provides the basis for human sensemaking. The chapter then provides an in-depth review of the literature related to sensemaking itself, from three perspectives: Organisational, enacted and computational sensemaking. The chapter synthesises this literature into a conceptual framework that fuses the core processes (cycle) of sensemaking dynamically, to show how, together, they continually form and reform a Frame of Reference (FoR).

Chapter 3 describes the research methodology adopted to conduct this study. As mentioned before, the DSR paradigm is employed in this study as the methodological framework. The processes and outputs of the DSR iterations are then explained. Additionally, the case study used in this research is introduced in this chapter. The chapter concludes by explaining the phases and tasks performed to fulfil the aim and objectives of this study.

Chapter 4 presents the first iteration of this project. During the analysis and design in this phase, further insights are acquired into the problem domain. In addition, the specifications of the required solutions are defined. The resulting artefact, embodied in a Connected Customer Lifetime Value (CCLV) model, represents the outcome of the first iteration which is an analytical solution that is based on network analysis methods and techniques and can meet part of the analytical requirements that are based on the transactional data of BankCo customers. This model can help BankCo value their customers based on their spend in the BankCo network of customers.

Chapter 5 represents and explains the second iteration of this research, which is about improving the model developed in the first iteration by understanding the transactional data in more detail and constructing the model in a time series format that can help monitor the changes in CCLV score over the time.

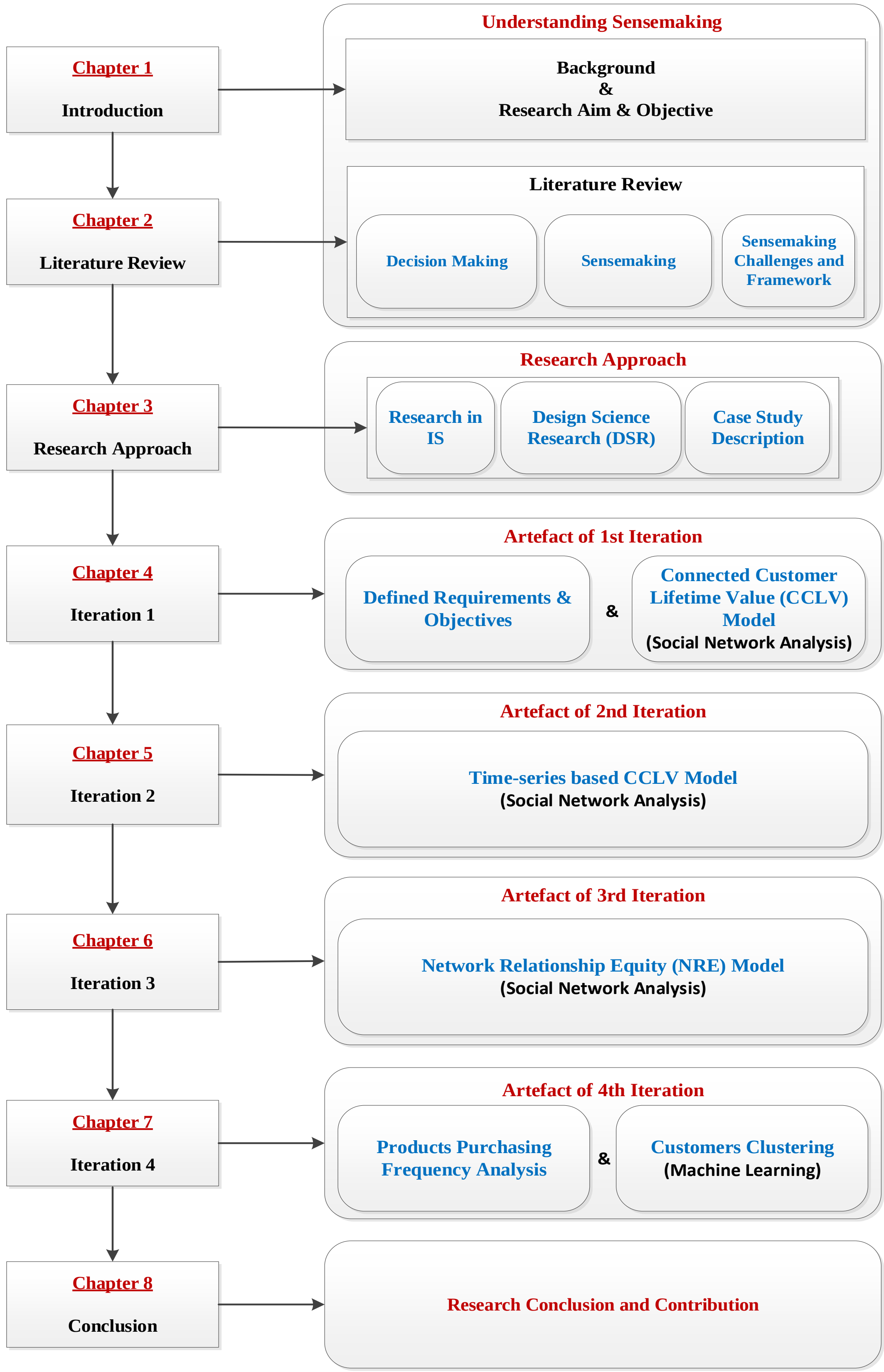
Chapter 6 discusses the third iteration of this research. This iteration is about designing, developing, implementing and evaluating a second analytical model – Network Relationship Equity (NRE) model – that can meet further requirements presented by BankCo. This model can help BankCo monitor its relationship with its customers in terms of increase and decrease in customers' activity over time and act upon it retrospectively in order to prevent losing valuable customers.

Chapter 7 presents the fourth iteration of this research. This iteration focuses on combining network analysis and machine learning techniques in order to address the last part of the requirements that are important for BankCo. First, Machine Learning (ML) is used to perform customer clustering. Second, an analysis of banking products purchasing frequency is performed. Third, the two analyses are combined in order to study the purchasing behaviour within each resulting cluster.

Core sensemaking processes (noticing, interpretation and action), which form the sensemaking framework synthesized in Chapter 3, are demonstrated throughout the four iterations reported in Chapters 4, 5, 6 and 7.

Chapter 8 presents the research conclusions and findings. Additionally, it highlights the key contributions made by this study and, finally, discusses its limitations in order to draw up future research directions.

Figure 1.3: Thesis Structure



Chapter 2: Literature Review

2.1 Overview

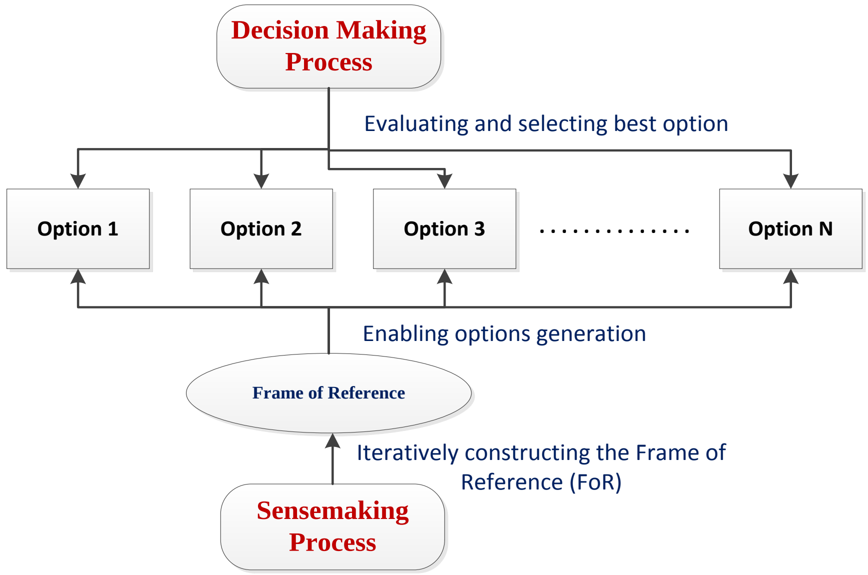
This chapter provides a literature review related to this research and covers a contextual background that span topics related to decision-making process strategies in business contexts paving the way for a detailed review of sensemaking in organisations and the different perspectives that relate to sensemaking.

The chapter is structured as follows. Section 2.2 introduces the difference between decision-making and sensemaking. Section 2.3 reviews decision-making process within organisations including the cognitive nature of decision-making and naturalistic decision-making. Section 2.4 provides a detailed review of the literature about the different perspectives or school of thoughts related to sensemaking within organisations. Section 2.5 explains the importance and challenges of sensemaking in modern business environments. Section 2.6 presents the conceptual sensemaking framework that helps in defining the core processes of sensemaking. Finally, Section 2.5 provides an overall summary of the chapter.

2.2 Decision Making and Sensemaking

The comprehension of data and information to make decisions is based on human-based cognitive processes of sensemaking and decision-making, which both have different perspectives when it comes to human behaviour (Boland, 2008). What brings them together, however, is the human being – since he or she is the performing actor in both processes. Decision-making is concerned with evaluating alternative approaches and making a choice among them based on predefined criteria. Sensemaking, on the other hand, is concerned with using past experiences as meaningful resources for future reference in order to understand and make sense of unknown and ambiguous facts and data (Lycett, 2013). The sensemaking process is based on the prior experiences encountered by a human being that help to form a frame of reference (Boland, 2008). Sensemaking can be perceived as the process that precedes and overlaps with decision-making to generate the satisficing options that need to be analysed and judged during the decision-making process to choose the best possible one, see Figure (2.1).

Figure 2.1: Decision making versus Sensemaking



The following section discusses the decision-making process and the cognitive nature of this process leading to the study of Naturalistic Decision Making (NDM).

2.3 Decision-Making Process in Organisations

Decision making is a continuous process that requires decision makers to evaluate the circumstances or problems and to study the alternatives in order to choose the best possible option (Pugh and Hickson, 2007). This mental process depends on having the correct information at the right time in order to pick the right choice

(Pugh and Hickson, 2007). Todd (2007) classifies the different schools of thought on decision-making into threefold:

- First, the view of ***unbounded rationality*** emphasises the importance of having as much information as possible, despite a human's limited ability to process the gathered information.
- Second, the view of ***bounded rationality*** is considered as the foundation of how people make reasonable decisions based on simple heuristics, given the limitations in time, information, and processing power.
- Finally, driven by the role of the environment in limiting and/or empowering human cognition capacity, the view of ***ecological rationality*** studies how humans adapt to the limitations imposed by the environment and use simple mental and cognitive mechanisms to make rational and sensible decisions by exploiting the structure and constraints inherent in the data and environment where the decisions are used.

The view of ecological rationality highlights the need to study the cognitive nature embedded in humans' decision-making process, which is covered in the next sub-section

2.3.1 The Cognitive Nature of Decision Making Process

Decision-making involves several cognitive abilities such as attention, memory, judgement, evaluation, computation, problem solving, comprehension and sensemaking. Decision makers use these abilities in a recursive manner until reaching the best possible decision (Wang and Ruhe, 2007; Allard-Poesi, 2005). These abilities can be categorised into four groups that represent decision-making strategies, which are intuitive, empirical, heuristic and rational. Table (2.1) summarises these categories, their strategies and the selection criteria.

Different decision strategies may be selected in the same situation or environment based on the decision makers' values and attitudes towards risk as well as their predictions of future outcomes based on their perceived experiences, judgments and biases (Wang and Ruhe, 2007). Despite the chosen strategy, however, Wang and Ruhe (2007) highlight the significance of preserving the decision-making experience in a form of heuristic feedback of known solutions in each of the categories of decision strategies, which leads to simpler cognitive-based decision-making tasks in the future (Wang and Ruhe, 2007). In other words, past experiences help in creating a frame of reference that can be recalled when facing new ambiguous situations that require human sensemaking.

Table 2.1: Taxonomy of strategies and criteria for decision-making (adapted from (Wang and Ruhe, 2007))

Category	Strategies	Criteria
Intuitive		
	Arbitrary	Based on the most easy or familiar choice
	Preference	Based on propensity, hobby, tendency, expectation
	Common senses	Based on axioms and judgment
Empirical		
	Trial and error	Based on exhaustive trial
	Experiment	Based on experiment results
	Experience	Based on existing knowledge
	Consultant	Based on professional consultation
	Estimation	Based on rough evaluation
Heuristic		
	Principles	Based on scientific theories
	Ethics	Based on philosophical judgment and belief
	Representative	Based on common rules of thumb
	Availability	Based on limited information or local maximum
	Anchoring	Based on presumption or bias and their justification
Rational		
Static		
	Minimum cost	Based on minimizing energy, time, money
	Maximum benefit	Based on maximizing gain of usability, functionality, reliability, quality, dependability
	Maximum utility	Based on cost-benefit ratio
	- Certainty	Based on maximum probability, statistic data
	- Risks	Based on minimum loss or regret
	- Uncertainty	
	* Pessimist	Based on maximin
	* Optimist	Based on maximax
	* Regretist	Based on minimax of regrets
Dynamic		
	Interactive events	Based on automata
	Games	Based on conflict
	- Zero sum	Based on Σ (gain + loss) = 0
	- Non zero sum	Based on Σ (gain + loss) $\neq$ 0
	Decision grids	Based on a series of choices in a decision grid

Considering the amount of information required to make good decisions, Wang and Ruhe (2007) put emphasis on the importance of having Cartesian Product phenomena. This means that the decision-making process would result in a better performance by having as many alternatives and selection criteria as possible. Conversely, Todd (2007) argues that the plentiness of information is not the key to a better decision since it might overcomplicate the decision-making process by having too much irrelevant or trivial data to consider. Rather, he provided a taxonomy of four classes of simple heuristics that use limited information, which are: recognition-based heuristics; one-reason decision mechanism; multiple-cue elimination strategies; and quick sequential search mechanism. Additionally, he supported this categorisation with many examples where people can use simple, fast and frugal heuristics to make good decisions with little information in a variety of domains. Todd (2007) emphasised the importance of the structure of the information in different decision environments and how this can be mapped to the structure of the decision heuristics and consequently the sensemaking process. In other words, sensemaking is about comprehending the norms of the rational decisions and actions taken by people in complex real-world situations. This point of view is further supported by the study presented by Bryant (2002) in which he also explained how properly structured information can enable fast and frugal heuristics to be employed to make better and more accurate decisions with fewer cues and fewer computations compared to complex statistical models with much more data and information (Gigerenzer and Todd, 1999).

The different approaches people use information to make decisions in complex real-world environments relate to the way they make sense of the context, the environment and the data related to the ambiguous situations they are trying to understand and make sense of. This leads to the natural way people make rational decisions in dynamic settings exemplified in today’s business world. Thus, the next sub-section focuses on Naturalist Decision Making (NDM), which is mainly related to the ecological rationality school of thought discussed in the previous section and the cognitive ability of humans to adapt to the limitations imposed by the data and the environment.

2.3.2 Naturalistic Decision Making and Sensemaking

Naturalistic Decision Making (NDM) studies provide approaches to understand how complex, uncertain, organisational-bounded and time-restricted situations affect people's ability to perform cognitive tasks and make rational decisions (Klein, 2008). In addition to the NDM perspective, the psychological and Human Centred Computing (HCC) perspectives were examined by Klein, Moon and Hoffman (2006a) as the viewpoints that define sensemaking. From the psychological viewpoint, sensemaking can be defined as how people make sense out of their experience in the world. In this viewpoint, the usual common terms such as creativity, curiosity, comprehension and mental modelling can participate in the formation of sense. None of them, however, is considered enough to provide interpretation of sensemaking as a non-stop work to understand information, cues and data, and to link them together, so it would be possible to act upon them. In the Human Centred Computing (HCC) viewpoint, Klein, Moon and Hoffman, (2006a) argue that features like data fusion might not play a good role in sensemaking since the human decision-making process needs more information to be as precise as possible. Finally, Naturalistic Decision Making (NDM) offers an empirical foundation that can be used to catch and explain theoretical reflections on the sensemaking process in the form of examples and results. In this perspective, Sensemaking is considered as the agent who delivers the following functions:

1- Leads and helps to create an overall comprehension of the situation
2- Helps people to test and improve their explanation of situations with uncertainty
3- Helps to perform a retrospective analysis to get insights from past events and decisions for better ones in the future
4- Helps to predict difficulties which could happen in the future as well as to realise problems and concerns
5- Gives options to contemplate various alternatives
6- Helps with information exploration
7- Helps to promote collaborative work

Naturalistic Decision Making (NDM) has become established as the methodological and theoretical perspective that looks at how rational decisions are taken and how sensemaking happens in a human context and how to have intelligent systems with sensemaking capability. Many fields of practice have adopted NDM as a doctrinal framework because it describes how practitioners actually comprehend information and make decisions in complex and dynamic domains rather than how they ought to make decisions (Shattuck and Miller, 2006).

NDM research focuses on the context of the situation under investigation and how to utilise past experiences of complex conditions in order to provide more insights for those interested in studying the factors that affect the human decision-making process. As a result, NDM research aims to have an improved performance in a number of fields including military doctrine as well as the development of decision support information technologies (Klein, 2008). NDM emphasizes the importance of the accumulated experience which enables people to rapidly classify situations and reach an effective verdict in a complicated and constantly-changing set of events.

Furthermore, NDM is considered as a descriptive approach that is employed to explain human sensemaking based on performances from previous experiences (Bryant, 2002). NDM deals with constraints related to limited time, stress and insufficient knowledge about the complex environment. This approach has a strong partnership with many cognitive processes such as Sensemaking, recognition, situational awareness, pattern matching, story building and mental simulation. However, since understanding NDM from a scientific point of view has many limitations, bounded rationality can offer an approach to make use of fast and frugal heuristics to develop computational models that conform to the general framework of NDM. Considering the strong relationship between NDM research and sensemaking studies, the NDM-based approaches provide means to understand sensemaking and its processes within organisations, which will be covered in the following section.

2.4 Sensemaking in Organisations

Sensemaking can be seen as the process that generates the heuristics, cues, options and hypotheses, which together form a *frame of reference* that is (implicitly) consulted when a new (ambiguous) situation triggers the sensemaking process (Llinas, 2014). Sensemaking provides the ability to deal with rapidly emerging threats as well as asymmetric, unfamiliar, and dynamic situations. Though Weick's work (Weick, Sutcliffe and Obstfeld, 2005; Weick, 1995) is pivotal in the organisational realm, sensemaking has reared its head in other domains. Thus, in order to formulate a comprehensive understanding of sensemaking, the following subsections discuss sensemaking from business, enacted cognition and computational perspectives.

2.4.1 The Perspective of Organisational (Strategic) Sensemaking

The prominent perspective of sensemaking in the business-related literature is presented by the research of Weick (1995). His work provided a fundamental set of basics and perceptions of sensemaking within organisations. He defines sensemaking as the process where people generate their own understanding and interpretations of certain situations by utilising their past experiences and interacting with their environments. Therefore, Sensemaking can be seen as a continuous retrospection, where beliefs, implicit assumptions, stories from the past, unspoken premises for decision, action, and ideas about what will happen as a result of what can be done are gathered to form an acceptable understanding or sense that is described with clear words. Nevertheless, the generated interpretation (sense) would be affected by selective perception since only some aspects of the human mind would be considered while others would be forgotten or neglected.

According to Weick (1995), organisational Sensemaking has at least seven distinguishing features, which are:

1- ***Sensemaking is grounded in identity construction*** because sense makers continually redefine their perception and awareness of themselves.
2- ***It is a retrospective process***, in which, a never-ending reconstruction of experience occurs. This characteristic is considered the most distinguishing one of Sensemaking.
3- ***Enactive of sensible environments***, because people interact with the environment that is based on knowledge gained through physical actions and individuals' skills. By doing so, they create (enact) a part of the very environment they face, including opportunities and constraints. Such action, however, becomes more difficult in crisis conditions, since the attempts to understand the circumstance surrounding the crisis often deepens the crisis (Weick, 1988).
4- ***Social***, because it occurs with and in relation to other people inside and outside the organisation.
5- ***On-going***, because Sensemaking is a continuous process.
6- ***Focused on extracted cues***, hence, it is shaping from familiar points of reference for other people. Controlling these cues is a source of power because controlling what others respond to helps in framing the view they will have and what they will do.
7- ***Driven by plausibility rather than accuracy***. People tend to go along with what is plausible and credible to them even if it cannot be checked. In other words, they follow their hunch. Because an equivocal and changing world has always moved on before a precise account of it can be formulated, however, absolute accuracy is impossible. Hence, accuracy takes second place to acceptability, to a version good enough to guide action for the time being.

Weick provides an example of the knitwear industry in Scotland to explain the seven characteristics of Sensemaking given above in more detail. This example is about small manufacturers of cashmere sweaters, where the managers of each one of them consider their products as having a unique identity that is distinguished by the colour and design of the product. The industry claims to have a strategy that has developed retrospectively from the experience of sales agents, who acquire feedback information about the products they sell and provide manufacturers with it. Thus, those agents constantly re-enact their environment, affirming this by social contacts in and between firms; forming a continual on-going process during which cues from designers, trade shows, and shops, as well as from the agents, reinforce the particular way in which the situation is perceived and so sustain its plausibility (Weick, 1995).

However, Weick argued that in younger organisations with qualified professionals, sensemaking has more flexible range, especially when innovative, non-routine decisions are to be made. Nevertheless, Sensemaking becomes less generic and more fragmented in organisations that follow hierarchical structures with separated departments and functional units that form self-managed teams. Though, whatever the form of organisation, some of its elements will be tightly coupled together, whereas the coupling of others will be comparatively loose. This means that if some parts or activities in an organisation change, the effect of this on other parts or activities will be limited, slow to show, or both. The mutual influence of loosely coupled systems is low (Weick, 1995). Loose coupling facilitates adaptation. In a loosely coupled organisation, there can be differential change, which means that some aspects change faster or more frequently than others, so that there is a flexible response by the organisation. This behaviour happens because bonds within loosely coupled subassemblies are stronger than those between them (e.g., within workgroups or departments, as against between workgroups or departments). Consequently, there is both stability and flexibility (Weick, 1995).

Moreover, Weick debated that whatever the form of business, it will have to work in circumstances that are characterised with ambiguous, uncertain, equivocal, and changing information where organisations and those who manage them go through guesswork, subjectivity, and arbitrariness. Weick believes that language could better reflect these ambiguous circumstances by making more use of verbs and less of nouns. In other words, he gives more importance to managing and organising rather than management and organisation (Weick, 1995).

Another conceptualisation of sensemaking within the organisational realm is provided by Jeong and Brower (2008), who describe it as the three interrelated processes of noticing, interpretation and action. Noticing is a process where people determine the causes for problems in different situations based on their previous experiences and use them as cues for further conscious processing. Interpretation is the process of connecting the discovered cues to a frame-of-reference of a generalised point of view about how the cue is constructed. Last, action represents the work of evaluating the cues and hypotheses generated from the interpretation process to achieve the future analytical goal. These processes form an important mechanism to generate the raw materials (data and cues) necessary for next iteration of meaning construction and reframing (Jeong and Brower, 2008).

Recent studies have concentrated on the social nature of organisational sensemaking. These studies have focused on language, rather than cognition as the enabler of sensemaking (Maitlis and Christianson, 2014; Maitlis, Vogus and Lawrence, 2013; Colville, Brown and Pye, 2012; Maitlis and Sonenshein, 2010; Maitlis, 2005). Similarly, emotion as a dimension is also receiving some attention (Maitlis, Vogus and Lawrence, 2013), where negative emotion, in particular, is considered of particular importance in crisis situations as well as organisational change (Maitlis and Sonenshein, 2010; Dougherty and Drumheller, 2006). Negative emotion can take the form of fear, desperation, anxiety and panic, which can significantly affect individual cognitive information-processing ability as well as the capability to notice and extract cues (Maitlis and Sonenshein, 2010; Stein, 2004). Conversely, it has also been demonstrated that planned organisational change can generate positive emotions that help the involved stakeholders to understand the initiative behind the change (Maitlis and Sonenshein, 2010). The important aspect here, however, is that the significance of emotion stems from its role as the necessary triggering factor that initiates the sensemaking process (Dougherty and Drumheller, 2006).

Additionally, other works in the area focus on characteristics, features and functionalities such as the recurring process of creating and modifying views and visions about ambiguous issues and situations (Grimson, 2008), the importance of the past experiences in shaping the primary assessments (Grimson, 2008; Louis, 1980), providing explanations and attributes for the emergent events and predicting the following ones (Grimson, 2008; Hill and Levenhagen, 1995; Louis, 1980), the ability to build relations between views, expectations and actions (Louis, 1980), the ability to create rational accounts of the world that enable actions (Hill and Levenhagen, 1995; Starbuck and Milliken, 1988), and finally, the ability to extract , interpret and explain cues from people's environments (Maitlis and Sonenshein, 2010; Stein, 2004; Weick, 1995).

2.4.2 The Perspective of Enacted Sensemaking

The enactive sensemaking literature examines how individuals use interaction to perceive and then shape their world to create meaning and value. In order to comprehend this understanding of Sensemaking, however, it is important first to understand the enactive approach for cognition. This point of view is based on a number of mutual core principles such as the autonomous nature of the individuals, the emergence of individuals' worlds through the interaction with the original world, the embodiment of the learning process, and finally the importance of the experience to form enough comprehension of the situation (Thompson and Stapleton, 2009; Varela, Rosch and Thompson, 1991). However, despite that Sensemaking process involves emotion as much as cognition as inseparable (Thompson and Stapleton, 2009), the first is mainly overlooked in the enactive sensemaking literature.

Nevertheless, the most important aspects of Sensemaking from the enactive perspective are the autonomy, where agents continuously regenerate their own understanding of their environment they are interacting with and through (De Jaegher and Di Paolo, 2007), and adaptivity, which provides the tolerance to face and deal with the emergent and varying challenges during the communication in the environment (Di Paolo, 2005). Consequently, sensemaking can be defined as the interaction and the engagement of a system, typically the individual, with its environment where a relational process happens. This interaction is explained by

De Jaegher and Di Paolo (2007) from a cognitive science perspective as a sustainable coordination between individuals, where additional meaning and value is created through interaction.

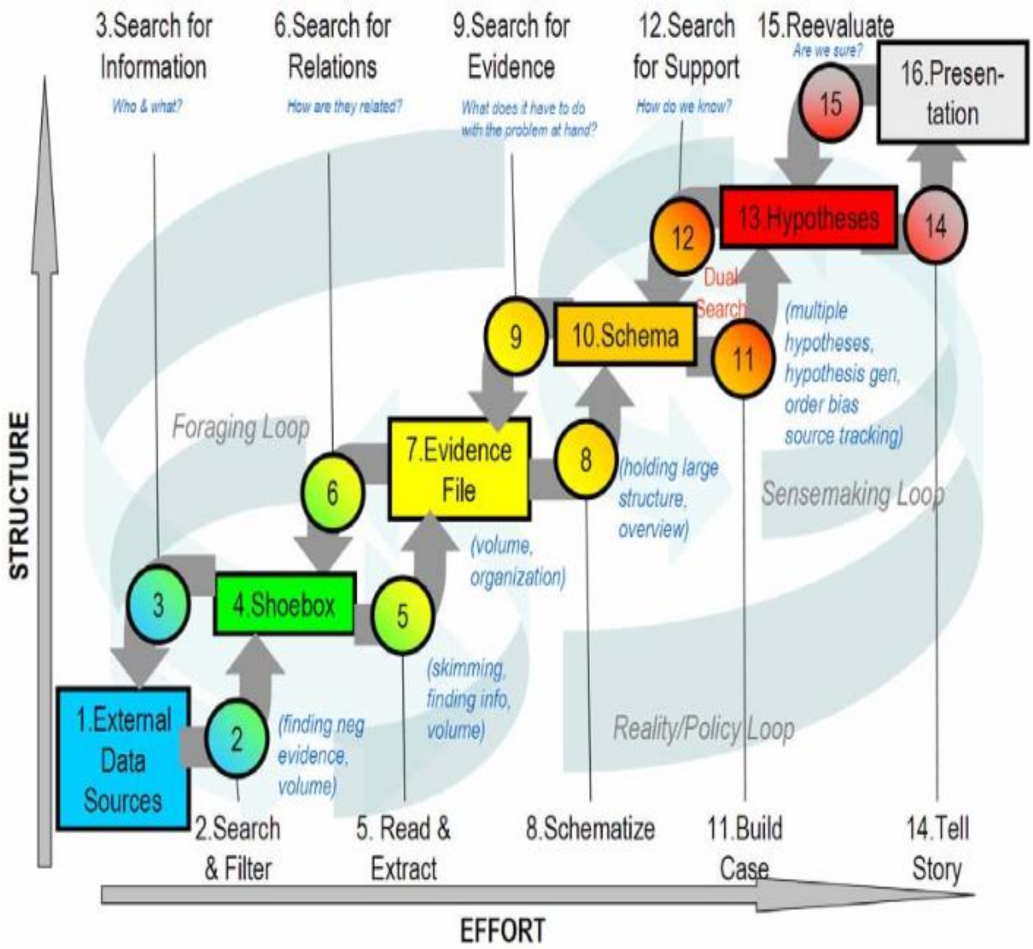
2.4.3 The Perspective of Computational Sensemaking

The computational sensemaking literature examines the phenomena in a way that seeks to operationalise it. One stream centres on situational awareness, exploring how the cognitive capability of the human mind uses non-stop creativity, curiosity, perception, comprehension, projection and mental modelling to make sense out of people's experiences – typically in the form of heuristics, cues and hypotheses (Llinas, 2014; Klein, Moon and Hoffman, 2006a; Klein, Moon and Hoffman, 2006b; Adams, Tenney and Pew, 1995). Much of the background here comes from studies focused on Naturalistic Decision Making (NDM), studying cognition in real-world environments that are characterised by ill-structured problems, uncertain and dynamic environments, ill-defined and competing goals, time stress, high stakes, multiple participants and important organisational goals (Orasanu and Connolly, 1993).

The naturalistic approach seeks to empirically trace the 'paths' that humans take in making sense of the world. In other words, using the past experiences as meaningful resources for future reference in order to understand and make sense of unknown and ambiguous facts and data. One school of thought structures this process in two main iterative loops derived from Cognitive Task Analysis (CTA) studies, see Figure (2.2). The first loop is a foraging loop, where sense makers seek information about the ambiguous problem, read through discovered resources, extract cues that can be used as evidence, create inferences and schematize them, and finally, build a story that exemplifies a case to confirm or disconfirm hypotheses that can provide schema (frame of reference) for the situation under study. The second loop is a sensemaking one, in which sense makers use the schema developed in the foraging loop to iteratively construct (generate, explore and manage) a mental model that best fits the examples and hypotheses generated in the first loop (Llinas, 2014; Pirolli and Card, 2005).

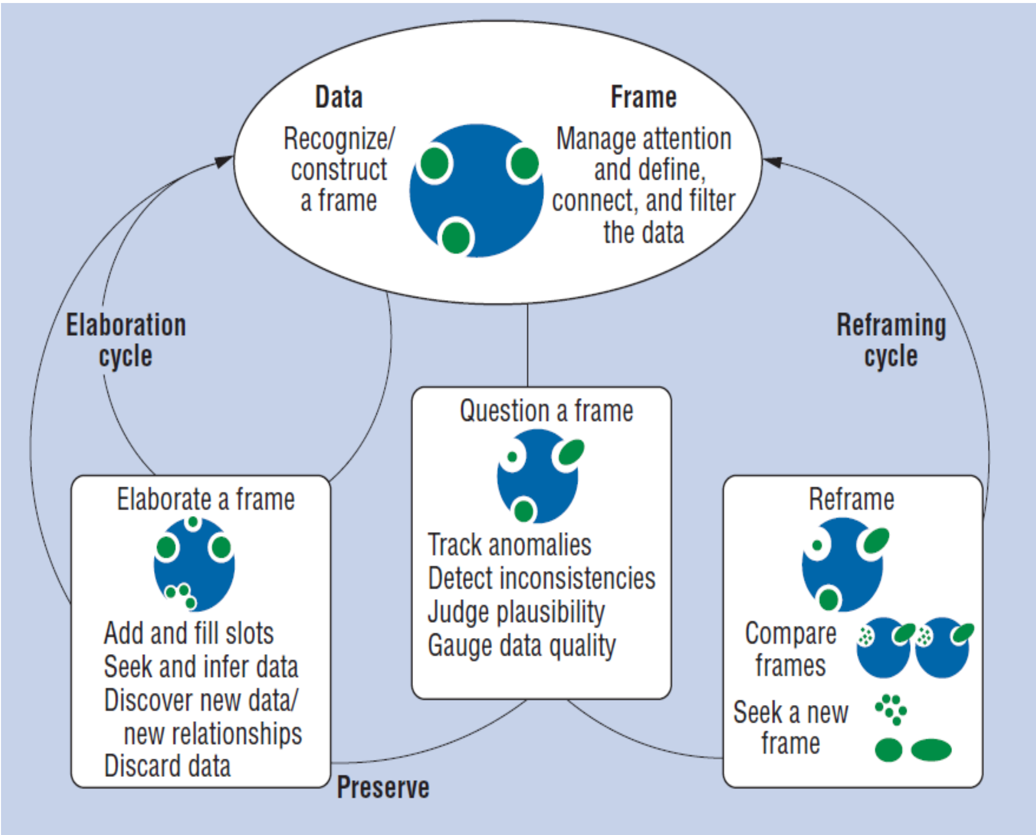
NDM utilises Cognitive Task Analysis (CTA) methods to comprehend how decision makers make complex decisions in dynamic environments (Salas and Klein, 2001). CTA is concerned with characterising decision making and reasoning skills that are related to a certain subject or situation, as well as understanding the requirements for complex information processing by eliciting and representing the necessary knowledge for this task (Li, 2005).

Figure 2.2: Notional sensemaking model (adapted from (Pirolli and Card, 2005))



A second (related) school examines how people start the sensemaking process by constructing a basic and minimal frame, which reflects a hypothesis concerning the connections among data and allows people to have some basic view on the problem at hand (Orasanu and Connolly, 1993). This frame can be further developed by the addition of details, seeking new data and relationships, and discarding useless data. Then, questioning the explanations the initial frame delivers happens by detecting anomalies and inconsistencies, measuring data quality and judging the plausibility the initial frame provides. Ultimately, this can lead to preserving or elaborating the frame as part of elaboration cycle. Alternatively, questioning and comparing the current frame with a new one can lead to a reframing cycle, in which a new, better and enhanced frame is created (Klein, Moon and Hoffman, 2006b). This model, see Figure (2.3), was examined in crisis situations and demonstrated how the loop of framing, questioning and reframing has proved its potential to model human intelligence and the sensemaking process (Moore and Hoffman, 2011).

Figure 2.3: Data/Frame Theory of Sensemaking (adapted from (Klein, Moon and Hoffman, 2006b))



As explained before, NDM is the study of cognition in real-world environments that are characterised by ill-structured problems, uncertainty and ambiguity (Orasanu and Connolly, 1993). Consequently, it provides a potential means for understanding the 'paths' that humans take in complex data analysis. The significance of having an initial frame for the sensemaking process stems from its ability to clarify what counts as cues, data and/or hypotheses. Consequently, constructing alternative reasoning paths, leading to questioning the frame followed by an elaboration of the frame or the creation of a new frame (re-framing) (Moore and Hoffman, 2011; Klein, Moon and Hoffman, 2006b). The importance of the initial frame is proved by the empirical findings resulting from testing this Data/Frame while considering casual reasoning, consideration of hypotheses, feedback and learning. Additionally, the ability to analyse feedback during the sensemaking process helps to deal with the bias associated with human sensemaking (Moore and Hoffman, 2011).

2.5 The Importance and Challenges of Sensemaking

Sensemaking can be seen as the process that generates the heuristics, cues, options and hypotheses, which together form a frame of reference that is (implicitly) consulted when a new (ambiguous) situation triggers the sensemaking process (Llinas, 2014). Consequently, sensemaking support systems are qualitatively different from (traditional) decision support systems. The latter type helps decision makers and management with known situations by facilitating the comparison of alternative solutions or decisions – effectively optimising the decision space. Sensemaking support systems, on other hand, help actors with equivocal and ambiguous problems that require constructing (and/or reconstructing) frames of reference, in order first to understand the factors that trigger the sensemaking process (Muhren and Van de Walle, 2010).

Zack (2007) argues that traditional Decision Support Systems are valuable in the context of uncertainty and complexity, but that they are lacking in the context of ambiguity and equivocality. Given that the last two are prevalent in the social and organisational realm, there is a need to help managers and decision-makers to better deal with ambiguous and equivocal (emergent) challenges – especially in rapidly changing conditions within organisations (Grimson, 2008). The terms, uncertainty, complexity, ambiguity and equivocality, however, are used interchangeably to describe difficult and unclear circumstances. Therefore, Zack (2007) distinguished them according to two dimensions: 1) the nature of what is being processed which could be either information or frames of reference, and 2) the constitution of the processing problem. Information refers to the observations that have been cognitively processed and punctuated into coherent messages. In contrast, frames of reference, presented by (Choo, 2006), are the interpretative frames which provide the context for creating and understanding information. The lack or abundance of information or frames constitutes the four different processing challenges which they are uncertainty, complexity, ambiguity and equivocality and are illustrated in Table (2.2)

Table 2.2: Information Processing Challenges (adapted from (Zack, 2007))

	Information	Frame(s) of Reference
Lack of…	Uncertainty	Ambiguity
Variety /Diversity of…	Complexity	Equivocality

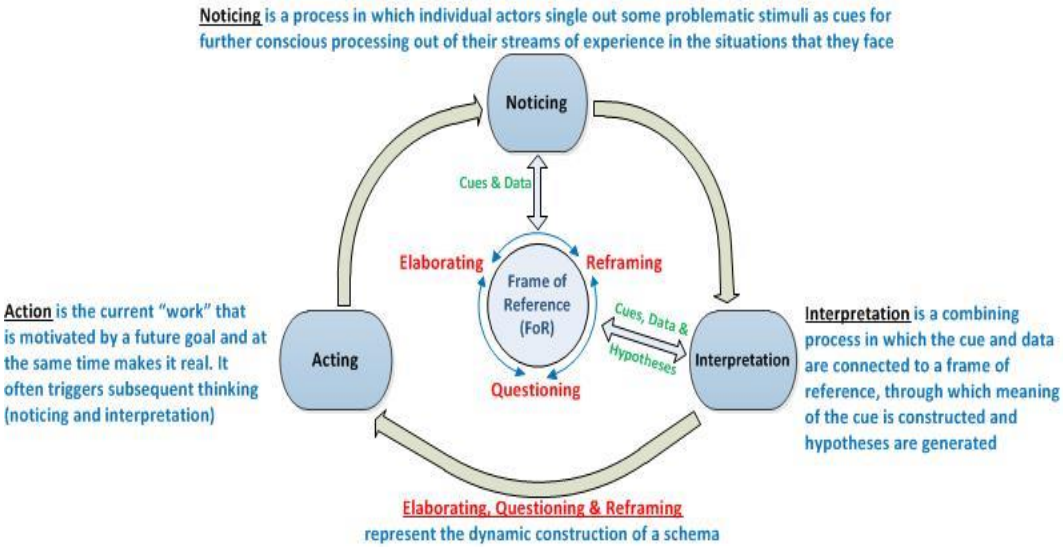
Basic human sensemaking process follows the approach explained by Bryant (2002) and Todd (2007), in which sense makers use less information and simple heuristics as well as a combination of decision-making strategies discussed earlier: intuitive, empirical, heuristic and rational to make sense of the situation they encounter as well as the data they have and/or generate throughout the sensemaking process. Given the complexity of today's businesses, the lack of Frame of Reference (FoR) in social and organisational realm, and the amount of data generated from various information systems, however, the context of ambiguity and equivocality are arguably prevalent in the social and organisational realm. Thus, there is a need to help managers and decision-makers to better deal with ambiguous and equivocal (emergent) challenges – especially in rapidly changing conditions within organisations (Grimson, 2008; Louis, 1980).

2.6 Conceptual Sensemaking Framework

Despite the fact that sensemaking accounts discussed before are generated in different domains at different times, a number of similarities arise. First, all accounts study the retrospective nature of how individuals shape their world via recurring interaction as well as the creation of meaning and value. Second, both organisational and enactive literature, examine how the new knowledge is created via a continuous (social) interaction with the environment they interact with and/or exist within. Third, the enactive and computational approaches, particularly, seek to put some 'flesh on the bones' in identifying constructs generated from the sensemaking process such as frames of reference, cues, hypotheses, etc. Thus, the computational literature proposes sensemaking models by identifying concepts such as foraging, framing, elaboration and reframing. As the state-of-the-art stands, however, there is little in the way of research that examines how current/emerging computational techniques such as Social Network Analysis (SNA) and Machine Learning (ML) techniques can enhance sensemaking process to generate the constructs discussed above.

The synthesis of the current literature is shown diagrammatically in Figure (2.4). The processes of noticing, interpreting and acting (Jeong and Brower, 2008) are considered as the core processes that inform/reform a Frame of Reference (FoR); this is a schema that contains and connects cues and, in addition, contains the hypothesis/hypotheses about the problem under investigation (Llinas, 2014; Klein, Moon and Hoffman, 2006b; Pirolli and Card, 2005). In this model, sensemaking process starts with noticing, which is the process of cue extraction in a form of questions and requirements. Then, interpretation, which is the process of structuring cues, developing hypotheses (solutions) and selecting the most likely as a precursor to take action based on the interpretations of these hypotheses. An inner loop represents the ongoing dynamics of that process via the questioning of a frame (as a reflection on action), elaboration of a frame and, ultimately, reframing it.

Figure 2.4: Conceptual Sensemaking Framework



Intelligence analysts try to identify relevant information within massive data and to synthesise these fragments into a coherent understanding of the entities, events, and the relations between them in order to characterize the unknown. This process of flexibly filtering and interpreting data and establishing relationships between scattered pieces of information is described as Sensemaking. The main goal here is to investigate how automated analytical techniques facilitate the processes of noticing and interpretation to support the generation of insights in order to make better-informed actions. The synthesised framework will be demonstrated in the following chapters of this thesis.

2.7 Summary

This chapter has explained the difference as well as similarities between decision making and sensemaking in the organisational realm. Also, it has provided a thorough review of the different types of decision-making strategies within organisations as well as the cognitive nature of this mental process and its relation to sensemaking. Additionally, it discussed naturalistic decision making as the basis for human sensemaking. Furthermore, it offered an in-depth review of the literature related to sensemaking from three perspectives: organisational, enacted and computational sensemaking. Moreover, the importance and the challenges that hinder sensemaking process have been covered.

Finally, this chapter has proposed a conceptual framework that explains the core processes (cycle) of sensemaking: noticing, interpretation and action and how they, together, form and reform the Frame of Reference (FoR). Within the context of data analysis, the sensemaking cycle begins with ***noticing***, which is about extracting cues in a form of questions and requirements that need to be addressed. Then, ***interpretation*** is the process that focuses on understanding the cues (questions and requirements) and developing hypotheses in a form of analytical solutions that can address the generated questions. Finally, ***action*** is the process that is concerned with testing the developed solution on real data. Concurrently, the inner loop within this framework is about continuously questioning (evaluating) the findings from testing the developed analytical solutions and updating (reframing) the frame of reference, ultimately, generating additional cues for the next sensemaking cycle. This cycle has a recurring nature since it recursively happens until reaching a “satisficing” answer for the generated questions and requirements and fits the real-world problem while considering the limitations in data accessibility, technology, functionality, time, budget and resources.

Chapter 3: Research Design and Approach

3.1 Overview

This chapter presents and justifies the research methodology adopted to carry out this research, the Design Science Research (DSR) methodology. Also, it sets the scene for the data analysis in the following chapters by describing the case study, the data sets used and the analytical requirements. Additionally, a summarised description of the DSR iterations is provided in this chapter.

This chapter is structured as follows: Section 3.2 discusses and highlights research paradigms used for research within the Information Systems (IS) domain. Section 3.3 provides a detailed review of the Design Science Research (DSR) paradigm employed in this research. Also, it provides a broad outline of the development research phases that exemplify DSR. Section 3.4 presents the case study and data sets used in this research. Section 3.5 explains the DSR approach for demonstrating sensemaking processes during data analysis. Finally, Section 3.6 summarises the key parts of this chapter.

3.2 Researching in the Information System (IS) Domain

The diversity that characterises the Information Systems (IS) domain makes it a multidisciplinary research field with inherited complexity from the disciplines that lend their knowledge to the IS domain (Mingers, 2001). The researchers in the IS domain, thus, need to consider multiple paradigms, methodological approaches and techniques to conduct IS studies (Baskerville and Myers, 2002; Frank, 1999; Couger, 1996). A research paradigm is described as a set of basic philosophies or beliefs that guide the researcher’s actions throughout the research life cycle

(Mingers, 2001; Guba and Lincoln, 1994). Several authors note that researching in the field of IS can be classified into three main categories or paradigms (Vaishnavi and Kuechler, 2013; Xinping, 2002; Klein and Myers, 1999; Orlikowski and Baroudi, 1991):

- **Positivist Research**: This research paradigm serves the role of testing theories with the aim of understanding a problem or phenomena under research in a better way. In such research, the researcher would produce some hypotheses based on previous research or observations, and then, the collected data is used to support the theories and assumptions made prior to investigation. The basic process in the positivist research can be described as following:

  Problem → Hypothesis → Proposition → Verification → Conclusion
- **Interpretive Research**: Generally, this research category tries to understand certain phenomena by understanding people's subjective and intersubjective meanings they create while interacting with their environments. This happens by providing interpretation for views, values, meanings, beliefs, thoughts, events and contexts. Unlike the positivist research, the collected data is used to conclude and create knowledge without making assumptions.
- **Design Science Research (DSR)**: where creating novel artefacts and evaluating them are utilised to explain and enhance the performance of the information system. The emphasis here is on generating knowledge or developing the desired artefact through several iterations (Purao, 2002).

Moreover, literature concerned with researching within the IS domain (Vaishnavi and Kuechler, 2013; Zahedi and Sinha, 2010; Mingers, 2001; Guba and Lincoln, 1994) has defined four philosophical theories or beliefs that can be used to view the previous three research perspective. These are:

- **Ontology**, which is concerned with finding a formal definition of objects in a certain domain of interest (the theory of reality or existence).

Ontological belief or theory focuses on studying the truth behind the reality as well as the originality of the reality.

- **Epistemology**, which focuses on the theory of knowledge. Epistemology theory represents one of the fundamental philosophical research areas, in which the nature, sources, limits and the certainty of the knowledge are studied.
- **Methodology**, which is the systematic set of steps and methods that can be applied to perform research. Methodology theory comprises the tools and methods that are necessary for data collection and analysis. This theory is about reasoning and creating relations between theory and practice with the aim of finding the best processes or steps that help in setting the path for generating the desired outcome.
- **Axiology**, which relates to two types of research values – ethics and aesthetics. This theory is concerned with understanding the values that affects researchers' motivations to conduct research in a specific community.

Table (3.1) represents a matrix that illustrates the relationship between the philosophical theories and research paradigms discussed above. The highlighted cell in this table represents the basic belief, "Methodology", and research paradigm, "Design", where this research falls. This is because DSR methodology involves incremental and iterative design and the development of novel or innovative artefacts with the aim to improve a current situation associated with organisational or social systems to make them more desirable through the creativity of the novel artefacts (Hevner *et al.*, 2004). This approach is consistent with the aim of this research, which is about exploring how human sensemaking can be enhanced using automated analytical techniques; leading to produce new artefacts in the form of analytical models that improve the generation of insights from the use of business analytics.

Table 3.1: Philosophical assumptions of three research perspectives (adapted from (Vaishnavi and Kuechler, 2013))

Basic Belief	Research Paradigm		
	Positivist	Interpretive	Design
Ontology	A single reality, knowledge, Probabilistic	Multiple realities, Socially constructed	Multiple, contextually situated alternative world-states. Socio-technologically enabled
Epistemology	Objective; dispassionate. Detached observer of truth	Subjective; values and knowledge emerge from the researcher-participant interaction	Knowledge through making: objectively constrained construction within a context. Iterative circumscription reveals meaning
Methodology	Observation; quantitative, statistical	Participation; qualitative. Hermeneutical, dialectical	Developmental: Measure artefactual impacts on the composite system
Axiology	Truth: universal and beautiful; prediction	Understanding: situated and description	Control; creation; progress; understanding

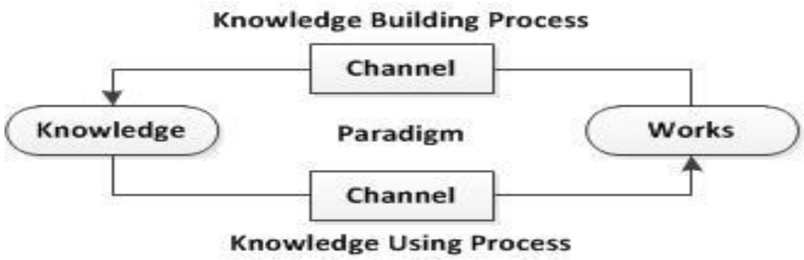
3.3 Design Science Research (DSR) Methodology

The process of design is about the creation of an applicable solution that represents an object or system that addresses a problem or requirement (Peffers *et al.*, 2007); with a focus on the technological perspective of the developed artefact and less attention to the organisational context (Sein *et al.*, 2011). The design process, in its nature, is non-linear (Gregor, 2009), reflecting a situation where the problem space defines the characteristics of the solution space, which, in turn, reframes the problem space. Thus, the essence of the design process is the repetitive (and reflective) actions of learning and acquiring knowledge while evaluating and redesigning the generated artefact. In the design process, the designer aims to reach a realistic - “satisficing” - result by responding to the evaluation of the artefact, the sudden changes in the requirements and the unpleasant results by reflecting and modifying the original design (Simon, 1996).

The "satisficing" result in the design process represents the best outcome in combinatorial problems that can be expressed not only quantitatively but also symbolically. The "Satisficing" product is preferred in such situations because of the high cost and computing capabilities required to reach to the "Optimised" one, considering the cost of fitting the real-world problem into the computational criteria and requirements (Simon, 1996). This explanation of the design process provides an indication of the difference between designing and engineering where, for example, the required artefact is to construct a building with pre-planned and known milestones in the project management plan. Thus, Simon (1996) concluded that in a design process, envisioning the design steps or events is very difficult in most of the time where design science is employed; rather, the design process should have the characteristic of adaptability to better deal with changes throughout the course of design.

The iterative process of generating and accumulating knowledge is illustrated in the general model that is introduced by Owen (1998), see Figure (3.1). This model helps to visualise the design process as a repetitive cycle of using knowledge to create works and evaluating the works to generate and accumulate knowledge.

Figure 3.1: A General Model for Generating and Accumulating Knowledge (adapted from (Owen, 1998))



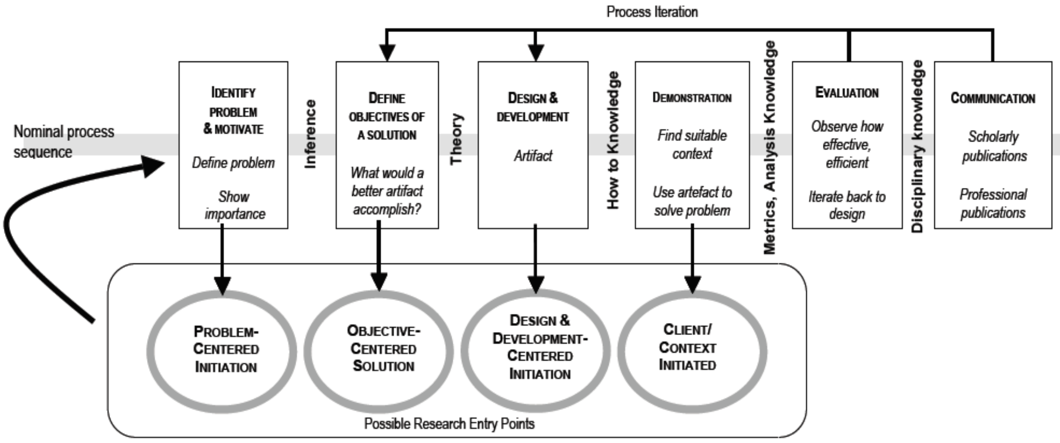
In this model, the channels are the systems of conventions and rules that govern how the discipline operates and they play the role of the empirically matured measures, values and policies that control how to do and judge the works in order to generate new and improved knowledge.

Design Science Research (DSR) involves developing, capturing, accumulating and communicating the knowledge acquired during the course of the design process (Vaishnavi and Kuechler, 2007). DSR methodology is a combination of the iterations, processes and analytical methods for conducting research in the domain of information systems with the aim of understanding, explaining and enhancing the behaviour of one or more aspects of the information system under study (Vaishnavi and Kuechler, 2013). This methodology helps in providing the path to successfully understand the problem or the required improvement as well as to design, develop, implement, and evaluate the technology-oriented solution that is characterised as novel, innovative, and purposeful (Peffers *et al.*, 2007). This indicates that the produced solution would possibly provide organisations and stakeholders with a useful tool or technique that is capable to address unsolved difficulties (Hevner *et al.*, 2004), or provide better solutions with enhanced and improved practices (Vaishnavi and Kuechler, 2007). Moreover, according to Gregor and Jones (2007) and Simon (1996), it is important for the final artefact to respond to its environment changes properly, or it will risk failing its intended purpose. This shows the importance of having a clear boundary for the envisioned outcome (Gregor and Jones, 2007).

3.3.1 Design Science Research (DSR) Processes

DSR methodology involves several iterations of repeated processes until reaching the satisficing artefact. There are several DSR models introduced in the literature (Peffers *et al.*, 2007; Vaishnavi and Kuechler, 2007; Takeda, Veerkamp and Yoshikawa, 1990). The number and definitions of the DSR processes (phases) differ from one model to another and selecting the most suitable model for any research depends on the desired outcome. For example, the models proposed by Vaishnavi and Kuechler, (2007) and by Takeda *et al*, (1990) do not specify a design phase. Thus, since the emphasis in this research is on designing and developing an artefact with the focus on the IS domain among other disciplines, the model that is considered in this research is the one that is introduced by Peffers *et al.* (2007), see Figure (3.2). In this model, the steps of design, develop, implement and evaluate provide a relevant research environment to practice the sensemaking core processes of noticing, interpretation and action.

Figure 3.2: DSRM Process Model (adapted from (Peffers et al., 2007))



The processes and their descriptions are explained below:

- **Problem Identification and Motivation:** The process that initiates the DSR lifecycle and where several motivators from different resources could participate in discovering the research problem as well as the value that could come of solving such an issue.
- **Define the Objectives of the Solution:** The process where concluding the aims of the new artefacts happens. Also, it involves evaluating the possibility and feasibility of each goal.
- **Design and Development:** The process that is concerned with designing and developing the provisional artefact. Additionally, in this phase, new knowledge is generated and accumulated to be utilised in the next iteration of the DSR.
- **Demonstration:** The process that explains how the developed artefact can be used to solve the identified problem in the first phase. This can happen by utilising the new product in experimentation or in a real-life scenario.
- **Evaluation:** Once the artefact is developed and put under testing, an evaluation can be done against the constructed objectives and criteria resulted from phase 2. This stage is important because it specifies whether

a new DSR iteration is needed or not depending on the outcome of the evaluation.

- **Communication:** This step is not actually part of the researching, designing and development loop. Instead, it is about publicising the problem and its importance, and the developed artefact and its novelty, effectiveness and significance for other researchers and practising professionals. The generated knowledge can be categorised as either '*Firm*' to indicate that the facts and the experience can be applied or utilised in similar scenarios, or as 'loose ends' to indicate that might need further research (Vaishnavi and Kuechler, 2013).

3.3.2 Design Science Research (DSR) Outputs

A review of prior studies provided by Vaishnavi and Kuechler (2013), Hevner *et al.*, (2004) and March and Smith (1995) shows the five most important artefacts generated as DSR outputs to be:

- **Constructs:** The conceptual vocabularies and symbols that are used to describe the problems and solutions.
- **Models:** The set of statements or constructs that are used to express or describe the problem at hand. Also, they help to establish the connections and relationships between the problem and the solution proposed for that problem.
- **Methods:** The set of processes that can be followed to perform a specific task or to find a solution for a problem.
- **Instantiations:** The combination of constructs, models and methods that form an operationalised system that aims to solve the problem.
- **Better Theories:** It is argued that design science research can contribute to the process of positively constructing and building theories by adding to the knowledge through the artefact construction as a result of experimental scientific investigation.

In fact, the outputs described above embody important prescriptive knowledge (Gregor and Hevner, 2013) generated from the design science process in IS that is called the Design Theory. The prescriptive knowledge that represents the abstract level of definitions for the methods, research, analysis and the communication of the design practice. When these abstractions are instantiated, then they would represent a physical representation of the artefact that plays the roles of explanatory or testing objects (Gregor and Jones, 2007). This explains the difference between the contributions made at the abstract level of knowledge and those made by the exemplified artefacts (Gregor and Hevner, 2013), see Table (3.2).

Table 3.2: Design Science Research Contribution Types (Adapted from (Gregor and Hevner, 2013))

	Contribution Types	Example Artefacts
More abstract, complete, and mature knowledge ↕ More specific, limited, and less mature knowledge	Level 3. Well-Developed design theory about embedded phenomena	Design theories (mid-range and grand theories)
	Level 2. Nascent design theory – knowledge as operational principles/architecture	Constructs, methods, models, design principles, technological rules
	Level 1. Situated Implementation of the artefact	Instantiations (software products or implemented process)

In addition to the Design theory explained above, Kernel theory and Utility theory are two important theories that can be refined (Peffers *et al.*, 2007) or be developed (Kuechler and Vaishnavi, 2008) from Design Science Research. In the IS context, Kernel theory (justificatory knowledge) is defined as the novel theories, techniques and approaches that originate from natural, social or design sciences to help in IS design problems (Gregor and Hevner, 2013; Fischer, Winter and Wortmann, 2010; Gregor and Jones, 2007). Utility theory, on the other hand, refers to and discusses the generated link between a problem and the technology that addresses it (Fischer, Winter and Wortmann, 2010), as well as the usefulness of the link (Gill and Hevner, 2013). The specific outputs that resulted from the iterations of this research are summarised in sub Section (8.3.1).

3.4 Case Study Description

This research is conducted in cooperation with one of the four major commercial banks within the UK market, BankCo, with over 2,000 branches across the UK and serves more than 30 million customers including 11 million online banking users. BankCo offers a comprehensive range of financial products and services including insurance and banking products that cover current accounts, savings, mortgages, loans and credit cards.

The Client Asset Management Department at BankCo has a focus on portfolio analytics, aiming to understand client needs by analysing the data they have related to client business-to-business transactions and banking products. The anonymised data provided by BankCo relates primarily to Small to Medium Enterprises (SME) and, less so, Mid-Market clients who are customers. The data spans a 30-month period (from January 2014 to the end-of-June 2016) and covers four categories of data:

- ***Business-to-business transactions*** initially comprising around 900 million inter-firm transactions mediated by BankCo, covering SMEs and Mid-Market firms. In this data, both sides of the transaction are known when both firms are clients of the bank. When either one of them is not a client of BankCo, then the information about that client or his ID is unknown.
- ***ValClass,*** which is BankCo's risk-based standard value classification of its clients and it is categorical ranging from D to A+.
- ***Firmographic,*** which contains 185,586 records relating to 140,585 BankCo clients. This dataset includes information about the firms' annual turnover, net worth, size and assets among others.
- ***Products,*** which include information about the banking products acquired by BankCo's clients such as lending, asset finance, and invoice finance products among others.

The computational analysis of the data was undertaken on a Spark server with a 10 nodes Hadoop cluster. The analysis was performed using Spark SQL and GraphX libraries alongside more mainstream statistical programs (e.g., R). Preparing the data for analysis was performed by following the next steps:

1. The first step was to only include the transactions where only the customers of BankCo are involved in the transaction.
2. Second, it was necessary to exclude transactions where customers were moving money between their own accounts (self-paying).
3. Third, utility firms such as electricity and telecommunication companies were removed as they represent ‘false hubs’, since every business has to transact with them and they skew outcomes.

Approximately 335 million inter-firm business transactions remained, representing companies interacting with each other and exchanging goods and services, represented by the financial transaction. In this data, the output of some firms (sub-contractors) are input for some other firms (supply-chain links between companies). Thus, the failure of a firm is likely to increase the probability of failure in connected firms; causing fluctuations in the number of failed firms. Therefore, the transactional data is modelled as a scale-free network where the nodes (vertices) represent the companies and the directed weighted edges represent transactions among those organisations to reflect the dynamics and effects resulting from firms transacting with each other.

The analytical requirements that are expected from the work are to develop methods for understanding client value within the network, understanding their long-term relationship with BankCo and understanding their needs for banking products as well as their purchasing behaviour. These requirements were formulated throughout the DSR iterations and the sensemaking cycles that happened in each iteration. Quarterly meetings were held with the analysts at BankCo to discuss and evaluate the analytical models developed to meet the analytical requirements. Additionally, written reports and ad-hoc virtual meetings in-between the quarterly meetings were used to demonstrate ideas and get feedback from the analysts at BankCo.

3.5 DSR Approach for Enhanced Sensemaking

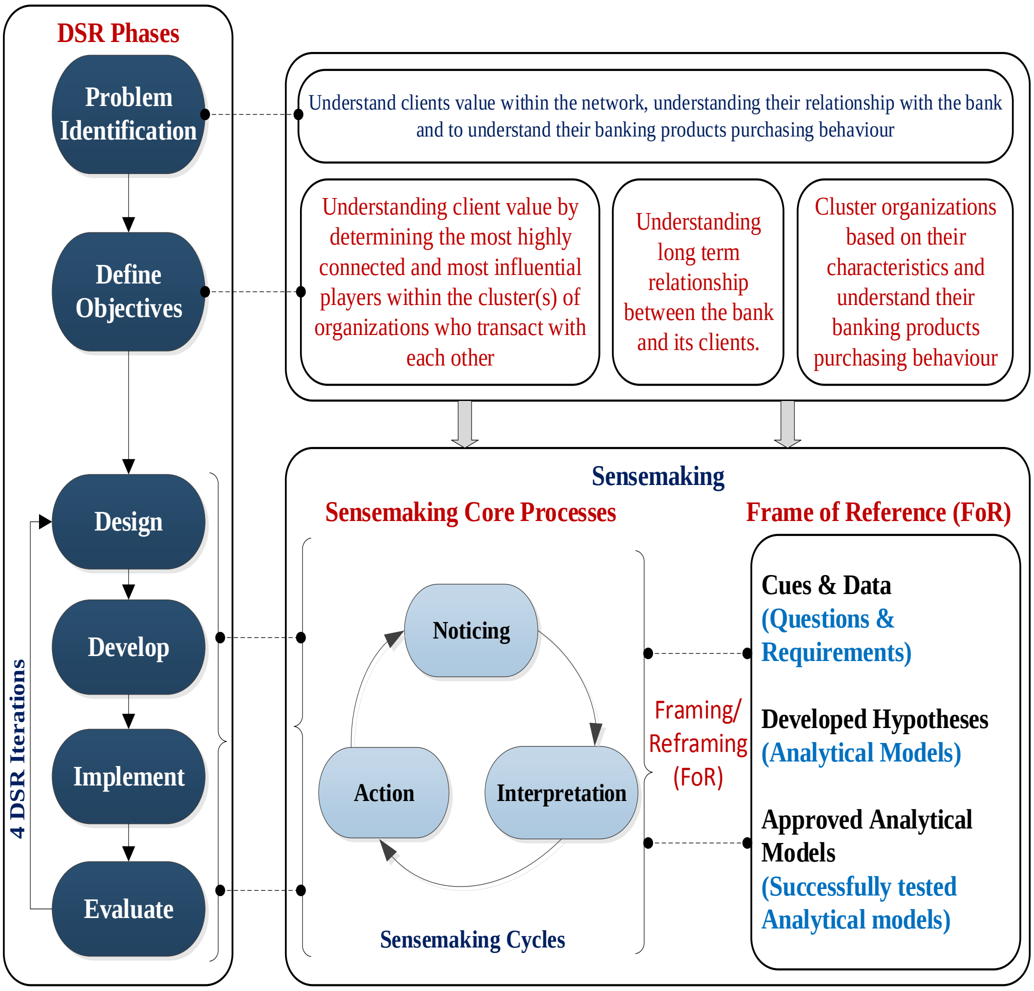
This research demonstrates how human sensemaking, conceptualised in the sensemaking framework proposed in the previous chapter, can be enhanced by using emergent techniques such as Social Network Analysis (SNA) and Machine Learning (ML) to improve business analytics and the generation of valuable insights. The DSR methodology provided by Peffers *et al.* (2007) and illustrated in Figure (3.2) is employed in this research as a general methodological framework, or the research approach umbrella, to meet the aim and to carry out the objectives of this research. This research is composed of four iterations, in which the core processes of sensemaking cycle represented in Figure (2.4) are demonstrated in each iteration. Looking at the DSR process model in Figure (3.2), it can be noticed that the actual iterative part of the model, in essence, is composed of design, develop, implement and evaluate steps, thus, the focus in the four iterations that are reflected in Chapters 4, 5, 6 and 7 is on these four steps. The following four subsections briefly describe the four iterations.

- **DSR Iteration One (Connected Customer Lifetime Value (CCLV) Model)**: The first iteration of this research investigates the problem domain in more detail in order to have a better understanding of the requirements requested by BankCo. Network analysis concepts are employed in this iteration in order to design, develop, implement and evaluate the model that can meet the first part of the requirements, which is about understanding clients' value within the network of customers explained above. Therefore, Customer Lifetime Value (CLV) is considered as the basis for the model intended to meet the first analytical requirement. A network-based customer lifetime value, namely Connected Customer Lifetime Value (CCLV), thus, is designed, developed, implemented and evaluated using the transactional data provided by BankCo. This model can help BankCo value its customers based on their spend in BankCo's universe of clients.

- **DSR Iteration Two (Timeseries CCLV Model):** The second iteration of this research, is about improving the CCLV model, developed in the first iteration, by having an in-depth understanding of the transactional data in terms of network stability over time. Consequently, a timeseries-based CCLV model is constructed. The improved model can assist BankCo to monitor the changes in the CCLV score over the time and act accordingly.
- **DSR Iteration Three (Network Relationship Equity (NRE) Model)**: This iteration is about designing, developing and evaluating a second network-based analytical model. In the business world, relationship equity represents a measure of customer's loyalty to the business. Consequently, a network-based model, called Network Relationship Equity (NRE), is designed, developed, implemented and evaluated in this iteration and can meet additional analytical requirement requested by BankCo. This model can help BankCo monitor the increase and/or decrease in customers' activity over time and act upon it retrospectively in order to prevent losing valuable customers.
- **DSR Iteration Four (Clustering and Products Purchasing Frequency Analyses)**: This iteration focuses on combining network analysis and machine learning techniques in order to address the last part of the analytical requirements that are important for BankCo and based on the data sets under investigation. First, an analysis of banking products purchasing frequency is performed. Second Machine Learning (ML) is used to perform customer clustering. Third, the two analyses are combined in order to uncover banking product purchasing behaviour within each resulting cluster.

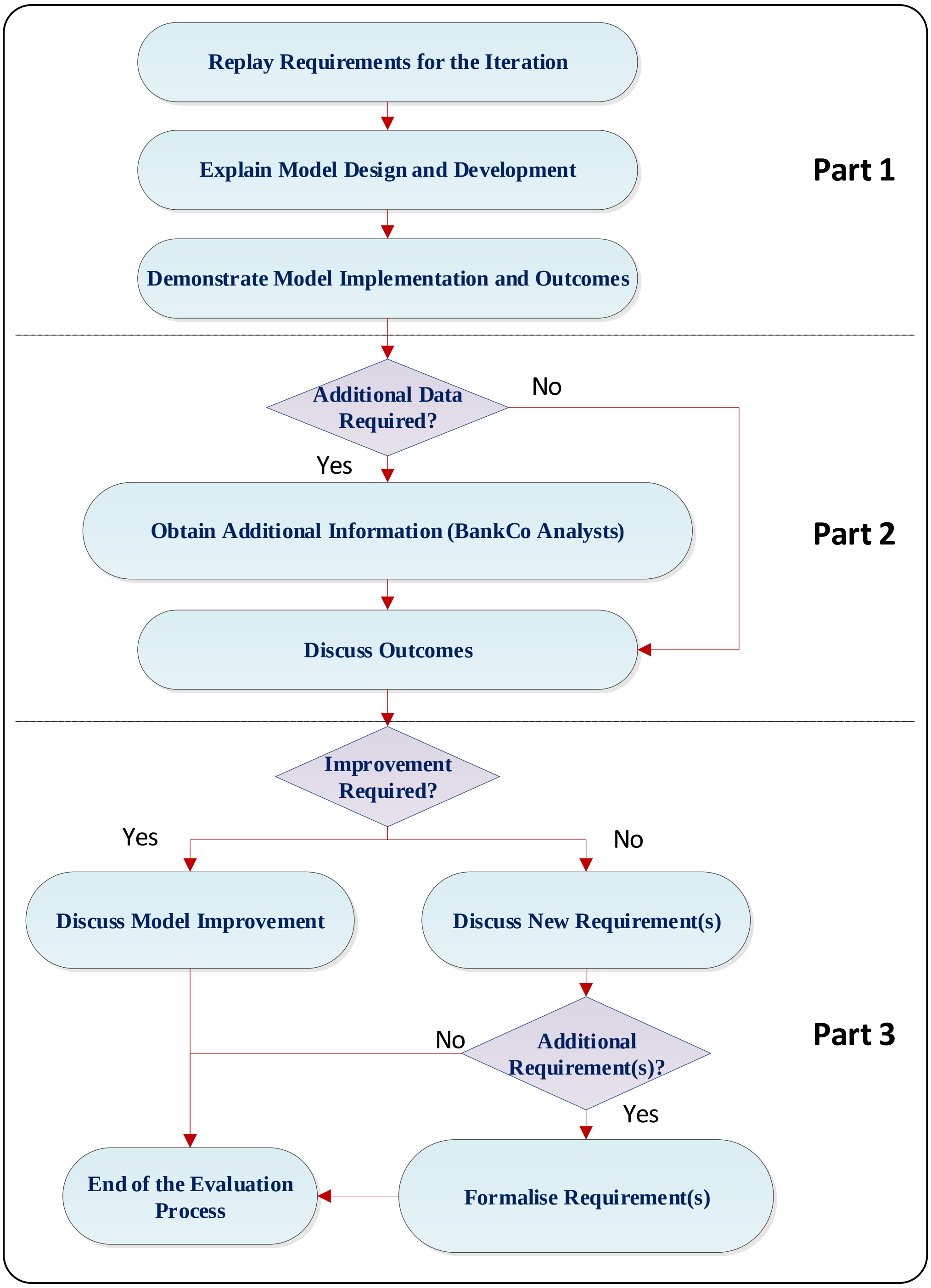
Figure (3.3) illustrates the general research approach employed in each iteration of this research. This figure shows how sensemaking cycles that occur within each DSR iteration participate in framing/reframing the Frame of Reference (FoR); leading to meeting the objective of the iteration.

Figure 3.3: General Research Approach



As noted, quarterly meetings with portfolio analytical team at BankCo were the primary means of discussing and evaluating the models developed. These meetings were supplemented with phone conversations and BankCo doing additional work themselves (e.g., to de-anonymise and drill into additional data) between meetings. Figure (3.4) illustrates this process as a precursor to examples of the sensemaking that follows in subsequent chapters.

Figure 3.4: Evaluation process with stakeholders



3.6 Summary

This chapter has provided an overview of research approaches in IS domain. Also, it has discussed DSR methodology in more details. Additionally, it has presented the case study (a commercial bank within the UK market) as well as data sets used for demonstrating sensemaking core processes during data analysis. It can be concluded that the iterative nature of the DSR methodology suits the aim of this study for two reasons. First, the artefacts that meet the analytical requirements are designed and developed in an incremental manner, in which successive iterations are dependent on the outcomes and evaluation resulting from the ones before. Second, the main DSR iterative phases: design, develop, implement and evaluate form a suitable environment to go through sensemaking cycles and its processes: noticing, interpretation and action. The following 4 chapters will explain the 4 iterations considered for this research.

Chapter 4: Connected Customer Lifetime Value (CCLV) Model (1st Iteration)

4.1 Overview

This chapter presents the first iteration of this study, which considers the banking transactional data described in the methodology chapter as an implementation context for the sensemaking conceptual framework developed in this research in Chapter 2. This iteration focuses on demonstrating how Social Network Analysis (SNA) techniques can help to get valuable insights and enhance the human sensemaking process. To do so, SNA techniques will be used to model Customer Lifetime Value (CLV) while considering the element of connectivity among the customers of BankCo in order to introduce Connected Customer Lifetime Value (CCLV) model. The CCLV model helps to discover the most influential customers for BankCo. Identifying those clients would help BankCo with its marketing and strategic plans.

This chapter is structured as follows: Section 4.2 reviews Social Network Analysis (SNA) techniques used in this iteration. Section 4.3 describes the design phase of this iteration. Section 4.4 explains the development process of this DSR iteration. Section 4.5 presents the implementation phase and Section 4.6 provides the technical evaluation of the developed model as well as the evaluation of the sensemaking framework. Finally, Section 4.7 provides an overall summary of the chapter.

4.2 Social Network Analysis (SNA) Techniques

A social network consists of a set of ‘entities’ and the ‘relation(s)’ between those entities (Butts, 2008). Social Network Analysis (SNA) is the field of study that is concerned with mapping and measuring relationships and the flow of values between the nodes that form the network. Nodes can be people, groups or organisations (etc.), while the edges (connections) represent relationships between the nodes (Butts, 2008; Kiss and Bichler, 2008). Relationships can also take many forms including transactions, information flows, friendships, etc. Social Network Analysis (SNA) techniques can help in determining the influential nodes within a network. By presenting a business network using a graph, SNA techniques have the capability to determine highly-connected centralized hubs, hence, influential nodes (Chen, Fu and Shang, 2009).

Of importance here, social network analysis encompasses the study of centrality and topological ranking measures, which are used to identify the most valuable vertices within a network (Kiss and Bichler, 2008). Centrality measures that are relevant to identifying the most central and influential customers include:

- ***Degree Centrality***: In its simplest definition, this represents the number of edges attached to a node. This measure can be divided into in-degree, representing the number of edges arriving at the node, and out-degree, representing the number of edges initiated from the node.

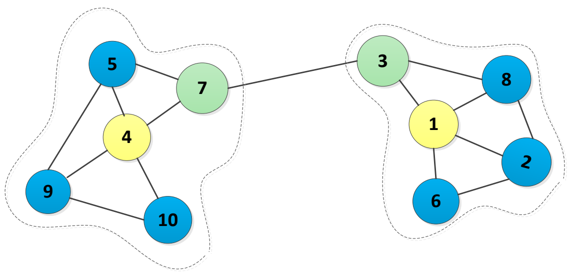
- ***Closeness Centrality***: A node is considered as central if it has the shortest path to all other nodes in a network. Identifying the central nodes can improve communications with a network.

- ***Betweenness Centrality***: This measure of centrality quantifies the importance of a node (A) as a bridge between two other nodes (B) and (C) when the shortest path between (B) and (C) goes through (A). This means that node (A) can control the interaction between nodes (B) and (C).

- ***Eigenvector Centrality***: Here, the importance of a node (A) is measured by the importance or centrality of the nodes connected to (A). In other words, this measure of centrality means that links with influential people make you more powerful than links with powerless people.

Additionally, PageRank, an algorithm originally developed by Google to rank websites according to their importance considering their link structure, is used to rank the nodes in any network that represents a business or social domain (Gleich, 2015). Moreover, community structure is an important feature of complex networks, which occurs when nodes cluster into tightly-knit groups with a high density of within-group connections and a lower density of between-group connections (Chen, Fu and Shang, 2009). Figure (4.1) illustrates a sample network with two community structures. The importance of community detection stems from the fact that it helps in understanding complex systems by de-structuring complex networks into smaller ones with centralised hubs (e.g., nodes (1) and (4) in the sample network in Figure 1). Moreover, using community detection algorithms enables the detection of nodes that act as a bridge between one community and another (e.g., nodes (3) and (7) in Figure (4.1)).

Figure 4.1: Community Structure



Many community detection algorithms exist (Blondel *et al.*, 2008; Ronhovde and Nussinov, 2009; Wang, Chen and Lu, 2007; Chen, Fu and Shang, 2009; Chen, Fu and Shang, 2009). Lancichinetti and Fortunato, (2009) have conducted a study to compare the various community detection algorithms available and concluded that the structural algorithm introduced by Rosvall and Bergstrom (2008), Fast modularity optimization *algorithm* introduced by Blondel *et al.* (2008) and *Potts model approach* introduced by Ronhovde and Nussinov (2009) provide the best performance in detecting communities within complex networks. Additionally, other community detection algorithms have been introduced by Wang, Chen and Lu (2007) and Chen, Fu and Shang (2009); claiming these algorithms to be the fastest and most efficient. Choosing the best algorithm for the context in hand, a large network of BankCo's customers, however, depends on the programmability of the algorithm and its performance on large data sets (big data).

4.3 Iteration 1: Designing the CCLV Model

The design phase in this iteration of the DSR methodology focuses on exploring the data and investigating the analytical requirements requested by BankCo and noticing cues that help to define the analytical objectives that address these requirements. The objective of this iteration is to determine the most highly influential clients within the cluster(s) of companies who transact with each other. Thus, it can be noticed that Customer Lifetime Value (CLV) is such a marketing measure that can be used to determine customers' importance for the organisation. However, traditional CLV models define it as assessing the present value of the cash flow generated by a customer minus the cost of acquisition or retention of that customer – costs such as discounts or promotions (Zhang, Liang and Wang, 2016; Kumar, 2010). Therefore, it can be noticed as well, that traditional CLV models do not count for clients' connectedness with each other within the network that represents the clients and the transactions among them.

Consequently, since the SNA techniques are used to measure the relationships and the flow of values (transactions) between the nodes that form a network, there is a need to investigate how these techniques can help to model and calculate Customer Lifetime Value (CLV) while considering the element of connectivity among BankCo's customers, which represents the outcome of the noticing process of sensemaking.

4.3.1 Customer Lifetime Value (CLV)

This section reviews traditional CLV models as a platform to discuss recent advances in the state-of-the-art that models 'connectedness', including the impact of social media and the case of referrals (Weinberg and Berger, 2011). Customer Lifetime Value (CLV) is one such marketing measure that has become increasingly important (Mzoughia and Limam, 2015; Gupta *et al.*, 2006). CLV is a disaggregate metric that can be used to identify profitable customers and allocate resources accordingly, providing more customer's specific insights (Gupta *et al.*, 2006; Kumar, Lemon and Parasuraman, 2006). Prior studies have argued CLV to be a key concept for any business that can positively affect its current and future performance (Feiz, RamezaniGhotbabadi and BteKhalifah, 2016; Feiz, RamezaniGhotbabadi and BteKhalifah, 2016; Baum and Singh, 2008; Kumar, Lemon and Parasuraman, 2006). Unsurprisingly, therefore, CLV is a fundamental concept in many customer relationship management approaches, such as one-to-one, loyalty, and database marketing (Blattberg, Malthouse and Neslin, 2009; Borle, Singh and Jain, 2008). One of the main strengths of CLV analysis is that it can be used to predict the future profitability of clients, leading to more accurate marketing strategies and decisions relating to customers (Chang, Chang and Li, 2012). In summary, the usefulness of CLV models for selecting and targeting specific customers is justified by "customers who are selected on the basis of their lifetime value provide higher profits in future periods than do customers selected on the basis of several other customer-based metrics" (Venkatesan and Kumar, 2004).

Crudely speaking, traditional CLV models consider the profit generated by the customer while subtracting the cost of customer acquisition and retention. Though valuable, this approach does not reflect the dynamic networks that reflect the business relationships of our time. Thus, the concept of the 'value network' is important to consider, since it potentially reflects both tangible and intangible dynamic value exchange in a network of customers transacting among each other (Hosseini and Albadvi, 2010). Fledgling work has been done in this area (Klier *et al.*, 2014), which combines the notion of CLV with Social Network Analysis (SNA) techniques. Existing models, however, are idiosyncratic in nature, driven in good part by the data available to them (Klier *et al.*, 2014). The intended CCLV model, however, takes the spirit of the work of Klier *et al.* (2014), developing a more general model applicable to transaction-oriented environments and focusing on firm-to-firm relationships, which is novel in the literature to date.

4.3.2 Customer Lifetime Value without Connectedness

Customer Lifetime Value (CLV) as a financial measure that assesses the value of customers based on their cash flow minus promotion costs is generally considered as a trusted metric to measure customer performance in the Customer Relationship Management (CRM) field (Venkatesan and Kumar, 2004). The noted benefits of CLV are: (a) The ability to identify the importance of each customer to the organisation; (b) the prediction of whether or not it is profitable to acquire new customers or retain existing ones (Feiz, RamezaniGhotbabadi and BteKhalifah, 2016; Blattberg, Malthouse and Neslin, 2009); (c) more effective allocation of resources to customers; (d) more and improved information on how to develop long-term customer relationships (Tavakolijou, 2012); and (e) the ability to predict the probability of customers defecting to competitors in the future (Ferrentino, Cuomo and Boniello, 2016).

In measuring CLV, the standard approach is to estimate the present value of the net benefit to the firm from the customer – generally taken as the revenues from the customer minus the firm's costs in maintaining and developing the relationship with the customer over time (Borle, Singh and Jain, 2008). Many studies have proposed variations on CLV, but the underlying structure is similar (Hosseini and Albadvi, 2010). As a representative example, Berger and Nasr (1998) proposed the following variations to calculate CLV:

$$\boldsymbol{CLV} = \sum_{i=0}^{n} \pi(t) \frac{1}{(1+d)^i} \qquad (1)$$

Where:

$\boldsymbol{\pi(t)}$: is the function of customer profits according to time t

i: is the period of cash flow from customer transaction

n: is the total number of periods of customer transactions

d: is the discount rate

$$\boldsymbol{CLV} = \left\{GC * \sum_{i=0}^{n}\left[\frac{r^{i}}{(1+d)^{i}}\right]\right\} - \left\{M * \sum_{i=1}^{n}\left[\frac{r^{i-1}}{(1+d)^{i-0.5}}\right]\right\} \quad (2)$$

Where:

GC: is the (expected) yearly gross contribution margin per customer. It is, therefore, equal to revenues minus cost of sales

M: is the (relevant) promotion costs per customer per year

n: is the length, in years, of the period over which cash flows are to be projected

r: is the yearly retention rate, i.e., the proportion of customers expected to continue buying the company's goods or services in the subsequent year

d: is the yearly discount rate (appropriate for marketing investments)

4.3.3 Customer Lifetime Value with Connectedness

Though CLV is of immense value and widespread in use, the concept does not consider the effects (and potential value) of networking among firms (Klier *et al.*, 2014). This gap has motivated researchers to investigate the significance of considering the customer's surrounding network. Neighbours in a network of customers can refer products to each other. Also, social influence can help companies acquire new customers at relatively low acquisition costs (Klier *et al.*, 2014), and more profitable customers in terms of long-term relationship (Schmitt, Skiera and Van den Bulte, 2011; Villanueva, Yoo and Hanssens, 2008). Additionally, purchase decision and loyalty can be highly affected by social influence (Nitzan and Libai, 2011; Weinberg and Berger, 2011). Consequently, discovering the influencers (customers with high connectivity) in a company's network is considered as a crucial task prior to any marketing-related decision making.

Recent models that consider the connectivity among customers have been proposed by Nitzan and Libai (2011) and Weinberg and Berger (2011), identifying two kinds of social influence – Customer Referral Value (CRV) and Customer Social Media Value (CSMV). CRV is an important aspect of social influence that affects CLV, as it is the word-of-mouth referral that can lead new customers to buy a product/take-up a service. CSMV is another factor that can cause non-direct cash flow in the network through social media engagement depicted in a form of Twitter tweets, Facebook posts or communities' discussion; thus, this can affect and change the value of CLV (Nitzan and Libai, 2011; Weinberg and Berger, 2011). Considering these two aspects, Weinberg and Berger (2011) have proposed that Connected Customer Lifetime Value can be calculated as:

$$\boldsymbol{CCLV} = CLV + CRV + CSMV \tag{3}$$

Where:

$$CSMV_i = CLV_i * \left([1 + SM_{i1}] * [1 + SM_{i2}] * \ldots * \left[1 + SM_{ij}\right] * \ldots * \left[1 + SM_{iJ}\right] - 1\right)$$

Where:

$\boldsymbol{SM_{ij}}$ is the impact of social media j (Twitter, Facebook, Forums, Communities and Blogs) on customer i

Another approach for calculating CLV while considering the value of networking is presented by Hosseini and Albadvi (2010). Here, the value network is divided into tangible (goods and services, etc.) and intangible exchanges (knowledge and information that supports the take-up of goods/services). In this research, Network Customer Lifetime Value (NCLV) is defined as:

$$\boldsymbol{NCLV_i} = CLV_i + \sum_{j=1}^{n} \alpha_{ij}\, NRV_{ij} \qquad i \neq j, \qquad j = 1, \ldots, n, \qquad 0 \leq \alpha_{ij} \leq 1 \tag{4}$$

Where:

$\boldsymbol{NCLV_i}$ is the network customer lifetime value of customer i

$\boldsymbol{CLV_i}$ is the customer lifetime value of customer i

$\boldsymbol{NRV_{ij}}$ is the network relationship between value between customers I and j

$\boldsymbol{\alpha_{ij}}$ is the importance of NRV_{ij} from focal company's point of view

$\boldsymbol{n}$ is the total number of customers

Last, studies by Klier *et al.* (2014) have introduced a different method to calculate CLV while accounting for mutual network effects among the members of a network. The approach can be summarised thus as follows:

$$\boldsymbol{CLNV} = Present\ value\ of\ individual\ cash\ flow + present\ value\ of\ \Delta\, network\ contribution$$

Where:

$\Delta\, \boldsymbol{network\ contribution}$ can be positive or negative depending on the customer contribution to the network. A customer can have positive $\Delta\, network\ contribution$ when the cash flow induced to the network by other members is depending on the cash flow generated by this customer. Conversely, when a cash flow of certain customer is dependent on other members, then $\Delta\, network\ contribution$ for this customer is negative.

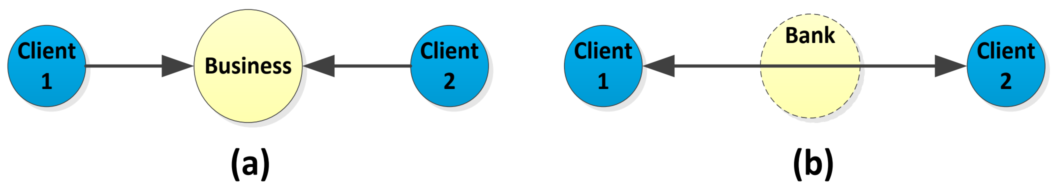
The calculation of $\Delta\, network\ contribution$ is dependent on certain variables being computed first. For example, it is necessary to compute the probability that the influence (cash flow) exerted by customer (i) on customer (j) will actually lead customer j to make a purchase. Additionally, a discount rate (d) and a weighing factor α are important for the calculation of the final value of CLNV.

Calculating CLV while considering network perspectives can lead to additional revenue for the network members, increased customer referrals and improved network relationships among network members. Finding more efficient ways to manage and invest in a business network, therefore, is considered as a crucial task for organisations. Little research, however, has studied the networking effect while measuring CLV. The development phase of this iteration utilises Social Network Analysis (SNA) techniques to model and calculate the Connected Customer Lifetime Value (CCLV), which is defined as the present value of the first-order neighbours' influence on the cash flow generated by the focal customer within a network.

4.4 Iteration 1: Developing the CCLV Model

The development phase of the first DSR iteration builds on the state-of-the-art reviewed about the different CLV models and utilises SNA concepts to develop a network-based CLV model, namely Connected Customer Lifetime Value (CCLV) model. Traditional customer lifetime value models can be seen from the network perspective as the study of the relations between a firm and its clients – as illustrated in Figure (4.2.a). In such a case, the source node will be one of the clients and the destination node will always be the firm valuing its clients (e.g., ego networks). In the context of this research, however, the focus is on inter-firm relationships, where the form of the relationship (a transaction) is mediated by a third party – BankCo, which is essentially facilitating the financial manifestation of the inter-firm relationship as illustrated at Figure (4.2.b). For this study, the mediation of the third-party is immaterial, aside from the fact that all the firms in the dataset are clients of BankCo.

Figure 4.2: Different types of transactions among firms



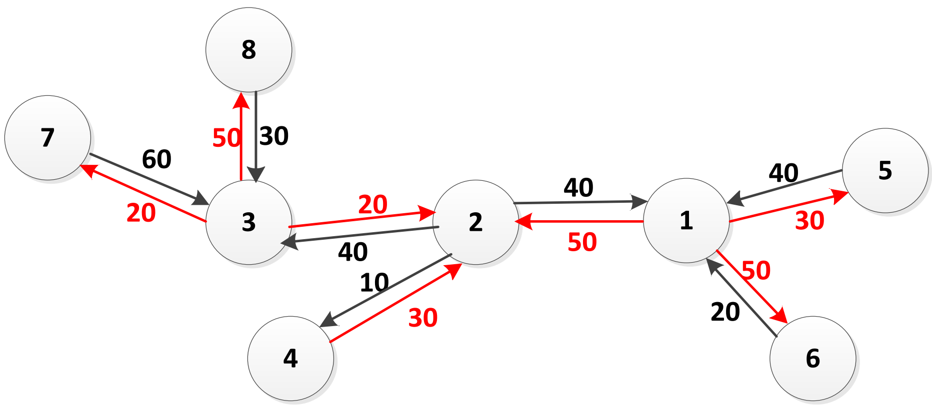
As noted before, the variables used in prior studies vary in accordance with what is available to researchers and what is suitable for the focus of study. In addition, given that the focus here is to look at inter-firm relations that are mediated by BankCo, variables that are traditionally used in CLV calculations are neither transparent nor available – e.g., costs for promotion, customer acquisition and retention and referral and/or social media factors. Consequently, the development work has taken the spirit of emerging models (building on the work of Klier *et al.* (2014) in particular) and developed a CLV model that is particularized to the domain under investigation.

The following simple fictitious but representative example explains the steps for model development. Table (4.1) lists the sample transactional data (the source, the destination and the amount of money transacted (cash flow) among 8 customers, and Figure (4.3) shows the network representation of the data. BankCo, which can be seen as 'a view from nowhere, values its customers by quantifying the cash flow generated by the customer while considering the influence of the customers on each other.

Table 4.1: Sample transactional data

From customer	To customer	Amount
1	2	50
1	5	30
1	6	50
2	4	10
2	3	40
2	1	40
3	7	20
3	2	20
3	8	50
4	2	30
5	1	40
6	1	20
7	3	60
8	3	30

Figure 4.3: Network representation of the transactional data



In this network, the average weighted network influence of customer (i) on customer (j) compared to other influencers on customer (j) is calculated by dividing the individual influence of the customer (i) by the total average influence on customer (j). Table (4.2) presents an example of how to calculate the average weighted network influence between customers (1) and (2).

Table 4.2: Average weighted network influence

From customer (1) on customer (2)	From customer (2) on customer (1)
50+20+30 = 100/3 = 33.33 50/33.33 = **1.5**	40+40+20 = 100/3 = 33.33 40/33.33 = **1.2**

Table (4.3) shows the matrix representation of influence among all the customers in the sample data. In addition, the table also displays the sum of the columns, which is used to standardize the values in the matrix by dividing the value in the matrix by the sum of the column it belongs to as displayed in Table (4.4).

Table 4.3: Matrix representation of the influence

From \ To	1	2	3	4	5	6	7	8
1		1.5			1	1		
2	1.2		0.92	1				
3		0.66					1	1
4		0.66						
5	1.2							
6	0.6							
7			1.38					
8			0.69					
Sum of the columns	**3**	**2.82**	**2.99**	**1**	**1**	**1**	**1**	**1**

Table (4.4) shows how the most influential node (customer) is calculated by summing the rows and dividing the individual results by the total of the sum of the rows. The values in the rightmost column of Table 4 represent the Connected Customer Lifetime Value (CCLV).

Table 4.4: Standardized matrix of influence and calculating most influential node

From \ To	1	2	3	4	5	6	7	8	Sum of Rows	Standardized sum of row (Most influential Node) (CCLV)
1		0.532			1	1			2.53	2.53/ 8 = **0.32**
2	0.4		0.308	1					1.71	1.71/ 8 = **0.21**
3		0.234					1	1	2.23	2.23/ 8 = **0.28**
4		0.234							0.23	0.23/ 8 = **0.03**
5	0.4								0.40	0.40/ 8 = **0.05**
6	0.2								0.20	0.20/ 8 = **0.03**
7			0.462						0.46	0.46/ 8 = **0.06**
8			0.231						0.23	0.23/ 8 = **0.03**
					Total of the sum of the rows				**8.00**	

The values in Table (4.4) show that the influence is equal to 1 when there are no other influencers on the customer. Also, the rightmost column (CCLV) shows that customer (1) is the most influential in the network, then customer (3) and then customer (2).

4.5 Iteration 1: Implementing the CCLV Model

The CCLV model is implemented in the context of the 30-months transactional data provided by BankCo, which resemble the data presented in the example above. The following two subsections demonstrate how the CCLV model ranks the most and least influential customers based on their influence upon their first order neighbours in the network.

4.5.1 Most Influential Clients (Top 20 influencers)

In this subsection, Table (4.5) presents the Top 20 influential customers in the transactional data network based on the CCLV values. In addition, it provides a comparison between the CCLV score and BankCo's own valuation of its customer. For completeness, centrality measures are also calculated, including those originating from the customer (out-degree), coming to the customer (in-degree) and the sum of these numbers (degree centrality).

From a descriptive point of view, the points of interest are as follows. First, there is a visible correlation between the CCLV ranking and ValClass value – most of the top 20 nodes are valued by BankCo as A+ customers. Second, discrepancies exist, however, the firm ranked 9th provides a striking one. This firm is very highly connected with out-degree of 1,620, an in-degree of 26,732 and a centrality score of 28,352 – the firm also has a relatively high cash flow of £214,329,592. Third, the firm ranked 17th has a significant cash flow of more than 3 billion GBP but is valued as B. Degree centrality measures are modest. Last, it can be concluded here that ValClass, as a unidimensional measure that only considers the risk factor as the basis for evaluating the firms, can be complemented by using the CCLV score to improve how the customers are evaluated.

Table 4.5: CCLV for most influential nodes compared to ValClass and Degree centrality

	Node (Anonymised Customer ID)	CCLV (Most Influential Node)	Business Category & ValClass		Degree Centrality			No of Trans (In and Out)	Sum of Amount Cash Flow (In and Out)
			Category	ValClass	Out-degree	In-degree	Degree		
01	9737233578753977	0.0024	MM	A+	1,520	1,502	3,022	44,299	£363,778,533
02	9737233577051063	0.002	Missing*		1,477	14	1,491	19,803	£2,279,683,601
03	1964933484975297	0.0019	MM	A+	947	329	1,276	21,523	£33,154,237
04	9737233564601535	0.0019	MM	A+	803	657	1,460	27,256	£953,122,361
05	7760150167739034	0.0019	SME	A+	554	371	925	25,426	£7,166,569
06	1184727448856452	0.0018	SME	A+	1,870	1,913	3,783	119,32	£36,543,368
07	3316428431849140	0.0016	MM	C	950	146	1,096	3,150	£4,727,054
08	9737233574695254	0.0015	MM	A+	1,174	461	1,635	16,972	£126,336,918

09	3044772329461175	0.0013	MM	D	1,620	26,732	28,352	361,96	£214,329,592
10	9737233532427343	0.0012	MM	A+	1,087	10	1,097	9,578	£107,418,271
11	9737233592737223	0.0010	MM	A+	1128	300	1,428	26,953	£118,388,635
12	6008353554890705	0.0010	MM	C	202	18	220	11,973	£5,380,468
13	9737233555064670	0.0010	MM	A+	633	434	1,067	15,134	£110,365,060
14	9737233530665152	0.0009	MM	A+	1,041	723	1,764	17,037	£157,170,001
15	9556487459311823	0.0008	MM	A+	427	279	706	4,596	£13,918,953
16	5183006448623432	0.0008	MM	A+	271	61	332	9,249	£36,083,006
17	9737233562316476	0.0008	MM	B	635	118	753	11,768	£3,090,350,308
18	2715714432040724	0.0008	MM	A+	312	225	537	18,883	£116,269,852
19	7299138661748686	0.0008	MM	D	286	3	289	2,482	£70,690,695
20	3323635188960776	0.0007	Missing*		583	57	640	9,399	£204,722,136

* Missing observations from the ValClass data provided by BankCo

4.5.2 Least Influential Clients (Bottom 20 influencers)

In this subsection, Table (4.6) presents the bottom 20 influential customers in BankCo's transactional data network based on the CCLV values. For consistency, the comparison between the CCLV and ValClass is shown, as are the centrality measures. From a descriptive point of view, the points of interest here are as follows. First, there is much less of a visible correlation between the CCLV score (which is zero for all firms) and BankCo's valuation. Second, all firms here are Small-to-Medium Enterprises (SMEs) and, unsurprisingly, they are much less connected. Last, discrepancies again exist, however, and the firm ranked 11th provides an illustrative example. This firm is valued as A+ even though its out-degree is 3, in-degree is 6 and centrality is only 9 (very low connectivity). In addition, the total number of transactions for this firm is only 168 and the cash flow is very low across the 3 years span of the data.

Table 4.6: CCLV for least influential nodes compared to ValClass and Degree centrality

	Node (Anonymised Customer ID)	CCLV (Least Influential Node)	Business Category & ValClass		Degree Centrality			No of Trans (In and Out)	Sum of Amount Cash Flow (In and Out)
			Category	ValClass	Out-degree	In-degree	Degree		
01	7760150148859210	0.0	SME	C	5	1	6	38	£29,023
02	0888254039226130	0.0	Missing*		Did not calculate**			0	£0
03	0690272583058668	0.0	Missing*		Did not calculate**			0	£0
04	7913952738710923	0.0	SME	B	25	12	37	190	£1,086,750
05	8309135282123031	0.0	SME	A+	16	12	28	96	£46,750
06	5584821899727824	0.0	SME	A+	6	60	66	189	£274,065
07	3683757165980692	0.0	SME	C	9	12	21	201	£321,724
08	8241990545812517	0.0	SME	B	4	1	5	54	£642135
09	6746568044338914	0.0	SME	B	9	11	20	98	£23,280
10	2473614834465458	0.0	Missing*		Did not calculate**			4	£36,320
11	1739074866884483	0.0	SME	A+	3	6	9	168	£438,489
12	1960646785381956	0.0	SME	C	Did not calculate**			255	£423,721
13	9386193073253650	0.0	SME	A+	22	16	38	336	£395,101
14	8402089455937756	0.0	SME	C	5	8	13	238	£176,645
15	1741462932647747	0.0	SME	C	1	1	2	21	£283,400
16	7896916623965298	0.0	SME	C	14	6	20	30	£5,856
17	4035855558378531	0.0	SME	C	12	2	14	23	£27,792
18	6196310905978599	0.0	SME	A+	17	35	52	350	£951,235
19	3190225371870381	0.0	SME	A+	14	37	51	497	£1,775,666
20	2954239842028123	0.0	SME	C	7	32	39	140	£140,722

* Missing observations from the ValClass data provided by BankCo
** Degree centrality did not calculate on GraphX in Spark for those nodes

4.6 Iteration 1: Evaluating the CCLV Model

This phase of the DSR methodology focuses on evaluating the CCLV model developed in this iteration. The CCLV model is tested and evaluated in the context of the 30-months transactional data provided by BankCo, which resemble the data presented in the example above.

4.6.1 Stakeholder Evaluation

Following the process presented in Figure (3.4), the CCLV model and the implementation results of this iteration were presented to and discussed with the analysts from BankCo. During Part 1 of the process, the researcher explained the noticed cues and their interpretation, positioning how using Social Network Analysis (SNA) techniques can help in calculating Customer Lifetime Value (CLV) while considering the mutual influence among the clients in the network (as the foundation for valuing the customers in the network). The developed model in the preceding sections was explained to the analysts and the implementation examples (discussed in Section 4.5) were demonstrated – highlighting the interesting discrepancies between the ValClass and the CCLV model in doing so.

In Part 2 of the evaluation process, given differences between the ValClass and the CCLV score, the analysts at BankCo de-anonymised the data and looked up additional information related to the companies where there were big differences between the two valuation techniques (presented in Tables 4.5 and 4.6). the outcome of this analysis helped BankCo to understand the value that data may have in telling different 'stories' in a given context. Unlike ValClass, which is a risk-based valuation, the CCLV model allows for a ranking of importance based on the influence that a firm has amongst its first-order relations. Most importantly, they noticed that modelling in network form makes it clear that decisions made in relation to one firm have tangible knock-on effects for others, and that single firm calculations of risk ignore that aspect. The additional work done at BankCo clarified how the CCLV model might be useful for providing the opportunity for new information related to products and marketing strategies that, for example:

1. Make sense and strengthen synergies between firms;
2. Enable the better allocation of resources (e.g., Relationship Managers), depending on a customer's individual network value;
3. Exploit a different form of customer segmentation based on relations.

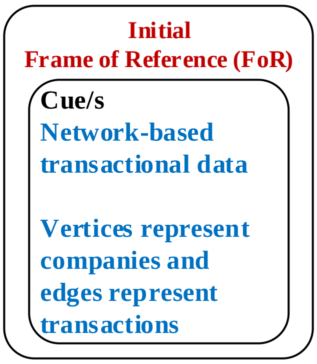
The understanding developed, however, sparked questions from BankCo that concentrated discussion of the following (potential) new requirements in Part 3 of the process:

- The CCLV model must consider firmographic data such as size as a controlling variable;
- The CCLV model should account for the sector to which the firm belongs;
- Given the fact that firmographic data changes over the time, the model needs to be reconstructed in order to provide a timeseries-based CCLV model that makes changes over time apparent.

4.6.2 Evaluation of the Sensemaking Framework

This subsection offers an evaluation of how sensemaking processes of noticing, interpretation and action (as illustrated in the sensemaking framework introduced in Chapter 2) took place in evolving the Frame of Reference (FoR) that led to the development of the CCLV Model (via framing and re-framing). As explained in Chapter 3, the transactional data is modelled as a scale-free network where the nodes (vertices) represent the organisations (clients) and the directed weighted edges represent transactions among those organisations. The network-based approach taken to the transactional data was given at the start of the work, so must be considered as the initial FoR for the first iteration, with network concepts such as nodes and edges forming the early cues - see Figure (4.4).

Figure 4.4: Initial FoR for the 1st iteration



In the noticing process, the initial FoR was informed by marketing analytics-based work on Customer Lifetime Value (CLV) – adding additional cues related to calculating CLV such as cash-flow generated by the customer and the cost of acquiring or retaining the customer. Interpreting and understanding the limitations in the information imposed by the data, however, helped the researcher to realise that not all aspects of the traditional CLV models could be calculated in the context of the banking transactional data (or indeed, would have been of value). Such information includes information about promotions offered to customers or the costs of customer acquisition or retention, which are not available in the data sets provided by BankCo. On the other hand, having the transactional data structured as a network of customers transacting with each other, as illustrated in the sample data in Figure (4.3), is noticed as a cue that is interpreted by the researcher as an opportunity for considering the influence (connectivity) of the customers on each other while calculating the CLV of the customers. Hence, as an action process, the Connected Customer Lifetime Value (CCLV) model is developed to address the analytical requirement.

To explain the inner cycle in the sensemaking framework, the initial Frame of Reference (FoR) that contained the network-based transactional data is questioned to investigate its relevance for the analytical objective of this iteration, and consequently, elaborated by hypothesising that a network-based CLV model can be the answer for the analytical requirement. Consequently, the initial FoR is reframed with the noticed cues, theorised hypothesis and the developed CCLV model, see Figure (4.5), and questioned to examine its performance and fit for the analytical requirement based on the feedback from the research partner (BankCo) on the implementation results and examples discussed earlier above.

Figure 4.5: Reframed FoR in the 1st iteration

Reframed Frame of Reference (FoR)

Cues
- CLV and its aspects could be the answer for the analytical requirement
- Network-based transactional data could enhance traditional CLV

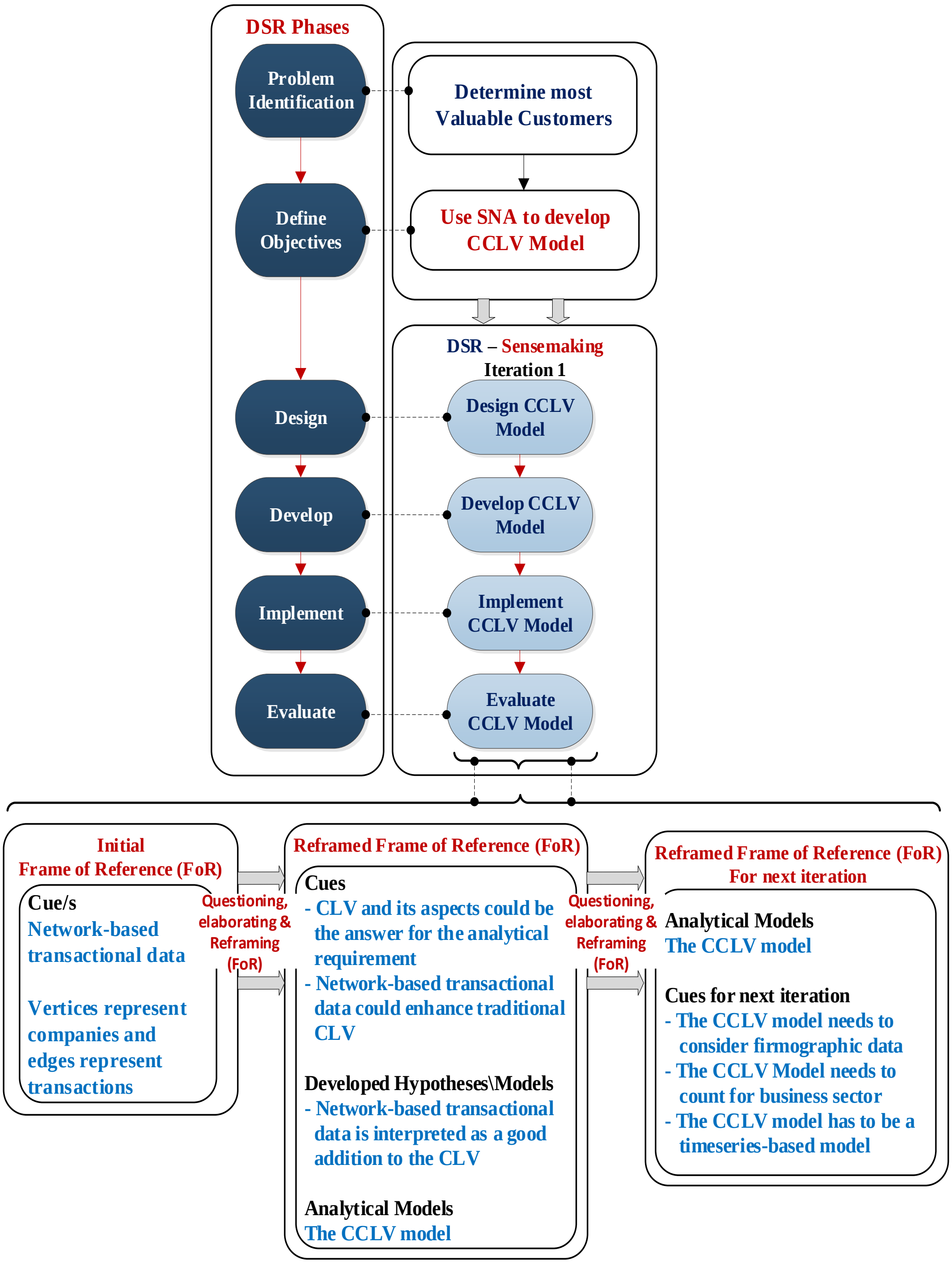
Developed Hypotheses\Models
- Network-based transactional data is interpreted as a good addition to the CLV

Analytical Models
The CCLV model

The CCLV model is considered as a good addition to the current risk-based customer valuation mechanism, ValClass, which BankCo currently has. The concerns highlighted in the stakeholders' evaluation, however, have led to a subsequent reframing of the FoR to reorient these points as cues for the next iteration explained in the following chapter.

Figure (4.6) illustrates an overall view of the first iteration presented in this chapter. This figure shows how the Frame of Reference (FoR) for next iteration is formed based on the outcomes of the first sensemaking cycle.

Figure 4.6: Overall view of the first iteration



4.7 Summary

This chapter has presented the first iteration of this work and demonstrated how sensemaking processes happen during this iteration, in which Customer Lifetime Value (CLV) models have been considered as the basis to design, develop, implement and evaluate the Connected Customer Lifetime Value (CCLV) model. Traditional CLV models have become increasingly important as a marketing metric – not least because it can positively affect current and future performance. A significant strength of CLV analysis is that it can be used to predict the future profitability of clients, potentially enabling more appropriate marketing strategies and decisions relating to customers. The downside of the approach is that it ego-based and does not reflect the (dynamic) networks that reflect the manner in which firms now operate. Recognizing this, fledgling research has emerged that seeks to add a 'connected' element to the more general CLV approach.

This iteration has sought to add to the 'connected' research base, developing a Connected Customer Lifetime Value (CCLV) model based on an analysis of transactions in the financial services domain. The model was applied to a significant number of transactions among firms using a modern open source computing infrastructure (Spark plus Hadoop). In positive terms, application of the model allows BankCo to see the network implications of the decisions they make with respect to customers. Compared to ValClass, which is a one-dimensional measure that only considers client's risk to calculate the ValClass score, the CCLV model allows addressing potential weaknesses in comparison to the ValClass by considering the connectivity of the clients within the network they belong to. Considering the feedback from the research partner, however, key areas for improvement arise. First, the model needs to count for the size of the firms and the sectors they belong to. Second, a timeseries-based CCLV model is required in order to consider the changes in firmographic data. The next iteration of this research will focus on addressing these points.

Chapter 5: Size-Controlled Timeseries-based CCLV Model (2nd Iteration)

5.1 Overview

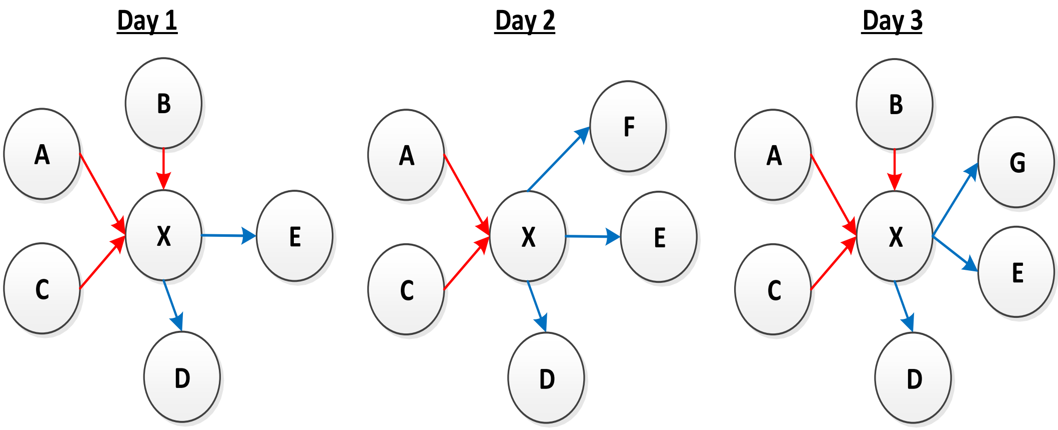
This chapter presents the second iteration of this study, which considers the evaluation of the CCLV model introduced in the first iteration in order to meet the analytical objective. The focus in this iteration is to re-construct the CCLV model as timeseries based model that can enable BankCo to monitor firms' CCLV score changes over time while using the size of the firms as a controlling variable.

This chapter is structured as follows: Section 5.2 describes the design phase of this iteration, which discusses the importance of having a stable network. Section 4.3 explains the development process of this DSR iteration and illustrates how the size-controlled timeseries-based CCLV analytical model is developed. Section 4.4 presents the implementation of the model and discusses a few examples that proved important for BankCo. Section 5.5 delivers a technical evaluation of the developed model as well as the evaluation of the sensemaking framework. Finally, Section 4.6 provides an overall summary of the chapter.

5.2 Iteration 2: Design an Improved CCLV Model

The design phase of the second iteration focuses on understanding the problems related to constructing a stable timeseries based network. A common problem that can be noticed in such cases is that of dealing with nodes and relationships that drop in and out of the network excessively across the time series. For example, Figure (5.1) illustrates how the network structure changes over time. In this example, it can be noticed that on the second day, client B does not transact with client X and the latter starts to transact with client F. On day 3, client B starts to transact with client X again and the latter does not transact with client F and starts to transact with client G for the first time. Developing a 'stable network', thus, is important for having reliable timeseries based results. This can be accomplished by studying the frequency of transactions among firms

Figure 5.1: Network structure changes over time



It is important as well, to consider the size of the firms as a controlling variable because that larger firms tend to be more connected; hence, causing biases in the CCLV score. Additionally, the sector, as another firmographic data, can cause the CCLV score to be unrealistic since clients from different sectors have different levels of business activity. For example, firms from the "retail" sector tend to have a higher number of transactions compared to clients from the "construction" sector. Consequently, accounting for the size of the firms and the sector they belong to is important to have a realistic timeseries-based CCLV score that can be useful for BankCo.

5.3 Iteration 2: Developing the Improved CCLV Model

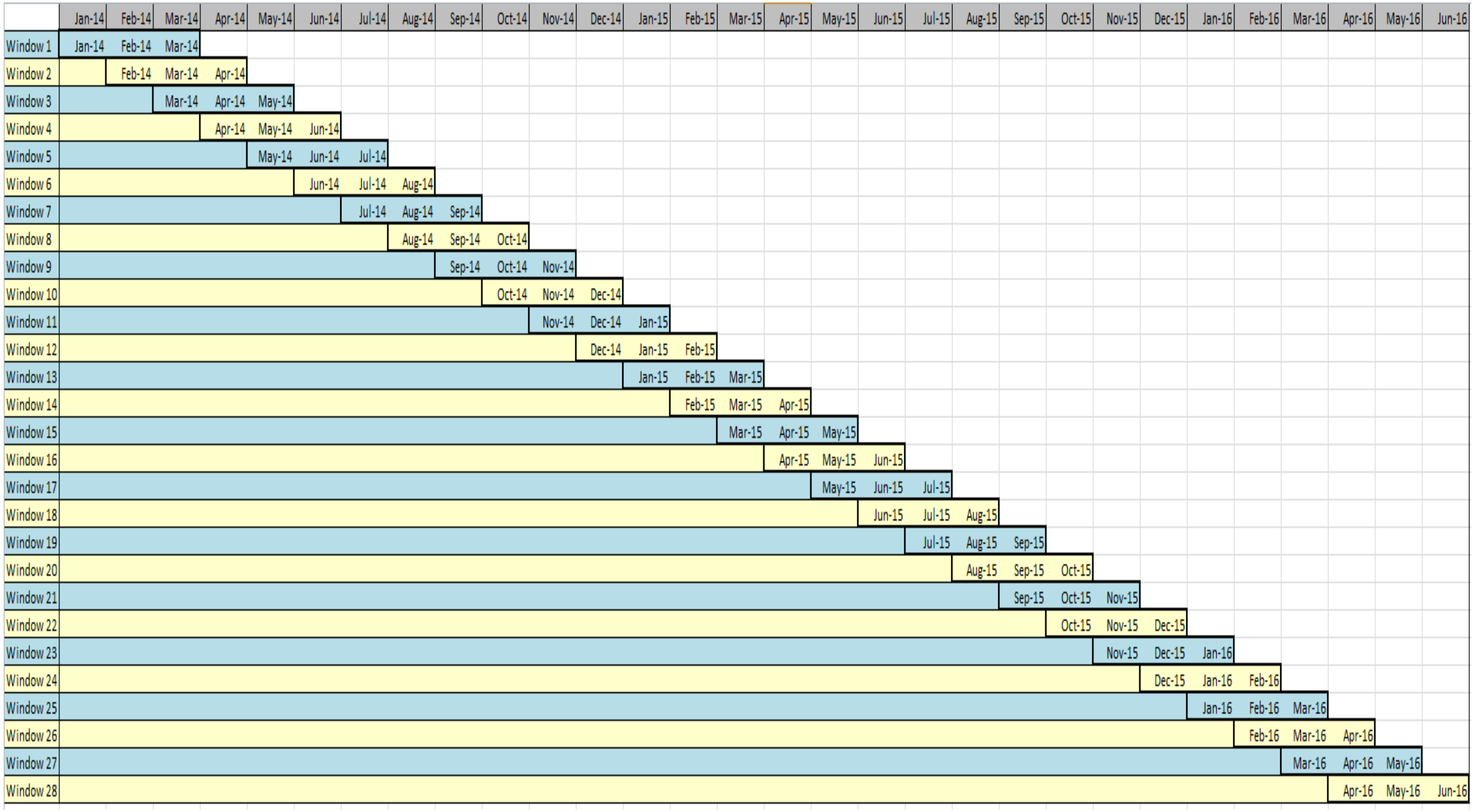
The development phase of this iteration focuses on understanding and interpreting the cues generated in the design phase and developing hypotheses in the form of analytical solutions. In order to address the network stability issue, a sliding window approach is adopted - set to a period which provides an acceptable level of network stability. To establish the optimal time-span for the windows, the transaction frequency was analysed. Table (5.1) show the transaction frequency analysis results.

Table 5.1: Transactions frequencies

Days	**Transaction Frequency %**
0 > 1	0.48
1 > 7	2.11
8 > 30	15.52
31 > 90	29.16
91 > 365	18.09
> 365	2.66
One-off	31.91

From Table (5.1), it can be noticed that approximately one-third of all transactions between nodes (firms) consisted of single (one-off) transactions. Thus, as an outcome of the frequency analysis, a period of 1 to 90 days was then selected as the basis for constructing a series of stable network, as approximately half of all transactions (47.27%) occurred in this period. A ‘sliding window’ approach was then implemented, consolidating 90 days into a window where the transactions were aggregated and edge weights calculated. So, months 0, 1 and 2 were consolidated into window period 1, months 1, 2 and 3 into window period 2, etc. This resulted in 28-window periods representing the timeseries and covers the 30 months. Figure 5.2 shows these 28 windows and the months that are covered in each window.

Figure 5.2: Times-Series Windows



The CCLV model developed in the previous iteration will be utilised and the CCLV scores for all the firms in the network will be calculated 28 times for each window illustrated above. Moreover, in order to control for the size of the firm and to consider the business sector, an approach presented by Larker (2013) is adopted. The following steps explain how the window-based sector-and-size-controlled CCLV is calculated:

1. Calculate the CCLV score for each window (28 times).
2. For each window, calculate the deciles for the “sizes” of the firms grouped by the business sector.
3. For each window, calculate the deciles for the “CCLV score” grouped by the business sector and the size deciles calculated step 2.
4. Repeat steps 2 and 3 for all 28 windows representing a time series CCLV.

Consequently, an ordinal (decile) timeseries CCLV is calculated for each firm within the network while using the size of the company as a control variable and grouping by the business sector. The use of the deciles provides a non-parametric approach to control for the effect of size on the original CCLV score. Additionally, a decile-based CCLV score is calculated for each sector which allows indicative performance comparison for the firm within its sector.

5.4 Iteration 2: Implementing the Improved CCLV Model

This section presents the implementation of the timeseries-based CCLV model on the transactional data provided by BankCo. To enable event spotting in the timeseries CCLV deciles, the standard deviation and variance were used to identify extreme changes over the time in the CCLV score. Table (5.2) presents a timeseries (window) view of the decile-based CCLV score representing the 28 windows constructed previously. It can be noticed from this table that firms sometimes do not have CCLV score in some windows. That is because those firms did not conduct in business transactions in those time windows. Thus, a filter has been placed to select only the clients with more than 14 decile-based

CCLV scores per client (more than 50% of the 28 windows). Also, it can be noticed that the data is ranked based on the standard deviation and variance values in a descending order to reflect the most extreme changes over the time.

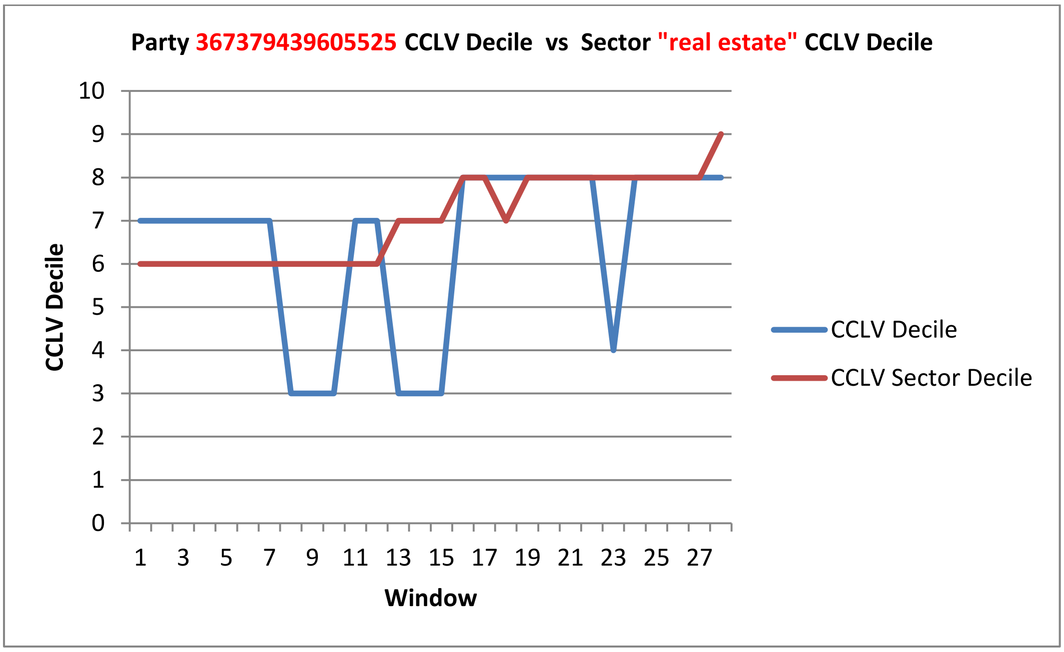
The highlighted lines in Table (5.2) represents two striking examples of the way the CCLV score changes over the time. In the first one with the client id “8461839533542971”, the CCLV score is increasing over the time in a drastic pattern, starting from 1 and rapidly ending up with 10. Such a pattern means that the number and value of business transactions generated by this client are growing rapidly over the time. In the second example, an opposite case appears because the CCLV score for client “8705044069419022” is decreasing over the time. Uncovering the amount of money spent and received by the client in the second example in the last 12 months (from July-2015 to Jun-2016), shows that the customer has spent £1.8M in the last 12 months and received £3.2M. This means that the customer started to be more influenced by other clients in the network than being an influencer, which conforms with the CCLV indicator.

Additionally, Figure (5.3) illustrates a comparison between the CCLV decile for one firm “367379439605525” against the CCLV decile for the sector it belongs to, namely “real estate”. Such comparison allows the monitoring of a firm’s performance against sector performance over the time and alerts for extreme changes.

Table 5.2: Clients with the highest standard deviation and variance in the timeseries-based CCLV score

pty_id	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	No of Entries	STD	Variance
952368066153341	1	1	1			1	1	1					1	1	1	1	6	10	10	10	10	10	10	10	10	10	10	10	22	4.50	20.21
4776111212192343		1	1	1	1	1	1	9	10	10	10	10	10	10	10	10	10	10	9	1									19	4.39	19.26
8461839533542971	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	6	10	9	9	9	10	10	10	10	10	10	10	10	28	4.36	19.03
8084738780218414							1	1	1	1	1	1	1	6	10	10	10	10	10	10	10	10	10	10	10	10	10	10	22	4.25	18.05
6860020656647354		1	1	1	1	1	1	1	1	1	1	1	1	1	1	9	9	9	9	9	6	9	10	10	10	10	10	9	27	4.21	17.76
88254075525865					10	9	9	1	1	1	1	1	1	10	10	10	1	1	1	1	1	1	1	1	9	9	9	1	24	4.19	17.54
1457604769039737	1	1	1	1	1	1	1	1	6	1	9	9	10	10	10	9	9	9	10	10	10	6	9	10	10	9	1	1	28	4.15	17.25
6348894751484781	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	10	10	10	10	10	9	10	10	27	4.13	17.09
9576537727959854	1	1	1	1	1		1	1	1	1	1	1	1	1	9	10	10	9	10	9	10	9	9	9	9	9	6	6	27	4.10	16.84
1045826343366054			1	1	1	1	1	1	1	1	1	1		1	1	1	1		1	1	1		10	10	10	10	10	10	23	4.04	16.33
5372180291267854					1	1	1	1	1	1	1	1	1	1	1	6	6	10	10	10	9	10							18	4.03	16.24
8705044069419022	10	10	10	10	9	10	10	10	10	10	10	9	9	9	9	9	6	1	6	1	6	1	1	1	1	1	1	1	28	4.00	16.04
5457445328813682	1	1	1	1	6	6	10	10	10	9	10	9	10	10	10	10	9	1	1	1	1	1	6	9	9	6	1	9	28	4.00	16.00
9315747176846036	10	10	9	10	6	10	10	10	9	1	1	1	6	10	10	6	1	6	6	1	1	1	1	1	1	6	1	1	28	3.98	15.88
1566783633221642	1	1	1	1	1	1	1	1	1	1	1	1	6	6	9	9	9	9	9	9	9	9	9	9	6	10	10	10	28	3.97	15.79
1457604710784095					1	1	1	1	1	10	10	10	6	1	1	1	1	1	1	6	6	10	10	10	1	1	1	1	24	3.97	15.77
9546314492117570			1	1	1			1	1	1	1	1	1	9	9	6	6	9	9	10	9	9	9	9	9	9	1	1	24	3.97	15.77
8083637903170881	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	6	1	9	9	10	10	10	9	9	9	10	6	6	28	3.96	15.68
9315747128529276	6	6			1	1	1	1	1	1	1	1	1	9	9	9	10	10	9	9	9	6							20	3.93	15.42
516360166443455	1	9	9	10	9	9	10	10	10	10	10	10	9	9	9	9	10	10	10	9	6	1	1	1	1	1	1	1	28	3.92	15.37

Figure 5.3: Firm and sector CCLV decile comparison



5.5 Iteration 2: Evaluating the Improved CCLV Model

This phase focuses on evaluating the timeseries-based size-controlled CCLV model developed in this iteration. The model is tested and evaluated in the context of the 28-window representing the timeseries and covers the 30 months of transactional data.

5.5.1 Stakeholder Evaluation

Following the process illustrated in Figure (3.4), the improved CCLV model and the implementation results of the second iteration were presented to and discussed with the BankCo team. During Part 1 of the evaluation process, it was explained how the concerns noticed during the evaluation stage in the previous iteration were considered as cues for this iteration. Also, it was explained how the interpretation of these cues led to the need to perform a transaction frequency analysis to construct a stable network over time. Then, the timeseries-based size-controlled CCLV model was demonstrated alongside the implementation examples (discussed in Section 5.4), highlighting the interesting patterns and extreme changes in the value of the CCLV score over time for two firms.

As before, in Part 2 of the evaluation process, the analysts at BankCo de-anonymised the data and looked up additional information related to the two contrasting examples highlighted in the implementation steps (as shown in Figure 5.2). Though the model outcomes were of concern to BankCo, additional information was required to understand cause – timeseries-based risk factors, liquidity and credit history, for example. If a client is flagged as problematic, because their influence is decreasing over time, then there is a need to investigate and clarify potential reasons that may cause or motivate such behaviour (such as a change of ownership or liquidity-related factors, which is information not available to the researcher).

The outcomes here were particularly interesting on two counts. First, it was apparent that BankCo had not considered their data in this way and realised they did not have an 'early warning' system to flag customers that they were at risk of losing. Second, the model-based analysis and consequent need for additional information provided an illustration of automated and human effort working together in sensemaking. A conclusion here was that BankCo's analysts need to consider more CCLV time series examples with extreme increases or decreases in the scores, study the reasons and examine whether patterns could be discerned (which might lead to further automated analysis at some future point). From BankCo's point of view, the modified CCLV model represents an analytical system with the following possible abilities:

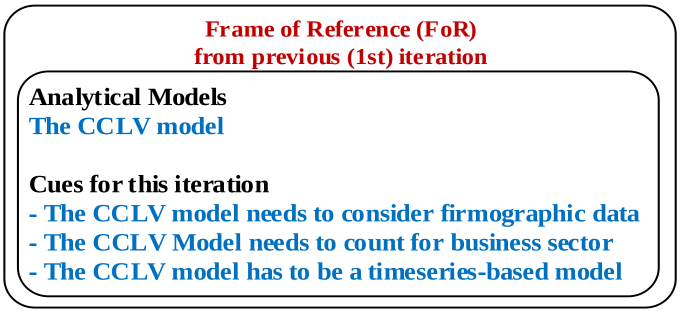
1- The relative (decile-based) ranking of all firms in a network;
2- Firm influence comparison with the sector to which the firm belongs;
3- Timeseries-based event spotting (and alerting) in relation to deviations from (agreed) norms.
4- A system to flag opportunities for BankCo, as a banking products recommendation system. For example, the CCLV model allows BankCo to target customers with a decreasing CCLV score with lending products.

During Part 3 of the evaluation process, the researcher and BankCo's analysts concluded that any additional improvement to the CCLV model required access to additional information that was not available to the researcher – thus, no further development was required of the CCLV model. Importantly, as has been noted, the work did trigger discussion of how the timeseries-based size-controlled CCLV model represents a long-term relationship (mutual influence) among the customers in the network. Their lack of understanding in that regard triggered the collective BankCo mind into wanting to investigate the long-term relationship between BankCo and its customers in the network over time (providing a new requirement for the next iteration).

5.5.2 Evaluating the Sensemaking Framework

This subsection demonstrates the sensemaking work that guided the development of the improved CCLV model. The initial Frame of Reference (FoR) for this iteration has resulted from the previous one and contained the original CCLV model as well as the issues raised regarding its performance, see Figure (5.4). These issues are based on the feedback from the research partner (BankCo) and include reconstructing the CCLV model in a timeseries format, using the size as a controlling variable and considering the business sector.

Figure 5.4: Initial FoR for the 2nd iteration



To address the first issue, the researcher has noticed (noticing process) a cue related to firms transacting irregularly, which means that transaction patterns and frequencies of the firms change over the time. Consequently, it was interpreted (interpretation process) that in order to have reliable timeseries based results, establishing a stable network was a necessity. Hence, as an action process, transaction frequency analysis was performed to settle on the most stable period of time. As explained in Section (5.3), interpreting the results of this analysis (interpretation process), see Table (5.1), has led the researcher to consider a period of 1 to 90 days as the base for the 'sliding window' approach, since approximately half of all transactions (47.27%) occurred in this period; thus, consolidating 90 days into a window where the transactions were aggregated, and edge weights calculated.

The size of the firms and the sector they belong to were additional cues that the researcher had to consider. These cues were interpreted (interpretation process) as concerns for the reliability of the CCLV score, as the goal is to look at the true effect of firm's connectedness rather than 'imperfect' proxies because the relationship between network variables and key outcome variables might be disturbed by the effect of size and the business sector. Consequently, as an action process, ordinal (decile) timeseries-based size-controlled CCLV scores are calculated for the firms in each business sector in the network to meet the concerns raised from the evaluation of the CCLV model developed in the previous DSR iteration.

To describe the questioning, elaborating and reframing cycle displayed in the sensemaking framework, see Figure (2.4), the initial Frame of Reference (FoR) for this iteration was questioned to investigate the issues included in that FoR, and elaborated by hypothesising that 1) transactions frequency analysis could help establishing stable network by defining the number of days for the window approach and 2) using the size of the firms, as a firmographic variable, can provide a non-parametric approach to control for the effect of size on the original CCLV score. Consequently, based on the feedback from BankCo on the implementation results and examples discussed in Section (5.4), the initial FoR was reframed with the noticed cues, theorised hypothesis and the improved CCLV model, see Figure (5.5).

Figure 5.5: Reframed FoR in the 2nd iteration

Reframed Frame of Reference (FoR)

Cues
- Constructing stable network is necessary because of the firms' irregular transaction patterns
- The size of the firm & the business sector are important to consider for unbiased CCLV score

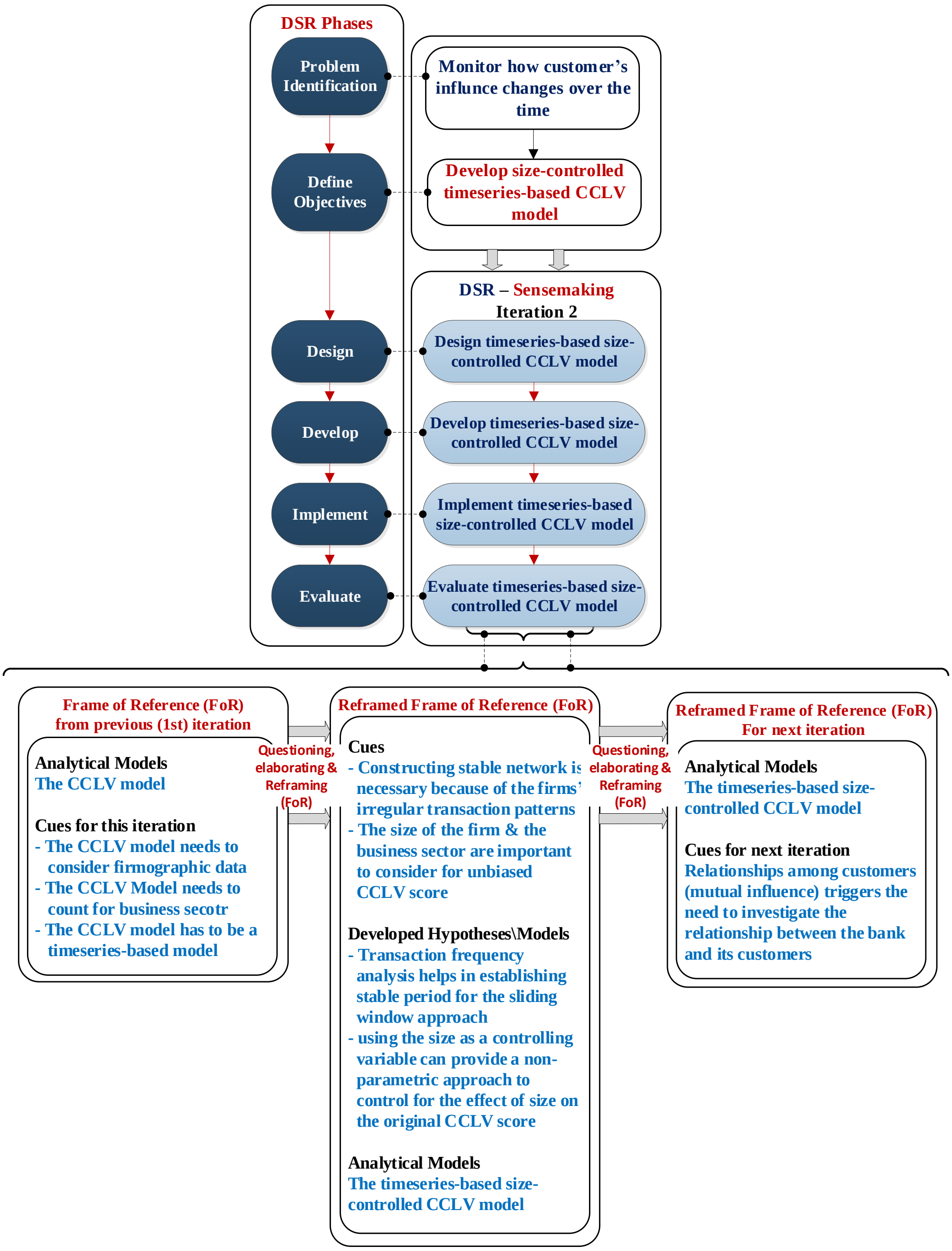
Developed Hypotheses\Models
- Transaction frequency analysis helps in establishing stable period for the sliding window approach
- using the size as a controlling variable can provide a non-parametric approach to control for the effect of size on the original CCLV score

Analytical Models
The timeseries-based size-controlled CCLV model

The modified CCLV model provides a useful tool to spot changes in the clients' influence on their neighbourhood over the time; enabling a dashboard-like mechanism to identify firms with critical transactional patterns and to understand the relationships (mutual influences) among the customers over time. The new requirement, which is about investigating the long-term relationship between the bank and its customers over time, has led to reframing the FoR again with the new requirement as a cue for the next iteration explained in the following chapter. Thus, the next iteration will focus on designing, developing, implementing and evaluating a timeseries-based model to understand the long-term relationship between BankCo and its the customers.

Figure (5.6) illustrates an overall view of the second iteration presented in this chapter. This figure shows how the work in this DSR iteration was built on the FoR resulting from the previous one; leading to a reframing process of the FoR that will be used in the following chapter, which reflects the third DSR iteration.

Figure 5.6: Overall view of the second iteration



5.6 Summary

This chapter has presented the second iteration of this work and demonstrated how sensemaking processes happen during this iteration. This chapter has explained how calculating and structuring CCLV results in a timeseries format allows BankCo to spot extreme changes in its score over the time and enable them to spot potential opportunities for recommending banking products. Constructing the optimal timespan for the timeslots (windows) for the timeseries, however, needs to consider the frequency of transactions conducted among firms. Additionally, considering the sector the firms belong to, and controlling for the size of the firms allows for having more realistic results that reflect the true effect of the firm’s connectedness within the network representing the transactional data.

Chapter 6: Network Relationship Equity (NRE) Model (3rd Iteration)

6.1 Overview

This chapter presents the third iteration of this study. The analytical objective in this iteration is about understanding and evaluating the long-term relationship between BankCo and its customer. This iteration investigates how Social Network Analysis (SNA) techniques can help to evaluate and quantify the long-term relationship between BankCo and its clients in a way that allows BankCo to get valuable insights and enhance the human sensemaking process. The aim here is to develop a Network Relationship Equity (NRE) model that addresses further requirements presented by the bank. This model can help BankCo monitor the increase and decrease in customers' activity within BankCo's network of customers compared to their transactional activity with customers from outside BankCo's network (the market) over the time and act upon it retrospectively. Such a model may help BankCo to prevent losing valuable customers by recommending appropriate products.

This chapter is structured as follows: Section 6.2 describes the design phase of this iteration, which discusses the design of the NRE model. This section defines the concept of Relationship Equity (RE) and explores how it can be used in the context of the banking transactional data. Section 6.3 explains the development process of this DSR iteration and illustrates how timeseries-based Network Relationship Equity (NRE) analytical model is developed. Section 6.4 presents the implementation of the developed model. Section 6.5 demonstrates the technical evaluation of the NRE model as well as the evaluation of the sensemaking framework. Finally, Section 6.6 provides an overall summary of the chapter.

6.2 Iteration 3: Designing the NRE Model

Similar to the previous two iterations, the design process of the third cycle is about understanding the analytical requirements, defining the analytical objective for this iteration and designing the solution that can meet the objective. The requirement here is to understand clients' long-term relationship with BankCo (loyalty to the bank). Thus, it can be noticed that Customer Equity (CE) and Relationship Equity (RE) can be considered as the research fields to address the requirement, since that they investigate what makes a client loyal to a brand or a business.

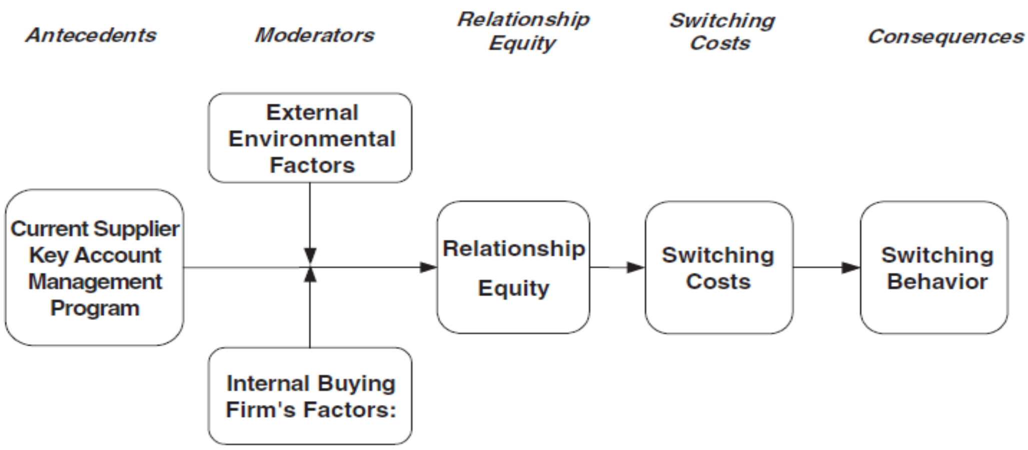
6.2.1 Customer Equity and Relationship Equity

Customer Equity (CE) can be defined as the sum of the lifetime values of a company's current and potential customers (Segarra-Moliner and Moliner-Tena, 2016). Measuring CE, thus, depends, predominantly, on the Customer Lifetime Value (CLV) metric (Segarra-Moliner and Moliner-Tena, 2016). This in turns depends on the revenues obtained from the customer as well as the costs related to customer acquisition and/or retention (Leone *et al.*, 2006). In case of customer equity, however, in order to optimise the calculation of the CLV, drivers such as value equity, brand equity, retention equity and relationship equity need to be considered (Segarra-Moliner and Moliner-Tena, 2016). That is because of the influence these drivers have on customer lifetime value (CLV) as well as customer behaviour (Segarra-Moliner and Moliner-Tena, 2016; Vogel, Evanschitzky and Ramaseshan, 2008).

Relationship Equity (RE), as one of CE drivers, measures a customer's loyalty over time to a firm or brand regardless of the customer's objective and subjective assessment of the purchased good and/or service (Zhang, van Doorn and Leeflang, 2014). Drivers for RE include trust, satisfaction and commitment (Zhang, van Doorn and Leeflang, 2014; Low and Johnston, 2006). Clients tend to stay loyal to the firm they are dealing with if they feel that they are being valued and that switching to a competitor would impose additional costs (Low and Johnston, 2006). Additionally, studies show that factors related to company image, brand image, intense price competition, external market competition, and relationship value and quality can affect customer decisions related to staying loyal or switching to competitors (Chen and Myagmarsuren, 2011; Low and Johnston, 2006).

The importance of Relationship Equity (RE) has motivated researchers to investigate the factors or dimensions that affect RE. In an attempt to conceptualise RE, Low and Johnston (2006) have proposed a model, see Figure (6.1), which illustrates the factors affecting RE and leads the customer to switch the supplier of the good and/or service. In this model, key account management related factors, considered as antecedents, include customer care, communication channels, meeting customers' expectations, trust, and conflict resolution skills (Chen and Myagmarsuren, 2011; Low and Johnston, 2006). While the supplier is able to control these antecedents, it is hard to control for moderators that can affect RE. These moderators can be external such as competition in the market and external technological uncertainty, and internal such as satisfaction with the rewards provided by the firm's suppliers. Comparing these rewards can lead a firm to a decision of switching supplier. Additionally, switching costs related to technological or relationship compatibility between the buyer and the new potential supplier is crucial to consider (Low and Johnston, 2006).

Figure 6.1: A conceptual model of relationship equity (Adapted from (Low and Johnston, 2006))



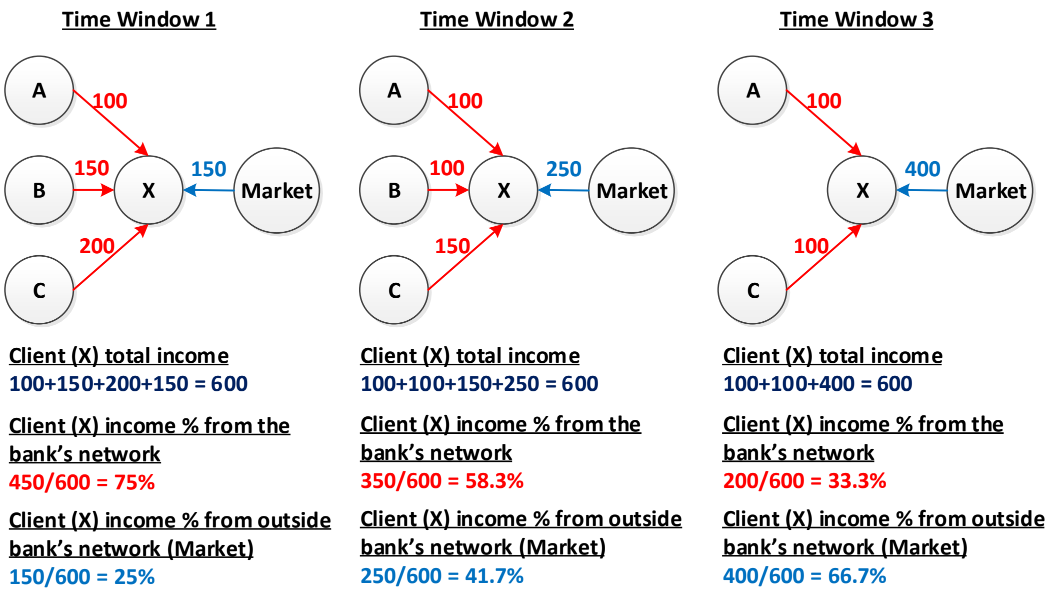
It can be concluded from the discussion above that the concept of Relationship Equity (RE) is related to a customer staying loyal to a firm over time. Though there are several factors to consider, the limitation imposed by the data provided by BankCo makes it difficult to investigate these factors. It can be noticed, however, that in case of the transactional data (network-based data), the Relationship Equity (RE) model that could reflect a limited version of the relationship between BankCo and its clients can be observed by monitoring firms spend and income (buyer and supplier) behaviour over time. The observation has to consider BankCo's 'transaction universe' (BankCo's network of customers), compared to outside of it (clients who are not customers of BankCo).

6.3 Iteration 3: Developing the NRE Model

The development phase (interpretation process) of this iteration focuses on building on the cues noticed during the design phase in order to develop an analytical solution that can address the objective of the iteration. Consequently, this phase introduces a network-based model, namely Network Relationship Equity (NRE) model, that can investigate the relationship between BankCo and its clients from the transactional data. The NRE model looks at the first-order network in the transactional data and investigates the income and spend for each client. Utilising the window approach introduced in Chapter 5, the following two examples for income and spend show a sample network of customer (X) at three different times and explain how the NRE model works.

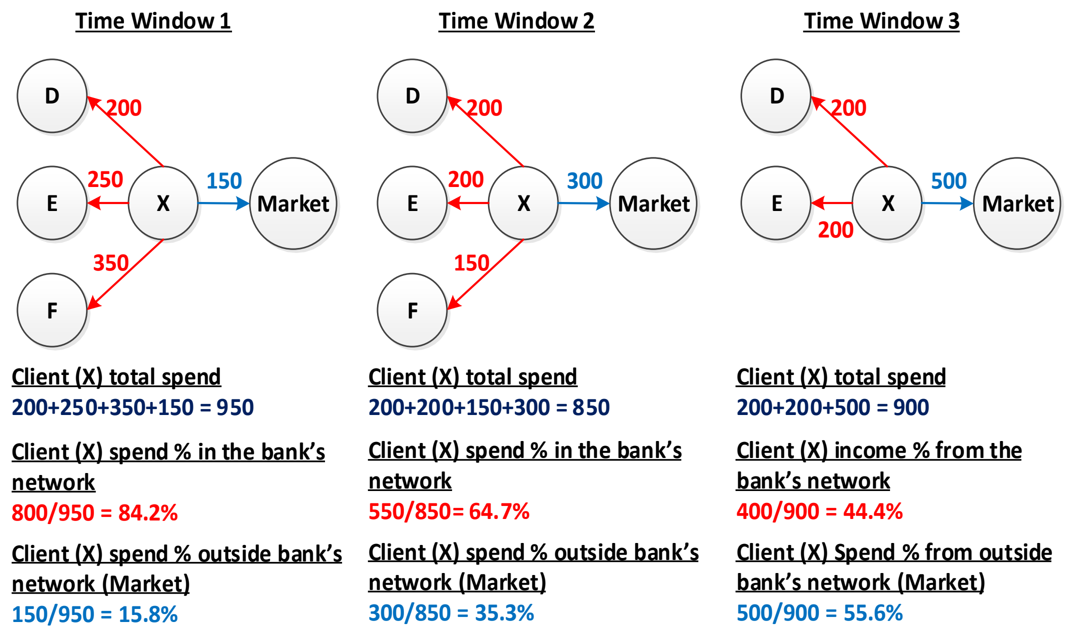
In the first case, the changes in the income can be monitored in the transactional data, see Figure (6.2). In this figure, it can be observed that the income from BankCo's network is decreasing (75%, 58.3% and 33.3%) in favour of the income from the market (25%, 41.7% and 66.7%).

Figure 6.2: Client (X) income from the BankCo's network vs income from the market



Similarly, the spend pattern for client (X) can be monitored over time in the transactional data, see Figure (6.3). Client (X) spend in BankCo's network is decreasing over the time (84.2%, 64.7% and 44.4%) compared to increases (15.8%, 35.3% and 55.6%) outside BankCo's network of customers.

Figure 6.3: Client (X) spend in BankCo's network vs spend in the market



Considering that it is in BankCo's interest for transactions to take place within its own network of customers, noticing such shifts in client's income and spend can be of great importance to BankCo to investigate the reason behind them. Similar to the timeseries-based CCLV, standard deviation and variance are used to identify patterns as a basis for action.

6.4 Iteration 3: Implementation the NRE Model

This section presents the implementation of the timeseries-based NRE model using the banking transactional data. To enable event spotting, the standard deviation and variance were used in descending order to identify extreme changes over time in the NRE score. Table (6.1) shows the percentage of income from BankCo's universe of transactions in 28 windows for 15 clients with the highest value of standard deviation and variance. It can be noticed from this table that when firms have income percentage equals to 0%, it means that their income in that window was completely from outside BankCo's network of customers. Thus, a filter has been placed to select only the clients with more than 14 entries (more than 50% of the 28 windows) for income percentage per client.

The highlighted lines in Table (6.1) represent two different examples of the way the NRE percentages change over the time. In the first one with the client id “6008353597410248”, the NRE percentage for income is decreasing over the time, starting from 99.3% and ending up with 0% in window 25 onwards. Such a pattern means that the client is switching their business to work with customers from outside BankCo’s network. In the second example, the NRE (income) score for the client “2574150494897542” is increasing throughout the 28 windows of the transactional data. Table (6.2), expands on the first example for the client “6008353597410248”, and provides the data at a granularity level that shows the movement of income across the time series for all parties transacting with this client including the clients from the market grouped in client “000000000001”.

Table 6.1: Timeseries-based NRE (income) percentages

Pty_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	No of Entries	STD	Variance
6926622069019719	3.69	3.69	3.69	0	0	0	0	99.51	98.97	98.56	98.6	98.67	98.7	98.73	98.68	98.57	98.45	98.45	98.5	98.51	98.58	5.71	5.65	4.49	4.79	4.87	5.94	98.42	24	48.47229964	2349.563832
4963909433774944	0	0	94.1	92.31	90.13	0	99.65	99.74	99.81	99.95	99.95	0	99.87	99.8	96.1	95.97	96.8	42.46	13.22	0.25	0.36	0.48	0.95	0.9	1.59	0.99	0.9	0.25	24	47.99988336	2303.988803
7373621352971413	0.03	0.02	0.01	0	0.19	0.33	0.28	0	0	0	4.35	17.96	75.39	95.6	96.89	97.13	98.05	97.84	97.9	97.57	97.53	95.55	95.09	93.49	96.11	97.21	95.93	95.55	24	47.28158458	2235.54824
6696807800181341	98.72	96.64	97.47	98.49	98.78	97.25	99.6	99.65	99.65	0	0	0	99.91	99.91	99.91	0	0	0	0	0	93.09	12.97	15.15	10.79	17.54	12.11	12.4	19.91	20	46.9053835	2200.115002
6681088130057194	0	99.15	99.13	99.2	99.99	99.99	99.86	99.45	99.05	98.19	96.05	95.85	96.77	0	59.27	32.92	7.35	0.66	0	3.15	2.11	2.52	0	0	8.93	10.26	9.86	0	22	46.87995747	2197.730412
6008353597410248	99.3	99.36	97.09	97.87	97.99	99.24	97.18	98.4	98.88	98.76	96.02	94.05	94.98	94.98	0	0	0	20.75	31.36	1.7	0.76	29.54	9.28	5.07	0	0	0	0	21	46.72996029	2183.689188
9546314454844356	2.48	2.32	1.74	0	0	0	0	0	0	0	0.23	0.47	31.9	43.68	53.54	0	99.71	99.64	97.88	99.71	99.94	99.94	94.24	89.21	89.21	96.06	98.05	98.96	20	46.48862071	2161.191855
6982001203115122	0	0	0	0	99.01	98.71	98.59	0	0	3	3.27	3.16	0	99.41	99.42	99.4	0	96.44	95.23	91.13	90.62	88.44	91.4	91.42	91.32	89.06	89.11	89	20	46.44745301	2157.365892
6074939073517698	99.68	99.54	99.43	0	0	0	0	90.58	87.33	73.16	98.31	98.17	99.59	99.46	99.51	0	0	0	0	72.5	64.65	97.72	99.51	99.52	95.33	13.3	3.73	3.14	20	46.04621128	2120.253574
3782026838331997	94.29	93.91	94.25	99.39	98.67	98.56	91.61	91.22	90.7	99.81	99.81	99.85	0	0	99.39	98.2	97.12	0	0	0	0	0	13.29	13.04	11.44	0	28	23.89	20	46.04038476	2119.717029
6074939012217425	35.21	39.6	99.83	99.95	99.94	99.94	0	0	0	0	99.92	99.83	99.83	99.9	0	99.95	99.9	99.91	99.96	0	0	99.94	99.94	99.87	99.87	99.9	0	99.93	20	46.01276746	2117.174769
7766588964606933	0	0	0	0	0	99.99	99.99	99.99	99.99	100	0	0	0	99.99	99.98	99.97	99.98	99.99	99.97	99.55	99.67	99.75	99.99	99.93	99.92	99.95	100	99.91	20	45.97021342	2113.260522
2574150494897542	0.29	1.78	4.75	4.82	3.49	0.5	0.81	0.57	0.61	0.13	0.85	19.75	59.72	97.35	96.45	96.88	90.09	90.48	88.12	95.53	95.19	99.37	99.07	99.07	99.07	93.35	93.35	93.35	28	45.89321624	2106.187297
7035158540326862	0	0	0	100	100	99.69	99.77	99.82	0	0	99.93	99.94	99.94	99.96	0	0	99.43	99.1	98.96	99.78	99.74	99.76	99.84	99.98	99.87	99.83	99.79	0	20	45.89296141	2106.163907
8747497109653015	99.06	97.65	98.47	98.83	0	96.52	96.7	95.23	99.79	99.51	0	0	98	98.33	98.74	0	99.68	99.78	99.77	0	0	0	99.22	97.8	15.94	15.6	18.26	99.57	21	45.87120606	2104.167545

Table 6.2: All clients that provide income to client "6008353597410248" over the 28 windows

To_pty_ID	From_pty_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
6008353597410248	0000000000000001	861.50	861.50	3,861.50	3,826.50	3,826.50	826.50	826.50	826.50	826.50	911.50	911.50	911.50	761.50	761.50
6008353597410248	6008353531128776	121,445.00	133,945.00	128,995.00	175,450.00	186,450.00	108,450.00	28,500.00	50,835.00	72,815.82	72,815.82	21,980.82	14,400.00	14,400.00	14,400.00
6008353597410248	6008353519943054	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6008353597410248	6008353594618216	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
To_pty_ID	From_pty_ID	15	16	17	18	19	20	21	22	23	24	25	26	27	28
6008353597410248	0000000000000001	761.50	955.00	955.00	955.00	984.99	25,983.99	25,983.99	25,968.50	106,413.83	203,801.83	251,937.41	193,881.08	143,881.08	160,733.04
6008353597410248	6008353531128776	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6008353597410248	6008353519943054	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5,444.33	5,444.33	5,444.33	0.00	0.00	0.00	0.00
6008353597410248	6008353594618216	0.00	0.00	0.00	250.00	450.00	450.00	200.00	5,444.43	5,444.43	5,444.43	0.00	0.00	0.00	0.00

Similarly, Table (6.3) shows the percentage of spend within BankCo's network of customers in 28 windows for the 15 clients with the highest value of standard deviation and variance. It can be noticed from this table that when firms have NRE scores equal to 0%, it means that they are only conducting business transactions and paying customers from outside BankCo's network. Thus, a filter has been placed to select only the clients with more than 14 entries (spend percentage) per client. Also, it can be noticed that the data is ranked based on the standard deviation and variance values in a descending order to reflect the most extreme changes over the time. The highlighted lines in Table (6.3) represents two opposite examples of the way the NRE percentages change over the time for the spend by these two clients.

In the first one with the client id "2432062256840480", the NRE percentage for spend by the client is increasing over the time, starting from 0.71% and reaching 96.23% in window 28. Such pattern means that the client is switching their business to be fully with customers from within BankCo's network. In the second example, the opposite case appears, where the NRE score for spend by the client "3015020579132517" was reasonably stable until window 20 and then suddenly it becomes 0% in window 21 onwards. Additionally, Table (6.4), expands on the second example for the client "3015020579132517", and provides the data at a granularity level that shows the movement of spend across the time series for all parties transacting with this client including the clients from the market grouped in client "000000000001". In this table, it can be noticed how the client "3015020579132517" is shifting their business transactions outside the bank's network until eventually stops using this bank account in window 22. Spotting such a pattern might be of a high importance for the bank in order to do preventive actions and to lose customers.

Table 6.3: Timeseries-based NRE (spend) percentages

Pty_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	No of Entries	STD	Variance
9737233563225942	0	0	0	0	0	0	0	2.27	2.36	2.21	2.25	1.5	1.68	1.25	83.31	91.89	95.61	96.73	98.35	98.27	98.63	96.73	96.78	94.41	96.91	95.97	97.31	94.23	21	48.1489374	2318.320173
8429118202421375	0.79	0.75	1.03	1.02	1.17	0.62	0.7	0.33	0.44	0.24	0.39	0.29	0.3	20.76	49.82	81.51	89.53	92.86	98.88	99.31	99.35	99.66	99.62	99.49	99.27	99.05	99.2	99.07	28	47.56808403	2262.722618
4502689428019697	86.07	0	0	0	0	0	0	0	97.75	98.47	98.79	98.31	98.44	96.8	91.44	95.32	95.29	96.12	1.06	0.63	3.65	90.23	91.95	92.55	3.74	3.14	2.85	11.12	21	47.45156883	2251.651384
6740241487710907	20.96	14.39	4.19	0	0	0	0	0	0	0.42	15.78	72.02	99.78	99.78	99.68	99.61	99.57	99.5	99.52	99.6	99.57	99.52	99.69	99.6	99.34	0	93.92	94.22	21	47.1804094	2225.991031
2506429240612427	2.94	2.39	28.28	95.52	0	0	0	96.98	98.14	97.79	99.43	99.37	99.41	98.95	97.29	72.17	96.54	97.7	97.68	0	0	0	0	0	99.05	98.79	98.54	92.31	20	46.87945156	2197.682978
2432062256840480	0.71	2.1	2.34	2.67	2.19	2.42	2.42	1.92	1.73	36.89	56.41	93.48	99.65	99.46	99.58	99.8	93.89	91.74	87.23	99.13	99.45	99.61	99.35	99.21	99.16	99.74	0	0	26	46.62953188	2174.313243
6492640738446420	0	0	0	0	0	0	0	98.65	97.11	95.01	61.85	52.61	4.89	6.35	5.76	99.52	96.65	6.04	4.16	3.94	98.55	97.28	96.18	95.73	94.36	94.84	90.23	95.51	21	46.11742002	2126.816429
7958748549306951	0	97.84	97.93	98	92.7	92.38	92.96	96.8	97.91	97.01	97.45	96.76	97.01	96.73	94.07	93.39	0	0	0	0	0	1.61	0.58	18.31	19.03	20.51	11.27	5.7	22	46.07008995	2122.453188
4056620389361128	99.99	99.99	99.99	99.99	99.93	99.93	99.92	99.87	99.9	99.92	99.99	99.98	99.78	98.36	99.63	99.82	99.88	0	0	0	0	98.67	40.63	40.63	0	0	0	0	20	45.79144023	2096.855998
9737233596016125	0	0	0	0	0	99.73	99.78	99.6	99.39	99.15	98.92	98.9	98.53	99.41	99.46	99.71	99.6	99.63	99.69	99.71	99.81	99.97	99.88	99.57	99.35	0	0	0	20	45.77055437	2094.943647
5457445350024594	0	0	0	99.6	99.94	99.93	99.95	0	99.83	0	99.41	20.62	20.6	0	99.42	99.71	99.71	99.55	99.55	99.55	68.55	59.26	37.34	27.32	9.48	0.58	0	0	20	45.73517242	2091.705996
1604424072159436	0.61	0.66	0.59	1.38	1.11	1.13	0.04	0.03	0.03	0.13	0.17	0.18	13.52	26.93	58.14	64.82	75.77	84.65	91.08	99.81	99.72	99.85	99	99.02	98.76	98.07	98.17	98.45	28	45.71297214	2089.675822
3015020579132517	99.98	99.94	99.91	99.91	99.91	99.82	99.12	99.19	99.27	97.63	97.23	97.76	99.87	99.87	99.73	99.75	99.8	99.44	99.02	97.81	0	0	0	0	0	0	0	0	20	45.66462313	2085.257805
6837257912344024	97.35	96.95	96.77	99.2	92.88	86.69	86.57	92.7	99.49	99.5	99.12	96.1	92.34	67.21	38.28	10.06	0.09	0.09	0.1	0.16	0.17	0	0	0	0	7.43	7.36	10.15	24	45.66261458	2085.07437
3180882098508050	97.94	98.94	99.54	99.51	99.51	99.01	99.37	99.12	99.3	99.38	99.61	99.65	99.52	98.76	0	99.4	99.63	99.69	99.37	98.27	0	0	0	0	0	0	0	98.25	20	45.6330751	2082.377543

Table 6.4: All clients that receive income from client “3015020579132517” over the 28 windows

From_pty_ID	To_pty_ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
3015020579132517	0000000000000001	45,460.44	47,330.01	47,846.57	48,186.08	54,276.09	51,942.01	51,657.50	56,158.49	63,873.67	60,198.67	54,050.67	63,964.38	60,842.38	58,059.38
3015020579132517	1184727407412267	0.00	0.00	0.00	0.00	0.00	0.00	0.00	490.00	490.00	490.00	100.00	660.00	660.00	560.00
3015020579132517	2886816973694169	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1,230.00
3015020579132517	5891027666080177	1,468.00	1,468.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3015020579132517	9144506691407609	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3015020579132517	9762541993884158	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

From_pty_ID	To_pty_ID	15	16	17	18	19	20	21	22	23	24	25	26	27	28
3015020579132517	0000000000000001	37,375.60	41,185.61	34,965.61	24,427.01	10,471.00	3,893.00	1,120.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3015020579132517	1184727407412267	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3015020579132517	2886816973694169	1,255.00	1,255.00	25.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3015020579132517	5891027666080177	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3015020579132517	9144506691407609	275.00	275.00	275.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3015020579132517	9762541993884158	700.00	700.00	700.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

6.5 Iteration 3: Evaluating the NRE Model

This section focuses on evaluating the NRE model developed in this iteration. The model is tested and evaluated in the context of the 28-window representing the timeseries and covers the 30 months of transactional data.

6.5.1 Stakeholder Evaluation

Following the process showed in Figure (3.4), the NRE model and implementation results of the third iteration were presented and discussed with the analysts from BankCo. During Part 1 of the evaluation process, the reasons for investigating the long-term relationship between BankCo and its customers were replayed and then related to the studies that cover Customer Equity (CE) and Relationship Equity (RE). Also, explanation was provided as to how the developed NRE model differentiates between the buyer and supplier relationship with the bank (spend and income) inside and outside BankCo's network over time (considering the limitations in the data, which prevents consideration of all of the factors related to CE and RE). Next, the implementation examples (discussed in Section 6.4) were demonstrated, highlighting the interesting patterns and extreme changes in the value of the NRE score over time for four firms in both the income and spend versions of the NRE model.

As per the previous iterations, in Part 2 of the evaluation process the analysts at BankCo de-anonymised the data and looked up additional information related to the four examples highlighted during the implementation step of this iteration (presented in Tables 6.1, 6.2, 6.3 and 6.4). The examples here were of key importance as they showed different types of pattern that were of direct import to business – the potential loss of a customer or an increase in business, for example. What was striking to the researcher, here, was that BankCo had no existing early warning system in this respect and were surprised themselves to see such patterns. As in the previous iteration, the causes underlying patterns were not readily available in the data and required further investigation on both counts. For example, a customer might shift to another bank (or shift the balance of business to another bank) because they are unhappy with their terms, start to deal with a new supplier that is not a customer of BankCo or because banking products offered by another bank are more innovative or favourable. Conversely, spend patterns showing an increase in the percentages that customers spend within BankCo's network, arguably represent opportunities for BankCo to recommend relevant banking products, cross-sell or up-sell products (depending on other factors that may be seen as causal). As things stand, these are missed opportunities.

Similar to the timeseries-based CCLV model, the NRE model thus presents an event-spotting mechanism to alert BankCo to patterns/deviations in the data. From BankCo's point of view, the spend perspective of the NRE model offers more interesting opportunities for the following reasons:

1- The timeseries-based NRE model provides 'trace points' from which an analyst/relationship manager can investigate related data to understand potential causes of change (e.g., price, credit status, sector-related factors) and act upon them (e.g., via product modification, cross-sell, etc.). For example, the value of the model was seen in its ability to spot the potential loss of a customer, allowing preventative measures to be put in place (where required) – such as the case where a customer holds long-term loans and their spend within BankCo's network decreases over time.

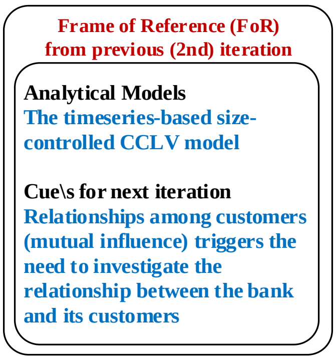
2- The model allows the uncovering of opportunities for BankCo to discuss and offer new banking products to customers based on their spending patterns.

In Part 3 of the evaluation process, it was concluded that additional information (not accessible to the researcher) needs to be investigated in order to explore the full capacity of the timeseries-based NRE model. Indeed, it may be the case that causal information has to be gleaned via a Relationship Manager having a conversation with a client. Thus, no further development was required on the NRE model at this stage. Interestingly, however, the discussion led one senior BankCo staff member to stress that a deeper understanding of customer behaviour was required (related to acquiring banking products in particular). Further, they reported that they had been looking at the concept of 'personas' as a means of doing this, but that: (a) the work was outside of the 'day job' and the team had been struggling to do anything meaningful; and (b) that the network perspective that had been worked up here provided one way of examining 'behaviour'. Consequently, this was formalised as a requirement for the next iteration. What was particularly interesting from the sensemaking perspective here was that it was very unlikely that this the requirement could have been formulated *a priori* – it came explicitly from the 'journey' of the previous iterations.

6.5.2 Evaluating the Sensemaking Framework

This subsection demonstrates the sensemaking work that guided the development of the NRE model. The initial Frame of Reference (FoR) for this iteration resulted from the previous one, see Figure (6.4). In that FoR, it is stated as a cue that studying relationships among customers (mutual influences) in the CCLV model has triggered the need to investigate the relationship between BankCo and its customers (loyalty to the bank).

Figure 6.4: Initial FoR for the 3rd iteration



Consequently, the noticing process drew attention to the research fields that study Customer Equity (CE) and Relationship Equity (RE), as they investigate what makes a client loyal to a brand or a business. As explained in Subsection (6.2.1), studying the literature that is related to CE and RE, however, has helped the researcher to notice that there are limitations imposed by the transactional data provided by the research partner that prevent investigating all the factors affecting CE and RE. Additionally, noticing and comparing firms' transactional patterns in the network-based data over the time has helped the researcher to interpret customers' income and spend inside and outside BankCo's network as an approach that reflects the relationship between BankCo and its customers over time. Consequently, as an action process, the Network-based Relationship Equity (NRE) model was designed, developed and implemented to address the objective set for this iteration.

To explain the questioning, elaborating and reframing cycle, the initial Frame of Reference (FoR) for this iteration was questioned to investigate the requirement regarding investigating the relationship between BankCo and its customers. Hence, the FoR was elaborated by hypothesising that Customer Equity (CE) and relationship Equity (RE) are the research fields to address the requirement since they investigate what makes a client loyal to a brand or a business. Consequently, after developing the NRE model and based on the feedback from the research partner on the implementation results and examples discussed in Section (6.4), the FoR was reframed with the noticed cues, theorised hypothesis and the developed NRE model as a solution for the requirement listed in the initial FoR for this iteration, see Figure (6.5).

Figure 6.5: Reframed FoR in the 3rd iteration

Reframed Frame of Reference (FoR)

Cues

-Investigating the relationship between the bank and its customers is related to the research fields that study Customer Equity (CE) and relationship Equity (RE), since they investigate what makes a client loyal to a brand or a business

-There are limitations imposed by the transactional data provided by the research partner that prevent investigating all the factors affecting the CE and RE

Developed Hypotheses\Models

-Monitoring customers' income and spend inside and outside the bank network over time can reflect the relationship between the bank and its customers

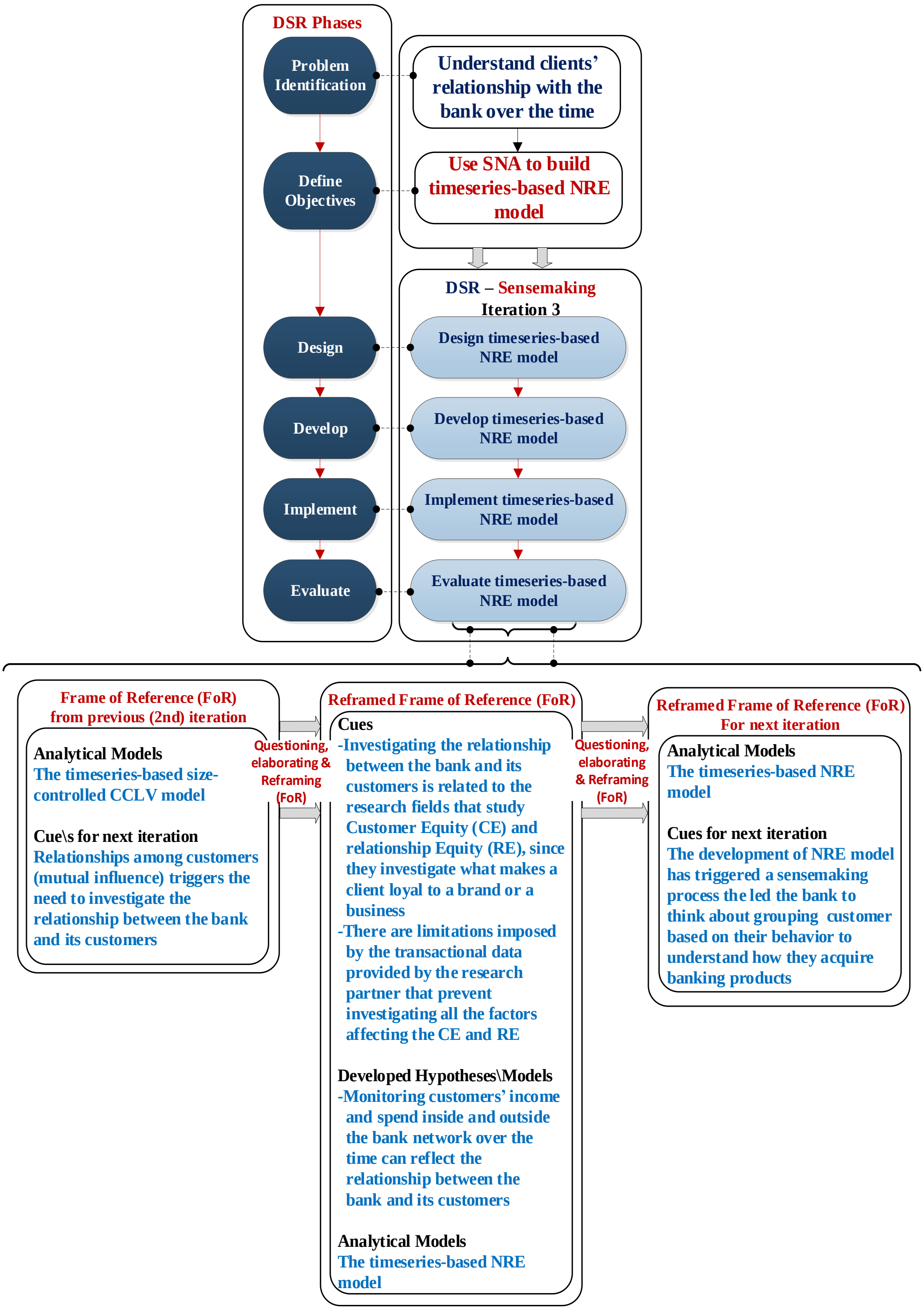
Analytical Models

The timeseries-based NRE model

As shown by the feedback from the analysts at BankCo, the NRE model has proved its ability to spot interesting events in the relationship between BankCo and its customers; facilitating the discovery of potential customer loss as well as the possibility to offer new products to the customers of BankCo depending on the type of the event and the factors leading to such an event. The NRE model provides a proxy to understand the customers' transactional behaviour inside and outside BankCo's network. The requirement emerged during the evaluation with the analysts from BankCo, which is about developing firm personas to understand their behaviour related to acquiring banking products, has led to reframing the FoR again to include the new requirement as a cue for the next iteration explained in the following chapter.

Figure (6.6) illustrates an overall view of the third iteration presented in this chapter. As a result of another sensemaking cycle (noticing, interpretation and action), the Frame of Reference (FoR) is updated (reframed) with new cues, and hypotheses (analytical model) that enable understanding of how a firm's relationship with BankCo changes over time.

Figure 6.6: Overall view of the third iteration



6.6 Summary

This chapter has presented the third iteration of this work and demonstrated how sensemaking processes happen during this iteration. The model proposed in this chapter is based on the concept of loyalty, which is the base for Relationship Equity (RE). The transactional data under investigation, however, limits the factors that can be considered for calculating RE score, thus, a timeseries network-based relationship equity model is developed. The evaluation of the timeseries-based Network Relationship Equity (NRE) model has shown its possible ability to support the analysts at BankCo to explore and spot interesting patterns in customers' income and spend, and consequently, act upon the threats or opportunities discovered by the assist of the NRE model.

Chapter 7: Persona-Products Purchasing Behaviour Model (4th Iteration)

7.1 Overview

This chapter presents the fourth iteration of this study that addresses the last objectives of this research which is about 1) profiling customers based on their features to create personas and 2) understanding how different personas acquire banking products. This iteration investigates how Machine Learning (ML) techniques can help to meet the analytical requirements. The aim is to develop an analytical model that can assist BankCo to understand how different types of customer acquire banking products, and ultimately, assist BankCo as a product recommendation system.

This chapter is structured as follows: Section 7.2 describes the design phase of this iteration, which explains the design phase of the persona-product purchasing frequency model. Additionally, this section discusses clustering algorithm as the machine learning technique to be used to profile the customers of BankCo. Section 7.3 explains the development process of this DSR iteration, which is about performing the clustering task twice, for inbound transactions and for outbound transactions. Also, it explains the development of product purchasing frequency analysis. Section 7.4 presents the implementation of the developed model. Section 7.5 describes the technical evaluation of the developed model as well as the evaluation of the sensemaking framework. Finally, Section 7.5 provides an overall summary of the chapter.

7.2 Iteration 4: Designing the Persona-Product Purchasing Frequency Model

The design phase in this iteration focuses on studying the analytical requirements requested by the research partner and set in the Frame of Reference (FoR) resulting from the previous iteration, see Figure (6.6). The objective of this iteration is to develop a model that helps in profiling the customers of BankCo based on their firmographic and network features (e.g., sum of the income, risk, size and degree centrality among others) in order to understand how different groups of customers acquire banking products. Thus, based on the data sets provided by BankCo (transactional, firmographic, and products tables) described in Chapter 3, it can be noticed that two types of analyses are required to address the objective of the iteration.

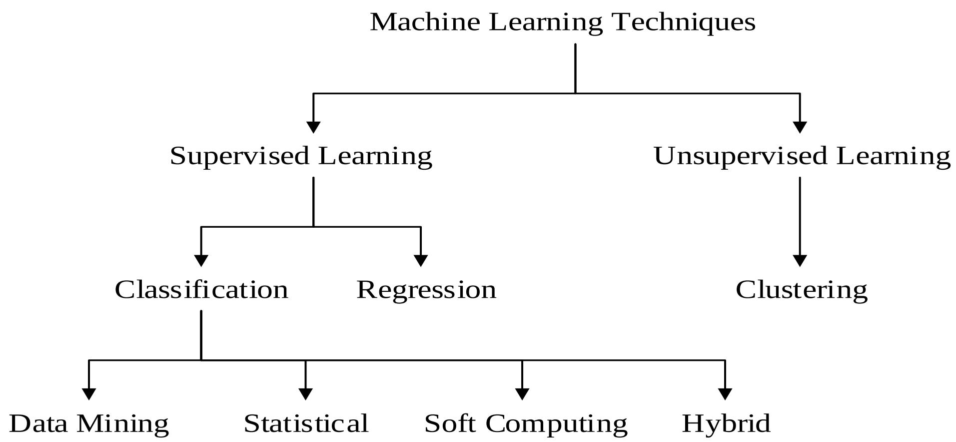
First, Machine Learning (ML) algorithms for cluster analysis can be utilised to group the customers into clusters with distinctive features. It can be noticed, however, that clients of BankCo show different characteristics that reflect their behaviours when they receive income from other customers in BankCo's network compared to when they pay suppliers. Consequently, cluster analysis needs to be performed twice: once for inbound features that reflect clients' income from the network and second for outbound features that reflect customers spending money in the network.

The second analysis is to discover information about how banking products are purchased such as the frequency and the interval of purchasing the product as well as the products that are purchased together. Then, combining these two analyses can uncover valuable information about how different types of customer acquire banking products.

7.2.1 Machine Learning (ML) – Cluster Analysis

Machine Learning (ML) is the branch of computer science that is concerned with developing algorithms to learn from data to predict outcomes, and they can be categorised into supervised algorithm for classification and regression tasks, and unsupervised algorithms for data clustering, see Figure (7.1) (Kolkur and Kalbande, 2016; Crisci, Ghattas and Perera, 2012). Supervised algorithms include linear and logistic regression, Support Vector Machine (SVM), decision trees and Artificial Neural Networks (ANNs) (Guzella and Caminhas, 2009). Training data sets used with supervised learning algorithms must be labelled in order for the algorithm to build and train a model based on the independent variables. This model, then, can be applied on the unlabelled testing data set to predict the values in target variables with the best possible accuracy (Crisci, Ghattas and Perera, 2012; Guzella and Caminhas, 2009).

Figure 7.1: Classification of Machine Learning Techniques (adapted from (Kulkor, 2016)**)**



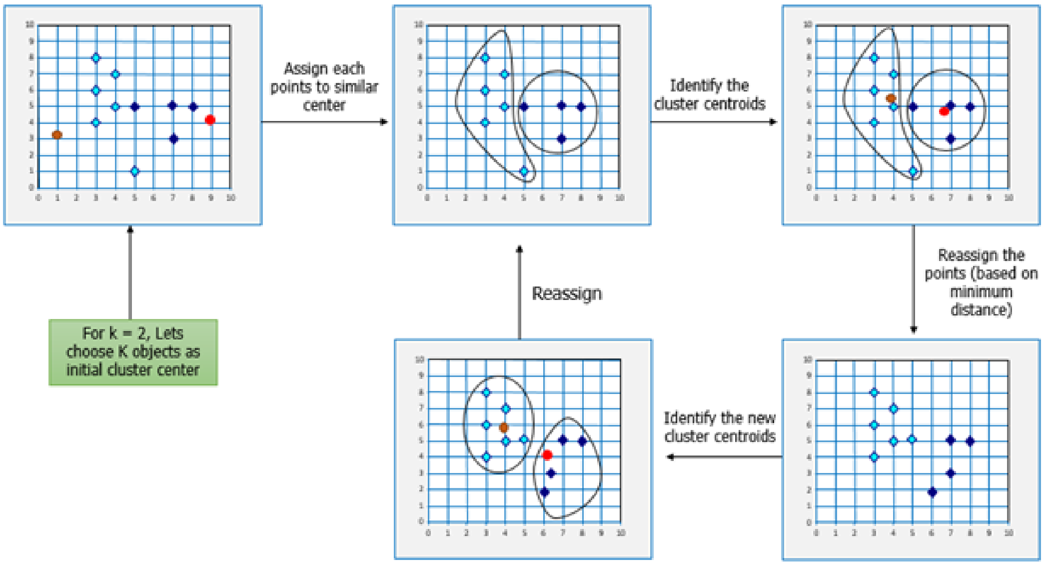
Data sets used in unsupervised learning algorithms, on the other hand, do not contain a target variable, and they are used to cluster data points into groups (cluster) based on the features. Cluster analysis is concerned with finding a structure in unlabelled data points, such as the clients in the case of BankCo's data sets. The ultimate goal from cluster analysis is to group the data into different clusters, in which the features of the data points within the same cluster are more similar compared to the data points in other clusters (Lanjewar and Yadav, 2013; Zadeh, Faraahi and Mastali, 2011).

Clustering algorithms are categorised into hierarchical, partitioning and density-based methods. Hierarchical clustering methods create a tree-like hierarchy of clusters, where data points are assigned to the branches of the tree (Shah and Jivani, 2013). Partitioning clustering methods organise data points into K clusters by randomly assigning centres for the clusters and then going through cycles of assigning the rest of the data points to the nearest centres, recalculating the new centres for the (K) clusters and re-assigning data points to the new centres to improve clustering results and until the centres no longer change (Bhat, 2014). Finally, unlike the hierarchical and partitioning clustering methods where the algorithms group data points in spherical-like clusters, density-based methods are able to create clusters with any shape (Shah and Jivani, 2013). The major disadvantage of the hierarchical and density-based clustering methods is their inability to handle large data sets (Shah and Jivani, 2013). In addition, Density-based algorithms such as DBSCAN do not perform well on data sets with varying densities, such as the clients' data set. Thus, the focus here will be on the partitioning clustering methods.

Hierarchical methods suffer from an inability to handle large data sets (Shah and Jivani 2013); and (b) Density-based algorithms do not perform well on data sets with varying densities, such as the data set yielded from the survey

Figure (7.2) illustrates the clustering process for the K-Means algorithm as one of the most famous partitioning algorithms. The K-Means clustering uses the mean/centroids to represent the clusters (Bhat, 2014). K-Means algorithm, however, is sensitive to normalization and scaling functions and it is very sensitive to noise or outliers such as data points with extreme values because it tries to minimise the square distance between the data points and the centroids of the clusters, hence, the clustering results will be impaired by the uneven distribution of the data points.

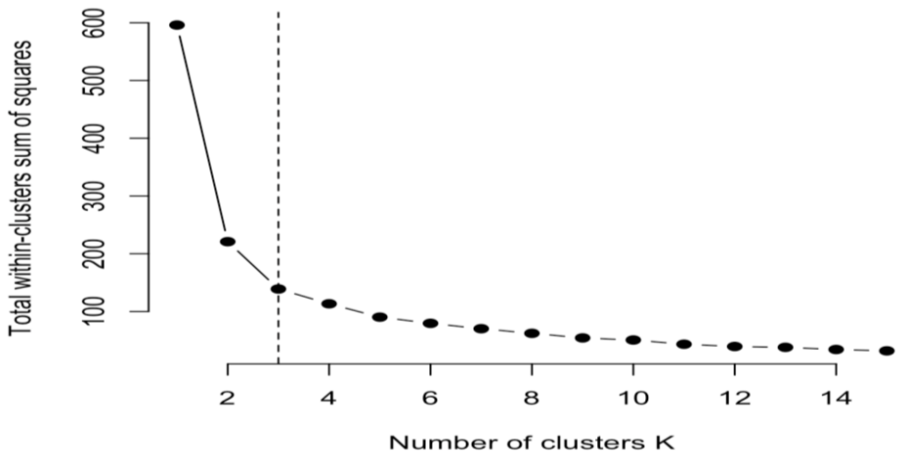
Figure 7.2: Partitioning clustering (K-Means algorithm)



Unlike the K-Means algorithm, the Partitioning Around Medoids (PAM) algorithm performs data clustering using K-Medoids to represent the cluster instead of K-Means (Li, Wang and He, 2017). A medoid can be defined as the centre point within the cluster whose average dissimilarity to all other data points in the cluster is minimal. Thus, the medoid is the most centrally-located point in the cluster. The PAM algorithm follows the same procedure as in the K-Means algorithm by randomly assigning initial medoids to represent the K clusters. Then, it goes through iterations of assigning the rest of the data points to the K clusters, selecting new data points as medoids and re-assigning the data points to the new medoids until having medoids with stabilised locations. The PAM algorithm aims to minimise the absolute distance between the data points and the medoids. Thus, it is more robust to noise and outliers compared to K-Means (Bhat, 2014).

Nevertheless, partitioning methods require the number of clusters (K) to be determined before running the clustering algorithms. Thus, in order to determine the optimal number of clusters K, the elbow method can be used (Othata and Pantaragphong, 2017; Kodinariya and Makwana, 2013). The idea here is to run the cluster analysis several times (starting from K=2 to K=15), and to calculate the cost of training the model, namely total Within-cluster Sum of Square (WSS). Then, by plotting the curve for WSS values against the number of clusters (K) as displayed in the sample curve in Figure (7.3), it can be noticed that the WSS curve goes down rapidly until reaching an elbow point where K=3, and then it starts to change very slowly. The bend point represents the approximate optimal number of clusters.

Figure 7.3: Sample curve



However, it is possible that sometimes, the elbow point is not clear on the curve. In such situations, the average silhouette width (Rousseeuw, 1987) can be used to determine the best option, noting that a higher average silhouette width indicates better clustering performance. The silhouette coefficient is calculated for each data point in each cluster. For example:

1- For each data point (X), the algorithm, first, calculates the average distance “DistA” between (X) and all data points in the same cluster (a measure of cohesion).
2- Then, the algorithm calculates the distance “DistB” between (X) and data points in the closest clusters (a measure of separation).
3- Then silhouette coefficient for (X) is calculated as the difference between “DistB” and “DistA” divided by the greater of the two distances.
4- Then, the overall average silhouette width is calculated by averaging the silhouette coefficient for all data points

7.3 Iteration 4: Developing the Persona-Product Purchasing Frequency Model

The development phase of this DSR iteration focuses on understanding the cues generated from the previous iteration and developing the analytical solutions that can address the analytical requirements. Thus, as discussed in the design phase, cluster analysis needs to be performed twice, because customers tend to have different financial and network features that reflect their income and spend in BankCo's network. Additionally, product purchasing frequency analysis needs to be performed to discover the patterns that reflect banking product purchase.

7.3.1 Cluster Analysis – Inbound Variables

The first clustering task considered financial and network variables that represent the inbound (income) transactions. These variables are listed in Table (7.1). The values of these variables are normalised in order to scale the data and improve convergence of the clustering algorithm.

Table 7.1: Variables for inbound clustering

Variable	Type	Variable Definition
Pty_id	Client Identifier	Unique identifier for the client
sum_income	Inbound financial variable	The total amount of money that the company receives from other companies in the network.
risk	Financial variable	Firm risk
indegree	Inbound Network Variable	A count of the number of ties directed to a company
percent_high_risk	Financial variable	% of high-risk companies in the 1st order neighbourhood
localrisk	Financial variable	First neighbour network risk

sicrisk	Financial variable	Sector risk
size	Financial variable	Firm size
Casht2	Financial variable	Total cash/total assets
deprec_yt	Financial variable	Total depreciation/fixed assets
Freq	Inbound Network Variable	Average number of days between transactions
to_lowfreqm	Inbound Network Variable	% of occasional customers who pay this customer
to_highfreqm	Inbound Network Variable	% of regular customers who pay this customer
to_lowquantitym	Inbound Network Variable	% of low income from inbound relations
to_highquantitym	Inbound Network Variable	% of high income from inbound relations

The original number of firms considered for clustering is 139,491 and because of data quality issues in the data provided by BankCo, some of the variables have missing data. Thus, after omitting the observations with missing values in any of the variables in the above table, the number of firms becomes 44,667. Next, feature selection using the Boruta package in R is utilised to specify the importance of the features in Table (7.1). This package uses the random forest classifier to examine the importance of the variables based on the mean decrease in accuracy measure, which evaluates how much the model fit decreases when a variable is dropped from the data set (Kursa and Rudnicki, 2010).

Figure (7.4) illustrate the outcome of running Boruta package in R on the data set after omitting missing values. Figure (7.4) shows that all variables are confirmed important, and it shows how the Boruta package ranks the variables based on their importance to the data set starting from “sum_income” as the most important one. The blue plot boxes represent the shadow variable that Boruta adds to the original data set to calculate variables’ importance.

Figure 7.4: Features importance using Boruta package in R (inbound variables)

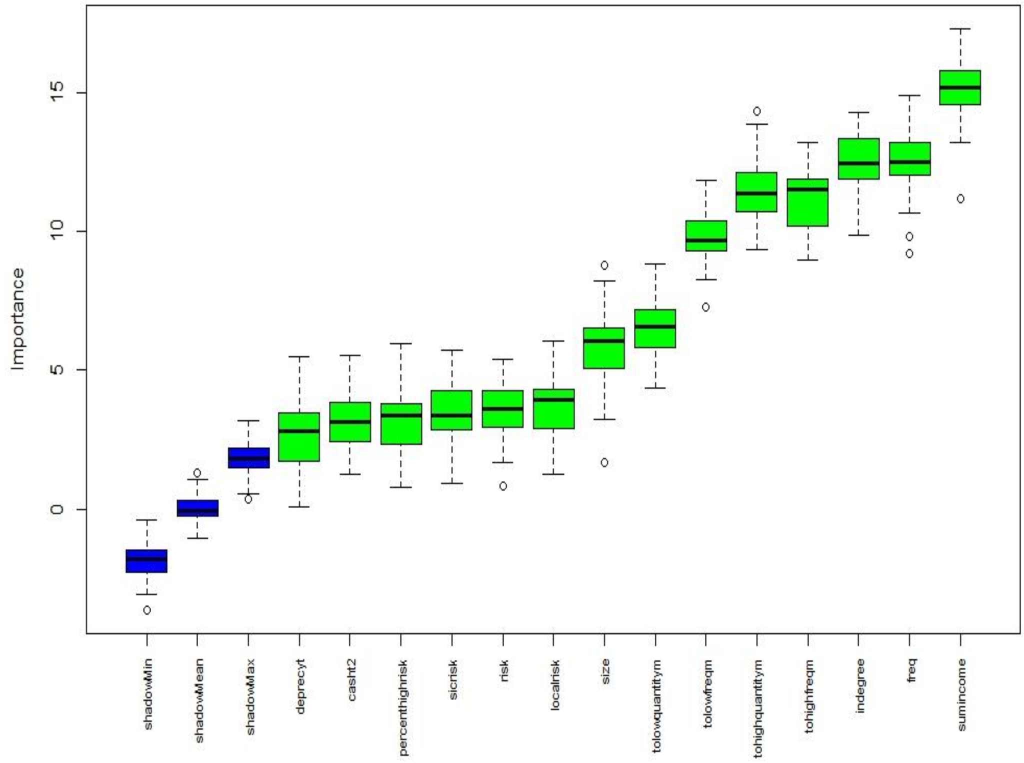


Figure (7.5) represents the WSS curve for the inbound variables, which does not provide a clear indication (bend) on the optimal number of clusters. Thus, several clustering trials (4, 5, 6, 7, 8, and 9) were performed and the average silhouette width is calculated (0.09, 0.08, 0.09, 0.09, **0.1**, and 0.09) for each clustering trial to determine the best option, see Figure (7.6). The outcome indicates that 8 clusters reflect the best clustering performance.

Figure 7.5: WSS curve (inbound variables)

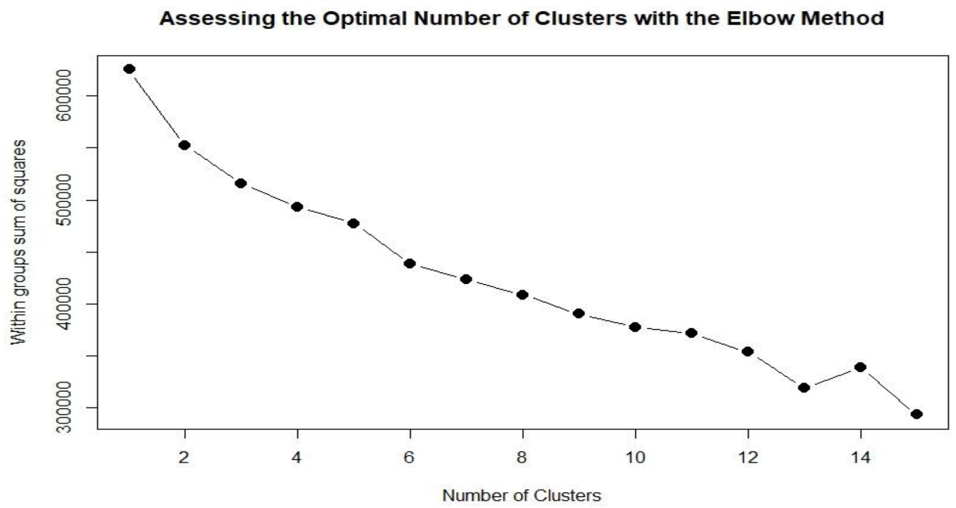
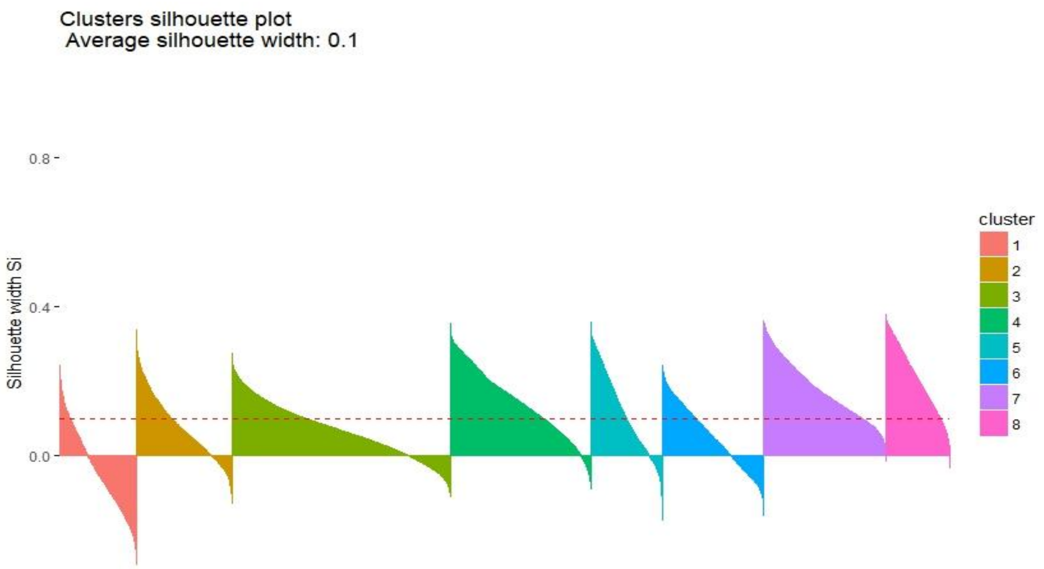


Figure 7.6: Average silhouette width (inbound variables)



7.3.2 Cluster Analysis – Outbound Variables

The second clustering task considered financial and network variables that represent the outbound (spend) transactions. These variables are listed in Table (7.2). Similar to the inbound variables, the values of these variables are normalised in order to scale the data and improve the convergence of the clustering algorithm.

Table 7.2: Variables for outbound clustering

Variable	Type	Variable Definition
Pty_id	Identifier	Identifier
sum_spend	Outbound financial variable	The total amount of money that the company spends on other companies in the network.
risk	Financial variable	Firm risk
outdegree	Outbound Network Variable	A count of the number of ties directed from a company
percent_high_risk	Financial variable	% of high-risk companies in the 1st order neighbourhood
localrisk	Financial variable	First neighbour network risk
sicrisk	Financial variable	Sector risk
Size	Financial variable	Firm size
Casht2	Financial variable	Total cash/total assets
deprec_yt	Financial variable	Total depreciation/fixed assets
Freq	Outbound Network Variable	Average number of days between transactions
from_lowfreqm	Outbound Network Variable	% of occasional customers who are getting paid by this customer
from_highfreqm	Outbound Network Variable	% of regular customers who are getting paid by this customer
from_lowquantitym	Outbound Network Variable	% of low spend in outbound relations
from_highquantitym	Outbound Network Variable	% of high spend in outbound relations

After omitting the observation with missing values in any of the variables in the above table, the number of firms becomes 40,482. Next, feature selection using the Boruta package in R is utilised to specify the importance of the features in Table (7.2). Figure (7.7) illustrates the outcome of running the Boruta package in R on the data set after omitting missing values. Figure (7.7) shows that all variables are ranked and confirmed important, except for "**deprec_yt**", "**percent_high_risk**", and "**Freq**", which were tentative and represented by the yellow plot boxes. This means that Boruta was not able to make a decision on these variables using the default number of random forest runs. Thus, after additional processing with the Boruta package, these variables are confirmed important but with lower importance compared to the other variables.

Figure 7.7: Features importance using Boruta package in R (outbound variables)

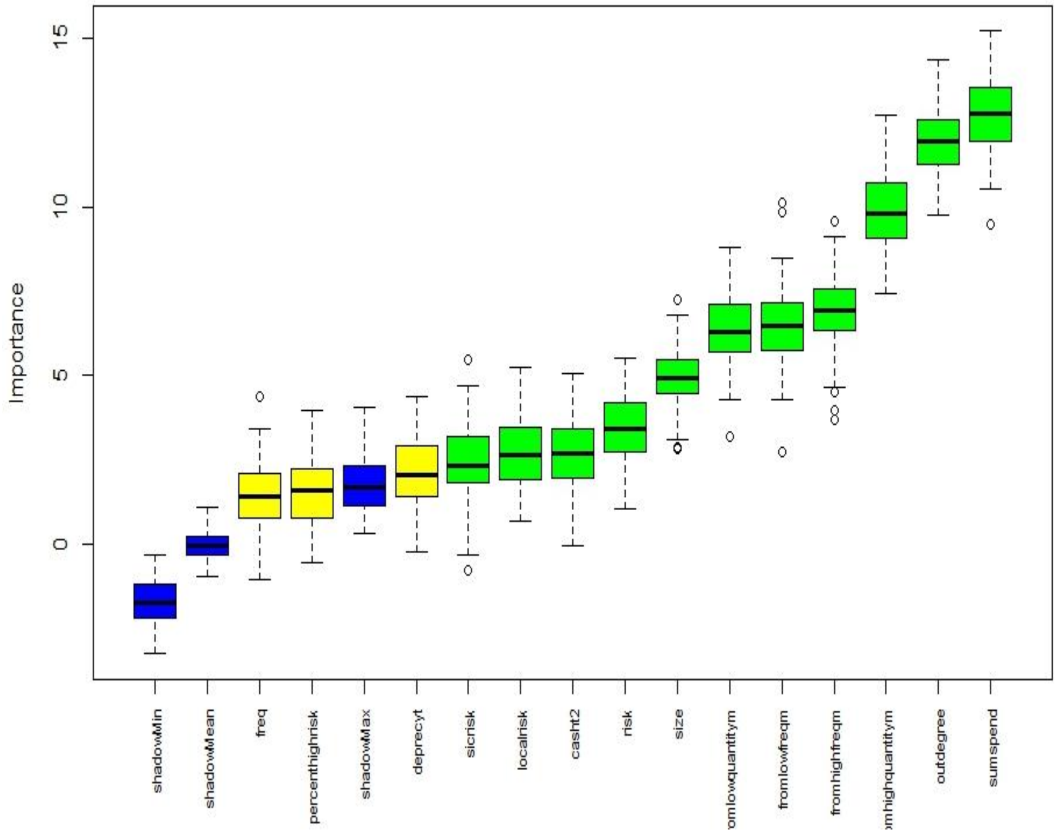


Figure (7.8) represents the WSS curve for the outbound variables, which does not provide a clear indication (bend) on the optimal number of clusters. Thus, several clustering trials (7, 8, 9, and 10) were performed and the average silhouette width is calculated (0.07, **0.08**, 0.07, and 0.07) for each clustering trial to determine the best option, see Figure (7.8). The outcome indicates that 8 clusters reflect the best clustering performance.

Figure 7.8: WSS curve (outbound variables)

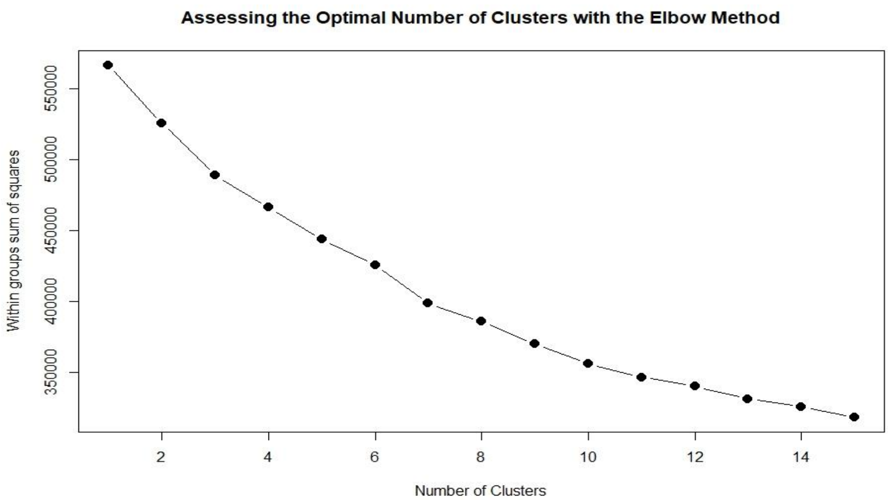
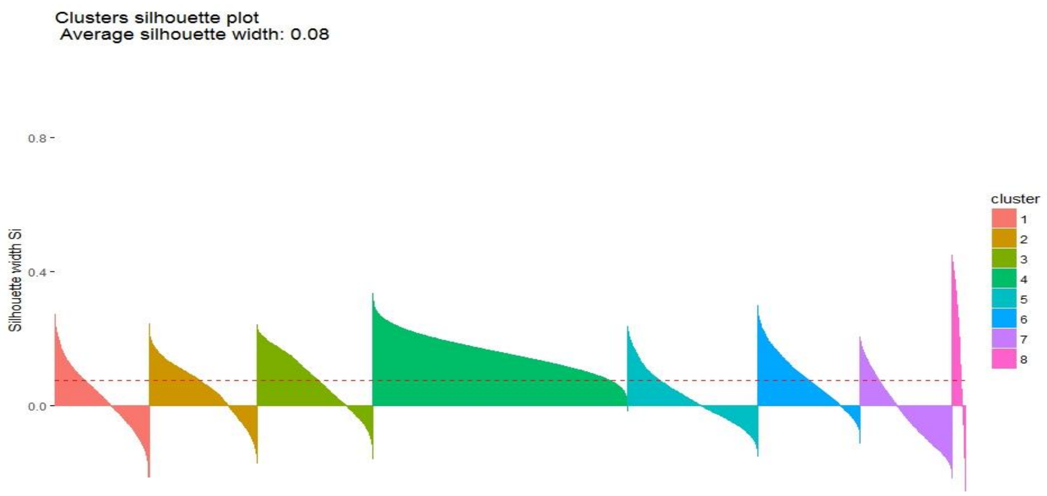


Figure 7.9: Average silhouette width (outbound variables)



7.3.3 Products Purchasing Frequency Analysis

This section explains the second analysis task required to meet the analytical objective defined in the design phase. Thus, the focus here is to analyse frequencies and intervals in which banking products purchased individually, together, or sequentially by each customer. From BankCo's perspective, the focus in this analysis is on the following products: Asset finance (AF), Invoice Finance (IF), Loans, Rates, Foreign Exchange (FX), Trade, Cash management and payments (CM&P) and Term Deposits. Table (7.3) shows the structure of the intended analysis.

Table 7.3: Structure of the products purchasing frequency analysis

Column	Definition
Customer ID	Unique customer identification
Products list	Products purchased by the customer. This could be one product, several products purchased at the same time, or several products purchased at different times.
Interval	Number of days between products purchase when several products are purchased at different times.

Product purchase frequency analysis results in sequences that reflect how products are purchased. Table (7.4) shows sample sequences that are produced by the analysis. In this table, any sequence could represent a single product or a set of products that are purchased together or separately, as well as the order of purchasing these products.

Table 7.4: Sample purchase frequency sequences

Customer ID	Interval list	Product list	Explanation
1002430114151590		AF	This party bought only one AF product
1037432172110480	76	AF- AF	This party bought two AF products with 76 days gap between the two purchases
1002430141886360	27; 78	AF- AF- AF	This party bought three AF products with 27 days gap between the first two and 78 days gap between the second two purchases
7426059650286240	191; 545	CM&P- AF- AF	This party bought one CM&P product and after 191 days they bought AF product and after 545 days they bought another AF product
7426059689264050	8; 28; 29; 35; 42; 214; 344	FX- CM&P- AF- AF- AF- AF- AF- AF	This party bought FX, CM&P, AF, AF, AF, AF, AF, AF with 8, 28, 29, 35, 42, 214, 344 days gap between each two consecutive products

The next step is to match the clustering results from inbound clustering and outbound clustering with the outcome of the product purchasing frequency analysis.

7.4 Iteration 4: Implementing the Persona-Product Purchasing Frequency Model

The implementation phase focuses on implementing the clustering and product purchasing frequency analyses explained above using the data sets provided by BankCo. The next two subsections explain how clustering results (inbound and outbound) are matched with product purchasing frequency analysis.

7.4.1 Matching (Inbound) Clustering and Products Purchase Frequency Analysis

Cluster analysis allows identifying homogeneous groups of customers that share similar features and have similar needs for products but are distinctively different from other clusters. The results of the inbound clustering are detailed in this subsection. Table (7.5) lists the number of firms within each of the 8 clusters, and Figure (7.10) illustrates the cluster plot for the 8 clusters generated by the PAM algorithm.

Table 7.5: Number of firms in each cluster (inbound clustering)

Cluster	Frequency (number of firms in the cluster)
Cluster 1	3,854
Cluster 2	4,800
Cluster 3	10,950
Cluster 4	7,047
Cluster 5	3,615
Cluster 6	5,072
Cluster 7	6,110
Cluster 8	3,219
Total	44,667

Figure 7.10: Clusters plot (inbound clustering)

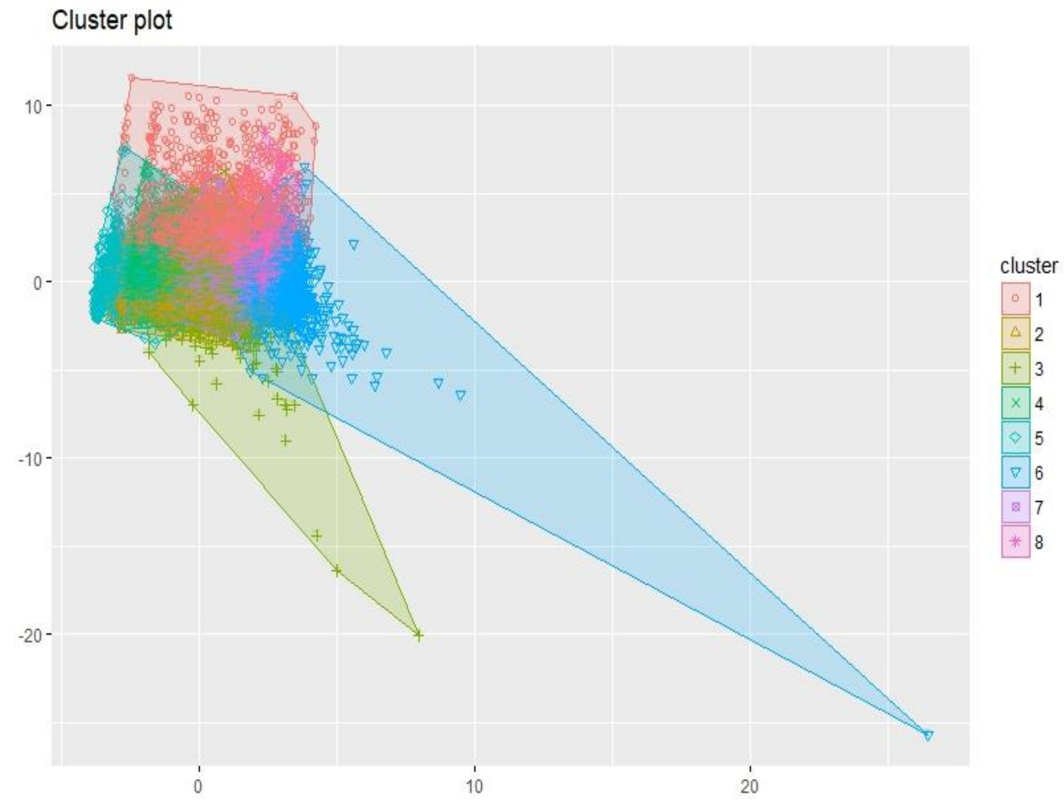


Figure (7.11) provides a comparison between the mean and standard deviation for the most important features that show the most variance in their values among the 8 clusters and thus provide reasonable 'descriptors' for the cluster. Also, table (7.6) provides statistical comparison (min, max, mean, and standard deviation) among the 8 clusters for each of the features displayed in Figure (7.11)

Figure 7.11: Features comparison among the 8 clusters (inbound clustering)

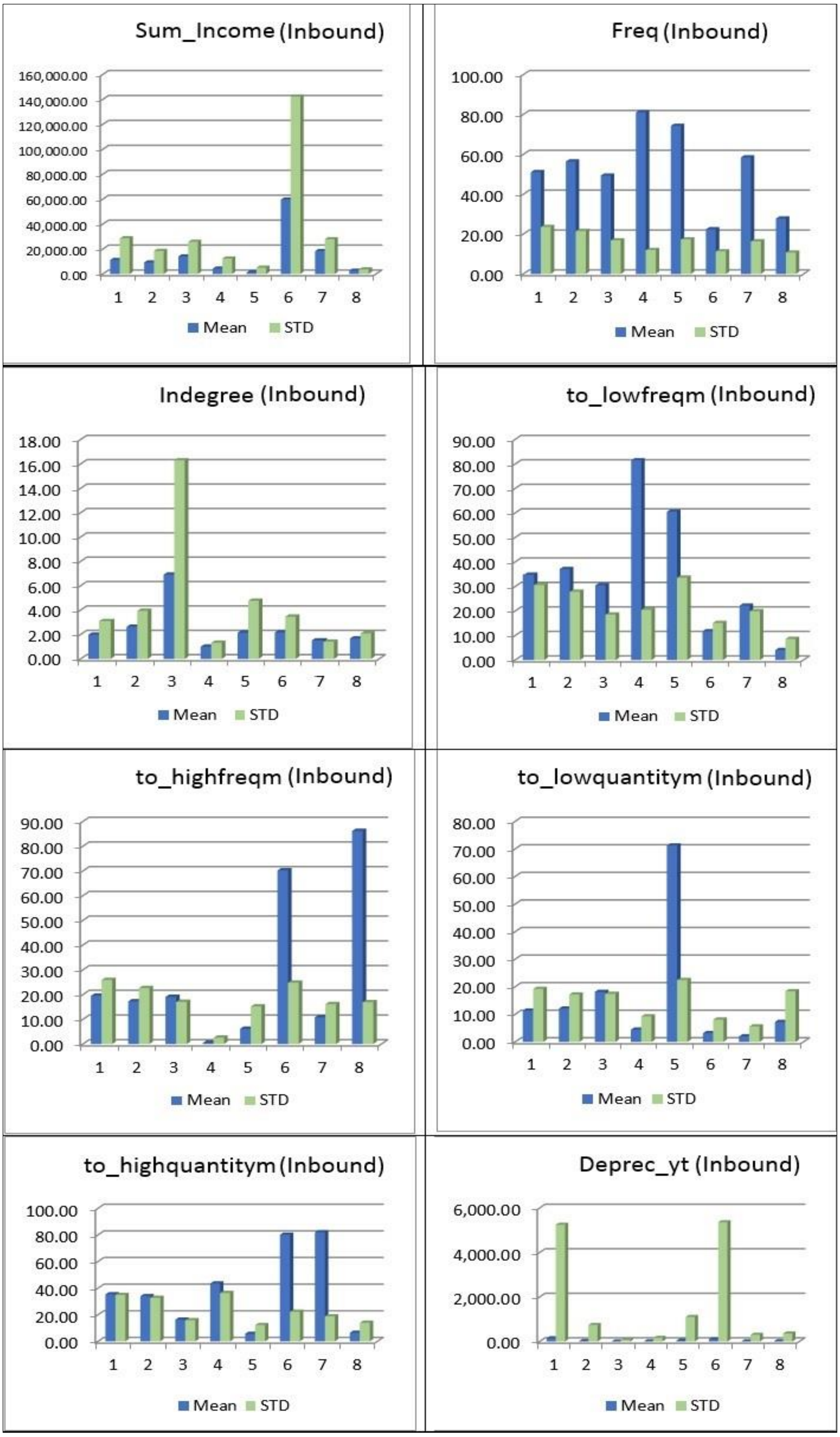


Table 7.6: Statistical comparison for the 8 clusters (inbound clustering)

	sum_income							
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Max	415,642.3	330,555.6	538,436.4	383,333.3	155,596.6	6,526,503.0	381,932.3	50,352.1
STD.Dev	28,589.16	18,546.47	25,786.31	12,407.28	5,044.61	142,267.18	27,806.71	3,608.68
Mean	11,204.36	9,291.63	14,096.75	4,308.54	1,264.95	59,672.26	18,396.39	2,587.42
Min	0.00	0.00	0.00	0.00	0.00	1,108.42	0.00	5.00
	Freq							
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Max	90.00	90.00	90.00	90.00	90.00	90.00	90.00	69.47
STD.Dev	23.57	21.50	16.84	12.00	17.30	11.33	16.35	10.72
Mean	51.06	56.43	49.34	81.07	74.28	22.41	58.43	27.78
Min	1.39	1.85	1.46	15.00	6.38	0.53	4.98	1.97
	Indegree							
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Max	62.35	77.82	701.03	23.03	114.50	109.53	15.82	43.14
STD.Dev	3.10	3.94	16.29	1.32	4.77	3.47	1.41	2.09
Mean	1.99	2.64	6.91	1.00	2.16	2.17	1.51	1.67
Min	0.03	0.03	0.03	0.03	0.03	0.28	0.07	0.28
	to_lowfreqm							
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Max	100.00	100.00	100.00	100.00	100.00	100.00	100.00	50.00
STD.Dev	30.59	27.86	18.56	20.45	33.60	15.11	19.91	8.59
Mean	34.79	37.11	30.49	81.46	60.52	11.73	22.16	4.04
Min	0.00	0.00	0.00	21.42	0.00	0.00	0.00	0.00
	to_highfreqm							
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Max	100.00	100.00	88.81	33.33	100.00	100.00	81.25	100.00
STD.Dev	25.89	22.60	17.04	2.53	15.17	24.79	16.09	16.89
Mean	19.52	17.23	19.04	0.48	6.13	70.24	10.78	86.18
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	31.25
	to_lowquantitym							
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Max	100.00	100.00	95.49	50.00	100.00	66.66	50.00	100.00
STD.Dev	19.21	17.16	17.41	9.25	22.42	8.06	5.60	18.35
Mean	11.35	12.08	18.12	4.43	71.20	3.17	1.97	7.22
Min	0.00	0.00	0.00	0.00	18.75	0.00	0.00	0.00

Cluster	to_highquantitym							
	1	2	3	4	5	6	7	8
Max	100.00	100.00	77.77	100.00	75.00	100.00	100.00	100.00
STD.Dev	34.90	32.69	15.88	36.36	12.20	22.25	18.78	13.97
Mean	35.40	33.96	16.27	43.49	5.62	80.06	81.98	6.48
Min	0.00	0.00	0.00	0.00	0.00	5.43	26.78	0.00
Cluster	**Deprec_yt**							
	1	2	3	4	5	6	7	8
Max	308,632.0	48,396.9	5,936.0	8,304.9	60,382.6	370,354.3	21,816.8	19,218.0
STD.Dev	5,248.21	740.58	76.11	161.97	1,104.46	5,360.84	303.19	359.41
Mean	146.84	18.80	2.31	4.88	35.97	97.05	7.36	11.70
Min	0.00	0.00	-0.48	0.00	0.00	0.00	-1.02	0.00

Matching clustering results with the results of product purchasing frequency analysis using Customer ID as the common field between both analyses, uncovers interesting patterns on how customers who belong to certain cluster purchase the products of interest for this research. Table (7.7) summarises the frequency and percentages of products purchased per cluster.

Table 7.7: Summary of the products purchase frequencies and percentages per cluster (inbound clustering)

	AF		IF		Loans		Rates		FX		Trade		CM&P		T&D	
Cluster	Cnt	%	Cnt	%	Cnt	%	Cnt	%	Cnt	%	Cnt	%	Cnt	%	Cnt	%
1	99	12.8	14	1.8	334	43.3	0	0	320	41.5	2	0.3	2	0.3	0	0
2	29	5.31	0	0	64	11.7	0	0	449	82.2	2	0.4	2	0.4	0	0
3	695	29.9	12	0.5	451	19.4	4	0.2	1142	49.1	18	0.8	4	0.2	0	0
4	240	18.1	11	0.8	340	25.7	0	0	718	54.3	10	0.8	4	0.3	0	0
5	96	18.3	4	0.8	131	25	0	0	294	56	0	0	0	0	0	0
6	177	23.1	2	0.3	267	34.9	0	0	314	41	4	0.5	2	0.3	0	0
7	211	21.9	2	0.2	291	30.2	0	0	448	46.4	9	0.9	4	0.4	0	0
8	27	12.2	1	0.5	167	75.6	1	0.5	25	11.3	0	0	0	0	0	0

Table (7.7) shows that lending products are on high demand in clusters 1, 4, 6, 7 and 8 with percentages (43.3%, 25.7%, 34.9%, 30.2% and 75.6%) respectively. Also, it shows that clusters 6 and 7 have more tendency to acquire Asset Finance (AF) products. In general, it can be noticed that loans, AF, IF and FX are the most acquired products by the customers.

Examining cluster 6, for example, shows that customers with higher income (**sum_income**) compared to clients in other clusters have more regular customers (**to_highfreqm**) and high income from inbound relations (**to_highquantitym**). Consequently, high demand for AF products in cluster 6 implies that businesses in this cluster are expanding and they need AF products to support their expansion plans. Thus, after assessing the risk factor, BankCo can make a decision on recommending or providing the AF products to the customers in this cluster.

7.4.2 Matching (Outbound) Clustering and Products Purchase Frequency Analysis

The results of the outbound clustering are detailed in this subsection. Table (7.8) lists the number of firms within each of the 8 clusters, and Figure (7.12) illustrates the cluster plot for the 8 clusters generated by the PAM algorithm.

Table 7.8: Number of firms in each cluster (outbound clustering)

Cluster	Frequency (number of firms in the cluster)
Cluster 1	4,183
Cluster 2	4,804
Cluster 3	5,166
Cluster 4	11,314
Cluster 5	5,785
Cluster 6	4,555
Cluster 7	4,084
Cluster 8	591
Total	40,482

Figure 7.12: Clusters plot (outbound clustering)

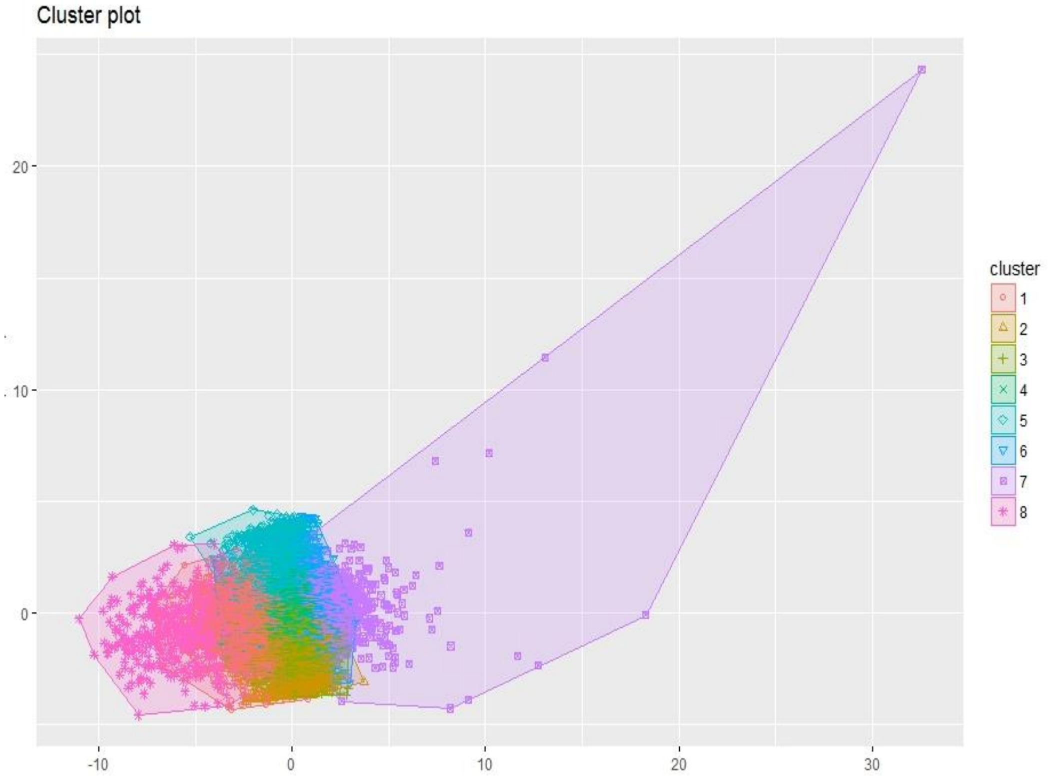


Figure (7.13) provides a comparison between the mean and standard deviation for the most important features that show the most variance in their values among the 8 clusters and thus provide reasonable ‘descriptors’ for the cluster. Also, table (7.9) provides statistical comparison (min, max, mean, and standard deviation) among the 8 clusters for each of the features displayed in Figure (7.13).

Figure 7.13: Features comparison among the 8 clusters (outbound clustering)

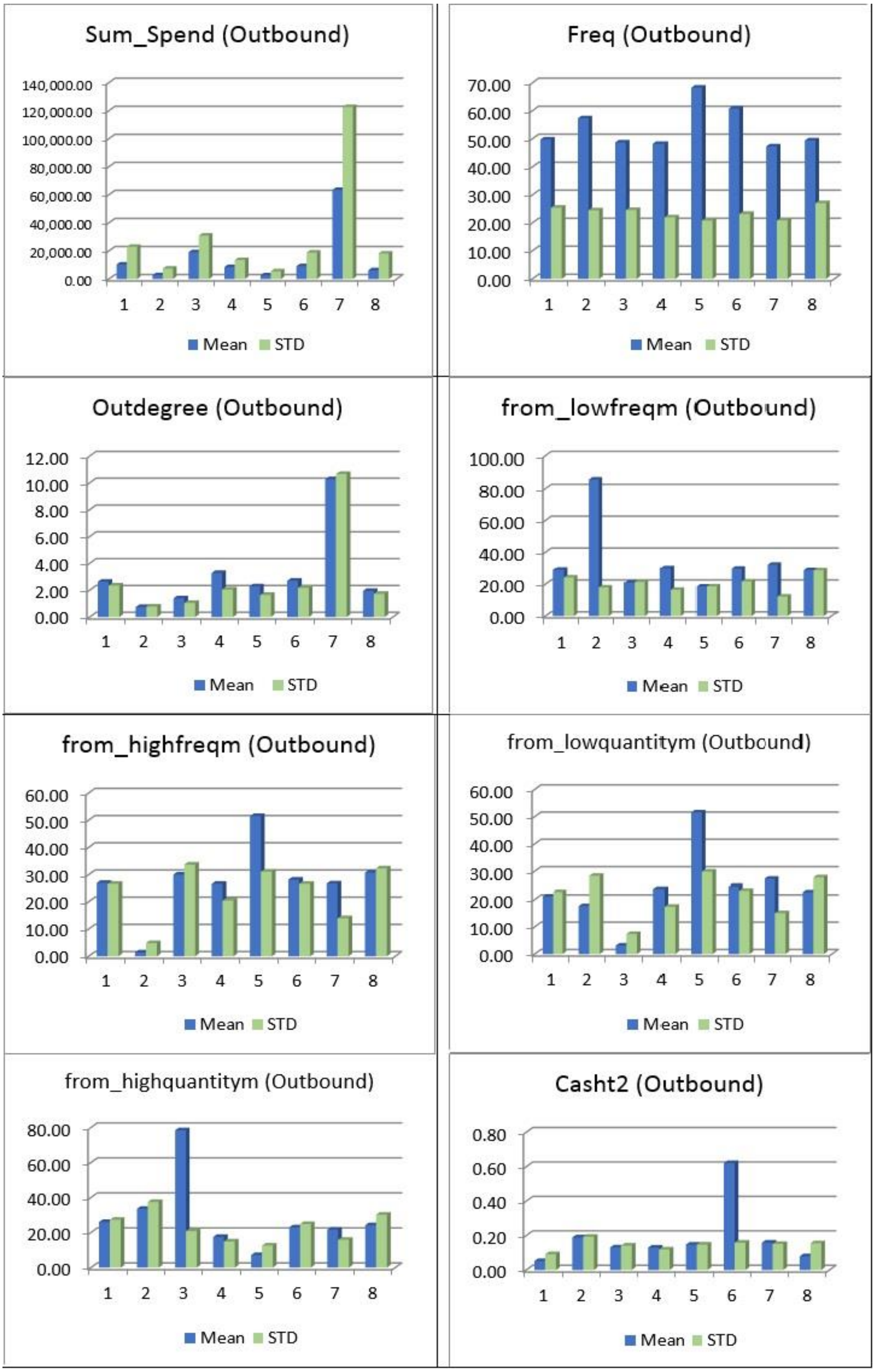


Table 7.9: Statistical comparison of the 8 clusters (outbound clustering)

	sum_spend							
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Max	380,923.4	166,666.7	443,767.8	214,637.8	110,343.8	302,582.9	2,665,771.0	313,456.5
STD.Dev	22,968.90	7,442.29	30,941.32	13,497.79	5,486.17	18,835.45	122,636.44	18,170.79
Mean	10,338.80	2,668.06	19,076.97	8,505.48	2,613.52	9,193.02	63,463.26	6,218.68
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
	Freq							
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Max	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00
STD.Dev	25.43	24.50	24.57	21.99	20.85	23.21	20.87	27.05
Mean	49.75	57.33	48.69	48.19	68.26	60.76	47.29	49.41
Min	0.53	2.14	0.89	0.86	1.65	0.91	0.82	2.13
	Casht2							
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Max	0.84	0.98	0.72	0.58	0.80	0.99	0.97	0.96
STD.Dev	0.09	0.19	0.14	0.12	0.15	0.16	0.15	0.15
Mean	0.05	0.19	0.13	0.13	0.15	0.62	0.16	0.08
Min	0.00	0.00	0.00	0.00	0.00	0.24	0.00	0.00
	to_lowfreqm							
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Max	100.00	100.00	100.00	75.00	100.00	100.00	100.00	100.00
STD.Dev	24.24	18.00	21.47	16.54	18.67	21.80	12.30	28.77
Mean	29.11	85.76	21.05	30.10	18.62	29.82	32.25	28.85
Min	0.00	25.00	0.00	0.00	0.00	0.00	0.00	0.00
	to_highfreqm							
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Max	100.00	40.00	100.00	100.00	100.00	100.00	100.00	100.00
STD.Dev	26.68	4.87	33.73	20.46	31.15	26.68	13.97	32.32
Mean	27.02	1.38	30.06	26.67	51.53	28.17	26.80	30.83
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	to_lowquantitym							
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Max	100.00	100.00	45.35	100.00	100.00	100.00	100.00	100.00
STD.Dev	22.73	28.31	7.38	17.35	29.88	23.16	14.99	27.76
Mean	21.06	17.56	3.14	23.78	51.56	24.64	27.23	22.63
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Cluster	to_highquantitym							
	1	2	3	4	5	6	7	8
Max	100.00	100.00	100.00	73.21	78.84	100.00	100.00	100.00
STD.Dev	27.33	37.49	20.88	14.93	12.62	24.92	15.87	30.23
Mean	26.03	33.60	78.46	17.50	7.19	22.99	21.74	24.23
Min	0.00	0.00	23.43	0.00	0.00	0.00	0.00	0.00
Cluster	**Casht2**							
	1	2	3	4	5	6	7	8
Max	0.84	0.98	0.72	0.58	0.80	0.99	0.97	0.96
STD.Dev	0.09	0.19	0.14	0.12	0.15	0.16	0.15	0.15
Mean	0.05	0.19	0.13	0.13	0.15	0.62	0.16	0.08
Min	0.00	0.00	0.00	0.00	0.00	0.24	0.00	0.00

Similar to the subsection before, matching clustering results with the results of product purchasing frequency analysis using the Customer ID as the common field between both analyses, provides insights on how customers who belong to certain clusters purchase the products under investigation. Table (7.10) summarises the frequency and percentages of products purchase per cluster.

Table 7.10: Summary of the products purchase frequencies and percentages per cluster (outbound clustering)

	AF		IF		Loans		Rates		FX		Trade		CM&P		T&D	
Cluster	Cnt	%	Cnt	%	Cnt	%	Cnt	%	Cnt	%	Cnt	%	Cnt	%	Cnt	%
1	94	10.4	11	1.21	442	48.8	0	0	355	39.2	2	0.2	2	0.2	0	0
2	84	15.5	3	0.55	204	37.7	3	0.6	246	45.5	1	0.2	0	0	0	0
3	113	20.5	3	0.54	232	42.1	1	0.2	200	36.3	0	0	2	0.4	0	0
4	558	28.2	17	0.86	485	24.5	0	0	911	46	9	0.5	1	0.1	0	0
5	149	24.2	7	1.13	226	36.6	1	0.2	234	37.9	0	0	0	0	0	0
6	26	4.69	0	0	61	11	0	0	460	83	3	0.5	4	0.7	0	0
7	528	25.6	1	0.05	201	9.73	0	0	1297	62.8	29	1.4	9	0.4	0	0
8	2	11.8	3	17.7	12	70.6	0	0	0	0	0	0	0	0	0	0

Table (7.10) shows that lending products are in high demand in clusters 1, 2, 3, and 5 with percentages (48.8%, 37.7%, 42.1%, and 36.6%) respectively. Also, it shows that clusters 5 and 7 acquire Asset Finance (AF) products more frequently compared to other clusters.

Examining cluster 5, for example, shows that customers with lower spend in the network (**sum_spend**) compared to customers in other clusters, have fewer occasional customers (**from_lowfreqm**) and pay their suppliers more regularly (**from_highfreqm**), which means that they pay their suppliers in small amounts (**from_lowquantitym**). The customers with these features tend to acquire more loans, AF and IF products. BankCo, thus, can assess the risk for such customers and recommend lending products to them if they represent a low risk for BankCo.

Additionally, Table (7.11) provides a sample comparison between the results of outbound clusters 3 and 4. This table shows how different combinations of products are acquired by the customers. Such comparison can provide a useful tool for the analysts in BankCo to make sense of the differences related to customers' needs for banking products. The same comparison is possible among all clusters.

Table 7.11: Sample comparison between (outbound) clusters 3 and 4

Cluster 3					Cluster 4				
product list	Prod Freq Per Cluster	Pord Seq % Per Cluster	Mean Interval List	Interval list	product list	Prod Freq Per Cluster	Pord Seq % Per Cluster	Mean Interval List	Interval list
AF	26	13.07	0		AF	94	17.28	0	
AF- AF	1	0.5	100	100	AF- AF	2	0.37	0	0
AF- AF- AF	1	0.5	258.5	194; 323	AF- AF- AF	1	0.18	420	108; 732
AF- AF- AF- AF	1	0.5	254.67	72; 345; 347	AF- AF- AF- AF	1	0.18	256.67	0; 0; 770
AF- AF- AF- AF- AF	1	0.5	45.5	14; 30; 58; 80	AF- AF- AF- AF- AF	1	0.18	209.75	25; 73; 182; 559
AF- AF- AF- AF- AF- AF- AF	1	0.5	85	46; 49; 71; 84; 93; 167	AF- AF- AF- AF- AF- AF- AF- AF- AF- AF- AF	1	0.18	49.45	3; 15; 17; 19; 33; 34; 39; 47; 90; 121; 126

AF- AF- AF- AF- AF- AF- AF- AF- AF- AF- AF- AF- AF	1	0.5	73.33	6; 20; 35; 42; 53; 58; 64; 84; 97; 111; 148; 162	AF- AF- AF- AF- AF- AF- AF- AF- AF- AF- AF- AF- AF- AF- AF	1	0.18	54.8	7; 21; 24; 24; 28; 32; 34; 46; 48; 57; 59; 83; 85; 88; 186
CM&P	2	1.01	0		CM&P	1	0.18	0	
FX- FX	1	0.5	55	55	FX- FX	1	0.18	161	161
FX- FX- FX- FX	1	0.5	170	61; 69; 380	FX- FX- FX- FX	1	0.18	38	1; 21; 92
IF	2	1.01	0		IF	6	1.1	0	
Loans	71	35.68	0		IF- AF- AF	1	0.18	304.5	149; 460
Loans- AF	1	0.5	104	104	IF- AF- AF	1	0.18	328	22; 634
Loans- AF	1	0.5	160	160	IF- IF- IF	1	0.18	1.5	0; 3
Loans- AF	1	0.5	257	257	Loans	159	29.23	0	
Loans- AF	1	0.5	29	29	Loans- AF	1	0.18	143	143
Loans- AF- AF- AF	1	0.5	164	28; 140; 324	Loans- AF- AF	1	0.18	366.5	105; 628
Loans- AF- AF- AF	1	0.5	175.67	93; 114; 320	Loans- AF- AF	1	0.18	180.5	107; 254

7.5 Iteration 4: Evaluating the Persona-Product Purchasing Frequency Model

This section focuses on evaluating the persona-product purchasing frequency model developed in this iteration. The model is tested and evaluated in the context of the 30 months transactional data.

7.5.1 Stakeholder Evaluation

The objective of this iteration is to develop a model that helps in profiling the customers of BankCo based on their firmographic and network features in order to uncover different products needs by different groups of customers. Following the process displayed in Figure (3.4), the persona-product purchasing frequency model and the implementation examples were presented and discussed with the analysts from BankCo. During Part 1 of the evaluation process, the use of cluster analysis was explained as a means to group customers, based on their firmographic and network features while differentiating between the buyer and supplier behaviours (i.e., different personas). Also, the foundation of the product purchase frequency analysis was explained as was how the results of both analyses were combined to explore the different product needs of the resulting groups of customers. Next, the implementation examples (discussed in Section 7.4) were demonstrated, highlighting the interesting observations in Tables (7.7) and (7.10) regarding the most-used products in each resulting cluster.

In Part 2 of the evaluation process, the analysts again de-anonymised the data and looked up additional information for some of the firms in the clusters where lending products were in high demand – such as Clusters 1 and 4 in the inbound clustering results (see Table 7.7) and Clusters 1 and 2 in the outbound clustering results (see Table 7.10). While the model shows some promise as a product recommendation system, more data such as risk, credit history and liquidity may be relevant to support the decision-making process regarding recommending banking products to clients. It was concluded, however, that there was enough 'sense' in the model as it stands for Relationship Managers to contact clients where the model indicated a propensity to purchase (i.e., where the majority of customers in a cluster had a product, but the focal company did not). From BankCo's perspectives, the personas-based modelling approach introduced in the iteration allows for:

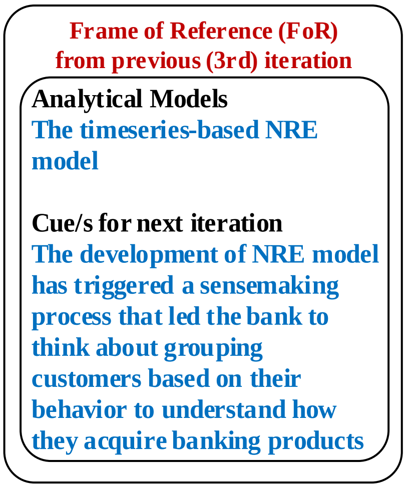
1- Behavioural customer profiling, where behaviour is based on the features included in the cluster analysis (i.e., multidimensional profiling of customers);
2- The discovery of the frequencies and percentages of products purchased within each cluster;
3- Prediction of customers who are likely to need a specific type of product based on the purchasing behaviour within the clusters to which the clients belong;
4- The design of new products that fit customer needs based on their purchasing behaviour;
5- Understanding of clients with multiple product needs.

In Part 3 of the evaluation process, additional requirements were discussed, including working-up the narrative detail of personas, applying the same clustering approach in different contexts and using a different set of variables in order to model the customers in those scenarios. The so called 'stopping conditions' were present at this point, however, on two counts: first, the researcher was reaching the end of his tenure in the PhD process; second, the contract period was at an end with BankCo. At this point, the models developed throughout the four iterations were deemed as satisficing for the purpose of exploring values from the transaction-based network data.

7.5.2 Evaluating the Sensemaking Framework

This subsection demonstrates how sensemaking processes of noticing, interpretation and action illustrated in the proposed sensemaking framework, see Figure (2.4), guided the development of the persona-product purchasing frequency model. The initial Frame of Reference (FoR) for this iteration resulted from the previous one, see Figure (7.14).

Figure 7.14: Initial FoR for 4th iteration



The cue in the initial FoR indicates that the development of NRE model, which provides an approach to understand customers' transactional behaviour inside and outside BankCo has triggered a sensemaking process to understand client behaviour regarding acquiring banking products. Consequently, noticing and interpreting the cues represented in the features (variables) available in the firmographic, products and network-based transactional data sets provided by BankCo, thus, has enabled the researcher to hypothesise that two types of analysis are required to address the objective of the iteration. First, cluster analysis needs to be performed to profile the customers. Second, product purchasing frequency analysis is required to discover the frequencies and intervals that reflect how banking products are purchased. It is hypothesised, also, that matching the results of both analyses can offer valuable insights on how customers' financial and network features affect how banking products are acquired by BankCo's clients.

As an action, and based on the hypotheses described above, the researcher has tried several clustering algorithms, aiming to identify homogeneous clusters of customers. The K-means algorithm was first applied using all of the financial and network features of the firms. As a sensemaking exercise, the results were interpreted as non-satisfactory since so many outliers were noticed, which resulted in having bad clustering results because the K-means algorithm is sensitive to scaling function and extremely sensitive to outliers due to unrealistic centroids of the clusters (Chawla and Gionis, 2013). Thus, in a second trial, the researcher has tried a density-based clustering algorithm called DBSCAN, and the results were interpreted as not acceptable as well since that the resulting clusters were unbalanced; having one big cluster containing 90% of the firms, and other small clusters for the rest of the clients.

During these two trials, the researcher has noticed that the clients of BankCo have different characteristics that reflect their behaviours when they receive income from other customers in BankCo’s network (receive money) compared to when they pay their suppliers (spend money). Consequently, the researcher has hypothesised that cluster analysis needs to be performed twice: once for inbound features that reflect clients’ income from the network and second for outbound features that reflect customers spending money in the network. Thus, as an action process, a third clustering exercise was performed twice using the Partitioning Around Medoids (PAM) clustering algorithm and the resulting clusters were balanced. Then, as discussed during the stakeholder evaluation, when the resulted clusters were matched with the product purchasing frequency analysis, useful insights were noticed by the analysts at BankCo.

To describe the questioning, elaborating and reframing cycle, the initial Frame of Reference (FoR) for this iteration was questioned to investigate the requirement regarding exploring how variances in customers’ financial and network characteristics affect customers’ tendencies for acquiring banking products. The initial FoR, thus, was elaborated by hypothesising that cluster and product purchasing frequency analyses can address the objective of the iteration. Consequently, after developing the personas-product purchasing frequency model and based on the feedback from the research partner on the implementation results and examples discussed in Section (7.4), the FoR was reframed with the noticed cues, theorised hypothesis and the developed product-purchasing frequency model as a solution for the requirements described in the initial FoR for this iteration, see Figure (7.15).

Figure 7.15: Reframed FoR in the 4th iteration

Reframed Frame of Reference (FoR)

Cues

-Studying customers' behavior is related to cluster analysis

-Studying customer tendencyy for acquiring products is related to studying products purchasing frequency analysis

-Companies' features in the data sets provided by the bank

-Clients' characteristics when they receive income is different from when they spend money

Developed Hypotheses\Models

-Two types of analysis are required to address the objective of the iteration:

1-cluster analysis to profile the customers (inbound & outbound), and

2-products purchasing frequency analysis to discover the frequencies and intervals that reflect how banking products are purchased.

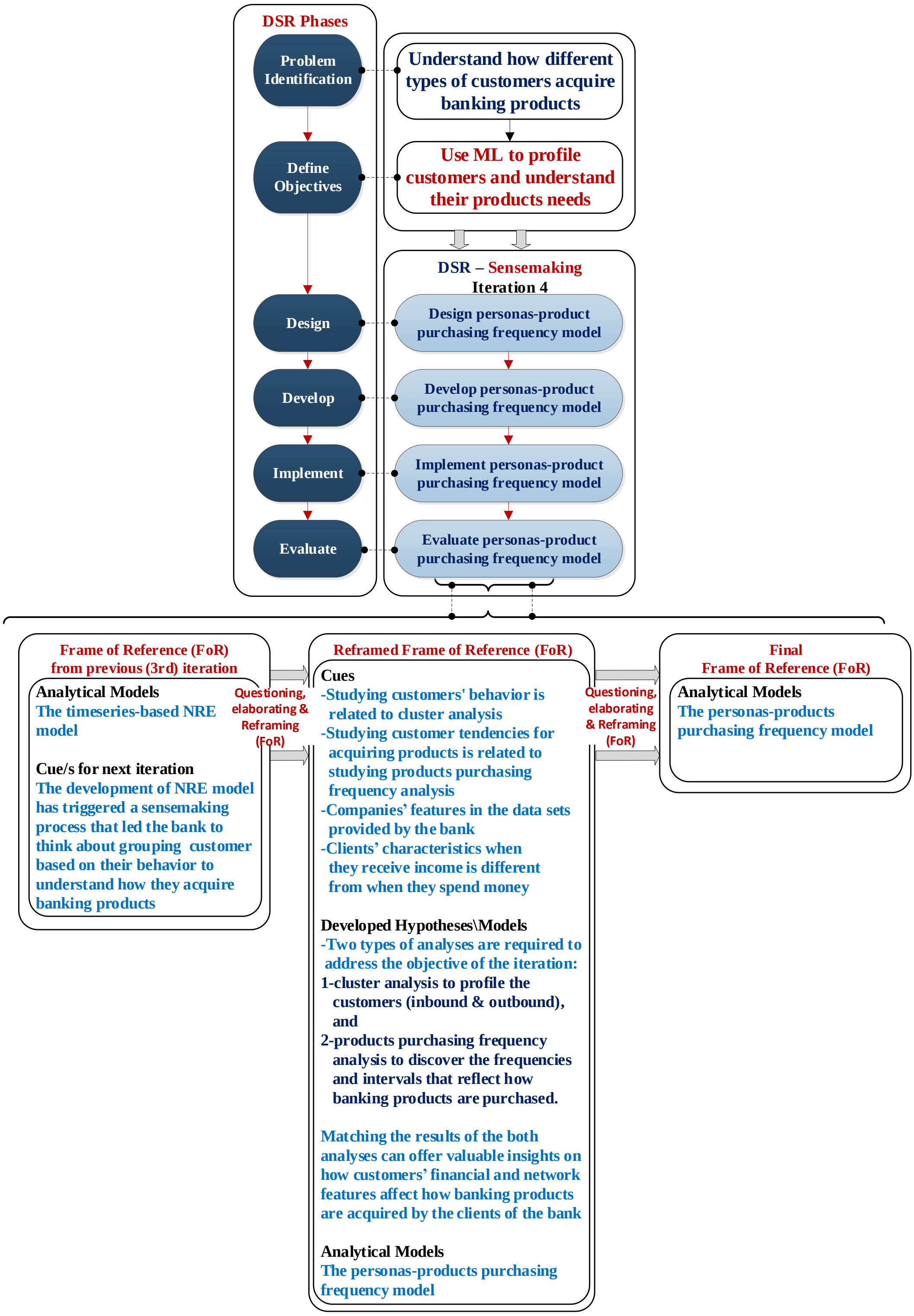
Matching the results of the both analyses can offer valuable insights on how customers' financial and network features affect how banking products are acquired by the clients of the bank

Analytical Models

The personas-products purchasing frequency model

Figure (7.16) illustrates an overall view of the fourth iteration presented in this chapter. The final Frame of Reference (FoR) includes the analytical model that enables BankCo to make sense of banking product purchasing patterns by different groups of customers.

Figure 7.16: Overall view of the fourth iteration



7.6 Summary

This chapter has explained the fourth iteration of this research and demonstrated how sensemaking processes happen during this iteration. The firm personas model designed and developed in this chapter is based on the use of clustering algorithm as an unsupervised Machine Learning technique. The concept of profiling the customers based on their features enables BankCo to understand its clients and their needs of the banking products. The evaluation of the analyses performed in this iteration has shown their possible ability to support the analysts at BankCo as a product recommendation system.

Chapter 8: Conclusion

8.1 Overview

This chapter summarises this research, presents its findings and highlights the main contributions and limitations of this study. This chapter consists of the following sections: Section 8.2 summarises the whole work and provides a brief description of each chapter. Section 8.3 identifies the research contribution to knowledge. Section 8.4 explains how the defined objectives were achieved. Finally, Section 8.5 discusses the limitations of the research as well as future work.

8.2 Research Overview and Outcomes

The aim of this research is to conceptualise the human sensemaking process and to investigate how emerging technologies such as Social Network Analysis (SNA) and Machine Learning (ML) techniques and algorithms can enhance human sensemaking to generate valuable insights during data analysis. Therefore, a conceptual sensemaking framework was introduced to explain the core sensemaking processes of noticing, interpretation and action. Also, the conceptual framework illustrates the cycle of framing and reframing a Frame of Reference (FoR), which holds the cues and hypotheses produced throughout the sensemaking cycles. The conceptual sensemaking framework was then demonstrated in the context of banking data analysis through four Design Science Research (DSR) iterations described in Chapters 4, 5, 6 and 7. The brief descriptions of the chapters of this research are as follows:

Chapter 1 presented an overview of the thesis alongside the aim and objectives of the work. The chapter sought to highlight the importance of the research as a response to the lack of research that aim to operationalise sensemaking, presenting the case for understanding how technology and humans work together in the sensemaking process.

Chapter 2 presented a detailed review of the literature related to decision making as a cognitive process paving the way to Naturalistic Decision Making (NDM) and its relation to sensemaking. A comprehensive review of the literature was presented, that discussed sensemaking from organisational, enactive and computational perspectives. These approaches all agree on the retrospective nature of how individuals shape their world via recurring interaction as well as the creation of meaning and value. Additionally, both organisational and enactive literature, examine how the new knowledge is created via a continuous (social) interaction with the environment they interact with and/or exist within. Finally, identifying constructs such as the Frame of Reference (FoR), cues and hypotheses as a result of the human sensemaking process was a common focal point for the enactive and computational approaches. Understanding these perspectives enabled the introduction of a framework that conceptualises the human sensemaking process. Given the complexity of the business environment and the amount of data generated from various information systems, however, the conclusion was drawn that there is a need to understand how the emerging technologies and human sensemaking can work together to support business analytics and the generation of useful insights.

Chapter 3 discussed research in the Information Systems (IS) arena and covered research paradigms including positivist, interpretivist and Design Science Research (DSR). Also, the chapter justified why the DSR methodology, with its main stages of design, develop, implement and evaluate, was adopted to demonstrate the sensemaking core processes of noticing, interpretation and action. Moreover, this chapter has introduced the case study for this research including the data sets provided by the research partner and their analytical requirements. Finally, this chapter has covered a brief description of the four DSR iterations that construct this research and demonstrate the application of proposed conceptual sensemaking framework in a real-life data analysis.

Chapter 4 explained the first iteration of this research, in which the Connected Customer Lifetime Value (CCLV) model is designed, developed, implemented and evaluated. This model is introduced to address the first analytical requirement requested by the research partner. The CCLV model is based on a marketing measure called Customer Lifetime Value (CLV), which assesses the present value of the cash flow generated by a customer minus the cost of acquisition or retention of that customer. The CCLV model improves upon the traditional CLV calculation by considering the connectivity of the customer in the network, hence, Social Network Analysis (SNA) was employed.

Chapter 5 the evaluation of the CCLV model developed in the first iteration has led to re-construct the model in a timeseries format while considering the size of the firms as a controlling variable. Thus, a second DSR iteration (sensemaking cycle) presented in this chapter and explained how the optimal time window was calculated based on analysing transactions frequency per day. Therefore, a 'sliding window' approach was then implemented, consolidating 90 days into a window where the transactions were aggregated and edge weights calculated. This resulted in 28 window periods representing the timeseries and covers the 30 months. A decile based CCLV score, thus, was calculated 28 times covering 28 windows while using the size as a controlling variable and allowing for the comparison with the sector CCLV decile. As proved by the evaluation phase in this iteration, calculating and structuring CCLV results in a timeseries format allows BankCo to spot extreme changes in its score over the time and enable them to spot potential opportunities for recommending banking products.

Chapter 6 explains the third DSR iteration, in which a new model, called the Network Relationship Model (NRE) was presented. This model aimed at addressing another requirement which is about understanding and monitoring the relationship between BankCo and its customer over the time. The NRE model is based on the concept of loyalty, which is the base for Relationship Equity (RE).

The transactional data under investigation, however, limits the factors that can be considered for calculating RE score, thus, a timeseries network-based relationship equity model is developed. The evaluation of the timeseries-based NRE model has proved its ability to support the analysts at BankCo to explore and spot interesting patterns in customers' income and spend from and within BankCo's universe of customers against their activity outside the bank. Consequently, the NRE model allows for BankCo to act upon the threats or opportunities discovered over the time

Chapter 7 described the fourth iteration of this research that aimed at utilising clustering algorithms as an unsupervised Machine Learning (ML) technique to profile the customers of BankCo into clusters based on their financial and network features. The clustering task was performed twice differentiating between the buyer and supplier behaviour for each firm. Then, the clustering results were matched with products purchasing frequency analysis to uncover how clients within each cluster acquire banking products. Additionally, the firm personas model allowed to discover the clients with multiple product needs, which would enable BankCo to design specialised products to meet customer expectations.

8.3 Research Contribution

This research has introduced several additions to the Information Systems (IS) knowledge base. Specifically, the contributions of this research are categorised into 1) the developed artefacts as contributions to Information Systems (IS) and, 2) the design construction of knowledge, represented in foundations.

8.3.1 The Developed Artefacts as a Contribution to IS

The CCLV model represents the ***first artefact*** from this research that enables BankCo to evaluate its customers based merely on transactional data. This model proved its ability as an alternative customer valuation technique for BankCo, which adds richness in comparison to their standard risk factor approach. Table (8.1) lists the DSR outputs related to this artefact. The definitions of these outputs were discussed in detail in Section (3.3.2) of the third chapter.

Table 8.1: Outputs from iteration 1

Artefact	Outcome from the artefact
Construct	The first iteration covered the traditional Customer Lifetime Value (CLV) and discussed how Social Network Analysis (SNA) techniques can add the element of the connectivity to the traditional CLV measure. Thus, as a construct, the CCLV model is developed in introduced in this iteration
Model	Iteration 1 has introduced a network-based CLV model, called the CCLV that evaluates BankCo customers based on the influence they have on their first-order neighbours in the network.
Method	The develop phase of the first iteration has provided a methodological approach to use SNA to understand a customer's influence within the network.
Instantiation	The CCLV model was implemented using real banking transactional data and proved its value to provide contribution in the form of applicability to transaction-oriented environments and a focus on firm-to-firm relationships (which is novel in the literature to date)

The timeseries-based size-controlled CCLV model represents the ***second artefact*** of this research. In the second iteration, the CCLV model was reconstructed in a timeseries based format and used the size of the firms as a controlling variable; allowing for more realistic and unbiased results. As demonstrated in the evaluation of this model, it allows BankCo to spot extreme changes in its score over time, enabling them to monitor the changes in customers' influence within the network and to spot potential opportunities for recommending banking products. Table (8.2) lists the DSR outputs related to this artefact.

Table 8.2: Outputs from iteration 2

Artefact	Outcome Description
Construct	The second iteration covered the shortcoming of the CCLV model and discussed how to construct a three-months sliding window based on the transactions frequency per day. Also, it

	introduced how the size of the firms can be used as a controlling variable for the modified CCLV model.
Model	Provision of a timeseries-based size-controlled CCLV model to enable the analysts at BankCo to monitor customer performance over the time.
Method	The iteration has provided a method that explains how to construct a sliding window-based timeseries.
Instantiation	The improved model is implemented using the same banking transactional data and proved its ability to spot extreme changes in customers influence within the network over the time.

The ***third artefact*** is the Network Relationship Equity (NRE) Model that is based on the definition of Relationship Equity (RE) as a measurement of customers loyalty to a brand or business. The NRE model, in both its forms, income and spend, offers a dashboard-like system to alert an analyst/relationship manager about extreme changes in customers relationship with BankCo. Additionally, it provides ‘trace points’ from which related data can be investigated to understand potential causes of change (e.g., price, credit status, sector related factors) and act upon them (e.g., via product modification, cross-sell etc.). Finally, the model allows uncovering opportunities for BankCo to discuss and offer new banking products to the customers based on their spending patterns. Table (8.3) lists the DSR outputs related to this artefact.

Table 8.3: Outputs from iteration 3

Artefact	Outcome Description
Construct	The third iteration has identified Relationship Equity (RE) as the measure to understand the relationship between BankCo and its customers compared to their relationship with the market. Thus, as a construct, the NRE model is developed, implemented and evaluated in this iteration.
Model	Iterations 2 has introduced NRE model as a network-based relationship equity model. This model enables the analysts to spot extreme changes over time and trace the reasons behind them.
Method	The iteration has provided a process to understand customers’ relationships with BankCo based on their spend and income patterns inside and outside BankCo’s universe of customers

Instantiation	This iteration has demonstrated an instantiation of the NRE model using BankCo's transactional data. The evaluation of the NRE model has proved its ability to uncover risk cases when there is a decline in customer's spend and spot potential product recommendation opportunities.

The ***fourth artefact*** of this research is presented in Iteration 4, in which a firm personas model is constructed to enable BankCo to understand how different types of customers acquire banking products. The model consists of two analyses, Machine Learning (ML) based clustering analysis and product purchasing frequency analysis. The evaluation of this model demonstrates its potential as a product recommendation system. Table (8.4) lists the DSR outputs related to this artefact.

Table 8.4: Outputs from iteration 4

Artefact	Outcome Description
Construct	Iteration 4 has identified the requirement as clustering related problem. Thus, clustering algorithms have been discussed as a potential solution. Consequently, the persona-purchasing frequency model was introduced as a construct resulted from this iteration.
Model	The fourth iteration has introduced clustering-based customers profiling approach, namely persona model, and combined it with product purchasing analysis.
Method	This iteration has offered a solution process to understand how customers features affect their needs for banking products.
Instantiation	This iteration has demonstrated an implementation of the personas-products purchasing behaviour model. This instantiation can be used by the analysts at BankCo to uncover opportunities for products recommendations.

8.3.2 Contributions to the IS Knowledge Base (Knowledge Construction)

In today's fierce market competition, information systems play a vital role in supporting humans' ability to manage and make sense of data and information. The choice of technology, however, is highly dependent on the past experiences and knowledge that reside in human minds, the limitations imposed by the situation and environment, the availability of the data and information, and finally humans' interpretation of the cues noticed in the data. This research, thus, has contributed to the IS knowledge base by demonstrating how human sensemaking and emerging research and analytical techniques can work together while considering the constraints discussed in the data provided by the research partner, BankCo. More specifically, the sensemaking evaluations and discussions provided at the end of each iteration have shown how the interpretation of the data and the noticed cues has guided the researcher to theorise the need to investigate specific research fields and to use specific technologies to support human sensemaking process in order to meet the analytical requirements. The formulation of these requirements was only possible through the sensemaking cycles that happened in each iteration, which demonstrate how one sensemaking cycle has led to another one until the satisficing answers were achieved.

As explained in Chapter 3, the design process, in its nature is non-linear (Gregor, 2009), in which the problem space defines the characteristics of the solution space, which, in turn, reframes the problem space. Both, the design and the sensemaking processes are iterative in nature, in which a bridge building process facilitates the identification of key concepts within the problem space in order to shape the solution space. This research, thus, has contributed to knowledge by demonstrating how the iterative nature of sensemaking enables the transition from the problem space to the solution space to design the required artefacts using the iterative processes of noticing, interpretation and action to collect cues and data, synthesise hypotheses and build analytical models. Then, questioning, elaborating and framing/reframing the Frame of Reference (FoR) helps in defining the satisficing answers that shape the solution space.

This research has adopted a novel design methodology to demonstrate sensemaking cycles and the core processes of noticing, interpretation, and action through the implementation of the DSR phases (design, develop, implement and evaluate). The approach has illustrated how human sensemaking can be enhanced using emerging and automated technologies such as Social Network Analysis (SNA) and Machine Learning (ML) techniques. Consequently, problem understanding and knowledge construction were improved by the adopted approach in the ways outlined in the practical iterations. The design construction knowledge was improved by providing detailed explanations of how the artefacts were developed, which is considered as another important contribution to the IS knowledge base.

The review of the literature shows variance in the way that Customer Lifetime Value (CLV) and Relationship Equity (RE) models are constructed – that variance stems, in good part, from the data that is available to support the construction of the model. Having developed the CCLV and NRE models by applying SNA concepts on the minimum information available, this research contributes to knowledge in the form of applicability to transaction-oriented environments and a focus on firm-to-firm relationships (which is novel in the literature to date).

Additionally, though the application of clustering algorithms on customers' data is not novel, combining the results of the clustering task (inbound and outbound clustering) with product purchasing frequency analysis provides a novel approach to product recommendation. Thus, the design process of the personas-products purchasing frequency model introduced in the fourth iteration is considered as a contribution to the knowledge base of IS.

Finally, the employment of sensemaking has enabled a fledgling study in the interaction between problem and solution spaces when using DSR alongside how technologies such as Social Network Analysis (SNA) and Machine Learning

(ML) can help mediate that interaction. This aspect has not been explored in detail in the literature to-date. The DSR approach proved to be a good fit to the processes that construct human sensemaking for two reasons: 1) The iterative nature of sensemaking processes is similar to the incremental and iterative development of the intended artefacts in DSR approach, and 2) the design, develop, implement and evaluate stages of the DSR paradigm provide suitable approach to practice the core sensemaking processes of noticing, interpretation and action; leading to a framing/reframing of the Frame of Reference (FoR).

8.4 Meeting Research Objectives

This section explains how the four research objectives of this research have been fulfilled.

Objective 1: Review the literature in order to understand the state-of-the-art in sense making.

This objective has been met in Chapter 2, the literature review chapter, which explained the similarities and differences between decision making and sensemaking in the organisational realm. Also, it has provided a thorough review of the different types of decision-making strategies within organisations as well as the cognitive nature of this mental process and its relation to sensemaking. Additionally, it discussed naturalistic decision making as the basis for human sensemaking. Furthermore, it offered an in-depth review of the literature related to sensemaking from three perspectives: organisational, enacted and computational sensemaking. Moreover, the importance and the challenges that hinder sensemaking process have been covered.

Objective 2: Develop a framework that captures the core sense making processes that support data exploration and analysis.

This objective was achieved in Chapter 2, which proposed a conceptual framework that explains the core processes (cycle) of sensemaking: noticing, interpretation and action and how they, together, form and reform the Frame of Reference (FoR).

Objective 3: To empirically examine that framework to explore the relationship between human sense making and (automated) analytical techniques (in a financial services context).

Objective 4: To evaluate the insights gained from empirical application.

Objectives 3 and 4 were addressed in Chapters 4, 5, 6 and 7. These chapters have demonstrated four sensemaking cycles through four DSR iterations. Each iteration has discussed sensemaking processes: noticing, interpretation, and action through design, develop, implement and evaluate phases. The resulting artefacts were evaluated and shown their possible ability to the research partner, thus, the efficiency of the selected techniques.

8.5 Research Limitations and Future Work

Though the outcomes of the work here provide contributions, there are several limitations that need to be accounted for that, for example, hinder the full implementations of the developed models (CCLV, NRE and firm personas models). The following points summarise these limitations:

- The first limitation is related to data accessibility, which restricted full evaluation of the models. For example, the CCLV and NRE models promise to be timeseries-based event spotting (and alerting) systems, however, these systems need to be complemented with additional data about the customers that allow the analysts to investigate the reasons behind the extreme changes. Access to this data was restricted by the research partner due to the sensitivity of the information. Thus, a full evaluation of the analytical models' potential was not achieved.

- The second limitation is caused by the missing firmographic data provided by BankCo. This limitation was particularly important for the firm personas model since missing data caused the number of firms considered for clustering to drop down to the third of the original number of firms. Consequently, matching the results of clustering and product purchasing frequency analyses was affected and it did not cover all of the firms that were originally considered for analysis.
- The third limitation is related to implementation time of the developed models. This is because monitoring timeseries-based events takes times to spot interesting patterns, and due to the time limitation of the research, it was not possible to spot several interesting patterns that help the analyst to fully make sense of the reasons that might cause a customer to acquire banking products or leave BankCo.

The analytical models introduced in this research have the potential to assist human sensemaking in the context of banking transactional data analysis. Bearing in mind the limitations discussed above, however, future research areas include:

1) To consider additional data sets that can further assist human sensemaking and can provide deeper insights important to BankCo;
2) To implement the models for a time that is long enough to spot and analyse extreme changes in the scores of the models;
3) Due to its potential to profile the customers based on their behaviours, implementing the firm personas model in different contexts other than products purchasing recommendation represents an interesting research area;
4) Information Technology (IT) artefact design is the focal point in Design Science Research (DSR), hence, scant attention is paid to the artefact shaping influenced by the organisational context. Combining Action Research (AR) and DSR methodologies may, thus, be a valuable addition

in: (i) capturing the contextual richness that enables the transition in the circular relationship between the problem space and the solution space; and (ii) understanding the stopping conditions in exploring that circular relationship.

References

Adams, M.J., Tenney, Y.J. and Pew, R.W. (1995) 'Situation awareness and the cognitive management of complex systems', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1), pp. 85-104.

Allard-Poesi, F. (2005) 'The paradox of sensemaking in organizational analysis', *Organization*, 12(2), pp. 169-196.

Baskerville, R.L. and Myers, M.D. (2002) 'Information systems as a reference discipline'. *Mis Quarterly,* 26(1), pp. 1-14.

Baum, M. and Singh, I. (2008) 'Winning the customer service battle', *Chicago: AT Kearney, Inc,* 5.

Berger, P.D. and Nasr, N.I. (1998) 'Customer lifetime value: Marketing models and applications', *Journal of interactive marketing,* 12(1), pp. 17-30.

Bhat, A. (2014) 'K-Medoids clustering using partitioning around medoids for performing face recognition', *International Journal of Soft Computing, Mathematics and Control (IJSCMC),* 3(3), pp. 1-12.

Blattberg, R.C., Malthouse, E.C. and Neslin, S.A. (2009) 'Customer lifetime value: Empirical generalizations and some conceptual questions', *Journal of Interactive Marketing,* 23(2), pp. 157-168.

Blondel, V.D., Guillaume, J., Lambiotte, R. and Lefebvre, E. (2008) 'Fast unfolding of communities in large networks', *Journal of statistical mechanics: theory and experiment,* 2008(10), pp. P10008.

Boland, R.J. (2008) 'Decision Making and Sensemaking', in *Handbook on Decision Support Systems 1.* Springer, pp. 55-63.

Borle, S., Singh, S.S. and Jain, D.C. (2008) 'Customer lifetime value measurement', *Management science,* 54(1), pp. 100-112.

Bryant, D.J. (2002) 'Making naturalistic decision making “fast and frugal”', *Proceedings of the 7th International Command and Control Research and Technology Symposium.*

Butts, C.T. (2008) 'Social network analysis: A methodological introduction', *Asian Journal of Social Psychology,* 11(1), pp. 13-41.

Chang, W., Chang, C. and Li, Q. (2012) 'Customer lifetime value: A review', *Social Behavior and Personality: an international journal,* 40(7), pp. 1057-1064.

Chawla, S. and Gionis, A. (2013) 'k-means–: A unified approach to clustering and outlier detection', *Proceedings of the 2013 SIAM International Conference on Data Mining.* SIAM, 189-197.

Chen, C. and Myagmarsuren, O. (2011) 'Brand equity, relationship quality, relationship value, and customer loyalty: Evidence from the telecommunications services', *Total Quality Management & Business Excellence,* 22(9), pp. 957-974.

Chen, D., Fu, Y. and Shang, M. (2009) 'A fast and efficient heuristic algorithm for detecting community structures in complex networks', *Physica A: Statistical Mechanics and its Applications,* 388(13), pp. 2741-2749.

Chen, H., Chiang, R.H.L. and Storey, V.C. (2012) 'Business Intelligence and Analytics: From Big Data to Big Impact.', *MIS Quarterly,* 36(4), pp. 1165-1188

Choo, C.W. (2006) 'The knowing organization: How organizations use information to construct meaning, create knowledge, and make decisions'. 2nd Edition edn. New York: Oxford university press.

Colville, I., Brown, A.D. and Pye, A. (2012) 'Simplexity: Sensemaking, organizing and storytelling for our time', *Human Relations,* 65(1), pp. 5-15.

Couger, J.D. (1996) 'Measurement of the climate for creativity in IS organizations', *Creativity and Innovation Management,* 5(4), pp. 273-279.

Crisci, C., Ghattas, B. and Perera, G. (2012) 'A review of supervised machine learning algorithms and their applications to ecological data', *Ecological Modelling,* 240, pp. 113-122.

Davenport, T.H., Barth, P. and Bean, R. (2012) 'How 'big data is different', *MIT Sloan Management Review,* 54(1), pp. 22-24.

De Jaegher, H. and Di Paolo, E. (2007) 'Participatory sense-making', *Phenomenology and the cognitive sciences,* 6(4), pp. 485-507.

Di Paolo, E.A. (2005) 'Autopoiesis, adaptivity, teleology, agency', *Phenomenology and the Cognitive Sciences,* 4(4), pp. 429-452.

Dougherty, D.S. and Drumheller, K. (2006) 'Sensemaking and emotions in organizations: Accounting for emotions in a rational(ized) context', *Communication Studies,* 57(2), pp. 215-238.

Feiz, S., RamezaniGhotbabadi, A. and BteKhalifah, Z. (2016) 'Customer Lifetime Value in Organizations', *Asian Journal of Research in Social Sciences and Humanities,* 6(5), pp. 53-58.

Ferrentino, R., Cuomo, M. and Boniello, C. (2016) 'On the customer lifetime value: a mathematical perspective', *Computational Management Science,* 13(4), pp. 521-539.

Fischer, C., Winter, R. and Wortmann, F. (2010) 'Design theory', *Business & Information Systems Engineering*, 2(6), pp. 387-390.

Frank, U. (1999) 'Conceptual modelling as the core of the information systems discipline-perspectives and epistemological challenges', *Milwaukee AMCIS Proceedings*, pp. 240.

Gigerenzer, G. and Todd, P.M. (1999) *Simple heuristics that make us smart.* Oxford University Press.

Gill, T.G. and Hevner, A.R. (2013) 'A fitness-utility model for design science research', *ACM Transactions on Management Information Systems (TMIS),* 4(2), pp. 5.

Gleich, D.F. (2015) 'PageRank beyond the Web', *SIAM Review,* 57(3), pp. 321-363.

Gregor, S. (2009) 'Building theory in the sciences of the artificial', *Proceedings of the 4th international conference on design science research in information systems and technology.* Pennsylvania, Philadelphia ACM, 4.

Gregor, S. and Hevner, A.R. (2013) 'Positioning and presenting design science research for maximum impact', *MIS quarterly,* 37(2), pp. 337-356.

Gregor, S. and Jones, D. (2007) 'The anatomy of a design theory', *Journal of the Association for Information Systems,* 8(5), pp. 312-335.

Grimson, A. (2008) 'Coaching: a change management tool: The importance of sensemaking', *Development and Learning in Organizations: An International Journal,* 22(4), pp. 24-26.

Guba, E.G. and Lincoln, Y.S. (1994) 'Competing paradigms in qualitative research', *Handbook of qualitative research,* Thousand Oaks, CA: Sage, 2(163-194).

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N. and Sriram, S. (2006) 'Modeling customer lifetime value', *Journal of service research,* 9(2), pp. 139-155.

Guzella, T.S. and Caminhas, W.M. (2009) 'A review of machine learning approaches to spam filtering', *Expert Systems with Applications,* 36(7), pp. 10206-10222.

Hevner, V.R., March, S.T., Park, J. and Ram, S. (2004) 'Design science in information systems research', *MIS quarterly,* 28(1), pp. 75-105.

Hill, R.C. and Levenhagen, M. (1995) 'Metaphors and mental models: Sensemaking and sensegiving in innovative and entrepreneurial activities', *Journal of Management,* 21(6), pp. 1057-1074.

Hosseini, M. and Albadvi, A. (2010) 'Customer Value Network Analysis for Improvement of Customer Life-Time Value Computation', *International Journal of Electronic Commerce Studies,* 1(1), pp. 15-24.

Jeong, H.S. and Brower, R.S. (2008) 'Extending the Present Understanding of Organizational Sensemaking Three Stages and Three Contexts', *Administration & Society,* 40(3), pp. 223-252.

Kiss, C. and Bichler, M. (2008) 'Identification of influencers—measuring influence in customer networks', *Decision Support Systems,* 46(1), pp. 233-253.

Klein, G., Moon, B.M. and Hoffman, R.R. (2006a) 'Making Sense of Sensemaking 1: Alternative Perspectives.', *IEEE intelligent systems,* 21(4), pp. 70-73.

Klein, G., Moon, B. and Hoffman, R.R. (2006b) 'Making Sense of Sensemaking 2: A macrocognitive model', *Intelligent Systems, IEEE,* 21(5), pp. 88-92.

Klein, H.K. and Myers, M.D. (1999) 'A set of principles for conducting and evaluating interpretive field studies in information systems', *MIS quarterly,* 23(1), pp. 67-93.

Klein, G. (2008) 'Naturalistic decision making', *Human factors,* 50(3), pp. 456-460.

Klier, J., Klier, M., Probst, F. and Thiel, L. (2014) 'Customer Lifetime Network Value', *In Proceedings of the 35th International Conference on Information System (ICIS),* pp. 1–21.

Kodagoda, N., Pontis, S., Simmie, D., Attfield, S., Wong, B.W., Blandford, A. and Hankin, C. (2017) 'Using Machine Learning to Infer Reasoning Provenance from User Interaction Log Data: Based on the Data/Frame Theory of Sensemaking', *Journal of Cognitive Engineering and Decision Making,* 11(1), pp. 23-41.

Kodinariya, T.M. and Makwana, P.R. (2013) 'Review on determining number of Cluster in K-Means Clustering', *International Journal of Advance Research in Computer Science and Management Studies (IJARCSMS),* 1(6), pp. 90-95.

Kolkur, S. and Kalbande, D. (2016) 'Review of Machine Learning Algorithms in R software for Diagnosis of ESD Diseases', *Proceedings of the ACM Symposium on Women in Research 2016.* ACM, 20-25.

Kuechler, B. and Vaishnavi, V. (2008) 'On theory development in design science research: anatomy of a research project', *European Journal of Information Systems*, 17(5), pp. 489-504.

Kumar, V., Lemon, K.N. and Parasuraman, A. (2006) 'Managing customers for value: An overview and research agenda', *Journal of Service Research*, 9(2), pp. 87-94.

Kumar, V. (2010) 'A customer lifetime value-based approach to marketing in the multichannel, multimedia retailing environment', *Journal of Interactive Marketing*, 24(2), pp. 71-85.

Kursa, M.B. and Rudnicki, W.R. (2010) 'Feature selection with the Boruta package', *Journal of Statistical Software*, 36(11), pp. 1-13.

Lancichinetti, A. and Fortunato, S. (2009) 'Community detection algorithms: a comparative analysis', *Physical review E,* 80(5), pp. 056117.

Lanjewar, R. and Yadav, O.P. (2013) 'Understanding of Customer Profiling and Segmentation Using K-Means Clustering Method for Raipur Sahkari Dugdh Sangh Milk Products', *International Journal of Research in Computer and Communication Technology* (*IJRCCT)*, 2(3), pp. 103-107.

Leone, R.P., Rao, V.R., Keller, K.L., Luo, A.M., McAlister, L. and Srivastava, R. (2006) 'Linking brand equity to customer equity', *Journal of service research*, 9(2), pp. 125-138.

Li, P. (2005) 'Cognitive task analysis: A cognitive approach to evaluate evidence-based nursing websites', *OCLC Systems & Services*, 21(3), pp. 252-256.

Li, Z., Wang, G. and He, G. (2017) 'Milling tool wear state recognition based on partitioning around medoids (PAM) clustering', *The International Journal of Advanced Manufacturing Technology*, 88(5-8), pp. 1203-1213.

Llinas, J. (2014) 'A survey of automated methods for sensemaking support', In *SPIE Sensing Technology Applications.* International Society for Optics and Photonics, 912206-912206-13.

Louis, M.R. (1980) 'Surprise and sense making: What newcomers experience in entering unfamiliar organizational settings', *Administrative Science Quarterly*, 25, pp. 226-251.

Low, B. and Johnston, W.J. (2006) 'Relationship equity and switching behavior in the adoption of new telecommunication services', *Industrial Marketing Management*, 35(6), pp. 676-689.

Lycett, M., (2013) ‘‘Datafication’: Making sense of (big) data in a complex world’, *European Journal of Information Systems*, 22, pp. 381-386.

Maitlis, S. (2005) 'The social processes of organizational sensemaking', *Academy of Management Journal*, 48(1), pp. 21-49.

Maitlis, S. and Christianson, M. (2014) 'Sensemaking in organizations: Taking stock and moving forward', *The Academy of Management Annals*, 8(1), pp. 57-125.

Maitlis, S. and Sonenshein, S. (2010) 'Sensemaking in crisis and change: Inspiration and insights from Weick (1988)', *Journal of Management Studies*, 47(3), pp. 551-580.

Maitlis, S., Vogus, T.J. and Lawrence, T.B. (2013) 'Sensemaking and emotion in organizations', *Organizational Psychology Review*, 3, pp. 222-247.

March, S.T. and Smith, G.F. (1995) 'Design and natural science research on information technology', *Decision Support Systems*, 15(4), pp. 251-266.

Mingers, J. (2001) 'Combining IS research methods: towards a pluralist methodology', *Information systems research*, 12(3), pp. 240-259.

Moore, D.T. and Hoffman, R.R. (2011) 'Data-Frame Theory of Sensemaking as a Best Model for Intelligence', *American Intelligence Journal*, 29(2), pp. 145-158.

Muhren, W.J. and Van de Walle, B. (2010) 'A call for sensemaking support systems in crisis management', in *Interactive collaborative information systems*. Springer, pp. 425-452.

Mzoughia, M.B. and Limam, M. (2015) 'An improved customer lifetime value model based on Markov chain', *Applied Stochastic Models in Business and Industry*, 31(4), pp. 528-535.

Nitzan, I. and Libai, B. (2011) 'Social effects on customer retention', *Journal of Marketing*, 75(6), pp. 24-38.

Orasanu, J. and Connolly, T. (1993) 'The reinvention of decision making.', in: G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok (Eds) Decision-making in action: models and methods (Norwood, NJ, Ablex).

Orlikowski, W.J. and Baroudi, J.J. (1991) 'Studying information technology in organizations: Research approaches and assumptions', *Information systems research,* 2(1), pp. 1-28.

Othata, P. and Pantaragphong, P. (2017) 'Number of cluster for k-means clustering by RCFDC method', In the 22nd Annual Meeting in Mathematics (AMM). Department of Mathematics, Faculty of Science. Chiang Mai University, Chiang Mai, Thailand.

Owen, C.L. (1998) 'Design research: building the knowledge base', *Design Studies,* 19(1), pp. 9-20.

Peffers, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S. (2007) 'A design science research methodology for information systems research', *Journal of Management Information Systems,* 24(3), pp. 45-77.

Peffers, K., Tuunanen, T., Gengler, C.E., Rossi, M., Hui, W., Virtanen, V. and Bragge, J. (2006) 'The design science research process: a model for producing and presenting information systems research', *Proceedings of the first international conference on design science research in information systems and technology (DESRIST 2006).* sn, 83-106.

Pirolli, P. and Card, S. (2005) 'The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis', *Proceedings of international conference on intelligence analysis,* pp. 2-4.

Pugh, D.S. and Hickson, D.J. (2007) *Writers on organizations.* Penguin UK.

Purao, S. (2002) 'Design research in the technology of information systems: Truth or dare', *GSU Department of CIS Working Paper.Atlanta*.

Ronhovde, P. and Nussinov, Z. (2009) 'Multiresolution community detection for megascale networks by information-based replica correlations', *Physical Review E,* 80(1), pp. 016109.

Rosvall, M. and Bergstrom, C.T. (2008) 'Maps of random walks on complex networks reveal community structure', *Proceedings of the National Academy of Sciences of the United States of America,* 105(4), pp. 1118-1123.

Rousseeuw, P.J. (1987) 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics,* 20, pp. 53-65.

Salas, E. and Klein, G.A. (2001) *Linking expertise and naturalistic decision making.* Psychology Press.

Schmitt, P., Skiera, B. and Van den Bulte, C. (2011) 'Referral programs and customer value', *Journal of Marketing,* 75(1), pp. 46-59.

Segarra-Moliner, J.R. and Moliner-Tena, M.Á. (2016) 'Customer equity and CLV in Spanish telecommunication services', *Journal of Business Research,* 69(10), pp. 4694-4705.

Seidel, S., Kruse, L.C., Székely, N., Gau, M. and Stieger, D. (2017) 'Design principles for sensemaking support systems in environmental sustainability transformations', *European Journal of Information Systems*, pp. 1-27.

Shah, C. and Jivani, A. (2013) 'Comparison of Data Mining Clustering Algorithms', *Nirma University International Conference on Engineering (NUiCONE),* Ahmedabad, India, IEEE.

Shattuck, L.G. and Miller, N.L. (2006) 'Extending naturalistic decision making to complex organizations: A dynamic model of situated cognition', *Organization Studies,* 27(7), pp. 989-1009.

Simon, H.A. (1996) *The sciences of the artificial.* MIT press.

Singh, S. and Singh, N. (2012) 'Big data analytics', *International Conference on Communication, Information & Computing Technology (ICCICT), 2012.* IEEE, 1-4.

Starbuck, W.H. and Milliken, F.J. (1988) 'Executives' perceptual filters: What they notice and how they make sense', *The executive effect: Concepts and methods for studying top managers,* 35, pp. 65.

Sein, M., Henfridsson, O., Purao, S., Rossi, M., & Lindgren, R. (2011). Action design research. MIS Quarterly, 35(1), 37-56.

Stein, M. (2004) 'The critical period of disasters: Insights from sense-making and psychoanalytic theory', *Human Relations,* 57(10), pp. 1243-1261.

Takeda, H., Veerkamp, P. and Yoshikawa, H. (1990) 'Modeling design process', *AI magazine,* 11(4), pp. 37.

Tavakolijou, M. (2012) 'A Model to Determine Customer Lifetime Value in Iranian Banking Industry'.

Thompson, E. and Stapleton, M. (2009) 'Making sense of sense-making: Reflections on enactive and extended mind theories', *Topoi,* 28(1), pp. 23-30.

Todd, P.M. (2007) 'How much information do we need?', *European Journal of Operational Research,* 177(3), pp. 1317-1332.

Vaishnavi, V.K. and Kuechler, J., W. (2007) *Design science research methods and patterns: innovating information and communication technology*. CRC Press.

Vaishnavi, V. and Kuechler, W. (2013) *Design Science Research in Information System*. Available at: http://desrist.org/desrist/ (Accessed: March/08 2015).

Varela, F.J., Rosch, E. and Thompson, E. (1991) *The embodied mind: Cognitive science and human experience.* MIT press.

Venkatesan, R. and Kumar, V. (2004) 'A customer lifetime value framework for customer selection and resource allocation strategy', *Journal of Marketing,* 68(4), pp. 106-125.

Villanueva, J., Yoo, S. and Hanssens, D.M. (2008) 'The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth', *Journal of Marketing Research*, 45(1), pp. 48-59.

Vogel, V., Evanschitzky, H. and Ramaseshan, B. (2008) 'Customer equity drivers and future sales', *Journal of Marketing,* 72(6), pp. 98-108.

Wang, X., Chen, G. and Lu, H. (2007) 'A very fast algorithm for detecting community structures in complex networks', *Physica A: Statistical Mechanics and its Applications,* 384(2), pp. 667-674.

Wang, Y. and Ruhe, G. (2007) 'The cognitive process of decision making', *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI),* 1(2), pp. 73-85.

Weick, K.E. (1995) *Sensemaking in organizations.* Sage Publications.

Weick, K.E. (1988) 'Enacted Sensemaking in Crisis Situations [1]', *Journal of management studies,* 25(4), pp. 305-317.

Weick, K.E., Sutcliffe, K.M. and Obstfeld, D. (2005) 'Organizing and the process of sensemaking', *Organization science,* 16(4), pp. 409-421.

Weinberg, B.D. and Berger, P.D. (2011) 'Connected customer lifetime value: The impact of social media', *Journal of Direct, Data and Digital Marketing Practice,* 12(4), pp. 328-344.

Xinping, Z. (2002) 'Interpretivist research, positivist research, and field research', *Chinese Education & Society,* 35(2), pp. 39-46.

Zack, M.H. (2007) 'The role of decision support systems in an indeterminate world', *Decision Support Systems,* 43(4), pp. 1664-1674.

Zadeh, R.B.K., Faraahi, A. and Mastali, A. (2011) 'Profiling bank customers behavior using cluster analysis for profitability', *International Conference on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia.*

Zahedi, F.M. and Sinha, A.P. (2010) 'Ontology design for strategies to metrics mapping', in *Global Perspectives on Design Science Research.* Springer, pp. 554-557.

Zhang, H., Liang, X. and Wang, S. (2016) 'Customer value anticipation, product innovativeness, and customer lifetime value: The moderating role of advertising strategy', *Journal of Business Research,* 69(9), pp. 3725-3730.

Zhang, S.S., van Doorn, J. and Leeflang, P.S. (2014) 'Does the importance of value, brand and relationship equity for customer loyalty differ between Eastern and Western cultures?', *International Business Review,* 23(1), pp. 284-292.