



A Data-driven Framework for  
Investigating Customer Retention

A thesis submitted for the degree of Doctor of  
Philosophy

by

Chidozie Simon Mgbemena

Department of Computer Science

Brunel University

July 2016

# Abstract

This study presents a data-driven simulation framework in order to understand customer behaviour and therefore improve customer retention. The overarching system design methodology used for this study is aligned with the design science paradigm. The Social Media Domain Analysis (SoMeDoA) approach is adopted and evaluated to build a model on the determinants of customer satisfaction in the mobile services industry. Furthermore, the most popular machine learning algorithms for analysing customer churn are applied to analyse customer retention based on the derived determinants. Finally, a data-driven approach for agent-based modelling is proposed to investigate the social effect of customer retention.

The key contribution of this study is the customer agent decision trees (CADET) approach and a data-driven approach for Agent-Based Modelling (ABM). The CADET approach is applied to a dataset provided by a UK mobile services company. One of the major findings of using the CADET approach to investigate customer retention is that social influence, specifically word of mouth has an impact on customer retention. The second contribution of this study is the method used to uncover customer satisfaction determinants. The SoMeDoA framework was applied to uncover determinants of customer satisfaction in the mobile services industry. Customer service, coverage quality and price are found to be key determinants of customer satisfaction in the mobile

services industry. The third contribution of this study is the approach used to build customer churn prediction models. The most popular machine learning techniques are used to build customer churn prediction models based on identified customer satisfaction determinants. Overall, for the identified determinants, decision trees have the highest accuracy scores for building customer churn prediction models.

# ABBREVIATIONS

**ABM:** Agent Based Modelling

**ABMS:** Agent Based Modelling and Simulation

**AUC:** Area Under ROC

**B2B:** Business to Business

**CADET:** Customer Agent Decision Trees

**CART:** Classification and Regression Trees

**CDR:** Call Detail Record

**CRISP-DM:** Cross-Industry Standard Process-Data Mining

**CRM:** Customer Relationship Management

**CSD:** Customer Satisfaction Determinant

**CSDO:** Customer Satisfaction Determinants Ontology

**DM:** Data Mining

**DSR:** Design Science Research

**DT:** Decision Trees

**FN:** False Negative

**FP:** False Positive

**GTM:** Grounded Theory Method

**IS:** Information Systems

**MNO:** Mobile Network Operator

**MSI:** Mobile Services Industry

**ROC:** Receiver Operating Characteristics

**OO:** Object Oriented

**SoMeDoA:** Social Media Domain Analysis

**SMS:** Short Message Service

**SNA:** Social Network Analysis

**SVM:**Support Vector Machine

**TN:** True Negative

**TP:** True Positive

# List of Figures

1.1	Overview of the Thesis . . . . .	18
3.1	DSR Phases . . . . .	43
3.2	Research Iterations in line with DSR . . . . .	47
4.1	Tweet frequency from 8am to 8pm on Wednesday 3rd July. . . . .	64
4.2	Tweet frequency from 8am to 8pm on Saturday 29th June . . . . .	64
4.3	Tweet frequency from 8am to 8pm on Monday 1st July . . . . .	65
4.4	Tweet frequency from 8am to 8pm on Tuesday 2nd July . . . . .	65
4.5	Daily Sentiment Analysis scores . . . . .	67
4.6	The process of importing and categorising Tweets into Nvivo10 . . . . .	69
4.7	Results derived from the first round of performing word frequency on tweets . . . . .	69
4.8	Results derived from the second round of performing word frequency on tweets . . . . .	70
4.9	The process of deriving CSDs . . . . .	72
4.10	Customer satisfaction determinants ontograph (CSDO)- An ontograph that shows the relationship between CSDs and related words . . . . .	73
4.11	Customer satisfaction determinant model . . . . .	78

5.1	Decision Tree Classification Process . . . . .	83
5.2	An example of the ROC Curve . . . . .	90
5.3	An example of the Lift Curve . . . . .	91
5.4	Graphic representation of CSDs derived from iteration one . . . . .	95
6.1	An example of deriving attributes for the CADET approach . . . . .	117
6.2	The CADET Model . . . . .	118
6.3	Conceptual Architecture for the CADET Approach and the TEA-SIM Tool . . . . .	121
6.4	Decision Tree Analysis . . . . .	125
6.5	Init.json File . . . . .	126
6.6	Model.json File . . . . .	126
6.7	Stepper function . . . . .	127
6.8	Agent Interaction Process . . . . .	127
7.1	Data-driven Simulation Framework . . . . .	141
7.2	CADET Approach in R . . . . .	142
7.3	ROC curves for Customer Service (Twitter Dataset) . . . . .	163
7.4	ROC curves for Coverage Quality (Twitter Dataset) . . . . .	163
7.5	ROC curves for Coverage Quality (Twitter Dataset) . . . . .	163
7.6	ROC curves for Customer Service (Twitter Dataset) . . . . .	164
7.7	ROC curves for Coverage Quality (Twitter Dataset) . . . . .	164
7.8	ROC curves for Coverage Quality (Twitter Dataset) . . . . .	164
7.9	ROC curves for Customer Service (Telco Dataset) . . . . .	165
7.10	ROC curves for Coverage Quality (Telco Dataset) . . . . .	165
7.11	ROC curves for Coverage Quality (Telco Dataset) . . . . .	166
7.12	ROC curves for Customer Service (Telco Dataset) . . . . .	166
7.13	ROC curves for Coverage Quality (Telco Dataset) . . . . .	167

*LIST OF FIGURES*

viii

7.14 ROC curves for Coverage Quality (Telco Dataset) . . . . .	167
7.15 Snapshot of R code for Churn Analysis . . . . .	168



# List of Tables

2.1	Review of significant studies on customer churn on mobile services	32
3.1	Research Framework (March and Smith, 1995)	40
3.2	DR Artefact Evaluation Criteria (von Alan et al., 2004)	45
3.3	Steps for conducting iteration 1	49
3.4	Steps for conducting iteration 2	52
3.5	Steps for conducting iteration 3	56
4.1	The SoMeDoA Research Framework (Bell and Shirzad, 2013)	61
4.2	Distribution of sentiment score for MNOs	66
4.3	Distribution daily sentiment scores	67
4.4	Senti-Average of the derived Themes (CSDs)	76
5.1	An Example of Logistic Regression on Customers and their probability to Churn	85
5.2	Confusion Matrix	87
5.3	Confusion Matrix for lift chart example	91
5.4	CSD and related terms	94
5.5	Telco Dataset Description	97
5.6	Comparison of Decision Trees techniques for predicting churn on the Twitter and Telco datasets	100

5.7	Logistic Regression Churn Prediction Results . . . . .	102
5.8	SVM churn prediction results . . . . .	103
5.9	Model Comparison on Twitter and Telco Datasets . . . . .	104
5.10	Lift scores on CSDs for Twitter and Telco datasets . . . . .	106
5.11	AUC scores on CSDs for Twitter and Telco datasets . . . . .	107
5.12	Limitations of Data for churn analysis (Hassouna et al., 2015) .	108
6.1	Steps to identifying Customer Types from DT . . . . .	116
6.2	Steps for Decision Tree Analysis . . . . .	124
7.1	Chapters Addressing Objectives of Study . . . . .	136

# Dedication

This thesis is dedicated to my Dad and my Mum for their endless support and prayers in making sure that I complete my PhD.

# Declaration

The following papers have been published as a result of the research conducted in this thesis.

- **Mgbemena C.**, Bell, D (2016) 'Data-driven Customer Behaviour Model Generation for Agent Based Exploration', *Proceedings of the 2016 Spring Computer Simulation Conference, Pasadena, CA, USA*.
- **Mgbemena C.**, Bell, D., Shirzad S.R. (2016) 'Social Media: A source for uncovering the determinants of customer satisfaction in the Mobile Services Industry', *Proceedings of the 2016 UKAIS conference, Oxford, UK*.
- **Mgbemena C.**, Bell D., Saleh N. (2016) 'A Data-driven Methodology for Agent Based Exploration of Customer Retention', *Proceedings of the 2016 DS-RT conference, London, UK*.
- **Bell D.**, Mgbemena C. (2016) 'A Data-driven Approach to Agent Based Exploration of Customer behaviour' *Submitted to the Transactions of the Society for Modelling and Simulation International*.
- **Mgbemena C.**, Bell D., (2016) 'A Novel Approach to Customer Churn Prediction in the Mobile Services Industry', *To be submitted to expert Systems with Applications*.

# Acknowledgments

I would like to thank God almighty for giving me life and good health throughout the process of my PhD. This was a challenging journey but I was able to finish the journey as a result of the endless support I received from my supervisors, colleagues and friends.

A big thank you goes to my first supervisor, Dr David Bell for his endless support throughout this journey. I'd also like to use this medium to thank my second supervisor, Prof. Mark Lycett for his support in making sure I was well prepared for my viva. Finally, I'd like to thank my siblings, Gloria Ikeaba and Emeka Mgbemena. My in-laws, Ejikeme Ikeaba, and Joy Mgbemena. My friends and colleagues, Fadzai Hezel Nkwenzi, Alla, Huda, Tomaso, Sara, Roja, Boyce, Fatima, Zainab, Neda, Masoud Fakhimi and all the students/staff who supported me through this process. May God bless you all.

# Contents

<b>1</b>	<b>Introduction</b>	<b>8</b>
1.1	Overview . . . . .	8
1.2	Background and Motivation . . . . .	9
1.3	Research Aims and Objectives . . . . .	11
1.4	Research Approach . . . . .	11
1.5	Thesis Structure . . . . .	15
<b>2</b>	<b>Literature Review</b>	<b>19</b>
2.1	Overview . . . . .	19
2.2	Customer Relationship Management . . . . .	19
2.2.1	Customer Satisfaction . . . . .	20
2.2.2	Customer Retention . . . . .	23
2.3	Customer Modelling Behaviour . . . . .	24
2.3.1	Traditional Modelling Approaches . . . . .	25
2.3.2	Social Impact on Customer Retention . . . . .	33
2.3.3	Agent Based Modelling and Simulation . . . . .	34
2.4	Summary . . . . .	36
<b>3</b>	<b>Research Methodology</b>	<b>38</b>
3.1	Overview . . . . .	38

<i>CONTENTS</i>	5
3.2 Research Approaches in Information Systems (IS) . . . . .	39
3.3 Design Research Background . . . . .	39
3.3.1 The Design Science Research Process . . . . .	41
3.3.2 Design Science Research Evaluation . . . . .	43
3.4 Applying Design Science Research . . . . .	44
3.5 Research Iterations . . . . .	47
3.5.1 Iteration 1 . . . . .	48
3.5.2 Iteration 2 . . . . .	52
3.5.3 Data Mining Development Cycle . . . . .	53
3.5.4 Iteration 3 . . . . .	54
3.6 Summary . . . . .	57
<b>4 Customer Satisfaction Determinants</b>	<b>59</b>
4.1 Overview . . . . .	59
4.2 Towards the DSR Output Artefact . . . . .	60
4.3 Analysis and Results . . . . .	60
4.3.1 Dataset Description . . . . .	62
4.3.2 Twitter Temporal Separation . . . . .	62
4.3.3 Tweets Per Day . . . . .	63
4.3.4 Sentimental Average Per Day . . . . .	66
4.4 Temporal Coding . . . . .	68
4.4.1 Tweet Per Word . . . . .	68
4.4.2 Reporting Determinants with an Ontology-Based Con- cept Network . . . . .	71
4.4.3 Sentimental Average Per Word . . . . .	74
4.5 Evaluation . . . . .	75
4.6 Summary . . . . .	78

<b>5</b>	<b>Machine Learning for Churn Analysis</b>	<b>80</b>
5.1	Overview . . . . .	80
5.2	Churn Modeling Techniques . . . . .	81
5.2.1	Decision Trees . . . . .	81
5.2.2	Logistic Regression . . . . .	84
5.2.3	Support Vector Machines . . . . .	86
5.3	Classification Model Evaluation . . . . .	86
5.3.1	Binary Classification . . . . .	87
5.3.2	Classification Accuracy . . . . .	88
5.3.3	Sensitivity and Specificity . . . . .	88
5.3.4	Receiver Operating Characteristics . . . . .	89
5.3.5	Lift Chart . . . . .	90
5.4	Churn Modelling Experiments and Results . . . . .	92
5.4.1	Data mining process . . . . .	92
5.4.2	Datasets Description . . . . .	92
5.4.3	Twitter Dataset Description . . . . .	93
5.4.4	Telco Dataset Description . . . . .	96
5.4.5	Data Preparation . . . . .	98
5.4.6	Data Modelling . . . . .	99
5.4.7	Decision Trees Analysis . . . . .	99
5.4.8	Logistic Regression Analysis . . . . .	101
5.4.9	SVM Analysis . . . . .	102
5.4.10	Evaluation of Models and Discussion . . . . .	103
5.4.11	Model Evaluation Metrics . . . . .	103
5.5	Limitations to Data Mining . . . . .	105
5.6	Summary . . . . .	109



<b>6</b>	<b>Data-Driven Approaches for ABM</b>	<b>111</b>
6.1	Introduction . . . . .	111
6.2	Purpose of Model . . . . .	112
6.3	Customer Behaviour Modelling . . . . .	113
6.4	The CADET Approach . . . . .	114
6.5	CADET Model Validation . . . . .	118
6.5.1	The TEA-SIM Tool . . . . .	119
6.5.2	Applying the CADET approach to the Telco dataset . .	120
6.5.3	Agent Attributes and Behaviour . . . . .	122
6.5.4	Dataset Description . . . . .	122
6.5.5	Model Structure . . . . .	122
6.6	Conclusion . . . . .	128
<b>7</b>	<b>Conclusion and Discussion</b>	<b>129</b>
7.1	Overview . . . . .	129
7.2	Research Summary . . . . .	129
7.3	Research Contributions and Conclusions . . . . .	135
7.4	Summary of Design Artefacts . . . . .	140
7.5	Research Limitation and Future Work . . . . .	142

# Chapter 1

## Introduction

### 1.1 Overview

This chapter presents a description of the whole thesis, which investigates customer retention in the mobile services industry (MSI) using selected tools and techniques. After providing a background on customer retention, the aims and objectives of this thesis are presented. The research approach applied to achieve the aims and objectives are explained, followed by a synopsis of the thesis chapters. A diagrammatic representation of this thesis is also presented as a guide for the reader.

This chapter is structured as follows; Section 1.2 presents the background of the research problem domain and the motivation for carrying out the study. Section 1.3 outlines the research aims and objectives derived from the research problem and motivation. Section 1.4 presents a description of the research approach utilised for conducting this study. Section 1.5 provides the structure of this thesis and a diagram that summarises the contents of the thesis.

## 1.2 Background and Motivation

Customer retention is key in customer relationship management (CRM). Customer retention primarily focuses on building a long-lasting relationship with customers. Customer retention strategies are essential for businesses to gain a competitive advantage and to survive in the market place. Therefore, developing an effective customer retention strategy is critical for businesses especially for mobile network operators (MNOs) because of the fierce competition in the market place. In addition, it is crucial for MNOs to develop an effective customer retention strategy due to the high churn rate (mobile customers leaving their MNO for a competitor) and the cost of gaining new customers (Gerpott and Ahmadi, 2015; Vafeiadis et al., 2015)

In developed countries, mobile services companies struggle to acquire new customers due to saturation in the market (Verbeke et al., 2012). As a result, customer retention receives a growing amount of attention from MNOs. The high penetration rates have also inspired many management scholars to focus on customer retention strategies in the MSI (Lee et al., 2015; Min et al., 2015). While the concept of customer retention has been previously presented in the literature, more research needs to be conducted to uncover effective strategies for addressing customer retention. This forms the motivation for this research. Many published studies have shown that customer retention is profitable to a company because: (1) acquiring new customers cost five times more than retaining existing customers (Keramati and Ardabili, 2011), (2) existing customers generate higher profits, become less costly to serve, and may provide new referrals by providing positive word-of-mouth while dissatisfied customers might spread negative word of mouth (Alexandrov et al., 2013; Jones et al., 2014), and (3) losing customers may lead to opportunity costs because of reduced sales (Verbeke et al., 2012). Importantly, a little improve-

ment in customer retention can lead to a significant increase in profit (Epstein and Westbrook, 2012).

The competition in the MSI makes the need to understand customer behaviour crucial because understanding customer behaviour is a fundamental element for the success of a business (Kang et al., 2012). Understanding customer behaviour will also help with the identification of dissatisfied customers who are likely to churn. However, the increasing number of customers in the mobile services market can make it even more challenging for MNOs to understand and effectively cater to customer needs (Schaarschmidt and Kilian, 2014). Customer churn is a challenging issue in the mobile services market (Seo et al., 2008). As a result, various tools and techniques have been utilised to address customer churn including data mining (DM) and statistical approaches.

Customer satisfaction and customer retention are often studied simultaneously because of the relationship between both conceptual elements (Johnson et al., 2012). In addition, customer satisfaction is often seen as a motivator for customer retention (Ribeiro Soriano et al., 2012). While it seems that satisfied customers will remain with their mobile service provider, this is not always the case because satisfied customers can defect while dissatisfied customers can be retained (Ribeiro Soriano et al., 2012). Many published studies investigating customer retention using customer data provided by MNOs primarily focused on the characteristics of customers and customer interaction with their MNO, while ignoring social effects that are likely to cause customer churn (Haenlein, 2013). In addition, many of these studies are predictive in nature and do not give possible reasons on customer decision to churn (Coussement et al., 2015).

### 1.3 Research Aims and Objectives

This research aims to develop a set of methods that enable evidence based, in-depth understanding of customer satisfaction, thereby enhancing customer retention. A mobile services industry context is used to test and evaluate the framework. In order to achieve this aim, the following objectives are set forth:

**Objective 1:** Analyse the normative literature to uncover the state-of-the-art of customer satisfaction and customer retention debates in the MSI.

**Objective 2:** Derive the determinants of customer satisfaction in the MSI using social media data.

**Objective 3:** Assess the current and most widely-used machine learning techniques for analysing customer retention in the MSI highlighting their capabilities and limitations.

**Objective 4:** Develop a data-driven framework for describing agents in agent based modelling and simulation.

**Objective 5:** Demonstrate the effectiveness of the data-driven framework by conducting experiments into mobile service customer retention.

### 1.4 Research Approach

The design science research (DSR) approach is utilised to carry out this study as the study seeks to provide insights into customer retention. The application of DSR seeks to provide insights in forms of models, methods and instantiations (Peppers et al., 2007). Furthermore, the design science approach is suitable for carrying out this study as a set of analytical techniques are utilised to understand the problem domain better. Thereby, providing insights while seeking to make a valid contribution in the problem domain. The DSR process is different compared to other design activities because it involves building,

capturing and communicating knowledge acquired during the design process (Vaishnavi and Kuechler, 2015). March and Smith (1995) describe the DSR approach as both a product and a process. The process integrates a number of design and behavioural science activities i.e. build, evaluate, justify and theorise (March and Smith, 1995), while the products are classified according to the following product classification points (March and Smith, 1995):

- **Constructs** involve concepts that are used to describe problems within the domain of interest and to specify their solutions.
- **Models** represent real-world problems in specific domains.
- **Methods** are a sequence of steps taken to solve specific problems. These steps are based on constructs and models (March and Smith, 1995).
- **Instantiations** are the execution of constructs, models and methods and evaluating the effectiveness of the design research artefact.

DSR must be carried out as a process of establishing the most effective solution to a problem while applying the laws of a problem domain. To establish the effectiveness of a DSR solution, a rigorous validity measure must be carried out to evaluate the effectiveness of the artefact (von Alan et al., 2004). DSR seeks to accomplish an effective solution to a design problem in an iterative manner whereby each iteration accomplishes the build and execute cycle, contributing new knowledge that is utilised to execute further iterations (Peffer et al., 2007).

Four consecutive design research phases were applied to the design and implementation process to achieve the aims and objectives of this study. These phases are:

1. Problem awareness and motivation involves conducting an extensive review into the customer retention domain, identifying a problem and providing a justification for addressing the selected problem.
2. Solutions selection and suggestion involves introducing possible ideas for solving the identified problems with the chosen approaches or frameworks. This phase is carried out in the three iterations of this theses. Different approaches are used to address the objectives of each iteration in this study and the knowledge derived from each iteration is fed into the next iteration while attempting to solve the problem identified in phase one.
3. Development is accomplished by building research artefacts. Artefacts developed in this study are the customer satisfaction determinants ontology (CSDO) and the customer agent decision trees (CADET) framework. CSDO comprises of the identified determinants of customer satisfaction in iteration one. The CADET framework is a data driven approach for describing agents in agent based modelling. The CADET framework is explained in chapter 6.
4. Evaluation is performed by validating the effectiveness of the chosen frameworks adopted to address a research problem. In this research, various evaluation techniques are utilised to validate the chosen approaches used to achieve the objectives of this research. These evaluation techniques are presented in chapters 4 to 6.
5. Conclusion presents the research outputs. In addition, the evaluation results are presented. The limitations to the research are presented and areas for future work are also suggested.

In order to achieve the main artefact for this research, the following activities are conducted in an iterative DSR manner. The activities consist of three iterations and these iterations are briefly summarised below.

**Iteration 1:** This iteration is achieved by synthesising and analysing the customer satisfaction literature. Furthermore, the SoMeDoA framework (Bell and Shirzad, 2013) is used to address customer satisfaction and an evaluation method is used to validate the SoMeDoA framework for identifying the determinants of customer satisfaction in the MSI. The customer satisfaction determinants ontology (CSDO) is derived from this iteration.

**Iteration 2:** This iteration seeks to investigate and analyse churn using the most popular churn modelling techniques. The analysis is carried out based on the established determinants of customer satisfaction derived from iteration one. The popular techniques used for analysing each determinant are compared to obtain the best technique for predicting churn based on each customer satisfaction determinant (CSD). This iteration contributes to the literature by highlighting the best technique for building churn prediction models based on customer satisfaction determinants. The scores derived from the validation tests are also presented.

**Iteration 3:** As a means of addressing the limitations highlighted in iteration 2, iteration 3 seeks to provide a novel approach for conducting agent based modelling. The CADET approach for Agent Based Modelling (ABM) is derived from automated decision trees. The CADET framework is validated by conducting an ABMS experiment. The experiment was conducted to understand the influence of word of mouth (WOM) from close friends and family to customer retention. This iteration contributes to DSR by providing the CADET framework. The CADET framework is discussed in chapter 6.



## 1.5 Thesis Structure

The rest of this thesis is structured as follows:

**Chapter 2** presents a literature review focusing on the key areas of this study: customer satisfaction, customer retention and customer behaviour modelling. This chapter is organised into four main sections. The first section introduces the concept of customer relationship management (CRM) and highlights the relationship between customer satisfaction and customer retention to CRM. The second section provides a review of customer retention in the MSI, exploring the main factors that drive churn. The third section presents a critical review of the key studies in customer retention, highlighting their benefits and limitations to this study. This section also provides the most popular techniques used in customer retention analysis. The final section explores studies on the use of Agent-Based modelling and simulation (ABMS) in customer retention analysis and how ABMS can overcome the highlighted limitations. A popular study is that of Gimblett (2002) who applied ABMS to model decision making while incorporating heterogeneity and interaction/feedback.

**Chapter 3** proposes and describes the use of the DSR approach to address the research problem. This chapter presents a background on design research. In addition, methods and techniques adopted to conduct the study, along with the justification for adopting these techniques are presented. The chapter concludes by explaining the steps and phases carried out to meet the aim and objectives of this study.

**Chapter 4** presents iteration one of this study, which focuses on identifying customers' perception about their MNO using social media. Social media is selected as a means to gather more information from a larger audience. The SoMeDoA approach is adopted for the purpose of extracting and analysing data from the selected domain as a means to uncover the determinants of cus-

customer satisfaction. During the analysis of this phase, more knowledge is gained about the problem domain. In addition, the knowledge gathered during the execution of this iteration was fed into subsequent iterations. The SoMeDoA approach, which was used to develop the CSDO artefact in this iteration is evaluated by conducting interviews with a random selection of mobile network customers who represent all the companies included in the analysis. The findings of the interviews were positive and confirmatory in terms of the significance of the determinants of customer satisfaction. However, the dominant view of the respondents was that the characteristics of the people should also be investigated.

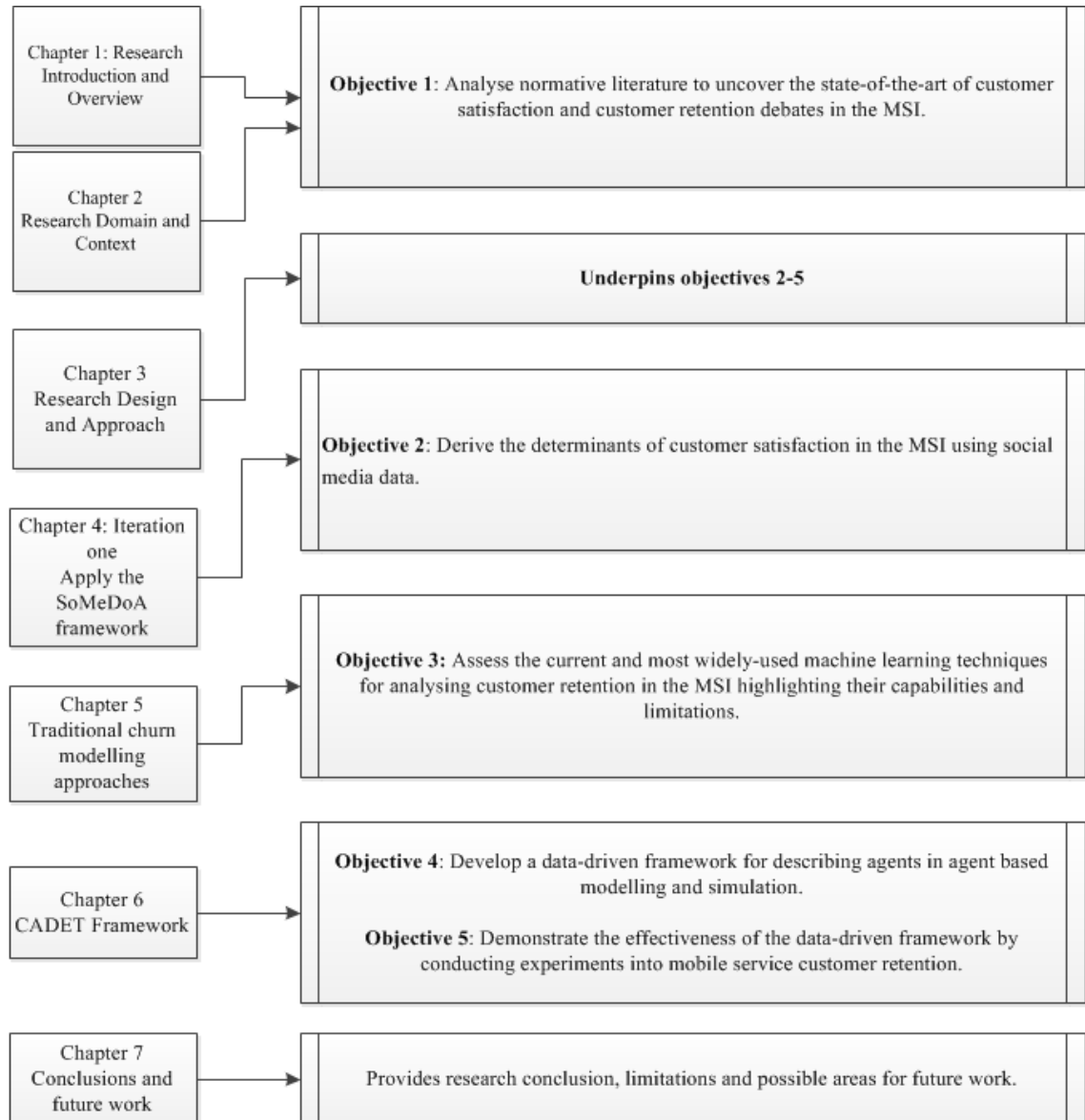
**Chapter 5** presents an empirical evaluation of four widely used techniques for analysing customer retention: classification and regression trees (CART), random forests, logistic regression and support vector machines (SVM). A theoretical background on the selected customer retention analysis techniques are also presented. Subsequently, a classification model evaluation section detailing the evaluation metrics used in this phase is presented. Furthermore, the experiments conducted are reported. Finally, evaluation of the experiments conducted are presented. The results derived in this chapter provide a motivation for the final findings chapter.

**Chapter 6** proposes a data-driven framework for ABM. This chapter focuses on describing the components of this framework, how it can be applied and justification for proposing the framework. A validation of the CADET framework is also presented. The validation is carried out using the TEA-SIM tool. Details about the TEA-SIM tool are also presented in this chapter.

**Chapter 7** concludes the research and presents key contributions and findings. It summarises how the research achieves the aims and objectives. In addition, the contributions made by the study are presented, along with the

limitations to drive further research.

Figure 1.1: Overview of the Thesis



# Chapter 2

## Literature Review

### 2.1 Overview

This chapter seeks to explore the state-of-the-art of studies in customer satisfaction and customer retention in the MSI. The chapter is organised as follows. Section 2.2 presents background on CRM, customer satisfaction and customer retention. Section 2.3 reviews the literature on consumer modelling behaviour, focusing on the traditional methods for modelling customer behaviour, social impacts of customer retention and ABM. Finally, Section 2.4 provides a summary for the chapter.

### 2.2 Customer Relationship Management

Historically, organisations had close relationships with their customers. This was as a result of knowing each customer individually and offering them personal customised service (Benoit and Van den Poel, 2012). The provision of these customised services earned such organisations greater levels of loyalty (Coelho and Henseler, 2012). Over the years, as the market became more com-

petitive, customers interchanged personalised services for anonymity, reduced variety and lower prices (Peppard, 2000). The current business environment is characterised by fierce competition and saturated markets.

Mutanen et al. (2006) argue that the mass marketing approach, where each individual customer receives equivalent treatment from the company, cannot succeed in today's diversity of consumer business. Hence, organisations are practising an approach to marketing described by Peppard (2000) as the customer relationship management (CRM) approach. This approach uses continuously refined information about current and potential customers for businesses to anticipate and respond to their needs. CRM entails structuring and managing relationships with customers (Kim et al., 2003), covering the processes related to customer acquisition, customer cultivation, customer retention and the reactivation of defected customers (Benoit and Van den Poel, 2012).

### **2.2.1 Customer Satisfaction**

Customer satisfaction is regarded as the central element in maintaining long-term customer relationship in the literature of relationship marketing (Siu et al., 2013). Therefore, when customers experience a bad service, it is a crucial challenge to re-establish customer satisfaction and retain unsatisfied customers. Researchers have explored customer satisfaction in various fields and have come up with various definitions for customer satisfaction. A widely used definition in the literature is that of Kotler (2009) who defined customer satisfaction as a person's feeling of pleasure or displeasure resulting from capturing a product's perceived performance (or outcome) in relation to his or her expectations. Kotler (1994) further stated that the key to customer retention is customer satisfaction.

Numerous studies have addressed customer satisfaction in various indus-

tries. Hanif et al. (2010) investigated the factors affecting customer satisfaction in the telecommunications sector in Pakistan using price fairness and customer services as predicting variables. They found that both variables are not only independently important for satisfying customers but they also complement each other. Bamfo (2009) reports that factors such as: friendly, courteous, knowledgeable and helpful employees, accuracy of bills, competitive pricing, and service quality enhance customer satisfaction. Rahman et al. (2011) found that price plays an important role in the choice criteria for mobile telephone operators in Malaysia.

In the telecommunications sector, a number of researchers have found that satisfied customers tend to increase their usage and have the intention to purchase that particular product in the future (Santouridis and Trivellas, 2010; Lee, 2013; Choi et al., 2008). Satisfied customers will repeat the purchase, be brand loyal, and convey positive word of mouth. These will potentially enhance sales and bring about more profit (Oliver, 2010). Leelakulthanit and Hongcharu (2011) investigated the determinants of customer satisfaction by interviewing 400 mobile phone users in Thailand. They found that promotional value, quality of customer service at shops and corporate image play the most important role in determining customer satisfaction.

Mobile service providers have to monitor the market continuously to ensure that their offers, charges, signal coverage, and quality of services are better than their competitors (Almossawi, 2012). Das Gupta and Sharma (2009) found that in order to retain customers and attract new customers, mobile service providers must provide service with reasonable quality without any hidden price, as these are the two most important determinants of customer satisfaction. Despite the cost and difficulty in measuring customer satisfaction, it is still considered an important method of securing a competitive advantage

(Mittal et al., 2005). In order to gain a competitive edge, organisations need to seek the determinants of customer satisfaction and provide them.

The aforementioned studies investigated customer satisfaction with traditional data collection methods such as interviews and questionnaires. However, the area of customer satisfaction is increasingly gaining more audience and researchers are seeking alternative approaches to investigating customer satisfaction.

Agnihotri et al. (2015) investigated the impact of social media on influencing customer satisfaction in business to business (B2B) sales. They found that social media use has an influence on customer satisfaction and customers value interaction with companies on social media. In addition, they found a positive relationship between responsiveness and customer satisfaction, implying that customers also value timely responses from companies.

The use of social media by companies is gradually increasing. As a result, social media is becoming an integral part of companies success. Social media provides an opportunity for companies to gain more exposure, increase traffic and gain more insights in the marketplace (Stelzner, 2011). From a sales force perspective, social media allows sales people to engage customers and develop social grounds where companies can encourage customers to interact and establish relationships with their customers (Agnihotri et al., 2012). Social media offers platforms where companies can communicate with customers enabling a better sales person responsiveness. The extensive use of social media platforms by companies, with increased interaction with companies is causing buyers to gain equal power with sellers in the market place (Prahalad and Ramaswamy, 2004). Consequently, customers have a higher expectation from sellers and sellers are at risk of losing their customer base if an effort is not made to satisfy customers (Agnihotri et al., 2015). The next section presents



a review of some key studies on customer retention.

### 2.2.2 Customer Retention

Over the last decade, the number of mobile phone users has increased reaching an overwhelming number of 7 billion (Ozcan, 2014). In developed countries, telecommunication companies have mobile penetration rates above 100% with no new customers (Verbeke et al., 2012). As a result, customer retention receives a growing amount of attention from MNOs, and management scholars are increasingly conducting more research in customer retention (Backiel et al., 2016; Jahromi et al., 2016). It has been shown in the literature that customer retention is profitable to a company because:

1. Acquiring new customers cost five times more than retaining existing customers (Keramati and Ardabili, 2011).
2. Existing customers generate higher profits, become less costly to serve, and may provide new referrals through providing positive WOM while dissatisfied customers might spread negative word of mouth (Jones et al., 2014).
3. Loosing customers may lead to opportunity costs because of reduced sales (Itzkowitz, 2013).

A number of studies in the area of customer retention have revealed that customer satisfaction is a strong predictor of customer retention (Baumann et al., 2012).

Customer churn is a term widely used in the area of customer retention to describe customers who switch to a different mobile service provider or leave the market entirely. The main factors that influence customer churn in the

mobile services market are: (1) customer satisfaction, (2) switching costs, (3) relationship quality and (4) price (Hanif et al., 2010; Silva and Yapa, 2013; Svendsen and Prebensen, 2013). In addition, social influence is another key driver to customer churn in the MSI (Phadke et al., 2013; Verbeke et al., 2014). Price is the most important factor for customer churn, followed by customer service, service quality and coverage quality (Chu et al., 2007). Most MNOs already have a customer churn prediction model that gives information on the customers with the highest propensity to churn (Verbeke et al., 2012). This enables efficient customer management, and a better allocation of resources for customer retention campaign. There are two basic approaches that can be used to address customer churn namely, untargeted and targeted approaches (Khan et al., 2010). Untargeted approaches rely on outstanding product and mass advertising to increase brand loyalty and retain customers while targeted approaches rely on identifying customers who are likely to churn and then providing them with either a direct incentive or a customised plan for them to stay (Khan et al., 2010). Various types of information can also be used to predict customer churn, such as information on socio-demographic data (e.g. sex, age, or post code) and call behaviour statistics (e.g. the number of international calls, billing information, or the number of calls to the customer helpdesk). Alternatively, social networks information extracted from call detail records can also provide insights on churn prediction (Dasgupta et al., 2008).

## **2.3 Customer Modelling Behaviour**

Customer retention can be achieved if a company is able to understand patterns in which customers behave and the likely triggers for such behaviour. Understanding customers, managing interactions and relationships with them

is a vital part of CRM. A company with a good CRM should be able to predict customer behaviour. Predicting customer behaviour can be achieved through customer behaviour modelling (Ottar Olsen et al., 2013). Customer behaviour modelling entails applying tools and techniques to gain a better insight on customer behavioural patterns and in turn predict future behaviour. Neslin et al. (2006) characterised CRM models as either analytical or behavioural models. Analytical models involve large datasets that are stored in data warehouses. These datasets require models that can easily scale the dataset and provide results to increase company revenue. Behavioural models make use of surveys to analyse cognitive responses to services provided (Neslin et al., 2006). Furness (2001) classified customer behaviour modelling into (1) descriptive modelling, (2) predictive modelling and (3) a combination of descriptive and predictive modelling. Descriptive modelling describes models that attempt to answer ‘why’ questions. An example of descriptive modelling is clustering. When a customer clustering exercise is conducted, customers belong to a certain cluster because they collectively possess similar attributes or behaviours. Predictive modelling describes models that answer the ‘who’ questions. For example who will buy a product or service? In the context of customer churn, predictive models can give insight on who is likely to churn. Predictive models work by predicting future customer behaviour based on their past behaviour. Finally, the combination of descriptive and predictive modelling addresses problems by integrating both descriptive and predictive models to provide a more concise answer on ‘who’ and ‘why’ questions at the same time (Verbeke et al., 2012).

### **2.3.1 Traditional Modelling Approaches**

Numerous methods have been applied to analyse and investigate customer retention. These methods can generally be classified as statistical analysis

methods and data discovery methods. Statistical analysis for modelling customer retention includes correlation analysis, ANOVA analysis and chi-square test. Data discovery methods for modelling customer retention include data mining techniques. Data mining simply means extracting hidden knowledge from data. The use of data mining techniques is popular in the area of customer retention to understand customer behaviour from raw data. Data mining methods are widely used in the literature for analysing and investigating customer churn for two main reasons: they have better prediction results than traditional statistical techniques (Ye et al., 2008) and they are more suitable for analysing large data sets (Lemmens and Croux, 2006). Hence, this study focuses on exploring data mining techniques for customer churn analysis.

Numerous industries have attempted to study customer churn. These industries include banking (He et al., 2014; Raju et al., 2014), insurance (Sundarkumar and Ravi, 2015) and retail (Miguéis et al., 2012, 2013). Customer churn has also been applied into different areas such as economics, and behavioural studies (Neslin et al., 2006). Consequently, there is substantial literature uncovering factors that drive customer retention. These factors have been investigated using various tools and techniques (Ahn et al., 2006).

Gerpott and Ahmadi (2015) explored the capability of socio-demographic, contract and service usage characteristics of MNO subscribers as well as their stated reason for contract termination to predict the likelihood of an entirely successful win back. How customer service usage characteristics and the cause for termination reflect on the MNO's effort to win back customers. This study also explored the outcomes of company programs attempting to motivate customers to withdraw contract cancellation request. The data utilised to conduct this study was collected from the billing system of one of the four mobile network operators that were competing in the German market, with a sample of

305,466 post-paid residential customers. The primary objective of the study is to understand the factors that are associated with the success of an MNO's effort to regain customers who terminated their mobile contract but are still legally bound to their MNO until the end of their contract. The results derived from the study enabled MNO representatives to develop profiles of two distinct target groups. The first group consist of customers who are likely to withdraw their prior termination notice. The MNO's win-back practices appear to fit well with specific needs of the target audience. Hence, MNO executives can make a decision to focus on win-back activities.

Vafeiadis et al. (2015) carried out a comparison of machine learning techniques for customer churn prediction. The machine learning techniques covered in this study are artificial neural networks, SVMs, decision tree learning, naïve bayes and logistic regression analysis. The churn dataset is a telecom dataset obtained from the UCI machine learning repository which is included in package C5.0 and used to evaluate the performance of the tested classifiers. The dataset contains 19 predictors and a binary(yes/no) churn variable. The dataset contained 5000 samples of telecom customer data. The experiments in this study were carried out in two phases; without boosting and with boosting. The top performing method without boosting is the decision trees with an accuracy of 94%. The SVMs classifiers (RBF and POLY kernels) obtained accuracy of 93%. Boosting was carried out using the AdaBoost.M1 algorithm. Naïve bayes and logistic regression classifiers were unable to be boosted because they lack free parameters to be tuned. Hence, the results show that for the three remaining classifiers, artificial neural networks, decision trees and SVMs, accuracy was improved between 1% and 4%. Overall, the boosted SVM was the best classifier with accuracy of almost 97%. The primary benefit of this work is that it provided additional insight on the performance of the most

popular churn prediction techniques for predicting churn in the telecommunication sector. It also sheds light on the application of boosting techniques. The limitations of this work are: 1) more customer data should be tested with the boosting method, 2) other boosting methods other than AdaBoost should be explored, and 3) there is need for a larger and more detailed dataset from the telecom industry to maximise the statistical significance of the result.

Huang and Kechadi (2013) carried out a study for predicting customer behaviour by utilising a novel-hybrid based learning system which integrates supervised and unsupervised techniques for predicting customer behaviour. The applied system combines a modified k-means clustering algorithm and a classic inductive rule technique (FOIL). Three experiments were conducted in this study. The aim of experiment 1 was to verify if the weighted k-means clustering is able to lead to a better data partitioning result. Experiment 2 involved evaluating classification results and comparing them to other well-known modelling techniques. Experiment 3 compared the proposed-hybrid model system with several other recently proposed hybrid classification approaches. The experiments were conducted using a telecom dataset consisting of 104,199 customer records with 6,056 churners and 98,143 non-churners. The attributes in the dataset mainly consist of demographic profiles, account information and call details. The 5-fold cross validation model was used in this research. The results show that the hybrid model-based learning system is very promising and outperforms other models. This work was carried out to understand customer behaviour, however, the limitations of this work are that it did not consider eliminating redundant data and outliers before performing the experiments. This could affect the effectiveness of the model. Furthermore, other clustering algorithms can also be applied to gain more information about customers in several clusters.

Kirui et al. (2013) predicted churn using a dataset that was provided by a European telecommunications company. The dataset was collected for three months with 112 attributes and 106,405 instances. 5.6% of the customers on the dataset were churners while the rest of the dataset consists of active users. In order to make the churn customer data recognised by the mining algorithms, additional features were derived and then added to the dataset. Stratified random sampling was carried out and then added to the original dataset. The problem of imbalance was handled by applying stratified random sampling to both the original and modified datasets. Random sampling was used in each stratum independently to the required data sample size. A new set of features were proposed in this study to improve prediction rates for churn recognition. The results show that most of the feature subsets perform almost equally using Naïve Bayes. The results also show that as sample size increases, the ratio of churn to active also increases. The probabilistic classifiers (Naïve Bayes and Bayesian network) achieved better results when compared to the a decision tree (C4.5). The results also show that the prediction rates on the modified dataset are better than the original dataset. The improvement in the results is attributed to the features of the new dataset which signifies the benefits of new features. The new features added to the dataset are contracted related features, call pattern description features and calls pattern changes. The initial features of the dataset are call profiles and call traffic details. The limitations of this study is that the authors did not attempt to understand customer behaviour, and the imbalance in data may have affected the performance of the modelling results.

Verbraken et al. (2013) proposed a new framework, the cost benefit analysis framework to define performance measures associated with profit maximisation. The authors applied this framework to customer churn with its particular

cost benefit structure. Companies benefit from this approach as it provides an insight on selecting the best classifier for maximising profit. The proposed framework also provides guidance on which customer base should be involved in the customer retention campaign. A case study was carried out by applying 21 techniques to 10 different customer churn datasets. The results show that area under the ROC (AUC) makes incorrect implied assumptions about misclassification costs. Hence, the use of this performance evaluation metric in a business environment may lead to suboptimal profits. The benefit of the EMPC measure is that it maximises profit and provides an insight on the customer base that needs to be targeted for customer retention campaigns. The limitation of this framework is that it is only applicable to a specific business problem, maximising profit. The authors created the H-measure to address the limitation of the AUC measure in their study. The H-measure benefits from the ability to select the most profitable classifier. However, the H-measure does not provide guidance on the optimal fraction of customer base to be included in the retention campaign.

Kim et al. (2014) proposed a new method for churn prediction by analysing how subscribers communicate and considering a propagation process in a network based on call detail records which transfers churning information from churners to non-churners. The dataset used to conduct this study was obtained from a telecommunications company and it includes customer personal information and call detail records (CDR) data. In addition, the dataset consisted of 89,412 customers and 36 variables that include customer personal variables such as demographic information, product details (cell phone types and performance, duration from the last change of cell phone), service satisfaction factors (number of complaints, service quality score, and loyalty score), propensity to telephone calls (proportion of non-voice calls and proportion



of calls during the day time). The authors introduced network analysis and accomplished further improvement in churn prediction compared with the traditional machine learning approach that handles personal information stored in companies. The network variable was generated from SPA, one of the propagation processes, and combined with traditional personal variables to train a model. The authors highlight that for the purpose of future study, a larger dataset which spans across a longer period should be used with the proposed method for churn prediction. Table 2.1 presents a review of some significant studies on customer churn in the MSI.

Table 2.1: Review of significant studies on customer churn on mobile services

Study	Author(s)	Method	Model type(s)	Factors covered
Regaining drifting mobile communication customers: Predicting odds of success of win back efforts with competing risks regression.	Gerpott and Ahmadi (2015)	Survival analysis techniques	Predictive	Customer, contract, usage characteristics and termination reason
A comparison of machine learning techniques for customer churn prediction	Vafeiadis et. al., (2015)	Artificial neural networks, decision trees, logistic regression, Naïve Bayes and SVMs	Predictive	Daily usage characteristics, international plan
An effective hybrid learning system for telecommunications churn prediction.	Huang, Kachadi, Ying and Tahar (2013)	K-means, FOIL	Descriptive and Predictive	Customer demographic profiles, account information and call details
Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining	Kirui et al., (2013)	Naïve Bayes and Bayesian Network	Predictive	Call details, customer profiles and call pattern description
A novel profit maximising metric for measuring classification performance of customer churn prediction	Verbraken T. (2013)	Decision Tree approaches, ensemble methods and SVM based techniques	Predictive	Customer demographic profiles, account information and call details
Improved churn prediction in telecommunications by analysing a large network	Kim, Jun and Lee (2014)	Logistic Regression, Multilayer perceptron (MLP)	Predictive	Demographic information, product details, service satisfaction factors, propensity of telephone calls

### 2.3.2 Social Impact on Customer Retention

More recently, social networks' influence have been found to be a key contributing factor to customer retention (Zhang et al., 2010; Risselada et al., 2014). Social networks' influence is typically carried out with the use of word of mouth. word of mouth is the informal communication between private parties regarding evaluations of goods and services (Anderson, 1998). 75% of defected customers spread a negative word of mouth to one or more customers (Keaveney, 1995). The following section presents some relevant studies on the impact of social influence on customer retention.

Phadke et al. (2013) developed a model that integrates social network analysis with traditional churn modelling concepts. The model was applied to a dataset of over half a million subscribers, provided by a large mobile network provider. The dataset contained customer call detail records (CDR). To compute social tie strength, the authors used three attributes namely: 1) the number of calls placed between two users, 2) the total duration of calls between two users and 3) neighbourhood overlap of the two connected users. This study found users that make phone calls to numbers on a different network are likely to churn in future to save costs.

Similarly, Verbeke et al. (2014) conducted a study investigating the impact of social networks on customer retention. However, the latter differs from the former in that it uses both networked and non-networked (customer-related) information about millions of users. The key finding of this study was that churn not only had an impact on customers' friends, it also had an impact on friends of friends.

Although the studies mentioned above have contributed to the knowledge of customer retention, however, they do not capture the possible factors as to why customers made their decision to churn. This study applies the SoMe-

DoA framework to understand the key factors of customer satisfaction. Furthermore, the classification churn prediction models are built based on these factors. In addition, a novel framework for describing agents in ABM is presented and the model is evaluated with an ABMS experiment.

### 2.3.3 Agent Based Modelling and Simulation

Over the years, researchers and industry practitioners have attempted to apply different techniques to understand customer behaviour in the market place. The ABMS approach is a typical example of one of these techniques (Twomey and Cadman, 2002). ABMS provides an understanding of how systems work under certain conditions. ABMS works by creating scenarios that imitate real life conditions for example carrying out an ABMS exercise with customers who walk into a retail store. ABMS can be used in this context to derive insights on customer behaviour in the retail store. This process provides an explanation of the relationship between elements in a complex system. ABMS is made up of two main parts: modelling and simulation. Modelling is the process of representing real life events into a model while simulation is the process of executing represented models such that they imitate the proposed system. Agent based models are composed of agents and a structure for agent based interaction.

Agents can represent anything from a number of patients in a hospital to consumers of a product or service. ABMs are often characterised by rules and these rules define the behaviour of agents in the system (Macal and North, 2010). These behaviours are often influenced by agent interactions with other agents in the system, making the outcome difficult to predict. In such cases, a balance may be difficult to reach, making the ability to study the underlying system and the dynamics of the behaviour imperative. ABM is distinct from

traditional modelling approaches where characteristics are often aggregated and manipulated (Baxter et al., 2003). Traditional modelling techniques for modelling maybe suitable for their purposes but they may not be able to provide adequate level of details in regards to the independent behaviours of agents. Although the commonly used traditional techniques for modeling consumer markets (discussed in the previous section) are powerful with regard to their purposes, they are generally not able to provide sufficient levels of detail with regard to the interdependent behaviors of consumers, retailers, and manufacturers.

In addition, ABM is able to sufficiently represent interdependent systems even on a large scale i.e. incorporating a high number of factors, with each factor's level of detail and the behavioral complexity of those factors North et al. (2010). ABM is a widely used for modelling complex systems that are composed of interacting independent elements (Macal and North, 2010). The ABM technique can be applied to any aspect of a phenomena. ABM has been applied in various areas including economics (Farmer and Foley, 2009), health-care (Epstein, 2009), management science (Macal and North, 2010) and geography (Heppenstall et al., 2011). In business, ABM has been applied to help decision makers understand underlying market structures and anticipate dynamics in the market place (Macal and North, 2010). ABM has also been utilised in artificial life research, to explore life in order to uncover how it might be, rather than how it actually is (Adami, 1998). ABM is also used in consumer modelling to understand and to predict consumer modelling process (Zhang and Zhang, 2007). Consumers are represented as independent agents with individual characteristics and an independent decision making process (Gilbert, 2008). Sellers are represented as agents who present their products with different characteristics into the market (Roozmand et al., 2011).

In the study of social behaviour and interactions, ABM starts with a set of assumptions derived from the real world (deductive), and produces simulation-based data that can be analysed (inductive) (An, 2012). ABM must create a clear representation of what happens in reality so that every agent performs a task of an individual as if it is happening in social reality (Roosmand et al., 2011).

ABM has a number of benefits such as the ability to model individual decision making while incorporating heterogeneity and interaction/feedback (Gimblett, 2002). Additionally, ABM has the ability to incorporate social/ecological processes, structures, norms and institutional factors (Deadman et al., 2004). These advantages make it possible to couple human and natural systems in an ABM.

Despite the strengths of ABM, it also has some limitations. These limitations include lack of predictive power, difficulty in validation and verification (Matthews et al., 2007).

## **2.4 Summary**

This chapter provides the foundation for this research. Customer satisfaction is vital to the success of any business. Customer satisfaction is one of the key drivers of customer retention, and customer retention is one of the fundamental elements of CRM. Traditional approaches and ABM approaches have been applied to study customer churn in the literature. As reported in the literature, data mining is a popular technique for analysing customer churn in the MSI. This chapter provides some relevant studies of churn analysis using traditional methods, data mining techniques and ABM. This chapter also shows that majority of the studies that applied data mining to address customer

churn did so using decision trees, logistic regression and SVMs. The majority of the studies on customer retention have two common limitations; firstly, they focused on investigating certain factors such as customer satisfaction and customer retentions while neglecting the possible effects of social ties on customer retention (Hanif et al., 2010; Edward and Sahadev, 2011). Secondly, they focus on customer interaction with the network operator (Richter et al., 2010; Günther et al., 2014), neglecting the fact that customers often interact amongst themselves. To overcome these limitations, this study proposes an approach to analysing churn and a novel approach to agent-based modelling. The next chapter presents the methodology for conducting this research.

# Chapter 3

## Research Methodology

### 3.1 Overview

This chapter describes the research approach chosen to investigate customer retention in the MSI. It also provides a description of methods utilised to carryout studies in iterations in this study. Furthermore, an explanation of the research development phases are presented. Design Science Research DSR is utilised as the general methodological framework to conduct this study. DSR is selected because it allows knowledge gained from one iteration to be applied to subsequent iterations. This chapter is organised as follows. It begins by highlighting research approaches in information systems (IS) in section 3.2. Section 3.3 provides a background knowledge on design research. Section 3.4 explains how DSR is applied to achieve the aim and objectives of this research. Section 3.5 provides a description of how the iterations in this study are carried out and how the artefacts generation in each iteration are achieved. Finally, section 3.6 presents an overall summary for this chapter.



## 3.2 Research Approaches in Information Systems (IS)

Information systems is a multidisciplinary field which spans across disciplines such as computer science, management science, engineering and others (Baskerville and Myers, 2002). IS research can be conducted using a variety of research approaches, techniques, methods, methodologies and paradigms. The design research is a popular research approach in conducting IS research because it provides a means for IS researchers to create or improve existing artefacts (Hevner and Chatterjee, 2010). The next section presents a background on design research.

## 3.3 Design Research Background

Design research (DR) is an innovative means of problem solving (von Alan et al., 2004). It is a broad area of research that spans across all design fields but importantly does not have the distinct feature of DSR; learning through building thereby creating an artefact (Kuechler and Vaishnavi, 2008). The term ‘science’ was incorporated into design research when IS researchers discovered that the term ‘design research’ had a past history as the study of design and designers, including, their methods, cognition and education (Kuechler and Vaishnavi, 2008). DR is a research into design or about design while DSR primarily uses design as a research method or technique to solve a problem and learn from the process of solving that problem (Kuechler and Vaishnavi, 2008). DSR is also widely adopted in other fields such as education, engineering, and health-care (Hevner and Chatterjee, 2010). March and Smith (1995) describe DSR as a research methodology that allows research to produce relevant and

	<b>Build</b>	<b>Evaluate</b>	<b>Theorise</b>	<b>Justify</b>
<b>Construct</b>				
<b>Model</b>				
<b>Method</b>				
<b>Instantiation</b>				

Table 3.1: Research Framework (March and Smith, 1995)

improved effectiveness by strategically combining research output (product) and research processing (activities) from both natural and design science in a two-dimensional framework. The design science output or artefacts includes constructs, models, methods and instantiations while the natural science activities include build, evaluate, theorise and justify March and Smith (1995) (see table 3.1). The application of the two-dimensional framework can be described as applying natural science activities to produce design science artefacts; constructs, models, methods and instantiations. Design science achieves satisfactory results to a design problem by process of iterative knowledge refinement (Hevner and Chatterjee, 2010).

The DSR output classification defined by March and Smith (1995) can help establish appropriate measures to build, evaluate, theorise and justify a DSR. The four research outputs are described below:

**Constructs**: Constructs are a set of concepts that are used to describe problems within a domain and specify their solutions. Constructs also form the vocabulary of a discipline.

**Models**: Are a set of statements which express relationships among constructs and represent real-world design activities in a domain (March and Smith, 1995). Models can also be used to suggest solutions to problems in a solution space.

**Methods**: Are a sequence of steps used to execute a task. These steps provide guidelines on how to solve problems with the use of constructs and models. Furthermore, methods can be described as a set of methodological tools that are created by design science and applied by natural science (March and Smith, 1995).

**Instantiations**: Are the utilisation of constructs, models and methods to showcase an artefact in a domain. They demonstrate the effectiveness of the constructs, models and methods (March and Smith, 1995). Newell et al. (1972) describe the importance of instantiations in computer science by explaining how they offer a better understanding of a problem domain, and as a result, provide improved solutions. Instantiations provide working artefacts that can drive significant advancement improvement in both design and natural sciences. A DSR methodology incorporates five stages of a design cycle to address design research problem. These phases are designed to aid sustainable development during the research and transfer of knowledge from one iteration to the next iteration until a desired result is achieved. The next section explains the DSR processes.

### 3.3.1 The Design Science Research Process

The DSR process follows a stepwise approach structured as five phases.

**Awareness of problem**: The DSR process begins by identifying the problem under study. The identified problem may arise from multiple sources such as the literature or current problems in the industry. The research problem needs to be clearly defined and articulated. The output of this phase is a formal or informal proposal for new research.

**Suggestion**: This phase is explored when a research proposal has been presented. Possible solutions about the research problem are explored and evalu-

ated, leading to the acquisition of further insights to the domain under study. The specifications of the appropriate solutions to the research problem are defined. The output of this phase is a conditional design or representation of proposed solutions.

**Development:** This phase involves further developing and implementing the DSR artefacts based on the suggestions from the previous phases. The outputs of this phase are the artefacts, which are core elements of the DSR process. March and Smith (1995) described DSR artefacts into four categories: Constructs, models, methods and instantiations.

**Evaluation:** The developed artefacts are analysed and evaluated according to the criteria set in phase 1 (awareness of problem phase). Deviations and expectations should be noted and explained in this phase. If the outcomes derived from the development or evaluation phase do not meet the objectives of the problem, the design cycle returns to the first phase, along with the knowledge gained from the process of the first round of work. These phases may be iterated until the evaluation of the artefacts meets the solution requirements. The outputs of this phase are performance management that should improve the efficiency and effectiveness of the artefact.

**Conclusion:** This is the last phase of the DSR cycle. The results of the research are written up and communicated to a wider audience in forms of professional publications and scholarly publications (Peffer et al., 2007). Kuechler and Vaishnavi (2008) categorise the knowledge gained in this phase as either firm or loose ends. Firm knowledge are facts that have been learned and can be repeatably applied or behaviour that can be repeatably invoked, while loose ends are anomalous behaviour that defies explanation and may well service as the subject of further research (Kuechler and Vaishnavi, 2008)

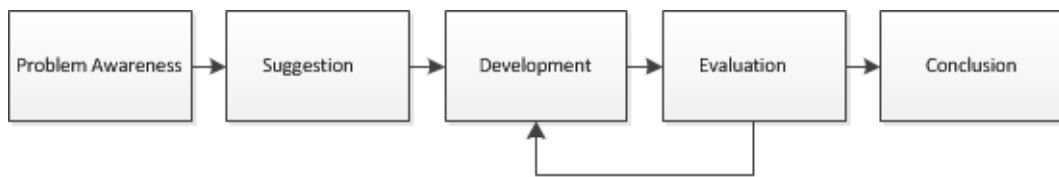


Figure 3.1: DSR Phases

### 3.3.2 Design Science Research Evaluation

Evaluation is an integral part of the DSR process. It is an avenue to validate the performance of an artefact and measure progress according to the defined metrics (March and Smith, 1995).

Artefacts are constructed to carryout specific problems, thereby, demonstrating their effectiveness in solving the problems. The process of developing an artefact may result in deviations from expectations. In this case, these deviations should be properly explained (Kuechler and Vaishnavi, 2008). Knowledge gained from the evaluation phase of one iteration can be applied into further iterations. Evaluation plays an essential role in DSR as it is iterative in manner. Hence, it is important to develop appropriate evaluation metrics to assess artefact performance and to measure the efficiency and effectiveness of artefacts (March and Smith, 1995). The criteria for evaluating the quality of an artefact depends on the artefact type (March and Smith, 1995). Table 3.2 presents types of artefacts as described by (March and Smith, 1995) and their evaluation criteria.

Usually, evaluation is concerned with answering the question 'How well does the artefact work?' (March and Smith, 1995). This question can be answered by identifying a suitable evaluation measure to validate the effectiveness of the artefact. An appropriate evaluation method is selected once the evaluation metrics and criteria are identified (March and Smith, 1995). The selection of an evaluation method should be carefully considered, and when suited with an

artefact and metric evaluation, evaluation methodologies are typically drawn from the knowledge base (von Alan et al., 2004). A design artefact should be comprehensive and effective such that it fulfils the requirements and restrictions of the problem it aims to solve (Simon, 1996). The use of real-life cases is a popular evaluation method used in DSR (Pries-Heje and Baskerville, 2008). Table 3.2 summarises a set of the most popular evaluation methods based on artefact types (von Alan et al., 2004).

### 3.4 Applying Design Science Research

To achieve the aim and objectives of this work, DSR described by Kuechler and Vaishnavi (2008) will be adopted as the overarching methodology. Artefacts will be produced in forms of constructs, models, methods and instantiations

**Problem Awareness** will be derived by conducting a comprehensive review and analysis of the related literature. The literature review is conducted to identify state of the art of the research problem and to understand the viability of the approaches that have been adopted to address the problem. Furthermore, the problem awareness phase serves as a means to explore the domain of interest and understand the concepts that have been applied to solve problems in the chosen domain. The result of exploring the problem awareness phase shows that the MSI needs more research and analysis concerning customer retention. Therefore, a pilot study will be conducted to seek an in-depth understanding of customer retention in the MSI.

**Suggestion** involves proposing possible ideas of how to address a research problem by designing an appropriate framework for addressing the problem. The idea in this phase is suggested after exploring the literature on customer retention in the MSI and realising a possible approach to address customer

<b>Artefact</b>	<b>Brief Description</b>	<b>Evaluation Criteria</b>
Construct	The conceptual vocabulary and symbols describing a problem within a domain	Completeness, clarity, elegance, ease of understanding and ease of use.
Model	A set of propositions or statements expressing relationships between constructs. Models represent situation as problem and solution statements.	Precision with real-world phenomena, completeness, level of detail, robustness, and internal consistency.
Method	A sequence of steps used to perform a task. A method can be tied to a particular model. A method may not be articulated explicitly but represents tasks and results.	Operationality (ability for the method to be reused), efficiency, generality and ease of use.
Instantiations	Application of constructs, models and methods to provide working artefacts.	Efficiency and effectiveness of an artefacts. Also, the influence of the artefact on its users and on the environment at large.

Table 3.2: DR Artefact Evaluation Criteria (von Alan et al., 2004)

retention in the MSI. This step emerges in iteration one with the adoption of a framework for understanding customer retention in the mobile services sector. More suggestions emerge in further iterations; for example when the SoMeDoA framework is applied to gain a broader understanding of customer retention in the first iteration, new knowledge (CSDs) was derived from the development and evaluation of adopting social media as a tool to understand customer retention. The derived knowledge was fed into further iterations.

**Development** is carried out by developing research artefacts in form of models, methods, and instantiations. The customer satisfaction determinants ontology (CSDO) is developed in iteration one which will lead to the development of further artefacts in subsequent iterations. The CSDO consists of the determinants of customer satisfaction in the MSI . The aim of the CSDO is to identify the determinants of customer satisfaction by applying techniques that will possibly provide an in-depth understanding of this domain. All artefacts that are derived from this research will be explained in section 3.5

**Evaluation** is conducted using evaluation techniques that measure the validity of the artefacts developed for each iteration. Validity approaches measure the effectiveness of the chosen frameworks utilised to analyse the research problem and determine how suitable the framework and approaches are in addressing the problem in the chosen domain. DSR evaluation seeks to examine the efficiency and generality of an approach utilised to address a research problem. Evaluating the performance of social media analysis to analyse customer retention resulted in further insights into the area of customer retention. These insights were fed into iteration 2 . Selected evaluation metrics are also used to validate the findings in further iterations in order to derive a better understanding of customer retention in the MSI.

**Conclusion** provides a summary of the iterations carried out to deal with



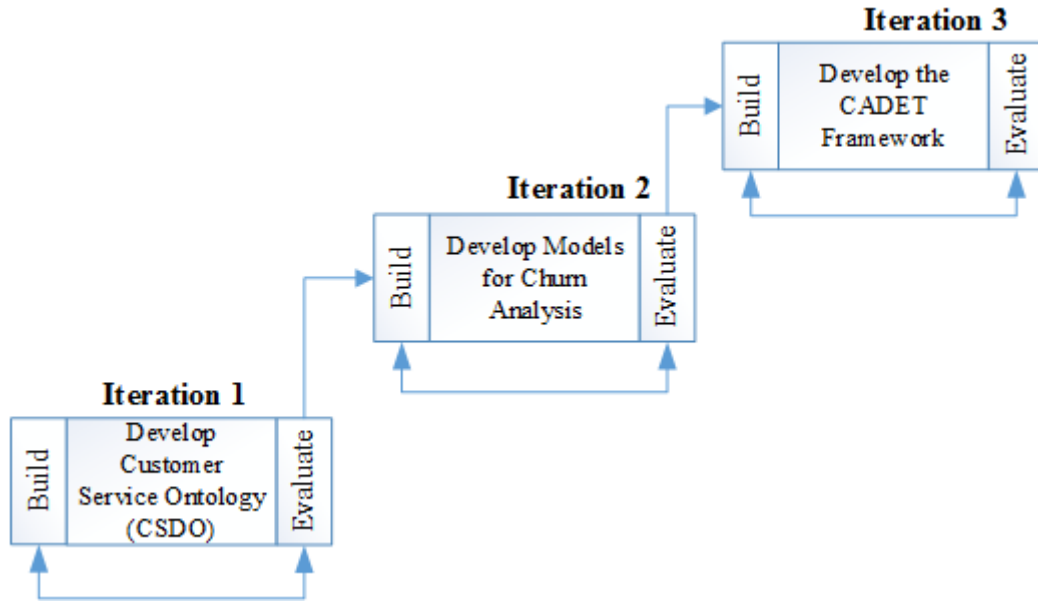


Figure 3.2: Research Iterations in line with DSR

the research problem. The research outputs along with the evaluation results are outlined. In addition, the limitations to the research and areas for improvement are also identified.

### 3.5 Research Iterations

DSR is conducted through iterative design cycles, which can be improvement iterations or improvement and additional iterations (von Alan et al., 2004). This research is conducted as improvement and additional iterations by which each iteration is carried out to extend and refine the design solution. The research is carried out in three design iterations. Each iteration provides knowledge that is applied to the next iteration. The three design iterations are used to deliver the final artefact which is the CADET approach to ABM.

The objective of iteration one is to uncover the determinants of customer satisfaction using social media data. The objective of iteration two is to investi-

gate and analyse the most popular data mining techniques for churn prediction in the MSI. The final iteration (iteration 3) provides a data-driven approach to modelling customer behaviour using decision trees and validating the approach with an experiment. Figure 3.2 presents an illustration of the three iterations and how they are used to produce the final design artefact. Each iteration involves a DSR cycle of building and evaluating the developed artefact in each iteration. The next section sheds light on the iterations conducted in this study.

### 3.5.1 Iteration 1

To meet objective two of this research, this iteration seeks to derive an in-depth understanding on the determinants of customer satisfaction in the MSI. This iteration will analyse customer data from social media specifically Twitter. Customer tweets will be captured and downloaded for a given period. A technique derived from the suggestion phase of Kuechler and Vaishnavi (2008) design science process will be applied to analyse the data. Traditional methods such as questionnaires and interviews, for analysing customer satisfaction were not selected for this study because these methods restrict participants to some questions; thereby not allowing them the flexibility to express in detail what they feel about their mobile services provider. The utilisation of the selected framework in this study provides more insights on customers perception about their mobile service provider. The selected framework gives customers the chance to express words that denote negativity or positivity to their mobile services provider in their chosen words. Customer data will be analysed by carefully scrutinising these words to derive the best possible results in regards to the objective of this iteration.

In order to investigate the factors that drive customer satisfaction, an em-

<b>Iteration 1</b>	<b>Aim-</b> Derive the determinants of customer satisfaction in the MSI using social media data.
Steps	Work done
Step 1	Capture and download tweets on selected companies: O2, T-mobile, ThreeMobile, Virgin and Vodafone.
Step 2	Import data into Nvivo10.
Step 3	Perform word frequency.
Step 4	Perform dataset processing by taking out words that are not useful to the research (see figure 4.9) and perform temporal separation (see section 4.3.2). Separate tweets as published daily and separate tweets published for each MNO.
Step 5	Visualise daily sentiments of tweets.
Step 6	Tabulate the daily sentiment scores for tweets (see table 4.3) and visualise the sentiment scores of the selected MNOs (see figure 4.10).
Step 7	Import processed dataset and perform temporal coding (section 4.4).
Step 8	Identify CSDs based on frequency of tweets.
Step 9	Conduct interview to validate findings.

Table 3.3: Steps for conducting iteration 1

pirical study will be conducted, using Twitter as a data source. Twitter is selected as it covers a wider audience than utilising the traditional methods of data collection. Also, Twitter is selected because it overcomes the limitations of restraining customers to a given set of questions using traditional methods. Data captured from Twitter will be analysed using the Grounded Theory Method (GTM) coding approach. GTM is the process of generating theory from data (Glaser and Strauss, 1971). Hence, the analysis will classify textual material (raw tweets) semantically and provide more significant and manageable data (Weber, 1990).

A thematic coding process will be used when analysing the data. Strauss and Corbin (1998) highlight that the coding process will help a comparison and relationship between elements with the aim of supporting the final model. The coding process is explained further by Strauss and Corbin (1998) as they describe the coding process into three categories; open coding, axial coding and selective coding. Open coding is the initial coding of the original dataset. Axial coding is putting together categories and subcategories into a hierarchy. Selective coding is the process of combining and refining categories to derive a theory (Strauss and Corbin, 1998). Strauss and Corbin (1998) summarised this whole process of coding as a series of activities that entail conceiving and formulating ideas with a logical, systematic and explanatory scheme.

Nvivo10, a software for qualitative analysis, will be used for organising and categorising raw tweets collected for this study. Nvivo10 is selected for analysing this study because it is deemed useful for carrying out content analysis, it is easy to use, and it proved to be error-free. In addition, it is stable in its operations and has export facilities which is required for this study. In addition, Nvivo10 proved suitable for manipulating and analysing the dataset captured for this study. The dataset will be captured into an excel spread

sheet. Nvivo10 is able to analyse data in excel format thus all data files will be imported from excel into Nvivo10 for analysis. Open coding is the first activity to be carried out using the GTM coding approach. All imported files will be analysed and significant tweets will be given a code (a base code in Nvivo terms). These base codes will be reviewed through a process of consolidation and words that have or appear to have the same meaning will be merged. Axial coding is the next step after open coding. Axial coding is used to review the remaining codes (free nodes in Nvivo10) and the codes that are found to be related will be merged with a parent node. The axial coding process is iterative in manner and will take account of changes in patterns and emergence of new relationships. This is referred to as 'constant comparison' which is a key feature of GTM as defined by Glaser and Strauss (1971). The key feature of GTM, 'constant comparison' relates to the design research process defined by Kuechler and Vaishnavi (2008).

This iteration will provide an artefact in the form of an onto-graph of the factors that drive customer satisfaction in the MSI. Also, the analysis in this iteration will demonstrate how customers perceive the services provided to them by their mobile services provider. From the literature, customer service, and price fairness were seen as the factors that drive customer satisfaction (Hanif et al., 2010). However, conducting this study using the selected methodology seeks to provide more insights into the problem domain. The artefact developed in this iteration is named the customer satisfaction determinants ontology (CSDO). This artefact is evaluated using interviews. Interview questions are structured to validate the findings derived from the DSR cycle. The next iteration analyses and investigates churn using three popular data mining techniques. Table 3.3 presents a tabular description of the steps taken to achieve iteration one.

<b>Iteration 2</b>	<b>Aim-</b> Assess the current and most widely-used machine learning techniques for analysing customer retention in the MSI highlighting their capabilities and limitations
Steps	Work done
Step 1	Data preparation (see sections 5.4.3 and 5.4.4).
Step 2	Apply selected machine learning techniques to datasets.
Step 3	Calculate accuracy score
Step 4	Validate by plotting ROC and lift curves.

Table 3.4: Steps for conducting iteration 2

### 3.5.2 Iteration 2

Data mining is a popular tool that is widely used in the MSI to study customer behaviour and predict future customer behaviour. Hung et al. (2006) defined data mining as the application of analytic and discovery algorithms to provide insights from raw data. Data mining is applied in this study to develop artefact in the domain of customer retention. Furthermore, data mining techniques can be applied to build models in selected domains. These models can be categorised into seven techniques (Ngai et al., 2009): association, classification, clustering, forecasting, regression, sequence discovery and visualisation. Decision trees, regression models and support vector machines (SVM) are widely used to study customer retention (Hung et al., 2006). Hence, decision trees, regression models and SVM are utilised in this study. Section 5.2 provides an overview of decision trees and regression models. The next section discusses data mining as a methodology used to conduct this study. Table 3.4 presents a tabular description of the steps taken to achieve iteration two.

### 3.5.3 Data Mining Development Cycle

To meet objective 3 of this research, three widely used data mining techniques were used to analyse and investigate churn. The data mining techniques are decision trees (CART and Random forest), logistic regression and SVMs. The details of these data mining techniques and the experimental results are presented in chapter 5. The study follows the widely used data mining development cycle CRISP-DM (Cross-Industry Standard Process Data Mining) proposed by (Chapman et al., 2000). The CRISP-DM methodology provides instructions on the use of data mining algorithms to address problems in selected domains of interest. The CRISP-DM is made up of six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. Business understanding involves understanding a company's business needs and applying suitable data mining applications to address the company's needs. Data understanding involves data extraction and understanding the variables and observations in the dataset. Data preparation involves processing the data such that it is in an appropriate format for applying the chosen data mining algorithm(s). Data modelling is the process of deriving a suitable model to solve the business problem. Evaluation involves applying selected measures or metrics to uncover how the designed model suits the business problem. The decision to adopt a data mining model should be taken after the evaluation phase. Finally deployment is the process of applying the developed model to the business problem.

Iteration two builds on the research carried out in iteration one. Iteration one uncovers CSDs in the MSI. CSDs identified from iteration one are customer service, coverage quality and price. The data mining churn prediction models were built specifically for each determinant i.e decision trees, logistic regression and SVM models were built for customer service, coverage quality and price.

The results of the models built for each determinant are compared to other data mining churn prediction models. This process was conducted for the two datasets used for this study, the twitter dataset and the dataset extracted from a telecommunications company's data warehouse (Telco dataset). The artefact derived from this iteration is the comparison of models built for each determinant. The results show the model that is most suitable to predict churn on each dataset, based on the derived determinants.

### 3.5.4 Iteration 3

Agent-based models comprise of three elements (Macal and North, 2010)

- A set of agents, including their attributes and behaviours.
- Agents relationship and the methods of interaction.
- Agent environment.

The elements enable agent interaction in an environment and provide details of the outcome of agents' interaction with the environment. Agents are autonomous and self-directed entities that are able to act independently without a need for an external force. Therefore, they are able to make independent decisions.

Agent-based modelling provides a way to model social systems that are composed of agents who interact with and influence each other, learn from their experiences and adapt to their behaviours in a way to better suit the environment (North, 2014). Agents have behaviours and these behaviours are often described by rules and interactions with other agents (Macal and North, 2010). Agent-based modelling provides a chance to observe the diversity of the rules and behaviours that apply to individual agents. In such experiments, agent-based models are run and agents within the ABM execute their



behaviours while adhering to the set rules. ABMS can reveal patterns and behaviours that may not be present during the ABM but arise during interaction with other agents. There are various methods for designing and implementing agent-based models. Macal and North (2007) in their paper discussed a variety of design methodologies and environments for implementation. Also, Marsh and Hill (2008) proposed a methodology for defining the behaviour of agents in an application for autonomous vehicles. In addition, Grimm et al. (2006) proposed a standard protocol for describing ABMs as a first step for implementing ABMS. ABMs should be designed in such a way that they can be easily interpreted, understood and re-used. To achieve objective 4 and 5 of this research, this iteration proposes a novel data-driven framework for describing ABMs. The proposed framework is named the CADET framework. The CADET framework is derived by applying a data mining technique known as decision trees. The decision tree algorithm is applied to one of the datasets used in this study, the telco dataset. The application of the decision tree algorithms derived from automated decision trees are used to describe the ABM (see figure 6.5). Section 6.6 sheds more light on the CADET framework. The CART algorithm implemented in iteration 2 is used to derive the decision trees used in this iteration. Subsequently, the decision tree produced using the CART algorithm is used to automate the description of agents in ABM. Agent-based modelling can be conducted using a variety of software (such as Netlogo and Repast) and programming languages (such as C, C++, Java and Python), or by using a specially designed software and toolkit that address the requirements and functionality of agent based modelling (North, 2014).

This phase of research also presents a novel tool, the TEA-SIM tool for conducting ABMS. The TEA-SIM tool was developed as part of the TEA-POCT project (Bell et al., 2016). It is a proprietary tool that was written in

<b>Iteration 3</b>	<b>Aim-</b> Demonstrate the effectiveness of the data-driven framework by conducting experiments into mobile service customer retention.
Steps	Workdone
Step 1	Develop a novel framework for Agent-Based modelling
Step 2	Validate the approach by carrying our an ABMS experiment using the TEA-SIM tool. see 6.6.1 for more details about the TEA-SIM tool

Table 3.5: Steps for conducting iteration 3

PHP. Section 6.4.3 sheds more light on the TEA-SIM tool. To achieve objective 5 of this research, the CADET framework is utilised along with the TEA-SIM tool to validate the CADET framework, while conducting an experiment on customer retention. The experiment is carried out by first, automating a decision tree to get customer values. These customer values are used to create agents and their attributes. To validate the CADET framework, 10 customer agents were fed into the TEA-SIM tool. These agents are either happy or sad customers. Happy customers become sad under certain conditions and vice-versa. Chapter 6 elaborates on the CADET framework and how it is used to address customer retention. Table 3.5 presents a tabular description of the steps taken to conduct iteration three. The next section presents the summary of this chapter.

### 3.6 Summary

This chapter provides a detailed description of the approaches and methods utilised to develop the artefacts in this study. This chapter also allows flexibility for a review of the methods used to derive the models in this research. A variety of datasets were used to achieve the artefacts developed in this study due to the multidisciplinary nature of this research. This research follows the DSR methodology, which facilitated the creation and evaluation of the artefacts developed to address the problem of customer retention.

The SoMEDoA framework, data mining and ABMS are the key research techniques used in the iterations that make up this study. The selection of these techniques to conduct this research is clearly articulated and justified. Each iteration in this study utilises the build and evaluate phases as described by (March and Smith, 1995) as the basic activities of design science. The build and evaluate phases are used to address customer retention in the MSI. In the first iteration, the SoMeDoA framework is used to understand the determinants of customer satisfaction in the mobile services sector. The artefact developed in this iteration is the customer satisfaction determinants ontograph (CSDO). Iteration two analyses and investigates churn based on the determinants derived from iteration one. This iteration compares data mining models and provides the best model for predicting churn based on each determinant. The final iteration proposes a novel framework for designing agent-based models. Also, it contributes to the field of customer retention by demonstrating the impact of social influence on customers decision to churn or remain with their MNO. In summary, this chapter illustrates how customer retention in the MSI is investigated using a set of techniques along with the guidelines for conducting DSR. The next chapter presents iteration one of this study which focuses on uncovering the determinants of customer satisfaction using twitter

data.

# Chapter 4

## Uncovering the Determinants of Customer Satisfaction

### 4.1 Overview

This chapter presents the first iteration of this study, which identifies customers' perception about their MNO using social media. Social media was chosen as a means to gather insights from a larger audience. While investigating the determinants of customer satisfaction using social media, this chapter seeks to make a contribution to the customer satisfaction literature by uncovering information that traditional methods such as interviews and questionnaires are unable to identify. Further insights into the problem domain were acquired during the analysis of this iteration. These insights were considered during subsequent iterations in order to achieve the overall aim of this research.

This chapter is structured as follows: Section 4.2 provides details on how the output artefacts for this study will be achieved. Section 4.3 describes the analysis carried out to generate customer satisfaction determinants (CSDs). Section 4.4 discusses temporal coding on the dataset, themes and sub-themes

derived from using the SoMeDoA approach. Section 4.5 presents the evaluation of the output artefacts. Finally, Section 4.6 provides an overall summary of the chapter.

## 4.2 Towards the DSR Output Artefact

To achieve the objectives set forth in this iteration, the SoMeDoA framework (Bell and Shirzad, 2013) is adopted to uncover the determinants of customer satisfaction in the MSI. The SoMeDoA framework was chosen because of its proven efficiency for extracting and analysing social media data (Bell and Shirzad, 2013). The SoMeDoA framework is based on the following steps (see figure 4.1). These steps make up the fundamental elements for the process of capturing and analysing data. Data from Twitter was extracted using specific search terms that relate to the MNOs under study. The data files generated from the search conducted on Twitter are analysed using a number of selected analytical and visualisation tools which will be discussed in further sections of this chapter.

## 4.3 Analysis and Results

This section reports on steps taken to derive CSDs in the MSI. The SoMeDoA approach developed by Bell and Shirzad (2013) was selected to conduct this study because it provides a comprehensive stepwise procedure for conducting social media data analysis.

<b>Phase</b>	<b>Description</b>	<b>Resulting Output</b>
Data Selection	Social media Web sites are selected as suitable sources for the domain of study	List of social media platforms and associated search terms.
Data Gathering	Data gathering tools are selected and run against the selected Social Media sites.	List of software tools.
Temporal Separation	Public information, news and communications are extracted in order to determine the public activities of organisations (with associated timelines)	DateTime lists files for each organisation.
Temporal Coding	Further analysis of temporal data to uncover topics of importance (with timeline)	Keywords lists and domain ontology. DateTime datalists for each keyword, list or category.

Table 4.1: The SoMeDoA Research Framework (Bell and Shirzad, 2013)

### 4.3.1 Dataset Description

Twitter was selected to carry out this study in order to efficiently detect customers' real time activities within the chosen domain. Twitter users are able to post tweets of up to 140 characters. As a result, users can express their feelings about products and services in a short text. The first part of this research begins by identifying the leading companies in the domain of interest. Data about the companies of interest were obtained by using their Twitter IDs as search/query terms (e.g. '@Vodafone OR #Vodafone, OR Vodafone'). Tweets about the following companies were captured: Vodafone, Virgin, Three, T-Mobile and O2. Tweetcatcher2 (an application developed as part of the MATCH project in Brunel University) was used to gather tweets and related data such as published date, user, number of followers, re-tweet count and sentiment analysis. Tweets were monitored and captured from the 27th of June 2013 to the 3rd of July, 2013.

### 4.3.2 Twitter Temporal Separation

In order to discover if there were changes in customers' feelings towards their mobile provider over time, tweets were categorised as they were published. A total number of 246,160 tweets posted about the selected organisations were captured and used for analysis. The extracted dataset about the selected companies were separated according to the days they were published to uncover the rate at which customers posted tweets about their MNO. Section 4.3.3 presents a discussion of how tweets were analysed daily, as this is the core element of temporal separation according to the SoMeDoA framework.



### 4.3.3 Tweets Per Day

The analysis of tweets per day was carried out by analysing and studying tweets to derive an insight on how frequent customers post tweets about their MNO. Figure 4.1 presents a graph of July 3rd, 2013 which had the highest number of tweets. The diagram displays the day tweets were published and the number of tweets published. The height of the column displays the number of tweets published from the chosen data set while the breadth indicates the date and the time of each tweet. The analysis shows that the highest rise in tweets occurred on Thursday 3rd July 2013, where a total number of 35,770 tweets were published. Subsequently, 35,752 tweets were captured on Tuesday, 2nd July 2013 (figure 4.4). Many of these tweets were about a popular musical artist who was performing in London. The day recorded with the least number of tweets is Saturday, 29th June 2013 (figure 4.2), where a total of 32,450 tweets were captured. The number of tweets may be as a result of poor coverage quality as many tweets captured on this day emphasised problems with network quality. Some Three mobile customers complained about their inability to make calls in certain locations. Poor coverage quality with respect to location was not investigated as that is not the objective of this study. The analysis also shows an interesting observation that occurred on the 1st of July 2013. O2 asked their customers to tweet about their overall customer experience. Many customers responded to this tweet by tweeting about the problems they have experienced with the network. These problems included lack of support and long waiting time in the queue to speak with a customer service adviser. O2 responded to these tweets by sending an apology tweet later that day. Vodafone customers experienced a good day as there were few complaints about their services. The 3rd of July had the highest number of negative tweets on most of the investigated network providers. The majority of negative tweets

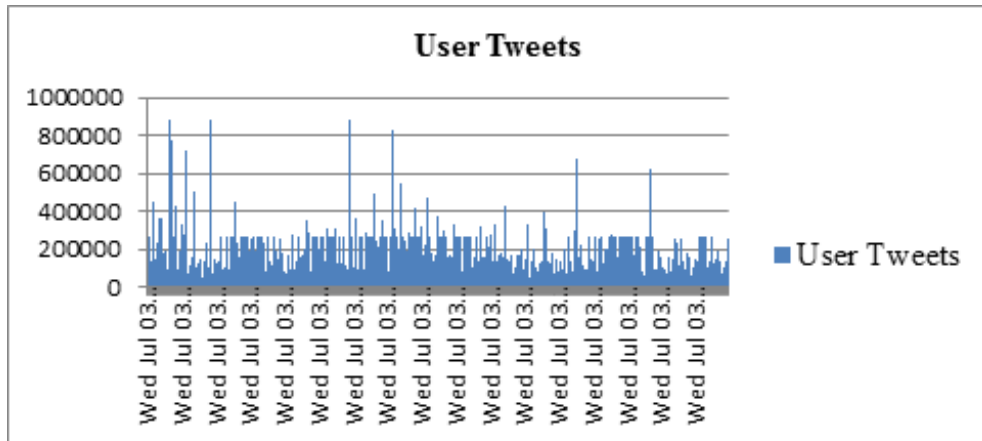


Figure 4.1: Tweet frequency from 8am to 8pm on Wednesday 3rd July.

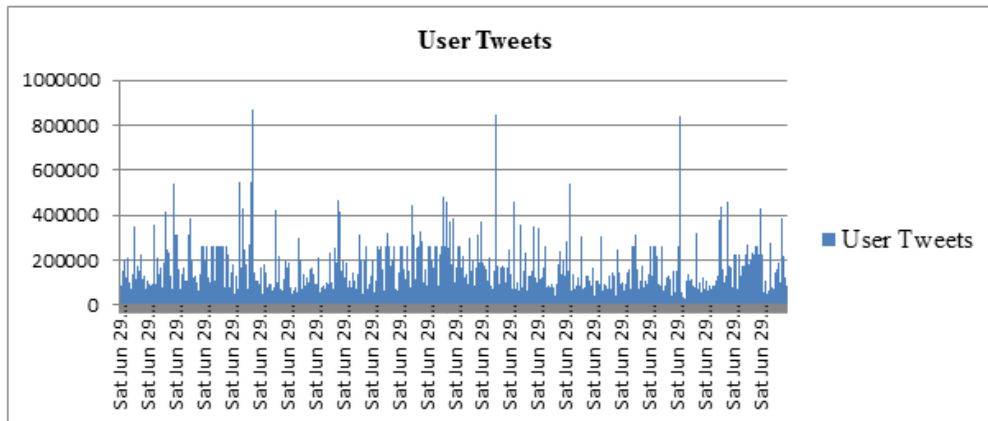


Figure 4.2: Tweet frequency from 8am to 8pm on Saturday 29th June

emphasised on poor customer service from T-Mobile and O2.

Table 4.2 displays a table of senti-positive, senti-neutral and senti-negative scores of the MNOs that were studied. The table shows that senti-neutral has the highest scores while senti-negative has the lowest scores. This result shows that majority of tweets captured have a neutral sentiment. The next section presents a brief summary of sentiment analysis and sentimental average per day for the dataset under study.

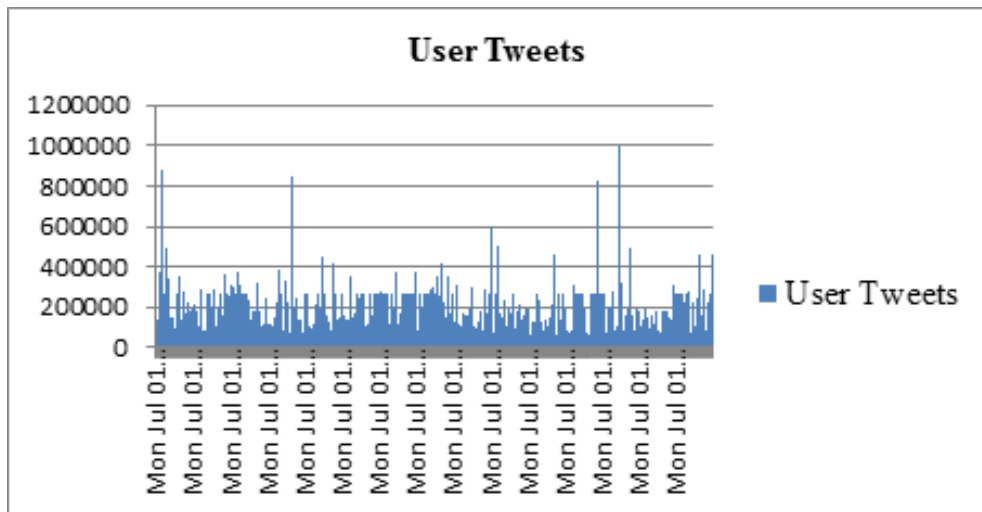


Figure 4.3: Tweet frequency from 8am to 8pm on Monday 1st July

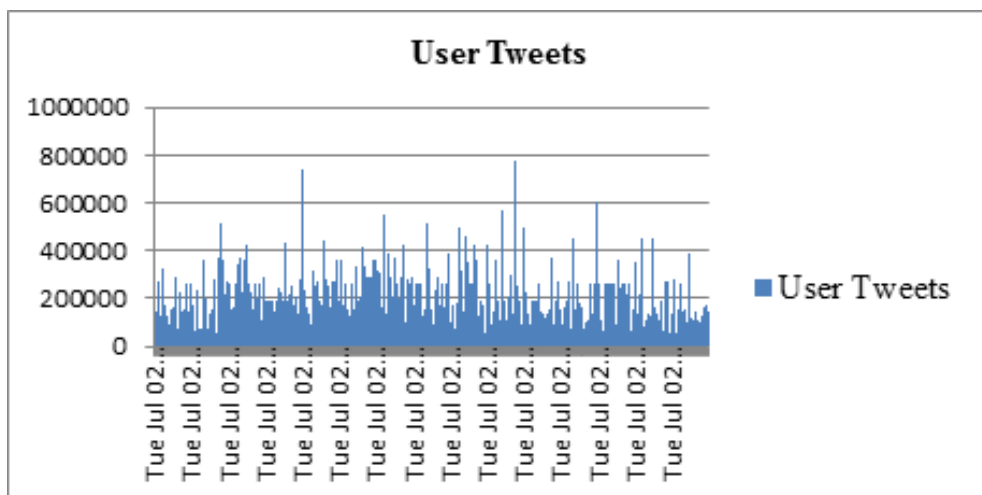


Figure 4.4: Tweet frequency from 8am to 8pm on Tuesday 2nd July

Mobile Provider	Senti-Positive	Senti-Neutral	Senti-Negative
T-Mobile	31%	41%	28%
O2	29%	48%	23%
Three	31%	47%	22%
Virgin	31%	46%	26%
Vodafone	28%	50%	22%

Table 4.2: Distribution of sentiment score for MNOs

#### 4.3.4 Sentimental Average Per Day

Sentimental average per day is the next phase for temporal separation according to the SoMeDoA framework. The sentiStrength 7 tool developed by (Thelwall et al., 2010) and implemented in Brunel University's Tweetcatcher was used to assign sentiment scores to each tweet. This tool simultaneously assigns positive and negative scores to words in a tweet with the idea that users can express both types of sentiments at the same time; for example, 'I love you, but I also hate you' (Kucuktunc et al., 2012). Positive sentiment scores range from +1 to +5, indicating not positive to extremely positive. Similarly, negative sentiment scores range from -1 to -5, indicating not negative to extremely negative (Kucuktunc et al., 2012). In order to get the final positive or negative sentiment score for a piece of text, the final positive or negative score is calculated by extracting the maximum score from all individual positive scores. The negative sentiment strength is similarly calculated (Kucuktunc et al., 2012). From the analysis, the majority of tweets captured were assigned a neutral sentiment score (0). Also, the analysis show a high percentage of slightly negative and slightly positive scores. Tables 4.2 present the percentage scores for each service provider within the chosen time slot.

Timeslot	Senti-Positive	Senti-Neutral	Senti-Negative
27th June 2013	28%	50%	22%
28th June 2013	30%	66%	24%
29th June 2013	28%	51%	21%
30th June 2013	29%	47%	24%
1st July 2013	31%	48%	21%
2nd July 2013	30%	47%	23%
3rd July 2013	28%	47%	25%

Table 4.3: Distribution daily sentiment scores

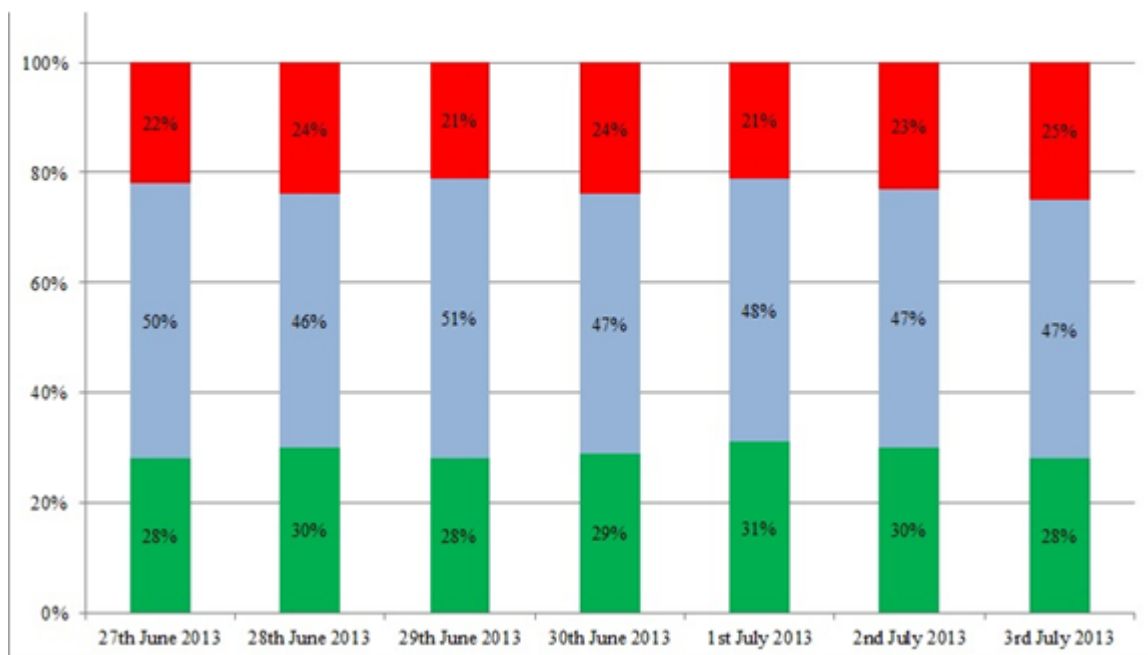


Figure 4.5: Daily Sentiment Analysis scores

Table 4.3 displays a tabular description of the distribution of daily sentiment scores in percentage for the period of study. Overall, there were more neutral sentiments (0). These sentiments were between 47% and 51%. Positive sentiments were between 28% and 31% while negative sentiments were between 21% and 25%. Figure 4.5 presents a graph of the daily sentiment analysis for the dataset. The colours green, blue and red represent positive, neutral and negative sentiment scores respectively. To gain a deeper understanding of the contents of the tweets, temporal coding, which is the next phase after temporal separation was conducted. Temporal coding was carried out by using Nvivo10 to uncover CSDs

## 4.4 Temporal Coding

Temporal coding is the final phase of the SoMeDoA framework. This phase incorporates analysis of tweets per word and sentiment analysis of the key words found in tweets. Figure 4.6 presents a graphical representation of the process of importing and categorising tweets. Subsequently, the next section provides a detailed description of tweets per word.

### 4.4.1 Tweet Per Word

The next step for carrying out temporal coding on the dataset is analysing tweets per word. Tweets captured for this study were scrutinised and interpreted by undergoing a coding process to achieve the best result. The SoMeDoA approach, which was adopted to analyse tweets makes use of GTM coding approach (see Chapter 3).

Analysis of tweets was carried out by counting the word frequency of tweets using Nvivo10. In the first round of Nvivo10 analysis, the most frequent words

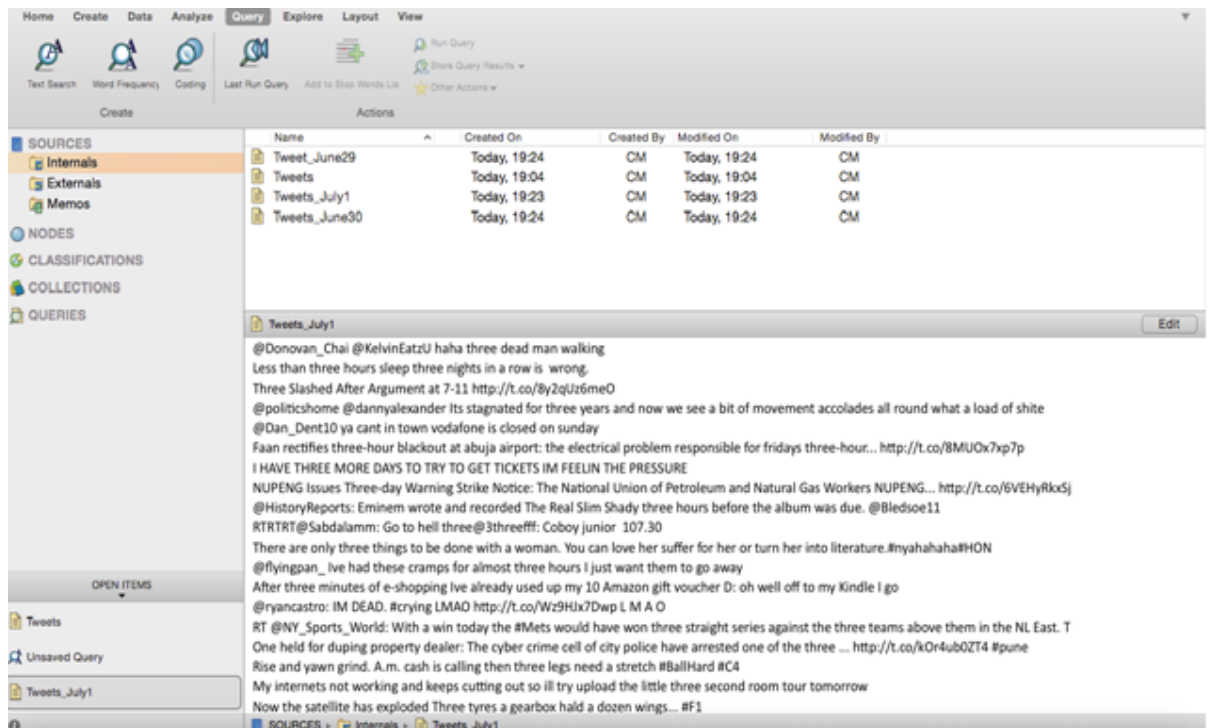


Figure 4.6: The process of importing and categorising Tweets into Nvivo10

about active acts all am pamp associate avi awards been being between cena change  
 characters cities **CO** communication completely **contest** could days  
**direction** dont elements embarrassingly fluid from get go god got  
 had haiku has have her hieroglyph his hours how **i im its** ive just later leader letter like  
 lol make **me** months more much **my** never now o2 old our period point politician raise  
 royalmumble **rt** ryback same section she sheamus so **state** syllables text than thats  
 them those **three** times tonight u united up **virgin** we week were what when  
 who why words would wrestlemania yeah **you** your

Figure 4.7: Results derived from the first round of performing word frequency on tweets



Figure 4.8: Results derived from the second round of performing word frequency on tweets

in the dataset included ‘rt’, ‘direction’, ‘three’, ‘you’, ‘contest’, ‘my’, ‘virgin’, and ‘days’ (see figure 4.7). From these list of words, only ‘three’ and ‘virgin’ were relevant to this study. Hence, tweets about ‘three’ and ‘virgin’ are reviewed. Interestingly, many of these tweets are not about Three mobile and Virgin as a service provider. Thus, they are not relevant to this study and they were separated from other tweets. The irrelevant ‘three’ and ‘virgin’ tweets were identified by carrying out a word search on Microsoft Excel and reading through the tweets. During this process, tweets about each MNO were saved in a separate file. Sentiment analysis was further carried out on the tweets about the selected MNOs to uncover the satisfaction level of customers with their MNO (see table 4.2)

After separating the Twitter files according to each MNO, the files for each MNO were loaded into Nvivo10 and a word frequency count was conducted. The second round of the word frequency count using Nvivo10 revealed more



words that are related to the topic and domain under study. These words include ‘digital’, ‘O2’, ‘direct’, ‘communication’, and ‘4G’. The bolder the word, the more frequent it is in the dataset (see figure 4.8). Based on the frequency of words in this analysis, the themes and subthemes, which determine the CSDs in this study, are derived. Themes in this context refer to possible words that denote satisfaction or dissatisfaction, and subthemes are words that are similar in meaning to themes. The derivation of CSDs from themes and subthemes are explained in the next section.

#### **4.4.2 Reporting Determinants with an Ontology-Based Concept Network**

The SoMeDoA approach was used to derive the determinants of customer satisfaction in this study. The SoMeDoA approach incorporates GTM, which is used to analyse the textual content of tweets, following the open coding and axial coding rule discussed in chapter 3.5.1. The result derived from the axial coding content resulted in some categories and subcategories. The categories and subcategories are referred to as themes and subthemes respectively in this study. The themes and subthemes include payment, bills, support, data, upgrade, annoying, appalling and bad experience. For example, a price category was created with associations to payment, bills, credit and extra charges. Figure 4.9 presents the process of storing and categorising datasets. Subsequently, the categories and subcategories are put together and an ontology was created using the Protégé 4.3 Ontograf (see figure 4.10).

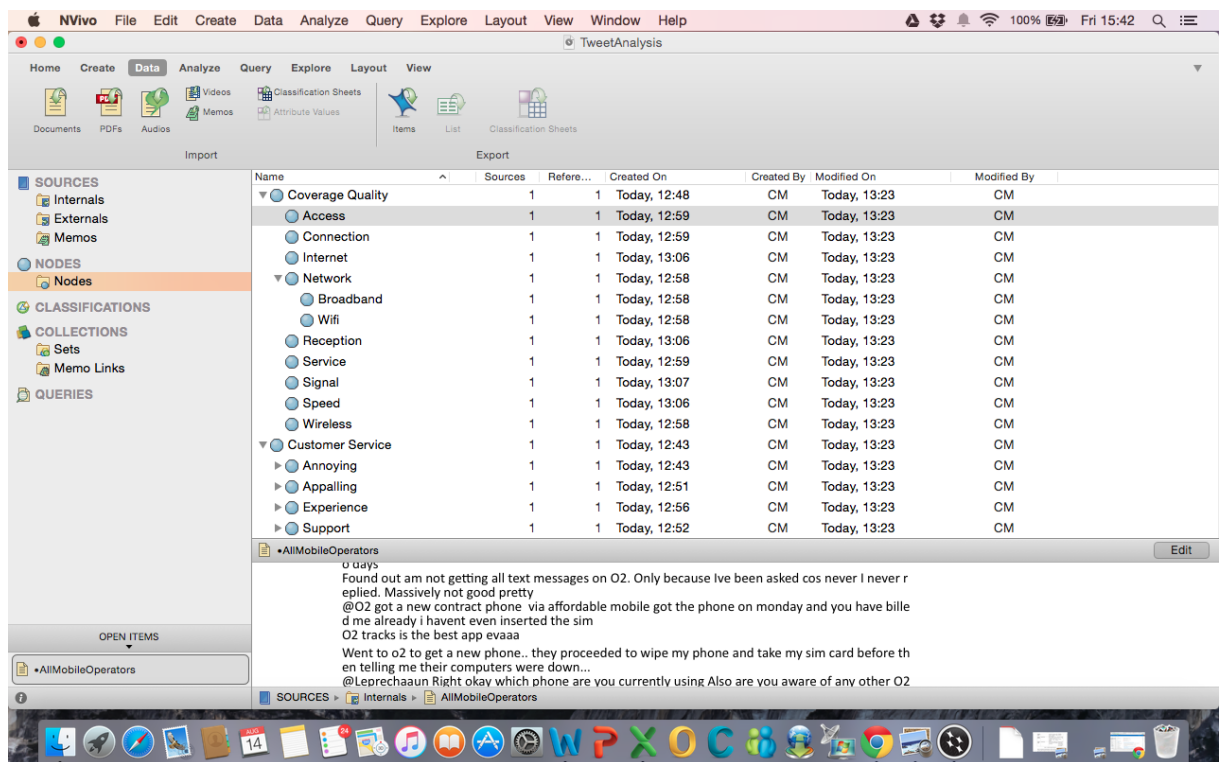


Figure 4.9: The process of deriving CSDs

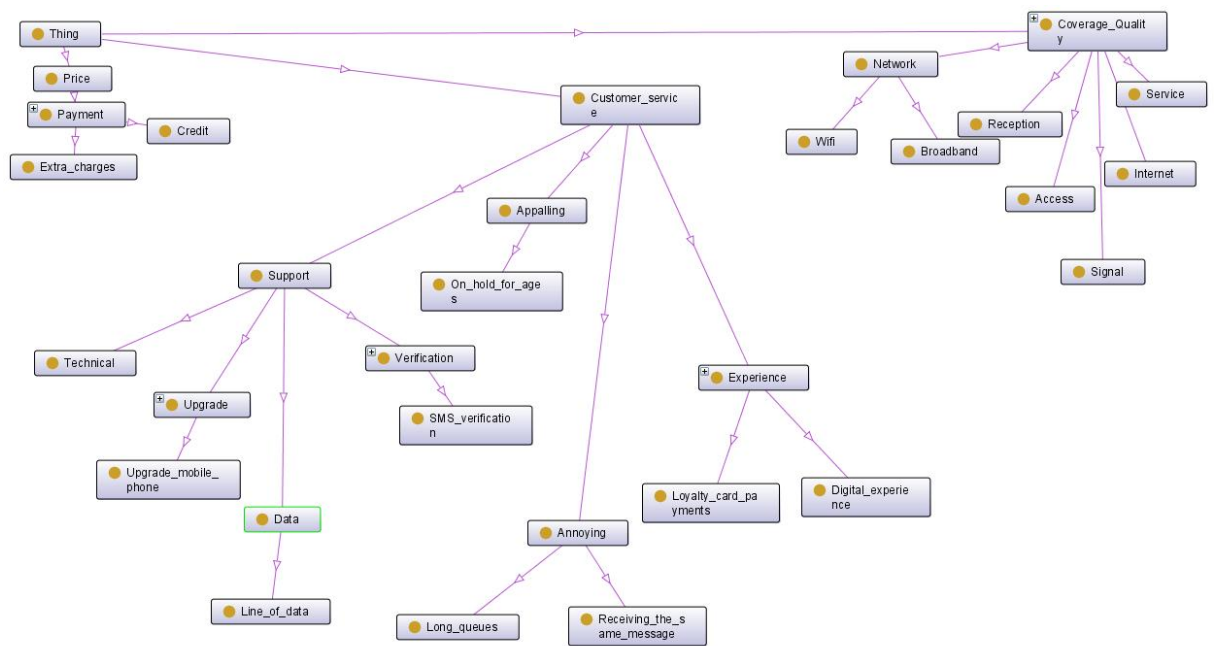


Figure 4.10: Customer satisfaction determinants ontograph (CSDO)- An ontograph that shows the relationship between CSDs and related words

### 4.4.3 Sentimental Average Per Word

Table 4.4 presents the main nodes that are classified as ‘themes’ in this analysis. The main nodes in Nvivo10 term are referred to as key codes. Coverage quality and customer service are the key codes displayed on figure 4.9. The themes are derived from the key codes and they make up the CSDs. Sentiment analysis was carried out on the identified CSDs to determine the number of positive and negative tweets that were published under each determinant. Table 4.4 shows that price has the highest percentage of positive sentiment. The high percentage could be because customers who are on a contract with their MNO are already aware of the monthly price plan of the package they have chosen. The senti-negative percentage for price is 42%. This percentile score is the lowest amongst the determinants, but it is also a relatively high score. From the sentiment analysis of each CSD, customer service and coverage quality are of high importance because they both have a higher number of tweets than price. As seen in Table 4.4, customer service has a positive percentage of 55%, and coverage quality has a positive percentage of 56%. These scores do not indicate that customers had a better coverage during the period of analysis. This could be due to the number of tweets published about each CSD. A total of 21,742 tweets were published regarding customer service while a total of 19,055 tweets were published about coverage quality. From these published tweets, customer service had a total of 12,055 positive tweets, coverage quality had a total of 10662 published tweets. Subsequently, customer service had a total of 9,687 negative tweets and coverage quality had a total of 8,393 negative tweets. In order to determine how reliable the SoMeDoA approach is in uncovering the determinants of customer satisfaction, an evaluation procedure was conducted. The aim of the evaluation procedure is to validate the findings derived from the application of the SoMeDoA framework to uncover CSDs. The evaluation

procedure also seeks to uncover more determinants of customer satisfaction

## 4.5 Evaluation

In line with the DSR methodology, this study has to be evaluated. The SoMe-DoA approach is a novel approach for investigating the determinants of customer satisfaction in the MSI, and it is important to validate the findings derived by using this approach. As displayed in figure 4.4, customer services, coverage quality and price were found to be the determinants of customer satisfaction. In order to validate the findings derived in this study, a set of interviews were conducted. The three basic approaches to conducting interviews include structured, semi-structured and unstructured (Oates, 2005). Structured interviews are based on planned, standardised and identical questions for every interviewee. Semi-structured interviews are conducted with a fairly open framework, which allows for focused, conversational, two-way communication. Unstructured interviews are not specifically limited to a set of questions to allow the conversation flow freely. In this study, the semi-structured interview approach was adopted. A random sample of four customers for each MNO used for this study was interviewed. A total of 20 customers were interviewed. Before commencing with the interviews, the interview questions were tested with three participants. These participants shared their thoughts on how to make the interview questions more focused to the research objective. The interview questions were re-written in regards to the thoughts shared by the participants. After reviewing interview questions, two interviews were further conducted to test the new set of questions. Again, more thoughts were shared on how to make the questions better and more focused to the research objective. Interview questions were designed to validate the determinants of

Table 4.4: Senti-Average of the derived Themes (CSDs)

Name	#Senti- Pos.	#Senti- Neg.	Senti- Pos. Percentage	Senti- Neg. Percentage
Customer service	12055	9687	55%	45%
Coverage Quality	10662	8393	56%	44%
Price	9292	6716	58%	42%

customer satisfaction identified using the SoMeDoA approach; customer service, coverage quality and price. The questions were asked in such a way that the participants could express their general views on the determinants of customer satisfaction. Overall, the interviewed participants emphasised on customer service and coverage quality to be their main determinants of customer satisfaction. When a participant was asked why they have been with their mobile service operator for so long, they responded ‘They have a brilliant customer service, and they also have good network quality’. Another participant talked about their experience with the customer service department of their MNO. They highlighted that ‘Customer service is very important as one can get upset if their problems are not resolved’. Apart from customer service and coverage quality, participants also emphasised the importance of price as a factor for customer satisfaction. One major finding of this factor is that most of the interviewed participants already knew the cost of the package they signed up for. However, a few participants complained about incurring unexpected charges. When a participant was asked if price is a major factor for them to be satisfied with their MNO, they responded by saying, ‘Price is a major issue for me with my mobile operator. I do not like when I am charged, and I do not know why I am charged. I chose my mobile operator because they offered a cheaper price than other competitors. I guess I made a mistake, I should have considered their reputation before choosing them’. Another customer also emphasised on reputation as one of the factors that make them satisfied with their mobile provider. When they were asked ‘To the best of your knowledge which mobile operator stands out at providing the best network/coverage quality and what has the company done to make you feel that way?’ Their response was, ‘O2. My friend has been with them for 5 years. He always talks about their good customer service and network quality.

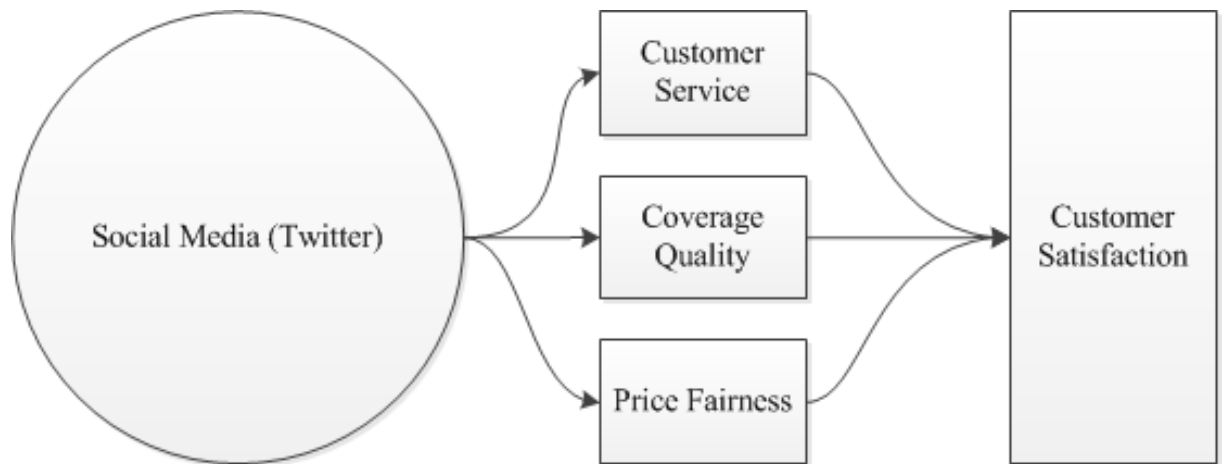


Figure 4.11: Customer satisfaction determinant model

He also praises them because of their reputation and their advertisements'. Value added service is another factor found to be a determinant of customer satisfaction. A Virgin customer highlighted that he is satisfied with his MNO because they have given him discounted prices and incentives (such as, Virgin broadband and TV). Although, majority of interviewed participants' emphasised on customer service and coverage quality to be their major determinants of customer satisfaction, the interviews conducted also found reputation, and value added services to be part of the determinants of customer satisfaction in the MSI.

## 4.6 Summary

This chapter utilises the SoMeDoA framework for uncovering the determinants of customer satisfaction in the MSI. The SoMeDoA framework was chosen to conduct this study because its application is capable of improving the domain knowledge of customer satisfaction in the MSI. The SoMeDoA framework's capability to incorporate concepts and semantic relations from social networks



assists in uncovering hidden patterns from the area under study. This chapter contributes by employing the SoMeDoA approach to uncover the determinants of customer satisfaction in the MSI. Another contribution derived from this chapter is the evaluation of the SoMeDoA framework adopted for uncovering CSDs. The evaluation process conducted using semi-structured interviews uncovered more underlying factors that determine customer satisfaction. The SoMeDoA approach proved efficient in extracting domain concepts, linking them together and providing reasonable precision and coverage. Finally, the learning and concepts that emerged from this chapter can be further researched, and the process by which this research was conducted can be applied in another domain. The next chapter seeks to apply the most widely-used data mining techniques for analysing churn to analyse customer churn based on the established CSDs.

# Chapter 5

## Machine Learning Techniques for Churn Analysis

### 5.1 Overview

This chapter presents the second iteration of this study, which aims to examine and evaluate the current customer retention analysis techniques. A novel approach is taken to analyse churn based on the CSDs derived from chapter 4. Analysis of CSDs is carried-out using the most widely-used churn modelling techniques. Decision trees, logistic regression and support vector machines (SVM) are used to build customer churn models on two datasets. Details of datasets used are presented in this chapter. The results derived from building churn prediction models using decision trees, logistic regression and SVM are compared to uncover the best model for predicting churn on the identified CSDs. A discussion of the limitations of the techniques used for churn prediction are presented.

This chapter is structured as follows. Section 5.2 provides brief explanation of the selected churn modelling techniques: decision trees, logistic regression

and SVM. Section 5.3 describes selected classification model evaluation metrics. Section 5.4 presents churn modeling experiments and results Section 5.5 presents the limitations of data mining. Finally, section 5.6 provides a summary of the chapter.

## 5.2 Churn Modeling Techniques

Different approaches have been adopted to investigate customer churn in the MSI (see section 2.3). Data mining techniques have emerged as powerful platforms for analysing and investigating customer churn. Decision trees, regression analysis and SVM are among the most widely used techniques for predicting customer churn (Huang et al., 2012). As a result of the widespread usage of these three techniques for investigating churn in research and academia, they are applied in this study for empirical analysis. The next three sections discuss decision trees, logistic regression and SVM.

### 5.2.1 Decision Trees

Decision trees are widely used prediction models. They are sequential models that logically combine a series of simple tests; every individual test compares a numeric attribute against a threshold value or a nominal attribute against a set of possible values (Kotsiantis, 2013). Decision trees consist of many nodes and branches in different stages and various conditions. It is a popular technique for solving classification problems because it can reveal rules that make up an outcome or a result to a problem (Tsai and Lu, 2010). Decision tree models are commonly used to solve classification and prediction problems where instances are classified into positive and negative instances. In the case of this study, instances are classified into churn or no-churn. Decision tree models are usually

represented in a top-down manner (see figure 5.1). The main aim of a decision tree is to derive a tree which solves a particular business problem and is easy to understand. To achieve such result, the decision tree undergoes two stages; tree building and tree pruning. Tree building is carried out using top-down strategy (also known as a divide and conquer strategy). The process of tree building involves:

- Selecting the attribute for the root node.
- Splitting instances into subsets.
- Repeat recursively for each branch.
- Stop if all instances have achieved the same class.

A root node is selected by comparing the number of bits (splitting based on information gain) for possible root nodes and choosing the node that has the most bits of information. After selecting the root node, the next step is to look at the branches that emanate from the root node. The process of selecting the node with the most bits of information is repeated. This process continues until all instances have the same number of bits i.e. when there are no more classes to split on (accuracy is 100%). The tree pruning process involves eliminating error-prone branches. A pruned tree can improve a classifier's performance and can facilitate further analysis of the model for the purpose of knowledge acquisition. The pruning process should never remove predictive parts of the classifier. For better understanding of how decision trees work, see the example below:

If (Contract length  $\geq 24$  months, Location = midlands, priceplan  $< 30$  and data  $< 4$ gb), then Churn = Yes.

A number of algorithms can be used to build decision trees. These algorithms

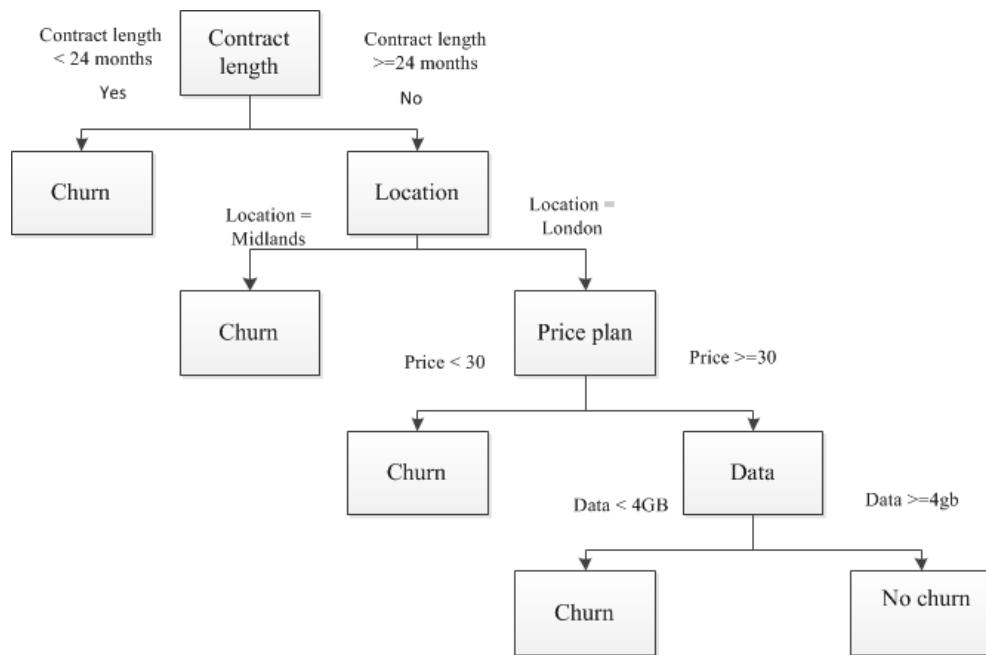


Figure 5.1: Decision Tree Classification Process

include CART (Classification and Regression Trees) and random Forest.

The random forests algorithm was introduced as a solution to one of the limitations of decision trees i.e. a change in data may affect the overall quality of the prediction model (Breiman, 2001; Breiman et al., 2001). Random forest uses a subset of 'x' randomly chosen predictors to grow each tree on a bootstrap sample of the training data. Ideally, this number of selected variables (i.e. x) is much lower than the total number of variables in the model. Each tree votes the most popular class from the large number of trees generated. By accumulating the votes over the trees that participated in voting, each tree votes for the most popular class. The random forest classification technique is popularly used when there are more variables than observations. Decision trees have many advantages; in the binary classification context, they are popular because they are easy to interpret and understand (Duda et al., 2012). They have the ability to handle covariates, which are measured at different measurement

levels (Burez and Van den Poel, 2008) and they can process both numeric and categorical data. However, decision trees have some limitations; they have a high degree of instability (Friedman et al., 2001). A small change in data can result in different series of splits, which often affects the quality of prediction when validating the trained model (Burez and Van den Poel, 2008).

### 5.2.2 Logistic Regression

Logistic regression is used to analyse a dataset by providing a probability score for the predictor variable based on the variables and observations of a dataset. Logistic regression has been applied extensively to study customer churn and it has also proven to be efficient in individual customer marketing and churn prediction (Neslin et al., 2006). In order to apply logistic regression to churn analysis, a logistic regression equation needs to be derived from the dataset. An evaluation process for each customer dataset is performed once a logistic regression equation is obtained. If the probability value for a given set of customers in a dataset is greater than a defined threshold (e.g. 0.5), these customers are likely to churn. Conversely, if the probability value of customers in that same dataset is less than (e.g 0.5), the customers are not likely to churn. An example of applying logistic regression to churn analysis is shown below:

A MNO's dataset consists of the following elements: target variable (churn) and predictor variable (age, gender, location, problem occurrence with dataset, and tenure). Table 5.1 presents the records of 7 customers with their associated target variable values, predator variable values and p.value. The table shows that customers with a probability value greater than 0.50 churned and customers with a probability value less than 0.50 did not churn.

Age	Gender	Location	Problem occurrence with handset	Prob. Value	Churn
25	M	London	2	0.72	Yes
28	F	Birmingham	2	0.78	Yes
30	M	Surrey	1	0.44	No
32	M	Staines	0	0.66	Yes
33	F	High Wycombe	0	0.59	Yes
27	F	Exeter	1	0.43	No
26	M	Durham	0	0.42	No

Table 5.1: An Example of Logistic Regression on Customers and their probability to Churn

### 5.2.3 Support Vector Machines

SVMs are supervised models with associated learning algorithms that analyse data and recognise patterns used for classification and regression analysis. An SVM classifier is trained by finding a maximal margin hyperplane in terms of a linear combination of subsets (support vectors) of the training set (Huang et al., 2012). Given a set of training examples (for example the Twitter dataset used in this study) as points in space, SVM maps these datasets into separate categories and divides them by a clear margin within the given space. When new examples are loaded into the model, the SVM training algorithm builds a model that predicts each example to belong to one of the categories already created based on the side of space they fall on. SVM is efficient for performing linear classification. SVM can also effectively perform non-linear classification using the 'kernel trick', implicitly mapping their inputs into high-dimensional features. The accuracy of the SVM model depends on the selection of the kernel parameters. In this study, the kernel parameter was selected by k-fold cross validation technique. SVM is a powerful prediction tool which has many benefits over traditional prediction tools. One of the most important benefits of SVM is that the solution of any problem relies on a small subset of the dataset which gives SVM a great computational advantage (Rodan et al., 2014). In addition, SVM aims to minimize the upper bound of generalization error instead of minimising the training error (Vapnik and Vapnik, 1998).

## 5.3 Classification Model Evaluation

There are several existing methods for modelling churn (see section 5.2). However, one challenge for using these methods are how efficient the performance methods are and how to compare the performance amongst competing models.



<b>Model</b>	<b>Target</b>	
	<b>Churn</b>	<b>No-Churn</b>
Churn	TP(true Positive)	FN (False Negative)
No-Churn	FP(False Positive)	TN (True Negative)

Table 5.2: Confusion Matrix

It is often difficult to build a perfect classification model that will correctly classify all the values from the test set. Therefore, it is important to choose a model that best works for the problem domain under study. This section sheds light on some classification evaluation metrics and comparison methods of models.

### 5.3.1 Binary Classification

A confusion matrix can be used to measure the accuracy of a binary classification model. When attempting to solve a binary classification problem, one class can be labelled as positive, and the other class can be labelled as negative. For the purpose of this study, the positive class would be labelled as churners (TP), while the negative class would be labelled as non-churners (FP). The objective of applying a confusion matrix is to get a high TP and a low FP. Table 5.2 shows a confusion matrix, where TP is the number of correctly predicted churners and FP is the number of incorrectly predicted churners. FN is the number of incorrectly predicted non-churners and TN is the number of correctly predicted non-churners. A confusion matrix (or a contingency table) is a clear and unambiguous way to present the prediction results of a classifier.

A number of accuracy metrics (such as classification accuracy, sensitivity and specificity) can be derived from a binary classifier confusion metrics.

### 5.3.2 Classification Accuracy

Classification accuracy is the percentage of correct predictions obtained when a model is applied to a dataset. To further explain classification accuracy, an example is given below:

A mobile services company has a customer base of 20,000 customers. 80% of customers from the customer base are customers who have churned and the other 20% of the customer base are customers who have not churned i.e. customers who stayed with their MNO. A classification model is applied to predict churn with this customer base and the model predicts correctly. This implies that the model has a classification accuracy of 99.9%. This accuracy rate means that the model predicts churn correctly. However, this can be misleading because the model did not predict the customers who did not churn. Predicting churn correctly is more beneficial for a company than predicting churn wrongly. If a model correctly predicts customers who are likely to churn, resources can be invested to those customers in order to retain them. The above example illustrates one of the limitations of applying classification accuracy alone to imbalanced data. Sensitivity and specificity are able to overcome some of the limitations of accuracy metrics.

### 5.3.3 Sensitivity and Specificity

In churn prediction, sensitivity is the percentage of churners that are predicted as churners and specificity is the percentage of non-churners that are predicted as non-churners. MNOs prefer models that have high sensitivity percentage over models that have a high specificity percentage. The reason behind the preference is because; the cost of detecting a churner is higher than the cost of detecting a non-churner in a customer retention campaign (Verbeke et al.,

2011). However, it is also important to consider models with high specificity because customers who fall into this category may eventually churn. Reaching a compromise on the percentage of sensitivity and specificity of a model is helpful to retain as many customers as possible while effectively managing the marketing budget in the customer retention campaign.

### 5.3.4 Receiver Operating Characteristics

The ROC curve is an effective way to visualize a classifier's performance in order to select a suitable operating point or decision threshold. The ROC graph is defined by Burez and Van den Poel (2009) as  $x = 1 - \text{specificity}(t)$ ,  $y = \text{sensitivity}(t)$ . Every binary classifier (for a given test set of values) is defined by a point (1-specificity, sensitivity) on the graph. By altering the threshold of the probabilistic classifier, a set of binary classifiers represented with a set of points on the graph are obtained. In the case of churners and non-churners, ROC shows the relationship between correctly predicted churners (TP), and non-churners predicted as churners (FP). Generally, a model with a good performance has its curve passing through the top left side of the ROC curve. The closer the curve is to the top left corner, the better the model. The best model is one which passes through or is close to (0 and 1). This indicates that the model has a sensitivity of 100% (no FN) and specificity of 100% (no FP). Probabilistic models such as the logistic regression model are incapable of providing binary class decisions (for example churn or no-churn), instead they produce a rank or a score. In this case, thresholds need to be used to provide a binary classifier. If the classifier's output is greater than a 'set threshold', the classification class will be churn. Otherwise, the classification class will be no churn. It is difficult to use the ROC technique to evaluate the set of pairs that are from different prediction modeling techniques or feature subsets

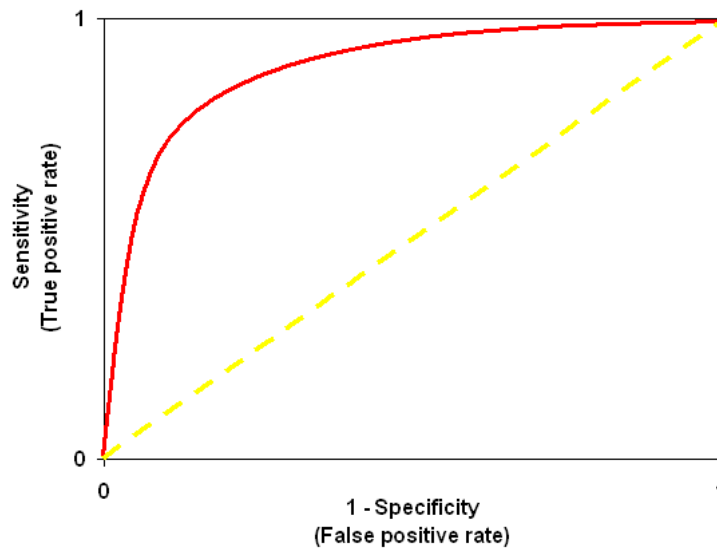


Figure 5.2: An example of the ROC Curve

of data (Huang et al., 2012). The area under the ROC curve (AUC) was introduced by (Bradley, 1997) to overcome this problem. The AUC is also a popular technique that is used to calculate the performance metrics of models in the MSI. The AUC is equal to the probability that the prediction model will correctly differentiate between instances of churners and non-churners. The AUC score varies from 0 to 1, and models with a higher score are usually considered to have a better performance. Figure 5.2 presents an example of an ROC curve. The next section presents a description of the lift chart.

### 5.3.5 Lift Chart

The cumulative lift chart is a popular technique used to calculate the effectiveness of a prediction model. The lift chart is explained with an example below:

The marketing department of a telecoms company is planning to send advertisements to selected customers with the goal of boosting sales of the service.

Service	Subscribe	No-Subscribe
Subscribe	TP(benefits)	FN
No-subscribe	FP(costs)	TN

Table 5.3: Confusion Matrix for lift chart example

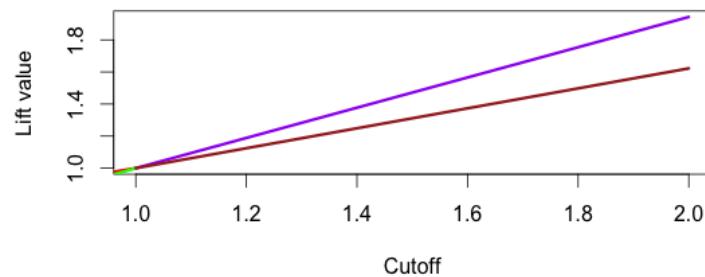


Figure 5.3: An example of the Lift Curve

The department has a list of all customers where each customer is described by a set of attributes. Each advertisement sent costs a few pounds, but the company will be at profit if the customer subscribes for the service. Therefore, the marketing department wants to minimise the number of advertisement sent, while maximising the number of sold services by reaching only customers who will subscribe for the service. The marketing company develops a classifier that predicts the probability that a household will subscribe to the proposed service. The lift chart can be used in this scenario to fit classifiers and to express the dependency between the costs and the expected benefit. Table 5.3 presents a confusion matrix that explains this scenario.

In practice, TP is difficult to measure but the lift chart can help the company identify customers that fit in this bracket i.e. households where the advertisements should be sent.

## 5.4 Churn Modelling Experiments and Results

This section presents experiments on churn analysis and prediction in the MSI. The experiments were conducted using three modelling methods: decision trees, logistic regression and SVM. The experiments follow the CRISP-DM. The CRISP-DM methodology was chosen because it provides sequential instructions on solving real world issues using data mining techniques.

### 5.4.1 Data mining process

The methodology for data mining can be described as a four step process (Berry and Linoff, 2004). The first step of the data mining process is identifying the business problem. The second step is data processing. Data processing can be carried out by methods of data transformation and data cleaning. The third step of the data mining process is applying specific algorithm(s) over the processed data. In this study, classification algorithms were used to investigate churn in the MSI. The final step of the data mining process is evaluating the results derived by applying a specific algorithm to solve a particular problem. Evaluation can be done using techniques described in section 5.3.

### 5.4.2 Datasets Description

This section explains the activities associated with the data understanding phase. Two different MNOs datasets were utilised in this study to showcase the selected widely-used churn modelling techniques for churn analysis. Furthermore, datasets were chosen so that performance of the churn modelling techniques can be compared. Section 5.4.3 and section 5.4.4 provide a more comprehensive description of the datasets used in this study.

### 5.4.3 Twitter Dataset Description

As described in section 4.3.1, Tweetcatcher was used to capture tweets that were posted about MNOs of interest (see section 4.3.1 for further details about the dataset). The aim of this search query was to identify positive and negative sentiment associated with the selected MNOs. The geo-coordinates related to the tweets helped in identifying regions with a good or bad sentiment. The overall aim of capturing tweets was to identify satisfied, and dissatisfied customers' as satisfied customers often stay with their MNO and dissatisfied customers churn (Henkel et al., 2006). To predict churn using the Twitter dataset, sentiment analysis was used to describe customers who churned and the customers who did not churn. Each tweet was associated with a senti-score. The senti-scores are either negative, positive or neutral. For the purpose of this study, tweets with a positive senti-score represent customers who did not churn (i.e. they described a positive sentiment because they are happy with their MNO) and tweets with a negative senti-score represent customers who churned. The neutral sentiments were not used in this study because they neither represent satisfied nor dissatisfied customers. Figure 5.4 presents a representation of CSDs derived from chapter 4.

In order to separate tweets according to CSDs (customer service, coverage quality, price), the process of extracting similar words of each customer satisfaction determinant was conducted. This process was carried out using online lexical resources such as thesaurus and dictionaries. Dictionaries are generally considered as a valuable and reliable source containing information about the relationships among terms (e.g synonyms) (García-Crespo et al., 2010). Table 5.4 presents CSDs and the terms related to each CSD.

The evaluation carried out in the first iteration of this study also provided

<b>Customer Service</b>	<b>Coverage Quality</b>	<b>Price</b>
Information	Network	Cost
Helpline	Coverage Quality	Monetary Value
Advice	Coverage	Terms
Information	Signal	Toll
Helpdesk	Service	Value
After-sale service	3g	Worth
Service	4g	Fee
Client service	2g	Charge
CS	LTE	Amount
Product service	Signals	Bill
Maintenance	Signal strength	Rate
Provision	Weak	Expense
Care	Network quality	Outlay
Backing	Edge	Rate
Upkeep	Wi-Fi	Discount
Technical		Pay
Technical team		Paid
Repair		Payment
Replacement		Tariff
Issues		
Courtesy phone		
Troubleshooting		

Table 5.4: CSD and related terms



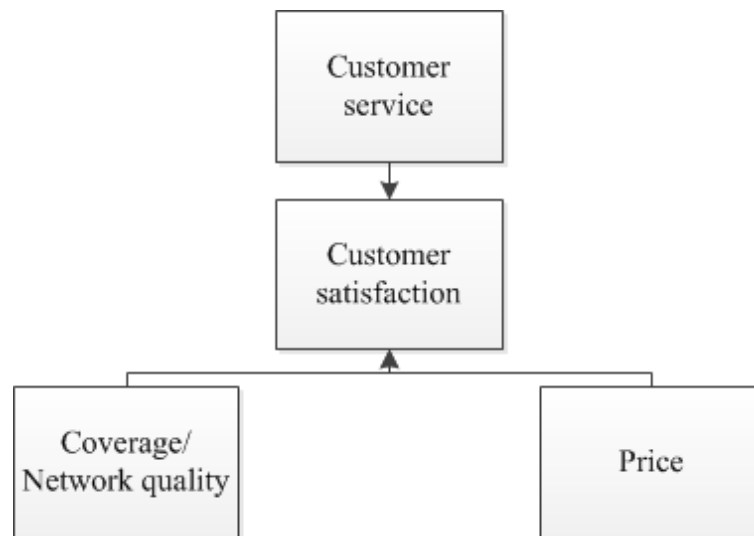


Figure 5.4: Graphic representation of CSDs derived from iteration one

insights regarding the CSDs and related terms. Participants mentioned words for example pay, bill and rate when they were asked questions with regards to price. In order to prepare the tweets for churn modelling, the following steps were taken:

1. As described in section 5.4.3, the process of extracting similar words of each CSD was conducted.
2. Each CSD and similar words were saved in separate files.
3. The negative and positive tweets were separated using sentiment analyses. The negative and positive tweets of each determinant were saved in different files.
4. Word frequency was conducted for each determinant and related words against all the tweets for each determinant. The number of occurrence of each word per tweet was recorded.
5. The negative file for each determinant was classified as customers who

churned while the positive files for each determinant were classified as customers who did not churn.

6. The negative and positive datasets for each determinant were combined into a single file.
7. Churn modelling algorithms were run on the datasets.

#### 5.4.4 Telco Dataset Description

The Telco dataset was extracted from a UK MNOs data warehouse. The Telco dataset comprises of 19,919 observations and 23 variables. In addition, the dataset contains a dependent variable (churn/stay) and predictor variables i.e. customer data (such as type of device, price plan and region). The dataset contains 50% of customers who churned and 50% of customers who stayed with their MNO until the end of their contract. The dataset is based on an 24 month contract. Some customers stayed with their MNO after the end of the 24 month period (i.e they renewed their contract) while other customers left their mobile operator after the 24 month contact. In order to investigate churn with respect to the determinants of customer satisfaction derived from chapter 4 (coverage quality, customer service and price). The dataset is separated such that each variable has a link to the already established CSD. Hence, the dataset is separated into three parts (see table 5.5). The models for predicting churn are tested on each of the CSD and a comparison is carried out to establish which model best predicts churn based on the CSD. Table 5.5 provides a brief description of the input variables for the dataset.

Demographics	Variable name	Description
	Contract length	Length of contract
	Gender	Customers gender
	Sales channel	Company that delivered contract
	Post code	Post code in which customer lives
	County Name	The name of county where customer lives
	Region	The region where customer lives
	Devices	Name and model of device used by customer
	Tenure	Number of months with the present mobile service provider
	Lifestage segment	Customer age
Customer Service	Number of complaints	Number of customer complaints regarding billing in the 18 month contract period
	Problems with Handset	Number of times customer has reported problem with handset
Coverage Quality	Q2 Voice	Data usage in second quarter
	Q3 Voice	Voice usage in third quarter
	Q2 bytes	Voice usage in second quarter
	Q3 bytes	Data usage in third quarter
Price	Price plan	Type of plan chosen by customer (for example 500mins..)
	Billing Queries	No of times customer has queried their monthly bill
	Cost	Monthly cost of customer plan
	Price plan	Type of plan chosen by customer (for example 500mins..)

Table 5.5: Telco Dataset Description

### 5.4.5 Data Preparation

Data captured from social media can be incomplete, inconsistent and noisy. Therefore, data has to be prepared to improve the quality and the performance of data analysis. The major tasks in data preparation include data cleaning, data integration, data transformation, data reduction, and data discretisation (Han et al., 2011). Data cleaning as the name states implies, cleaning the data by filling in missing values, identifying or removing outliers, and resolving inconsistencies. Data cleaning is usually performed in two iterative phases: discrepancy detection and data transformation. Data integration entails merging data from multiple sources in analysis, for example integrating multiple databases, data cubes or data files. Careful integration can help reduce and avoid inconsistencies. Data reduction techniques can be applied to a dataset to obtain a reduced representation of the dataset that is much smaller in volume, yet closely maintaining integrity of the original data. Mining the reduced dataset should produce the same (or almost the same) result. Data transformation is converting data to a form that is suitable for mining. Data discretisation is the process of transforming continuous data attributes values into a finite set of intervals with minimal loss of information. In this study, data cleaning was conducted with the following steps:

1. Filling in missing values.
2. Resolving inconsistencies.
3. Identifying and removing outliers.

Filling in missing values and resolving inconsistencies was achieved by predicting missing values with the rest of the dataset. The process of identifying and removing outliers involve removing unimportant and unreliable variable.

The flag variable, along with three other variables were removed from the list of variables because they displayed little variation and would not help get the best model. As modelling algorithms may have specific requirements for data preparation, data preparation should be carried out such that it suits the requirements for selected algorithms.

### 5.4.6 Data Modelling

This section presents the process and result used to conduct churn analysis with both the Twitter dataset and the Telco dataset. Three modelling techniques (Decision Trees, Logistic regression and SVM) were applied for churn analysis. R studio (Version 0.98.1103) statistical language was used to carryout experiments with the chosen modelling techniques. A split design was used to carryout analysis for the CSDs on the Twitter dataset and the Telco dataset. Eighty percent of the data in each dataset was used for training the model and twenty percent of the data in each dataset was used to test the trained model. Customer churn is modelled on the training dataset and the model is applied on the testing dataset in order to validate the model. The essence of these experiments is to build churn prediction models with the selected algorithms and to compare the results. The results will determine which model is most suitable for predicting churn based on each CSD.

### 5.4.7 Decision Trees Analysis

This section presents a comparison of results derived from using decision tree analysis techniques for building customer churn prediction models using two datasets- Twitter and Telco. The decision tree algorithms to be implemented on the datasets are CART and random forest. The decision tree accuracy rates

Table 5.6: Comparison of Decision Trees techniques for predicting churn on the Twitter and Telco datasets

<b>Twitter Dataset</b>		
<b>Determinants</b>	<b>CART</b>	<b>Random Forest</b>
Customer Service	93.1	94.48
Coverage Quality	52.64	50
Price Fairness	70.58	73.53
<b>Telco Dataset</b>		
<b>Determinants</b>		
Customer Service	85.45	86.62
Coverage Quality	85.41	85.19
Price Fairness	87.76	87.58

for predicting churn based on customer service on the Twitter dataset are as 93.10% for the CART model and 94.48% for the random forest model.

The accuracy rates for predicting customer churn on coverage quality are 52.64% for the CART model and 50.00% for the random forest model. Finally, the random forest model produced a higher accuracy rate for predicting churn based on price fairness. The accuracy rate for the random forest model was 73.53% while the accuracy rate for the CART model was 70.58%. Table 5.6 presents the results of the CART model and the Random forest model for predicting customer churn based on CSDs with the Twitter dataset.

The Telco dataset has some similar results with the Twitter dataset (see also table 5.6). The random forest model performed better than the CART model for predicting churn based on customer service. The accuracy rates are 85.45% for the CART model and 86.62% for the random forest model. The accuracy rates for predicting churn based on coverage quality are 85.41% for

the CART model and 85.19% for the random forest model. The accuracy rates for predicting churn based on price fairness are 87.76% for the CART model and 87.58% for the random forest model. Generally, the CART algorithm outperformed the random forest model for predicting churn based on the dataset used. Although the random forest model outperformed the CART model in predicting churn based on customer service, the CART model performed better than the random forest model for predicting coverage quality and price fairness.

#### 5.4.8 Logistic Regression Analysis

This section presents logistic regression results for predicting churn on the CSDs. Again, the logistic regression technique is used to predict churn on both the Twitter dataset and the Telco dataset. As explained in section 5.2.2, an evaluation process was carried out to establish probability value for all datasets. The cross-validation method was adopted to establish the best cut-off point for all the datasets. The cutoff point was then used as the probability value for analysis. This process was carried out for the training dataset and the testing data set. The scores for the logistic regression prediction are 91.03% for predicting churn based on customer service, 64.19% for predicting churn based on coverage quality and 74.26% for predicting churn based on price. The logistic regression model performed best for predicting churn based customer service for the Twitter dataset. The lowest predicting score was on coverage quality with an accuracy score of 64.19%. The logistic regression model also performed best for predicting churn based on customer service for the Telco dataset. The scores for the predictions on CSDs for the Telco dataset are 78.14%, 76.33% and 77.30% for customer service, coverage quality and price respectively. Table 5.7 presents the logistic regression accuracy scores for CSDs

<b>Determinants</b>	<b>Twitter Dataset</b>	<b>Telco Dataset</b>
Customer Service	91.03	78.14
Coverage Quality	64.19	76.33
Price	74.26	77.3

Table 5.7: Logistic Regression Churn Prediction Results

on the Twitter dataset and the Telco dataset. See appendix for the ROC and LIFT curves for prediction models on both the Twitter dataset and the Telco dataset.

#### 5.4.9 SVM Analysis

Similar to section 5.4.8, this section presents the results of using the SVM technique for churn prediction on the Twitter dataset and the Telco dataset. The SVM model produced an accuracy rate of 47.66% and 55.88% for predicting churn based on coverage quality and price fairness respectively. The results of applying the SVM model to CSDs on the Telco dataset are 86.25%, 68.29% and 87.98% for customer service, coverage quality and price fairness respectively. The SVM model performed better with the Telco dataset because it produced higher accuracy rates. This may be as a result of the better quality of the Telco dataset. Table 5.8 presents a table of the accuracy rate in percentage for predicting each customer satisfaction determinant on the Twitter dataset and the Telco dataset. In addition to the results already presented, tables 5.10 and 5.11 present a comparison of results for all the modelling techniques selected.



<b>Determinants</b>	<b>Twitter Dataset</b>	<b>Telco Dataset</b>
Customer Service	63.45	86.25
Coverage Quality	47.66	68.29
Price	55.88	87.98

Table 5.8: SVM churn prediction results

### 5.4.10 Evaluation of Models and Discussion

The prediction models described in section 5.2 were selected based on their quality, and widespread usage for data analysis. As stated in section 5.4.1, the evaluation phase is the final phase of the data mining process. Models built are compared to uncover the best model for solving the business problem. There are several evaluation metrics that can be used to compare the performance of different predictive modelling techniques (see section 5.3). In this study, AUC, the lift curve and the ROC curve are used as performance evaluation metrics. Section 5.4.11 presents the evaluation of the models built. Subsequently, a discussion on the limitations of applying data mining for churn analysis is presented.

### 5.4.11 Model Evaluation Metrics

Tables 5.9 presents scores of the three evaluation metrics adopted in this chapter. The random forest model outperforms all other models in predicting churn based on customer service on Twitter data. Also on the Twitter dataset, the logistic regression model outperforms other models for predicting churn based on coverage quality. Finally on the Twitter dataset, the logistic regression model outperforms other models for predicting churn based on price fairness. The prediction results of the Telco data have some similarities with the Twitter

Table 5.9: Model Comparison on Twitter and Telco Datasets

<b>Twitter Dataset</b>	<b>CART</b>	<b>RandomForest</b>	<b>SVM</b>	<b>Logistic Regression</b>
<b>Customer satisfaction determinant</b>				
Customer Service	93.1	94.48	63.45	91.03
Coverage Quality	52.34	50	47.66	64.19
Price	70.5	73.53	55.88	74.26
<b>Telco Dataset</b>				
<b>Customer satisfaction determinant</b>				
Customer Service	86.08	86.08	86.25	78.14
Coverage Quality	85.41	85.19	68.29	76.33
Price	87.76	87.58	87.96	77.3

data. On the Telco dataset, the random forest model also outperforms other models for predicting churn based on customer service. The CART model outperforms other models for predicting churn based on coverage quality. The CART model also outperforms other models in predicting churn based on price fairness. Although, the SVM model was not significantly outperformed, it performed well in predicting churn based on customer service and price. Overall, the results show that decision trees are best for predicting churn in the MSI and the results confirm studies in the literature, such as Verbeke et al. (2012).

It is impossible to make a fair comparison with studies in the literature because different datasets are used for analysis. Nevertheless, based on the results derived from evaluating this study, decision tree analysis are suggested to be a possible valuable technique for churn prediction. Tables 5.10 to 5.11 present the lift scores and the AUC scores obtained from the prediction models under study. See Appendix (7.1 to 7.12) for ROC and Lift curves for the models used in this study.

## 5.5 Limitations to Data Mining

This study has shown that data mining techniques can be used to explore customer churn. However, there are some limitations to applying data mining techniques to understanding customer behavior. This section highlights three limitations of data mining for churn analysis in the MSI. The first limitation is data mining models perform better with good quality datasets. The MSI is an evolving industry where new standards and technologies are introduced every day. Over time certain techniques may not be sufficient to solve similar business problems because of the dynamics of the market. The second limitation is regarding the dataset used in this study. Besides the factors revealed

Table 5.10: Lift scores on CSDs for Twitter and Telco datasets

<b>Lift Scores for Twitter Dataset</b>						
<b>CSDs</b>		<b>CART</b>	<b>Random Forest</b>	<b>SVM</b>	<b>Logistic Regression</b>	
Customer Service Model		1.55914	1.5934	1.563	1.5469	
Coverage Quality Model		1.0426	1.0036	0.9707	1.2172	
Price Model		1.9428	1.9833	1	1.6227	
<b>Lift Scores for Telco Dataset</b>						
Customer Service Model		1.338	1.343	1.37	1.262	
Coverage Quality Model		1.336	1.344	1.024	1.248	
Price Model		1.395	1.371	1.382	1.363	

Table 5.11: AUC scores on CSDs for Twitter and Telco datasets

<b>AUC Scores for Twitter Dataset</b>		<b>CART</b>	<b>Random Forest</b>	<b>SVM</b>	<b>Logistic Regression</b>
<b>Determinants</b>					
Customer Service Model		0.9193	0.9354	0.6725	0.8908
Coverage Quality Model		0.5429	0.5023	0.9707	0.4797
Price Model		0.6736	0.707	0.5	0.746
<b>AUC Scores for Telco Dataset</b>					
Customer Service Model		0.836	0.86	0.831	0.74
Coverage Quality Model		0.8298	0.8275	0.5259	0.7
Price Model		0.8754	0.8669	0.8587	0.7035

No.	Limitation	Possible solutions
1	Issue with quality of data.	Provide tools that can minimize the dependency of data quality.
2	Limited customer data for churn analysis.	Provide tools that can deal with a variety of customer data.
3	Inability to provide possible causes of customer churn.	Offer tools that can better explain churn models.

Table 5.12: Limitations of Data for churn analysis (Hassouna et al., 2015)

as CSDs in this study, customers can churn due to other reasons such as social influence. Since only limited data was available for analysis, it is impossible to capture these factors accurately and produce more effective results. For instance, for one of the determinants (customer service), churn on customer service was predicted using the available attributes on the dataset. There are other factors that influence a good customer service and a bad customer service. These factors can be included as variables to the dataset to produce a more accurate result. The third limitation lies in the inability of data mining to provide possible reasons why customers churned or stayed with their MNO. In this study, data mining was helpful in predicting churn based on the identified CSDs but the models developed were unable to provide further insights on the causes of customer churn. Table 5.12 provides a list of the highlighted limitations of using data mining techniques in churn analysis and possible measures of overcoming them.

## 5.6 Summary

Churn is both an expensive and a crucial problem in the MSI. As a result, MNOs have adopted techniques such as decision trees, regression analysis and SVM to analyse and predict churn. The results of the experiments presented in this study show that the decision tree models (CART and Random Forest model) outperform other models for predicting churn using the Twitter dataset. The lift scores for the decision tree models (CART and Random forest) are **1.55914** and **1.5934** respectively for predicting churn based on customer service. The lift scores produced by the decision tree (CART and random Forest) models for predicting churn based on price are **1.9428** and **1.9833** respectively. However, the logistic regression model outperforms other models in predicting churn based on coverage quality. The logistic regression model attains a lift score of **1.2172** for predicting churn based on coverage quality. This study recommends the use of decision trees and logistic regression for predicting churn using unstructured dataset.

The prediction results for the Telco data are different from the prediction results obtained from the Twitter dataset. The CART and random forest models produced a lift score of **1.338** and **1.343** respectively for predicting churn based on customer service while the SVM model produced a lift score of **1.370**. The decision tree models outperformed other models for predicting churn based on coverage quality. The lift score produced for the CART model is **1.336** in predicting churn based on coverage quality. The random forest model produced a lift score of **1.344**. Similarly, one of the decision tree models, the CART model outperformed other models for predicting churn based on price fairness. The lift score for the CART model is **1.395**. The random forest model also produced an impressive lift score of **1.371** with an AUC score of **0.8669**. However, this model was outperformed by the SVM model, which

produced a lift score of **1.382**. Generally, the decision tree models produced the best AUC and lift scores for predicting churn using the Telco dataset. Therefore, for similar datasets, this study recommends the use of decision trees for analysing customer churn. This chapter shows that data mining can be useful for churn analysis and it also provides some limitations to the use of data mining for churn analysis. As a means to overcome the limitations, the next chapter attempts to study the impact of social effect on customer retention by applying a data-driven approach to ABM.



# Chapter 6

## Data-Driven Approach For Agent Based Modelling

### 6.1 Introduction

This chapter presents the third iteration of this study. It seeks to provide insight into the social effects of customer retention by examining agent behaviours. In order to address the limitation of understanding the impact of social effects on customer retention, a novel approach is proposed to design agents in an ABM exercise to uncover how customers social network influences customer retention. The novel approach was built using automated decision trees and the experiment for investigating the social influence of customer churn was conducted using a novel tool for understanding customer behaviour in a wider market place.

This chapter is structured as follows. Section 6.2 presents the purpose of the model. Section 6.3 provides an overview of modelling customer behaviour. Section 6.4 presents the CADET approach, a novel approach for describing agents in ABM. Section 6.5 presents the validation of the CADET model with

the TEA-SIM tool. Finally section 6.6 provides the conclusion of the chapter

## 6.2 Purpose of Model

The Customer Agent Decision Trees (CADET) approach for conducting ABM is a novel approach for building agent based models using automated decision trees. Decision trees are used to describe customers final decision regarding their mobile contract. At the end of a mobile contract, each customer makes a decision to renew their mobile contract or to churn and move to a different MNO. The details of the mobile contracts are provided in section 5.2. The CADET approach aims to provide insight on why customers have made certain decisions and the driving forces behind those decisions. The drivers of customer decision regarding the renewal of contract is derived from the set of attributes provided with the dataset. The CADET approach is able to show attributes that drive customers final contract renewal decision. In addition, the CADET approach provides a clear diagrammatic analysis of customers decision which is driven by a set of subjective attributes. These attributes make up the nodes on the decision tree. In this study, mobile network customers' attributes include mobile phone type, customer location, price plan and data usage. Customers decision to churn or renew their contract is composed of a number of phases represented on the decision tree. The final node of the decision tree is customers' final decision to stay with/ or leave their MNO. Decision trees consist of dependent and independent variables. The dependent variables are the nodes that make up the final node. These variables influence customers final contract renewal decision.

The primary purpose of adopting the CADET approach to building a model is to understand the social influence on customer retention i.e. the influence

of the environment, customers family and friends network on customer retention. The CADET approach utilised with the TEA-SIM model (See section 6.5.1 for the description of the TEA-SIM model) provides information on the possible decisions a customer might make when they interact with other customers within their network or environment. If a customer meets another customer within the network or environment, and they share the same MNO, they may have a conversation on their experience and that conversation may be an influence on customer final re-purchase decision. The CADET approach to ABM seeks to provide an understanding on how customer variables along with customer network can influence customer retention.

### 6.3 Customer Behaviour Modelling

Customers often interact with other customers about products and services that they purchase. In addition to the advertisement campaigns carried out by MNOs, interaction among customers can also influence customer decision to either purchase a product or service or not. There are various theories in social science and marketing concerned with understanding and modelling customer behaviour (Richard and Chebat, 2016). Customer behaviour may change as a result of an act of a consumer changing preference on a product or service. A number of approaches have been applied to understand the concept behind consumers changing preferences (Vag, 2007). These concepts can be summarised as: (1) multi-agent simulation, (2) WOM research, (3) consumer behaviour model, (4) social network analysis (Vag, 2007).

1. Multi-agent systems are computerised systems that consist of independent agents within an environment. These agents interact with one another as a means of imitating social processes. Multi-agent models can

simulate behaviour that emerge from motivation.

2. WOM is the action of a customer spreading negative or positive remarks to other customers about a product or service. Organisations use WOM marketing to affect the WOM process (Vag, 2007).
3. Consumer behaviour models provide a description and explanation to how consumers act when purchasing a product or a service. The description and explanation are carried out using standard scientific terms and approaches. In consumer behaviour models, consumers go through four phases namely, awareness, interest, decision, and action (Vag, 2007). The awareness phase entails discovering a service. The interest phase is the act of showing interest in the service. The decision phase is accepting or declining the service and the action phase is acting on the decision taken.
4. Social network analysis (SNA) is a means of exploring social structures using network and graph theories. It describes networks as nodes and the ties or edges that connect each node on the network. This study presents the application of SNA to ABM with the CADET approach. The CADET approach is discussed in details in section 6.4.

## 6.4 The CADET Approach

A number of approaches can be used to describe an ABM process. Agent oriented methodologies have been extended into two areas; object oriented (OO) methodologies and knowledge engineering methodologies (Iglesias et al., 1999). The three common views in OO technologies for describing ABMs are static for describing the structure of objects; dynamic for describing object interac-

tion and functional for describing the data flow of the methods of the objects (Iglesias et al., 1999). Flow chart is another popular example for representing the process flow of an ABM exercise (Grimm et al., 2006)

This study presents the CADET approach, a novel approach for ABM using results derived from decision tree analysis. The agent attributes are derived from the decision tree flow process. The CADET approach can be applied to various domains and industries. To further explain the structure and functionality of the CADET approach, a scenario is presented below.

A Telco company provides a number of products and services. In order to understand customer purchase behaviour, a decision tree analysis is conducted with historical data. The decision tree approach is effective in understanding trends and patterns of consumer decision making. In addition, the analysis is carried out to improve understanding of customer purchase behaviour. The analysis shows that customers who have certain attributes purchase the iPhone from the Telco company. The variables used for this analysis are age, gender, salary, location, make of current phone and number of complaints. The analysis shows that customers who have the attributes described below have purchased an iPhone from the Telco provider while other customers have purchased other brands.

if (age <30, gender = female, salary >30000, location = midlands, make of current phone = iPhone, number of complaints/year <2) then purchase an iPhone, otherwise purchase another brand. Table 6.1 presents a tabular description of the steps to identify customers who either purchased an iPhone or another brand.

Figure 6.1 displays the decision tree analysis process described above. The CADET approach works best with Decision tree's CART algorithm. The process of performing ABM using the CADET approach is displayed in figure 6.2.

Steps	Description
Step 1	If age $\not>$ 30, purchase another brand, else consider the next node.
Step 2	If age $>$ 30, and gender = female, purchase another brand, else consider the next node.
Step 3	If age $>$ 30 and gender $\neq$ female, and salary $\not>$ 30000, purchase another brand, else consider the next variable.
Step 4	If age $>$ 30 and gender $\neq$ female, and salary $>$ 30000, and location $\neq$ midlands, then purchase another brand, else consider the next variable.
Step 5	If age $>$ 30, and gender $\neq$ female, and salary $>$ 30000, and location = midlands, and make of current phone $\neq$ iphone, then purchase another brand else consider the next variable.
Step 6	If age $>$ 30, and gender $\neq$ female, and salary $>$ 30000, and location = midlands, and make of current phone = iphone, and number of complaints/year $\not<$ 2, then purchase another brand, else purchase an iphone.

Table 6.1: Steps to identifying Customer Types from DT

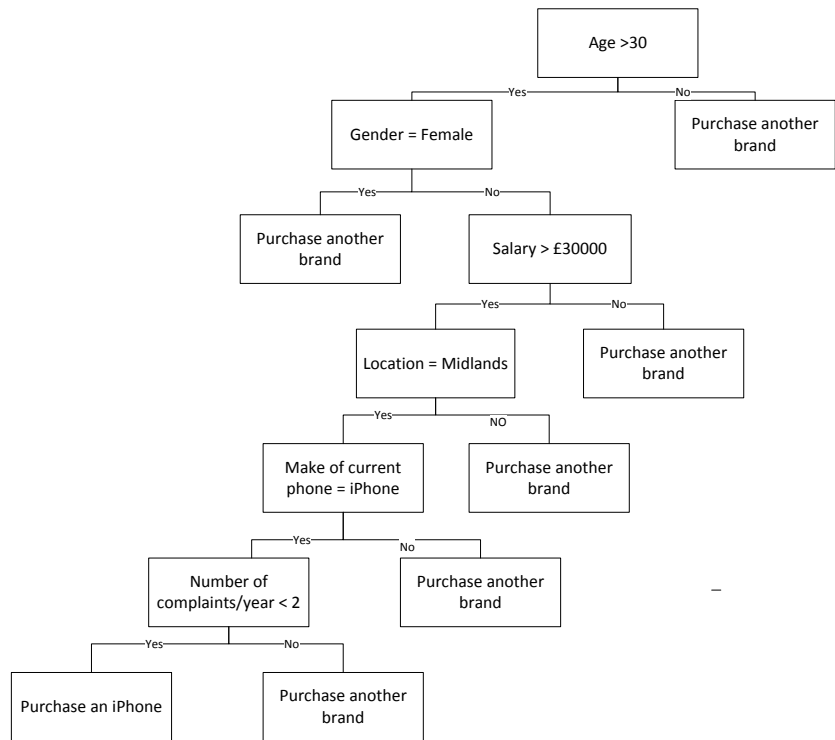


Figure 6.1: An example of deriving attributes for the CADET approach



Figure 6.2: The CADET Model

Firstly, a data source is selected. The CADET approach can be applied to both structured and unstructured data. Secondly, a decision tree analysis is performed on the dataset. This allows the data-driven process to ABM. Finally, ABM is performed with the results derived from the decision tree analysis. The next section presents validation process of the CADET model.

## 6.5 CADET Model Validation

Validation is the process of testing a model to see how it conforms to reality. Validation is a standard process in modelling. There are four major steps for effectively validating a model Rand and Rust (2011): micro-face validation, macro-face validation, empirical input validation and empirical output validation. Micro-face validation is the process of ensuring that the properties of the model correspond to the real world properties that are modelled. Macro-face validation is the process of ensuring that the patterns described in the model correspond to real-world patterns. Empirical input validation is the process of ensuring that the results derived from applying the model can be obtained using a different dataset. This borrows from the concept of training and testing datasets. Empirical output validation involves showing that the output of the model corresponds to the real world. To provide a validation for the CADET model, the TEA-SIM tool, an environment used to validate the CADET model is described. Subsequently, the dataset used to perform validation of the CADET model is then described.



### 6.5.1 The TEA-SIM Tool

Over the years, MNOs have spent a large amount of money on strategies that enable them to manage and understand their customer churn behaviour more effectively. While advertisement and WOM can be a powerful tools for customer retention, customers are often skeptical about advertisement. As a result, customers may turn to their family or friends within their network to seek advice before deciding whether to churn or remain with their mobile network provider. Some companies have manipulated WOM by running schemes that offer customers benefits for expressing positivity about their product or service. As positive WOM is a great tool for marketing, negative WOM can also be a damaging tool for businesses. Dissatisfied customers of a product or service may share their experience about their dissatisfaction with members of their social network which can include family and friends.

The TEA-SIM tool is an agent-based customer modelling tool built by incorporating a cognitive process for understanding how members of a small-world network make decisions. In addition, the TEA-SIM tool is a decision support tool that can be adopted, not only by MNOs but also by various industries to model various entities such as customers, products and services. The TEA-SIM tool also provides a medium for companies to see the interaction process between agents and how the interaction influences the decision of agents. The result derived from this process can be used to provide information to organisations so that they can strengthen their CRM strategies by exploring the effect of WOM to improve customer retention.

The dynamic nature of the TEA-SIM tool provides a unique approach to understanding customer behaviour. It can be used to observe the patterns of customer interaction and perform further analysis to explore the possibilities of incorporating those patterns into marketing strategies in order to increase

revenue. The TEA-SIM tool also works as a generic model that captures the key drivers behind customer change of behaviour and it can also work well in a consultancy environment. It is not a precise prediction tool. As a result, ABM is used to provide insight into the behaviour of a population of customers. The next section presents an application of the CADET approach to a selected dataset.

### 6.5.2 Applying the CADET approach to the Telco dataset

The CADET approach utilises decision trees for ABM. Decision trees are used to describe customers that either remain with their MNO after their contract period and customers that leave their MNO after their contract period. The CADET approach shows that customers decision to churn or remain with their MNO is driven by a set of subjective attributes. These attributes are the individual nodes on the decision tree. Mobile network customers may have attributes such as mobile phone type, customer location, price plan and data usage. The decision to churn or stay with an MNO is composed of a number of phases represented on the decision tree. The final node of the decision tree is a customer's final decision to either churn or stay with their MNO. Decision trees consist of dependent and independent variables. The dependent variables are the nodes that make up the final node. These variables influence customers final decision. The primary purpose of applying the CADET approach for ABMS using the TEA-SIM tool is to understand how much influence a customer's environment, family, and friends within their network have on customer retention. In addition, the CADET approach utilised with the TEA-SIM model provides information on the possible decisions a customer might make when they interact with other customers within their network or environment. If a customer meets another customer within the environment,

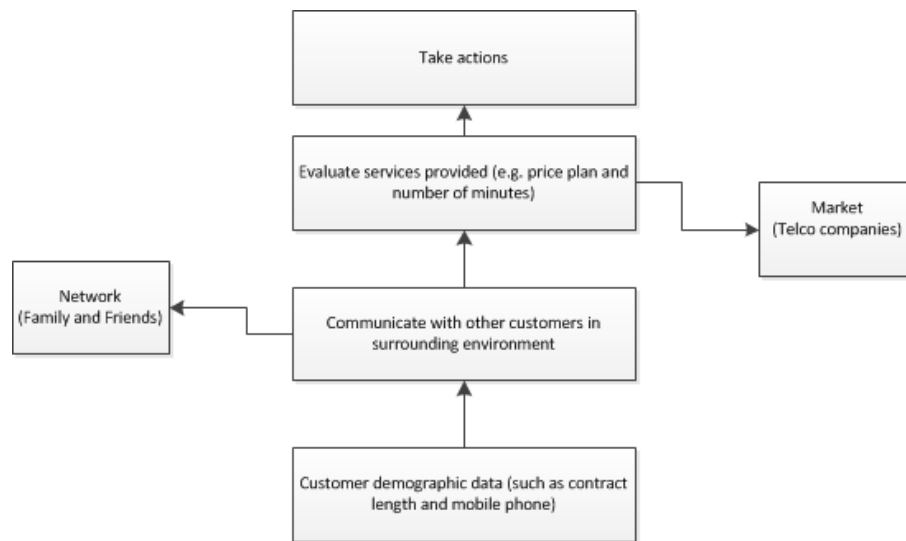


Figure 6.3: Conceptual Architecture for the CADET Approach and the TEA-SIM Tool

and they share the same MNO, they may have a conversation about their overall customer experience and that conversation may influence a customer's decision to churn or renew their contract. The CADET approach to ABM seeks to provide an understanding on how customer variables along with customer network can influence customer retention. Figure 6.3 shows the interaction process of customers by applying the CADET approach and the TEA-SIM tool.

Applying the CADET approach with the TEA-SIM tool works by evaluating customer satisfaction derived from the interaction with other customers in the surrounding environment. This interaction can lead customers to evaluate the services provided by their MNO and make a final decision to churn or stay with their MNO.

### 6.5.3 Agent Attributes and Behaviour

There are two major components derived from experimenting with the CADET model and the TEA-SIM tool. The first component is customer agents, which represents mobile services users. The second component is the environment that represents mobile services operators and the mobile service market place. Modelling customer behaviour can provide a better understanding of customer behavioural intentions. Modelling customer behaviour involves applying interdisciplinary processes and theory from sociology, psychology, economics and marketing (Shah et al., 2006). In addition, modelling and simulating agent behaviour provides a means to understanding customer demands and needs so companies can provide their customers with better services that meet their needs. The CADET model applied with the TEA-SIM tool represents how customer interaction can influence customer behaviour. Figure 6.3 shows a conceptual model of the CADET model and the TEA-SIM tool.

### 6.5.4 Dataset Description

To demonstrate the usability of the CADET approach, the dataset described in section 5.4.4 is utilised. The dataset is based on a 24 month contract. Some customers renewed their contract at the end of the 24 month contract period while other customers churned. See table 5.5 for the dataset description.

### 6.5.5 Model Structure

To apply the CADET model with the TEA-SIM tool, the CART decision tree algorithm is run on the dataset described in section 5.4.4 and the result of the decision tree is visualised. The decision tree shows the flow of the decision process for customers who either churned or remained with their MNO. The

top of decision tree analysis shows that a customer's tenure is either greater than 24 months or not. If tenure is greater than 24 months then move to the next variable on the left. However, if tenure is not greater than 24 months, then move to the next variable on the right-side of the tree. This process continues until the end of the tree. The end of the tree displays the end result of the process i.e. either customer churned or renewed their contract at the end of the 24-month contract period. Figure 6.4 displays the decision tree analysis diagram and table 6.2 further explains the the decision tree analysis.

To simulate this process, agents are fed into the TEA-SIM tool by undergoing the following process: Three files are created to run simulation on the TEA-SIM tool. The files are `init.json`, `model.json` and `steps.php`. The json files describe agent attributes and the ABMS environment. The php file describes the interaction process of the agents.

The `init.json` file (figure 6.5) is composed of the following attributes: `grid`, `agent` and `simulation`. `n` and `m` under "grid" represent the number of rows and columns required for the ABMS experiment. "Agents" is composed of agent type, instances and the position of the agents. The `init.json` files show that the instances are six in number and the position is random. "Simulation" on the `init.json` file represent the number of runs for the experiment. In this experiment, "start" is 0 and "end" is 25. This means that the agents in the grid should move 25 times.

The `model.json` file (figure 6.6) describes customer type and attributes. There are two types of customers with customer id 1 and customer id 2. Customer id 1 represent churn customers while customer id 2 represent stay customers. From the decision tree analysis carried out (figure 6.4), churn customers have the following attributes:

tenure >14, device = samsung, contract >12 months, region = midlands,

Steps	Description
Step 1	If tenure is $>$ or $<$ 24 consider the next node.
Step 2	If tenure is $\not>$ 24 and tenure $<$ 14, churn, else consider the next node.
Step 3	If tenure $\not>$ 24, and tenure $\not<$ 24, and device = samsung then churn, else renew contract.
Step 4	If tenure $>$ 24, and contract length $\not>$ 12, then churn else consider the next node.
Step 5	If tenure $>$ 24, and contract length $>$ 12, and region = midlands, then churn, else consider the next node.
Step 6	If tenure $>$ 24, and contract length $>$ 12, and region $\neq$ midlands, and region $\neq$ London, then churn else consider the next node.
Step 7	If tenure $>$ 24, and contract length $>$ 12, and region $\neq$ midlands, and region = London, and device = iphone 5, then churn, else move to the next node.
Step 8	If tenure $>$ 24, and contract length $>$ 12, and region $\neq$ midlands, and region = London, and device = iphone 5, and problem with handset $>$ 3, and gender = male, then churn else renew contract.
Step 9	If tenure $>$ 24, and contract length $>$ 12, and region $\neq$ midlands, and region = London, and device = iphone 5, and problem with handset $\not>$ 3, and number of complaints $>$ 2, then churn, else renew contract.

Table 6.2: Steps for Decision Tree Analysis

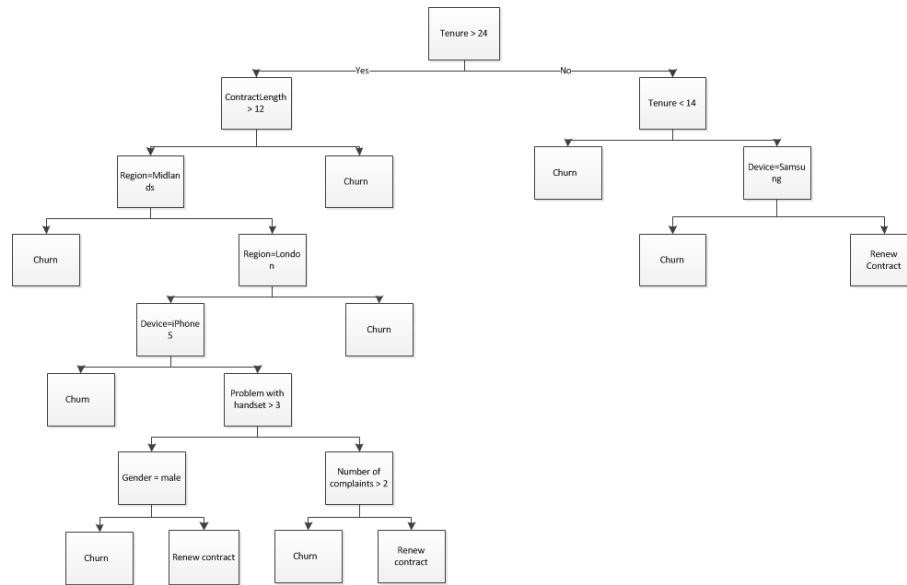


Figure 6.4: Decision Tree Analysis

region  $\neq$  London, device = iPhone5, problem with handset  $>3$ , gender = male, number of complaints  $>2$ .

Furthermore, stay customers have the following attributes: tenure  $>24$ , contract length  $> 12$ , region  $\neq$  Midlands, region = London, device  $\neq$  iPhone5, problem with handset  $< 3$ , gender = female, number of complaints  $< 2$ .

The steps.php file (figure 6.7) describes agent interaction. To summarise the code, if a churn customer meets a stay customer, the stay morphs to a churn customer. The crying face in figure 6.8 represents churn customers and the smiley face represent stay customers. Overall, this experiment shows how customer behaviour can be influenced by the environment. Figure 6.8 shows the process in which agents interact on the TEA-SIM model. The agents move from one grid to another and their decisions are based on the interaction with family/friends as represented in figure 6.8.

```

1 {
2   "grid":
3     [
4       { "n" : 10, "m" : 10 }, "agents" : [ { "type" : "customers.1", "instances" : 6, "position" : "rnd"},
5         { "type" : "customers.2", "instances" : 5, "position" : "rnd" } ], "simulation" : { "start" : 0, "end" : 25 } }

```

Figure 6.5: Init.json File

```

1 {
2   "customers" : [
3     { "id" : 1, "name": "ChurnCustomers",
4       "attributes":[ { "name" : "tenure", "value" : < 14 months } ,
5         { "name" : "device", "value" : samsung } ,
6         { "name" : "contract", "value" : < 12 months } ,
7         { "name" : "region", "value" : midlands } ,
8         { "name" : "region", "value" : !London } ,
9         { "name" : "device", "value" : iPhone5 } ,
10        { "name" : "problemsWithHandset", "value" : > 3 } ,
11        { "name" : "gender", "value" : male } ,
12        { "name" : "number of complaints", "value" : > 2 } ,
13        { "name" : "_img" , "value" : "img/ChurnCustomer.jpeg" } ] } ,
14
15
16
17     { "id": 2, "name": "StayCustomers",
18       "attributes":[ { "name" : "tenure", "value" : > 24 } ,
19         { "name" : "churn", "value" : false } ,
20         { "name" : "contract length", "value" : > 12 } ,
21         { "name" : "region", "value" = !Midlands } ,
22         { "name" : "region", "value" : London } ,
23         { "name" : "device", "value" : !iphone5 } ,
24         { "name" : "problemsWithHandset" , "value" : < 3 } ,
25         { "name" : "gender", "value" : female } ,
26         { "name" : "number of complaints", "value": < 2 } ,
27         { "name" : "_img" , "value" : "img/StayCustomer.jpeg" } ] } ] }
28

```

Figure 6.6: Model.json File



```

<?php

class customers_ChurnCustomers extends Agent {
    function step($step) {
        if (!$this->churn && $this->anyNeighbour (1, 2, array("churn"=>false),
array("churn" => true, "_img" => "img/ChurnCustomer.jpeg"))) {
            $this->churn = true;
            $this->_img = "img/ChurnCustomer.jpeg";
        }
        $this->move(1);
        $this->morph(2);
    }
}

class customers_StayCustomers extends Agent {
    function step($step) {
        $this->move(1);
    }
}

```

Figure 6.7: Stepper function

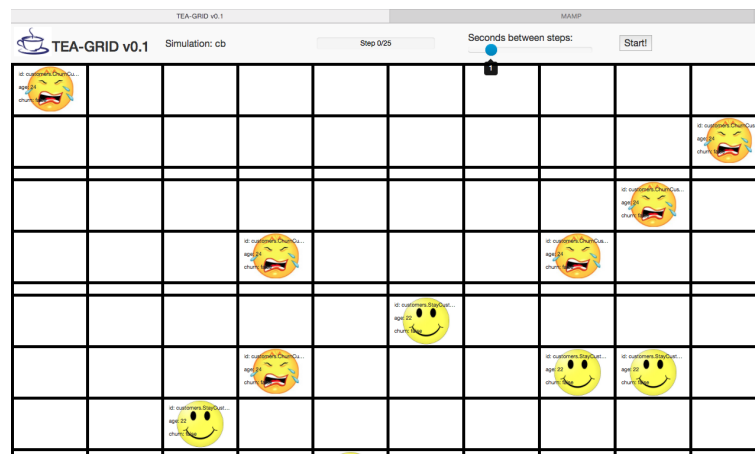


Figure 6.8: Agent Interaction Process

## 6.6 Conclusion

Numerous factors can influence the decision for customers to purchase or adopt a product or service. However, customers are likely to trust the WOM of someone in their network. This chapter introduced a novel approach to agent-based modelling with the use of the TEA-SIM ABMS tool.

The TEA-SIM tool provides a generic approach for companies across various industries to better understand their customers. The tool can help decision makers in organisations work on better strategies to understand customer behaviour, enhance customer retention, and subsequently improve CRM.

This chapter also presents an experiment of implementing the CADET model with the Telco dataset, and the TEA-SIM ABMS tool. The CADET model was used to effectively provide an Agent-Based model using the results derived from decision trees. Although the CADET model was applied to a dataset extracted from an MNO, the concept of the CADET approach can be extended and implemented in other industries such as health-care, manufacturing and financial services.

# Chapter 7

## Conclusion and Discussion

### 7.1 Overview

This chapter presents the research findings and conclusion of this study. It presents the contributions made by this study and highlights the limitations which will provide insights for further research. The remaining sections of this chapter are structured as follows: section 7.2 presents a summary of the thesis by providing the rationale of each chapter. Section 7.3 highlights the research contributions. Section 7.4 presents a summary of design artefacts achieved from this study. Section 7.5 presents the limitations to this study. In addition, some possible future research that will provide more insights into the customer retention domain are suggested.

### 7.2 Research Summary

This thesis presents some the techniques for understanding customer behaviour and improving customer retention in the MSI. Many tools and techniques for accessing customer retention have evolved over the past few decades. However,

these tools and techniques are equally relevant to understanding recent customer behaviour and can be applied to a modern-day dataset. Understanding customer retention in the MSI is determined by a broad range of attributes such as location of customer, type of service purchased and customer demographics. These attributes were used in this thesis to; identify the determinants of customer satisfaction in the MSI, analyse and develop models for predicting churn and to propose a novel approach for ABM, focusing on modelling the effect of customer social network (family and friends) on customer retention.

Customer retention is crucial to every business and it is even more critical for fiercely competitive industries such as the MSI where acquiring new customers is expensive (i.e. advertising cost, setting up new accounts and educating customers). MNOs try to avoid these expenses by deriving strategies to retain their customers. One approach that has been taken by most MNOs is the use of social media to interact with their customers. The evolution of social media (Facebook, Twitter, and on-line reviews) and its increasing adoption by customers is beneficial to companies to better understand customer needs. This has attributed to new and innovative ways to better satisfy customers and improve customer retention. Given the fundamental role played by customers who publish their thoughts about their MNO on social media, it is essential for companies to maximise the utility of social networks to businesses. In addition, it is critical for decision makers in companies to broaden their understanding on how to appropriately utilise these mediums to improve customer satisfaction and, in turn, increase customer retention. The aim of this research is to provide a methodological approach to understanding customer behaviour, focusing on customer retention in the MSI. The aim of this research was achieved in three phases. Firstly, the SoMeDoA approach was applied to uncover the determinants of customer satisfaction in the MSI. Sec-

ondly, a novel approach was applied to build churn prediction models. Lastly, the CADET approach was proposed for describing agents and their behaviour in ABM. The objectives set out for achieving the aim of this research are summarised below:

**Objective 1:** Analyse the normative literature to uncover the state-of-the-art of customer satisfaction and customer retention debates in the MSI.

**Objective 2:** Derive the determinants of customer satisfaction in the MSI using social media data.

**Objective 3:** Assess the current and most widely-used machine learning techniques for analysing customer retention in the MSI highlighting their capabilities and limitations.

**Objective 4:** Develop a novel data-driven framework for describing agents in agent based modelling and simulation.

**Objective 5:** Demonstrate the effectiveness of the data-driven framework by conducting experiments into mobile service customer retention.

In order to achieve the aim and objectives set forth for this research, Chapter 2 reviewed the state of the art of core areas of study in this research: customer satisfaction, customer retention and customer behaviour modelling. The aim of the review is to gain a deeper understanding of the state-of-the-art of the aforementioned areas. The two key sections in chapter two focused on customer relationship management and customer behaviour modelling. In addition, chapter 2 examined the interrelationship of customer satisfaction, customer retention and CRM. These areas were studied in order to analyse the debate on the possible drivers for customer churn in the MSI. Several techniques for customer retention were discussed, focusing on data mining and ABMS techniques. Traditional analysis techniques have been applied to study customer retention in the literature. These techniques were explored and their

limitations were outlined. Consequently, ABMS was presented as a possible solution to addressing the outlined limitations. Researchers have used both traditional techniques and ABMS techniques to investigate customer retention because of how critical customer retention is in the MSI (Hassouna et al., 2015). Furthermore, the importance of the web in addressing customer retention was also highlighted. The accessibility of the web has allowed free communication between businesses and customers. Web technologies such as Twitter are increasingly used by businesses to engage and support customers, by listening to their needs and providing solutions to those needs.

Decision trees, logistic regression and SVM are among the most popular tools for investigating customer retention in the literature. While these tools have been widely used to investigate customer retention in the literature, many of the studies focused on; (1) analysing a few specific factors such as customer satisfaction and customer loyalty, paying less attention to other important aspects such as social ties, and (2) addressing customer characteristics and their interaction with MNOs, neglecting the importance of studying the effects of customer to customer interaction on customer retention.

Chapter 3 establishes the process of achieving the aim and objectives of this study using DSR. DSR provides a pathway in engaging in a design research problem by providing necessary learning from each phase to improve the proposed solution of the defined problem. The main design artefacts created in this study are the customer satisfaction determinants ontology (CSDO), customer satisfaction determinants (CSD) approach to predicting churn and the CADET approach for describing agent-based models. Following the fundamental DSR guidelines, this study was accomplished in three phases and carried out in an iterative manner. The knowledge gained from each iteration was used to facilitate the next iteration. This research was composed of

three build iterations (i.e.- build to evaluate) (March and Smith, 1995). The CSDO artefact was derived by applying the SoMeDoA framework (Bell and Shirzad, 2013) to investigate the determinants of customer satisfaction in the MSI. Interviews were used as a method for validating the findings derived from this phase of work. The findings of this iteration are represented on an onto graph. The CSD approach to predicting churn was derived by applying the most popular churn prediction techniques in the MSI, to each of the determinants derived from iteration one. This approach was carried out using two datasets, the Twitter dataset and the Telco datasets. Both datasets are described in chapters 4 and 5 respectively. The findings reveal that decision tree models are the best models for predicting churn in the MSI. This is in line with some prior study in the literature e.g (Verbeke et al., 2012). The final artefact developed in this study is the CADET approach to ABM. The CADET approach to ABM is a novel approach to describing agents and their behaviour using results generated from decision trees. Chapter 6 sheds more light on the CADET approach and its implementation.

Chapter 4 described the first iteration of this study. This phase of the study focused on utilising the SoMeDoA framework to investigate the determinants of customer satisfaction in the MSI. The aim of the iteration is to extract data from the web to identify the most crucial customer satisfaction factors and examine their valuable temporal components. This iteration demonstrated the utility and practicality of using social network data to gain an in-depth understanding of customer needs. The SoMeDoA approach was utilised to extract and analyse domain specific data about selected MNOs from a social networking site (Twitter). The primary focus of this chapter was utilising the SoMeDoA framework to provide some useful insights on the factors that determine customer satisfaction in the MSI. The rationale for selecting social

networks are: (a) to access and analyse a broader and wider range of dataset for addressing the problem of customer satisfaction in the MSI (b) to analyse the information generated from Twitter with the aim of advising MNOs on the key factors that can help improve customer satisfaction. Interviews were used to evaluate the results achieved by applying the SoMeDoA framework as a means to assess its validity in the MSI. The evaluation process proved that the factors identified as the determinants for customer satisfaction are true and some additional learning was derived in the interview process. The learning derived from this iteration was fed into iteration 2 to provide more understanding about the problem domain.

Chapter 5 established the capabilities and benefits of applying data mining techniques to investigating customer retention by empirically developing churn prediction models. The churn prediction models were built based on the already established determinants of customer satisfaction derived from chapter 4. Customer satisfaction determinants are customer service, coverage quality and price. The churn prediction models built on these determinants were developed using decision trees (CART and random forest), logistic regression and SVM. The results of the models developed are compared and they reveal that decision trees are the best techniques for building data mining models for churn prediction.

Chapter 6 describes the final iteration of this research. The CADET framework which is a novel framework for describing agent-based models is proposed. This chapter begins by explaining the purpose of the model. Afterwards, approaches to customer behaviour modelling are briefly explained and social network analysis is introduced. Furthermore, the CADET approach is introduced and an example is presented to help understand the CADET approach. The TEA-SIM tool which is also a novel tool built for understanding agent



behaviour within an environment is briefly explained as it was used to carry out the ABMS experiment.

In order to validate the CADET approach, an experiment was carried out with the Telco dataset. The agent architecture, agent attributes and behaviour, and the model structure are explained. The evaluation process was carried out to establish the effectiveness of the CADET approach to ABMS. In the experiment, agents represent mobile subscribers. Mobile subscribers were modelled to interact with other customers in the environment thereby establishing the effect of social influence on customer retention. The TEA-SIM tool provided a social environment, replicating how customers can be influenced by their social circle. Validating the CADET approach with the TEA-SIM tool provides insights on the impact of social influence of customer retention. In addition, it demonstrates how WOM can be a vital tool for customer retention. Table 7.1 presents a table that matches each chapter to the objective it addressed.

### **7.3 Research Contributions and Conclusions**

According to DSR, it is fundamental to make a contribution to a domain of study. The contributions in this study follow the DSR guidelines (March and Smith, 1995; von Alan et al., 2004; Vaishnavi and Kuechler, 2015). Contributions in a DSR study are in forms of artefacts i.e. constructs, methods, models, and instantiations (March and Smith, 1995; von Alan et al., 2004; Peffers et al., 2007). The artefacts derived from this research are summarised below:

#### **1. The CADET Framework**

The CADET framework, as far as the author is aware, is the first artefact to present an automated data-driven approach for describing agents in an

<b>Objective</b>	<b>Chapter</b>	<b>Outcome</b>
Objective 1- Analyse the normative literature to uncover the state-of-the-art of customer satisfaction and customer retention debates in the MSI	Chapter 2,3	Studying the area of customer satisfaction and customer retention in the MSI. In addition, studying the methodologies that have been utilised to analyse and investigate satisfaction and customer retention in the MSI.
Objective 2- Derive the determinants of customer satisfaction in the MSI using social media data.	Chapter 4	Applying the SoMeDoA approach to uncover the determinants of customer satisfaction in the MSI. A set of interviews were conducted to validate the findings derived by applying the SoMeDoA framework to the problem domain. The outcome of this objective is the customer satisfaction determinants ontology (CSDO).
Objective 3- Assess the current and most widely-used techniques for analysing customer retention in the MSI highlighting their capabilities and limitations	Chapter 5	Proposed a novel approach for predicting churn based on CSD.
Objective 4- Develop a data-driven framework for describing agents in agent based modelling and simulation.	Chapter 6	Developed a novel approach, the CADET approach for describing agents and their attributes in an ABMS experiment.
Objective 5- Demonstrate the effectiveness of the data-driven framework by providing viable insights into customer retention.	Chapter 6	Validated the usability of the CADET approach for ABMS.

Table 7.1: Chapters Addressing Objectives of Study

ABMS experiment. Hence, the development process of the CADET framework is a contribution to DSR. The CADET framework can be applied to both structured and unstructured datasets. The CADET framework was used in this thesis to describe agents and their attributes in an ABMS experiment. The ABMS experiment was conducted as validation for utilising the CADET framework as a data-driven approach to ABM.

The TEA-SIM tool is an ABMS platform that was used to conduct the ABMS experiment. The application of the CADET approach with the TEA-SIM tool can explicitly capture customer interactions with their network to enable better forecasts and better business decisions. In this study, the CADET framework and the TEA-SIM tool were used to capture customer interactions in the market place in order to provide useful insights on the impact of social influence in the market place. The application of the CADET framework and the TEA-SIM tool proves that the the CADET framework can account for the dynamics of customer behaviour in the MSI.

Furthermore, the CADET framework provided concrete insight into customer retention in this study. In particular, the experimental results derived from the evaluation of the CADET framework shows that WOM is an important factor for customer retention. In terms of describing ABMs, the CADET approach has the following advantages: 1) It provides a clear and intuitive way to describing agents and their attributes while explaining the significance of relationships between variables. 2) It can take into account the heterogeneity of customers to capture emergent phenomena. 3) It is flexible and can easily adapt new constraints. 4) In addition to its flexibility, the CADET approach can also be applied to various industries as seen in section 6.5.2.

Applying the CADET framework to an ABMS tool such as the TEA-SIM tool can provide actionable insights into the social effects of customer retention.

This work clarifies on the need to incorporate the study of social effects into the widely used churn-prediction models (Neslin et al., 2006; Phadke et al., 2013), which have so far ignored inter-customer dynamics. In addition, this study confirms the findings of Dasgupta et al. (2008) who found that the probability for a customer to churn increases when the number of neighbours who churn increase. The findings of this thesis also contribute to the knowledge of drivers of customer loyalty from a social effects perspective. In addition, the results derived from this study highlights the need to examine various social effects when addressing customer retention. Consequently, providing empirical evidence on how social effects influence customer retention over time.

### **2. The customer satisfaction ontology determinants (CSDO)**

The Customer Satisfaction Determinants Ontology (CSDO) was achieved by applying the SoMeDoA framework to investigate the determinants of customer satisfaction using a dataset obtained from Twitter. The results show that customer service, coverage quality and price are the key determinants of customer satisfaction in the MSI. These findings confirm some key studies in customer satisfaction domain such as Hanif et al. (2010), and Emerah et al. (2013). Furthermore, the results derived from using SoMeDoA to examine customer satisfaction was evaluated using interviews. The interviews added some valid opinions that could not be derived from using the SoMeDoA framework in isolation (see section 4.5 for the interviews conducted to validate the of SoMeDoA).

### **3. Novel approach for predicting churn**

Another contribution derived from this study is the application of a novel approach for building churn prediction models for structured (i.e. Telco data) and unstructured datasets (Twitter data) datasets. As far as the author is aware, no study in the customer retention literature has attempted to build

churn prediction models based on CSDs. This study applies the most widely used churn prediction techniques to build churn prediction models for the CSDs identified in chapter 4 (see 4.13 for the identified CSDs).

The approach applied to building churn prediction models reveal that decision trees (CART and Random Forest) are the best techniques for building churn prediction models on Twitter data. The lift scores for decision tree models (CART and Random forest) predicting churn based on customer service are 1.55914 and 1.5934 respectively. The lift scores for predicting churn based on price are 1.9428 for the CART model and 1.9833 for the random forest model. The logistic regression model is the best model for predicting churn based on coverage quality. The lift score for the logistic regression model is 1.2172 (see table 5.13).

The same approach for building churn prediction was utilised with the Telco dataset. Majority of studies in the literature that compared models for predicting churn did not consider comparing models for predicting the causes of churn. This study builds and compares churn prediction models using a similar approach with the Twitter dataset. The Telco dataset was split into the identified CSDs and the selected data mining techniques were applied to the CSDs to build churn prediction models. The results show that the SVM model outperforms other models for predicting churn based on customer service with a lift score of 1.370. The decision tree models (CART and Random Forest) outperform other models for predicting churn based on coverage quality with a lift score of 1.336 and 1.344 respectively (see table 5.16). Similarly, the CART model outperforms other models for predicting churn based on price with a lift score of 1.395. The findings in this study confirms studies in the literature such as Huang et al. (2012) and Verbeke et al. (2012).

## 7.4 Summary of Design Artefacts

This section presents a summary of the artefacts developed in this research. The artefacts were developed in forms of models, methods and instantiation.

In iteration one of this study, a **model** was built from the process of uncovering the determinants of customer satisfaction. The customer satisfaction determinants ontology (CSDO) was derived by identifying the CSDs and associating similar words with each CSD. In addition, this research also benefits from proposing the CADET model. The CADET **model** incorporates the process of gathering structured or unstructured data, performing a decision tree analysis on the selected dataset and then describing the agents in an ABM exercise with the results derived from the decision tree analysis.

Furthermore, this research benefits from two **methods**: The first **method** is the CADET approach, which is a data-driven method for describing agents for ABM. The second **method** is the process of conducting this research. The overall purpose of this study is to derive a methodological approach to understand customer behaviour, focusing on customer retention. To achieve this, three iterations of study were conducted. First, the factors that drive customer satisfaction are identified, secondly, some churn models were built to uncover the model that best predicts churn based on each of the identified CSD. Finally, based on one of the techniques applied in iteration two, a data-driven approach to ABM is derived. Figure 7.1 presents a diagrammatic representation of the research process.

Finally, this research also benefits from two **instantiations**. The first **instantiation** is the applied CADET approach in R statistical language. The CADET approach is applied to capture the dynamics of customer behaviour using data. Figure 7.2 presents a snippet of the code used to derive the CADET approach in R.

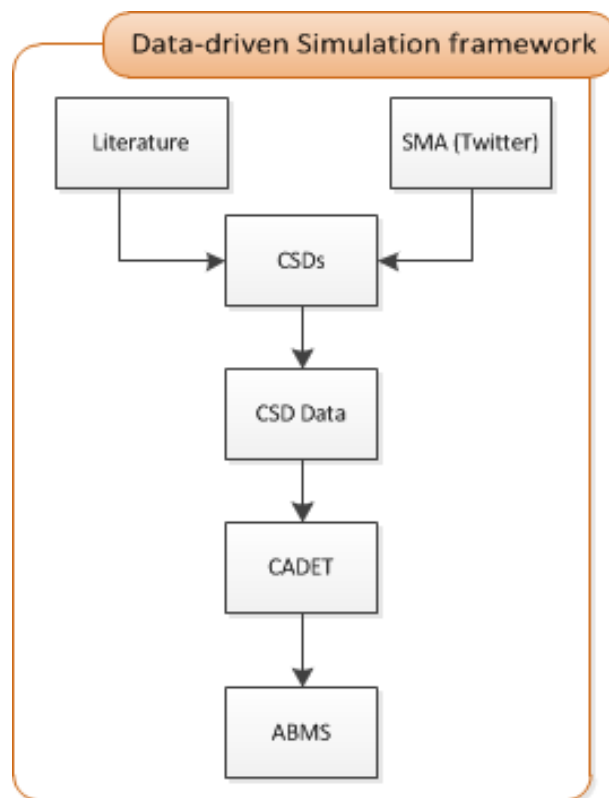


Figure 7.1: Data-driven Simulation Framework

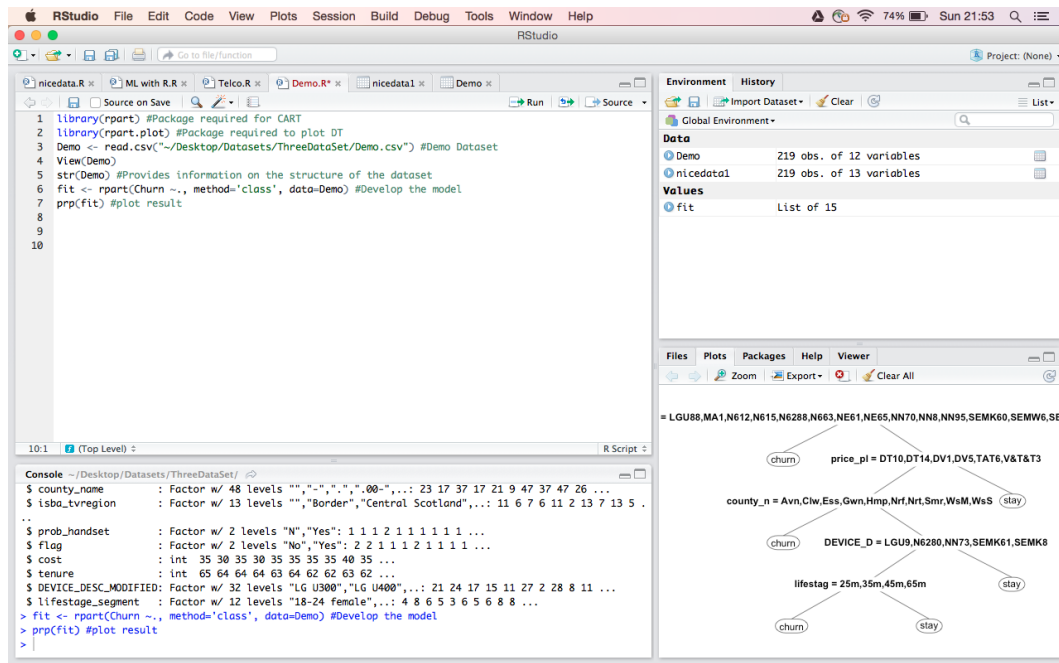


Figure 7.2: CADET Approach in R

The packages used to achieve the results presented on the screen shot are `rpart` and `rpart.plot`. The tree like structure on the right side of the screen represents the behaviour of customers driven from the dataset. The second **instantiation** is the ABMS experiment conducted with the CADET approach and the TEA-SIM tool. The ABMS experiment shows that social influence which has been commonly ignored is an important factor for customer retention Haenlein (2013).

## 7.5 Research Limitation and Future Work

Although the research presented has made some valuable contributions to understanding customer behaviour in the MSI, some challenges are noted.

While adopting the SoMeDoA framework provides insights on key CSDs in the MSI, the author has identified three limitations to its use to address



the research problem. Firstly, the Twitter dataset used for the study was collected for a period of one week. Analysing data for a longer period may bring about some useful insights into CSDs in the MSI. Secondly, comparing results with other weeks may showcase some more CSDs. Thirdly, extracting and analysing data from other social media websites such as Facebook and Instagram may bring about more useful insights in CSDs in the MSI. This is because some customers are more active using other social media sites other than Twitter. Analysing data from a single social media platform may mean that some customers opinions are neglected. In addition, analysing data from other platforms may create a competitive advantage for MNOs.

Another major limitation of this research is the limited amount of customer information used to carryout analysis and to develop the artefacts derived from this research. The Twitter dataset does not include variables such as customer complain information, contract information and fault reports. Analysing the dataset with these features may reveal some interesting insights about CSDs and the data mining model built using the Twitter dataset.

Another limitation lies in the second iteration of this study. Boosting the selected algorithms used to develop churn prediction models may bring about new insights that will contribute to the state of the art of customer retention in the MSI. Furthermore, it will be interesting to detect customers who are likely to churn. This study compares the most popular and widely used techniques for developing churn prediction models. Uncovering customers who are likely to churn will provide more insights on whom to send targeted advertisements to.

Finally, the CADET approach uses automatically generated attributes to describe agents in an ABM exercise. As a result, some variables may be ignored in an ABM exercise. Applying a technique that considers most or

all the variables in a dataset may bring about a more interesting result and may contribute to the state-of-the-art of customer retention and DSR.

# Bibliography

- Adami, C. (1998), *Introduction to artificial life*, Vol. 1, Springer Science & Business Media.
- Agnihotri, R., Dingus, R., Hu, M. Y. and Krush, M. T. (2015), 'Social media: Influencing customer satisfaction in b2b sales', *Industrial Marketing Management* .
- Agnihotri, R., Kothandaraman, P., Kashyap, R. and Singh, R. (2012), 'Bringing "social" into sales: The impact of salespeople's social media use on service behaviors and value creation', *Journal of Personal Selling & Sales Management* **32**(3), 333–348.
- Ahn, J.-H., Han, S.-P. and Lee, Y.-S. (2006), 'Customer churn analysis: Churn determinants and mediation effects of partial defection in the korean mobile telecommunications service industry', *Telecommunications policy* **30**(10), 552–568.
- Alexandrov, A., Lilly, B. and Babakus, E. (2013), 'The effects of social-and self-motives on the intentions to share positive and negative word of mouth', *Journal of the Academy of Marketing Science* **41**(5), 531–546.
- Almossawi, M. M. (2012), 'Customer satisfaction in the mobile telecom in-

- dustry in bahrain: Antecedents and consequences', *International Journal of Marketing Studies* **4**(6), p139.
- An, L. (2012), 'Modeling human decisions in coupled human and natural systems: review of agent-based models', *Ecological Modelling* **229**, 25–36.
- Anderson, E. W. (1998), 'Customer satisfaction and word of mouth', *Journal of service research* **1**(1), 5–17.
- Backiel, A., Baesens, B. and Claeskens, G. (2016), 'Predicting the time-to-churn of prepaid mobile telephone customers using social network analysis', *Journal of the Operational Research Society* .
- Bamfo, B. A. (2009), 'Exploring the relationship between customer satisfaction and loyalty in the mobile telecommunication industry in ghana.', *Indian Journal of Economics & Business* **8**(2).
- Baskerville, R. L. and Myers, M. D. (2002), 'Information systems as a reference discipline', *Mis Quarterly* pp. 1–14.
- Baumann, C., Elliott, G. and Burton, S. (2012), 'Modeling customer satisfaction and loyalty: survey data versus data mining', *Journal of services marketing* **26**(3), 148–157.
- Baxter, N., Collings, D. and Adjali, I. (2003), 'Agent-based modelling—intelligent customer relationship management', *BT Technology Journal* **21**(2), 126–132.
- Bell, D., Kashefi, A., Saleh, N. and Turchi, T. (2016), 'A data-driven agent based simulation platform for early health economics device evaluation', *to appear in Proceedings of the SpringSim'16* .

- Bell, D. and Shirzad, S. R. (2013), 'Social media business intelligence: A pharmaceutical domain analysis study', *International Journal of Sociotechnology and Knowledge Development (IJSKD)* **5**(3), 51–73.
- Benoit, D. F. and Van den Poel, D. (2012), 'Improving customer retention in financial services using kinship network information', *Expert Systems with Applications* **39**(13), 11435–11442.
- Berry, M. J. and Linoff, G. S. (2004), 'Data mining techniques second edition-for marketing, sales, and customer relationship management'.
- Bradley, A. P. (1997), 'The use of the area under the roc curve in the evaluation of machine learning algorithms', *Pattern recognition* **30**(7), 1145–1159.
- Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.
- Breiman, L. et al. (2001), 'Statistical modeling: The two cultures (with comments and a rejoinder by the author)', *Statistical Science* **16**(3), 199–231.
- Burez, J. and Van den Poel, D. (2008), 'Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department', *Expert Systems with Applications* **35**(1), 497–514.
- Burez, J. and Van den Poel, D. (2009), 'Handling class imbalance in customer churn prediction', *Expert Systems with Applications* **36**(3), 4626–4636.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000), 'Crisp-dm 1.0 step-by-step data mining guide'.
- Choi, J., Seol, H., Lee, S., Cho, H. and Park, Y. (2008), 'Customer satisfaction factors of mobile commerce in korea', *Internet research* **18**(3), 313–335.

- Chu, B.-H., Tsai, M.-S. and Ho, C.-S. (2007), 'Toward a hybrid data mining model for customer retention', *Knowledge-Based Systems* **20**(8), 703–718.
- Coelho, P. S. and Henseler, J. (2012), 'Creating customer loyalty through service customization', *European Journal of Marketing* **46**(3/4), 331–356.
- Coussement, K., Benoit, D. F. and Van den Poel, D. (2015), Preventing customers from running away! exploring generalized additive models for customer churn prediction, *in* 'The Sustainable Global Marketplace', Springer, pp. 238–238.
- Das Gupta, D. and Sharma, A. (2009), 'Customer loyalty and approach of service providers: An empirical study of mobile airtime service industry in india', *Services Marketing Quarterly* **30**(4), 342–364.
- Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A. and Joshi, A. (2008), Social ties and their relevance to churn in mobile telecom networks, *in* 'Proceedings of the 11th international conference on Extending database technology: Advances in database technology', ACM, pp. 668–677.
- Deadman, P., Robinson, D., Moran, E. and Brondizio, E. (2004), 'Colonist household decisionmaking and land-use change in the amazon rainforest: an agent-based simulation', *Environment and Planning B* **31**, 693–710.
- Duda, R. O., Hart, P. E. and Stork, D. G. (2012), *Pattern classification*, John Wiley & Sons.
- Edward, M. and Sahadev, S. (2011), 'Role of switching costs in the service quality, perceived value, customer satisfaction and customer retention linkage', *Asia Pacific Journal of Marketing and Logistics* **23**(3), 327–345.

- Emerah, A. A., Oyedele, S. O. and David, J. O. (2013), ‘Determinants of customer satisfaction in the nigerian telecommunication industry: an empirical evidence’, *International Journal of Management and Strategy* **4**(6), 1–12.
- Epstein, J. M. (2009), ‘Modelling to contain pandemics’, *Nature* **460**(7256), 687–687.
- Epstein, M. J. and Westbrook, R. A. (2012), ‘Linking actions to profits in strategic decision making’, *Image* .
- Farmer, J. D. and Foley, D. (2009), ‘The economy needs agent-based modelling’, *Nature* **460**(7256), 685–686.
- Friedman, J., Hastie, T. and Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics Springer, Berlin.
- Furness, P. (2001), ‘Techniques for customer modelling in crm’, *Journal of Financial Services Marketing* **5**(4), 293–307.
- García-Crespo, Á., Colomo-Palacios, R., Gómez-Berbís, J. M. and Ruiz-Mezcua, B. (2010), ‘Semo: a framework for customer social networks analysis based on semantics’, *Journal of Information Technology* **25**(2), 178–188.
- Gerpott, T. J. and Ahmadi, N. (2015), ‘Regaining drifting mobile communication customers: Predicting the odds of success of winback efforts with competing risks regression’, *Expert Systems with Applications* .
- Gilbert, G. N. (2008), *Agent-based models*, number 153, Sage.
- Gimblett, H. R. (2002), *Integrating geographic information systems and agent-based modeling techniques for simulating social and ecological processes*, Vol. 168, Oxford University Press New York.

- Glaser, B. S. and Strauss, A. (1971), ‘A. 1967, the discovery of grounded theory’, *New york* .
- Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S. K., Huse, G. et al. (2006), ‘A standard protocol for describing individual-based and agent-based models’, *Ecological modelling* **198**(1), 115–126.
- Günther, C.-C., Tvette, I. F., Aas, K., Sandnes, G. I. and Borgan, Ø. (2014), ‘Modelling and predicting customer churn from an insurance company’, *Scandinavian Actuarial Journal* **2014**(1), 58–71.
- Haenlein, M. (2013), ‘Social interactions in customer churn decisions: The impact of relationship directionality’, *International Journal of Research in Marketing* **30**(3), 236–248.
- Han, J., Kamber, M. and Pei, J. (2011), *Data mining: concepts and techniques: concepts and techniques*, Elsevier.
- Hanif, M., Hafeez, S. and Riaz, A. (2010), ‘Factors affecting customer satisfaction’, *International Research Journal of Finance and Economics* **60**, 44–52.
- Hassouna, M., Tarhini, A., Elyas, T. and AbouTrab, M. S. (2015), ‘Customer churn in mobile markets: a comparison of techniques’, *International Business Research* **8**(6), 224.
- He, B., Shi, Y., Wan, Q. and Zhao, X. (2014), ‘Prediction of customer attrition of commercial banks based on svm model’, *Procedia Computer Science* **31**, 423–430.
- Henkel, D., Houchaime, N., Locatelli, N., Singh, S., Zeithaml, V. and Bittner,



- M. (2006), 'The impact of emerging wlangs on incumbent cellular service providers in the usmj services marketing'.
- Heppenstall, A. J., Crooks, A. T., See, L. M. and Batty, M. (2011), *Agent-based models of geographical systems*, Springer Science & Business Media.
- Hevner, A. and Chatterjee, S. (2010), *Design science research in information systems*, Springer.
- Huang, B., Kechadi, M. T. and Buckley, B. (2012), 'Customer churn prediction in telecommunications', *Expert Systems with Applications* **39**(1), 1414–1425.
- Huang, Y. and Kechadi, T. (2013), 'An effective hybrid learning system for telecommunication churn prediction', *Expert Systems with Applications* **40**(14), 5635–5647.
- Hung, S.-Y., Yen, D. C. and Wang, H.-Y. (2006), 'Applying data mining to telecom churn management', *Expert Systems with Applications* **31**(3), 515–524.
- Iglesias, C. A., Garijo, M. and González, J. C. (1999), A survey of agent-oriented methodologies, *in* 'Intelligent Agents V: Agents Theories, Architectures, and Languages', Springer, pp. 317–330.
- Itzkowitz, J. (2013), 'Customers and cash: How relationships affect suppliers' cash holdings', *Journal of Corporate Finance* **19**, 159–180.
- Jahromi, A. T., Stakhovych, S. and Ewing, M. (2016), Customer churn models: A comparison of probability and data mining approaches, *in* 'Looking Forward, Looking Back: Drawing on the Past to Shape the Future of Marketing', Springer, pp. 144–148.

- Johnson, M. D., Herrmann, A., Huber, F. and Gustafsson, A. (2012), *Customer retention in the automotive industry: quality, satisfaction and loyalty*, Springer Science & Business Media.
- Jones, M. A., Taylor, V. A. and Reynolds, K. E. (2014), 'The effect of requests for positive evaluations on customer satisfaction ratings', *Psychology & Marketing* **31**(3), 161–170.
- Kang, J., Tang, L., Lee, J. Y. and Bosselman, R. H. (2012), 'Understanding customer behavior in name-brand korean coffee shops: The role of self-congruity and functional congruity', *International Journal of Hospitality Management* **31**(3), 809–818.
- Keaveney, S. M. (1995), 'Customer switching behavior in service industries: An exploratory study', *The Journal of Marketing* pp. 71–82.
- Keramati, A. and Ardabili, S. M. (2011), 'Churn analysis for an iranian mobile operator', *Telecommunications Policy* **35**(4), 344–356.
- Khan, A. A., Jamwal, S. and Sepehri, M. (2010), 'Applying data mining to customer churn prediction in an internet service provider', *International Journal of Computer Applications* **9**(7), 8–14.
- Kim, J., Suh, E. and Hwang, H. (2003), 'A model for evaluating the effectiveness of crm using the balanced scorecard', *Journal of interactive Marketing* **17**(2), 5–19.
- Kim, K., Jun, C.-H. and Lee, J. (2014), 'Improved churn prediction in telecommunication industry by analyzing a large network', *Expert Systems with Applications* **41**(15), 6575–6584.

- Kirui, C., Hong, L., Cheruiyot, W. and Kirui, H. (2013), 'Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining', *IJCS* **10**.
- Kotler, P. (1994), 'Marketing management, analysis, planning, implementation, and control, philip kotler'.
- Kotler, P. (2009), *Marketing management: A south Asian perspective*, Pearson Education India.
- Kotsiantis, S. B. (2013), 'Decision trees: a recent overview', *Artificial Intelligence Review* **39**(4), 261–283.
- Kucuktunc, O., Cambazoglu, B. B., Weber, I. and Ferhatosmanoglu, H. (2012), A large-scale sentiment analysis for yahoo! answers, *in* 'Proceedings of the fifth ACM international conference on Web search and data mining', ACM, pp. 633–642.
- 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), 489–504.
- Lee, C., Kwak, N. and Lee, C. (2015), 'Understanding consumer churning behaviors in mobile telecommunication service industry: Cross-national comparison between korea and china'.
- Lee, H. S. (2013), 'Major moderators influencing the relationships of service quality, customer satisfaction and customer loyalty', *Asian Social Science* **9**(2), p1.
- Leelakulthanit, O. and Hongcharu, B. (2011), 'Factors that impact customer

- satisfaction: Evidence from the thailand mobile cellular network industry', *International Journal of Management and Marketing Research* **4**(2), 67–76.
- Lemmens, A. and Croux, C. (2006), 'Bagging and boosting classification trees to predict churn', *Journal of Marketing Research* **43**(2), 276–286.
- Macal, C. M. and North, M. J. (2007), Agent-based modeling and simulation: Desktop abms, in 'Proceedings of the 39th conference on Winter simulation: 40 years! The best is yet to come', IEEE Press, pp. 95–106.
- Macal, C. M. and North, M. J. (2010), 'Tutorial on agent-based modelling and simulation', *Journal of simulation* **4**(3), 151–162.
- March, S. T. and Smith, G. F. (1995), 'Design and natural science research on information technology', *Decision support systems* **15**(4), 251–266.
- Marsh, W. E. and Hill, R. R. (2008), 'An initial agent behaviour modelling and definition methodology as applied to unmanned aerial vehicle simulations', *International Journal of Simulation and Process Modelling* **4**(2), 119–129.
- Miguéis, V. L., Camanho, A. and e Cunha, J. F. (2013), 'Customer attrition in retailing: An application of multivariate adaptive regression splines', *Expert Systems with Applications* **40**(16), 6225–6232.
- Miguéis, V. L., Van den Poel, D., Camanho, A. S. and e Cunha, J. F. (2012), 'Modeling partial customer churn: On the value of first product-category purchase sequences', *Expert systems with applications* **39**(12), 11250–11256.
- Min, S., Zhang, X., Kim, N. and Strivastava, R. K. (2015), 'Customer acquisition and retention spending: An analytical model and empirical investigation in wireless telecommunications markets', *Journal of Marketing Research* .

- Mittal, V., Anderson, E. W., Sayrak, A. and Tadikamalla, P. (2005), 'Dual emphasis and the long-term financial impact of customer satisfaction', *Marketing Science* **24**(4), 544–555.
- Mutanen, T., Ahola, J. and Nousiainen, S. (2006), 'Customer churn prediction—a case study in retail banking', *Practical Data Mining: Applications, Experiences and Challenges* p. 13.
- Neslin, S. A., Gupta, S., Kamakura, W., Lu, J. and Mason, C. H. (2006), 'Defection detection: Measuring and understanding the predictive accuracy of customer churn models', *Journal of marketing research* **43**(2), 204–211.
- Newell, A., Simon, H. A. et al. (1972), *Human problem solving*, Vol. 104, Prentice-Hall Englewood Cliffs, NJ.
- Ngai, E. W., Xiu, L. and Chau, D. C. (2009), 'Application of data mining techniques in customer relationship management: A literature review and classification', *Expert systems with applications* **36**(2), 2592–2602.
- North, M. J. (2014), 'A theoretical formalism for analyzing agent-based models', *Complex Adaptive Systems Modeling* **2**(1), 3.
- North, M. J., Macal, C. M., Aubin, J. S., Thimmapuram, P., Bragen, M., Hahn, J., Karr, J., Brigham, N., Lacy, M. E. and Hampton, D. (2010), 'Multiscale agent-based consumer market modeling', *Complexity* **15**(5), 37–47.
- Oates, B. J. (2005), *Researching information systems and computing*, Sage.
- Oliver, R. L. (2010), *Satisfaction: A behavioral perspective on the consumer*, ME sharpe.

- Ottar Olsen, S., Alina Tudoran, A., Brunsø, K. and Verbeke, W. (2013), 'Extending the prevalent consumer loyalty modelling: the role of habit strength', *European Journal of Marketing* **47**(1/2), 303–323.
- Ozcan, A. (2014), 'Mobile phones democratize and cultivate next-generation imaging, diagnostics and measurement tools', *Lab on a chip* **14**(17), 3187–3194.
- Peppers, 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), 45–77.
- Peppard, J. (2000), 'Customer relationship management (crm) in financial services', *European Management Journal* **18**(3), 312–327.
- Phadke, C., Uzunalioglu, H., Mendiratta, V. B., Kushnir, D. and Doran, D. (2013), 'Prediction of subscriber churn using social network analysis', *Bell Labs Technical Journal* **17**(4), 63–75.
- Prahalad, C. K. and Ramaswamy, V. (2004), 'Co-creation experiences: The next practice in value creation', *Journal of interactive marketing* **18**(3), 5–14.
- Pries-Heje, J. and Baskerville, R. (2008), 'The design theory nexus', *Mis Quarterly* pp. 731–755.
- Rahman, S., Haque, A. and Ahmad, M. I. S. (2011), 'Choice criteria for mobile telecom operator: empirical investigation among malaysian customers', *International Management Review* **7**(1), 50–57.
- Raju, P. S., Bai, V. R. and Chaitanya, G. K. (2014), 'Data mining: Techniques for enhancing customer relationship management in banking and retail in-

- dustries', *International Journal of Innovative Research in Computer and Communication Engineering* **2**(1), 2650–2657.
- Rand, W. and Rust, R. T. (2011), 'Agent-based modeling in marketing: Guidelines for rigor', *International Journal of Research in Marketing* **28**(3), 181–193.
- Ribeiro Soriano, D., Jyh-Fu Jeng, D. and Bailey, T. (2012), 'Assessing customer retention strategies in mobile telecommunications: Hybrid mcdm approach', *Management Decision* **50**(9), 1570–1595.
- Richard, M.-O. and Chebat, J.-C. (2016), 'Modeling online consumer behavior: Preeminence of emotions and moderating influences of need for cognition and optimal stimulation level', *Journal of Business Research* **69**(2), 541–553.
- Richter, Y., Yom-Tov, E. and Slonim, N. (2010), Predicting customer churn in mobile networks through analysis of social groups., *in* 'SDM', Vol. 2010, pp. 732–741.
- Risselada, H., Verhoef, P. C. and Bijmolt, T. H. (2014), 'Dynamic effects of social influence and direct marketing on the adoption of high-technology products', *Journal of Marketing* **78**(2), 52–68.
- Rodan, A., Faris, H., Alsakran, J. and Al-Kadi, O. (2014), 'A support vector machine approach for churn prediction in telecom industry', *International journal on information* **17**(8), 3961–3970.
- Roosmand, O., Ghasem-Aghaei, N., Hofstede, G. J., Nematbakhsh, M. A., Baraani, A. and Verwaart, T. (2011), 'Agent-based modeling of consumer decision making process based on power distance and personality', *Knowledge-Based Systems* **24**(7), 1075–1095.

- Santouridis, I. and Trivellas, P. (2010), 'Investigating the impact of service quality and customer satisfaction on customer loyalty in mobile telephony in greece', *The TQM Journal* **22**(3), 330–343.
- Schaarschmidt, M. and Kilian, T. (2014), 'Impediments to customer integration into the innovation process: A case study in the telecommunications industry', *European Management Journal* **32**(2), 350–361.
- Seo, D., Ranganathan, C. and Babad, Y. (2008), 'Two-level model of customer retention in the us mobile telecommunications service market', *Telecommunications Policy* **32**(3), 182–196.
- Shah, S. R., Roy, R. and Tiwari, A. (2006), 'Technology selection for human behaviour modelling in contact centres'.
- Silva, K. and Yapa, S. (2013), 'Customer retention: with special reference to telecommunication industry in sri lanka'.
- Simon, H. A. (1996), *The sciences of the artificial*, Vol. 136, MIT press.
- Siu, N. Y.-M., Zhang, T. J.-F. and Yau, C.-Y. J. (2013), 'The roles of justice and customer satisfaction in customer retention: A lesson from service recovery', *Journal of business ethics* **114**(4), 675–686.
- Stelzner, M. A. (2011), 'Social media marketing industry report', *Social Media Examiner* **41**.
- Strauss, A. and Corbin, J. (1998), 'Basics of qualitative research: Procedures and techniques for developing grounded theory', ed: *Thousand Oaks, CA: Sage* .



- Sundarkumar, G. G. and Ravi, V. (2015), 'A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance', *Engineering Applications of Artificial Intelligence* **37**, 368–377.
- Svendsen, G. B. and Prebensen, N. K. (2013), 'The effect of brand on churn in the telecommunications sector', *European Journal of Marketing* **47**(8), 1177–1189.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. (2010), 'Sentiment strength detection in short informal text', *Journal of the American Society for Information Science and Technology* **61**(12), 2544–2558.
- Tsai, C.-F. and Lu, Y.-H. (2010), 'Data mining techniques in customer churn prediction', *Recent Patents on Computer Science* **3**(1), 28–32.
- Twomey, P. and Cadman, R. (2002), 'Agent-based modelling of customer behaviour in the telecoms and media markets', *info* **4**(1), 56–63.
- Vafeiadis, T., Diamantaras, K., Sarigiannidis, G. and Chatzisavvas, K. C. (2015), 'A comparison of machine learning techniques for customer churn prediction', *Simulation Modelling Practice and Theory* **55**, 1–9.
- Vag, A. (2007), 'Simulating changing consumer preferences: a dynamic conjoint model', *Journal of business research* **60**(8), 904–911.
- Vaishnavi, V. K. and Kuechler, W. (2015), *Design science research methods and patterns: innovating information and communication technology*, CRC Press.
- Vapnik, V. N. and Vapnik, V. (1998), *Statistical learning theory*, Vol. 1, Wiley New York.

- Verbeke, W., Dejaeger, K., Martens, D., Hur, J. and Baesens, B. (2012), ‘New insights into churn prediction in the telecommunication sector: A profit driven data mining approach’, *European Journal of Operational Research* **218**(1), 211–229.
- Verbeke, W., Martens, D. and Baesens, B. (2014), ‘Social network analysis for customer churn prediction’, *Applied Soft Computing* **14**, 431–446.
- Verbeke, W., Martens, D., Mues, C. and Baesens, B. (2011), ‘Building comprehensible customer churn prediction models with advanced rule induction techniques’, *Expert Systems with Applications* **38**(3), 2354–2364.
- Verbraken, T., Verbeke, W. and Baesens, B. (2013), ‘A novel profit maximizing metric for measuring classification performance of customer churn prediction models’, *Knowledge and Data Engineering, IEEE Transactions on* **25**(5), 961–973.
- von Alan, R. H., March, S. T., Park, J. and Ram, S. (2004), ‘Design science in information systems research’, *MIS quarterly* **28**(1), 75–105.
- Weber, R. P. (1990), *Basic content analysis*, number 49, Sage.
- Ye, Y., Liu, S. and Li, J. (2008), A multiclass machine learning approach to credit rating prediction, *in* ‘Information Processing (ISIP), 2008 International Symposium on’, IEEE, pp. 57–61.
- Zhang, T. and Zhang, D. (2007), ‘Agent-based simulation of consumer purchase decision-making and the decoy effect’, *Journal of Business Research* **60**(8), 912–922.
- Zhang, X., Liu, Z., Yang, X., Shi, W. and Wang, Q. (2010), Predicting customer churn by integrating the effect of the customer contact network, *in*

'Service Operations and Logistics and Informatics (SOLI), 2010 IEEE International Conference on', IEEE, pp. 392–397.

# Appendix

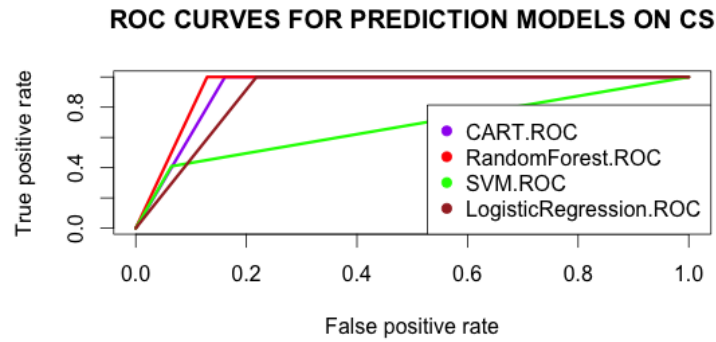


Figure 7.3: ROC curves for Customer Service (Twitter Dataset)

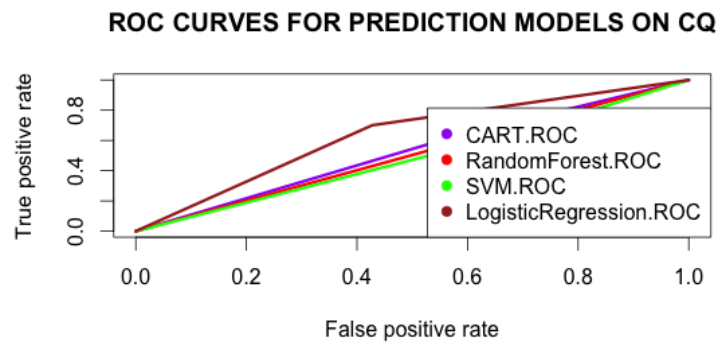


Figure 7.4: ROC curves for Coverage Quality (Twitter Dataset)

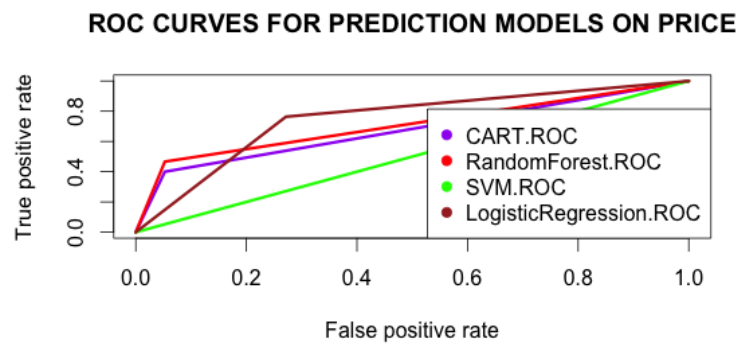


Figure 7.5: ROC curves for Coverage Quality (Twitter Dataset)

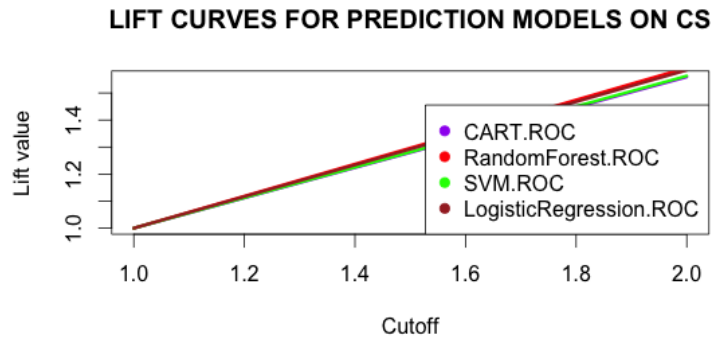


Figure 7.6: ROC curves for Customer Service (Twitter Dataset)

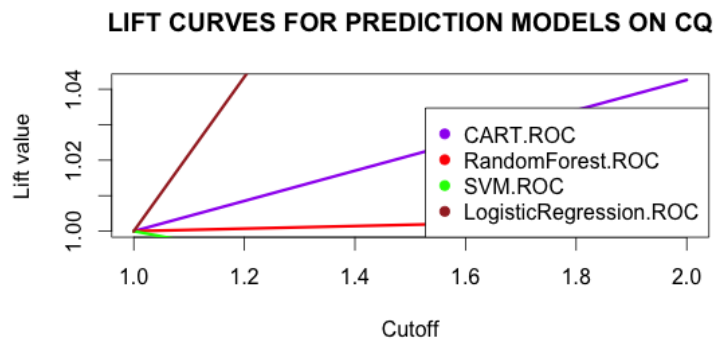


Figure 7.7: ROC curves for Coverage Quality (Twitter Dataset)

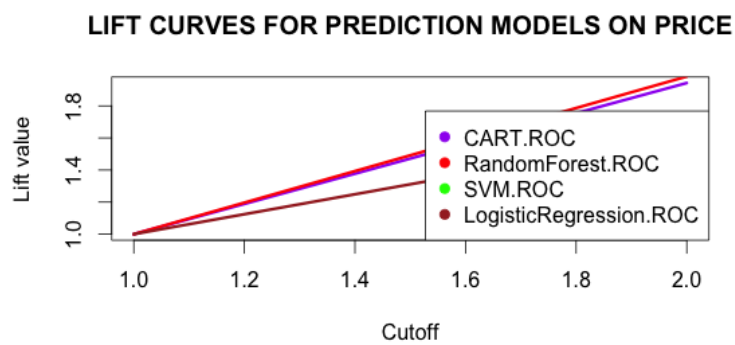


Figure 7.8: ROC curves for Coverage Quality (Twitter Dataset)

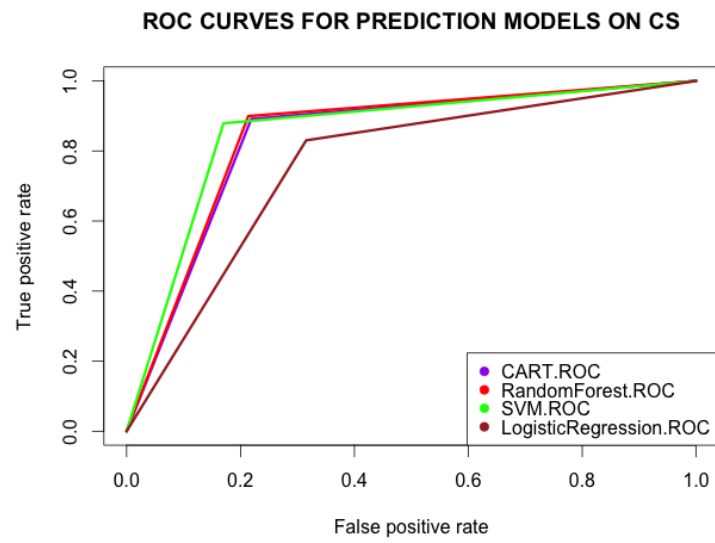


Figure 7.9: ROC curves for Customer Service (Telco Dataset)

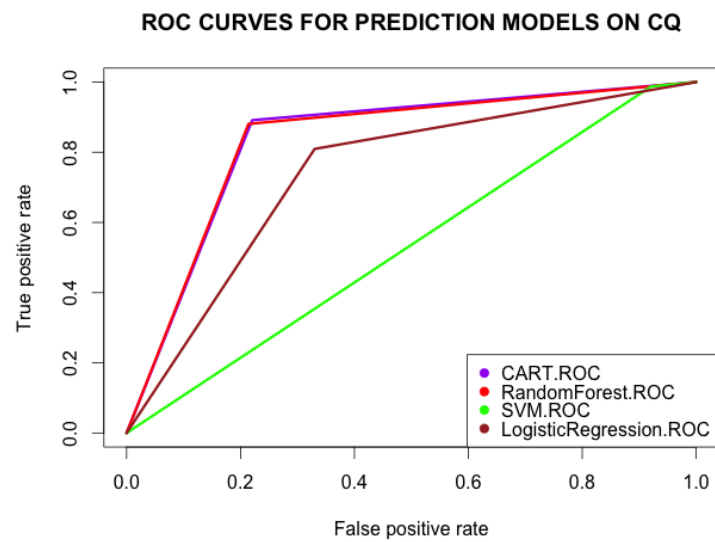


Figure 7.10: ROC curves for Coverage Quality (Telco Dataset)

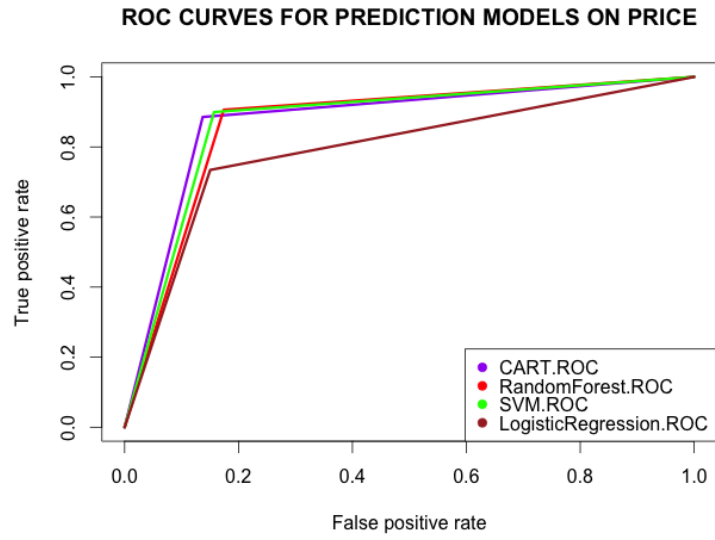


Figure 7.11: ROC curves for Coverage Quality (Telco Dataset)

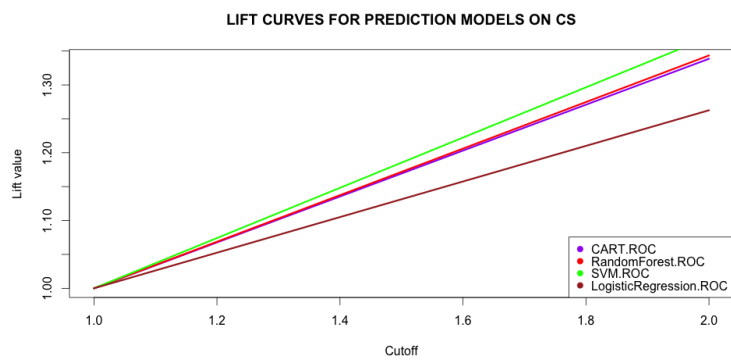


Figure 7.12: ROC curves for Customer Service (Telco Dataset)



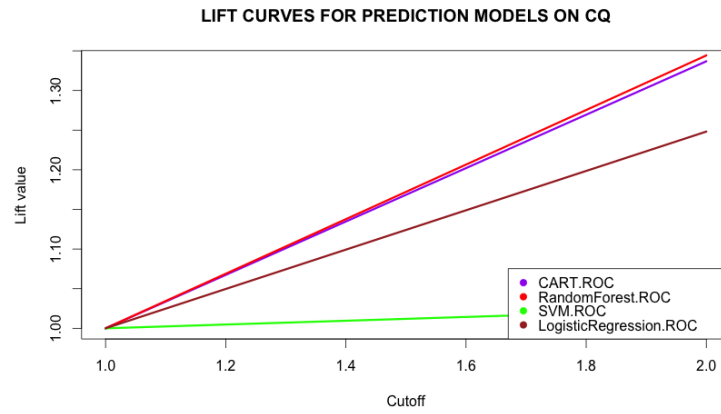


Figure 7.13: ROC curves for Coverage Quality (Telco Dataset)

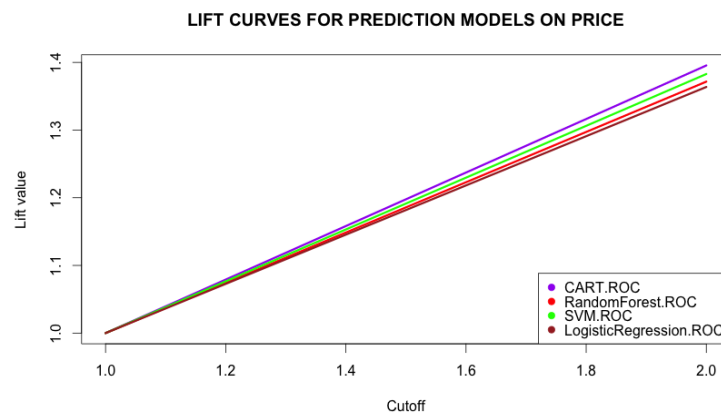
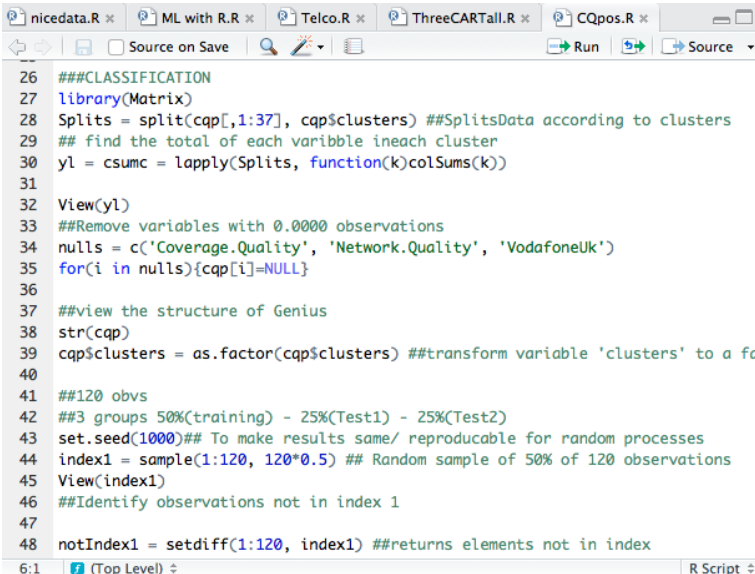


Figure 7.14: ROC curves for Coverage Quality (Telco Dataset)



```
26 ##CLASSIFICATION
27 library(Matrix)
28 Splits = split(cqp[,1:37], cqp$clusters) ##SplitsData according to clusters
29 ## find the total of each varibble ineach cluster
30 yl = csumc = lapply(Splits, function(k)colSums(k))
31
32 View(yl)
33 ##Remove variables with 0.0000 observations
34 nulls = c('Coverage.Quality', 'Network.Quality', 'VodafoneUk')
35 for(i in nulls){cqp[i]=NULL}
36
37 ##view the structure of Genius
38 str(cqp)
39 cqp$clusters = as.factor(cqp$clusters) ##transform variable 'clusters' to a fa
40
41 ##120 obvs
42 ##3 groups 50%(training) - 25%(Test1) - 25%(Test2)
43 set.seed(1000)## To make results same/ reproducable for random processes
44 index1 = sample(1:120, 120*0.5) ## Random sample of 50% of 120 observations
45 View(index1)
46 ##Identify observations not in index 1
47
48 notIndex1 = setdiff(1:120, index1) ##returns elements not in index
6:1 (Top Level) ↕ R Script ↕
```

Figure 7.15: Snapshot of R code for Churn Analysis