8-30-2006

Customer lifetime value : an integrated data mining approach

Chen XU

CUSTOMER LIFETIME VALUE:

AN INTEGRATED DATA MINING APPROACH

XU CHEN

MPHIL

LINGNAN UNIVERSITY

2006

CUSTOMER LIFETIME VALUE:

AN INTEGRATED DATA MINING APPROACH

by
XU CHEN

A thesis
submitted in partial fulfillment
of the requirements for the Degree of
Master of Philosophy

Lingnan University

2006

ABSTRACT

Customer Lifetime Value: an integrated data mining approach

by
XU Chen

Master of Philosophy

Customer Lifetime Value (CLV) ---which is a measure of the profit generating potential, or value, of a customer---is increasingly being considered a touchstone for customer relationship management. As the guide and benchmark for Customer Relationship Management (CRM) applications, CLV analysis has received increasing attention from both the marketing practitioners and researchers from different domains. Furthermore, the central challenge in predicting CLV is the precise calculation of customer's length of service (LOS). There are several statistical approaches for this problem and several researchers have used these approaches to perform survival analysis in different domains. However, classical survival analysis techniques like Kaplan-Meier approach which offers a fully non-parametric estimate ignores the covariates completely and assumes stationary of churn behavior along time, which makes it less practical. Further, segments of customers, whose lifetimes and covariate effects can vary widely, are not necessarily easy to detect. Like many other applications, data mining is emerging as a compelling analysis tool for the CLV application recently. Comparatively, data mining methods offer an interesting alternative with the fact that they are less limited than the conventional statistical approaches.

Customer databases contain histories of vital events such as the acquisition and cancellation of products and services. The historical data is used to build predictive models for customer retention, cross-selling, and other database marketing endeavors. In this research project we discuss and investigate the possibility of combining these statistical approaches with data mining methods to improve the performance for the CLV problem in a real business context. Part of the research effort is placed on the precise prediction of LOS of the customers in concentration of a real world business. Using the conventional statistical approaches and data mining methods in tandem, we demonstrate how data mining tools can be apt complements of the classical statistical models ---resulting in a CLV prediction model that is both accurate and understandable. We also evaluate the proposed integrated method to extract interesting business domain knowledge within the scope of CLV problem.

In particular, several data mining methods are discussed and evaluated according to their accuracy of prediction and interpretability of results. The research findings will lead us to a data mining method combined with survival analysis approaches as a robust tool for modeling CLV and for assisting management decision-making. A calling plan strategy is

designed based on the predicted survival time and calculated CLV for the telecommunication industry. The calling plan strategy further investigates potential business knowledge assisted by the CLV calculated.

DECLARATION

I declare that this is an original work based primarily on my own research, and I warrant that all citations of previous research, published or unpublished, have been duly acknowledged.

(Xu Chen)
2006.08.30

CERTIFICATE OF APPROVAL OF THESIS

CUSTOMER LIFETIME VALUE:
AN INTEGRATED DATA MINING APPROACH

by
XU CHEN
Master of Philosophy

Panel of Examiners:

	Chairman
	External Member
	Internal Member
	Internal Member

Chief Supervisor: Dr. Wong man-leung

	Approved for the Senate:
	Chairman, Research and Postgraduate Studies Committee
	Date

TABLE OF CONTENTS

LIST OF TABLES iii
LIST OF FIGURES iv
LIST OF ABBREVIATIONS v
ACKNOWLEDGEMENT vi
CHAPTER 1 INTROCUDTION 1
1.1 Overview 1
1.2 Research Motivation 2
1.3 Organization of the Thesis 4
CHAPTER 2 BACKGROUND AND LITERATURE REVIEW 5
2.1 Customer Relationship Management (CRM) 5
2.1.1 CRM Awareness and Development 5
2.1.2 CRM Process and Approaches 6
2.2 CRM and Data Mining 7
2.2.1 Applying Data Mining to CRM 9
2.3 Background of Data Mining Methods Applied 14
2.3.1 Linear Regression 14
2.3.2 Neural Networks 14
2.3.3 Classification and Regression Tree 18
2.3.4 Latent Class Regression 22
2.4 Background on Customer Lifetime Value analysis 25
2.4.1 Customer Lifetime Value---A Marketing Perspective 26
2.4.2 Customer Lifetime Value and Survival Analysis 34
2.4.3 Research in Survival Neural Networks 46
CHAPTER 3 RESEARCH FRAMEWORK 53
3.1 Research Model 53
3.1.1 Research Model for Method Development 53
3.1.2 Research Model for Business Meaning Generation 58
CHAPTER 4 SURVIVAL ANALYSIS AND DATA MINING APPROACHES 61
4.1 Data Description 61
4.2 Statistical Survival Analysis 66
4.2.1 Survival Analysis Assumptions 66
4.2.2 Facts of Censored Data 66
4.2.3 Survival Analysis Application 68
4.3 Survival Data Mining Approaches 72
4.3.1 Linear Regression 72
4.3.2 Regression Tree 74
4.3.3 Neural Networks 76
4.3.4 Latent Class Regression 78
CHAPTER 5 COMPARISIONS OF ALTERNATIVE METHODS 83
5.1 Accuracy Comparisons 83
5.2 Extreme Value Estimation Comparison 84

5.3 CLV Calculation 87
CHAPTER 6 FURTHER ANALYSIS 90
6.1 Motivation of Further Analysis... 90
6.2 A Calling Plan Recommendation Strategy... 91
6.2.1 Test Design (A Supervised Machine Learning Schema)....................................... 91
6.2.2 Experiments Schema and Test Results 94
CHAPTER 7 IMPLICATIONS AND CONTRIBUTION 98
7.1 Making Good Prediction Method for CLV Problem in Marketing Practice 98
7.1.1 Results from Combined Survival and Data Mining Method 98
7.1.2 Results from Further Analysis 99
7.2 Contribution of the Study 100
7.3 Further Research.. 101
BIBLIOGRAPHY 102

LIST OF TABLES

Table 1 Definition of Terms in Censoring Problem..35

Table 2 Variable Type Description..62

Table 3 Call Behavior Data Summary..62

Table 4 Variable Description and Relative Variable Importance 65

Table 5 Month Statistics for All Cases before Survival Analysis 67

Table 6 Month Statistics for Censored Cases before Survival Analysis 67

Table 7 Month Statistics for Uncensored Cases 68

Table 8 Survival Month Statistics for All cases after Kaplan-Meier Analysis..............................70

Table 9 Survival Month Statistics for Censored Cases after Kaplan-Meier Analysis...................71

Table 10 Survival Month Statistics for Uncensored Cases...71

Table 11 Full Regression Result...73

Table 12 Variable Importance of Regression Tree Modeling 75

Table 13 Detailed Statistics of Regression Models Generation...79

Table 14 Customer Profile among Four Latent Classes..80

Table 15 Month Summary among four latent classes..81

Table 16 Average monthly revenue, usage and contract value among each latent class................81

Table 17 Survival Time Predication Accuracy Comparison for Alternative Methods...................83

Table 18 CLV Calculation Comparison among Methods 88

Table 19 Contract Renewal Suggestions and Predefined Groups..93

Table 20 Misclassification Cost Matrix..94

Table 21 Classification Accuracy among Groups between Two Methods....................................96

LIST OF FIGURES

Figure 1 Full-Connected Network Model........16
Figure 2 Decision Tree Splitting Process Illustration........20
Figure 3 Type I Censoring........35
Figure 4 Right Censoring........37
Figure 5 Interval Censoring........38
Figure 6 Left Censoring........39
Figure 7 Example of Modular Neural Network for Survival Analysis.........47
Figure 8 Research Model for Survival Analysis and Data Mining Methods Selection........55
Figure 9 Two-Level CLV Calculation Method........57
Figure 10 Research Model for Predefining Classification Groups........59
Figure 11 Research Model for Further Analysis........60
Figure 12 Distribution of Month for All Cases........67
Figure 13 Distribution of Month for Censored Cases........67
Figure 14 Distribution of Month for Uncensored Cases........68
Figure 15 Distribution of Survival Month for All Cases........69
Figure 16 Distribution of Survival Month for All Cases........70
Figure 17 Distribution of Survival Month for Censored Cases........71
Figure 18 Distribution of Survival Month for Uncensored Cases........72
Figure 19 Distribution of Survival Time for Uncensored Cases of Linear Regression Model VS. Actual Survival Time Distribution. (a) Linear Regression, (b) Actual Survival Time.........74
Figure 20 Distribution of Survival Time for Uncensored Cases of Regression Tree Model VS. Actual Survival Time Distribution. (a) Regression Tree, (b) Actual Survival Time.........76
Figure 21 RMSE Comparison for Multi-Nodes Neural Networks........77
Figure 22 Distribution of Survival Time for Uncensored Cases of Neural Networks Model VS. Actual Survival Time Distribution. (a) Neural Networks Model, (b) Actual Survival Time.........78
Figure 23 Distribution of Survival Time for Uncensored Cases of Latent Class Regression VS. Actual Survival Time Distribution. (a) Latent Class Regression, (b) Actual Survival Time.........80
Figure 24 Accuracy Comparison for Alternative Methods (MAE, MSE, RMSE)........84
Figure 25 Scatter Plot for Predicted Survival Time of Linear Regression Model........85
Figure 26 Scatter Plot for Predicted Survival Time of Regression Tree Model........85
Figure 27 Scatter Plot for Predicted Survival Time of Neural Networks Model........85
Figure 28 Scatter Plot for Predicted Survival Time of Latent Class Regression Model........86
Figure 29 MAE Comparison of CLV Calculation........88
Figure 30 MSE Comparison of CLV Calculation........89
Figure 31 RMSE Comparison of CLV Calculation........89

LIST OF ABBREVIATIONS

B2B	Business to Business
B2C	Business to Customer
BIC	Bayesian Information Criterion
CART	Classification and Relationship Tree
CLV	Customer Lifetime Value
CRM	Customer Relationship Management
EW	Error derivative for the weight
GLM	General Linear Models
LC	Latent Class
LL	log-likelihood
LMS	Least mean square
LOS	Customer’s length of service
LR	Linear Regression
MAE	Mean absolute error
MSE	Mean squared error
NNs	Neural Networks
PH	Proportional hazards
RMSE	Root mean squared error
ROI	Return on investment

ACKNOWLEDGEMENT

Two years is just a blink of eyes compared to the life-long time, however something can be easily change with just a blink of eyes, something, say, me. At the time I am about to leave this campus I still cannot imagine my days to be with no protection of the identity as a "student". Luckily, there are a lot of people helped my on my way here, put faith on me and get me prepared for the future. I am quite grateful to Dr. WONG man-leung, a gifted researcher and teacher for his insightful guidance and support. Dr. WONG never ceased his concern on my study and campus life. Besides teaching me a lot about how to do the research, Dr. WONG allow me the free space to learn by myself. That is the fortune of my whole life. As a very responsible man, I can feel the high expectations from Dr. WONG. The pressure keeps stimulating me to improve.

Two years study changed me and will affect the rest of my life. A simple "Thank you" is not enough to express my grateful heart. I should also give my best regards to Prof. Daning SUN and all the colleagues in Department of Computing and Decision Sciences in Lingnan University for making the past two year incredibly rewarding.

And, I appreciate the care and help from my friends in the past two years, thanks especially to Carol Guo, Janet Kong, Bennet Yu, Eva Cheung, Elsa Ma, George Chen, Mingzhi Liu, and Wengjing Bao (name regardless of importance sequence). Their long-lasting friendships are the fortune of my life.

Last, all my growth is with my parents and my grandma. I hereby present this piece of work to them for their deep understanding and endless love.

CHAPTER 1 INTROCUDTION

1.1 Overview

The past few years have seen dramatic growth of Customer Lifetime Value (CLV) usage in business world. CLV---which is a measure of the profit generating potential, or value, of a customer---is increasingly being considered a touchstone for customer relationship management (CRM). As the guide and benchmark for Customer Relationship Management applications, CLV analysis has received increasing attention from both the marketing practitioners and academic researchers from different domains. Furthermore, the central challenge in predicting CLV is the precise calculation of customer's length of service (LOS). There are several statistical approaches for this problem and several researchers have used these approaches to perform survival analysis in different domains. However, classical survival analysis techniques suffer from certain drawbacks. For example, Kaplan-Meier approach (Kaplan and Meier, 1958), which offers a fully non-parametric estimate, ignores the covariates completely and assumes stationarity of churn behavior along time. This makes it a less practical method. Moreover, segments of customers, whose lifetimes and covariate effects can vary widely, are not necessarily (or respectively) easy to detect. As in many other applications, data mining is emerging as a compelling analysis tool for the CLV application recently (Monik, 2004; Neal, 2004; Rosset et al., 2003). Comparatively, data mining methods offer an interesting alternative with the fact that they are less limited than the conventional statistical approaches.

Customer databases contain histories of vital events such as the acquisition and cancellation of products and services. The historical data is used to build predictive models for customer retention, cross-selling, and other database marketing endeavors. In this research study, we discuss and investigate the possibility of combining these statistical approaches with data mining methods to improve the performance for the CLV problem within the scope of a business context. Specifically, in this study we take into consideration the telecommunications industry which represents a good example of service-centered and subscription-based business where customer and customer relation is the vital factor in success. Part of the research effort is placed on the precise prediction of LOS of the customers in concentration of a real world business, and CLV calculation afterwards. Using the conventional statistical approaches and data mining methods in tandem, we demonstrate how data mining tools can be apt complements of the classical statistical methods---resulting in an accurate CLV prediction model. We also evaluate the proposed method to extract interesting business domain knowledge for the CLV problem.

1.2 Research Motivation

Customer lifetime value is rapidly gaining acceptance as a metric to acquire, grow, and retain the "right" customer in CRM (Rajkumar and Kumar, 2004). However, many companies do not use CLV measurements judiciously, either they work with undesirable customers to begin with, or they do not know how to customize the customer's experience to create the highest value. The challenge that most marketing managers currently face is to achieve convergence between marketing actions (e.g., contacts across various channels, promotion strategy) and CLV. Specifically, they need to take all the data they have collected about customers and integrate them with how the firm interacts with its customers. In other words, the managers need to know how to calculate the CLV and maximize it under real business settings.

As stated above, there are both statistic and data mining approaches for the CLV problem. However, they both have limitations for the purpose of academic research or a company's practical application. For example, in pure-parametric approach spikes will be generated; in non-parametric approach covariate effects are totally ignored; and Neural Network models might encounter the over-fitting problem (Mani et. al., 1999). Thereby the idea of combining the conventional statistical methods and data mining approaches to get over these disadvantages could be potentially justified.

Through literature review, we believe that by combining the conventional statistical methods and data mining approaches, the accuracy and efficiency of the CLV calculation can be improved. Besides yielding good prediction accuracy, the combined method also provides us space for further exploration of the CLV problem for marketing practice purpose.

The main objective of this study is to investigate the possibility of combining the conventional statistical survival approaches with the data mining methods to handle the CLV problem within the scope of a real world scenario, namely, the telecommunication industry. This requires a study on both the conventional survival approaches and different data mining methods. It also requires a discussion of the results of the combined method, which might produce useful business meaning. In this research study, we present an integrated data mining method in a way that combines different data mining methods and conventional survival analysis approaches together. By conducting survival analysis and data mining method in tandem, a procedure has been designed for better predicting the survival time thus CLV for censored cases. In detail, because of their ability to impute the "partially missing" survival time for censored cases, we intend to appeal conventional survival analyses in this simple integration procedure first. However, as noted by many researchers that conventional survival analyses have their drawbacks and can not generate an accurate prediction result. Thus in the next step of this integration procedure, we apply data mining methods to provide better prediction results based on the imputed survival time derived from survival analysis.

Another aspect of the integrated method is shown in this research study. Based on the predicted survival time and calculated CLV, in this integration procedure, we apply Decision Tree Model to further investigate the business insights within the scope of telecommunication industry. The survival time is imputed and further predicted by conventional survival analysis and Latent Class Regression respectively. Moreover, we add more business knowledge into the original dataset and apply Decision Tree Model to generate decision rule sets accordingly. By conducting the strategy in this way, we add some extra advantage of decision rule sets into the CLV calculation thus to generate potential business knowledge. Overall, the research findings will lead us to an integrated data mining method combined with survival analysis approaches as a robust tool for modeling CLV and for assisting management decision-making. As a result of our discussions we will conclude that proper use of the combination of the statistical survival analysis and data mining methods will lead to a better support in building and maintaining the customer relationships based on CLV calculation in a marketing perspective.

1.3 Organization of the Thesis

This thesis is organized as follows. Chapter 1 gives an overview of the study. It identifies the general CLV problem in business world and stress the importance for this problem. It also presents the main idea of this thesis. Chapter 2 is a literature review including the prior studies on CRM, CLV, Conventional Survival Analysis and Data Mining techniques. The censored data problem is mentioned in this chapter. Chapter 3 elaborates the research methodology in detail. An integrated research model is given out in the chapter. Chapter 4 discusses the detailed experiments design. Data mining techniques including Linear Regression, Neural Networks, Regression Tree and Latent Class Regression are applied in this chapter. Chapter 5 mainly discusses the results based on both the statistics analysis as well as the data mining approaches. In Chapter 6, we will conduct a calling plan strategy to evaluate the proposed method to help marketing practice. We conclude the thesis by pointing out the implication for further studies in Chapter 7.

CHAPTER 2 BACKGROUND AND LITERATURE REVIEW

2.1 Customer Relationship Management (CRM)

2.1.1 CRM Awareness and Development

Fundamental changes in the market environment force marketers to reconsider marketing strategies. In today's increasingly global and competitive marketplace, customers have more options available to them than ever before. Many analysts, in fact, are calling this a "customer economy".

Customer relationship management (CRM) is the answer to the competition in this new "customer economy". It helps companies improve the profitability of their interactions with customer, while at the same time; makes the interactions appear friendlier through individualization. In addition, the growing awareness of retaining customers is becoming more profitable than acquiring new customers (Reichheld and Sasser, 1990) which leads companies to shift their focus from transaction-oriented to relationship-oriented or customer-based marketing.

Marketing theory and practice, accordingly, have become increasingly customer-centered during the past 40 years (Vavra, 1997). For example, marketing has decreased its emphasis on short-term transactions and has increased its focus on long-term customer relationships (e.g., Håkansson, 1982; Storbacka, 1994). The customer-centered viewpoint is reflected in the concepts and metrics that drive marketing management, including such metrics as customer satisfaction (Oliver, 1980), market orientation (Narver and Slater, 1990) and customer value (Bolton and Drew, 1991).

Attracting customers effectively and meeting their expectations for selection, price, quality, and service are essential to a customer value strategy. It is equally important, however, to identify and retain profitable customers, and increase their value over time. This requires the ability to anticipate customer needs and present attractive offers in the right way, and at the right time. The new requirements of customer-based marketing call for new management tools and/or the adaptation of existing tools such as CLV (Hoekstra and Huizingh, 1999). In recent years, customer lifetime value (CLV) and its implications have received increasing attention (Berger and Nada, 1998; Mulhern, 1999; Reinartz and Kumar, 2000).

2.1.2 CRM Process and Approaches

There has been considerable academic interest in Customer Relationship Management (CRM) strategies, applications and processes, with some 600 papers published in the year between 1997 to 2001 years (Romano, 2001). CRM is an integrated approach to identifying, acquiring, and retaining customers. Customer relationship management (CRM) in its broadest sense simply means managing all customer interactions. By enabling organizations to manage and coordinate customer interactions across multiple channels, departments, lines of business, and geographies, CRM helps organizations maximize the value of every customer interaction and drive superior corporate performance (SIEBEL Systems, 2002). To succeed with CRM, companies need to match products and campaigns to prospects and customers – in other words, to intelligently manage the customer life cycle. The customer life cycle has three stages:

■ Acquiring customers

The first step in CRM is to identify prospects and convert them to customers. This requires the understanding of current customers and a right projecting of future situations.

■ Increasing the value of customers

CRM can dramatically change the selling situation for a company by investigating potential selling opportunity and correctly recommending cross-selling and up-selling to the right customers. CRM helps company better understand their customers' needs and thereby increase its profitability. A good example is the personalization via web customer profiles.

■ Retaining good customers

As an industry enters the mature stage of its product life cycle, competition increases and it is more costly for firms to acquire new customers. For almost every company, the cost of acquiring a new customer exceeds the cost of keeping good customers. Generally, CRM applies churn project made use of three steps. The first thing a churn project needs to do is predicting which customers will leave the company. Next, the project needs to identify who are "good" customers (who are worth the effort of retaining). Predicting who will leave and who is profitable is not enough. Based on the results of the above two steps, a churn project needs to come up a marketing program to match the potential churners with the most appropriate offer.

2.2 CRM and Data Mining

Until recently, most CRM software focused on simplifying the organization and management of customer information. Such software, called operational CRM, focuses on creating a customer database that presents a consistent picture of the customer's relationship with the company and providing that information in specific applications. These include sales force automation and customer service applications, in which the company "touches" the customer. However, the sheer volumes of customer information and increasingly complex interactions with customers have propelled data mining to the forefront of making customer relationships profitable.

Data mining is a non-trivial extraction of novel, implicit, and actionable knowledge from large datasets. The techniques deal with extremely large datasets, discover the non-obvious, useful knowledge that can improve processes. They are technologies to enable data exploration, data analysis, and data visualization of very large databases at a high level of abstraction, without a specific hypothesis in mind. What is more, data mining methods possess sophisticated data search capability that uses statistical and machine learning algorithms to discover patterns and correlations in data.

Data mining process uses a variety of data analysis and modeling techniques to discover patterns and relationships in data that are used to understand what your customers want and predict what they will do. Data mining can help companies to select the right prospects on whom to focus, offer the right additional products to company's existing customers and identify good customers who may be about to leave. This results in improved revenue because of a greatly improved ability to respond to each individual contact in the best way and reduced costs due to properly allocated resources. CRM applications that use data mining are called analytic CRM.

In CRM, data mining is frequently used to assign a score to a particular customer or prospect indicating the likelihood that the individual behaves the way they are supposed to. For example, a score could measure the propensity to respond to a particular offer or to switch to a competitor's service. It is also frequently used to identify a set of characteristics (called a customer profile) that segments customers into groups with similar behaviors, such as buying a particular product. Particularly, data mining can improve the profitability in each stage of the customer life cycle when it is integrated with operational CRM systems or implement it as independent applications (acquiring, retaining, personalization, up-selling, and cross-selling). The route to a successful business requires the understanding of customers and their requirements, and data mining is the essential guide.

2.2.1 Applying Data Mining to CRM

In order to build good models for the CRM system, there are a number of steps must to be followed. The data mining process model described below is similar to other process models, differing mostly in the emphasis it places on the different steps.

As Werner and Kumar (2003) mentioned that data mining process is not linear – and will inevitably need to be looped back to previous steps. For example, in the explore data step (step 3) the information learned by the data exploration process itself may require adding new data to the data mining database.

The initial models built may provide insights that lead to create new variables. What is more, the basic steps of data mining for effective CRM are:

1. Define business problem

Each CRM application has one or more business objectives for which to build the appropriate model. Depending on the specific goal, such as "increasing the response rate" or "increasing the value of a customer," a very different model will be developed. An effective statement of the problem includes a way to measure the results of the CRM project.

2. Build marketing database

Steps two through four constitute the core of the data preparation. Together, they take more time and effort than all the other steps combined. There may be repeated iterations of the data preparation and model building steps as learning something from the model that suggests modifying the data. These data preparation steps may take anywhere from 50 to 90 percent of the time and effort for the entire data mining process (Werner and Kumar, 2003).

"You will need to build a marketing database because your operational databases and corporate data warehouse often don't contain the data you need in the form you need it. Furthermore, your CRM applications may interfere with the speedy and effective execution of these systems." (Werner and Kumar, 2003)

When building the marketing database we need to clean it up – if we want good models we must have clean data. The data that is needed may reside in multiple databases such as the customer database, product database, and transaction databases. This means we need to integrate and consolidate the data into a single marketing database and reconcile differences in data values from the various sources. Improperly reconciled data is a major source of quality problems. There are often large differences in the way that data is defined and used in different databases. Some inconsistencies may be easy to uncover, such as different addresses for the same customer. However, these problems are often subtle. For example, the same customer may have different names or, even worse, multiple customer identification numbers.

3. Explore data

Before we can build good predictive models, we must understand the data. Start by gathering a variety of numerical summaries (including descriptive statistics such as averages, standard deviations, and so forth) and looking at the distribution of the data. We may want to produce cross tabulations (pivot tables) for multi-dimensional data.

Graphing and visualization tools are a vital aid in data preparation and their importance for effective data analysis can't be overemphasized. Data visualization most often provides new perspectives that lead to new insights and success. Some common and very useful visual inspections are histograms or box plots that display distributions of values. We may also want to look at scatter plots in two or three dimensions of different pairs of variables. The ability to add a third, overlay variable greatly increases the usefulness of some types of graphs.

4. Prepare data for modeling

This is the final data preparation step before building models and the step where the most experiences and skills comes in. There are four main sub-steps to this step:

⇒ First, we need to select the variables on which to build the model. Ideally, we take all the variables we have, feed them to the data mining tool and let the data mining tool finds the best predictors. In practice, this is very involved. One reason is that the time it takes to build a model increases with the number of variables. Another reason is that blindly including extraneous variables can lead to models with less, rather than more, predictive power.

⇒ The next sub-step is to construct new predictors derived from the raw data. For example, forecasting credit risk using a debt-to-income ratio rather than just debt and income as predictor variables may yield results that are more accurate and easier to understand.

⇒ Next, decision on selecting a subset or sample of the data on which to build models needs to be made. If we have a lot of data, however, using all these data may take too long or require too much resource. Given a choice of either investigating a few models built on all the data or investigating more models built on a sample, the latter approach usually helps to develop a more accurate and robust model of the problem.

⇒ Last, variables need to be transformed in accordance with the requirements of the algorithm that is chosen to build the model.

5. Build model

The most important thing to remember about model building is that it is an iterative process. We need to explore alternative models to find the one that is most useful in solving our business problem. What we learn when searching for a good model may lead us to go back and make some changes to the data we are using or even modify the problem statement.

Most CRM applications are based on a protocol called supervised learning. The model starts with customer information for which the desired outcome is already known. For example, we may have historical data from a previous mailing list that is very similar to the one we are currently using. Or, we may have to conduct a test mailing to determine how people will respond to an offer. We then split this data into two groups. On the first group, we train or estimate the model on the first group. Then it is tested on the second group of data. A model is built when the cycle of training and testing is completed.

6. Evaluate model

"Perhaps the most overrated metric for evaluating the results is accuracy" (Bell et al., 2002). Suppose we have an offer to which only one percent of the people respond. A model that predicts "nobody will respond" is 99 percent accurate and 100 percent useless. Another measure that is frequently used is lift. Lift measures the improvement achieved by a predictive model. However, lift does not take into account cost and revenue so it is often preferable to look at profit or Return on Investment (ROI). Depending on whether the company chooses to maximize lift, profit or ROI, it may choose a different percentage of the mailing list to whom they send solicitations.

7. Deploy model and results

In building a CRM application, data mining is often a small, albeit critical, part of the final product. For example, predictive patterns through data mining may be combined with the knowledge of domain experts and incorporated in a large application used by many different kinds of people.

The way data mining is actually built into the application is determined by the nature of the customer interaction. There are two main ways to interact with the customers: the contact is inbound or outbound. The deployment requirements are quite different.

Outbound interactions are characterized by the company, which originates the contact, such as through a direct mail campaign. Thus, the company selects people to contact by applying the model to the customer database. Another type of outbound campaign is an advertising campaign. In this case, the company may match the profiles of good prospects shown by the developed model to the profile of the people advertisement would reach.

For inbound transactions, such as a telephone order, an Internet order or a customer service call, the application must respond in real time. Therefore, the data mining model is embedded in the application and actively recommends an action.

In either case, one key issue we must deal with in applying a model to new data is the transformations we used in building the model. Thus if the input data (whether from a transaction or a database) contains age, income, and gender fields, but the model requires the age-to-income ratio and gender has been changed into two binary variables, the input data must be transformed accordingly.

2.3 Background of Data Mining Methods Applied

2.3.1 Linear Regression

The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. Linear regression is an extension of the concept of simple regression. Rather than using values on one predictor variable to estimate values on a criterion variable, it uses values on several predictor variables. According to Harris (1975), linear-regression is a measure of the overall degree of relationship between the set of predictor variables and the criterion measure.

The linear regression equation is listed as follows:

$$Y=\beta_0+\beta_1X_1+ \beta_2X_2+ \beta_3X_3+\ldots+ \beta_{n-1}X_{n-1}+ \beta_nX_n$$

where n is the number of predictors, Y denotes the dependent variable, X_i, $1 <= i <= n$, represent the predictors, β_0 is a constant, and β_1, $1 <= i <= n$, are the coefficients of the corresponding predictors. The amount of each β_i means how much the relative importance of the corresponding predictive variable for a case's dependent variable value.

2.3.2 Neural Networks

Neural Network (NN), a procedure that mimics the processes of the human brain, have seen an explosion of interest over the last few years, and are being successfully applied across an extraordinary range of problem domains, in areas as diverse as finance, medicine, engineering, geology, and physics. Indeed, anywhere that there are problems of prediction and classification, neural networks are being introduced. This sweeping success can be attributed to two key factors:

- **Power.** Neural Networks are very sophisticated modeling techniques capable of modeling extremely complex functions. In particular, neural networks are nonlinear.

For many years, linear modeling has been the commonly used technique in most modeling domains since linear models have well-known optimization strategies. Where the linear approximation was not valid the models suffered accordingly. Neural networks also can model nonlinear functions with large numbers of variables.

- **Ease of use.** Neural Networks *learn by example*. An NN is configured for a specific application, such as pattern recognition or data classification, through a learning process. The neural network user gathers representative data, and then invokes *training algorithms* to automatically learn the structure of the data. Although the user does need to have some heuristic knowledge of how to select and prepare data, how to select an appropriate neural network, and how to interpret the results, the level of user knowledge needed to successfully apply neural networks is much lower than that required in using some other traditional nonlinear statistical methods.

The Basic Network Model:

To further the introduction of feed-forward neural networks, some terms must be defined first.

- Inputs neuron: Inputs neuron represents the raw information that is fed into the network.
- Hidden neuron: Neurons that play an internal role in the network. The activity of each hidden neuron is determined by the activities of the input neurons and the weights on the connections between the input and the hidden neurons.
- Outputs neuron: The behavior of the output neurons depends on the activity of the hidden neurons and the weights between the hidden and output neurons. The input, hidden, and output neurons need to be connected together.
- Weights: Each input comes via a connection that has a strength (or weight); these weights correspond to synaptic efficacy in a neuron.
- Neuron: The artificial neuron receives a number of inputs (either from original data, or from the output of other neurons in the neural network). Each neuron also has a

single threshold value. The weighted sum of the inputs is formed, and the threshold subtracted, to compose the activation of the neuron.

- The activation signal is passed through an activation function (also known as a transfer function) to produce the output of the neuron. The activation function is a function used to transform the activation level of a neuron into an output signal. Typically, activation functions have a "squashing" effect (i.e., through the function, it makes the neuron's output 0 if the input is less than zero, and 1 if the input is greater than or equal to 0).
- Epoch: During iterative training of a neural network, an Epoch is a single pass through the entire training set, followed by testing of the *testing* set.

The feed-forward structures have proved most useful in solving real problems. A simple network has a feed-forward structure: signals flow from inputs, forwards through any hidden neuron, eventually reaching the output units. Such a structure has stable behavior. A typical network has neurons arranged in a distinct layered topology. The input layer is not really neural at all. These neurons simply serve to introduce the values of the input variables. The hidden and output layer neurons are each connected to all of the units in the preceding layer. Figure 1 shows a fully connected neural network; from input to output, each neuron is connected to every neuron on the adjacent layers. Again, it is possible to define networks that are partially-connected to only some neuron in the preceding layer; however, for most applications fully-connected networks are more suitable.

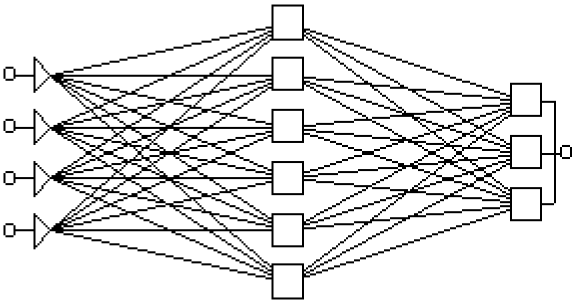


Figure 1 Full-Connected Network Model

As we can see from Figure 1, there are 3 layers in the network. There are *N* neurons in the first layer, where *N* equals number of inputs. There are *M* neurons in the output layer, where *M* equals number of outputs. For example, when you are building the network capable of predicting the stock price, you might want the (yesterday's) highest, lowest, and market close value as inputs and today's market close value as the output. What is more, there may have any number of neurons in the inner (also called "hidden") layers.

When the network is executed, the input variable values are placed in the input neurons, and then the hidden and output layer neurons are progressively executed (feed-forward the data into next layer). Each of them calculates its activation value by taking the weighted sum of the outputs of the neurons in the preceding layer, and subtracting the threshold. The activation value is passed through the activation function to produce the output of the neuron. Once the output neurons produce the predicted value, the results are compared with the actual value of expected outputs. What is more, the error, which is defined as the square of the difference between the actual and the desired outputs, is calculated. The weight of each connection is then changed so as to reduce the error. This training process is repeated until the network generate outputs in which error convergence reaches, i.e. the minimization of error between the desired and computed unit values. The aim is to determine a set of weights which minimizes the error. One well-known method, which is common to many learning paradigms, is the least mean square (LMS) convergence. To implement this procedure we need to calculate the error derivative for the weight (EW) in order to change the weight by an amount that is proportional to the rate at which the error changes as the weight is changed. One way to calculate the EW is to slightly perturb a weight and to observe how the error changes. But that method is inefficient because it requires a separate perturbation for each of the many weights. Another way to calculate the EW is to use the Back-propagation algorithm which has become nowadays one of the most important tools for training neural networks. It was developed independently by two teams, one (Fogelman-Soulie et al., 1987) in France and the other (Rumelhart et al., 1986) in U.S.

Once the adjustments are applied to the neurons in the output layer, the Neural Network can back-propagate the changes to the previous layers of the network. Adjustment will change weights of the input neurons of the neurons in the output layer. The system therefore progresses iteratively, through a number of epochs. When the entire network has been executed, the outputs of the output layer act as the output of the entire network. The detail discussion of backpropagation algorithm can be seen in (Rumelhart and McClelland, 1986).

2.3.3 Classification and Regression Tree

CART builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). The classical CART algorithm was popularized by Breiman et al. (1984).

There are two types of simple binary decision trees; regression and classification. Regression trees are appropriate where the dependent variable is a ratio scale data type. In other words, if the dependent variable can assume any value over the range of observations, and if the differences are quantitative and consistent, then we want a model that can predict these values and one that is not constrained to particular members. An example is number of sales per day. A regression model will predict somewhere between zero and a reasonable maximum number of sales for a given day based on the independent variables.

A classification tree is appropriate where the independent variable itself belongs to the data types nominal (named) or ordinal (ordered). For example, you may be interested in predicting who will or will not renew a subscription. These would be examples of simple binary classification problems, where the categorical dependent variable can only assume two distinct and mutually exclusive values. In other cases one might be interested in predicting which one of multiple different alternative consumer products a person decides to purchase, or which type of coffee most ordered a day. In those cases there are multiple categories or classes for the categorical dependent variable. A regression tree would not make sense in this case because it would predict unsuitable results such as a magnitude 2.76 or 4.89 event.

The type of model chosen, regression or classification, depends in part on the dependent variable type. A regression tree model can not be applied to classification data. However, a classification tree model can be applied to continuous data by generalizing the data into classes. Then a classification tree model could be used on number of events observed, where the observations have been re-expressed into nominal data.

Decision trees are an efficient form for representing decision processes for classifying patterns in data or piecewise constant functions in nonlinear regression. A tree functions in a hierarchical arrangement; data flowing “down” a tree encounters one decision at a time until a terminal node is reached. In most general terms, the purpose of the analyses via tree-building algorithms is to determine a set of “if-then” logical (split) conditions that permit accurate prediction or classification of cases. A particular variable enters the calculation only when it is required at a particular decision node, and only one variable is used at each decision node. These characteristics contrast with multivariate analyses and clustering, which use all critical variables for each case.

Decision tree analysis is a form of binary recursive partitioning. The term “binary” implies that each group of patients, represented by a “node” in a decision tree, can only be split into two groups. Thus, each node can be split into two child nodes, in which case the original node is called a parent node. The term “recursive” refers to the fact that the binary partitioning process can be applied over and over again. Thus, each parent node can give rise to two child nodes and, in turn, each of these child nodes may themselves be split, forming additional children. The term "partitioning" refers to the fact that the dataset is split into sections or partitioned.

Initially, tree modeling starts with a training set in which the classification label (say, "purchaser" or "non-purchaser") is known (pre-classified) for each record. Firstly, all of the records in the training set are grouped together at root node before further classification. The algorithm then systematically tries breaking up the records into two parts, examining one variable at a time and splitting the records on the basis of a dividing line in that variable (say, income > $75,000 or income <= $75,000). The object is to attain as homogeneous set of labels (say, "purchaser" or "non-purchaser") as possible in each partition. This splitting or partitioning is then applied to each of the new partitions. The process continues until no more useful splits can be found. Figure 2 illustrates the splitting process.

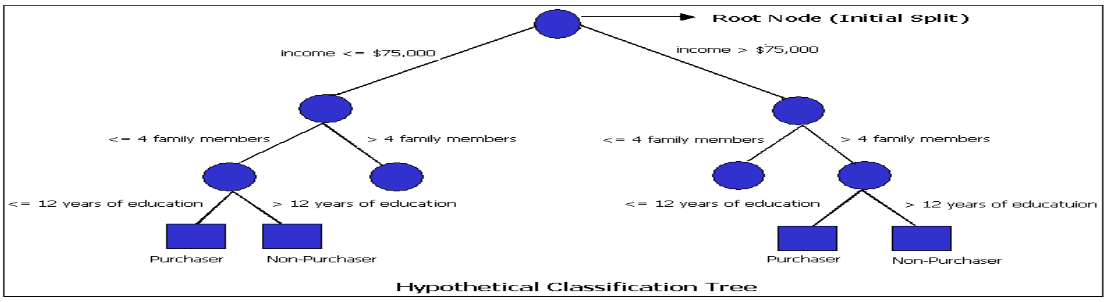


Figure 2 Decision Tree Splitting Process Illustration

In detail, decision tree analysis consists of four basic steps. The first step consists of tree building, during which a tree is built using recursive splitting of nodes. Each resulting node is assigned a predicted class, based on the distribution of classes in the learning dataset which would occur in that node and the decision cost matrix. The assignment of a predicted class to each node occurs whether or not that node is subsequently split into child nodes. The second step consists of stopping the tree building process. At this point a "maximal" tree has been produced, which probably greatly overfits the information contained within the learning dataset. The third step consists of tree "pruning," which results in the creation of a sequence of simpler and simpler trees, through the cutting off of increasingly important nodes. The fourth step consists of optimal tree selection, during which the tree which fits the information in the learning dataset, but does not overfit the information, is selected from among the sequence of pruned trees. Each of these steps will be discussed in more detail below.

1. Tree Building

Tree building begins at the root node, which includes all records in the learning dataset. Beginning with this node, the decision tree model finds the best possible variable to split the node into two child nodes. In order to find the best variable, the model checks all possible splitting variables (called splitters), as well as all possible values of the variable to be used to split the node. In choosing the best splitter, the model seeks to maximize the average "purity" of the two child nodes. A number of different measures of purity can be selected, loosely called "splitting criteria" or "splitting functions." The most common splitting function is the "Gini Index" for classification tree, and "Least Squares" for regression tree.

2. Stopping Tree Building

As mentioned above, the tree building process goes on until it is impossible to continue. The process is stopped when: (1) there is only one observation in each of the child nodes; (2) all observations within each child node have the identical distribution of predictor variables, making splitting impossible; or (3) an external limit on the number of levels in the maximal tree ("depth" option) has been reached.

3. Tree Pruning

In order to generate a sequence of simpler and simpler trees, each of which is a candidate for the appropriately-fit final tree, the method of "cost-complexity" pruning is used. This method relies on a complexity parameter, denoted α, whose initial value is zero and is gradually increased during the pruning process. Beginning at the last level (i.e., the terminal nodes) the child nodes are pruned away if the resulting change in the predicted misclassification cost is less than α times the change in tree complexity. Thus, α is a measure of how much additional accuracy a split must add to the entire tree to warrant the additional complexity. As α is increased, more and more nodes (of increasing importance) are pruned away, resulting in simpler and simpler trees.

4. Optimal Tree Selection

The maximal tree will always fit the learning dataset with higher accuracy than any other tree. The performance of the maximal tree on the original training set, termed the "resubstitution cost," generally greatly overestimates the performance of the tree on an independent set of data. This occurs because the maximal tree fits the patterns in the training set, which are unlikely to occur with the same pattern in a different set of data. The goal in selecting the optimal tree, defined with respect to expected performance on an independent set of data, is to find the correct complexity parameter α so that the information in the learning dataset is fit but not overfit. Usually this is done by setting up an independent testing set to evaluate a set of optimal tree candidates. The decision tree structures are evaluated on the testing set to ensure the optimal tree will accurately classify existing data and predict results. Thus the optimal tree which produces the lowest overall misclassification cost is selected.

As we can see, based on its complex model development schema, classification and regression tree method demonstrates itself a powerful and useful model for both classification and regression tasks.

2.3.4 Latent Class Regression

Over the past several years more significant books have been published on latent class (LC) and finite mixture models than any other class of statistical models (Agresti, 2002; Bartholomew and Knott, 1999; Skrondal and Rabe-Hesketh, 2004) The recent increase in interest in LC models is due to the development of extended computer algorithms, which allow today's computers to perform latent class analysis on data containing more than just a few variables. In addition, researchers are realizing that the use of latent class models can yield powerful improvements over traditional approaches to cluster, factor, regression, segmentation, and neural network applications.

Latent class analysis was originally introduced by Lazarsfeld (1950) as a way of explaining respondent heterogeneity in survey response patterns involving dichotomous items. Latent classes are unobservable (latent) subgroups or segments. It represents the dimensions which structure the cases with respect to a set of variables. Cases within the same latent class are homogeneous on certain criteria, while cases in different latent classes are dissimilar from each other in certain important ways. That is, latent class analysis divides the cases into latent classes which are "conditionally independent", meaning that the variables of interest are uncorrelated within any one class. Formally, latent classes are represented by K distinct categories of a nominal latent variable X. Since the latent variable is categorical, LC modeling differs from more traditional latent variable approaches such as factor analysis, structural equation models, and random-effects regression models that are based on continuous latent variables. In the context of marketing research, one will typically interpret the categories of these latent variables, the latent classes, as clusters or segments (Dillon and Kumar 1994; Wedel and Kamakura 1998).

LC models do not rely on the traditional modeling assumptions which are often violated in practice (linear relationship, normal distribution, homogeneity). Hence, they are less subject to biases associated with data not conforming to model assumptions. Also, for improved cluster or segment description (and prediction), the relationship between the latent classes and external variables (covariates) can be assessed simultaneously with the identification of the classes (clusters, segments). This eliminates the need for the usual second stage of analysis where a discriminant analysis is performed to relate the resulting clusters or factors obtained from a traditional cluster or factor analysis to demographic and other variables.

The LC Regression model, also known as the LC Segmentation model, is used to predict a dependent variable as a function of predictors. It includes a K-category latent variable, each category representing a homogeneous population (class, segment). What is more, different regressions are estimated for each population (for each latent segment), and it further classifies each case into segments and develops regression models simultaneously. Latent regression uses predictor and covariate variables to estimate a statistical model which can assign cases to each latent class, all without the traditional regression assumptions of normally distributed prediction error or of homogeneity.

The Latent Class Regression model is a model with (Vermunt and Van Dijk, 2001):

1. a single nominal latent variable x,
2. T_i replications or repeated observations of a single dependent variable y_{it} (the value of the dependent variable for case i at replication t) which may be nominal, ordinal, continuous, or a binomial or Poisson count,
3. Q numeric or nominal predictors z_{itq}^{pred} affecting y_{it} via a GLM (General Linear Models), where parameters may differ across latent classes,
4. zero, equality, fixed value, and order restrictions on regression coefficients,
5. R numeric or nominal covariates z_{ir}^{cov} affecting x.

The value of the dependent variable for case i at replication t is denoted by y_{it}, and its total number of replications by T_i. Covariates are variables influencing the latent variable. Predictors are variables that influence the dependent variable. Covariates are denoted by z_{ir}^{cov} and predictors by z_{itq}^{pred}, where the index t in z_{itq}^{pred} reflects that the value of a predictor may change across replications. A covariate, on the other hand, has the same value across all replications of a particular case. Note that, in fact, we are dealing with a two-level data set, where *t* indexes the lower-level observations within higher-level observation *i*. Covariates serve as higher-level exogenous variables and predictors as lower-level exogenous variables.

Based on the parameters discussed above, the most general probability structure that can be used in the Regression Module takes on the following form:

$$f(y_i \mid z_i^{\mathrm{cov}}, z_i^{pred}) = \sum_{x=1}^{K} P(x \mid z_i^{\mathrm{cov}}) \prod_{t=1}^{T_i} f(y_{it} \mid x, z_{it}^{pred})$$

As can be seen, we are specifying a model for $f(y_i \mid z_i^{\mathrm{cov}}, z_i^{pred})$, which is the probability density corresponding to a particular set of y_i values given a particular set of z_i^{pred} and z_i^{cov} values. The equation shows that the unobserved variable *x* intervenes between z_i^{pred} , z_i^{cov} and the y_i variables. Here, $P(x \mid z_i^{\mathrm{cov}})$ is the probability of belonging to a certain latent class given an individual's realized covariate values (the mixture weights), and $f(y_{it} \mid x, z_i^{pred})$ is the probability density of y_i given *x* and z_i^{pred} (the mixture densities). Thus, *x* variables may be influenced by z_i^{cov} variables, and y_i variables may be influenced by *x* and z_i^{pred} , z_i^{cov} variables.

2.4 Background on Customer Lifetime Value analysis

In the real business world, CRM tailors analytical solutions to business problems in many different industries: churn and retention analysis, fraud analysis, campaign management, credit and collection risk management, and more. CLV plays a major role in several of these applications, in particular churn analysis and retention campaign management (for communication industry especially), customer selection and marketing resource allocation (for B2B business). It can help to select the right prospects on whom to focus, offer the right additional products to your existing customers and identify good customers who may be about to leave.

Nowadays, CLV is widely used as the basis for evaluating CRM and Database Marketing initiatives, and is now identified as a standard by the Database Marketing Institute (Hughes, 2002). The idea is that the worth of a customer relationship to an organization can be evaluated by adding up the revenues and costs associated with servicing that customer over the lifetime of the relationship, taking into account future behaviors (such as churn and survival time) and the time value of money (Berger and Nada, 1998).

2.4.1 Customer Lifetime Value---A Marketing Perspective

2.4.1.1 Definition of CLV

Customer lifetime value (CLV) is usually defined as the total net income a company can expect from a customer (Novo, 2001). This is a long-term marketing principle which looks at the value of a customer over the entire time that they relate with the company. CLV in customer-based marketing reflects the value of a customer during the entire customer life cycle.

Courtheoux (1995) states that estimation of the lifetime value of customers enables marketers to evaluate ongoing programs to existing customers in terms of the changes wrought in the lifetime value of those customers. However literature shows that CLV calculations are mainly used to support decisions about acquiring new customers, for instance about the budget that can be spent on acquiring a new (comparable) group of customers (Jackson, 1994; Keane and Wang, 1995; Hughes, 1996). In customer-based marketing, CLV can be used for a wide range of marketing decisions with the assist by data mining techniques. Examples are creating, developing, and maintaining relationships.

Although there are several definitions of CLV, the ambiguousness still lies between these definitions and impedes the CLV application. For instance, an unhealthy amount of confusion today over the meanings of two of the most important terms in interactive marketing: *customer lifetime value* and *customer profitability*. An example can be seen in Reichheld and Sasser (1990). What is more, the acquisition cost, the scope of CLV analysis and some other factors challenge the researchers.

First, whether acquisition costs should be included in measures of CLV or not is a confusion discussed in many research studies.

Second, the time scope of CLV analysis needs to be clarified. The implication of time to CLV being the present value of future cash flows should be taken into serious consideration (Pfeifer et al., 2004). To make sense, the user must specify when "now" is. Not only does "now" determine how many periods to discount each cash flow, but "now" also determines which cash flows are in the future (and to be included) and which are in the past (to be excluded). And this becomes particularly important with respect to acquisition spending (Pfeifer et al., 2004).

Third, something is hard to be defined in the CLV analysis. In the application of acquiring a new customer, the CLV at this particular time can be defined relatively concise, however consider the situation of retaining an old customer, something becomes hard in progress. First of all, the referrals from the old customers can be very hard to define. Some factors like option leader, the word-of-mouth effect is notoriously difficult to measure on a practical basis (for example double counting). One example can be seen in The Jiffy Lube example provided by Hughes (1997). Secondly, the network effect (the cash flows expected from one customer depend on how many other customer relationships the firm has) is also hard to be materialized. Network effects create dependencies among the firm's customer relationships that will make it difficult to attribute the firm's cash flows to individual customer relationships (Pfeifer et al., 2004). Finally, the firm's relationships with the other players (suppliers, competitors, etc.) in this environment would be easily ignored and hard to be tracked.

2.4.1.2 CLV Calculation

When modeling CLV, different context (different business application) needs different issues to be taken into account. For instance, for the retention management, the CLV needs to be calculated before and after the retention effort, in other words, manager would need to calculate several CLV for each customer, each incentive in each different business period (Rosset et al., 2003).

In general, a CLV model has three basic components: customer's value over time, customer's length of service and a discounting factor, besides many other specific elements (Rosset et al., 2003).

- **Customer's value over time:**

The customer's value over time: $V(t)$ for $t \geq 0$, where t is time and $t = 0$ is the present. In practice, the customer's future value has to be estimated for current data, using business knowledge and analytical tools.

- **Customer's length of service:**

While the customer monthly margin can be obtained from an accounting model, the major problem is customer survival time (Mohammed and Kotze, 2005). The central challenge in predicting CLV is the production of estimated customer tenures with a given service supplier i.e., customer's length of service, based on information contained in company databases.

A length of service (LOS) model describes the customer's churn probability over time. This is usually described by a "survival" function $S(t)$ for $t \geq 0$, which describes the probability that the customer will still be active at time t. we can then define $f(t)$ as the customer's "instantaneous" probability of churn at time t: $f(t) = -dS/dt$. The quantity most commonly modeled, however is the hazard function $h(t) = -f(t)/S(t)$. Helsen and Schmittlein (1993) discussed why $h(t)$ is a more appropriate quantity to estimate than $f(t)$. The LOS model has to be estimated from current and historical data as well.

- **Discounting factor:**

As argued by Pfeifer et al. (2004), the definition of CLV should connect the value in CLV to the finance concept of present value. In so doing, most of the CLV calculation equations are consistent with the definitions of Customer Lifetime Value that mention "discounted," "present value" or "taking into account the time value of money." Thus a discounting factor is added in CLV calculation to project future profit into present value.

A discounting factor $D(t)$, which describes how much each $1 gained in some future time t is worth for us right now. This function is usually given based on business knowledge. Two popular choices are:

- Exponential decay: $D(t) = \exp(-at)$ for some $a \geq 0$ (a = 0 means no discounting)
- Threshold function: $D(t) = I(t \leq T)$ for some $T > 0$ (where I is the indicator function).

Given these three components, we can write the explicit formula for a customer's CLV as follows:

$$CLV = \int_0^\infty S(t)V(t)D(t)dt$$

While this formula is attractive and straight-forward, the essence of the challenge lies, of course, in estimating the $V(t)$ and $S(t)$ components in a reasonable way. In different business settings, for each different marketing objective, the CLV can be calculated differently. And for other factors, like competitors, referrals, and relationship scenario can add incremental trouble for this CLV application.

We can build models of varying structural and computational complexity for these quantities, for example, for LOS we can use a highly simplistic model assuming constant churn rate—so if we observe 5% churn rate in the current month, we can set $S(t) = 0.95^t$. This model ignores the different factors that can affect churn—a customer's individual characteristics, contracts and commitments, etc. On the other hand, we can build a complex proportional hazards model, using hundreds of customer properties as predictors. Such a model can turn out to be too complex and elaborate, either because it is modeling "local" effects relevant for the present only and not for the future, or because there is not enough data to estimate it properly. Thus to build practical and useful analytical models we have to find the one, which makes effective and relevant use of the data available to us.

2.4.1.3 Practical Value Calculations

Calculating a customer's current value is usually a straightforward calculation based on the customer's current or recent information: usage, price plan, payments, collection efforts, calls center contacts, etc.

The statistical techniques for modeling customer value along time include forecasting, trend analysis and time series modeling. However, the complexity of modeling and predicting the various factors that affect future value: seasonality, business cycles, economic situation, competitors, personal profiles and more, make future value prediction a highly complex problem. The solution in CLV applications is usually to concentrate on modeling LOS, while either leaving the whole value issue to the experts (Mani et al., 1999), or considering customers' current value as their future value (Novo, 2001).

Some side effects of the B2C or B2B communication can raise huge difficulty to calculate CLV in a more precise way. These factors are network effect, option leader, word-of-mouth, penetration ration, budget constraint, seasonality, business cycle, economic situation, personal profiles, channel communications, share-of-wallet, competitor effect, size of the market etc. Relatively speaking, these factors are hard to calculate in a precise way since their nature of ambiguousness.

The simplest way to deal with these factors is just considering the current value as the future value, thus exclude those factors completely (Novo, 2001). Despite the difficulty, many researchers use managerial methods to estimate those factors. For example, manager interviews (Rust et al., 2004), judgment-based estimate (Blattberg and Deighton, 1996). We will not elaborate them further since it is out of the scope of this research project.

2.4.1.4 Current Use of CLV in Practice

In the empirical study conducted by Hoekstra and Huizingh (1999), they investigated the CLV practice in the real world business, whether or not customer CLV was calculated in their company and the level of management used CLV information. Some of their study results are listed below.

CLV was calculated by 24% of the companies. When comparing between industries, the highest scores were found for the publishing companies (35%) and the lowest for the automobile dealers (12%). Performing CLV analyses seems to be a quite recent phenomenon. Almost all companies (90%) have calculated CLV for five years or less; 45% of these companies have calculated for three years or less.

With respect to the management level at which the CLV information is used they distinguished between three levels: top-management, marketing management, and operational management. Not surprisingly, CLV information is used mostly by marketing management (91%). Top management follows with 50%, and CLV analyses are used by operational management in only 43% of the companies. On average 1.8 management levels use CLV information, while in five companies (23%) all three management levels use CLV information.

Either calculating or not calculating CLV is not related to the general background variables such as the number of employees, revenues, market share, and number of customers, the average yearly number of contracts per customer, and the average yearly number of transactions per customer. However, two more specific characteristics, the importance of direct marketing and the sophistication of the customer information system, showed significant results. The larger the proportion of revenues following from direct marketing, the more often companies calculate customer CLV (Mann-Whitney test, $p <.00$). The more sophisticated the customer information system, the more likely that CLV are calculated (t - test, $p =.055$). Finally, they measured the relationship between the degree of market orientation and either calculating or not calculating CLV. For measuring market orientation they applied a slightly modified version of the Narver and Slater (1990) scale. Although the organizations that do calculate CLV have a higher mean for each of the three components of market orientation (customer orientation, competitor orientation, and interfunctional coordination), the differences are not significant.

2.4.1.5 Applications of CLV

As Mani et al. (1999) said "In the realm of CRM, modeling CLV has a wide range of applications", since it deals with the issue managers feel most interesting, the value. To be specific, these application areas can be churn and retention analysis, fraud analysis, campaign management, credit and collection risk management, cross-and-up selling, customer selection, and resource allocation. And these application lies in all kinds of industries, especially in direct marketing, telecommunication company, marketing oriented company (both manufactures and retailers) and so forth. Some of the applications of CLV analysis are listed below.

- Special services (e.g., premium call centers and elite service) and offers (concessions, upgrades, etc.) based on CLV- the more valuable a customer, the more irresistible services and offers could be, subject to satisfactory profit margins for the business.
- Targeting and managing unprofitable customers.
- Segmenting customers, marketing, pricing and promotional analysis based on CLV.
- Sizing and planning for future market opportunity based on cumulative CLV.

Some of these applications would use a single CLV score computed for every customer. Other applications require a separation of the survival time and value component for effective implementation, while even others would sue either the survival or value term and ignore the other components of CLV.

2.4.1.6 Strategy Based on CLV Calculation

One method of customer segmentation concerning the two dimensions: the customer's present value and its lifetime value. Compare those two dimensions will give the company some business insights to the marketing group. For instance, some customer might have small present value but high values in the future, who is very likely to be ignored because his low present value.

Ultimately, a company wants to know the financial impacts that will result from various marketing actions, this knowledge is essential if competing marketing initiatives are to be evaluated. A company may attempt to improve its customer value by making improvements in the drivers, or it may drill down further to improve sub-drivers that influence the drivers (e.g., improving dimensions of advertising awareness). This requires the measurement of customer perceptions of the sub-drivers about which the firm wanted to know more, the problem here is the difficulty to calculate the change on the individual level for each customer. Therefore, the CLV concept is applied.

The strategy ought to be applied lies through the company operation channel from marketing to Research and Development (R&D). Some of them are listed below:

- Attracting prospects with the highest potential lifetime value
- Forging stronger, more profitable relationships with existing customers
- Allocating the proper resources to those customers who are most likely to drive revenue and profit growth
- Supporting decisions for acquiring, retaining, growing and reactivating profitable customers
- Improve the channel communication
- Better the product design strategies

2.4.2 Customer Lifetime Value and Survival Analysis

2.4.2.1 Problem of Censoring

Imagine that you are a researcher in a telecommunication company who is studying the effectiveness of a new calling plan for extending customers' duration of relationship. The major variable of interest is the number of days that the respective customer stays in the company. In principle, one could use the standard parametric and nonparametric statistics for describing the average duration time, and for comparing the new calling plan with traditional marketing practice. However, at the end of the study there will be customers who stay over the entire study period; there will be other customers with whom we will have lost contact. Surely, one would not want to exclude all of those customers from the study by declaring them to be missing data (since most of them are "survivors" and, therefore, they reflect on the success of the new calling plan strategy). Those observations, which contain only partial information, are called censored observations (e.g., "customer 'A' stays over the entire study period".) The term censoring was first used by Hald (1949).

Normally, censored observations can be roughly divided into two censored types, which are based on their way of censoring and testing schema. Before further discussion, some of the terms are defined in Table1.

Terms	Notes
Event	the exact time point of interest (failure) was observed for an event
Time origin	admission and randomization to the study, birth, and contract start etc.
Endpoint	occurrence of an event
	— The time to the failure of a physical component (mechanical or electrical)
Time to failure	— The time to the death of a biological unit (patient, animal, etc.)
	— The time to the stop of a business relationship (contract, service subscription, etc.)

Table 1 Definition of Terms in Censoring Problem

Censored Type I Data:

During the T hours of test, we observe r failures (where r can be any number from 0 to n, which is the number of units to be tested). The (exact) failure times are t_1, t_2, ..., t_r and there are (n - r) units that survived the entire T-hour test without failing. Note that T is fixed in advance and r is random, since we don’t know how many failures will occur until the test is run. Note also that we assume the exact times of failure are recorded when there are failures. Figure 3 elaborates this censoring condition in details below.

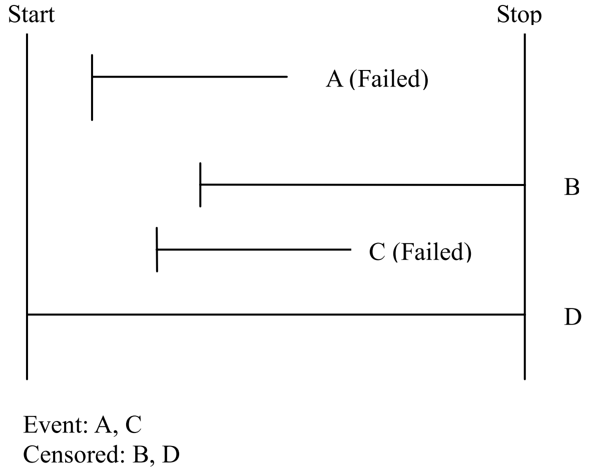


Figure 3 Type I Censoring

Another (much less common) way to test is to decide in advance that you want to see exactly r failure times and then test until they occur. For example, you might put 100 events on test and decide you want to see at least half of them fail. Then r = 50, but T is unknown until the 50th fail occurs. This is called **Censored Type II data**.

Censored Type II Data:

We observe t_1, t_2, ..., t_r, where r is specified in advance. The test ends at time T = t_r, and (n-r) units have survived. Again we assume it is possible to observe the exact time of failure for failed events.

Type II censoring has the significant advantage that how many failure times the test will yield is known in advance - this helps enormously when planning adequate tests. However, an open-ended random test time is generally impractical from a management point of view and this type of testing is rarely seen.

What is more, in detail, there are three types of possible censoring schemes, right censored (also called suspended data), interval censored, and left censored.

Right Censored (Suspended):

The most common case of censoring is what is referred to as right censored data, or suspended data. In the case of life data, these data sets are composed of units that did not fail. For example, if we tested four units and only two had failed by the end of the test, we would have suspended data (or right censored data) for the two not failed units. The term “right censored” implies that the unit of interest (i.e. the time-to-failure) is to the right of our data point. In other words, if the units were to keep on testing, the failure would occur at some time after our data point (or to the right on the time scale). Figure 4 elaborates this censoring condition in details below.

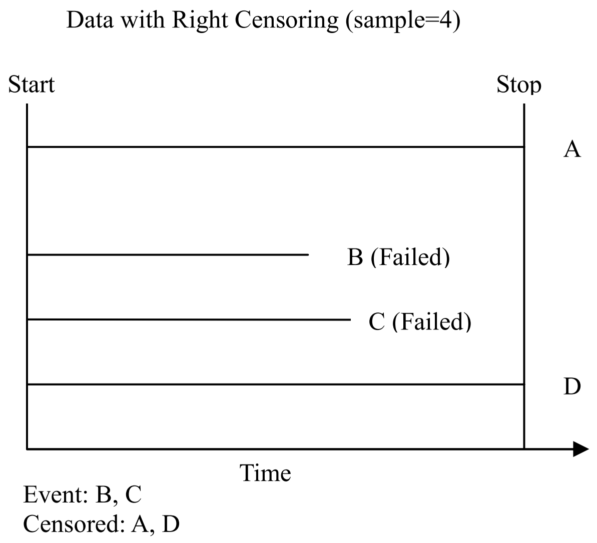


Figure 4 Right Censoring

Interval Censored:

The second type of censoring is commonly called interval censored data. Interval censored data reflects uncertainty as to the exact times the units failed within an interval. This type of data frequently comes from tests or situations where the objects of interest are not constantly monitored. If we are running a test on four units and inspecting them every 10 days, we only know that a unit failed or did not fail between inspections. More specifically, if we inspect a certain unit at 10 days and find it is operating and then perform another inspection at 20 days to find that the unit is no longer operating, we know that a failure occurred in the interval between 10 and 20 days. In other words, the only information we have is that it failed in a certain interval of time. This is also called inspection data by some authors. It must be noted that the interval censored cases are not failed events since their exact failure time is not known for sure. We only know the interval of their failure time. Figure 5 elaborates this censoring condition in details below.

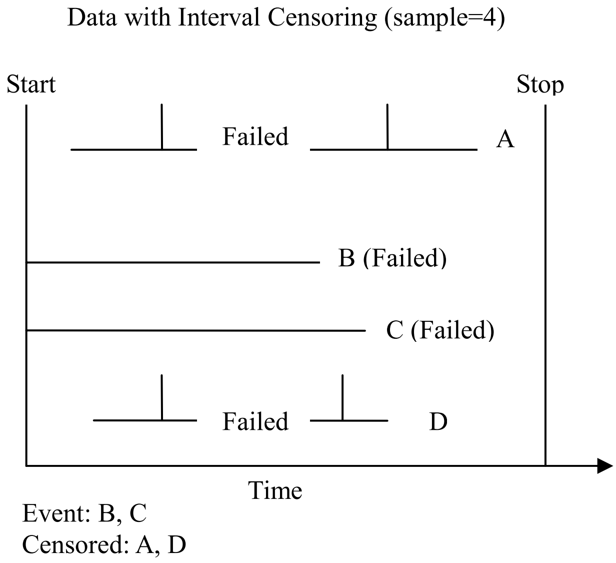


Figure 5 Interval Censoring

Left Censored:

The third type of censoring is similar to the interval censoring and is called left censored data. In left censored data, a failure time is only known to be before a certain time. For instance, we may know that a certain unit failed sometime before 10 days but not exactly when. In other words, it could have failed any time between 0 and 10 days. This is identical to interval censored data in which the starting time for the interval is zero. Same as interval censored cases, the left censored cases are not failed events since their exact failure time are not known. We only know their failures happen before a certain time point. Figure 6 elaborates this censoring condition in details below.

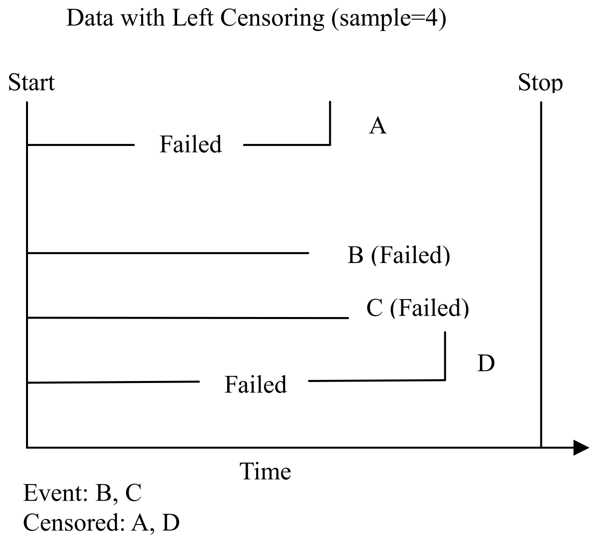


Figure 6 Left Censoring

2.4.2.2 Research in Handling Censored Data

As mentioned above, CLV has three components, namely, value, length and discounting rate. Somehow, in the literature, there are different focuses from different sectors of the company and different research area personnel; however, since the data for the value of a customer can be extracted from the data base, the major focus has been shifted to how long a customer will be in the business with the company. Moreover, faced with the censored data problem mentioned above, approaches called “Survival Analysis” were introduced to produce imputation for those “partially missing” censored data.

Conventional Survival Analysis

Survival analysis is a loosely defined statistical term that encompasses a variety of statistical techniques for analyzing censored data. The techniques of conventional survival analysis were primarily developed from the medical and biological sciences, but they are also widely used in the social and economic sciences, as well as in engineering (reliability and failure time analysis). There are four classical statistical approaches for the analysis of survival data, largely distinguished by the assumptions they make about the parameters of the distribution(s) generating the observed survival times. This is usually done by specifying two other mathematically equivalent functions, the survival function $S(t)$, the instantaneous probability function $f(t)$ and the hazard function $h(t)$ defined as follows:

- $S(t) = P(T > t), where\ P(T > t)$ is the probability that a subject will survive time period *t*.
- $f(t) = -dS/dt$ is the customer's "instantaneous" probability of churn at time *t.*
- $h(t) = -f(t)/S(t)$ The hazard function tries to quantify the instantaneous risk that an event will occur at time *t* given that the subject has survived to time *t*.

We now present a brief description of common Survival Analysis approaches and their possible use in LOS modeling. Detailed discussion of prevalent Survival Analysis approaches can be found in the literature, (Venables and Ripley, 1999).

Life Table Analysis:

The most straightforward way to describe the survival in a sample is to compute the Life Table (also called a mortality table or actuarial table). The life table technique is one of the oldest methods for analyzing survival (failure time) data (Berkson and Gage, 1950; Cutler and Ederer, 1958; Gehan, 1969). This table can be thought of as an "enhanced" frequency distribution table. The distribution of survival times is divided into a certain number of intervals. For each interval, the number and proportion of cases or objects that entered the respective interval "alive", the number and proportion of cases that failed in the respective interval (i.e., number of terminal events, or number of cases that "died"), and the number of cases that were lost or censored in the respective interval can be computed then.

Based on those numbers and proportions, several additional statistics can be computed:

- **Number of Cases at Risk:** This is the number of cases that entered the respective interval alive, minus half of the number of cases lost or censored in the respective interval.

- **Proportion Failing:** This proportion is computed as the ratio of the number of cases failing in the respective interval, divided by the number of cases at risk in the interval.

- **Proportion Surviving:** This proportion is computed as 1 minus the proportion failing.

- **Cumulative Proportion Surviving (Survival Function):** This is the cumulative proportion of cases surviving up to the respective interval. Since the probabilities of survival are assumed to be independent across the intervals, this probability is computed by multiplying out the probabilities of survival across all previous intervals. The resulting function is also called the survivorship or survival function.

- **Probability Density:** This is the estimated probability of failure in the respective interval, computed per unit of time, that is:

  $$F_i = \frac{(P_i - P_{i+1})}{w_i}$$

  In this formula, F_i is the respective probability density in the i^{th} interval, P_i is the estimated cumulative proportion surviving at the beginning of the i^{th} interval (at the end of interval i-1), P_{i+1} is the cumulative proportion surviving at the end of the i^{th} interval, and w_i is the width of the respective interval.

- **Hazard Rate:** The hazard rate (the term was first used by Barlow et al. (1963)) is defined as the probability per time unit that a case that has survived to the beginning of the respective interval will fail in that interval. Specifically, it is computed as the number of failures per time units in the respective interval, divided by the average number of surviving cases at the mid-point of the interval.

- **Median Survival Time:** This is the survival time at which the cumulative survival function is equal to 0.5. Other percentiles (25th and 75th percentile) of the cumulative survival function can be computed accordingly. Note that the 50th percentile (median) for the cumulative survival function is usually not the same as the point in time up to which 50% of the sample survived. (This would only be the case if there were no censored observations prior to this time).

As we can see from the discussion above, Life table analysis is an effective way to present and evaluate survival data in a number of circumstances. The summary tables and survival curves that are commonly used in other more complex survival analysis are frequently derived through life table analysis. A basic understanding of life table construction is of benefit to our further discussion of other more advance survival analysis approaches.

Non-Parametric Method:

Rather than classifying the observed survival times into a life table, we can estimate the survival function directly from the continuous survival or failure times. Intuitively, imagine that we create a life table so that each time interval contains exactly one case. The Kaplan-Meier estimator (Kaplan and Meier, 1958) offers a fully non-parametric estimate for *S*(*t*) by averaging over the data:

$$S(t) = S(t-1) \times (1 - h(t))$$

Where

- *S*(*t-1*) is the survival probability for the time period *t-1.*
- $h(t) = {}^{d_t}\!/_{n_t}$, where $\mathrm{d_t}$ represents the number of deaths in period *t*, and n_t represents the subjects at risk in that period.

The advantage of the Kaplan-Meier method over the life table method for analyzing survival and failure time data is that the resulting estimates do not depend on the grouping of the data (into a certain number of time intervals). Actually, the Kaplan-Meier method and the life table method are identical if the intervals of the life table contain at most one observation.

However, the problem may arise when it is hard to obtain the historical data of the customers who has already left the company. What is more, the Kaplan-Meier approach ignores the covariates completely and assumes stationarity of churn behavior along time, which makes it less practical.

Pure Parametric Method:

In summary, the life table gives us a good indication of the distribution of failures over time. However, for predictive purposes it is often desirable to understand the shape of the underlying survival function in the population. Pure parametric survival models (Gamel and Vogel, 1997) estimate the effects of covariates (subject variables whose values influence lifetimes, i.e., independent variables) by presuming a lifetime distribution $S(t)$ of a known parametric form with the parameters depending on the covariates, including t. The major distributions that have been proposed for modeling survival or failure times are the exponential (and linear exponential) distribution, the Weibull distribution of extreme events, and so on.

Semi-Parametric Method:

Common research question in medical, biological, or engineering (failure time) research is to determine whether or not certain continuous (independent) variables are correlated with the survival or failure times. There are two major reasons why this research issue cannot be addressed via straightforward multiple regression techniques:

First, the dependent variable of interest (survival/failure time) is most likely not normally distributed -- a serious violation of an assumption for ordinary least squares multiple regressions. Survival times usually follow an exponential or Weibull distribution. Second, there is the problem of censoring, that is, some observations will be incomplete.

Semi-parametric approaches, such as the Cox proportional hazards (PH) model (Cox, 1972) is the most general of the regression models because it is not based on any assumptions concerning the nature or shape of the underlying survival distribution. The model assumes that the underlying hazard rate (rather than survival time) is a function of the independent variables (covariates); no assumptions are made about the nature or shape of the hazard function. Thus, in a sense, Cox's regression model may be considered to be a semi-parametric method. The Cox PH model assumes a model for the hazard function *h*(*t*) of the form:

$$h_i(t) = [h_0(t)]e^{b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}}$$

where

- $h_i(t)$ is the hazard rate for the i^{th} case at time t
- $h_0(t)$ is the baseline hazard at time t
- p is the number of covariates
- b_j is the value of the j^{th} regression coefficient
- x_{ij} is the value of the j^{th} covariate of the i^{th} case

Therefore, we can see there is a fixed parametric linear effect (in the exponent) for all covariates, except time, which is accounted for in the time-varying "baseline" hazard $h_0(t)$. This equation says that the hazard for a subject i at time t is the product of an unspecified, positive baseline hazard function $h_0(t)$, and a linear function of a vector of inputs x_i which is exponentiated. The baseline hazard $h_0(t)$ is a function of time only, and is assumed to be the same for all subjects. The name proportional hazard stems from the fact that the hazard of any individual is a fixed proportion of the hazard of any other individual over time. The b_j coefficients of the proportional hazards model can be estimated without having to specify the baseline hazard function $h_0(t)$. Therefore, the proportional hazards model is often called a semi-parametric model. The estimation of the covariate coefficients is done by using the partial likelihood principle.

As we can see from above, while no assumptions are made about the shape of the underlying hazard function, the model equations shown above do imply two assumptions. First, they specify a multiplicative relationship between the underlying hazard function and the log-linear function of the covariates. This assumption is also called the proportionality assumption. In practical terms, it is assumed that, given two observations with different values for the independent variables, the ratio of the hazard functions for those two observations does not depend on time. The second assumption of course, is that there is a log-linear relationship between the independent variables and the underlying hazard function. A different semi-parametric approach is taken by Mani et al. (1999), who build a Neural Network semi-parametric model, where each possible survival time *t* has its own output node (the survival time is discretized to the monthly level). They illustrated that the more elaborate NN model performs better than the PH model on their data.

Critics on Conventional Survival Analysis

Although conventional survival analyses are very useful approaches to deal with censored data, they have their shortages. First, note that this estimator cannot easily estimate the effects of covariates on the hazard and survival functions. Subsets of customers can generate separate Kaplan-Meier estimates, but sample size considerations generally require substantial aggregation in the data, so that many customers are assigned the same hazard and survival functions, regardless of their variation on many potential covariates (Mani et al., 1999). What is more, while popular for some applications, as Mani et al. (1999) mentioned, pure parametric approaches are generally not appropriate for CLV modeling, since the survival function tends to be “spiky” and non-smooth, with spikes at the contract end dates. Last but not least, for the semi-parametric approach, although it appeals the power of covariates, as it assumes a baseline hazard function (proportionality assumption); it still has limitations that other methods like data mining are not restrained from.

2.4.3 Research in Survival Neural Networks

Besides the conventional models we listed above, several researchers have explored the possibility of using Neural Networks (NNs) for survival analysis in the context of medical prognosis, personal loan analysis (Stepanova and Thomas, 2002), etc. By comparing the conventional statistic approaches and Neural Network models, they illustrated that the more elaborate NN model performs better on their data. Moreover, Neal also studied the possibility of the Bayesian Neural Network application in this area (Neal, 2001).

Compared to the conventional survival analysis, Neural Networks may offer an interesting alternative because of their universal approximation property and the fact that presumed survival distribution or baseline hazard assumption is needed. Several Neural Network survival analysis models are discussed and evaluated according to their way of dealing with censored observations, time-varying inputs, the monotonic of the generated survival curves and their scalability. In the following section, we will present a literature overview on the use of Neural Networks and Bayesian Neural Networks for survival analysis.

The simplest method considers survival status for a fixed time scale, thus consequently gives a problem of binary classification. Censored observations are totally excluded from the experimental dataset. The Neural Network then outputs an estimation of the probability that a subject will survive the time period or not. By given a predefined threshold value, the subject is assumed to survive the period. There is no denying that this approach is rather basic and does not allow producing survival curves or hazard functions for individual cases. Furthermore, it complete ignores the problem of censoring and time-varying inputs.

In spite of direct classifying survival time status by using Neural Networks, some researchers built Survival Neural Networks Models to generating individual survival curves or hazard functions for each censored cases.

Ohno-Machado (1996) applied multiple neural networks to solve the survival analysis problem. In this multiple level networks, each Neural Network has a single output predicting survival at a certain time point. Censored cases are not included in the network until the time censoring happens. However, this method makes the number of training instances gradually decreases for the later time intervals which makes the predictions less reliable. The author argues that when using these neural networks isolation, non-monotonic survival curves may result. As a result, the probability of a person surviving two periods could be greater than the probability to survive one period because the interdependencies of the survival probabilities over time are not properly taken into account when isolated neural networks are used.

The author combined the neural networks to decrease the frequency of non-monotonic curves. Survival predictions of one Neural Network are then used as an additional input to another Neural Network as illustrated in Figure 7. However, it is not a simple task to combine the networks. Although not presented in the original paper, the approach allows to easily including time-dependent inputs into the different data subsets. However, since the method requires the combination of a multiple neural networks it results in an important scalability problem which makes the method less suitable for handling large data sets.

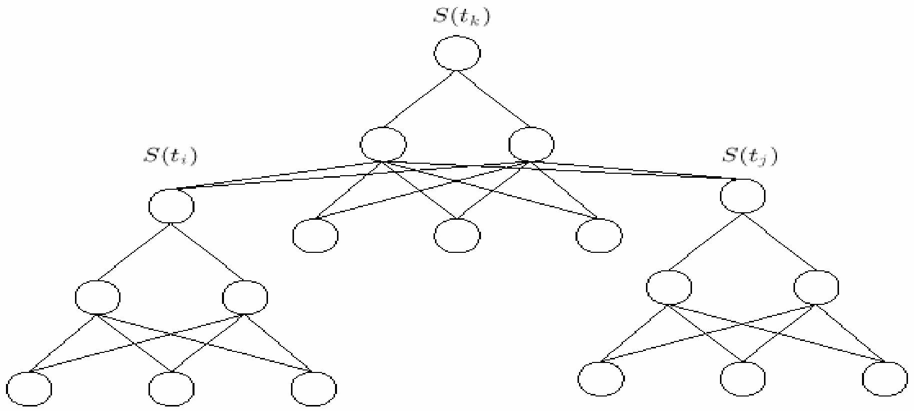


Figure 7 Example of Modular Neural Network for Survival Analysis.

where, the output of the networks predicting S(ti) and S(tj) are used as additional inputs for the network predicting S(tk).

Ravdin and Clark (1992) used a multi-layer feed-forward Neural Network with a single output unit representing the survival status. A time indicator and a survival status indicator are appended to each individual record. The time indicator records the successive time periods [1, T_{max}] for which a prediction is to be made. T_{max} is the maximum time of test follow-up. An uncensored input is then replicated T_{max} times whereas a censored input is replicated t times where t is the time of censoring. The survival status is the target of the network and is set to zero as long as the subject is alive and to 1 otherwise.

Although time dependent inputs were not discussed in the original study, they can be easily included into the corresponding data records. The authors stated that the output of the Neural Network (referred to as the survival probability index) is roughly proportional to the Kaplan-Meier estimation of the survival probability.

However, there is no guarantee that the generated survival curves will be monotonically decreasing. Furthermore, there are two problems brought in because of the replication of records. First, it results in large biases because the number of deaths in the late time intervals will be overrepresented. The authors suggested handling this by selective sampling such that the proportion of deaths matches the Kaplan-Meier estimation. Second, the replication of records results in very large data sets, which will cause severe scalability problems.

A variation on the approach of Ravdin and Clark was proposed by Biganzoli et al. (1998). They also trained a Neural Network with one output and an additional time indicator input. However, unlike Ravdin and Clark, uncensored subjects are only replicated for the time intervals in which they were actually observed. Hence, cases that have died are not included after the time period of death. Again, since each subject has multiple input vectors which may change across the intervals of observation there is possible to include time dependent variables. The Neural Network predicts discrete hazard rates which may be easily converted to monotone survival probabilities. However, like other methods this approach requires enormous data replication thus is not scalable.

Lapuerta et al. (1995) proposed a multi-network strategy to impute the survival times for censored cases. They built a separate neural network for each time interval. These networks are trained with only the observation whose survival status for the corresponding time period is known. Subsequently, they trained networks to predict the outcome for the censored cases. Then the principal Neural Network (referred to as the Predictor network in the original paper) is trained with uncensored and imputed censored observations. The principal Neural Network predicts the probability of survival for each time period considered. The author reported that the proposed method compares favorably to the Cox proportional hazards model. However, there is no guarantee that the derived survival probabilities are monotonically decreasing and time varying inputs are also not allowed. Furthermore, it is clear that this approach is not suitable for large-scale applications since one need to train as many neural networks as there are time periods considered.

Faraggi and Simon (1995) suggested a Neural Network extension of the Cox proportional hazards model by replacing the linear function $\beta^T x_i$ in Cox Regression model by the output $g(x_i, \theta)$ of a Neural Network with a single, logistic hidden layer and a linear output layer

$$h(t, x_i) = h_0(t)e^{g(x_i, \theta)}$$

Analogous to the Cox model, no bias input is considered for the output layer since this is implicitly incorporated into the baseline hazard $h_0(t)$. Then the model uses the partial likelihood principle and Newton-Raphson optimization to estimate the θ parameters. This method preserves the advantages of the classical proportional hazards model. However, the standard approach still assumes that the hazard functions are proportional. However, time-varying covariates might allow for non-proportionality, which may not be the best way to model the baseline variation.

Street (1998) used a multilayer perceptron with T_{max} output neurons to handle the survival analysis problem within the scope of medical study, where T_{max} represents the maximum time horizon of the study. He used a hyperbolic tangent activation function in the output layer such that all output neurons take on values between -1 and +1. The output neuron that predicts the event time is selected as the first output neuron having a value < 0. If all output neurons have values > 1, then the observation is considered to survive the entire time period of the study. The output neurons thus represent the survival probability for the corresponding time period.

For the uncensored cases, the output values are set to +1 as long as the observation is alive and to -1 thereafter. For the censored cases, the output units are also set to +1 until the time censoring happens. After this pre-processing effort, Street used the Kaplan-Meier hazards to compute the survival probability of the censored observations after the censoring occurred. Note that these probabilities are then scaled to the range of the hyperbolic tangent function in the following way: activation = 2 × probability - 1. In summary, the outputs of the training set observations are encoded as follows:

$$S(t) = \begin{cases} 1 & 1 \leq t \leq L \\ -1 & \text{D} = 1 \text{ and } L < t \leq T_{max} \\ S(t - 1) \times (1 - h(t)) & \text{D} = 0 \text{ and } L < t \leq T_{max} \end{cases}$$

where T_{max} represents the maximum number of time periods involved in the study, L the subject lifetime or censoring time, and D indicates if the subject is censored (D = 0) or not (D = 1). The individual survival curve of an observation can then be derived based on the activation values of the output units. Since the Neural Network cannot be forced to generate monotonically decreasing output units, there is still possibility of generating a potential non-monotone survival curve, which complicates its interpretation (Mani et al., 1999).

Furthermore, no extension is provided to deal with time-varying inputs.

A variation on the method of Street was developed by Mani et al. (1999). Again, for every observation in the training set, T_{max} output units are computed. Nevertheless, these output units now represent the hazard rate instead of the survival probabilities that were used in the approach of Street. The outputs are then computed as follows:

$$h(t) = \begin{cases} 0 & 1 \leq t \leq L \\ 1 & \text{D} = 1 \text{ and } L < t \leq T_{max} \\ \dfrac{d_t}{n_t} & \text{D} = 0 \text{ and L} < t \leq T_{max} \end{cases}$$

Again, T_{max} represents the maximum number of periods involved in the study, L the subject life-time or censoring time, and D indicates if the subject is censored (D = 0) or not (D = 1). For uncensored observations, the hazard is set to zero until the time of death and 1 thereafter. For censored observations, the hazard is set to zero until censoring time and to the Kaplan-Meier estimate thereafter. The survival probabilities may then be estimated by using $S(t) = S(t-1) \times (1 - h(t))$. The generated survival curves will thus be monotonically decreasing which simplifies the interpretation and increases robustness. However, the issues of time-varying inputs have not been left unaddressed.

Analogous to Mani et al.'s approach, Brown et al. (1997) proposed a single Neural Network with multiple outputs to predict hazard rates. For the uncensored observations, the network output is set to 0 as long as the subject is alive and to 1 when the subject undergoes the event. For the time intervals following the event, the hazard function is unconstrained. The output values for the censored observations are set to 0 until the time of censoring and are unconstrained for all subsequent time intervals. The authors suggested to train the Neural Network to minimize the sum of squared error criterion and to perform no weight updates when the hazard function is unconstrained by setting the corresponding errors to 0. The approach presented is scalable and results in monotonic survival curves. However, no extension is presented to deal with time-varying inputs.

Analogous to Faraggi and Simon's approach, Neal proposed a Bayesian Neural Network extension of the Cox proportional hazards model by setting the input as time and covariates and targeting the log of the hazard function (Neal, 2001). The approach was applied in a clinical trial with 312 subjects, testing a drug for treating primary biliary cirrhosis on an example from the book of Fleming and Harrington (1991). The author concluded that the more complex models seem to perform better on their test cases, but none of the differences are statistically significant by using a paired t-test. What is more, over-fitting was avoided by using Bayesian method. However, the model training process is computationally demanding. The author claimed that it takes several hours to a day to learn a Bayesian Neural Network.

From the literature review above, it becomes clear that for large scale data sets, the approaches of Faraggi and Simon, Neal, Mani et al., and Brown et al. seem to be the most interesting. They allow to generate monotonically decreasing survival curves and only one neural network needs to be trained. Although the first two approaches allow for time-varying inputs, they are less flexible in modeling the baseline variation. On the other hand, while the later two approaches allow for flexible baseline modeling, they do not solve the problem of time-varying inputs.

CHAPTER 3 RESEARCH FRAMEWORK

The research develops and extends ideas from the marketing CLV problem framework and conventional survival analysis in a way that allows a more accurate prediction for survival time and therefore a better CLV calculation.

The research proceeds in two phases: firstly, method is developed by means of combining conventional statistical survival analysis with data mining techniques; secondly, method is validated with a further analysis process, followed by an evaluation on marketing practice.

3.1 Research Model

3.1.1 Research Model for Method Development

As mentioned in Chapter 1, in this study we investigate the possibility of combining conventional survival analysis and some data mining techniques to provide a better method for survival time prediction and therefore a better Customer Lifetime Value calculation. In the first step, due to the fact of censored data problem, we apply the conventional survival analysis to impute the "partially missing" survival time for each censored cases. In the second step, we seek a better prediction of imputed survival time by appealing different data mining techniques. By conducting the conventional statistical approaches and data mining methods in tandem, we demonstrate how data mining tools can be apt complements of the classical statistical models---resulting in a prediction method that is both accurate and understandable.

In the first step of the analysis, we appeal the power of conventional survival analysis to generate a reasonable imputation of "partially missing" survival time for all the censored cases. We apply a non-parametric survival approach, Kaplan-Meier estimator because of its simplicity and relatively strong estimation power. In detail, the system will first produce a lifetime table and a hazard rate for every time interval. Then we use the equation $S(t) = S(t-1) \times h(t)\ and\ S(0) = 1$ to produce a survival time vector for every time interval. Finally, the median survival time t_median (where $S(t) = 0.5$) are selected as the predicted survival time for every censored cases in each time interval.

In the second step, different statistical and data mining models are developed to further predict the imputed survival time for seeking a better estimation. The guidelines for selecting the data mining methods are effectiveness and efficiency. Moreover, since the purpose of this research is to generate useful and concise knowledge of real world business, the data mining methods selected must be finally evaluated on a real business scenario. However, some other criteria, such as over-fitting and model complexity are also considered. Based on these guidelines, we conduct the experiments on different statistical and data mining methods to find the best candidate for the proposed CLV problem. These methods are Linear Regression, Neural Networks, Regress Tree, and Latent Class Regression models. Through our experiments, we finally create an integrated data mining method which can successfully predict the survival time for a censored customer and generate useful business meaning within the scope of telecommunication industry.

This method is developed by complying with the training and testing schema for conducting the experiments. First, a training set containing both censored and uncensored cases and a testing set, which contains only uncensored cases, are randomly selected from the population. It should be noted that there is no overlapping between the training and testing sets and the testing set contains only uncensored cases whose survival time are known. Next, we apply the statistical and data mining methods on the training set. To ensure a fair comparison, all the statistical and data mining methods are finally evaluated on the testing set only. Through this way, the most suitable method for survival time estimation can be founded. The method selection process is illustrated in Figure 8.

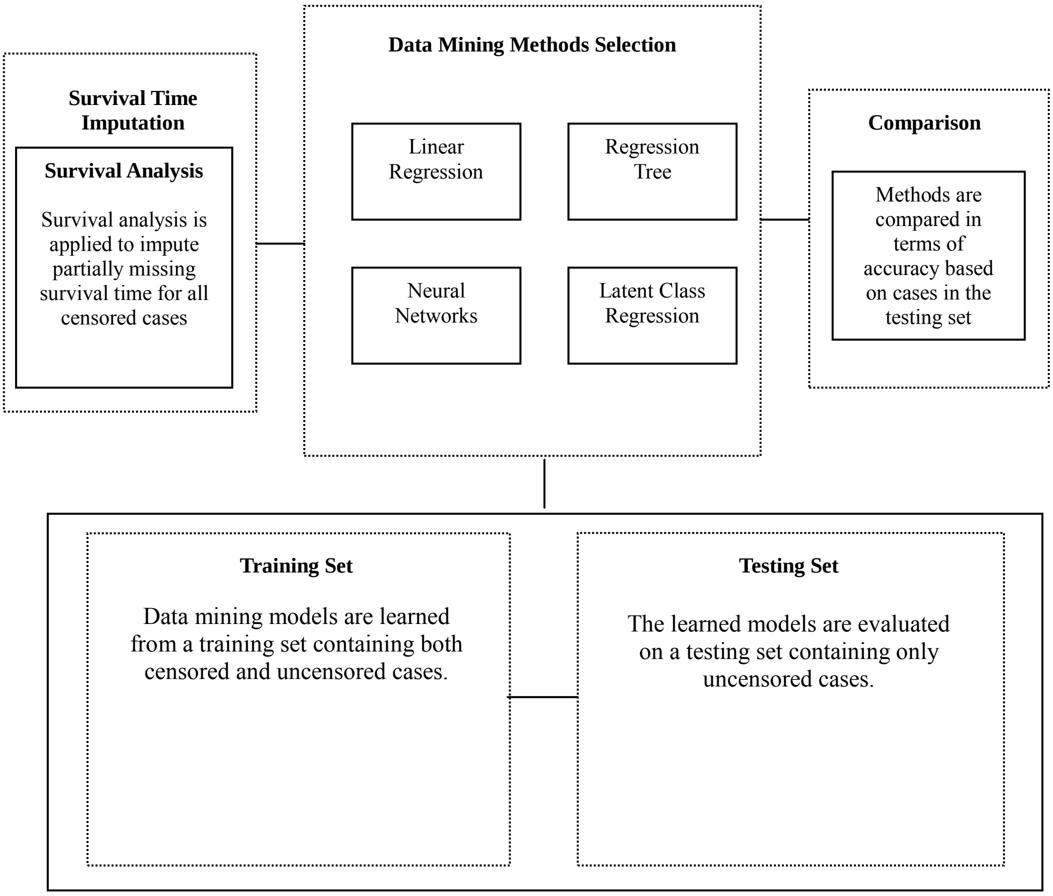

Data Mining Methods Selection
Survival Time Imputation
Survival Analysis
Survival analysis is applied to impute partially missing survival time for all censored cases
Linear Regression
Regression Tree
Neural Networks
Latent Class Regression
Comparison
Methods are compared in terms of accuracy based on cases in the testing set
Training Set
Data mining models are learned from a training set containing both censored and uncensored cases.
Testing Set
The learned models are evaluated on a testing set containing only uncensored cases.

Figure 8 Research Model for Survival Analysis and Data Mining Methods Selection

After selecting the most suitable method to impute survival time for censored cases, we intend to further calculate CLV based on the more accurate survival time. As mentioned in Chapter 2, a CLV model has three basic components: customer's value over time, customer's length of service, and a discounting factor, besides many other specific elements (Rosset et al., 2003). Given these three components, we can write the explicit formula for a customer's CLV as follows:

$$CLV = \int_0^\infty S(t)V(t)D(t)dt$$

While this formula is attractive and straight-forward, the essence of the challenge lies, of course, in estimating the $V(t)$ and $S(t)$ in a reasonable way. In different business settings and marketing objectives, the CLV can be calculated differently. The other factors, such as competitors, referrals, and relationship scenario, can add incremental trouble for this CLV application.

.

As mentioned in Chapter 2, the simplest way to deal with these factors is just considering the current value as the future value, thus excluding the factors completely (Novo, 2001; Mani et al., 1999). What is more, most researchers usually concentrate on modeling LOS alone when they intend to provide a solution for CLV application, while leaving the value issue to the experts (Mani et al., 1999; Rosset et al., 2003; Mohammed and Kotze, 2005). Based on this idea, we make some simple but logical assumptions so that our calculation could be more practicable. The assumptions are given as follows.

- Assumption 1:

  We assume that time interval is not continuous but discrete. This makes the integral function along with the continuous time scale into a summation function based on discrete time intervals.

- Assumption 2:

  We assume that a customer's billing behavior doesn't change a lot from month to month, which makes us able to replace the $V(t)$ with her/his average monthly value.

- Assumption 3:

  Follow the assumption 2, we further assume that customer's value over time is not dependent on time t, i.e., $V(t)$ is a constant function along time for each time interval.

- Assumption 4:

  Instead of calculating survival rate for each time interval, we can simply assume that the churn status of a customer doesn't change from month to month, i.e., instead of summating value term for each time interval individually, we can simply calculate CLV by multiplying average value and predicted survival time together.

- Assumption 5:

  Marketing practitioners normally apply a single constant discounting factor instead of applying a discounting function to calculate it for each time interval. In this research, we assume there is no difference between future value and current value, and thus the discounting factor is 1.0. In other words, we presume that there is not discounting effect.

Based on the above assumptions, the calculation equation can be rewritten as follows,

$$CLV = Mean\_V \times S\_pred$$

where, *Mean_V* indicates the average monthly value and *S_pred* indicates the predicted survival time.

Given the CLV calculation equation, we propose a two level CLV calculation method. In the first level we use the combination of the survival analysis and data mining methods to predict the survival time. In the second level we directly calculate CLV based on the predicted survival time. All the four methods, namely, Linear Regression, Neural Networks, Regression Tree, and Latent Class Regression are applied and compared thereafter. This process is illustrated in Figure 9.

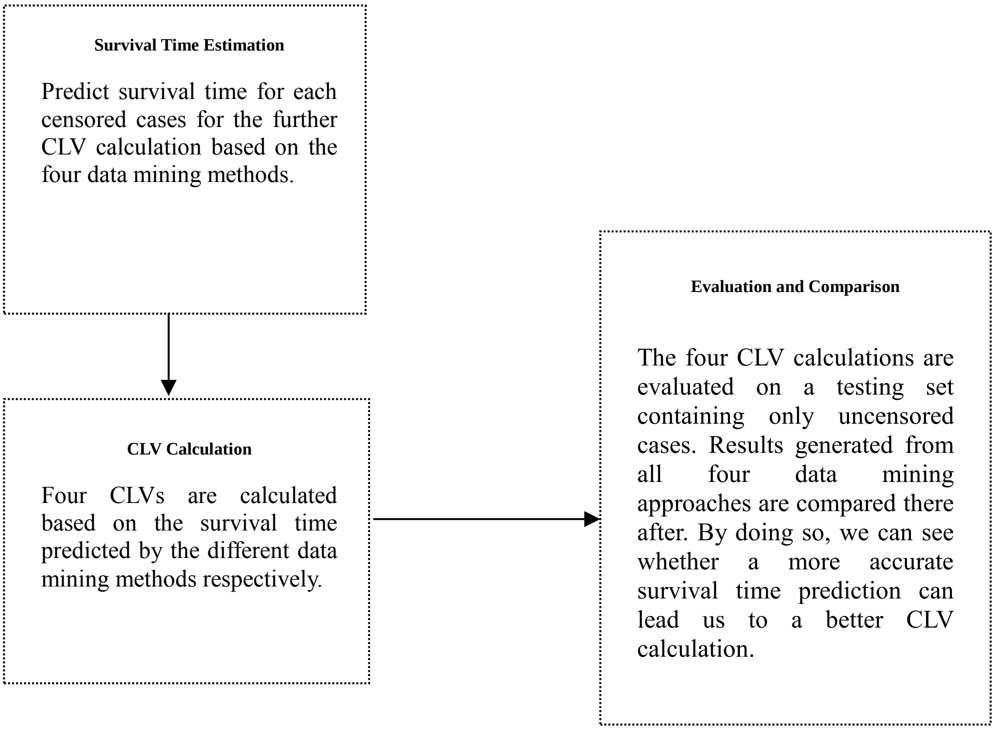

Survival Time Estimation
Predict survival time for each censored cases for the further CLV calculation based on the four data mining methods.
CLV Calculation
Four CLVs are calculated based on the survival time predicted by the different data mining methods respectively.
Evaluation and Comparison
The four CLV calculations are evaluated on a testing set containing only uncensored cases. Results generated from all four data mining approaches are compared there after. By doing so, we can see whether a more accurate survival time prediction can lead us to a better CLV calculation.

Figure 9 Two-Level CLV Calculation Method

3.1.2 Research Model for Business Meaning Generation

Mainly, the proposed data mining method will be applied on a large real-world telecommunication dataset published on the Churn Response Modeling Tournament 2003 for modeling defections. The reasons for choosing this dataset are simple. Firstly, this is a real world telecommunication situation which contains one hundred thousand records and 171 variables. This dataset gives us two challenges, a real scenario and large volume of data. Secondly, this dataset contains time indictor and churn information which is a must for the survival analysis.

Through the application on this dataset, we generate business knowledge from a general industry prospect to detailed individual information. Examples are, what is the number and percentage of customers who will leave the company within one year, two years or even longer; how long a specific group of customers will stay in the company, and when is the right time for the company to spend marketing efforts such as mass promotion on this type of customers. Thus, the real-world business knowledge justifies that this research is practicable in taking marketing actions. Moreover, because we emphasize this study on the telecommunication industry, the results of this study can have a strong proximity and relevance that is meaningful and useful for the telecommunication industry.

Business practice does not stop even though we already obtain the knowledge of customer's survival time or Customer Lifetime Value (CLV). However, a more accurate estimation of survival time and CLV can better the further analysis, thereby indirectly benefits the marketing performance in the real business world. Following the test experiments on data mining method selection, we conduct further analysis to evaluate the proposed data mining method to help companies' marketing practice. To be specific, a calling plan recommendation strategy is made based on the estimated customer lifetime value to evaluate the combined survival and data mining method on a real world business scenario. Moreover, by carrying out this further analysis, we can have a measure on how good is our proposed integrated data mining approach on real CLV applications compared to the conventional survival analysis.

In details, a supervised classification method which requires the construction of two sets of decision trees is applied. Initial classification knowledge is generated based on the facts from the telecommunication dataset. Figure 10 illustrates the research model for classifying customer groups.

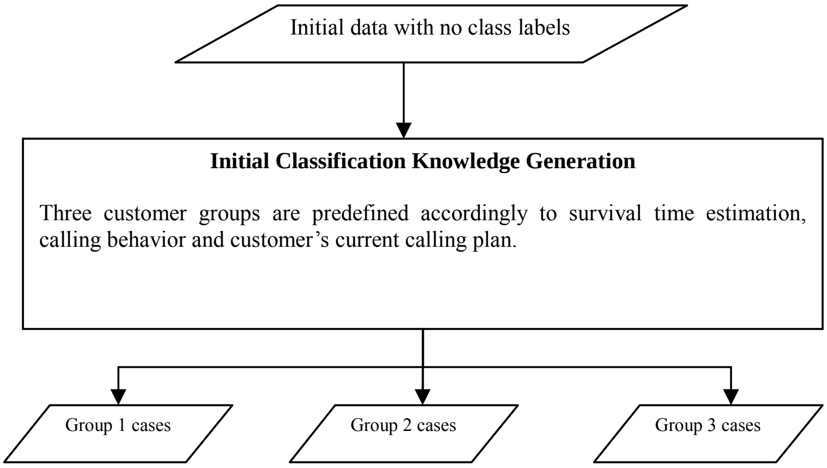
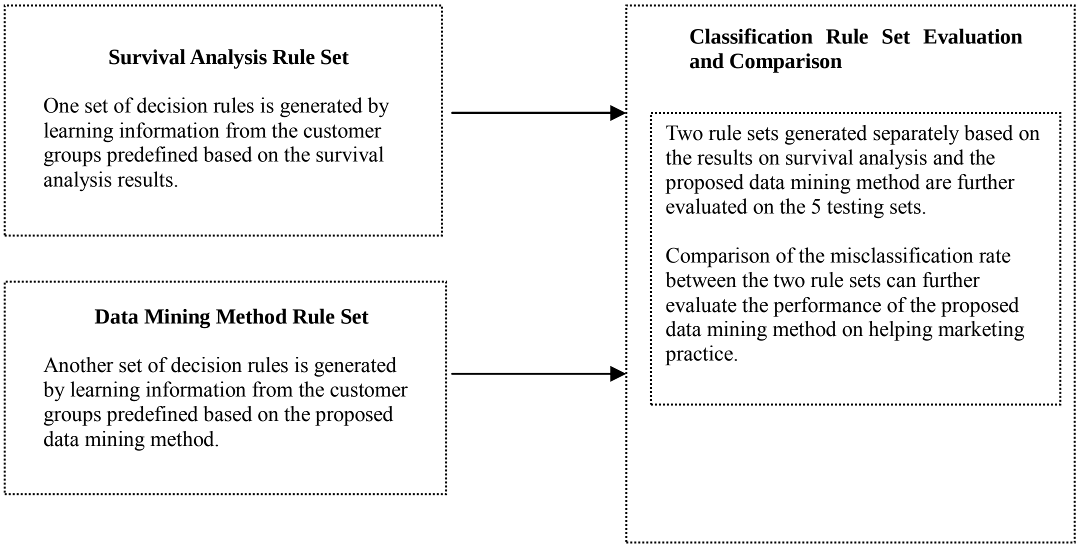

Initial data with no class labels
Initial Classification Knowledge Generation
Three customer groups are predefined accordingly to survival time estimation, calling behavior and customer's current calling plan.
Group 1 cases
Group 2 cases
Group 3 cases

Figure 10 Research Model for Predefining Classification Groups

In order to evaluate the rule sets on all the uncensored cases. Five training sets and five testing sets are generated accordingly. The rule sets are generated from the information in training sets and then evaluated based on the corresponding testing set. We present the averaged results derived from those 5 testing sets. Figure 11 elaborates the research model for the further analysis.



Survival Analysis Rule Set

One set of decision rules is generated by learning information from the customer groups predefined based on the survival analysis results.

Data Mining Method Rule Set

Another set of decision rules is generated by learning information from the customer groups predefined based on the proposed data mining method.

Classification Rule Set Evaluation and Comparison

Two rule sets generated separately based on the results on survival analysis and the proposed data mining method are further evaluated on the 5 testing sets.

Comparison of the misclassification rate between the two rule sets can further evaluate the performance of the proposed data mining method on helping marketing practice.

Figure 11 Research Model for Further Analysis

CHAPTER 4 SURVIVAL ANALYSIS AND DATA MINING APPROACHES

In this chapter, we compare different statistical and data mining methods that are known for their ability of solving prediction problems. They include Linear Regression, Neural Networks, Regression Tree of CART (Breiman et al., 1984), and Latent Class Regression (Vermunt and Dijk, 2001). Recently, latent class models, which control for unobserved heterogeneity among subjects using latent or unknown groups, have become increasingly popular in marketing research. Thus, it serves a method for comparison. We examine the performance of these prediction methods by testing the predictive ability of the models learned from the training data on unseen cases in the testing data.

4.1 Data Description

We perform experiments on a large real world telecommunication industry dataset. The data, which was provided by a major wireless carrier, is released by Teradata Center in the year 2003. Account summary data was provided for 100,000 customers who had been the subscribers of the company for at least 6 months. This dataset is used in the statistical and data mining approaches, and facilitates comparison of the various techniques.

To assist in the modeling process the churners are oversampled in the way that one half of the sample consists of uncensored customers (churners, those who left the company by the end of the test) and the other half are censored customers who did not left the company (the exact number is 49,562 churners and 50,438 censors). A broad range of 171 potential predictors are made available, spanning all the types of data a typical service provider would routinely have available. Specifically, Table 2 presents the four types of variables:

Types of Variables	**Notes**
Demographics	Age, Location, Number, and ages of children, etc.
Financial	Credit score, Credit card ownership.
Product details	Handset price, Handset capabilities, etc.
Phone usage	Number and duration of various categories of calls, etc.

Table 2 Variable Type Description

Moreover, the nature of the call behavior data includes summary statistics describing number, duration, etc., which is listed in Table 3.

Call Behavior Data
Completed calls
Failed calls
Voice calls
Data calls
Call forwarding
Customer care calls

Table 3 Call Behavior Data Summary

Furthermore, these historical data include the statistics of mean and range for the prior 3 and 6 months of the relationship and the lifetime (at least 6 months, as much as 5 years) information.

Data Preprocessing:

Customer data for CLV modeling should have, in addition to a variety of independent attributes, two important attributes: tenure (time with the company), and a censoring flag. In our dataset, the attribute “MONTH” describes the customer tenure in months, and the attribute “CHURN” indicates if the customer is still active or has cancelled her/his service. If

CHURN=0, the customer is still active and MONTH indicates the number of months the customer has had service; if CHURN=1, the customer has cancelled the service and MONTH is her/his age in months at the time of cancellation. Intuitively, we set the survival time of the customer to MONTH when a customer is churned (uncensored) and to the Kaplan-Meier estimator if the data is censored.

Usually, it is necessary to reduce the dimension of the data set by selecting the attributes that are relevant and necessary. Towards this feature selection process, there are many possible options. In this study, we intend to apply a Neural Network method to handle this data reduction problem because of its strong prediction power. However, due to the extremely large variable dimension and data volume, the Neural Networks method can not be directly applied for this case. Thus we first use the forward selection criteria of linear regression to reduce the variables dimension into 33 numerical variables. After that, the Neural Networks method will be applied to further select the most important variables correlated to the dependent variable, i.e., survival time.

Once a Neural Network has been trained, examination of the variable's network connection weights can be used to identify prospective variables for elimination. Variables with small connection weights (i.e., less prediction power towards survival time) are good candidates of elimination. Moreover, the Neural Network method can order variables with respect to the information they provide on the output variable, i.e., survival time. Neural Networks can then be built by adding one variable once at a time and examining if there is improved performance. We use the software provided by Tiberius Neural Networks to handle this problem automatically. Instead of directly selecting the important variables, the software removes the least important variables from the independent variables pool generated by the linear regression method. We apply the Automatic Pruning function built within the Neural Network software to handle this problem automatically. As mentioned in Chapter 2, Neural Networks assigns "weight" to each input neuron (in this case, the independent variables) according to its correlated importance with the output neuron (in this case, the dependent variable, survival time). Automatic Pruning is a process whereby the variables with the least significant weights are automatically removed from the network. Here, we set to prune one weighted link every 50 epochs (an epoch is one sweep through all the records in the training set). After 800 epochs the system automatically prunes 16 variables off during the network training section, which leaves 17 numerical variables. The reduced attribute set includes, the three factors identified by the Regency-Frequency-Monetary model (Petrison et al., 1997), personal demographic variables, phone usage variables, and some other contract related variables. Specifically, we display the description notes and relative importance (i.e., the ability of predicting survival time by using that certain variable) of the variables selected by the method in Table 4.

Rank	Variable	Relative importance	Variable Description	Notes
1	MODELS	1.00	Model issued	
2	EQPDAYS	0.649	Number of days of current equipment	
3	AVGMOU	0.197	Average monthly minutes of use over the life of the customer	
4	MOU_MEAN	0.087	Mean number of monthly minutes of use	
5	PEAK_VCE	0.038	Mean number of inbound and outbound peak voice calls	
6	UNIQSUBS	0.033	Number of Unique Subscribers	Number of individuals listed with the account.
7	VCEOVR_R	0.023	Range of revenue of voice overage	Overage represents calls or minutes of use over the number of minutes allowed by that customer's calling plan.
8	AVGREV	0.015	Average monthly revenue over the lifetime of the customer	
9	ACTVSUBS	0.012	Number of Active Subscribers	Number of individuals listed with the account who actively use the service.
10	DROP_VCE	0.005	Mean number of dropped voice calls	
11	V82	0.004	Well care medical expenditures	Data derived from annual population survey.
12	BLCK_VCE	0.003	Mean number of blocked voice calls	
13	CHANGE_M	0.002	Percentage change in monthly minutes of use vs. previous three month average	
14	ROAM_MEA	0.002	Mean number of roaming calls	
15	V70	0.002	Well care medical expenditures	Data derived from annual population survey.
16	TOTMRC_R	0.001	Range of total monthly recurring charge	
17	RETDAYS	0.001	Number of days since last retention call	The retention calls are those calls from customers considering whether to renew, reporting competitive offers, etc.

Table 4 Variable Description and Relative Variable Importance

4.2 Statistical Survival Analysis

4.2.1 Survival Analysis Assumptions

Before starting the Survival Analysis and data mining approaches, commonly, there are a few general assumptions to be followed.

1. Probabilities for the event of interest should depend only on time after the initial event---they are assumed stable with respect to absolute time. That is, cases that enter the study at different times (for example, patients who begin treatment at different times; customers who enter the company at different times) should behave similarly, known as the proportional hazards assumption.

2. There should also be no systematic differences between censored and uncensored cases. If, for example, many of the censored cases are patients with more serious conditions, the results may be biased.

3. There is an important assumption in Survival Analysis that at a specific time point, individuals who are censored are at the same risk of subsequent failure as those who are still alive. The risk set at any time point (the individuals still alive and uncensored) should be representative of the entire population alive at the same time. Statistically, this assumption is equivalent to the one that the censoring process is independent of the survival time.

4.2.2 Facts of Censored Data

In this CLV task, emphases are placed on providing a precise prediction of LOS of the customers in concentration of a real world business, which makes the censoring problem as our first priority.

As mentioned in section 4.1, since there are 50.4% censored cases in the dataset, firstly we intend to appeal conventional survival analysis to impute the missing survival month for those cases before the next process of our prediction method. To assess the effect of censoring, we first analyze the distribution of survival time for both the censored and the uncensored cases in Table 5 to Table 7 and Figure 12 to Figure 14.

- For all cases (censored and uncensored):

Number	Mean	Median	Minimum	Maximum
100000	18.83	16	6	61

Table 5 Month Statistics for All Cases before Survival Analysis

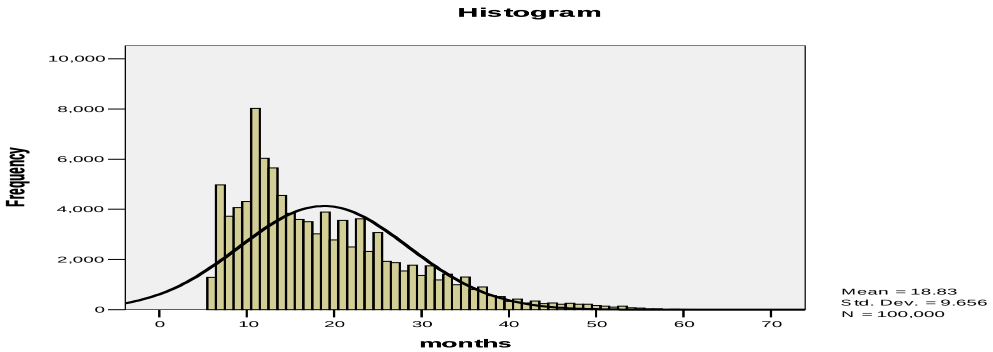


Figure 12 Distribution of Month for All Cases

- For censored cases:

Number	Mean	Median	Minimum	Maximum
50438	18.63	16	6	61

Table 6 Month Statistics for Censored Cases before Survival Analysis

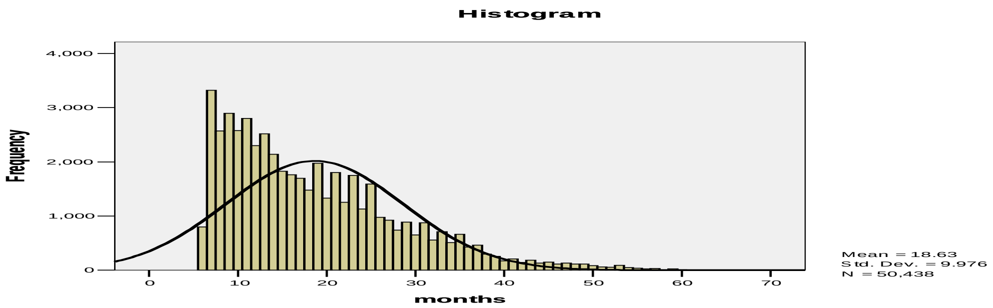


Figure 13 Distribution of Month for Censored Cases

- For uncensored cases:

Number	Mean	Median	Minimum	Maximum
49562	19.04	17	6	61

Table 7 Month Statistics for Uncensored Cases

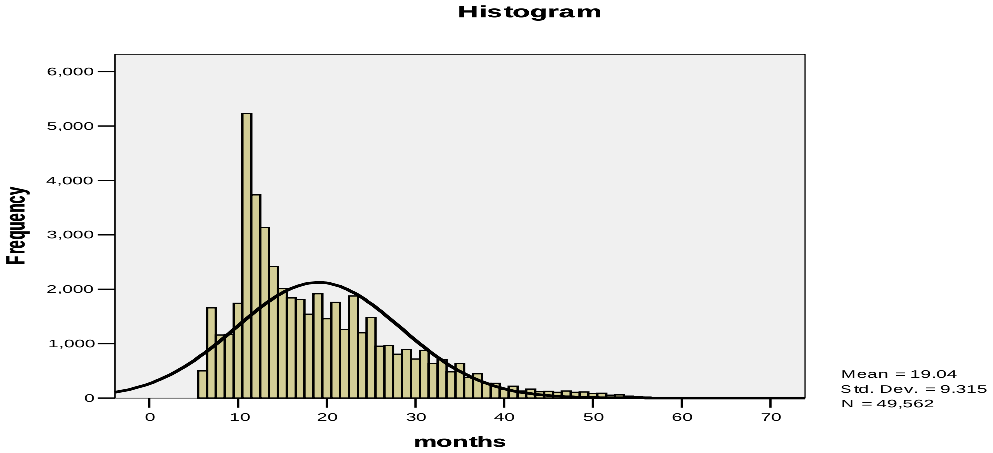


Figure 14 Distribution of Month for Uncensored Cases

4.2.3 Survival Analysis Application

Moreover, as noted in Chapter 2, generally speaking, there are four types of classic statistical survival approaches for the analysis of censored data, largely distinguished by the assumptions they make about the parameters of the distribution(s) generating the observed survival times. Based on the discussion in Chapter 2, we choose a non-parametric method, Kaplan-Meier estimator for its simplicity and relatively strong estimation power.

Following the work of Mani et al. (1999), we apply the “median survival time” as the predicted survival time for all the censored cases in a certain time interval. However, standard package like SAS (Monik, 2004) will not directly produce the “median survival time” that we need for each individual case. Even though, there are still useful measurements can be used to generate median survival time indirectly. Examples are the hazard rate for each time interval. Further, we apply the function $S(t) = S(t-1) \times h(t)$ *where* $S(0) = 1$ to produce a survival time vector for each time interval. Finally, the median survival time (the median survival time is the time *t* in which survival rate of the case $S(t) = 0.5$) can be computed after searching these vectors.

Survival Month Summarization for All Cases:

After applying the Kaplan-Meier estimator, the results show that the survival time for all cases ranges from 24 to 61 month, with a mean of 31.95 months. In the 25th month, we have the largest number of customers leaving the company, with a number of 20136, 20.14% out of all cases. Moreover, the survival analysis estimates that only 12% of all the customers will still be with the company after the 40th month. Figure 15 displays the distribution of survival time for all cases.

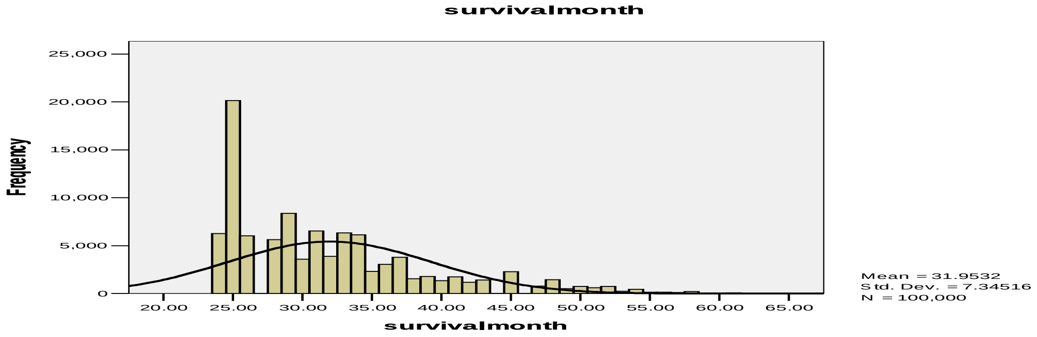


Figure 15 Distribution of Survival Month for All Cases

What is more, to observe the effect of survival analysis, we further create two variables to compare the customer survival time before and after survival analysis. Details can be seen in Table 8. In this table, "survivalmonth" stands for the survival month estimated by the Kaplan-Meier method and "survivalandmonth" stands for the replaced month value which consists of Kaplan-Meier estimation for censored cases and the exact original month value in the dataset for those uncensored cases.

Time Variable	survivalmonth	survivalandmonth
Number	100000	100000
Mean	31.95	25.52
Median	30.00	25.00
Minimum	24.00	6.00
Maximum	61.00	61.00

Table 8 Survival Month Statistics for All cases after Kaplan-Meier Analysis

After we replace the survival time with their exact date of death for those censored cases, we have a different picture as depicted in Figure 16. As we can see, their survival month ranges from 6 to 61 and with a smaller mean of 25.52 months. However, the month when the most of the customers leaving the company is also the 25th month, with a number of 12326 customers, 12.3% out of all the cases (the runner-up the 24th month is little less than a half of it with a number of 5320). Besides that, 41.5% of the customers will leave the company before a two-year relationship.

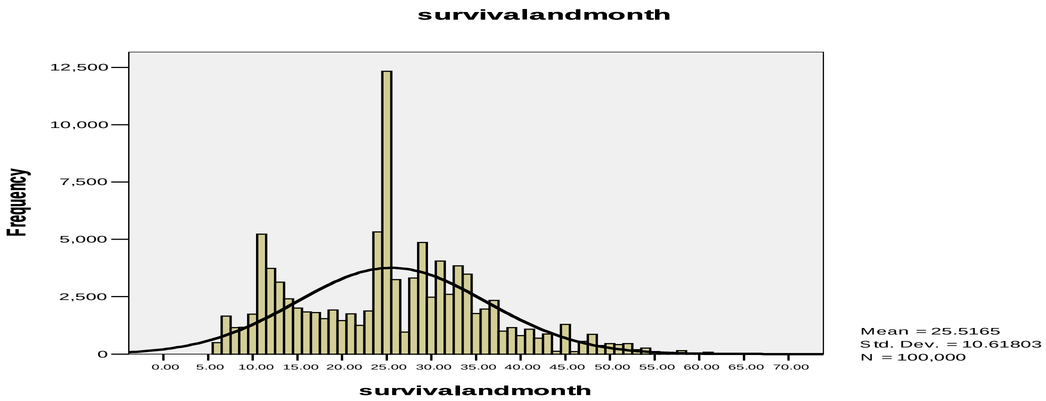


Figure 16 Distribution of Survival Month for All Cases

Survival Month Summarization for Censored Cases:

For all the censored cases, their survival time ranges from 24 to 61 month, with an average of 31.88 months. In the 25th month, we have the largest number of customers leaving the company, with a number of 10845, 21.5% out of all the censored cases (the runner-up the 24th month contains 4121 customers), right after a two years contract period. What is more, for those censored cases, the survival analysis estimates that only 10% of them will still be with the company after the 40th month. Detailed statistics are presented in Table 9 and Figure 17.

Number	Mean	Median	Minimum	Maximum
50438.00	31.88	30.00	24.00	61.00

Table 9 Survival Month Statistics for Censored Cases after Kaplan-Meier Analysis

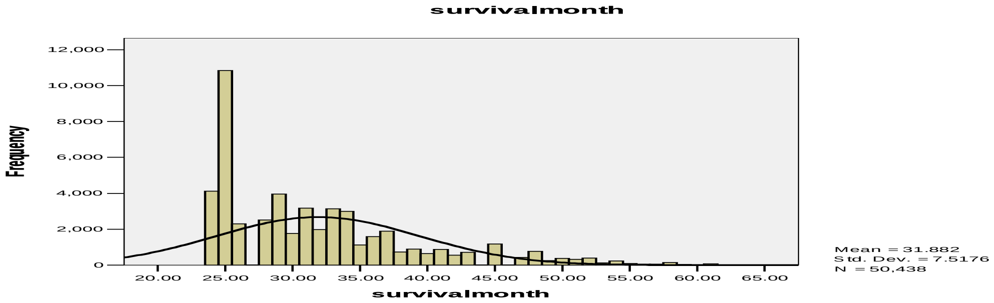


Figure 17 Distribution of Survival Month for Censored Cases

Survival Month Summarization for uncensored cases:

It ranges from 24 to 61 month scale, with a mean of 19.04 months. In the 25th month, we have the largest number of customers leaving the company, with a number of 9291, 19.74% out of all the uncensored cases (the runner-up the 29th month is little less half of it with a number of 4422), and right after a two-year contract period. Detailed statistics are presented in Table 10 and Figure 18.

Number	Mean	Median	Minimum	Maximum
49562	32.03	31	24	61

Table 10 Survival Month Statistics for Uncensored Cases

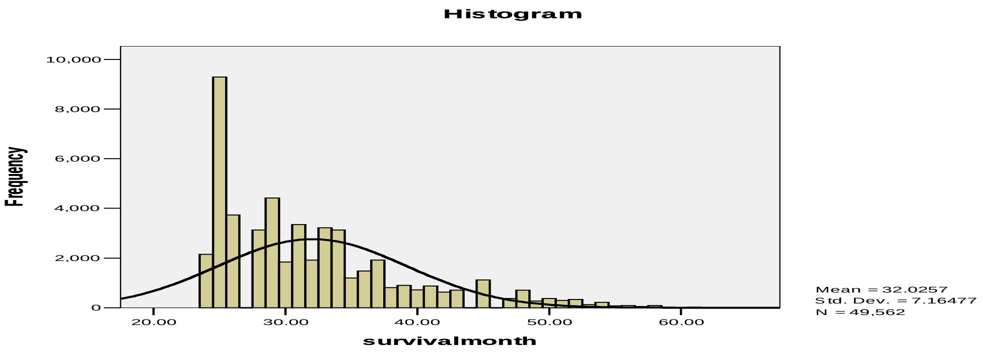


Figure 18 Distribution of Survival Month for Uncensored Cases

To sum up, we feel it is worthwhile to compare the number of customers before and after the 24th month for censored cases; the 2nd year and the 3rd year for uncensored cases. Moreover, practically speaking, the company can offer more attractive contracts to those who will churn after a one-year or two-year relationship.

4.3 Survival Data Mining Approaches

The data described in Section 4.1 is used in our experiments with the statistical and data mining techniques. For the Linear Regression, Neural Networks, Regression Tree, and Latent Class Regression, about 60% of the cases in the entire dataset (60083 cases) are used for training, and 19695 uncensored cases which contain about 20% of the cases in the entire dataset are used for testing the methods' performance. The mean value of survival month in the testing set is 18.96. During the dataset re-sampling process, we use stratified random sampling method to partition the dataset which keeps the churn rate at a 50-50 level. It must be noted that for the experiments, testing set is composed with only uncensored cases, whose exact survival time are known to us.

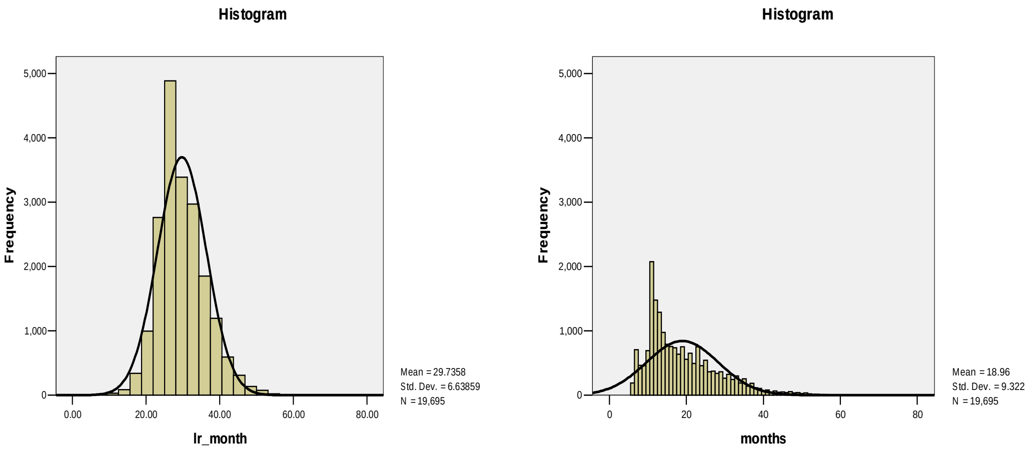
4.3.1 Linear Regression

The Linear Regression model is fit to the training data. The 17 independent variables selected by the Neural Networks method were entered as independent variables in linear regression model. Table 11 lists the results of linear regression analysis between the 17 independent variables and the dependent variable, survival time. The R-square is the coefficient of determination which is a measure of the proportion of the variance of the dependent variable about its mean that is explained by the independent, or, predictor variables. The adjusted R-square is 0.36, which means the full multiple regression equation with all 17 independent variables explains 36% of the variation in the predicted survival time. The beta coefficient is the standardized regression coefficient which allows for a direct comparison among coefficients as to their relative explanatory power of the dependent variable. The test statistic t is of importance for hypothesis testing: if the calculated t exceeds the critical value, then the null hypothesis is rejected. The significance of the calculated t exceeding the critical value is measured by the P-value. A P-value of less than 0.05 is considered significant.

	Full regression	Full regression	
Independent Variables	Beta	t-test	Significance
(Constant)	9.130	35.705	0.000
MOU_MEAN	0.004	13.697	0.000
ROAM_MEA	-0.046	-4.733	0.000
TOTMRC_R	0.004	1.512	0.131
VCEOVR_R	-0.014	-10.898	0.000
CHANGE_M	0.000	0.738	0.461
DROP_VCE	-0.033	-3.830	0.000
BLCK_VCE	-0.028	-4.526	0.000
PEAK_VCE	0.006	5.793	0.000
V70	-0.017	-1.409	0.159
V82	-0.002	-1.784	0.074
UNIQSUBS	-1.060	-8.876	0.000
ACTVSUBS	0.647	3.973	0.000
AVGREV	0.026	8.797	0.000
AVGMOU	-0.007	-20.236	0.000
MODELS	6.083	79.303	0.000
RETDAYS	0.002	2.131	0.033
EQPDAYS	0.020	75.517	0.000

Table 11 Full Regression Result

We compute the predicted survival month with the linear equation for the testing set. Following the experiment of survival analysis, first we generally analyze the predicted survival month. Here, the value of survival month ranges from -51.75 to 193.50. This result suggests that the survival time is overestimated for both the short-lived and long-lived customers. The mean value of the predicted survival month is 29.78, which is about 11 months larger than the original mean value derived from the survival analysis. What is more, according to this result, half of the customers will leave the company before the 33rd month. Figure 19 gives the predicted survival time distribution and the actual survival time distribution on the testing data.



(a) *(b)*

Figure 19 Distribution of Survival Time for Uncensored Cases of Linear Regression Model VS. Actual Survival Time Distribution. (a) Linear Regression, (b) Actual Survival Time.

4.3.2 Regression Tree

We apply CART 5.0 (Classification and Regression Tree version 5.0) to construct our Regression Tree models. As discussed in Chapter 2, Regression Tree is a recursive partitioning method that generates a tree model by “splitting the tree” at each node. It uses Least squares error to determine how well the splitting rule separates the cases contained in the node. Once the best split is found, CART repeats the search process for another node, and continues recursively until further splitting is impossible. Instead of deciding whether a given node is terminal, CART grows the tree to the maximum size and then starts “pruning” it to examine smaller trees. Finally, CART selects the best tree by testing its error rate. In this test, since we are dealing with a relatively large dataset, we decide to set a minimal terminal node size as 50 cases in order to make the tree structure simple and readable. Moreover, we also stop the tree from growing when it reaches the depth of 20. The resulting “optimal” tree has 78 nodes (splits) with a depth of 16. In fact, by giving a set of weight scores to each nodes of the tree, CART ranks the variables in terms of their overall explanatory power (weights) in classifying the cases. Table 12 displays the relative importance of splitters selected by the regression tree model.

Variable	Importance
EQPDAYS	100
MODELS	74.49
AVGREV	8.26
AVGMOU	6.85
PEAK_VCE	5.85
MOU_MEAN	3.85
V82	2.11
DROP_VCE	1.51
V70	1.22
UNIQSUBS	0.85
CHANGE_M	0.77
TOTMRC_R	0.68
RETDAYS	0.67
BLCK_VCE	0.48
VCEOVR_R	0.45
ACTVSUBS	0.36
ROAM_MEA	0.05

Table 12 Variable Importance of Regression Tree Modeling

We analyze the predicted survival month of the testing set. The predicted survival month ranges from 12.9 to 55.23, which indicates that the extreme survival time are marginally underestimated for both the short-lived and long-lived customers. The mean value of the predicted month is 25.46, which are about 6 months larger than the mean from the original testing set. According to the analysis result, half of the customers will leave the company before a two-year contract period. Figure 20 gives the predicted survival time distribution and the actual survival time distribution of the testing data.

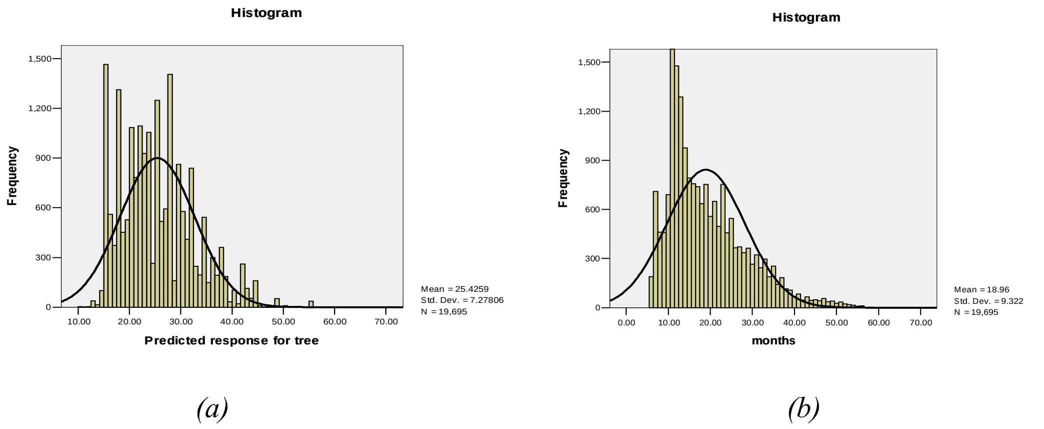
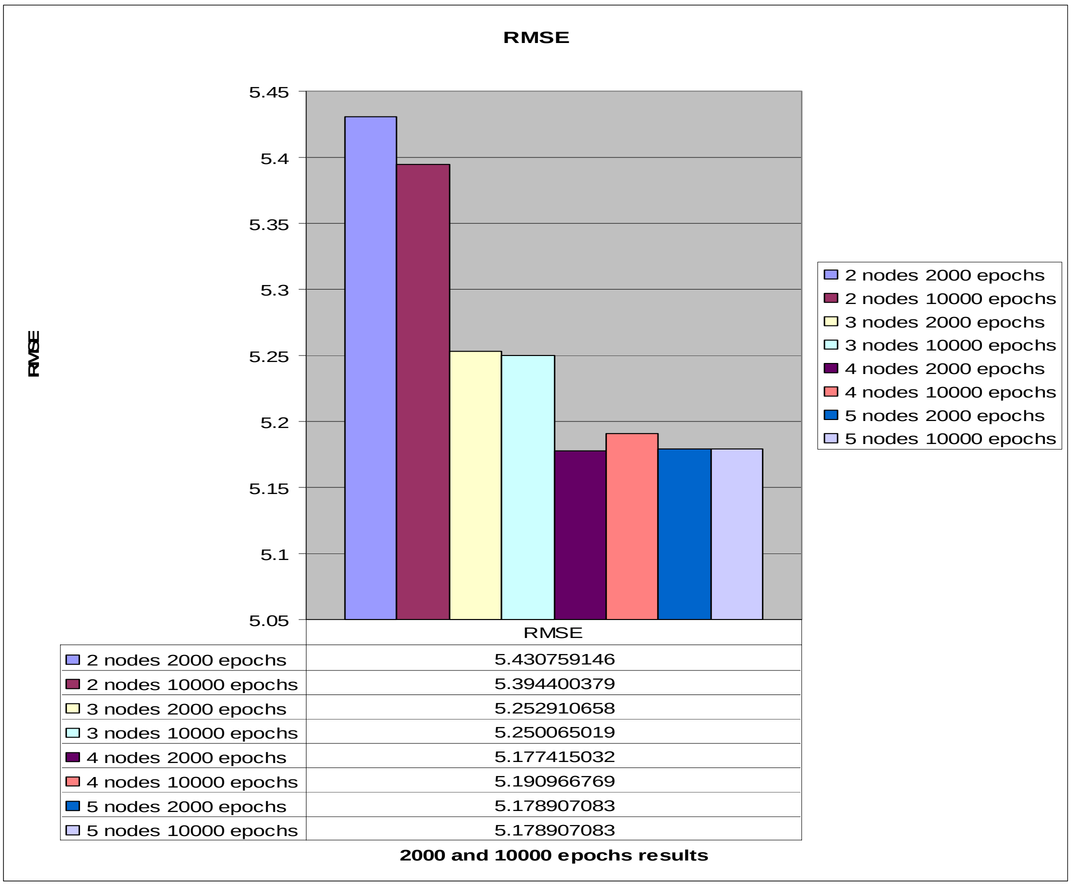


Figure 20 Distribution of Survival Time for Uncensored Cases of Regression Tree Model VS. Actual Survival Time Distribution. (a) Regression Tree, (b) Actual Survival Time.

4.3.3 Neural Networks

In this experiment, we learned a standard single hidden layer feed-forward Neural Network with a single output neuron to predict survival time for censored cases. The number of input units is dictated by the number of independent attributes. The Neural Network model depicts the uni-directional effects from the input variables to the hidden nodes and survival time. By adjusting the number of hidden neurons (2 neurons to 5 neurons), we can obtain the best network model in terms of accuracy. After repeated trials, an NN model with one hidden layer and four hidden nodes (neurons) returns the best result with the lowest error rate. Although NNs achieve a reasonable level of predictive accuracy, its empirical results do not formalize the relationships among the variables in a user-friendly and comprehensible way, nor does it provide the opportunity to gain fresh insight into the problem. Even for competent users of NNs, its lack of transparency and explanatory capability has been a major drawback (West et al., 1997).

In the NN experiment, we use the same 60-20 training and testing sets as described above. In order to gauge the fitness of each potential NNs some quantitative measures of the prediction error must be made. The most common procedure is to search for the weight set that gives the minimum mean absolute error (MAE), the minimum mean squared errors (MAE), or the minimum mean root of mean squared error (RMSE), over all the training cases, where the error is the difference between the required output and the network output. In this test, we choose RMSE. We test the performance of alternative NNs with 2 to 5 hidden neurons. Moreover, we train the NN model with two time schemas, 2000 epochs and 10000 epochs. We present the results of all the alternative Neural Network Models in Figure 21.

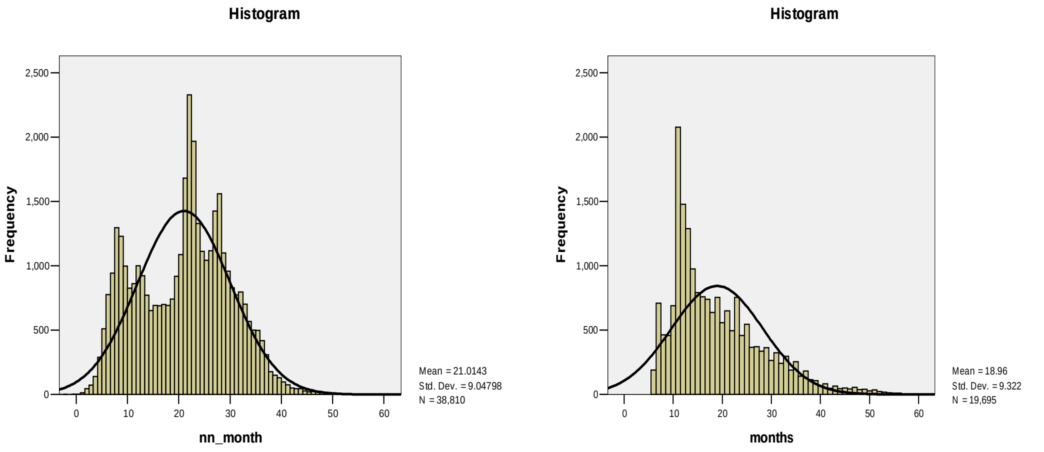


	RMSE
2 nodes 2000 epochs	5.430759146
2 nodes 10000 epochs	5.394400379
3 nodes 2000 epochs	5.252910658
3 nodes 10000 epochs	5.250065019
4 nodes 2000 epochs	5.177415032
4 nodes 10000 epochs	5.190966769
5 nodes 2000 epochs	5.178907083
5 nodes 10000 epochs	5.178907083

Figure 21 RMSE Comparison for Multi-Nodes Neural Networks

From the facts presented above, for all the NN models, their results do not differ too much between the models derived from two training schemas (2000 epochs 10000 epochs). Take the 4-nodes model as an example, its RMSE increases from 5.177415032 to 5.190967, which is a sign of potential over-fitting. Moreover, it is obviously that the 4-nodes, 2000-epochs model gives the best estimation compared to other models.

We generally analyze the predicted survival month of the testing set. The survival month ranges from 1.16 to 61.45, resulting in a marginal overestimation for the customers with short survival time. The mean value of the predicted month is 21.01, which is about 3 months larger than the mean value in the original testing set. What is more, according to the analysis result, half of the customer will leave the company before the 26th month. Figure 22 gives the predicted survival time distribution and the actual survival time distribution of the testing data.



(a) *(b)*

Figure 22 Distribution of Survival Time for Uncensored Cases of Neural Networks Model VS. Actual Survival Time Distribution. (a) Neural Networks Model, (b) Actual Survival Time.

4.3.4 Latent Class Regression

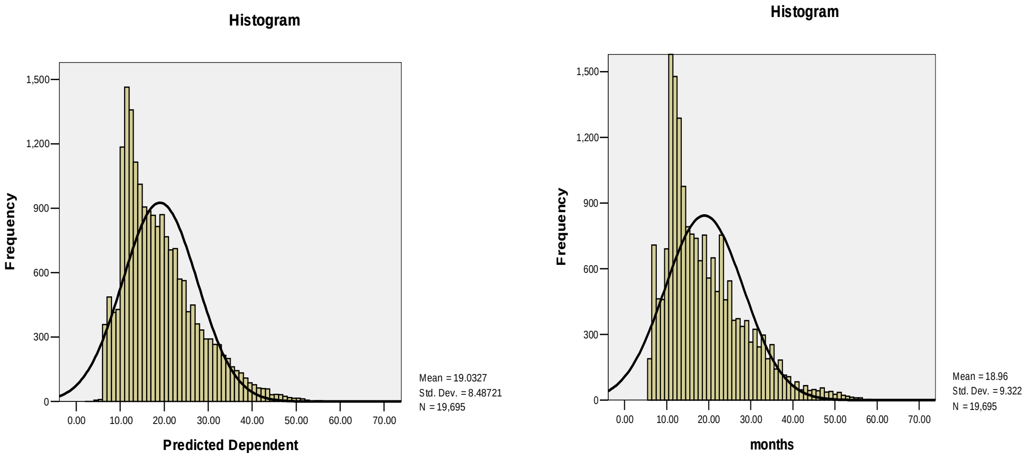
Latent Class Regression estimates a logit model based on the assumption that the coefficients of the predictors differ across unobserved latent segments, and executes separate regressions for each of the latent classes. In the experiment, the number of latent classes that we tested ranges from 1 to 5 and the same 17 independent variables selected by the Neural Networks model are used here. It appears that a 4-class model achieves the best model fit based on the log-likelihood (LL) value and the BIC (Bayesian Information Criterion) value. Detailed statistics of regression models generation are listed in Table 13.

	LL	BIC	Number of parameters	R²
1-Class Regression	-90359.68	180801.84	8	0.78
2-Class Regression	-76775.63	153726.51	17	0.88
3-Class Regression	-68218.04	136704.12	26	0.83
4-Class Regression	-66623.30	130607.42	35	0.86
5-Class Regression	-65354.34	131162.28	44	0.91

Table 13 Detailed Statistics of Regression Models Generation

This report of statistics in Table 13 assists in determining the correct number of latent classes. We report the log-likelihood (LL) value, the BIC value, the number of parameters in the estimated models, and the R-square value. The log-likelihood (LL) is the log-likelihood ratio goodness-of-fit value of the current model. In addition to model fit, the BIC statistic takes into account the parsimony (number of parameters) of the model. When comparing models, the lower the BIC value the better the model is. It is important to determine the right number of classes because specifying too few ignores class differences, while specifying too many may cause the model to be unstable. While the log-likelihood increases each time the number of classes is increased, the minimum BIC value occurs in the 4-Class Regression model, suggesting that the 4-class model is the best one among the five models. The R-square value (R^2) increases from .78 for the 1-class model to .86 for the 4-class model.

In the testing set, the survival month predicted by Latent Class Regression ranges from 2.49 to 65.63, which indicates that the method gives a better prediction for the customers with short or long survival time. Moreover, according to the analysis result, half of the customers will leave the company before the 18th month. Figure 23 gives the predicted survival time distribution and the actual survival time distribution of the testing data.



(a) *(b)*

Figure 23 Distribution of Survival Time for Uncensored Cases of Latent Class Regression VS. Actual Survival Time Distribution. (a) Latent Class Regression, (b) Actual Survival Time.

In our survival dataset, the Latent Class Regression method produces a relatively better result with the lowest RMES 1.299 on the testing dataset. What is more, the heterogeneity across these four clusters may improve predictive accuracy, and such unobserved clusters and their effect on the customer survival time might be meaningful and useful for the marketing decisions. Here, we display the customer profile among the four latent classes in Table 14.

	Class 1	Class 2	Class 3	Class 4
Class Size	36.22%	33%	19.12%	11.66%
Survival Time	19.62	30.92	35.75	20.82
Churn = 0	0.0%	99.39%	87.6%	0.08%
Churn = 1	100%	0.61%	1.24%	99.92%

Table 14 Customer Profile among Four Latent Classes

We can see that the first two classes contain more customers than the last two classes (36% and 33% vs. 19% and 11.6%). We can also see that class 1 and class 4 are composed mainly of the uncensored cases and class 2 and class 3 mainly contain censored cases. What is more, by looking at the predicted month distributions, we can find that uncensored cases are further divided by class 1 and class 4 according to that the customers in class 4 stayed with the company relatively longer than the customers in class 1. The censored cases are further stretch out in distributions, which indicates that the power of conventional survival analysis is consolidated by the latent class regression. Table 15 illustrates the details.

Month	6…12	12…24	24…36	36...60
Class1	33.10%	46.50%	20.10%	0.30%
Class2	0%	9.60%	76.60%	13.80%
Class3	0.10%	0.90%	45.80%	53.20%
Class4	20%	48.90%	29.00%	2.10%
All classes	19.70%	33.70%	35.10%	11.50%

Table 15 Month Summary among four latent classes

To further investigate the power of latent class profiling, we compare the monthly phone usage, service payment and current subscriber contract of customers among different latent classes in Table 16.

	REPLACED_MONTH	AVGREV	AVGMOU	TOTMRC_M	CONTRACT_DIFFER
Class1	17.99322204	45.59282	300.4035	38.56481	7.0280121
Class2	30.62203133	37.51148	172.1651	36.93422	0.5772672
Class3	37.7935	44.22849	233.8045	38.40653	5.8219577
Class4	20.55098061	59.0875	464.2859	45.45926	13.628234
All Classes	23.5039529	46.26443	296.2369	39.50047	6.7639548

Table 16 Average monthly revenue, usage and contract value among each latent class

Here, we introduce variable REPLACED_MONTH as the actual survival month for uncensored cases and Latent Class Regression estimated survival month for censored cases. AVGREV is the monthly service payment averaged over the lifetime time of a certain customer; AVGMOU is the monthly phone usage over the lifetime time of a customer; TOTMRC_M is the current subscriber contract of a customer; and CONCTACT_DIFFER is the difference between actual payment and contract plan.

From Table 16, we can see that the average tenure of customers in the dataset is around the 2-year contract period, with a relatively high usage and moderate revenue compared to all its subclasses. Moreover, we find that the customer who makes a lot of phone call thus paying a lot to the telecommunication company every month is the one who is very likely to leave the company before one or 2-year contract period (class 1 and class 4 respectively). On the other hand, in class 2, averagely, we have most people stay in the company for at least 2 years. In fact, those customers make relatively fewer calls and pay the least to the company every month compared to the others. Last but not least, customers in class 3 have moderate phone usage and revenue to the company; and comparatively they have the longest survival time, as up to 3 years.

Accordingly, some business insights can be generated based on the above observation. The company should act differently towards those four different classes. For class 1, promotion around the 1st year contract will be a good choice. For class 2, up-selling and cross-selling should be regularly carried out to increase their revenue but attention should also be paid since new contract may trigger the churn and make these customers leave the company earlier. For class 3, company should not offer renewal contract but keep moderate effort at the end of the 3rd year. For class 4, high intensity effort should be spent at both the end of the first and the second year. Moreover, company should continue to provide competitive offers to avoid these profitable customers leave the company.

CHAPTER 5 COMPARISIONS OF ALTERNATIVE METHODS

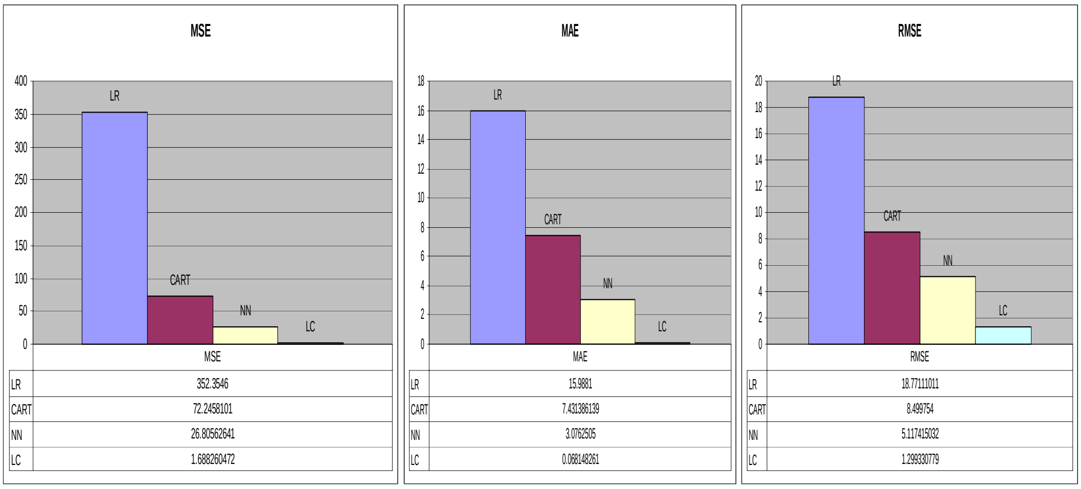
5.1 Accuracy Comparisons

Following the Survival Neural Networks research by Mani et al. (1999), we calculate the arithmetic mean differences (MAE), the arithmetic mean of the squared differences (MSE), and the rooted arithmetic mean of the squared differences (RMSE) between actual lifetimes and predicted lifetimes of the four prediction methods. Results for those uncensored cases in our testing set are summarized in Table 17. In the testing experiment, Latent Class Regression achieves the lowest value in RMSE (1.299); followed by Neural Networks (5.117), CART (8.4996), and Linear Regression (18.771). What is more, Latent Class Regression dominates the other methods among all the three accuracy measurements, which gives us an accurate and relatively stable result.

Accuracy Comparison of Alternative Methods				
ACCURACY	LR	CART	NN	LC
MAE	15.9881	7.431386	3.076251	0.068148
MSE	352.355	72.24581	26.80563	1.68826
RMSE	18.7711	8.499754	5.117415	1.299331

Table 17 Survival Time Predication Accuracy Comparison for Alternative Methods

To clearly illustrate the methods' prediction accuracy, the results of the methods are compared in Figure 24.



	MSE
LR	352.3546
CART	72.2458101
NN	26.80562641
LC	1.688260472

	MAE
LR	15.9881
CART	7.431386139
NN	3.0762505
LC	0.068148261

	RMSE
LR	18.77111011
CART	8.499754
NN	5.117415032
LC	1.299330779

Figure 24 Accuracy Comparison for Alternative Methods (MAE, MSE, RMSE)

5.2 Extreme Value Estimation Comparison

As noted in the research paper of Mani et al. (1999), a further step to compare the performance of different data mining models is to study how they treat extreme subjects, i.e. those subjects who have short or long survival times. In order to do so, we plot predicted survival time against actual survival time for the uncensored cases to visually inspect the performance of different methods.

In each scatter-plot, the X-axis is the predicted survival time by different methods, the Y-axis simply presents the actually survival time. For the purpose of visual inspection, we only plot 500 uncensored cases tested by each specific method. The extreme subjects, who have relatively short or long survival months are supposed to display themselves respectively in the low-left or up-right corner in the scatter-plots below. Thus, the clustering of points in the middle of a scatter-plot indicates that the method either overestimates short-lived subjects or underestimates long-lived subjects. On the other hand, a well-stretched scatter-plot indicates that the method gives an accurate estimation even for the cases with extreme survival time.

The scatter plots for Linear Regression Method, Regression Tree Method, Neural Networks Method, and Latent Class Regression Method are shown in Figure 25, 26, 27, and 28 respectively.

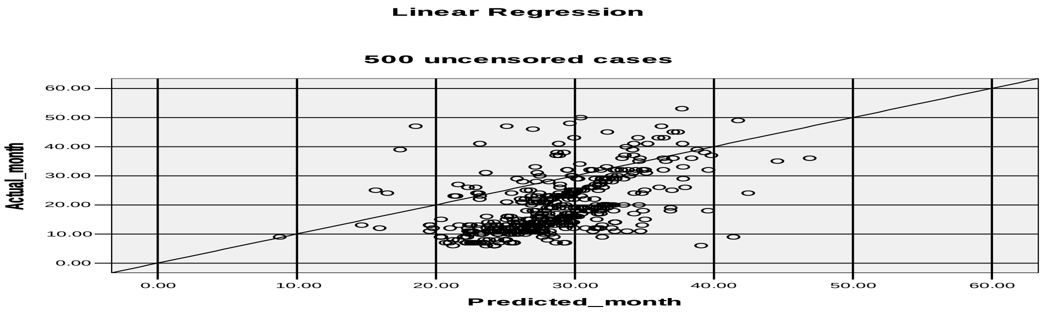


Figure 25 Scatter Plot for Predicted Survival Time of Linear Regression Model

As shown in Figure 25, cases group in the middle of the plot (around 30th month). On the other hand, few cases are plotted in the low-left or up-right corner.

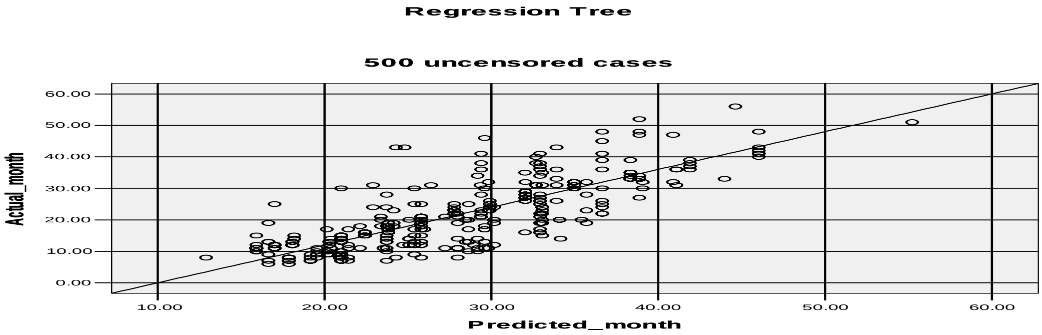


Figure 26 Scatter Plot for Predicted Survival Time of Regression Tree Model

Compared to the scatter plot of Linear Regression Model, the Regression Tree Model generates a relatively more accurate estimation for those long-lived or short-lived customers. However, Regression Tree Model still has outliers (those cases whose survival times are either underestimated or overestimated) as shown in Figure 26.

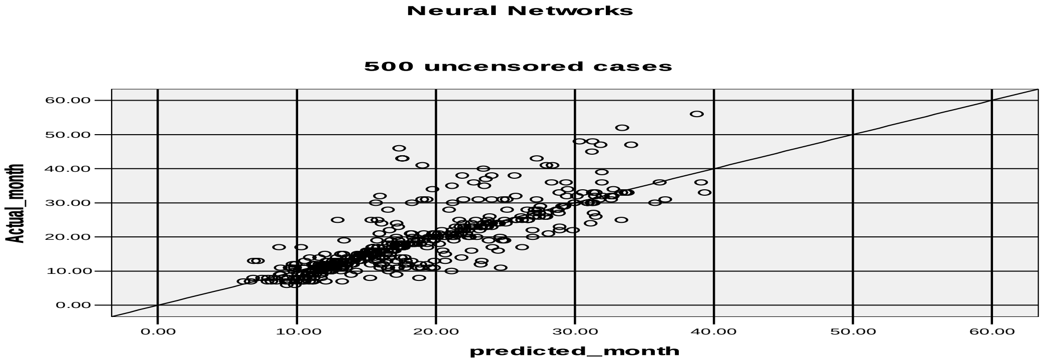


Figure 27 Scatter Plot for Predicted Survival Time of Neural Networks Model

From Figure 27, Neural Networks Model stretches along the fit-line, which indicates a better estimation for those subjects with extreme survival times. Still, like Regression Tree Model, it also generates outliers scattering in the middle of the plot.

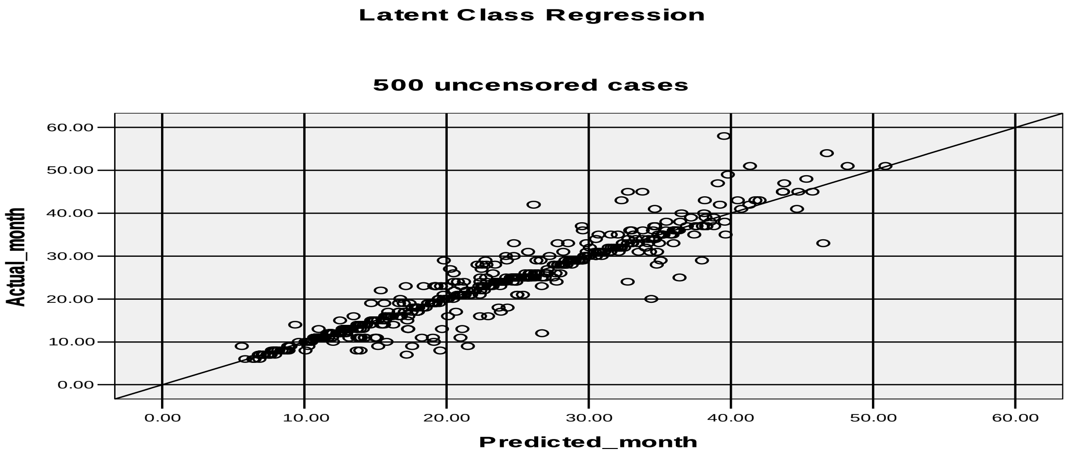


Figure 28 Scatter Plot for Predicted Survival Time of Latent Class Regression Model

Compared to the scatter plots for the other models listed above, Latent Class Regression gives the best estimation from the visual inspection of Figure 28. Moreover, this result is supported by the accuracy comparison in the Section 5.2.

From Figure 25 to Figure 28, it can be clearly observed that the Latent Class Regression is much better at predicting survival time for the cases with extreme survival time than the rest of the methods. Moreover, the points in the Linear Regression scatter-plot are clustered around 30th month (Figure 25), resulting in grossly overestimating CLV for short-lived customers and underestimating CLV for long-lived customers. On the other hand, Latent Class Regression has a more reasonable distribution of survival time (Figure 28), even for short-lived and long-lived customers. It is apparent that the Latent Class Regression method generally predicts lifetimes that are less extreme. It is probably because that unlike conventional survival analysis, Latent Class Regression does not use a presumed distribution of hazard function.

Through the discussion from Section 5.1 to Section 5.2, Latent Class Regression method generates the most accurate prediction results based on our experiments. Thus compared to other methods i.e., Neural Networks, CART, and Linear Regression, we can safely draw a conclusion that Latent Class Regression is the most suitable method for the prediction of survival time for censored cases.

5.3 CLV Calculation

Intuitively, having produced a more accurate survival time for those censored cases, we can have a more accurate Customer Lifetime Value calculation accordingly. In this section, we will further calculate CLV with the predicted survival time to prove this point.

As mentioned in Chapter 3, we follow the way that many researchers choose, to work out a solution for CLV application. We simply concentrate on modeling LOS alone, while assuming the current value as the future value (Mani et al., 1999; Rosset et al, 2003). Moreover, complying with this idea, some simple and logical assumptions are made to simplify the CLV calculation equation. Based on those assumptions and from the facts of our telecommunication dataset, we present the details of CLV calculation as below.

Facts from the datasets:

From the facts of our telecommunication dataset, we select the terms for value and cost as follows.

Value term variables:

- Avgrev: average monthly revenue through the lifelong of a customer.

Cost term variables:

- There is no specific cost term variable in this dataset, thus the only option for us is either to predict a fixed or non-fixed cost for each customer or just to ignore the cost term completely.

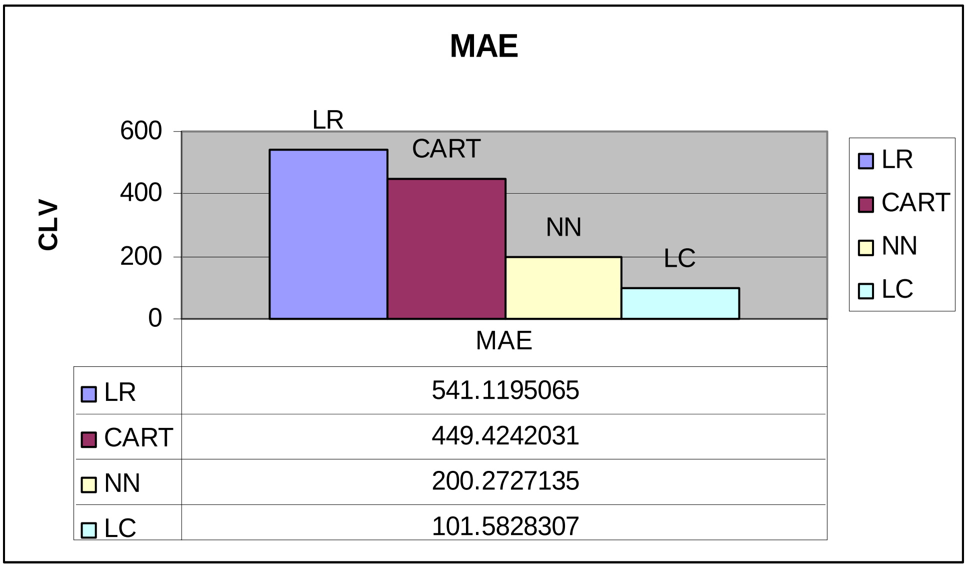
Thus from the facts of this dataset, the equation $CLV = Mean\_V \times S\_pred$ turns into this equation $CLV = Avgrev \times S\_pred$.

Test Results:

Following the discussion in Chapter 3, we conduct a two level CLV calculation process, and present the results as below. The results are summarized in Table 18, Figure 29, Figure 30, and Figure 31.

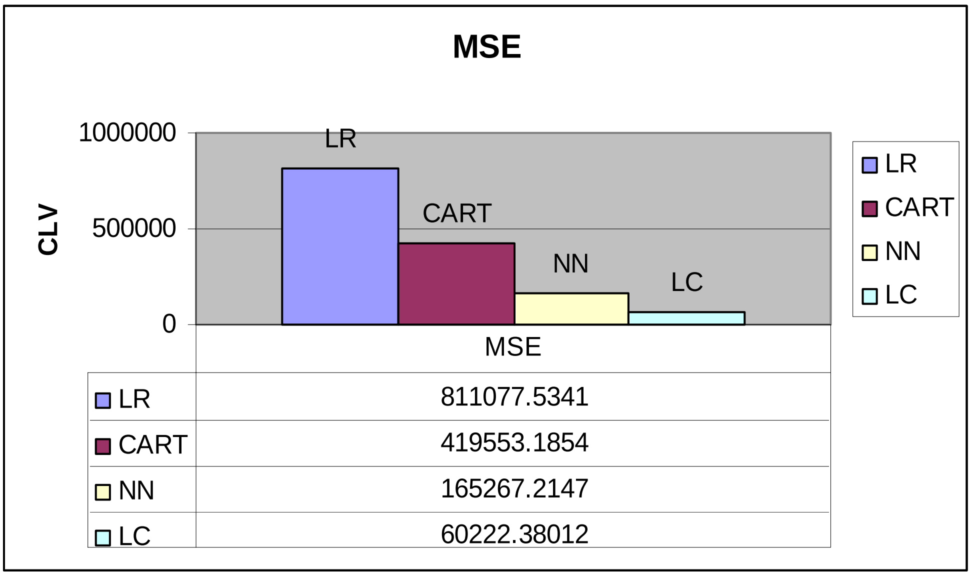
	MAE	MSE	RMSE
LR	541.1195	811077.5	900.5984
CART	449.4242	419553.2	647.7293
NN	200.2727	165267.2	406.5307
LC	101.5828	60222.38	245.4025

Table 18 CLV Calculation Comparison among Methods



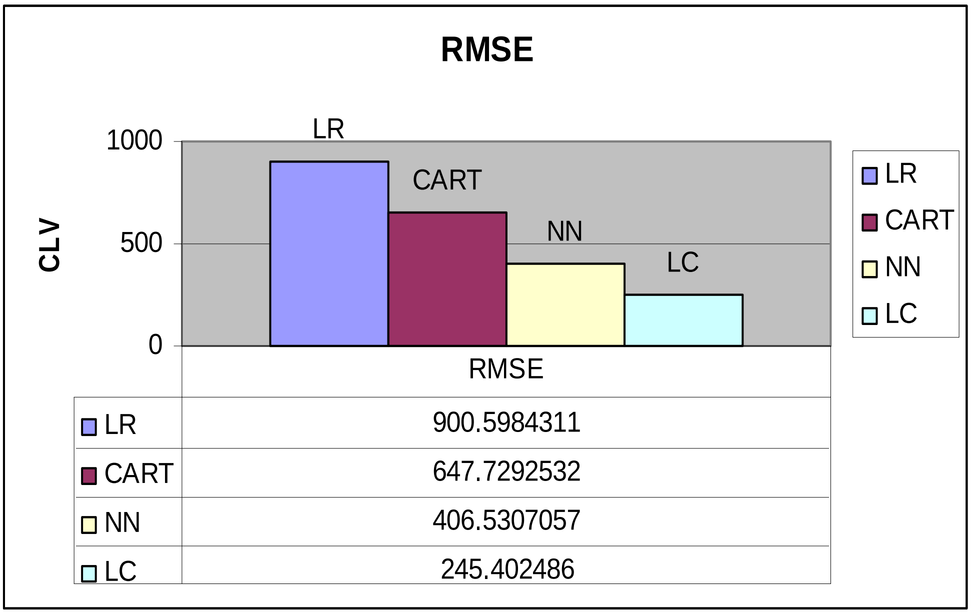
	MAE
LR	541.1195065
CART	449.4242031
NN	200.2727135
LC	101.5828307

Figure 29 MAE Comparison of CLV Calculation



	MSE
LR	811077.5341
CART	419553.1854
NN	165267.2147
LC	60222.38012

Figure 30 MSE Comparison of CLV Calculation



	RMSE
LR	900.5984311
CART	647.7292532
NN	406.5307057
LC	245.402486

Figure 31 RMSE Comparison of CLV Calculation

It is very clear that Latent Class Regression is again the best method for calculating customer's value term. This also indicates that a more accurate prediction of survival time can return a more accurate result of CLV calculation. Therefore, we can safely draw a conclusion that our proposed method is a better and more suitable approach for CLV calculation.

CHAPTER 6 FURTHER ANALYSIS

6.1 Motivation of Further Analysis

Business practice does not stop even though we already obtain the knowledge of customer's survival time or Customer Lifetime Value (CLV). However, a more accurate estimation of survival time and CLV can better further analyses; therefore indirectly benefits the marketing performance in business world. Thus, the focus of this chapter is to evaluate the proposed data mining method to help companies' marketing practice. Moreover, by carrying out some further actions, we can have a measure on how good is our proposed integrated data mining approach on CLV problem for real applications. By focusing this research study on the telecommunication industry, we choose to conduct one of the most popular analyses in this specific field, namely, a calling plan recommendation.

As proven by Mani and his colleagues (1999), the time around contract expiration is the time for most customers to churn (leaving the company). Thus, retention effort must be spent before that time in order to reduce the churn rate. However, we need to target the customers and their relative business strategies before we act. Still, we believe that information stays in the historical and analyzed data, such as the existing calling plan, customers' survival time, Customer Lifetime Value and so on. Further more, besides customer retention, up-selling and cross-selling should also be taken into consideration for the purpose of increasing both long-term and short-term profits for the company.

Keaveney (1995) proved the insight that pricing is the key factor for a customer to switch to competitors. In his paper he claimed that customers switch because of dissatisfaction with high price, price increases, unfair pricing practices, deceptive pricing practices and so forth. Because of that, companies might better off offering customer new contractual calling plan based on their CLV, survival time and their existing calling plans to avoid the customers from switching but to increase their profits instead. By encouraging customers to switch calling plans, a company is shown how to increase its long-term and short-term profits through increased retention rates and increased monthly service usage. Based on the above discussion, we suggest three contract renewal actions, namely, retaining, up-selling and offering more discounts. This gives us more information of the subscribers and makes it more efficient to carry out marketing activities accordingly.

6.2 A Calling Plan Recommendation Strategy

Based on the idea described above, in this section, we propose a calling plan recommendation strategy. It is based on the proposed data mining approach for survival analysis, in order to further evaluate the methods to help companies' marketing practice. The goal of the further analysis is to develop and assess the proposed data mining method. Moreover, this calling plan recommendation strategy is developed based on some widely accepted principles and measurement constructs from the fields of data mining (Hand, 1997). The proposed method for contract renewal is to build a decision tree model for classifying customers into different groups according to the relative renewal actions. In detail, two sets of decision rules will be generated according to the previous research results of statistical survival analysis and the proposed integrated data mining method respectively. By comparing the misclassification rate and misclassification cost of the two rule sets, we can further investigate the advantage of our integrated approach over other methods in terms of generating useful business knowledge.

6.2.1 Test Design (A Supervised Machine Learning Schema)

Before making any recommendation and seeking for the targeted customer groups, ground rules (guide lines) must be made about how to conduct contract renewal actions based on the historical and analyzed information. The purpose of this pre-processing action is to set a class label on each case. In this way, the proposed method can learn from the training cases with classification information and generate classification rules accordingly. Next, the decision rules will be evaluated on the testing cases with classification labels. Therefore, this procedure makes the method a supervised machine learning approach, in which classification information is defined before further analyses are conducted. This procedure makes two critical issues. First, how to define the initial groups for this supervised learning approach; and second, how to evaluate the rule-sets of calling plan recommendations generated based on the information in hand. For the first issue, customer groups are defined based on their actual calling behavior, their current calling plan, and the predicted survival time derived from the application in the previous chapters. For not making our judgment too subjectively, the initial classification rules are defined in a way such that they are easy to understand and logical in taking actions. First of all, the suggested actions of contract renewal are categorized as up-selling the customer, offering more discounts to the customers, and simply making no change at all. For each of the three renewal actions, we define three customer groups on which the actions are logical to be made accordingly. Moreover, these customer groups are classified based on customer's actual calling behavior, current calling plan and predicted survival time. For those customers who never make phone calls more than the volume limits stated in their current calling plans, we define them as group 1 customers. Accordingly, the calling recommendation for this group of customers is not to renew their contracts, since either up-selling or discounts offering is not rational (these customers will not upgrade their service because they do not even make full use of the existing service, and offering more discounts is just a waste of money for the company since they are not price sensitive customers). For the customers who make much more phone calls than the current calling plan and stay with the company longer than 12 months (the common mandatory contract period), they are defined as group 2 customers. To this group of customers, up-selling will be a good option since they are very loyal and profitable customers. The rest of the customers in the entire customer dataset are those who make more calls than the usage listed on the original calling plan but will leave the company within the first 12 months contract period. For these customers, retaining efforts are needed, thus offering more discounts is the right decision of contract renewal. Table 19 gives the detailed information of contract renewal actions and the rules that assign customers into these groups accordingly. In this table, the measurement "Contract_differ" indicates the difference between actually monthly phone usage and the standard charges stated in the customer's calling plan. The "survival time" is the one derived from two different applications (conventional survival analysis or integrated approach) respectively in the previous sections. To ensure a valid assessment of different rule sets, the variables used here for defining the pre-groups will not be used in the decision tree construction process again.

	Moves	Contract_differ	survival time
1	No change	<=0	any
2	Up-selling	>0	>1 year
3	Discount	>0	<1 year

Table 19 Contract Renewal Suggestions and Predefined Groups

While there are commonly accepted measures such as *misclassification rate*, describing the performance of the classifier, a simply misclassification rate comparison does not take into account the consequences of its performance. As such, there appears to be agreement within the literature that *misclassification costs* should be used (Hand, 1997) for evaluation. This is calculated by multiplying the pay-off of each outcome with its frequency. So for a three-treatment classifier like the contract renewal example in this study, there are six outcomes.

For the matter discussed above, we assigned different costs for different misclassification scenarios for each classification groups. First of all, losing customer is assumed to be unbearable for the company and costs twice as much as other losses. Table 20 gives the misclassification scenarios and their costs accordingly. For example, when the model misclassified a group 1 customer into group 2, which means an up-selling effort is spent on an uninterested customer, very likely this customer will not take the offer and leave for another service provider. On the other hand, when the model misclassified a group 1 customer into group 3, resulting in offering more discount to the customer who does not make enough phone calls each month. As a result, company will lose money. For example, the number 2 enclosed in a pair of parentheses indicates that the cost for misclassifying a group 3 customer into group 2 is twice as much as other misclassifications.

Misclassification Cost Matrix			
Predefined Group	1	2	3
1	0	lose customer (2)	lose money (1)
2	lose money (1)	0	lose money (1)
3	lose customer (2)	lose customer (2)	0

Table 20 Misclassification Cost Matrix

Another issue for this supervised learning approach is how to evaluate the generated rule sets according to the pre-defined groups. This requires us to know the exact calling behavior, bill details and survival time of the cases to be evaluated. Fortunately, this is the case for all the customers who have already churned. Therefore, we generate the rule sets from the information of all the customers (which includes both churners and censors) and evaluate them only on the churned cases.

6.2.2 Experiments Schema and Test Results

For the rule sets comparison, we apply the method on two groups of training and testing sets each containing five training sets and five testing sets respectively. One group of training and testing sets is constructed based on Conventional Survival Analysis and another one is derived from the data mining approach. These two groups of training and testing sets will be used to generate the decision rules of calling plan recommendation. Overlapping in the training and testing sets is avoided and a fair comparison is assured in the process of making the experiment sets. What is more, we average and return the classification results among those five training and testing sets. The detail design of training and testing schema is as follows.

In order to evaluate the decision rule sets and to ensure a fair comparison, we design a training and testing sets schema in the manner of one to one correspondence. Firstly, we randomly divide all the 50000 uncensored cases into five different groups as our testing sets. Each of the five groups has 10000 uncensored cases. The rule sets learned from the training sets will be evaluated on these five testing sets. For each testing set, two groups of training sets are constructed based on the results derived from Conventional Survival Analysis and the data mining approach respectively. The only difference between the cases in these two different groups of training sets is the survival times of censored cases. One is derived from conventional survival analysis and the other one is derived from the proposed data mining approach. Moreover, for the purpose of learning the tree models with the information of both uncensored and censored cases, we compose the training set of two parts. The first part is the 10000 censored cases randomly selected from all the 50000 censored cases. The second part includes 10000 uncensored cases randomly selected from the other four unrelated testing sets which contain 40000 uncensored cases. Based on this design, we have five testing sets each containing 10000 uncensored cases and five corresponding training sets each containing 10000 censored cases and 10000 uncensored cases. For example, to evaluate the testing set 1, we construct the corresponding training set. First, we randomly select 10000 cases from the 50000 censored cases. Then the other 10000 uncensored cases are randomly selected from the 40000 uncensored cases of testing sets 2, 3, 4, and 5. All other training and testing sets are generated similarly.

To construct the decision trees model, we choose the same 17 variables used in the regression models as the classification splitters to generate the two rule sets. After the decision tree models are constructed, all the parameters are shown to be statistically significant and intuitively make sense. Further more, we designate the cost matrix in Table 21 to the decision tree construction process to ensure the accuracy of the classification of groups 2 and 3.

The results for both rule sets on each testing set are averaged and listed in Table 21:

Customer Groups	Data Mining Approach	Survival Analysis	Difference
1	61.45%	61.19%	0.26%
2	52.50%	46.30%	6.20%
3	93.68%	89.65%	4.02%

Table 21 Classification Accuracy among Groups between Two Methods

It is clear that the rule set generated based on the data mining approach is better than the one based only on survival analysis in terms of classification accuracy. The difference between these two sets of results is significant at 0.05 level based on a Student's t-Test. Moreover, since the misclassification of group 2 and 3 can directly lead to the loss of existing subscribers, it makes the comparison of the misclassification rate between two groups a more important factor. For group 2, the rule set generated based on the data mining approach outperforms the other rule set by 6.2%. For group 3, the former one outperforms the latter by 4.02%.

We assume that customers will leave the company as long as they are misclassified into a wrong group, and inappropriate contract renewal recommendations are suggested. Based on the results above and the facts from our dataset, the rule set generated from the data mining approach can save about 10% more customers (about 10000 customers in our dataset) than the rule sets generated based only on survival analysis. Moreover, we investigate the profits generated from these 10% customers. The average monthly profits generated for group 2 customers and group 3 customers are 62 USD and 72 USD respectively. This indicates that the proposed data mining method can save up to about 262000 USD a month for the company when compared to the conventional survival analysis. To sum up, our proposed integrated data mining approach can not only produce a more accurate estimation for survival time and Customer Lifetime Value, but also can better company's marketing practice in a real business world.

CHAPTER 7 IMPLICATIONS AND CONTRIBUTION

7.1 Making Good Prediction Method for CLV Problem in Marketing Practice

Return to the research questions: “Will it be a better method to handle the CLV problem, especially the customer duration of relationship by combining conventional survival analysis and data mining methods together? How is the proposed method’s performance on helping the companies’ marketing practice in real business scenarios? ” They are the original objectives of this research. After exploring these questions in the previous chapters, we have obtained the answers. There are results from different stage of experiments and author’s general analysis.

7.1.1 Results from Combined Survival and Data Mining Method

Summarizing the test results in Chapter 4 and Chapter 5, we suggest that by combining conventional survival analysis and data mining approaches, an integrated method is built. This integrated method can produce a more accurate estimation for the survival time of a censored case, thus a more accurate CLV calculation. Moreover, through the comparison among different statistical and data mining methods, we can safely draw a conclusion that the Latent Class Regression model is the most suitable method for the problem of Customer Lifetime Value in terms of accuracy and interpretability among all methods.

In spite of their ability of imputing the “partially missing” data for censored cases, conventional survival analyses suffer from some disadvantages in their modeling mechanism and predication schema. For examples, the pure-parametric approach uses a presumed distribution function and the semi-parametric approach assumes a baseline hazard function dependent only on survival time. However, some data mining methods, as demonstrated in the previous chapters, can overcome these limitations of conventional survival analysis approaches and thus generate a more accurate estimation for survival time and a better CLV calculation as well. Despite of its simplicity during the process of building different prediction models, our proposed integrated method also gives fair prediction accuracy on the experiment datasets.

For the data mining methods comparison, we intend to find the most suitable method with the combination of survival analysis for survival time prediction and CLV calculation. The methods for comparison include Linear Regression, Neural Networks, Regression Tree, and Latent Class Regression. We apply these methods on a large real world telecommunication industry dataset and evaluate the methods in terms of prediction accuracy and their ability of generating business insights (or model interpretability). After a few experiments, the results on the testing dataset suggest that Latent Class Regression not dominates other methods with its prediction power.

7.1.2 Results from Further Analysis

Chapter 6 encompasses a calling plan strategy which is based on our data mining approach for survival analysis within the scope of a telecommunication industry. The benefit of this further analysis is that it provides a more detailed and direct measurement to further evaluate our proposed data mining method to help companies' marketing practice.

In this part of the study, we classify mobile service subscribers into 3 different customer groups according to the 3 contract renewal actions. We apply Decision Tree models to conduct the classification process based on the results derived from the proposed integrated method and the conventional survival analysis. Misclassification rate, number of customers saved, and amount of value saved are used to evaluate the two rule sets on our telecommunication dataset. The experiment results show that overall the decision rule sets derived from our data mining approach outperform conventional survival analysis by 10% in terms of misclassification rate. Therefore, complying with our assumption made in Chapter 6, at an average level, it can save up to about 10000 customers and thus about 262000 USD a month for the company when compared to the conventional survival analysis. Moreover, it proves that by using our proposed integrated data mining method, we could have better classification results than by applying survival analysis alone. Moreover, the method can retain more customers and increase company's long-term and short-term profits.

7.2 Contribution of the Study

There are three contributions of this study. Firstly, in this research, based on the combination of the conventional survival analysis and data mining methods we conduct experiments among different statistical and data mining methods and compare their prediction results in terms of accuracy and interpretability. To be noted, based on the literature review I introduced Latent Class Regression as a powerful prediction method which is never been applied in the area of survival analysis or CLV calculation. Secondly, we select the most suitable method for the estimation of customer duration of relationship and thus Customer Lifetime Value calculation based on the comparison among four statistical and data mining methods. Specifically, Latent Class Regression presents itself as the most suitable method for the CLV problem compared to other methods for its strong prediction power and interpretability potential. Thirdly, in this research study, we focus on the problem of CLV within the scope of telecommunication industry. This makes this study a good exploration of the CLV concept in the application of this specific industry which is facing with tough competition of customers among each other. Moreover, unlike other previous researches, in this study we develop an integrated method for investigating useful knowledge for real business applications. To be specific a calling strategy is developed for the telecommunication industry. It enlarges the proposed model's adaptability in real business scenario.

7.3 Further Research

There are still spaces for further study. For example, more advanced techniques such as Evolutionary Programming could be applied as a promising CLV estimation method. Though Evolutionary Programming has not been applied in this field yet, its strong prediction power suggests itself to be a competitive potential for the CLV problem. What is more, in Chapter 6, empirical studies need to be conducted to evaluate the proposed method through real-world marketing practice. In detail, classification rate should be calculated based on the real classification situation after the strategy is carried out on a certain group of customers. Company’s profits and number of customers saved by different rule sets need to be clarified based on the real data through the real-world marketing practice. In this way we can be truly convinced that our integrated method is a better approach to help company’s marketing practice.

Moreover, because of the constraints in the experiment dataset, in this research study we follow some researchers’ work and make a few assumptions, which results in a relatively simple CLV calculating equation. However, a more complex equation needs to be developed to generate a more practicable Customer Lifetime Value. Still, a dataset with more information of customers’ value and cost is required for the purpose of a complete CLV calculation.

Last but not least, our discussion, especially for the further analysis described in Chapter 6, lays concentration on telecommunication industry. We will further apply the proposed integrated data mining approach to other business domains, such as direct marketing, personal loan banking, and B2B transactions in the future.

BIBLIOGRAPHY

Agresti, A. 2002. Categorical Data Analysis. Second Edition, New York: Wiley.

Barlow, R. E., Marshall, A. W., and Proschan, F. 1963. Properties of Probability Distributions and Monotone Hazard Rate. Ann. Math. Statist, 34: 375-389.

Bartholomew, D. J. and Knott, M. 1999. Latent Variable Models and Factor Analysis. London: Arnold.

Bell, D., Deighton J., Werner, J. R., Roland T. R., and Gordon S. 2002. Seven Barriers to Customer Equity Management. Journal of Service Research, 5: 77-85.

Berger, P. D. and Nada, I. N. 1998. Customer Lifetime Value: Marketing Models and Applications. Journal of Interactive Marketing, 12: 17-30.

Berkson, J. and Gage, R. P. 1950. Calculation of Survival Rates for Cancer. Staff Meetings of the Mayo Clinic, 25:270-286.

Biganzoli, E., Boracchi, P., Mariani, L., and Marubini, E. 1998. Feed Forward Neural Networks for the Analysis of Censored Survival Data: A Partial Logistic Regression Approach. Statistics in Medicine, 17:1169-1186.

Blattberg, R. C. and Deighton J. 1996. Manage Marketing by the Customer Equity Test. Harvard Business Review, July/August.

Bolton, R., and Dew J. 1991. A Longitudinal Analysis of the Impact of Service Changes on Customer Attitudes. Journal of Marketing, 55: 1-10.

Breiman, L., Friedman, J. J., Olshen, R. A., and Stone, C. J. 1984. Classification and Regression Trees. Monterey, California: Wadsworth and Brooks Publishing.

Brown, S. F., Branford, A., and Moran, W. 1997. On the Use of Artificial Neural Networks for the Analysis of Survival Data. IEEE Transaction in Neural Networks, 8:1071-1077.

Courtheoux, R. 1995. Customer Retention: How Much to Invest. In K. Hartmann, J. Banslaben, and H. Seymour (Eds.), Research and the Customer Lifecycle, New York: DMA.

Cox, D. R. 1972. Regression Models and Life Tables. Journal of the Royal Statistical Society Series B, 34: 187-220.

Cutler, S. J. and Ederer, F. 1958. Maximum Utilization of the Life Table Method in Analyzing Survival. Journal of Chronic Disease, 8: 699-712.

De Laurentiis, M. and Ravdin, P. M. 1994. A Technique for Using Neural Network Analysis to Perform Survival Analysis of Censored Data. Cancer Letter, 77: 127-138.

Dillon, W. R. and Kumar, A. 1994. Latent Structure and Other Mixture Models in Marketing: An Integrative Survey and Overview. Chapter 9 in R. P. Bagozzi (Eds.), Advanced methods of Marketing Research, 352-388,Cambridge: Blackwell Publishers.

Faraggi, D. and Simon, R. 1995. A Neural Network Model for Survival Data. Statistics in Medicine, 14:73–82.

Fleming, T. R. and Harrington, D. P. 1991. Counting Processes and Survival Analysis. New York: John Wiley & Sons.

Fogelman-Soulie, F., Gallinari, P., LeCun, Y., and Thiria, S. 1987. Generalization Using Back-Propagation. The First International Conference on Neural Networks, (San Diego, California), IEEE.

Gamel J. W. and Vogel, R. L. 1997. Comparison of Parametric and Non-Parametric Survival Methods Using Simulated Clinical Data. Statistics in Medicine, 16:1629-43.

Gehan, E. A. 1969. Estimating Survival Functions from the Life Table. Journal of Chronic Diseases, 21: 629-644.

Håkansson, H. 1982. International Marketing and Purchasing of Industrial Goods: An Interaction Approach. Chichester: John Wiley & Sons.

Hald A. 1949. Maximum Likelihood Estimation of the Parameters of a Normal Distribution Which is Truncated at a Known Point. Skandinavisk Aktuarietidskrift, 32:119-134.

Hand, D. J. 1997. Construction and Assessment of Classification Rules. West Sussex, England, UK: John Wiley & Sons.

Harris, R. J. 1975. A Primer of Multivariate Statistics. New York: Academic Press.

Helsen, K. and Schmittlein, D. C. 1993. Analyzing Duration Times in Marketing: Evidence for the Effectiveness of Hazard Rate Models. Marketing Science, 11: 395-414.

Hoekstra, J. C. and Huizingh, E. K. R. E. 1999. The Lifetime Value Concept in Customer-Based Marketing. Journal of Market Focused Management, 3: 257-74.

Hughes, A. M. 1996. Database Marketing Has Arrived; Determining the Bottom Line. In Proceedings of the DMA 79th Annual Conference, New Orleans.

Hughes, A. M. 1997. Customer Retention: Integrating Lifetime Value into Marketing Strategies. Journal of Database Marketing, 5: 171-178.

Hughes, A. M. 2002. How Lifetime Value is Used to Evaluate Customer Relationship Management. Database Marketing Institute, http://www.dbmarketing.com/articles/Art194.htm.

Jackson, D. R. 1994. Strategic Application of Customer Lifetime Value in the Direct Marketing Environment. Journal of Targeting, Measurement and Analysis for Marketing, 3: 9–17.

Kaplan, E. L. and Meier, P. 1958. Nonparametric estimation from incomplete observations. Journal of American Statistical Association, 53: 457-481.

Keane, T. J., and Wang P. 1995. Applications for the Lifetime Value Model in Modern Newspaper Publishing. Journal of Direct Marketing, 2: 59-66.

Keaveney, S. M. 1995. Customer Switching Behavior in Service Industries: An Exploratory Study. Journal of Marketing, 59: 71-82.

Lapuerta, P., Azen, S. P., and LaBree, L. 1995. Use of Neural Networks in Predicting the Risk of Coronary Artery Disease. Computing Biomedical Research, 28: 38-52.

Lazarsfeld, P. F. 1950. The Interpretation and Computation of Some Latent Structures. Chapter 11 in Stouffer (Eds.), Measurement and Prediction, Princeton: Princeton University Press.

Mani, D. R., James, D., Andrew, B., and Piew, D. 1999. Statistics and Data Mining Techniques for Lifetime Value Modeling. Journal of Association for Computing Machinery, 1: 113-143.

Mohammed Z. and Kotze, D. 2005. Survival Data Mining in the Telecommunications Industries: Usefulness and Complications. WIT Transactions on Information and Communication Technologies, 35.

Monik, S. 2004. Customer Lifetime Value and its determination using the SAS Enterprise Miner and the SAS OROS Software. SAS institution: German.

Mulhern, F. J. 1999. Customer Profitability Analysis: Measurement, Concentration, and Research Directions. Journal of Interactive Marketing, 13: 25-40.

Narver, J. C. and Slater, S. F. 1990. The effect of market orientation on business profitability. Journal of Marketing, 54: 20-35.

Neal, R. M. 2001. Survival Analysis Using a Bayesian Neural Network. Joint Statistical Meetings Report, Atlanta.

Novo, J. 2001. Maximizing Marketing ROI with Customer Behavior Analysis. http://www.drilling-down.com.

Ohno-Machado, L. 1996. Sequential Use of Neural Networks for Survival Prediction in Aids. Journal of the American Medical Informatics Association, 3: 170-174.

Oliver,R. L. 1980. A Cognitive Model of the Antecedents and Consequences of Satisfaction Decision. Journal of Marketing Research ,17: 460-469.

Petrison, L., Blattberg, R. C., and Wang, P. 1997. Database Marketing – Past, Present and Future. Journal of Direct Marketing, 11: 109-125.

Pfeifer, P. E., Mark, E. H., and Robert, M. C. 2004. Customer Lifetime Value, Customer Profitability, and the Treatment of Acquisition Spending. Journal of Managerial Issues, 2:110-124.

Rajkumar, V. and Kumar, V. 2004. A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. Journal of Marketing, 68: 106-125.

Ravdin, P. M. and Clark, G. M. 1992. A Practical Application of Neural Network Analysis for Predicting Outcome of Individual Breast Cancer Patients. Breast Cancer Research, 22:

285-293.

Reichheld, F. F. and Sasser, W. E. 1990. Zero Defections: Quality Comes to Services. Harvard Business Review, 5: 105-111.

Reinartz, W. and Kumar, V. 2000. On the Profitability of Long Lifetime Customers: An Empirical Investigation and Implications for Marketing. Journal of Marketing, 64: 17-35.

Romano, N. C. 2001. Customer Relationship Management Research: An Assessment of Sub Field Development and Maturity. In Proceedings of the 34th Hawaii International Conference on System Sciences.

Rosset, S., Neumann, E., Eick, U., and Vatnik, N. 2003. Customer Lifetime Value Models for Decision Support. Data Ming and Knowledge Discovery, 7: 321-339.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. Learning Internal Representations by Error Propagation. In D. E. Rumelhart, and J. L. McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations (pp. 318--362). Cambridge, MA: MIT Press.

Rumelhart, D. and McClelland, J. 1986. Parallel Distributed Processing. Cambridge, Mass: MIT Press.

Rust R. T., Katherine N. L., and Zeithaml V. A. 2004. Return on Marketing: Using Customer Equity to Focus Marketing Strategy. Journal of Marketing, 68: 109-127.

SIEBEL white paper. 2002. Measuring and Managing Customer Lifetime Value. SIEBEL Systems, May 2002.

Skrondal A. and Rabe-Hesketh, S. 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. London: Chapman & Hall/CRC.

Stepanova, M. and Thomas, L.C. 2002. Survival Analysis Methods for Personal Loan Data. Journal of the Operational Research, 50: 277-289.

Storbacka, K. 1994. The Nature of Customer Relationship Profitability. Helsinki, Finland: Swedish.

Street, W. N. 1998. A Neural Network Model for Prognostic Prediction. In Proceedings of the Fifteenth International Conference on Machine Learning (ICML), Madison, Wisconsin, USA, 540-546.

Vavra, T. G. 1997. Improving Your Measurement of Customer Satisfaction. Milwaukee: ASQ Quality Press.

Venables, W. N. and Ripley, B. D. 1999. Modern Applied Statistics with S-PLUS, 3rd edition. Springer-Verlag.

Vermunt, J. K., and Van Dijk, L. 2001. A Nonparametric Random-Coefficients Approach: the Latent Class Regression Model. Multilevel Modelling Newsletter, 13: 6-13.

Wedel, M. and Kamakura, W. A. 1998. Market Segmentation: Concepts and Methodological Foundations. Boston: Kluwer Academic Publishers.

Werner, J. R and Kumar, V. 2003. The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. Journal of Marketing, 67: 77-99.

West, P. M., Brockett, P. L., and Golden, L. L. 1997. A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice. Marketing Sciences, 16: 370 –391.