Kwame Nkrumah University of Science and Technology, Kumasi

College of Science

Institute of Distance Learning

Department of Mathematics

Churn Model in the Mobile Telecommunications Industry in Ghana:

A Case Study of Vodafone Ghana

A Thesis Submitted to The Department of Mathematics in Partial Fulfilment of The Requirement For The Award of a Master of Science (MSc.) Degree in Industrial Mathematics.

BY

Godsway Roland Normeshie

Supervisor

Nana Kena Frempong

DECLARATION

I, Godsway Roland Normeshie therefore declare this submission is my own research work towards the award of Master of Science and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of other degree of the university, except where due acknowledgement dad been made in the text.

Godsway Roland Normeshie ……………………………
……………….

(PG 4068410) Signature Date

Certified by

Nana Kena Frempong …………………...........
………………..

Supervisor's Name Signature Date

F.K. Darkwa …………………...........
………………..

Head of Department's Name Signature Date

Prof. I.K. Dontwi ………………................
………………..

Dean of Institute of Distance Learning's Name Signature Date

ACKNOWLEDGEMENT

With my greatest great appreciation and sincere thankfulness, I would like to acknowledge the following people for their help, advice, encouragement and support which help me in finally complete of my study.

First of all my gratitude goes to God Almighty whose grace and loving kindness has brought me to the dawn of another era of my life. I am very much indebted to my parents whose strenuous effort and love have seen me through all these years of my academic career. To them I say, God richly bless you for your impact in my life.

I would like to express my gratitude to my supervisor Nana Kena Frempong for his kindness, guidance, comments, supervision and encouragement throughout this project. I also appreciate the effort of my colleague at Vodafone Ghana, Alfred Sako for his support.

Finally, I appreciate all those who in one way or the other contributed to this project. God bless you all.

DEDICATION

To my lovely and ever supportive wife, Mrs Rejoice Abui Normeshie and my children

ABSTRACT

In this study, the purpose was to propose an approach to assist Telecom Churn Management, use Kaplan Meier estimator to estimate survival probabilities for each of the value bands High Value Customer (HVC), Medium Value Customer (MVC) and Low Value Customer (LVC), to test whether there is differences in the survival curves of the three value bands and to model the effect of usage (Total MOU, GPRS Usage, Total SMS and Top up Amount) on the risk of churning using Cox Proportional Hazard Model.

Data on Prepaid customers was obtained from the start of SIM activation to date. The Kaplan Meier estimator and Cox Proportional Hazard model were the Statistical methods used in the analysis.

From the result, we observed that the median churn time for High Value Customers (HVC) is approximately 20 months, Medium Value Customers have a median churn time of 25 months and the Low Value Customers have 11 months of churn median time.

Generally, the Log-Rank test indicated significant differences among the survival function of the three value bands (HVC, MVC and LVC). From the Cox model total minute of use (Total MOU) and Top up Amount have significant effects on the survivorship.

We conclude that customers who have high top up amount have low risk of churning and have longer survival times.

LIST OF DEFINITIONS

(i) Call credit: prepaid customers pay their calls from this credit.

(ii) Churn: a term used by companies to denote the loss of customers.

(iii) Activation date: the date a new customer is registered.

(iv) Postpaid customer: a customer who is bound by a contract and pays a monthly sum in exchange for free call minutes.

(v) Prepaid customer: a customer who is not bound by a contract and who only pays for the calls he makes.

(vi) Top-up: the term used for raising the call credit.

(vii) Sim-card: a small electronic chip on which the mobile phone number is stored.

(viii) Total MOU: Total minutes of use

Table of Content

Title Page
Page

CHAPTER ONE

INTRODUCTION

1.1 Background of Study

In the telecommunication industry, customers are able to choose among multiple service providers and actively exercise their rights of switching from one service provider to another. In this fiercely competitive market, customers demand tailored products and better services at lesser prices, while service providers constantly focus on acquisitions as their business goals.

The last decade had seen rapid growth in the uptake of mobile communication technologies around the world. Mobile telephones in particular had demonstrated an exceptional ability to cut through hitherto obstinate barriers to adoption in developing countries.

The introduction of the wireless telecommunication and the wireless Internet broadband has changed the trend of the telecommunication industry in Ghana to become highly competitive. Customers opt to use wireless network rather than fixed network due to the mobility freedom. Furthermore, customers have the freedom to choose and switch from one package to another package offered by the service provider, or from one service provider to another service provider through the introduction of mobile number portability (MNP).

Recently, the mobile telecommunication market in Ghana has changed from a rapidly growing market, into a state of saturation and fierce competition. The focus of telecommunication companies has therefore shifted from building a large customer base into keeping customers in house. For that reason, it is valuable to know which customers are likely to switch to a competitor in the near future. Those customers are so called churned customers. Since acquiring new customers is more expensive than retaining existing customers, churn prevention can be regarded as a popular way of reducing the company’s costs. With this objective, this research is carried out for Vodafone Ghana. Vodafone is particularly interested in churn prediction of prepaid customers.

Prepaid customers are, as opposed to post-paid customers, not bound by a contract. With the introduction of the mobile number portability by the National Communication Authority, the competition among the service providers has grown so fiercely and the market is also getting saturated by the day. As a result of this, customers in the telecommunication industry are able to choose among multiple service providers and actively exercise their rights of switching from one service provider to another. In this fiercely competitive market, customers demand tailored products and better services at less prices, while service providers constantly focus on acquisitions as their business goals. Given the fact that the telecommunications industry experiences an average of 30-35 percent annual churn rate and it costs 5-10 times more to recruit a new customer than to retain an existing one. (Berson A., Smith S. and Thearling K. (2000). Building Data Mining Applications for Customer Relationship Management. New York: McGraw-Hill). Customer retention has now become even more important than customer acquisition.

For many incumbent operators, retaining high profitable customers is the number one business pain and has therefore shifted focus from building a large customer base to retaining existing ones. Customer churning has now become an issue of great concern to service providers in Ghana.

1.2Telecommunication Industry in Ghana

The country’s mobile phone penetration as captured on the National Communications Authority official website,has reached the 80.5 per cent mark as of the end of August, 2011. The major players in the industry are

(i) MTN

(ii) TIGO

(iii) VODAFONE

(iv) AIRTEL

(v) EXPRESSO

This means that out of a population of 24,722,485, there are19, 893,191 subscribers on at least one of the five active mobile networks in the country. The number of subscribers for January 2011 stood at 17.43 million.

From the latest industry statistics released on the National Communications Authority (NCA) official website, MTN is maintaining its lead as the country's biggest mobile phone company with 9.8 million subscribers, representing 48.49% of the market.

The company started the year with 8.9 million as of the end of January and has, in spite of stiff competition, managed to maintain its lead.

With this trend, MTN, which recorded its nine millionth subscribers before the end of the first quarter of this year, seems to be on track to register its 10 millionth customer by the end of the year. tiGO, operated by Millicom Ghana, also maintained its number two position having moved its subscriber base from 3.92 million in January to 4.2 million subscribers as of the end of August, to control 20.59 per cent of the market.

Vodafone, which is in the third slot, control 18.48% of the market with 3.73 million subscribers. Its subscriber base for January was 2.81 million.

The Indian-owned Airtel trails in the fourth position with approximately 10 per cent of the market with a total customer base of 2.0 million as of the end of the month under review, having held 1.6 million customers at the end of January.

The case of Expresso, which until recently, was known as Kasapa continues to worsen as its subscriber base dipped from 244,674 in January to 206,606 in August.

1.2.1Churn Prediction Modelling

“Churn” is a common phenomenon that occurs in telecom Industry. By “Churn” we mean those customers, who will be leaving in the near future. If we are able to predict in advance, the attributes of customers whom we are going to lose in near future one can take corrective action so that we can minimize this phenomenon.

In today’s competitive world, customer churn remains to be as one of the most pressing concerns for network providers. Through research, it is found that the cost of acquiring new customers is much higher than retaining the existing ones. However, predicting customer churn to tackle this main concern is an extremely difficult undertaking.

Churn prediction is currently a relevant subject in data mining and has been applied in the field of banking, mobile telecommunication, life insurances, and others. In fact, all companies who are dealing with long term customers can take advantage of churn prediction methods.

Models such as neural networks, logistic regression and decision trees are common choices of data miners to tackle this churn prediction problem. These models are trained by offering snapshots of churned customers and non- churned customers. The goal is to distinguish churners from non-churners as much as possible. When new customers are offered, the model attempts to predict to which class each customer belongs. Although in general this approach gives satisfying results, the time aspect often involved in these problems is neglected. For instance, in the case of life insurances, customers are more likely to churn after a year than after a period of 5 years. We propose an approach which makes it possible to incorporate this time aspect. In order to do so, a collection of statistical methods called survival analysis is used. Survival analysis is predominantly used in medical sciences to examine the influence of explanatory variables on the length of survival of patients, hence the name survival analysis. Survival analysis has different synonyms depending on its application.

For example, survival analysis is labelled ‘reliability analysis’ in engineering and is known under the term ‘duration analysis’ in economics. In addition, survival analysis applied in the field of data mining is also known as ‘survival data mining’.

1.2.2 Prepaid Versus Post-Paid

In this research paper, we deal with a different type of clients that is prepaid. In my opinion, modelling prepaid is far more challenging.

Post-paid customers are customers who are bound by a contract. In contrast to post paid customers, prepaid customers are far more difficult to deal with when it comes to churn prediction. In this research paper, we assume that prepaid clients do not sign any contract and are anonymous. So we do not have any personal data about them. All that we know is their tariff and usage. Prepaid customers does not pay the monthly subscription fee and their usage is less regular.

Secondly, post-paid customers are obligated to provide personal information like their name, gender and address. As a consequence, a post-paid phone number is linked to a single customer. Prepaid customers, on the other hand, do not necessarily need to provide this personal information and can therefore stay anonymous. Because of this, sim-cards, which contain the phone number of the customer, can easily be switched between individuals without having to inform the operator. Different calling behaviours are then registered on a single SIM-card and thus on a single phone number. Such numbers are misleading if we assume that a phone number is linked to a single and unique customer.

Finally, in most cases post-paid contracts have a fixed length of time say a year or two. Since people are likely to churn when their contract expires, the expiration date can be held as a very reliable predictor for churning. Unfortunately, this predictor is unavailable for prepaid churn.

We also do not have a strict definition of churn. Of course, there is a term called the expiration of the SIM card, but in my opinion, it is not a good definition of churn. This problem will be discussed in subsequent subsections.

1.3 Problem Statement

With a subscriber base of approximately 19.9 million representing 80.5% of the Ghanaian population as at August, 2011, the nation's mobile telephony industry is moving into a state of saturation. The National Communications Authority (NCA) recently implemented the Mobile Number Portability (MNP) which has added terribly to the woes of Service Providers. As a result of these developments, Service providers are gradually shifting focus from building a large customer base to retaining already existing clients and satisfy their needs.

Most of customers switch service provider as a result of lack of promotions on the network, network reliability and coverage (inability to make or receive calls) and expensive call rates.

1.4 Objectives of Study

With current developments in the telecom industry, the Service Provider is looking for a means to predict which customers are churning and which is not. The objective of this study therefore is

(i) To propose an approach to assist telecom churn management.

(ii) Use Kaplan Meier estimator to estimate survival probabilities for each of the value bands High Value Customer (HVC), Medium Value Customer (MVC) and Low Value Customer (LVC)

(iii) To test whether there is differences in the survival curves of the three value bands.

(iv) To model the effect of usage (TotalMOU, GPRSUsage, TotalSMS and Top upAmount) on the risk of churning using Cox Proportional Hazard Model.

1.5 Significanceof Study

As the estimation of marketing researchers, the average churn of mobile telecom industry is 2.2% per month. That is, about 27% of given carrier’s customers are lost each year. Thus, it is more essential to develop effective method to retain the existing customers for these companies, because the cost of acquiring a new customer is usually much higher than that of retaining an existing one**.** We know the problem confronting mobile telecommunications‟ management is that it is very difficult to determine which subscribers leave the company and why. It is therefore even more difficult to predict which customers are likely to leave the company, and more difficult still to devise cost-effective incentives that will convince likely “churners” to stay. Therefore, from the above description, we know the churn management is most critical issue in any telecom firm. Based on above mentioned, this research try to use Cox PH to fit a model and

KM estimator to estimator the survivorship of customers to prevent the customers turnover, further to promote their competitive advantage.

1.6 Methodology

In this study, two different predictive churn models are considered. The first model is extended Cox model which is based on the theory of survival analysis, whereas the second model, a decision tree, is commonly used in data mining.

In the present study, the focus will be on the extended cox model. However, in order to establish a direct comparison, we decided to consider the decision tree as well. Both models are ultimately tested on a selection of prepaid customers from the database provided by Vodafone. Models such as neural networks, logistic regression and decision trees are common choices of data miners to tackle this churn prediction problem. The goal is to distinguish churners from non-churners as much as possible. When new customers are offered, the model attempts to predict to which class each customer belongs.

1.7 Justification of Study

With the current trend in the Ghanaian mobile Telecommunication industry i.e. the saturation of the market as well as with introduction of the mobile number portability, the industry is faced with the problem of "churning".

Thus, it is more essential to develop effective method to retain the existing customers for these companies, because the cost of acquiring a new customer is usually much higher than that of retaining an existing one. We know the problem confronting mobile telecommunications management is that it is very difficult to determine which subscribers leave the company and why. It is therefore even more difficult to predict which customers are likely to leave the company, and more difficult still to devise cost-effective incentives that will convince likely "churners" to stay. Therefore, from the above description, we know the churn management is most critical issue in any telecom firm. Specifically, when CRM (customer relationship management) have been discussing more and more popularly in service industries, then how to manage customer churn is relatively important for telecom industry.

1.8 Limitation of Study

(i) Data was extracted from data warehouse of Vodafone Ghana for this study with sample around 15,000 subscribers for date of activation to date.

(ii) Time aspect involved in these problems is neglected.

1.9 Organisation of Study

Chapter one of the study deals with the introduction into the mobile communication industry in Ghana, the key players of the industry as well as discuss the current statistics in the market.

Chapter two talks about the concept of churn prediction and also discuss extensively the subject of survival analysis

Chapter three of the study discusses the methodology used in the study. It examines the mathematical treatment and presentation models, variant, formulation etc.

Chapter four basically deals with the data collection, analysis, as well as the computational algorithm used. It also discusses the results and significance of the results to the study.

Chapter five discuss the conclusion and summarizes the results obtained. Based on the finding, recommendations are also made.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Churn– In the telecommunications industry, the broad definition of churn is the action that a customer's telecommunications service is cancelled. This includes both service-provider initiated churn and customer initiated churn. An example of service-provider initiated churn is a customer's account being closed because of payment default. Customer initiated churn is more complicated and the reasons behind vary. In this study, only customer initiated churn is considered and it is defined by a series of cancel reason codes. Examples of reason codes are: unacceptable call quality, more favourable competitor's pricing plan, misinformation given by sales, customer expectation not met, billing problem, moving, and change in business among others.

Managing customer churn is of great concern to global telecommunications service companies and it is becoming a more serious problem as the market matures. Customer churn adversely affects these companies because they stand to lose a great deal of price premium, decreasing profit levels and a possible loss of referrals from continuing service customers (Reichheld&Sasser, 1990).

The review of literature summarizes a number of works in the field of churning, which have application to the telecom industry. This chapter consists of three individual sections. The first section aims at introducing Customer Relationship Management (CRM) and its basic concepts while it also tries to depict the contribution of machine learning techniques (especially data mining) to this realm. Section two is an introductory part to data mining and its significance role in CRM and ultimately the third section deals with the existing studies regarding the customer churn in different industries.

2.2 Customer Relationship Management: Basic Concepts

Eagerness toward Customer Relationship Management (CRM) began to grow in 1990 (Ling & Yen, 2001; Xu et al, 2002). A developed relationship with one's client can finally result in greater customer loyalty and retention and, also profitability (Ngai, 2005).

Despite the fact that CRM has become widely recognized, there is no comprehensive and universally accepted definition of CRM.

Swift (2001) defined CRM as an enterprise approach to understanding and influencing customer behaviour through meaningful communications in order to improve customer acquisition, customer retention, customer loyalty and customer profitability.

Koller and Keller (2006) have defined Customer relationship management (CRM) as the process of managing detailed information about individual customers and carefully managing all customer "touch points" to maximise customer loyalty.

Kincaid (2003) viewed CRM as "the strategic use of information, processes, technology, and people to manage the customer's relationship with your company (marketing, sales, services and support) across the whole customer life cycle"

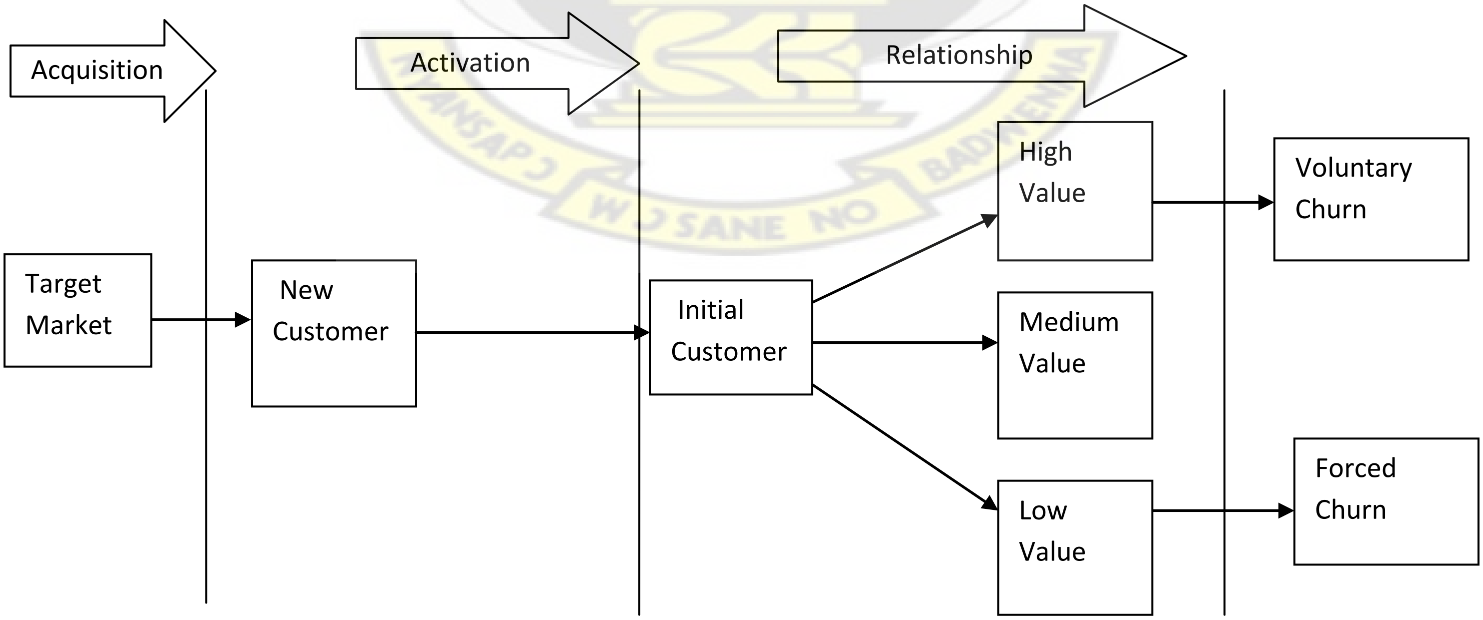
Bose (2002) viewed CRM as an integration of technologies and business processes used to satisfy the needs of a customer during any given interaction more specifically from his point of view Customer Relationship Management (CRM) involves acquisition, analysis and use of knowledge about customers in order to sell goods or services and to do it more efficiently.

Richards and Jones (2008) have defined CRM as "a set of business activities supported by both technology and processes that is directed by strategy and is designed to improve business performance in an area of customer management"

Having a glimpse to the above mentioned definitions of CRM one can understand that all the above authors' emphasis is on considering CRM as a "comprehensive strategy and process of acquiring, retaining and partnering with selective customers to create superior value for the company and the customer. It involves the integration of marketing, sales, customer service and supply-chain functions of the organisation to achieve greater efficiencies and effectiveness in delivering customer value." (Parvatiyar&Sheth, 2001).

Olafsson et al., (2008) believe that a valuable customer is usually dynamic and the relationship evolves and changes over time. Thus, a critical role of CRM is to understand this relationship. This is achievable by studying the customer life-cycle or customer lifetime, which refers to various stages of the relationship between customer and business.

Figure 2.1 depicts a typical customer life-cycle



As it is presented in the above figure, a prospect that responds to the marketing campaigns of the company in acquisition phase, becomes a customer and this “New Customer” becomes an established one once the relationship between him/her and the company has been established and this is the point that in which the company can benefit from its established customers by revenue that comes from cross – selling and up-selling, but the peril that threatens the company in this stage is that at some point established customers stop being customers (churn) (Olafsson et al., 2008). Thus, in simple words, the main goal of customer relationship management is to create satisfaction and delight among customers in order to prevent customer churn which is the most important threat that threatens all companies.

Van den Poel&Larivie’re, (2004) showed that a small change in retention rate can result in significant changes in contribution.

Ryals, (2005), classified CRM into two categories; attracting new customers what he referred to as offensive marketing and keeping existing customers known as defensive marketing. While acquiring a new customer is the first step for any business to start growing, the importance of retaining customers should not be over-looked.

Reinartz et al., (2005) showed that insufficient allocation to customer-retention efforts will have a greater impact on long-term customer profitability as compared to insufficient allocation to customer-acquisition efforts.

As Chu et al., (2007) have highlighted, the cost of acquiring a new customer is five (5) to ten (10) times greater than that of retaining existing subscribers. Even if we put aside the existing studies, which mentioned that it costs more to acquire new customers than to retain the existing customers, we can consider that customer retention is more important than customer acquisition because lack of information on new customers makes it difficult to select target customers and this will cause insufficient marketing efforts.

There exist different categorization approaches towards CRM (Teo et al., 2006; Ngai, 2005; He et al., 2004; Xu et al., 2002). From the architecture point of view, the CRM framework can be classified into operational and analytical (He et al., 2004; Teo et al., 2006). While operational CRM refers to the automation of business process, the analytical CRM refers to the analysis of the customer characteristics and attitudes in order to support the organisation's management strategies. Thus, it can help the company in more effective allocation of its resources (Ngai et al., 2009).

On the other hand Kincaid (2003), West (2001), Xu et al., (2002), and Ngai (2005) believe that CRM falls in the four following categories:

(i) Marketing

(ii) Sales

(iii) Service and support

(iv) IT and IS

According to what experts believe, the role of Information Technology (IT) and Information Systems (IS) in CRM can't be denied (Kincaid, 2003; Ling &Yen, 2001). Using IT and IS will make the companies capable of the collection of the necessary data to determine the economics of customer acquisition, retention and life-time value. This means involving the use of database, data warehouse and data mining to help organisations increase their customer retention rates and their own profitability.

2. 3 Data Mining and Its Application in CRM

Nowadays lack of data is no longer a problem, but the inability to extract useful information from data is (Lee &Siau, 2001). Due to the constant increase in the amount of data efficiently operable to managers and policy makers through high speed computers and rapid data communication, there has grown and will continue to grow a greater dependency on statistical methods as a means of extracting useful information from the abundant data sources. Statistical methods provide an organised and structured way of looking at and thinking about disorganised, unstructured appearing phenomena.

In fact the accelerated growth in data and databases resulted in the need of developing new techniques and tools to transform data into useful information and knowledge, intelligently and automatically. Thus data mining has become an area of research with an increasing importance (Weiss and Indurkhya, 1998 cited by Lee &Siau, 2001). Data mining techniques are the result of a long term research and product development and their origin have roots in the first storage of the on computers, which was followed by improvement in data access (Rygielski et al., 2002).

SAS Institute, (2000) defined data mining as "the process of selecting, exploring and modelling large amount of data to uncover previously unknown data patterns for business advantage."

Berry &Linoff, (2004) explained data mining as the "exploration and analysis of large quantities of data in order to discover meaningful patterns and rules"

Shaw et al., (2001) also said data mining involves selecting, exploring and modelling large amounts of data to uncover previously unknown patterns and finally comprehensible information from large databases.

Rygielski et al., (2002) noted that what data mining tools does is to take data and construct a model as a representation of reality. The authors also noted that the resulted model describes patterns and relationships present in the data.

According to Ngai, (2005), the broad application of data mining falls in two major categories namely;

(i) Descriptive Data Mining which aims at increasing the understanding of the data and their content.

(ii) Predictive or perspective data mining which aims at forecasting and devising at orienting the decision process.

Aiming at solving business problems, data mining can be used to build the following types of models (Ngai et al., 2009)

(i) Classification

(ii) Regression

(iii)Forecasting

(iv)Clustering

(v) Association analysis

(vi)Sequence discovery

(vii) Visualization

Among the above mentioned models the first three one are prediction tools while association analysis and sequence discovery are used for description and clustering is applicable to either prediction or description.

The applications of data mining ranges from evaluation of overall store performance, promotions' contribution to sales and determination of cross-selling strategies to segmentation of the customer base (Gomory et al., 1999).

Ngai et al., (2009) noted that the application of data mining tools in CRM is an emerging trend in global economy. Since most companies try to analyse and understand their customers' behaviour and characteristics, for developing a competitive CRM strategy, data mining tools has become of high popularity.

Beside the aforementioned roles for data mining in marketing, Rygielski et al., (2002) have identified a wide continuum of applications for data mining in marketing in different industries from retailing to banking and telecommunications industry.

According to Rygielski, Wang and Yen (2002) in retailing, data mining can be used to perform basket analysis, sales forecasting, database marketing and merchandise planning allocation. Besides, data mining-based CRM in banking industry can be utilised in card marketing, Cardholder pricing and profitability, fraud detection and predictive life-cycle management. In addition to the above mentioned realms, data mining possesses a significant role in telecommunications industry. To be more specific, using data mining, companies would be able to analyse call detail records and identify customer segments with similar use patterns and develop attractive pricing and feature promotions. Furthermore, data mining enables companies to identify the characteristics of customers who are likely to remain loyal and also determine the churners.

2.4 The Context of Churning and its Complexity

Chander et al., (2006) defined customer churn as "the propensity of customers to cease doing business with a company in a given time period"

Companies aim at getting more and more customers. Nevertheless, the ratio (new customers/ churners) tends towards one over time. The impact of churn becomes then markedly more sensitive (Lejeune, 2001).

Neslin et al., (2006) noted customer churn has become the main concern for firms in all industries and companies, regardless of the industry that they are active in, are dealing with the issue. Customer churn can blemish a company by reducing profit levels, losing a great deal of price premium, and losing referrals from continuing service customers (Reichheld&Sasser, 1990).

A research by Reichheld (1996) revealed that an increase of 5% in customer retention rate can increase the average net present value of customer by 35% for software companies and 95% for advertising agencies.

Considering the churn rate of different industries, one can find that the telecommunications industry is one of the main targets of this hazard such that the churn rate in this industry ranges from 20 to 40 annually (Berson et al., 1999;Madden et al., 1999)

According to Lejeune (2001) the concept of churn is often correlated with the industry life-cycle. When the industry is in the growth phase of its life-cycle, sales increases exponentially; the numbers of new customers largely exceeds the number of churners, but for products in the maturity phase of their life-cycle, companies put the focus on the churn rate reduction.

Customer churn figures directly how long a customer stays with a company and in turn, the customer's lifetime value (CLV) to that company (Neslin et al., 2006), which is the sum of the revenues gained from company's customers over the lifetime of transactions after the deduction of the total cost of attracting, selling and servicing customers, taking into account the time value of the money (Hwang et al., 2004)

Previous researches have examined the concepts of customer churn from different point of view. According to Olafsson et al., (2008), there are two different types of churn. The first is voluntary churn, which means that established customers choose to stop being customers. The other type is forced churn, which refers to those established customers who no longer are good customers and the company cancels the relationship.

Burez and Van den Poel (2008) have divided the voluntary churners into two groups; commercial churners and financial churners. According to their research customers who voluntary leave the company can be divided into two groups : those who do not renew their fixed term contract at the end of that contract and others who just stop paying during their contract to which they are legally bound.

Berson et al., noted customer churn is a term used in the mobile telecom industry to denote the movement of mobile customers from one provider to another. Churn management is to describe the process of ensuring that profitable customers stay with a particular company. However, customer churn is a serve problem for mobile telecom operators, in the meantime to increase its services base while controlling customer churn is essential to survival and growth of a mobile telecom company.

On the other hand, Kentritas (n.d.) thought that the term Churn Management in the telecommunications market is used to describe the procedure of securing the most important customers for a company. The proper customer management presupposes the ability to forecast the customer’s decision to move from one provider to another and the reception of proper measures in order to avoid the movement and retain their customers.

Wei et al., SAS institute investigated that Churn costs European and U.S. telecom close to US$4 billion each year, and the global cost of customer defection may well approach a staggering US$10 billion each year. Annual churn rates of 25 to 30 percent are the norm, and carriers at the upper end of this spectrum will get no return on investment on new subscribers. Why? Because it typically takes three years to recover the cost (approximately US$400 in the United States and US$700 in Europe) of replacing each lost customer with a new one (customer acquisition). Particularly, in the European and Asian markets, the number of new market entrants is adding to the churn phenomenon. In Europe, 30 new telecoms entered the market in 1998, seeking the 15 percent marketing share that analysts say they will need to survive. The growth in the number of subscribers has eased this situation in the past, but as market growth slows and average revenue-per-user declines, increased competition is likely. How serious is the churn problem in telecom industry?

(Thearling, K.1999), said Churn is expensive, not only is churn an inevitable part of doing business in mobile, it is also proving to be an extremely expensive experience. Churn has many consequences, and most of them carry a large price tag. The biggest consequences of churn are the loss of revenue, since lost customer is equal to lost revenue.

Thearling proposed that Data Mining is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their huge database. Data mining tools can answer business questions that traditionally were too time consuming to solve.

Miguel A.P.M. Lejeune (2001) addresses that Data Mining Techniques allow the transformation of raw data into business knowledge. The SAS institute defines data mining as “the process of selecting, exploring and modelling large amount of data to uncover previously unknown data patterns for business advantage”. Consequently, we would say that data mining is applying data analysis and discovery algorithms to enumerate patterns over data for prediction and description. Recently, data mining has been generally applied to many industries on solving business problems.

The empirical results shows that the response rate of the target marketing strategy supported by the proposed approach was more than three times that of the mass marketing strategy. Anand et al., have identified the cross-sales problem and analyzed this problem from the view of solving it using data mining techniques. The data mining solution is based on the discovery of characteristic rules, generalized to allow uncertainty in the rules, from the set of positive examples. The discussion has been applied to a real world data mining application in the financial sectors and results of its application are also provided. Moreover, the author mentioned that the methodology was developed to provide a data mining solution to the cross-sales problem and resulted in useful knowledge being discovered from the customer database, and it has great potential for exploitation within other service sectors.

In recent years, CRM has been discussing extensively, especially in customer-oriented industries. They emphasize CRM more than other aspects. Hence, from the above references, churn management seems much important to telecom's CRM, especially in the case of mobile market in Ghana moving into a state of saturation.

Therefore, how to keep existing customers is quite important for these mobile providers. After finding out the churn's predictive model, marketing department further needs to segment their customers, score customers, and to identify the profitable customers, retain the profitable ones who are likely to leave that has been predicted by using the predictive model. For all churn predictive models, customer segmentation and scoring can also be completed to use the data mining techniques.

Anand et al., mentioned data mining techniques can be applied to discover the customer's information for other service industries. Wei et al., also addressed data mining techniques can be developed for other telecommunication application. Consequently, this study is referring to these researches as a benchmark to develop similar approach to support telecom churn management. This research intends to base on these factors and concepts that addressed by above researchers to find a best predictive model for telecom churn management using data mining techniques.

In a highly competitive and maturing mobile telecommunications service market, a defensive marketing strategy is becoming more important. Instead of attempting to entice new customers or lure subscribers away from competitors, defensive marketing is concerned with reducing customer exit and brand switching (Fornell&Wernerfelt, 1987).

Reichheld (1996) estimated that, with an increase in customer retention rates of just 5%, the average net present value of a customer increases by 35% for software companies and 95% for advertising agencies. Therefore, in order to be successful in the maturing market, the strategic focus of a company ought to shift from acquiring customers to retaining customers by reducing customer churn.

In order to better manage customer churn, companies need to fully understand a customer's behavioural churn path and the factors pertaining to the customer churn; however, these problems have not been fully addressed in the literature.

First, previous studies mainly focused on finding a few specific factors (e.g., customer dissatisfaction, customer loyalty, etc.) pertaining to customer churn rather than investigating and empirically testing a comprehensive model encompassing relationships among various constructs, such as customer dissatisfaction, switching costs, service usage and other customer-related variables. For example, Keaveney (1995) only examined why customers switch their services and classified the reasons into eight general categories.

Bolton (1998) investigated the role of customer satisfaction in a dynamic model estimating the customer's duration with the service carrier. Bolton et al., (2000) found that members in loyalty reward programs overlook a negative evaluation of the company vis-à-vis its competitors in their re-patronage decisions.

Gerpott et al., (2001) analyzed a two-stage model where overall customer satisfaction has a significant impact on customer loyalty, which in turn influences customers' intentions to terminate their contractual relationship.

Lee et al., (2003) found some determinants of customer churn in the Korean broadband Internet access service market.

Kim et al., (2004) investigated the adjustment effect of switching barriers on customer satisfaction and customer loyalty.

Secondly, due to the proprietary nature of actual customer data, much of the research has dealt with consumer survey data asking consumers' perceptions of service experiences and intention to remain. However, the survey data rather than the actual customer transaction or billing data may not fully represent the customer's actual future re-patronage decision. Furthermore, due to cost concerns, most survey-based studies use a small sample of less than a thousand customer records (Keaveney, 1995; Bolton et al., 2000; Gerpott

et al., 2001; Lee et al., 2003; Kim et al., 2004), which may undermine the reliability and validity of analysis results.

In fact, there are several studies that are based on large-scale actual customer transaction and billing data.

However, their objectives mainly focus on predictive accuracy rather than descriptive explanation. For example, detailed call data is used to predict the probability of customer churn (Mozer, Wolniewicz, Grimes,Johnson, &Kaushansky, 2000; Ng and Liu, 2000; Wei and Chiu, 2002); a subscriber’s remaining tenure with the company is estimated using internal company databases (Drew, Mani, Betz, &Datta, 2001) and brand switching and adoption probabilities are forecasted using commercial databases (Weerahandi& Moitra,1995).

However, there are at least two exceptions, both of which use survival analysis to test hypotheses about churn predictors. One is a study by Bolton (1998) where actual customer transaction and survey data is used to analyze customer churn behaviour in the cellular service market.

Another is a customer attribution study where Poel and Lariviere (2004) analyzed an in-house data warehouse in the European financial service market. Compared with the previous studies, this paper has two distinct research objectives. The first objective is to develop a comprehensive churn model and empirically test it using a large sample of actual customer transaction and billing data, which is directly related to actual customer churn decisions. Identifying customer churn determinants, such as core service failures, customer complaints, loyalty programs, service usage, etc., may help managers improve company operations in terms of their marketing strategy, specifically customer churn prevention programs.

The churn rate in the South Korean mobile telecommunications service market was 16.9% in 2003, which was comparable with the churn rate in the USA (Parks Associates, 2003). However, the introduction of mobile number portability in 2004 has allowed mobile subscribers to switch operators without having to change numbers and thus the customer churn rate increased to 20.0% by the end of 2004.

Berson et al., noted that the annual churn rate ranges from 20% to 40% in most of the global mobile telecommunications service companies.(Reichheld&Sasser, 1990) said Customer churn adversely affects companies because they stand to lose a great deal of price premium, decreasing profit levels and a possible loss of referrals from continuing service customers

(Siber, 1997), found out that the cost of acquiring a new customer can substantially exceed the cost of retaining an existing customer.

Previous research suggests that network quality and call quality are key drivers of customer satisfaction/ dissatisfaction in the mobile communications services market (Gerpott et al., 2001; Lee et al., 2001; Kim & Yoon, 2004; Kim et al., 2004).

Keaveney's (1995) critical incident study of 835 customer-switching behaviours in service industries demonstrated that 44% switched their service providers because of core service failures. In addition, service failures have been ''triggers'' that accelerate a customer's decision to discontinue the service provider-customer relationship (Bolton, 1998; Bolton et al., 2000; Kim, 2000; Mozer et al., 2000). Therefore, some technical dimensions of service quality, such as the amount of call drops and call failures may be related to customer churn.

H1a: call drop rates are positively associated with the customer churn probability.

H1b: call failure rates are positively associated with the customer churn probability.

Though many factors are related to a customer's complaint behaviour, dissatisfaction with a service or a product is found to be positively related to the complainant's behaviour (Day & Landon, 1977; Bearden & Teel, 1983).

In fact, complaining customers can take private and/or public actions. For example, personal actions include no further purchase and negative word-of-mouth; public actions include redress-seeking efforts toward the provider/seller and appeals to third-party consumer affairs bodies (Day & Landon, 1977). That is, complaints lead to switching or customers churn.

On the other hand, Fornell and Wernerfelt (1987) argued that well-managed complaint management programs can lower the total marketing expenditure by substantially reducing customer churn. However, it should be noted that this is true only if a sufficiently large proportion of the complainants can be persuaded to remain as customers by the complaint management programs. If the customer complaint management programs in major service providers are not effective, customer complaints may lead to customer churn.

In a study of 306 subscribers in the Korean mobile service market, Kim et al., (2004)found that among factors constituting switching costs, loss of loyalty points has both a direct effect and an adjustment effect on customer loyalty. Because current loyalty points are forfeited as they switch, even dissatisfied customers may show a high level of ‘‘false’’ loyalty (Gerpott et al., 2001).

Bolton et al. (2000)argued, members in loyalty reward programs may overlook or discount negative evaluations of the company against competitors in terms of product, quality and price. Therefore such membership card programs establish switching costs for customers and thus they may inhibit customer churn

Mozer et al., (2000)conjectured that monthly charges and usage amounts are linked to churn. However, it is still unclear as to whether the relationship between service usage and customer churn is truly positive or negative.

Kim and Yoon, (2004) investigated the underlying elements of customer churn in mobile telecommunications service providers. From their study, they found that attrition of customers in this industry depends on the levels of satisfaction with alternative specific service attributes including call quality, tariff level, handset type, brand image as well as income and subscription duration, but only factors such as call quality, handset type and brand image affects customer loyalty as has been measured by the positive word of mouth in the form of recommendation. In other words, according to Kim and Yoon (2004) determinants of churn clearly differ from those of loyalty and in order to decrease the churn rate in the telecom industry, the companies are supposed to focus on boosting customer satisfaction levels rather than loyalty.

Gerpott et al., (2001) believe that retention, loyalty and satisfaction of customers in telecom industry are causally inter-correlated and that service price, perceived benefits and also lack of number portability have strong effects on customer retention. They investigated the influential factors on bringing superior economic success for telecommunication network operators in German markets and tested the hypotheses suggesting that Customer Retention (CR), Customer Loyalty (CL) and Customer Satisfaction (CS) should be treated as differential constructs which are causally inter-linked. The results showed that the overall CS has a significant positive impact on CL which in turn influences a customer's intention to terminate / extend the contractual relationship (CR). It also revealed that mobile service price and personal service benefit perceptions as well as lack of number portability between various cellular operators' perceived customer care performance had no considerable effect on CR.

Ahn et al., (2006) conducted an exploratory research in which they aimed at finding the most influential factors on customer churn. In their research they considered a mediator factor named "Customer's Status", between churn determiners and Customers Churn in their model and they mentioned that "Customer Status" (from active use to non-use or suspended) change is an early signal of total customer churn.

Furthermore, Seo et al., (2008) investigated about retention factors in telecommunications industry by examining other features and variables. The study focused on understanding the factors related to customer retention behaviour i.e. both behavioural factors such as switching costs and customer satisfaction and demographic factors and its two goals were to understand

(i) How factors that affect switching cost and customer satisfaction, such as length of association, service plan complexity, handset sophistication and the quality of connectivity, drive customer retention behaviour and

(ii) How customer demographics such as age and gender affect their choice of service plan complexity and handset sophistication, lending to differences in customer retention behaviour.

The methodologies they used were a binary logistic regression model and a two – level hierarchical linear model. The factors analysed consisted of: complexity of service plan, handset sophistication, length of association and connectivity.

The results show that:

(i) The more complex service plan, the more sophisticated handset, longer customer association, higher connectivity quality of wireless is positively related to customer retention behaviour.

(ii) Different age and gender groups revealed differences in wireless connectivity quality and service plan complexity, affecting their customer retention behaviour, while they did not experience differences in terms of length of customer association and headset sophistication.

Based on previous studies on churn prediction, Wei and Chiu (2002) developed a new model for customer churn prediction in telecommunication service providers by using data mining techniques. In that time past researches on churn prediction in the telecom industry mainly had employed classification analysis techniques for the construction of churn prediction models and they had used user demographics, contractual data, customer service logs and call patterns extracted from call details (e.g. average call duration, number of outgoing calls, etc), but Wei and Chiu believed that existing churn prediction models had several disadvantages. They listed the disadvantages in two groups; first, use of customer demographics in churn prediction renders the resulting churn analysis at the customer rather than contract (or subscriber) level. In other words, tendency of each customer towards churning was calculated on a per-customer basis rather than contract basis. It is quite common that a customer concurrently holds several mobile service contracts with particular carrier, with some contracts more likely to be churned than others. In this regard, customer-level-based churn prediction is considered inappropriate. Secondly, information on some of the input variables (features) was not readily available and this unavailability of customer profiles, had limited the applicability of existing churn-prediction systems.

In response to the described limitations of existing churn-prediction systems in that time, Wei and Chiu exploited the use of call pattern changes and contractual data for developing a churn-prediction technique that identifies potential churners at the contract level. They claimed that subscribers' churn is not an instantaneous occurrence that leaves no trace. Before an existing subscriber churn, his/her call patterns might change (e.g. the number of outgoing calls gradually gets reduced). In other words, changes in call patterns are likely to include warning signals pointing towards churning. Such call patterns pattern changes can be extracted from subscribers' call details and are valuable for constructing a churn prediction model based on a classification analysis technique. In their investigation, they used two types of available data: contractual data including length of services, payment type, contract type and call details such as Minutes of Use (MOU), Frequency of Use (FOU) and Sphere of Influence (SOI: refers to the total number of distinct receivers contracted by the subscriber over a specific period) in order to develop a churn prediction technique.

Using the data set Wei and Chiu (2002) randomly selected a prediction period (P) in order to generate an evaluation data set and also determine the churn status. According to them, churn status of a subscriber was the connected or disconnected status of the subscriber within the prediction period P, and subscribers who disconnected his/her mobile service during P were considered as churners while the ones who disconnected the service before P were not included in their evaluation data set. Furthermore, subscribers who were still connected to the service provider at the end of P were classified as non-churners.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This study will base on this framework and apply telecom's existing transaction data that is stored in data warehouse to find a churn predictive model. In this research, data mining techniques is selected to build predictive models and segment customer. We will introduce all procedures and related techniques that is used by this research.

(i) Data Collection

(ii) Exploring data analysis (benchmark, reference review)

(iii)Customer Experience Management

The commercial data mining tool Cordiant Predictive Analytics Director was used for automated data preparation (attribute binning, grouping, recoding, selection) and modelling (logistic regression, decision trees).

In telecom industry, the size of database usually is very large. It is because there are so many legacy systems are operating to record all kinds of customers data, such like Billing system, POS (point of sale) system, VAS (value added service system), pre-paid system, provision system and MSC (Mobile Switch Center), etc. Among of these, the stored CDRs (call detail records) database is the biggest one, hence how to collect data and integrate these systems need IT technology to solve it. Building a data warehouse of data mart is the solution that has now been widely using in India telecom industry. Due to time issue and limitation of data obtained, in this research we only focus on using Cox model data mining techniques to build churn predictive model to support telecom churn management issue. As for others, they would also be other research subjects if someone where interested in.

3.2 Operational Churn Definition

The preferred definition of churn is one which is able to indicate when the customer has permanently stopped using his/ her SIM- card as early as possible.

In this sub-section, such a churn definition is proposed. This is called the operational churn definition. An operational churn definition is necessary since the proposed models in this study are supervised models. Supervised models require a labeled dataset for training purposes.

The most obvious operational definition of prepaid churn is based on a number of successive months with zero usage. After all, this definition bares great similarity with our intended churn definition. For instance, a churned customer could be defined as not having had any incoming and outgoing calls in the past three months. Although this is an intuitive operational definition, we do have to take the following considerations into account. First, the definition of zero usage has to be refined. Incoming calls are also registered when the user does not answer his phone. Besides, incoming duration is registered when a message is recorded on voicemail. This means that a prepaid customer, who has already churned, still can have incoming calls and incoming duration from uninformed outsiders. Therefore, defining zero usage as zero outgoing duration plus zero incoming duration cannot be considered optimal. With regard to incoming duration, we manage a threshold. This threshold is set to the value of 30 seconds per month which turned out to be a practical value. In sum, zero usage can be defined as a total incoming duration of less than 30 seconds per month in combination with zero outgoing duration.

Secondly, the operational churn definition has to be refined. Most prepaid customers are irregular users that often display zero usage mainly because under our scope of study majority of prepaid customers are multi ‘SIMERS’. In addition, several successive months of zero usage often occurs as well. As a consequence, the above mentioned churn definition will not hold for every customer. Low usage customers could be quickly labeled as churned without actually being churned. To overcome this problem, a flexible churn definition is proposed.

This definition allows for a larger range of customer behaviours. The definition consists of two parts, α and β, where α is a fixed value and β is equal to the maximum number of successive months with zero usage. $\alpha + \beta$ is then used as a threshold to distinguish churn behaviour from non-churn behaviour. An example is provided in Figure 3.1.

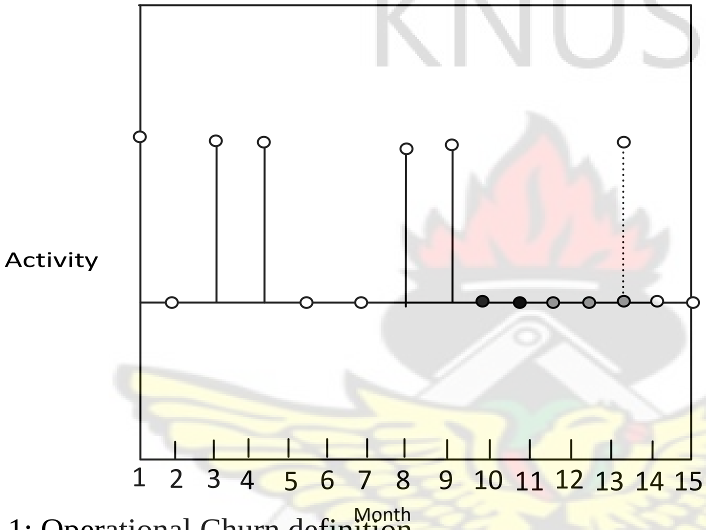


Figure 3.1: Operational Churn definition

In this example α is set to a value of 3 and is marked in grey. β is equal to2 and is marked in black. At month 13 the customer is labeled as churned since there are (2 + 3 = 5) successive months with zero usage. The churn definition can be interpreted as follows. Although the customer of Figure 3.1 did not show any usage during the fifth and sixth month, at month 7 he 'proved' that he was not churned. This means that another two successive months of zero usage would not guarantee any churn behaviour.

This is indicated with the variable β. The constant α is added to indicate a point where the customer would churn. Two different values for α are used to label the training set, that is α = 2 and α = 3. From now on these will be referred to as churn definition 1 and 2 respectively.

Customers with $\beta \geq 5$ are excluded, as the churn definition becomes less accurate and intuitive after longer durations.

3.3 Theory of Survival Analysis

Survival analysis is a collection of statistical methods which model time-to- event data. Central is the occurrence of a well-defined ‘event’. The variable of interest is the time until this event occurs. This is in contrast with approaches like regression methods and neural networks which model the probability of an event. Depending on its application, the event of interest can be the failure of a physical component or the time to death. In the context of data mining the event of interest is typically the time until churn or the time until the next purchase.

3.3.1 Survival and Hazard Functions

Let $T \geq 0$ be the random variable that denotes the time at which the event occurs and let it have density $f(t)$ and distribution function $F(t)$. The survival function is then defined as

$$S(t) = pr(T > t) = 1 - F(t) = \int_{t}^{\infty} f(t)dx \,. \tag{3.1}$$

It is important to note that $S(t)$ is a monotonic decreasing function $S(0) = 1$. The survival at time t is the probability that a subject will survive to that point in time. The hazard rate function $\lambda(t)$ is defined as

$$\lambda(t) = \frac{lim_{\triangle t \to 0} Pr(t < T < t + \triangle t | T > t)}{\triangle t} = \frac{f(t)}{1 - F(t)} \tag{3.2}$$

The hazard function is also known as the instantaneous failure rate, force of mortality and age specific failure rate. $\lambda(t)\Delta t$ can be interpreted as the instantaneous probability of having an event at time t given that one has survived (i.e. not had an event) up to time$t$. The functions $f(t)$, $F(t)$, $S(t)$and $\lambda(t)$give mathematically equivalent specifications of the distribution of T .

3.3.2 Recording Survival Data

Survival data is recorded in the following way. Subjects are observed for a certain period of time. During this period, the time of the event of interest is registered. However, it is possible that the event cannot be registered because it does not occur during the period of observation. This is called censoring and is discussed in the following sub-section.

Survival data requires both a well defined origin of time as well as a time scale. The origin of time is the moment from which the observation starts. This can be any particular event, for example birth, or in our case the commitment dates of a customer. The time scale is the frequency by which a subject is checked on the occurrence of the event. Common scales are based on years or months, depending on the nature of its application. In this study we apply a time scale based on months, since the data we use is monthly aggregated.

Let T_i denote the time at which the event occurs for the ithsubject and let T_i^* denote the observed time when dealing with censoring. Let σ_i be a 0/1 indicator which is 1 if T_i is observed and 0 if the observation is censored.

The pair (T_i^*, σ_i) is then used to describe a single observation.

3.3.3 Censoring

The concept of censoring entails an essential element of survival analysis.

Censoring occurs when there are incomplete observations. Three types of censoring can be distinguished:

(i) Right censoring

(ii) Left censoring

(iii)Interval censoring

These types mentioned are illustrated in Figure 3.2. The most common type of random censoring is right censoring. Right censoring occurs when the event time is larger than the period of observation. In that case, it is only known that the event time is larger than the last observed time. In the case of left censoring, the event occurs before the period of observation. Then, the event time is only known to be before a certain point in time. Finally, interval censoring occurs when the exact event times are not available, and are only considered to be in a certain interval. In the present study we are dealing with right censored data since not every customer churns during observation.

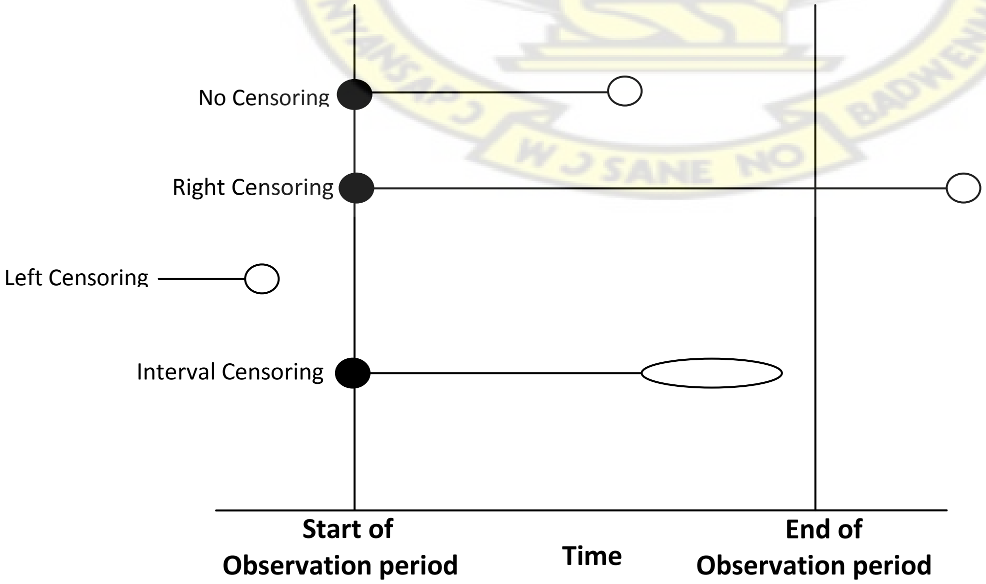


Figure 3.2: Illustration of the three different types of censoring

3.4 Survival Model as a Predictive Churn Model

In this sub-section, a survival model for churn prediction is proposed. There are many different types of survival models. In this approach, we are particularly interested in survival models that incorporate a regression component, since these regression models can be used to examine the influence of explanatory variables on the event time. In this context, such explanatory variables are often called covariates. There are two commonly used classes of regression models, namely so-called accelerated failure time models and proportional hazard models. Accelerated failure time models are based on a survival distribution. Common employed distributions are Weibull, exponential and log-logistic. In accelerated failure time models the regression component affects survival time by rescaling the time axis. This is illustrated in figure 3.3. The proportional hazard model is also known as the Cox model. It is the most popular survival regression model available since it does not make any assumptions on the survival function as opposed to accelerated failure time models. It was introduced by David Cox in 1972. In the Cox model, the regression component affects the hazard curve through multiplication. This is illustrated in figure 3.3. Many improvements and adjustments have been made to the Cox model since the introduction of the model. Because of these improvements, which will be discussed later on, the Cox model is a strong candidate for churn prediction.

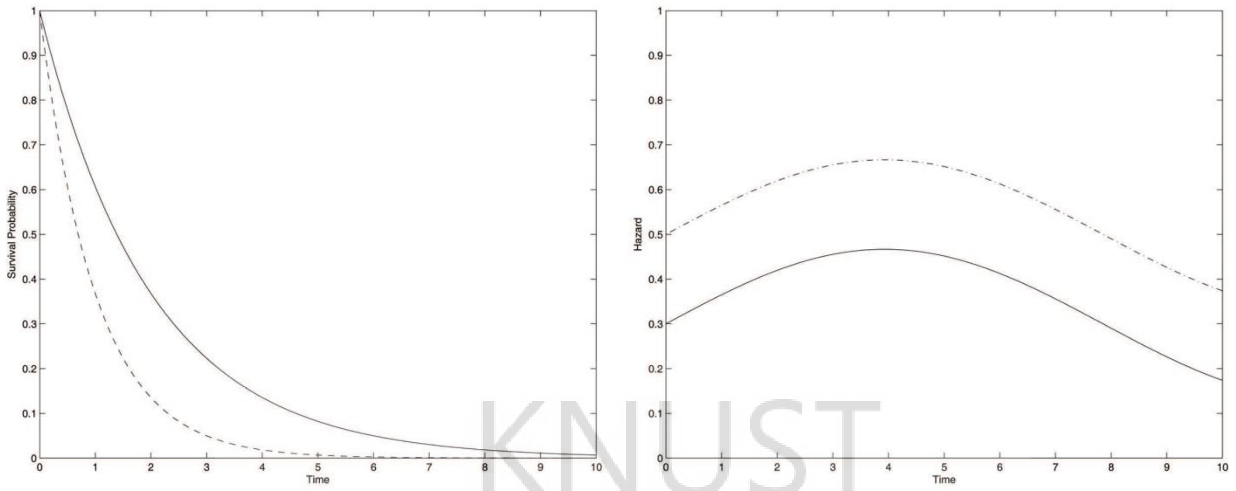


Figure 3.3: Left: The accelerated failure time model. Right: The proportional hazard time model.

3.4.1 Cox Model

Let X_{ij} denote the jth covariate of the ith subject, where $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$. One can think of as an $n \times p$ matrix where X_i denotes the covariate vector of subject i . The Cox model specifies the hazard rate as

$$\lambda_i(t) = \lambda_0(t)\ell^{X_i\beta}, \qquad (3.3)$$

where λ_0 is called the baseline hazard and β is a $p \times 1$ vector of regression coefficients. The latter is familiar from multivariate regression. The baseline hazard λ_0 is an unspecified nonnegative function over time which can be interpreted as the average hazard at time t. In contrast to accelerated failure time model, the baseline hazard does not have to follow a particular distribution. This is a significant advantage over parametric survival models which are restricted by the shape of a particular distribution. The hazard rate at time t is thus the product of a scalar, $e^{X_i\beta}$ and the baseline hazard at time t . To put it differently, the covariates increase or decrease the hazard function by a constant relative to the baseline hazard function. The Cox model is also referred to as the proportional hazard model since the hazard ratio for two subjects with covariate vector X_i and X_j , given by

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)e^{X_i\beta}}{\lambda_0(t)e^{X_j\beta}} = \frac{e^{X_i\beta}}{e^{X_j\beta}}, \qquad (3.4)$$

is constant over time. This implies that the covariates must have the same effect on the hazard at any point in time.

Estimation of parameter β is based on the partial likelihood function and is performed without specifying or even estimating the baseline hazard. The term partial likelihood is used because the likelihood formula considers probabilities only for those subjects who have an event, and does not explicitly consider probabilities for those subjects who are censored**.** The likelihood can be written as a product of several likelihoods, one for each event time.

The likelihood at time j denotes the likelihood of having an event at time j, given survival up to time j. The set of individuals at risk is called the risk set and is denoted by $R(j)$. The partial likelihood is then given by

$$L(\beta) = \prod_i \frac{e^{X_i\beta}}{\sum_{j\in R(i)} e^{X_j\beta}}, \qquad (3.5)$$

This equation holds as long as there are no tied event times, or ties in short. Ties are events that occur exactly at the same time. Adjustments to equation3.5 are necessary to deal with this. This will not be discussed here, but an extensive explanation can be found here.

3.4.2 Extended Cox Model

The Cox model, as discussed in the previous sub-section, is not suitable as a predictive model for prepaid churn. The main reason is that the explanatory variables, or covariates, are fixed over time. This implies that we can only use explanatory variables that are time-independent, such as address or sex.

Since we do not have these kind of variables to our disposal (due to reasons explained in Chapter 2) and, more importantly, because we want to use variables that capture the (current) behaviour of customers, the use of time-independent covariates is far from optimal.

Fortunately, several improvements and adjustments have been made to the Cox model in the last several years. With regard to the present work, we are particularly interested to the improvements made by Gill and Anderson**.**This adapted model is often referred to as the extended Cox model.

The extended Cox model includes the ability to accommodate left-censored data, time-varying covariates, and multiple events**.** The ability to include time-varying covariates makes it possible to use covariates that capture the behaviour of customers much better. Another advantage is that the proportionality assumption does not have to hold in this case, since the covariates are dependent on time. Equation3.3 is slightly adjusted to let Xi depend on time. The extended Cox model can be defined as

$$\lambda_i(t) = \lambda_0(t)\exp\left[\sum_{i=1}^{p1}\beta X_i + \sum_{j=1}^{p2}\beta X_j(t)\right]. \qquad (3.6)$$

Notice that covariates are now split in p1 time-independent covariates and p2 time-dependent covariates. In order to include time-varying covariates in the Cox model, a counting process formulation is required. A counting process is a stochastic process starting at 0 and whose sample paths are right continuous step functions with height 1**.** Recall that the pair (T_i^*, ∂_i) is normally used to represent an observation. The counting process formulation replaces this with $(N_i(t), Y_i(t))$ where $N_i(t)$ the number of observed events in [0, t] for subject i, whereas $Y_i(t)$ is 1 if subject i is under observation and at risk at time t and 0 otherwise. Right censored data is a special case of this formulation. Two examples are shown in figure 3.4. These correspond to (T^{*}, ∂) pairs (3, 0) and (4, 1). The counting process formulation makes it immediately possible to include multiple event times and multiple at-risk intervals. The application of this new formulation is not restricted to survival analysis anymore and can be used to model non-stationary Poisson processes, Markov processes and other processes**.**

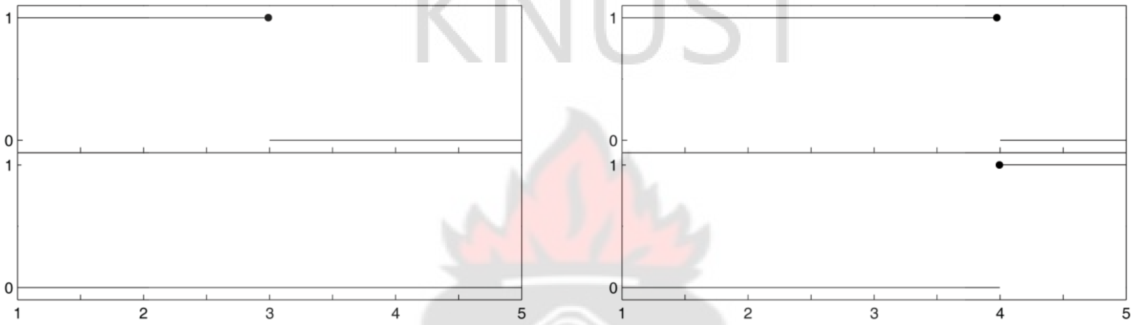


Figure 3.4: The counting process formulation. Left: Right censoring. Right: No censoring.

The extended Cox model is already implemented in the statistical modeling language R.

R features many additional packages and a large on-line community. The ‘event history analysis’ package is used here for modeling.

3.4.3Kaplan Meier Estimator

Let $S(t)$ denote the survival function for some population. Let $t_1 < t_2 < \ldots < t_m$denote the distinct ordered times of churning (not counting censoring times). Let $d_j$be the number of customers who churn at time t_j and $d_j$be the number of customers at risk at timet_j. The Kaplan

$$S(t) = \prod_{i;j<t}\left(1 - \frac{d_j}{r_j}\right) \qquad (3.7)$$

A heuristic justification of the estimate is as follows. To survive to time t you must first survive to $t(1)$. You must then survive from $t(1)$ to $t(2)$ given that you have already survived to$t(1)$. And so on. Because there are no churn between $t(i-1)$and $t(i)$, we take the probability of churning between these times to be zero.

3.5 Survival Data

Since we are dealing with survival analysis, the data must be represented as survival data. As discussed in previous sections, there are several recording properties that have to be defined:

(i) The origin of time is chosen to be equal to the commitment date of the customer.

(ii) The customers are followed for a maximum period of 15 months.

(iii)The time scale is set to months.

The proportion of churned and censored customers in the training set is made equal, because this appears to provide the best results.

3.6 Lagging

An important assumption of the extended Cox model is that the effect of a time-dependent covariate $X_i(t)$ on the survival probability at time t depends on the value of this covariate at that same time t, and not on a value at an earlier or later time. Since the covariate values $X_i(t)$ are not available until time t, the model can only predict the hazard for time t. In our churn application this can be easily interpreted. The data of a customer is only completely available when a month has ended. The hazard is then predicted for that month. In fact, the prediction is never up to date. If the customer would be churned during that month, it could only be predicted afterwards when it is already too late. Our goal is thus to forecast a future time t + L, where L is typically 1 or 2. Notice that this would not be an issue if we would not incorporate time dependent covariates. After all**,** equation 3.3 shows that the time aspect is only represented by the baseline hazard, and not by the covariates. To be able to forecast a future time using time-dependent covariates, the time-dependent covariates must be forecasted or lagged by L units. We have chosen for a lag-time of 1. This means that the data of the previous month is used to predict the hazard of this month. The extended Cox model can be written to allow for a lag-time modification of any time-dependent covariate as follows,

$$\lambda_i(t) = \lambda_0(t)\exp\left[\sum_{i=1}^{p1}\beta X_i + \sum_{j=1}^{p2}\beta X_j(t - L_j)\right], \qquad (3.7)$$

where L denotes the lag-time for time-dependent covariate j.

3.7 Predictive Score Method

Survival models are not designed to be used for classification or prediction.

Therefore, a specific procedure is required. A predictive score method is used to classify churn events. The lag-time extended Cox model as discussed above is used for predicting churn in the next month. The predictive score method is based on the hazard function λ . Recall that the $\lambda(t)$□(t) is the instantaneous probability of having an event at time t given that one has survived (i.e. not had an event) up to time t. The hazard is thus a good candidate for measuring churn. A threshold is set to distinguish churn behaviour from non-churn behaviour. This threshold is empirically determined. If the hazard exceeds the threshold, the customer is classified as churned.

Figure 3.5 shows the estimated survival curve and the estimated baseline hazard for churn definition 2. Notice that since we used α = 3 for churn definition 2, every customer survives the first three months. Figure 3.5 shows an example of a customer that churns at month 11 and its corresponding hazard values including the threshold.

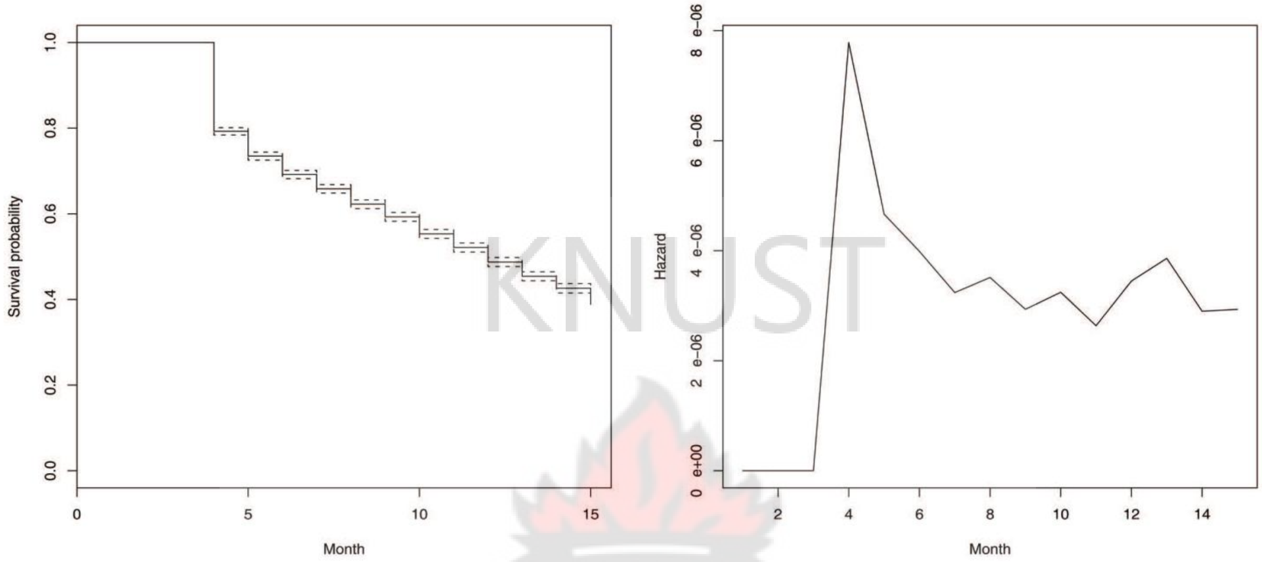


Figure 3.5: Left: The estimated survival curve Right: The estimated baseline hazard

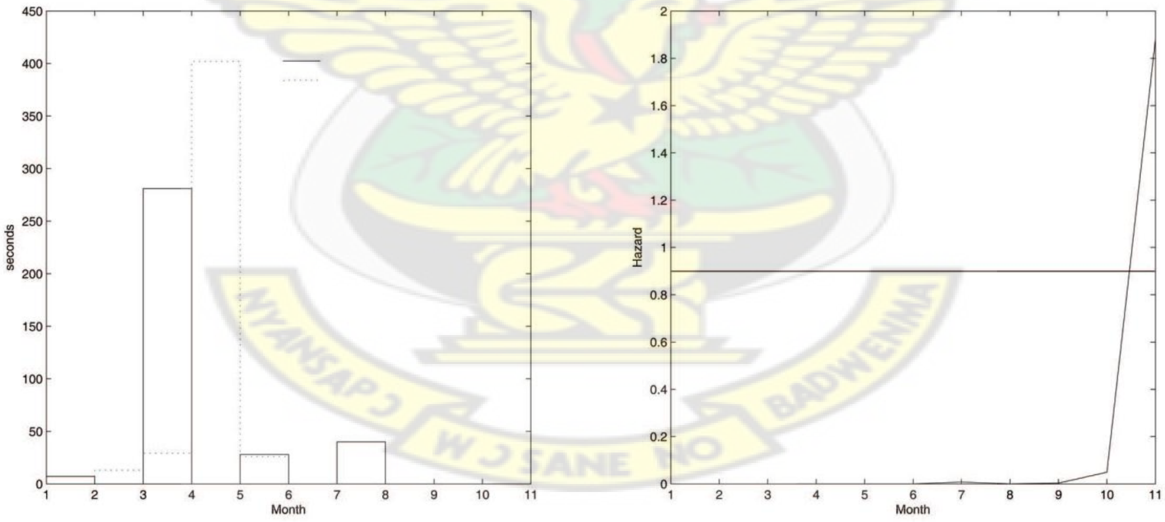


Figure 3.6: Left: illustration of customer behaviour. Right: The corresponding hazard value including the threshold.

3.8 Data Sources

Customer internal data is from the company’s data warehouse. It consists of two parts.

The first part is about customer information like market channel, plan type, payment, bill type, contract status, customer segmentation code, ownership of the company’s other products, dispute, late fee charge, discount, promotion/save promotion, additional accounts, rewards redemption, billing dispute, and so on. The second part of customer internal data is customer’s telecommunications usage data. Examples of customer usage variables are:

(i) Revenue

(ii) Length of calls (On-net/Off-net)

(iii)Call count

(iv)Number of SMS (On-net/Off-net)

(v) Top-up amount

(vi)GPRS usage

3.9 Summary

In this chapter, we have considered the methodological approach to model the problem of customer churn. In the following chapter, we put forward the data collection, analysis as well as the results of the study.

CHAPTER 4

DATA COLLECTION, ANALYSIS AND RESULTS

4.1 Introduction

Survival analysis was used to analyze the rate of churning among customers of the telecommunication industry in Ghana. The models are ultimately tested on a selection of prepaid customers from the database provided by Vodafone Ghana Limited.

The non-parametric estimation (Kaplan-Meier) and Parametric model (Cox’s Proportional Hazard’s Model) will both be used in the analysis. A thorough description of the survival curve is used to test the significance of the graph. During the Cox’s Proportional Hazard’s Model analysis, parameters in the model are estimated.

Althoughtherearewellknownmethodsforestimatingunconditionalsurvivaldistributions,mostinterestingsurvivalmodelingexaminestherelationshipbetweensurvivalandoneormorepredictors,usuallytermed covariatesinthesurvival-analysisliterature.

4.2 Data Collection

Customer related data for the research was collected from Vodafone Ghana Limited; a leading mobile telecom company in Ghana. According the protection law of consumer, we cannot catch any original data source, therefore we are used to the secondary data source included analysis results from company. The data source includes customer demography; Call Detail Records

(CDR) and billing data of churner and non-churner from July 2011 to December 2011. Since such kind of data is sensitive and confidential, they randomly sampled the data around 600 customers for this research that includes around 240 churners.

4.3 Data Analysis

PREMINARY ANALYSIS OF THE CHURN DATA

Summary Statistics for Time Variable SIM activation in months

Quartile Estimates			
Percent	Point Estimate	95% Confidence Interval	
		Lower	Upper
75	24.533	20.933	32.400
50	14.033	12.200	16.533
25	5.967	5.167	7.833

From the table above we find the estimated 75th, 50th, and 25th percentiles (labeled Quartiles). The 25th percentile (approximately 6months in this case) is the smallest event time such that the probability of churning earlier is greater than 0.25. Of greatest interest is the 50th percentile, which is, of course, the median churn time. Here, the median is approximately 14 months, with a 95-percent confidence interval of 12.2 to 16.5 months. An estimated mean time of churn is also reported. As noted on the output, the mean 23.5 months is biased downward when there are censoring times greater than the largest event time. Even when this is not the case, the upper tail of the distribution will be poorly estimated when a substantial number of the cases are censored, and this can greatly affect estimates of the mean. Consequently, the median is usually a much preferred measure of central tendency for censored survival data. Figure 1.1 below shows the overall estimated KM survivorship function for the churn data. At each churn time, the curve steps down to a lower value. It stops at the highest censoring time (136 months)

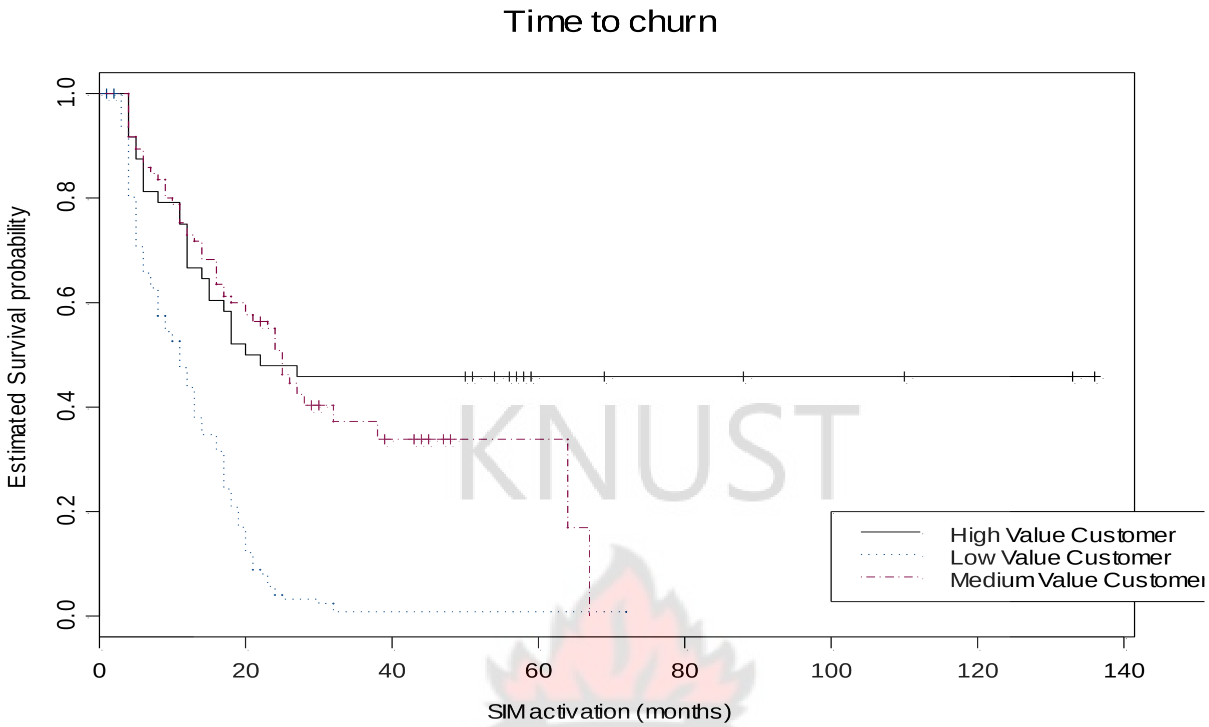
4.3.1 The Kaplan-Meier Curve

A plot of the Kaplan–Meier (**K-M**) estimate of the survival function is a series of horizontal steps of declining magnitude which, when a large enough sample is taken, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations is assumed to be constant.

Since the survivor function gives a complete accounting of the survival experience of each group, a natural approach to answering the question of equal probability of churning is to test the null hypothesis that the survivor functions are the same in the three groups; that is, $S_1(t) = S_2(t)= S_3(t)$ for all *t*,where the subscripts distinguish the three groups.

First, instead of a single table with KM estimates, separate tables are produced for each of the three value band groups. Second, corresponding to the two tables are three graphs of the survivor function, superimposed on the same axes for easy comparison

The graph of the estimated survivor function versus survival time to churn is shown in **Figure 4.1** below.



F

igure 4.1: Estimated KM survivor function for the three groups of value band.

The vertical axis gives the proportion of churning survival within the time frame. The values are fractions which run between 0 and 1, which is 100% survival at the extreme top to 0% survival at the extreme bottom. The horizontal axis gives the number of months for which the SIM has been activated. The small vertical tick-marks indicate losses which occur when a SIM is withdrawn from the study before the outcome is observed

From figure 4.1, before 5 months, the three survival curves are virtually indistinguishable, with little visual evidence of a group membership effect. The black horizontal line after 15 months reflects the fact that no additional customers churn occur in High Value Band after that time. For low and medium value customers, the last time for SIM activation churned. The figure also shows that the survival curve for low value customers drops precipitously, while the curve for high value customers declines much more gradually.

4.4 The Kaplan-Meier Analysis

Table 4.2: Kaplan-Meier Analysis on Value Band for HVC

Time(t)/months	No. risk	No. of Event	Survival	Std. err	95% LCI	95% LCI
0	48	0	1.000			
4	48	4	0.917	0.0399	0.824	0.998
5	44	2	0.875	0.0477	0.786	0.974
6	42	3	0.813	0.0563	0.709	0.931
8	39	1	0.792	0.0586	0.685	0.915
11	38	2	0.750	0.0625	0.637	0.883
12	36	4	0.667	0.0680	0.546	0.814
14	32	1	0.646	0.0690	0.524	0.796
15	31	2	0.604	0.0706	0.481	0.760
17	29	1	0.583	0.0712	0.459	0.741
18	28	3	0.521	0.0721	0.397	0.683
20	25	1	0.500	0.0722	0.377	0.663
22	24	1	0.479	0.0721	0.357	0.644
27	23	1	0.458	0.0719	0.357	0.623

Each line of numbers in table 4.2 corresponds to one of the distinct observed event times, arranged in ascending order (except for the first line, which is for time 0). The crucial column is the fourth labeled survival which gives the KM estimates. At 8 months, for example, the KM estimate is 0.792. We say, then, that the estimated probability that a customer will not churn for 39 months or more is 0.792 if he/she is a HVC subscriber. There are tied values (two or more cases with the same event time), as we have at 4 months and 11 months, the KM estimate is the same of the tied cases. The KM estimates are reported the same for the tied censored times. The last two columns present the lower and upper bounds of the 95% confidence intervals for the KM estimate.

Table4.3: Kaplan-Meier Analysis on Value Bandfor Medium Value Customers

Time(t)/months	No. risk	No. of Event	Survival	Std. err	95% LCI	95% LCI
0	85	0	1.000			
4	85	7	0.918	0.0298	0.861	0.978
5	78	2	0.894	0.0334	0.831	0.962
6	76	3	0.859	0.0378	0.788	0.936
7	73	1	0.847	0.0390	0.774	0.927
8	72	1	0.835	0.0402	0.760	0.918

9	71	3	0.800	0.0434	0.719	0.890
10	68	1	0.788	0.0443	0.706	0.880
11	67	3	0.753	0.0468	0.667	0.850
12	64	2	0.729	0.0482	0.641	0.830
13	62	1	0.718	0.0488	0.628	0.820
14	61	3	0.682	0.0505	0.590	0.789
16	58	4	0.635	0.0522	0.541	0.746
17	54	2	0.612	0.0529	0.516	0.725
18	52	1	0.600	0.0531	0.504	0.714
20	51	2	0.576	0.0536	0.480	0.692
21	46	1	0.564	0.0539	0.4676	0.680
23	43	1	0.551	0.0542	0.4542	0.668
24	39	3	0.508	0.0553	0.4109	0.629
25	33	3	0.462	0.0563	0.3640	0.587
26	28	1	0.446	0.0567	0.3474	0.572
27	21	1	0.424	0.0578	0.3250	0.554
28	20	1	0.403	0.0587	0.3032	0.536
32	13	1	0.372	0.0618	0.2688	0.516
38	11	1	0.338	0.0648	0.2325	0.493
64	2	1	0.169	0.1240	0.0403	0.711
67	1	1	0.000	NA	NA	NA

At 76 months, for example, the KM estimate is 0.859. We say, then, that the estimated probability that a customer will not churn after 76 months or more is 0.859 if he/she is a MVC subscriber. There are tied values (two or more cases with the same event time), as we have at 5 months and 7 months, the KM estimate is the same for the tied cases. The KM estimates are reported the same for the tied censored times. The last two columns present the lower and upper bounds of the 95% confidence intervals for the KM estimate.

Table 4.4: Kaplan-Meier Analysis on Value Band for Lower Value Customers

Time(t)/months	No. risk	No. of Event	Survival	Std. err	95% LCI	95% LCI
0	219	0	1.000			

3	219	14	0.936	0.0165	0.904	0.969
4	195	28	0.802	0.0274	0.750	0.857
5	152	18	0.707	0.0320	0.647	0.772
6	126	9	0.656	0.0339	0.593	0.726
7	116	5	0.628	0.0347	0.563	0.700
8	106	9	0.575	0.0360	0.508	0.650
9	95	5	0.544	0.0366	0.477	0.621
10	89	3	0.526	0.0368	0.459	0.603
11	83	8	0.475	0.0374	0.407	0.555
12	75	6	0.437	0.0375	0.370	0.517
13	68	9	0.379	0.0372	0.313	0.460
14	59	5	0.347	0.0367	0.282	0.427
16	54	5	0.315	0.0360	0.252	0.394
17	48	11	0.243	0.0337	0.185	0.319
18	35	5	0.208	0.0323	0.154	0.282
19	27	5	0.16965	0.03055	0.11920	0.2415
20	21	6	0.12118	0.02749	0.07768	0.1890
21	15	4	0.08886	0.02445	0.05182	0.1524
22	11	1	0.08079	0.02353	0.04565	0.1430
23	10	3	0.05655	0.02021	0.02807	0.1139
24	7	2	0.04039	0.01736	0.01739	0.0938
25	5	1	0.03231	0.01566	0.01250	0.0835
30	4	1	0.02424	0.01367	0.00802	0.0732
32	3	2	0.00808	0.00802	0.00116	0.0565
72	1	1	0.00000	NA	NA	NA

In the LVC category, however, the number of churns observed and the number of customers at the risk of churning was highest among all the three categories. The highest number of churns observed was 28 and occurred in the 4th month. There are no real trends in the number of churners observed but it was realized that the number of churns was very high within the 3rd, 4th and 5th months. The number of customers at risk is very high in the 3th month. This number then decreases sharply to 1 in the final month, which was the 72nd month in this category. The survival probability falls rather sharply from 0.936 in the 3rd month.

4.5 The Log-Rank test

The Log-Rank is one of the three pillars of modern survival analysis and it is commonly used to investigate whether or not survival curves differ from group to group. The main aim is to find an expression (based on the data) from which we know the distribution (or at least approximately) under the hypothesis.

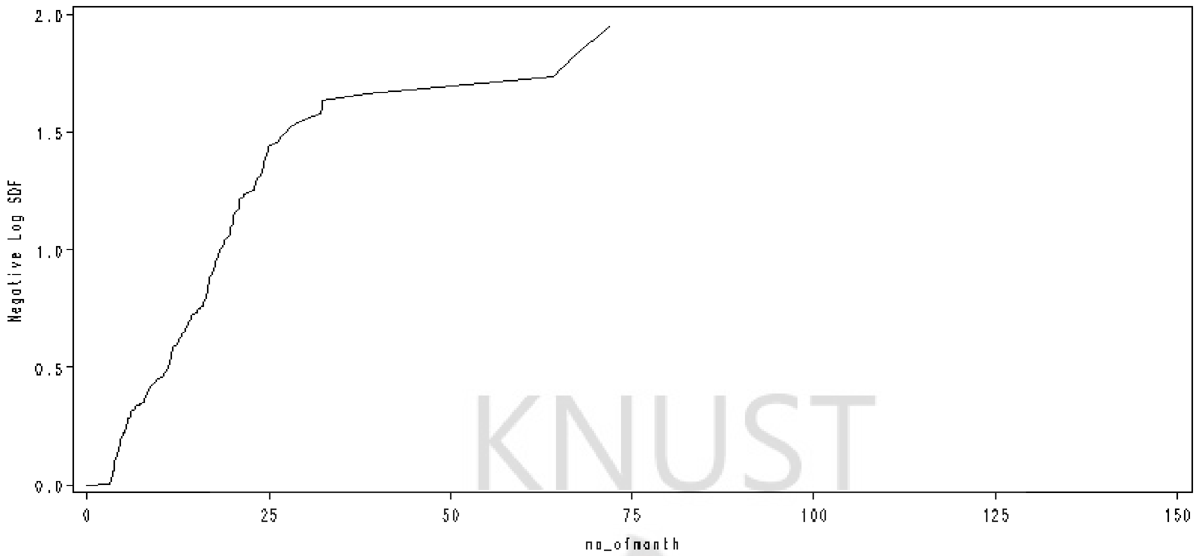
The Log-rank test of equality was conducted (using R) to investigate whether or not the differences observed in the Kaplan-Meier curves for the 3 groups are statistically significant. The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed.

Value Band	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
High Value Customer(HVC)	48	26	52.7	13.6	19.8
Medium Value Customer(MVC)	232	166103.2	38.2	79.6	
Low Value Customer (LVC)	85	51	87.0	14.9	26.0

Chisq= 80.2 on 2 degrees of freedom, p= < 0.0001

We do reject the null hypothesis of no difference between the three groups. There is statistical evidence that there exist differences in the survival probabilities among the value bands.

An examination of the log-survival (LS) plot can tell us whether the hazard is constant, increasing, or decreasing with time. The above table displays the log survival LS plot for the churn data. Instead of a straight line assuming a constant hazard, the graph appears to increase at a *decreasing* rate. This fact suggests that the hazard is *not* constant, but rather increase with time. If the plot had curved upward rather than downward, it would suggest that the hazard was *increasing* with time.



Fi

gure 4.5: Log survival Plot of Churn data

Model for usage and probability of churning

Par	Coef	Exp(coef)	Se(coef)	Z	P-value
TotalMOU	-0.000585	0.999	0.000293	-1.993	0.046
GPRSUsage	0.000145	1.000	0.000210	0.689	0.490
SMS	-0.000974	0.999	0.001154	-0.844	0.400
TopUpAmount	-0.008171	0.992	0.003918	-2.085	0.037

$\lambda_o(t)=\exp(-0.000585\textbf{TotalMOU}+0.000145\textbf{GPRSUsage}-0.000974\textbf{SMS}-0.008171\textbf{TopUpAmount})$

We use the p-value method to eliminate terms which are not significant at 95% confidence interval. Since "GPRSUsage" and "Total SMS" have relatively large p-values (>0.05), we conclude that their effects are not significant. However, "Total MOU" and "Top upAmount" have p-values which are less than 0.05. This shows that "Total MOU" and "Top upAmount" have significant effects on the survival probability. There is an overall significant relationship between these set of covariates ("Total MOU" and "Top upAmount") and survival time and they explain a significant portion of the variation observed.

The estimated coefficient for “TotalMOU” is -0.000585 with a p-value of 0.046. This means that fixing other covariates, customers with high “Total MOU” have lower churning probabilities and, hence, have longer survival time than those with low “TotalMOU” values.

Similarly, the estimated coefficient for “Top upAmount”is -0.008171 with a p-value of 0.037. Hence, if we fix all other covariates, customers with high Top upAmount have a decreased churning probability and, hence, have longer survival times than those with low “Top upAmount” values.

4.7Summary

In this chapter, we run a survival analysis on the data obtained. The non-parametric estimation (Kaplan-Meier) and Parametric model (Cox’s Proportional Hazard’s Model) were both used in the analysis. We also examined some factors that affect churning using the **PH** model.

Chapter 5

Summary, Conclusion and Recommendations

5.1 Overview

The conclusion of the present study is based on the research question andsub questions stated in Chapter 1. By answering these questions we do not onlyprovide an overview of the work presented in this study, but we consider theextent to which we succeeded answering these questions as well. Moreover,a number of recommendations for further research in this field are discussed.

5.2 Summary

In this research, survival analysis was used to analyze the rate of churning among customers of the telecommunication industry in Ghana and also to analyze the significance of some factors that are thought to affect churning. The models were ultimately tested on a selection of prepaid customers from the database provided by Vodafone Ghana Limited. The non-parametric estimation (Kaplan-Meier) and Parametric model (Cox's Proportional Hazard's Model) were both used in the analysis. A thorough description of the survival curve was used to test the significance of the graph.

From the analysis performed, it was found that there was a decrease in churning in all three categories analyzed. This rate of decrease was found to be highest among the LVC customers and lowest among the HVC customers.

The K-M Survival and Hazard Plot were used to predict the survival probabilities in all of the three categories and also to observe and explain the trends in the categories. Next, a Log-Rank test was performed to ascertain the fact that there exist significant differences between the three groups in context. Finally, the Cox Proportional Model was used to fit a model and also to find the factors that had significant effects on the rate of churning.

Non-parametric Kaplan-Meier estimates were performed to describe the survival characteristics. Log-rank tests were also carried out to test whether or not the differences observed in the Kaplan-Meier curves are statistically significant.

Next, the Cox PH test was used to examine some factors that have significant effect on the rates of churning and also to predict the ‘magnitude’ of this effect. The test was also used to predict how changes in these rates affected the rate of churning among the three categories.

5.3 Conclusion

The Log-rank test revealed that there was a significant difference between the churning rates of HVC, LVC, and MVC customers.

It was observed that the number at risk and the survival probability both decreases as time increases in all three categories (HVC, MVC, and LVC). This implies that the longer a person stays with a particular network, the lower the survival probability, which implies a high probability of churning. The survival curves for the three categories revealed that there were apparent differences in their rates of churning.

The median survival time is the time at which half of the customers in question have churned. By the survival curve in Figure 4.1, we observe that the median survival time is about 12 months in the LVC group whereas in both the HVC and MVC categories, it is almost 28 months. That means that the median survival time for an HVC or an MVC is over twice the median survival time for an LVC.

On the factors analyzed, the test revealed that “Total MOU” and “Top upAmount” had significant impacts on the probability of churning. However, the test also revealed that the effects of “GPRSUsage” and “Total SMS” were not significant in determining the survival probability of the customers.

Since “GPRS Usage” and “Total SMS” have relatively large p-values (>0.05), we conclude that their effects are not significant. However, “Total MOU” and “Top upAmount” have p-values which are less than 0.05. This shows that “Total MOU” and “Top upAmount” have significant effects on the survival probability. There is an overall significant relationship between these set of covariates (“Total MOU” and “Top upAmount”) and survival time and they explain a significant portion of the variation observed.

This means that fixing other covariates, customers with high “Total MOU” have decreased churning probabilities and, hence, have longer survival time than those with low “Total MOU” values.

Hence, if we fix all other covariates, customers with high Top upAmount have a decreased churning probability and, hence, have longer survival times than those with low “Top upAmount” values.

Moreover, customers who have higher “Total MOU” and “Top upAmount” are less likely to churn. Based on these factors, we can predict in advance, the attributes of customers whom we are going to lose in the near future so one can take corrective action.

5.4 Recommendations

Based on the analysis and findings, the following recommendations are made:

(i) In order to reduce churning in the telecommunication sector, the company should check the quality of their services, more especially their call rates. When a customer feels the utility he/she is getting is not worth the charges associated, he/she is more likely to churn. The charge for calls and SMS within the network should also be reduced. This will not only decrease the churn rate but will also prompt other customers to join the network.

(ii) More friendly service-help personnels should be trained in order to listen to and address problems that are reported by the customers. Conscious efforts should also be made to identify which category a customer belongs to. By this, it will be easier to provide them with services, promotions, and applications which will suit them and make them want to stay with the company even more.

(iii) Since post-paid customers are far easier to deal with when it comes to churn prediction, it is recommended that more pre-paid customers be encouraged to become post-paid customers. This could be achieved by making palatable offers for pre-paid customers who decide to become post-paid.

(iv) The coverage area of the network also accounts for the churn rate. This is due to the fact that when a customer relocates to an area without the coverage of the network, he/she has no choice than to churn to another network. Thus, an increase in the network coverage area would lead to a reduction in the churn rate.

(v) More conscious efforts should be made to predict possible churners based on their “Total MOU” and their “Top upAmount” rates. Promotions that are aimed at customer retention should therefore be steered in this direction.

REFERENCES

1. T. Therneau and P. Grambsch. Modeling Survival Data: Extending the Cox Model. Springer, New York, USA, 2000.
2. B. Ripley and R. Ripley. Neural networks as statistical methods in survival analysis. In Artificial Neural Networks: Prospects for Medicine. Landes Biosciences Publishers, 1998.
3. R. Miller. Survival Analysis. John Wiley and Sons, New York, USA, 1981.
4. F. Harrell. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer, New York, USA, 2001.
5. M. Halling and E. Hayden. Bank failure prediction: A two-step survival time approach. http://ssrn.com/abstract=904255, 2006.
6. S. Balcaen and H. Ooghe. Alternative methodologies in studies on business failure: do they produce better results than the classic statistical methods? Vlerick Leuven Gent Management School Working Paper Series, 2004.
7. David G. Kleinbaum Mitchel Klein. Survival Analysis A Self-Learning Text Second Edition
8. Prof S. Chandrasekhar B. Tech. M. Tech. (IIT, Kanpur) Ph. D (USA): Predicting the churn in the telecom industry
9. Ali TamaddoniJahromi. (2009). *Predicting Customer Churn in telecommunications Service Providers*',Lulea University of Technology.
10. MO Zan, ZHOA Shan, LI Li, LIU Ai-Jun. (2007).*A Predictive Model of Churn in Telecommunications*

11. MarcinOwczarczuk. Institute of Econometrics, Warsaw School of Economics Al. Niepodleglosci 164, 02-554 Warsaw, Poland: Churn models for prepaid customers in the cellular telecommunication industry using large data marts
12. National Communication Authority: http://www.nca.org.gh/
13. Fleming, T. H. and Harrington, D. P. (1984). Nonparametric estimation of the survival distribution in censored data.
14. Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York:Wiley.
15. Link, C. L. (1984). Confidence intervals for the survival function using Cox's proportional hazards model with covariates.
16. Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes, a large sample study.
17. Therneau, T., Grambsch, P., and Fleming. T. (1990). Martingale based residuals for survival models.
18. Aalen, O.O. (1978) Nonparametric inference for a family of counting processes. Annals of Statistics.
19. Altman, D.G. (1991) Practical Statistics for Medical Research, Chapman & Hall/ CRC, London.
20. Breslow, N.E.; Crowley, J. (1974) A large sample study of the life tables and Product Limit Estimator under random censorship.

APPENDIX

Table A1: ***Overall estimated KM survivorship for the sampled data***

Obs	SIM activation	CENSOR	SURVIVAL	LCL	UCL
1	0.000	.	1.00000	1.00000	1.00000
2	1.000	5	1.00000	.	.
8	2.000	7	1.00000	.	.
15	3.000	10	1.00000	.	.
25	3.033	0	0.99708	0.97943	0.99959
26	3.100	0	0.99415	0.97682	0.99853
27	3.167	0	0.99123	0.97305	0.99716
28	3.267	0	0.98538	0.96523	0.99389
29	3.300	0	0.97953	0.95755	0.99019
30	3.333	0	0.97661	0.95377	0.98823
31	3.367	0	0.96784	0.94267	0.98206
32	3.433	0	0.96491	0.93904	0.97992
33	3.467	0	0.95906	0.93186	0.97555
34	3.500	0	0.95614	0.92830	0.97332
35	3.533	0	0.94737	0.91776	0.96651
36	3.567	0	0.94444	0.91428	0.96420
37	3.600	0	0.94152	0.91082	0.96187
38	3.633	0	0.93275	0.90053	0.95479
39	3.667	0	0.92398	0.89035	0.94759
40	3.733	0	0.91228	0.87693	0.93783
41	3.767	0	0.90643	0.87028	0.93290
42	3.800	0	0.90351	0.86697	0.93041
43	3.833	0	0.89766	0.86037	0.92542
44	3.900	0	0.89474	0.85708	0.92292
45	3.967	0	0.89181	0.85380	0.92040
46	4.000	0	0.88889	0.85053	0.91788
47	4.000	15	0.88889	.	.
62	4.067	0	0.88581	0.84706	0.91524
63	4.133	0	0.87966	0.84014	0.90994
64	4.167	0	0.87043	0.82982	0.90193
65	4.333	0	0.86736	0.82640	0.89924
66	4.367	0	0.86428	0.82298	0.89655
67	4.400	0	0.85813	0.81617	0.89115
68	4.433	0	0.84890	0.80600	0.88301
69	4.467	0	0.84275	0.79925	0.87755
70	4.500	0	0.83660	0.79252	0.87208
71	4.567	0	0.83045	0.78581	0.86658
72	4.600	0	0.82430	0.77913	0.86106
73	4.767	0	0.81815	0.77246	0.85552
74	4.800	0	0.81507	0.76913	0.85275
75	4.867	0	0.81200	0.76581	0.84997
76	4.933	0	0.80892	0.76250	0.84719

77	4.967	0	0.80584	0.75918	0.84440
78	5.000	8	0.80584	.	.
86	5.033	0	0.80267	0.75575	0.84153
87	5.133	0	0.79950	0.75233	0.83866
88	5.167	0	0.79315	0.74549	0.83290
89	5.200	0	0.78998	0.74208	0.83001
90	5.300	0	0.78681	0.73868	0.82712
91	5.333	0	0.78364	0.73528	0.82423
92	5.400	0	0.77412	0.72510	0.81552
93	5.500	0	0.77095	0.72172	0.81261
94	5.533	0	0.76143	0.71159	0.80386
95	5.567	0	0.75825	0.70822	0.80093
96	5.633	0	0.75508	0.70485	0.79800
97	5.800	0	0.75191	0.70149	0.79507
98	5.967	0	0.74874	0.69814	0.79214
99	6.000	1	0.74874	.	.
100	6.033	0	0.74555	0.69477	0.78919
101	6.067	0	0.73599	0.68468	0.78032
102	6.200	0	0.72962	0.67797	0.77439
103	6.467	0	0.72643	0.67463	0.77142
104	6.600	0	0.72325	0.67128	0.76844
105	6.667	0	0.72006	0.66794	0.76547
106	6.700	0	0.71688	0.66460	0.76249
107	6.900	0	0.71369	0.66127	0.75950
108	7.000	5	0.71369	.	.
113	7.233	0	0.71043	0.65785	0.75646
114	7.300	0	0.70717	0.65444	0.75341
115	7.800	0	0.70391	0.65103	0.75036
116	7.833	0	0.69740	0.64421	0.74424
117	7.867	0	0.69414	0.64081	0.74118
118	7.900	0	0.69088	0.63742	0.73811
119	8.000	0	0.68762	0.63402	0.73505
120	8.000	2	0.68762	.	.
122	8.133	0	0.68433	0.63060	0.73195
123	8.233	0	0.68104	0.62717	0.72885
124	8.300	0	0.67446	0.62033	0.72264
125	8.467	0	0.67117	0.61692	0.71953
126	8.533	0	0.66788	0.61351	0.71641
127	8.567	0	0.66459	0.61010	0.71330
128	8.733	0	0.66130	0.60670	0.71018
129	8.767	0	0.65801	0.60330	0.70706
130	8.967	0	0.65472	0.59990	0.70393
131	9.000	1	0.65472	.	.
132	9.067	0	0.65141	0.59649	0.70079
133	9.300	0	0.64811	0.59308	0.69765
134	9.433	0	0.64480	0.58967	0.69450

135	9.600	0	0.64149	0.58627	0.69135
136	9.633	0	0.63819	0.58286	0.68819
137	9.733	0	0.63488	0.57947	0.68504
138	10.000	3	0.63488	.	.
141	10.400	0	0.63152	0.57601	0.68183
142	10.567	0	0.62816	0.57256	0.67863
143	10.600	0	0.62480	0.56911	0.67542
144	10.800	0	0.61808	0.56222	0.66899
145	10.900	0	0.61472	0.55878	0.66577
146	11.067	0	0.61136	0.55534	0.66255
147	11.167	0	0.60801	0.55191	0.65933
148	11.267	0	0.60465	0.54848	0.65610
149	11.300	0	0.59793	0.54163	0.64964
150	11.333	0	0.59457	0.53821	0.64640
151	11.400	0	0.59121	0.53479	0.64316
152	11.467	0	0.58785	0.53137	0.63992
153	11.500	0	0.58449	0.52796	0.63668
154	11.533	0	0.58113	0.52455	0.63343
155	11.600	0	0.57777	0.52115	0.63018
156	11.700	0	0.56770	0.51095	0.62041
157	11.833	0	0.56434	0.50755	0.61715
158	11.900	0	0.55762	0.50077	0.61063
159	12.000	1	0.55762	.	.
160	12.200	0	0.55424	0.49736	0.60734
161	12.367	0	0.55086	0.49396	0.60405
162	12.467	0	0.54748	0.49055	0.60076
163	12.600	0	0.54410	0.48715	0.59746
164	12.667	0	0.54072	0.48376	0.59417
165	12.833	0	0.53396	0.47697	0.58756
166	12.967	0	0.53058	0.47358	0.58426
167	13.033	0	0.52720	0.47020	0.58095
168	13.300	0	0.52382	0.46681	0.57764
169	13.367	0	0.52044	0.46343	0.57433
170	13.400	0	0.51706	0.46006	0.57101
171	13.467	0	0.51368	0.45668	0.56769
172	13.600	0	0.51031	0.45331	0.56437
173	13.800	0	0.50693	0.44994	0.56105
174	13.900	0	0.50355	0.44658	0.55772
175	13.967	0	0.50017	0.44322	0.55439
176	14.033	0	0.49679	0.43986	0.55106
177	14.200	0	0.49341	0.43650	0.54772
178	14.367	0	0.49003	0.43314	0.54438
179	14.400	0	0.48665	0.42979	0.54104
180	14.467	0	0.48327	0.42644	0.53770
181	15.100	0	0.47989	0.42310	0.53436
182	15.133	0	0.47651	0.41976	0.53101

183	15.567	0	0.47313	0.41642	0.52766
184	15.633	0	0.46975	0.41308	0.52430
185	15.900	0	0.46637	0.40974	0.52095
186	15.933	0	0.46299	0.40641	0.51759
187	15.967	0	0.45961	0.40308	0.51423
188	16.000	1	0.45961	.	.
189	16.267	0	0.45621	0.39973	0.51084
190	16.367	0	0.44940	0.39303	0.50406
191	16.433	0	0.44599	0.38969	0.50067
192	16.533	0	0.44259	0.38635	0.49727
193	16.633	0	0.43919	0.38301	0.49387
194	16.667	0	0.42897	0.37301	0.48365
195	16.700	0	0.42557	0.36968	0.48024
196	16.733	0	0.42216	0.36636	0.47683
197	16.800	0	0.41876	0.36304	0.47341
198	16.833	0	0.41535	0.35972	0.46999
199	17.000	0	0.41195	0.35640	0.46657
200	17.000	2	0.41195	.	.
202	17.200	0	0.40849	0.35303	0.46309
203	17.233	0	0.40503	0.34966	0.45961
204	17.333	0	0.40156	0.34629	0.45613
205	17.367	0	0.39810	0.34292	0.45264
206	17.533	0	0.39464	0.33956	0.44916
207	17.600	0	0.39118	0.33620	0.44566
208	17.700	0	0.38772	0.33285	0.44217
209	17.900	0	0.38425	0.32950	0.43867
210	18.000	3	0.38425	.	.
213	18.100	0	0.37714	0.32259	0.43150
214	18.200	0	0.37002	0.31571	0.42431
215	18.300	0	0.36647	0.31227	0.42071
216	18.567	0	0.36291	0.30883	0.41710
217	18.767	0	0.35935	0.30540	0.41350
218	18.867	0	0.35579	0.30198	0.40989
219	19.000	1	0.35579	.	.
220	19.067	0	0.35220	0.29851	0.40624
221	19.300	0	0.34860	0.29506	0.40259
222	19.600	0	0.34501	0.29160	0.39894
223	19.733	0	0.33782	0.28470	0.39162
224	19.767	0	0.33423	0.28126	0.38796
225	20.000	3	0.33423	.	.
228	20.033	0	0.33051	0.27770	0.38418
229	20.100	0	0.32309	0.27057	0.37662
230	20.200	0	0.31937	0.26702	0.37283
231	20.267	0	0.31566	0.26347	0.36904
232	20.567	0	0.31195	0.25993	0.36524
233	20.867	0	0.30823	0.25639	0.36144

234	20.900	0	0.30452	0.25286	0.35763
235	20.933	0	0.30081	0.24933	0.35382
236	21.033	0	0.29709	0.24580	0.35001
237	21.600	0	0.29338	0.24229	0.34619
238	21.667	0	0.28966	0.23877	0.34236
239	22.000	2	0.28966	.	.
241	22.967	0	0.28585	0.23516	0.33845
242	23.000	0	0.28204	0.23155	0.33453
243	23.000	3	0.28204	.	.
246	23.133	0	0.27807	0.22777	0.33046
247	23.300	0	0.27410	0.22400	0.32638
248	24.000	0	0.26615	0.21649	0.31822
249	24.000	3	0.26615	.	.
252	24.067	0	0.26199	0.21254	0.31396
253	24.200	0	0.25783	0.20859	0.30970
254	24.367	0	0.25368	0.20466	0.30542
255	24.533	0	0.24952	0.20074	0.30115
256	24.600	0	0.24536	0.19682	0.29686
257	24.900	0	0.24120	0.19292	0.29256
258	24.933	0	0.23704	0.18902	0.28826
259	25.000	2	0.23704	.	.
261	26.000	6	0.23704	.	.
267	26.333	0	0.23220	0.18438	0.28337
268	26.867	0	0.22737	0.17977	0.27847
269	27.333	0	0.22253	0.17516	0.27356
270	28.000	0	0.21769	0.17058	0.26863
271	28.000	2	0.21769	.	.
273	29.000	1	0.21769	.	.
274	29.767	0	0.21251	0.16563	0.26340
275	30.000	3	0.21251	.	.
278	32.000	1	0.21251	.	.
279	32.200	0	0.20676	0.16006	0.25771
280	32.400	0	0.20102	0.15452	0.25200
281	32.467	0	0.19528	0.14903	0.24625
282	38.367	0	0.18953	0.14356	0.24047
283	39.000	2	0.18953	.	.
285	43.000	1	0.18953	.	.
286	44.000	1	0.18953	.	.
287	45.000	1	0.18953	.	.
288	47.000	2	0.18953	.	.
290	48.000	1	0.18953	.	.
291	50.000	2	0.18953	.	.
293	51.000	1	0.18953	.	.
294	54.000	1	0.18953	.	.
295	56.000	2	0.18953	.	.
297	57.000	3	0.18953	.	.

300	58.000	1	0.18953	.	.
301	59.000	1	0.18953	.	.
302	64.167	0	0.17600	0.12759	0.23088
303	66.867	0	0.16246	0.11256	0.22046
304	69.000	4	0.16246	.	.
308	71.900	0	0.14215	0.08845	0.20820
309	88.000	1	.	.	.
310	110.000	1	.	.	.
311	133.000	1	.	.	.
312	136.000	4	.	.	.