Three Essays in Modeling Customer Retention

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy in Management

by

Zainab Jamal

2006
The dissertation of Zainab Jamal is approved.

Andrew Ainslie

William Mason

Donald G. Morrison

Randolph Bucklin, Committee Chair

University of California, Los Angeles

2006
To Pawan, the wind beneath my wings.

To Sophiya, my blessing.

To my unborn son, in anticipation.
# TABLE OF CONTENTS

1 Introduction 1

1.1 Overview of Research 1

1.2 References 8

2 Customer Fit and Customer Retention at an Internet Recommendation Site 9

2.1 Introduction 11

2.2 Modeling Framework 21

2.3 Data and Empirical Specification 26

2.4 Empirical Results and Discussion 33

2.5 Managerial Implications 39

2.6 Conclusions 42

2.7 Appendices 46

2.8 References 59

3 Improving the Diagnosis and Prediction of Customer Churn: A Heterogeneous Hazard Modeling Approach 66

3.1 Introduction 67

3.2 Modeling Approach 71

3.3 Data and Empirical Specification 74

3.4 Empirical Results and Discussion 81
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5 Conclusions</td>
<td>89</td>
</tr>
<tr>
<td>3.6 Appendices</td>
<td>93</td>
</tr>
<tr>
<td>3.7 References</td>
<td>101</td>
</tr>
<tr>
<td><strong>4 Hazards of Ignoring Involuntary Customer Churn</strong></td>
<td>105</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>107</td>
</tr>
<tr>
<td>4.2 Modeling Framework</td>
<td>110</td>
</tr>
<tr>
<td>4.3 Data and Empirical Specification</td>
<td>114</td>
</tr>
<tr>
<td>4.4 Empirical Results and Discussion</td>
<td>122</td>
</tr>
<tr>
<td>4.5 Conclusions</td>
<td>129</td>
</tr>
<tr>
<td>4.6 Appendices</td>
<td>132</td>
</tr>
<tr>
<td>4.7 References</td>
<td>148</td>
</tr>
<tr>
<td><strong>5 Conclusions</strong></td>
<td>152</td>
</tr>
<tr>
<td>5.1 Overall Summary</td>
<td>152</td>
</tr>
<tr>
<td>5.2 References</td>
<td>159</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

2.1 Illustration of a Visitor’s Event History 57
2.2 Baseline Hazard Function by Stratum 58
3.1 Cumulative Lift Chart 100
4.1 Cumulative Lift Chart 147
LIST OF TABLES

2.1 Number of Customers and Ratings for All Customers 51
2.2 Distribution of Customers by Number of Visits 52
2.3 Descriptive Statistics 53
2.4 Model Comparison in Estimation Sample 54
2.5 Parameter Estimates for Stratified Model (Model 5) 55
2.6 Model Performance in Customer Targeting 56
3.1 Summary Statistics 94
3.2 Parametric Hazard Model Selection 95
3.3 Latent Class Weibull Model Selection 96
3.4 Model Performance in Holdout Sample 97
3.5 Model Parameters – Three Segment Weibull 98
3.6 Segment Level Summary Statistics of Covariates 99
4.1 Summary Statistics 139
4.2 Model Selection (Voluntary and Involuntary Churn Sample) 140
4.3 Model Selection (Voluntary Churn Sample) 141
4.4 Model Selection (Involuntary Churn Sample) 142
4.5 Univariate Weibull Model – Parameter Estimates 143
4.6 Bivariate Weibull Model – Parameter Estimates 144
4.7 Model Comparison – Impact of Covariates on Voluntary Churn 145
4.8 Bivariate Weibull Model – Impact of Covariates on Voluntary and Involuntary Churn
ACKNOWLEDGEMENTS

I wish to thank all my colleagues and friends who have contributed to this dissertation, directly or indirectly, through their inputs and support. A special thanks to the committee members – Andrew Ainslie, William Mason, Don Morrison – and especially to my advisor Randy Bucklin for his understanding, guidance and support. I also wish to thank my family for their unconditional support in helping me achieve the balance between work life and family life.
VITA

April 25, 1973  Born, New Delhi, India

1990 – 1993  B.A., Economics
Hindu College, Delhi University
Delhi, India

Delhi School of Economics, Delhi University
Delhi, India

Indian Institute of Management (I.I.M.)
Ahmedabad, India

1997 – 1999  Assistant Brand Manager
SmithKline Beecham
Delhi, India

2000 – 2001  Product Manager
Go4i.com
Delhi, India

2001 – 2006  Research Assistant
UCLA Anderson School of Management
University of California, Los Angeles

2003  Entertainment Management Research Grant
UCLA Center for Communication Policy and
Entertainment Media
2004  Marketing Science Doctoral Consortium Fellow  
Erasmus University  
Rotterdam, Netherlands

2005  Marketing Science Doctoral Consortium Fellow  
Emory University  
Atlanta, Georgia

PUBLICATIONS AND PRESENTATIONS

Journal of Interactive Marketing, Forthcoming.


ABSTRACT OF THE DISSERTATION

Three Essays in Modeling Customer Retention

by

Zainab Jamal

Doctor of Philosophy in Management

University of California, Los Angeles, 2006

Professor Randolph Bucklin, Chair

The objective of this research was to improve the prediction of customer retention and customer churn and to establish empirically the impact on these rates of different factors in the context of the customer-firm relationship. In this dissertation I focus on three research problems. In the first problem, I model the prediction of customer retention in the context of repeat visits to a website. I incorporate heterogeneity in customer retention over time by allowing customer retention probabilities to vary across first and subsequent return visits. I show the effects of
customer fit (how well the product meets the customer’s requirement), switching costs and customer interactions on customer retention at an Internet recommendation site. In the second problem, I model the prediction of customer churn in the context of a continuous subscription product i.e., direct-to-home satellite television. I highlight the importance of incorporating heterogeneity in the customer churn rates across customers. I empirically link customer churn to customer service experience, failure recovery (how well the firm performs in its recovery efforts after a service failure) and payment equity. In the third problem, I model the prediction of different types of customer churn for a continuous subscription product. I show the importance of accounting for all types of customer churn – voluntary as well as involuntary when estimating the churn probabilities. Ignoring the existence of involuntary churn significantly impacts the prediction and diagnosis of voluntary churn.
CHAPTER 1

INTRODUCTION

1.1 Overview of Research

Understanding customer retention and its complement customer churn is an important focus area for managers. This importance has been highlighted in a number of studies. For example, small increases in retention rates have been shown to have a dramatic effect on the profits of a company because the cost of retaining an existing customer is less than the cost of acquiring a new customer, existing customers tend to purchase more than new customers, and there are efficiencies in dealing with existing customers than new customers (e.g., Fornell and Wernerfelt 1987, Reichheld and Sasser 1990). Customer retention has also been shown to have a significant impact on a company’s performance outcomes such as share price, market value and customer lifetime value. A 1% increase in retention rate has been shown to increase firm value on average by 5% (Gupta et al. 2004). A reduction in churn by 5 percentage points has been shown to double profits in some industries (Reichheld 1996).

A critical managerial issue is therefore how to influence these rates in order to improve the firm’s business performance. This is especially relevant in industries where the annual churn rate is quite high (e.g., U.S. long distance telephone (30%), cell phone (30%), Internet service providers (22%), Thomas et al. 2004, Bolton 1998).
Managers need models that can predict customer retention or churn accurately and which will empirically link these rates to factors under their control.

Further, the advent of customer-centric organizations has spurred the need among managers for models that help them to take actions at the customer level and in the context of a dynamic customer-firm relationship. This raises the need for a deeper understanding of issues not explored in detail in the past. For example, how valid is the assumption that customer retention rates are fixed across customers and over time? How does the type of churn – voluntary vis-à-vis involuntary – influence the prediction of customer churn rates? We hope to shed some light on these issues in our research.

The objective of my research is to improve the prediction of customer retention (churn) and to establish empirically the impact of different factors on these rates taking into consideration some of the issues pertaining to the nature of the customer-firm relationship. We focus on three research problems. In essay 1, we model the prediction of customer retention in the context of repeat visits to a website. We incorporate heterogeneity in customer retention over time by allowing customer retention probabilities to vary across first and subsequent return visits. We show the effects of customer fit (how well the product meets the customer’s requirement), switching costs and customer interactions on customer retention at an Internet recommendation site. In essay 2, we model the prediction of customer churn in the context of a continuous subscription product i.e., direct-to-home satellite television.
We highlight the importance of incorporating heterogeneity in the customer churn rates across customers. We empirically link customer churn to customer service experience, failure recovery (how well the firm performs in its recovery efforts after a service failure) and payment equity. In essay 3, we model prediction of different types of customer churn for a continuous subscription product. We show the importance of accounting for all types of customer churn – voluntary as well as involuntary when estimating the churn probabilities. We show that ignoring the existence of involuntary churn significantly impacts the prediction and diagnosis of voluntary churn.

We use hazard models to analyze the duration data in the three research studies. Hazard models are well established in the statistics literature and have been shown to be well suited for analysis of duration data and superior to other methods, such as logistic and least squares regressions, in terms of stability, face validity, and predictive accuracy (e.g., Helsen and Schmittlein 1993). They also are hardly new to the marketing or CRM literature (e.g., Jain and Vilcassim 1991, Bolton 1998).

An added dimension of our research is that we estimate our models using information which is readily available with managers. In essay 1 we use clickstream data which track a customer’s onsite behavior at an Internet site. In essays 2 and 3 we use secondary data from the customer relationship management software of a continuous service provider that tracks customer specific transaction and demographic data. We use this information to build modeling constructs identified in the extant literature as impacting customer retention or churn but typically measured with survey
and experimental data. We hope our research will help managers to more effectively use information which they collect on an ongoing basis to analyze customer retention and churn.

Essay 1: Customer Fit and Customer Retention at an Internet Recommendation Site

We study the effects of customer fit, switching costs and customer interactions on customer retention at an Internet recommendation site. We define ‘customer fit’ as the accuracy with which the site’s recommendation engine reproduces a customer’s stated preferences. We capture switching costs using the number of ratings given by a customer during a visit. We also capture customer interactions with the site using the average time spent giving a rating and the number of categories within the site visited by the customer.

We use a stratified conditional Cox proportional hazard model for multiple events where each return visit to the site is defined as an event. The stratification allows a different baseline hazard function for the first versus subsequent return visits, permitting customer retention probabilities to vary across first return versus subsequent return visits.

We estimate our model using clickstream data from the movies category of an Internet recommendation site. We find that the quality of the site’s recommendation engine, as captured by idiosyncratic customer fit, significantly affects the probability of a customer’s return visit. We also find a significant positive impact of switching costs on customer retention and that initial interactions during the first visit
significantly influence customer retention. Each of these factors had the most impact following the customer’s first visit, with effect sizes declining or becoming insignificant for subsequent visits.

*Essay 2: Improving the Prediction and Diagnosis of Customer Churn: A Heterogeneous Hazard Modeling Approach*

In this research we model the prediction of customer churn and establish empirically the link between customer churn and factors such as customer service experience, failure recovery and payment equity. We do this in the context of a continuous service provider – direct-to-home satellite television firm based in an emerging market country in South America.

We model the probability of a customer exiting from the firm in a given month (customer churn) as a function of customer satisfaction with the customer service experience, customer evaluation of the firm’s failure recovery efforts, and customer satisfaction with the cost-benefit tradeoff of the product. These factors are measured over the duration of the customer’s relationship with the firm using secondary data readily available from the firm’s customer relationship management software.

We use a latent class Weibull parametric hazard model which incorporates heterogeneity not only in the baseline hazard probabilities but also in the response parameters. Further, the model allows us to study the impact of time varying
covariates (billing and transaction variables which change with a monthly frequency) on the hazard probability.

We find significant links between customer service experience, failure recovery efforts, and payment equity on the customer’s churn probabilities. Further, there are differences across the customer segments in the magnitude and significance of the response parameters which would not have been highlighted without incorporating heterogeneity.

**Essay 3: Hazards of Ignoring Involuntary Customer Churn**

We study the impact of ignoring involuntary churn (when a firm terminates the subscription of a customer) on the estimation and diagnosis of voluntary churn (when the customer terminates the relationship). The presence of competing events has been shown to bias the estimates of the hazard rates and survival times of the main event. In our case, the main event is voluntary churn and the competing event is involuntary churn.

We estimate a bivariate Weibull survival model that captures the dependency between the two event times and has been proposed in the literature as an approach to the problem of dependent competing risk. We compare this model with a benchmark model – a univariate Weibull model with only voluntary churn. We estimate the models using maximum likelihood techniques.

We find significant differences in the prediction of voluntary churn rates. Also the impact of covariates on the voluntary churn rates is different across the two models.
Further, the bivariate Weibull survival model does better in predicting the customers who are more likely to churn. An added dimension of the study is that the key covariates that influence voluntary churn rate impact involuntary churn rate differently. Our study highlights the need to incorporate involuntary churn when modeling the voluntary churn process.
1.2 References


CHAPTER 2
CUSTOMER FIT AND CUSTOMER RETENTION AT AN INTERNET RECOMMENDATION SITE

Abstract

We study the effects of customer fit, switching costs and customer interactions on customer retention at an Internet recommendation site. We define ‘customer fit’ as the accuracy with which the site’s recommendation engine reproduces a customer’s stated preferences. We capture switching costs using the number of ratings given by a customer during a visit. We also capture customer interactions with the site using the average time spent giving a rating and the number of categories within the site visited by the customer.

We use a stratified conditional Cox proportional hazard model for multiple events where each return visit to the site is defined as an event. The stratification allows a different baseline hazard function for the first versus subsequent return visits, permitting customer retention probabilities to vary across first return versus subsequent return visits.

We estimate our model using clickstream data from the movies category of an Internet recommendation site. We find that the quality of the site’s recommendation engine, as captured by idiosyncratic customer fit, significantly affects the probability of a customer’s return visit. We also find a significant positive impact of switching
costs on customer retention and that initial interactions during the first visit significantly influence customer retention. Each of these factors had the most impact following the customer’s first visit, with effect sizes declining or becoming insignificant for subsequent visits.

Keywords: Customer Retention, Customer Lifetime Value, Hazard Models, Internet, Recommendation Systems, Stratification.
2.1 Introduction

Understanding customer retention – the ability of a company to retain its acquired customers – has become a focus area for marketing managers. Past research has underlined this importance. For example, small increases in retention rates have been shown to have a dramatic effect on the profits of a company. The cost of retaining an existing customer is less than the cost of acquiring a new customer, existing customers tend to purchase more than new customers, and there are efficiencies in dealing with existing customers than with new customers (e.g., Fornell and Wernerfelt 1987, Reichheld and Sasser 1990). Customer retention has also been shown to have a significant impact on a company’s performance outcomes such as share price, market value and customer lifetime value. A 1% increase in retention rate has been shown to increase firm value on average by 5% (Gupta et al. 2004). Further, long life customers can increase the reputation of the company and attract new customers through word of mouth (Villanueva 2002).

A number of factors intrinsic to the customer have been identified in the literature as affecting customer retention. These include customer satisfaction, loyalty, commitment, trust, service quality, perceived value, customer interactions and switching costs (Zeithaml and Gupta 2004, Blattberg et al. 2001, Johnson et al. 2003). But there is little empirical work that establishes and quantifies the impact of these factors on customer retention and other observable behavioral constructs (Zeithaml and Gupta 2004).
In the Internet setting, loyalty has been found to be higher than in the offline world (Brynjolfsson and Smith 2000). In fact, studies have shown that online customers exhibit lock-in effects and do not search as much as expected (Johnson et al. 2003). Thus, the Internet provides us with an opportunity to study customer retention at an individual level using tracking information from clickstream data. We can, thus, study directly the influence of customer fit – how well a product meets the customer’s requirement – on customer retention.

The objective of our research is to investigate the empirical link between customer retention, conditional on acquisition, at a website and its determinants such as customer fit, switching costs and customer interactions. Customer retention probability, in this context, is the probability of return visit of an acquired customer to the web site. In our research we are also interested in how customer retention probabilities may vary across first and subsequent return visits. To do this, we will use a stratified hazard modeling approach specifically designed to handle this potential variation. A key goal of our research is to better understand, quantitatively, the factors which drive customer retention in an Internet business setting.

In the past research literature, customer fit and customer retention have been shown to be linked to each other, albeit indirectly. Customer fit has been considered as one of the primary dimensions of quality (Garvin 1988, Juran 1988). Several studies have shown that perceived quality affects customer satisfaction (Anderson and Sullivan 1993, Bolton and Drew 1991). Customer satisfaction, in turn, has been shown
to have a strong positive link with customer retention (e.g., Anderson and Sullivan 1993, Boulding et al. 1993, Bolton 1998).

According to past research, online customers’ loyalty behavior could be explained by the costs they would experience if they switch to another site (e.g., Alba et al. 1997). One source for these switching costs could be learning or cognitive costs (e.g., better navigation skills) (Johnson et al. 2003). Another source for switching costs could be the use of proprietary customer information (browsing and purchase history) by the site to improve the value of its offering to the customer (Alba et al. 1997, Bakos 1991). For example, a site can use e-customization tools, as some news and e-commerce sites do, to present tailored content to the customer. Finally, switching costs could arise from the additional investment of time and effort by the customer at the new site in order to receive a similar level of service as the current site (Alba et al. 1997). For example, to obtain movie recommendations from a new web site, a customer would have to rate some minimum number of movies before the site returns any recommendations.

Customer interaction with a website is also considered to be an important factor influencing customer retention in the Internet environment. Research using Internet clickstream data has used time spent per page, duration of visit and number of page views to describe consumer visit behavior and has found them to be significantly linked to repeat visits (Johnson et al. 2003).
We estimate our model of customer retention, conditional on acquisition, using clickstream data from the movies category of an Internet recommendation site. The web site provides visitors with a list of recommended movies to help them make a decision on which movies to watch. Nearly 20% of online customers in North America use the Internet to read reviews of television shows and movies before they make a decision about what to watch (Forrester Research 2002). In fact, decision aids like recommendation systems have been shown to help customers make better and more efficient decisions (Haubl and Trifts 2000).

A visitor to the Internet recommendation web site first selects a category such as movies from among a list of categories which includes books, music etc. He then gives preference ratings on a scale of 1 – 10 (1 is a low and 10 is high) to a list of movies. The recommendation algorithm of the web site then generates a recommended list of movies from movies not yet rated by the customer. The recommended movies list has the movie name and the predicted value of the customer’s preference rating. The predictions are generated using the weighted average of linked movies – the movies which were rated along with the movie of interest by other customers in the database (see Appendix A for a brief overview of the algorithm.) A customer has to give at least 25 movie ratings before he can receive the recommendation list.
We define acquired customers\(^1\) as those who have had a real opportunity to try the web site’s product i.e., the recommended movies list. Customers who give at least 25 ratings during their first visit get the recommended movies list during the first visit. Hence, in our sample we use only acquired customers and study the differences in the probability of return visit for the first versus subsequent return visits. Acquired customers can be divided into retained customers\(^2\) and not-yet-retained customers. Acquired but not-yet-retained customers have visited the website once with the site capturing their first visit information. Acquired and retained customers have visited the site more than once and the site had information about their first and subsequent visits. (We use visitor and customer interchangeably to refer to a person who visits the Internet site, gives ratings, and receives recommendations.).

Looking at past research in marketing, the possibility that customer retention probabilities might vary across first and subsequent return visits is not new to marketing. A similar notion was incorporated into new product introduction models in consumer packaged goods. These models linked the impact of consumer attitudes about product quality and product satisfaction to the movement of consumers from trial to initial repeat to subsequent repeat. Modelers considered it desirable to specify

---

\(^1\) A customer is defined as acquired when he visits the site for the first time through the movies category and gives at least 25 ratings – the minimum number of ratings needed by the recommendation engine to generate a list of recommendations.

\(^2\) A customer is defined as retained if he revisits the site once he is acquired and gives at least 1 rating during that visit in the movies category.
different classes of repeat users since greater depth of repeat was expected to lead to a higher probability of repurchase (e.g., Blattberg and Golanty 1978, Narasimhan and Sen 1983).

The possibility of selection bias in the modeling of customer retention when customer acquisition and retention are assumed to be independent processes was raised in past research (Thomas 2001). The model is estimated using a random sample of customers from the membership database of a service organization who must renew their membership on a yearly basis. Customer acquisition process is defined as that part of the customer-firm relationship that begins with the customers’ first interaction with the firm through the first purchase until the first repeat purchase. Customer retention process is defined as beginning with the first repeat purchase until the termination of the relationship. Thus, the first year of membership is the acquisition phase with indicator variables for product usage during the first year of membership and demographic dummy variables included in the selection equation. The selection equation captures the probability of being acquired i.e., renewing the membership (the first repeat purchase).

In another study (Reinartz et al. 2005) a modeling framework for resource allocation between acquisition and retention is developed in the context of a B-to-B high-technology product company. Customer acquisition process is defined as beginning when the firm contacts a prospect for potential acquisition and ending during the first quarter of the study period when the prospect is acquired – makes the
first purchase. The customer retention process begins with the first purchase and ends when the relationship is terminated. Further, the ability to separate marketing expenditures and actions between customer acquisition and retention is more straightforward in a B-to-B context but can become quite complex in B-to-C settings where mass communication forms a major part of the marketing expenditure.

In our research we define the customer acquisition process as beginning when a customer becomes aware of the website and terminating when a customer receives the firm’s product i.e., a list of movie recommendations during their first website visit. We define the customer retention process as beginning when a customer is acquired and terminating with the end of the study when the duration observation is right-censored. The firm, a startup, did not have a planned acquisition strategy and nor did they use any marketing interventions to drive acquisitions. In lieu of the lack of detailed information on the acquisition process – the likelihood of an individual (visitor or non-visitor to the web site) receiving the firm’s product – we were unable to incorporate it in our model.

Further, the main focus of this research is to study the impact of customer fit (how well a product meets a customers requirements) on customer retention. Therefore, we estimate probability of customer retention conditional on acquisition with a random sample of acquired customers who received the firm’s product (the recommendations list) during their first visit. These are customers who rated at least 25 movie ratings during their first visit – the minimum number of ratings needed by
the recommendation algorithm to generate the recommendation list. This is similar to the selection criterion used by Lewis (2004) where they use a sample of customers with meaningful transaction histories to model the impact of loyalty programs on customer retention. They use a sample of customers from an internet grocer who made at least two purchases during the first five weeks in the database and had at least 30 weeks of observations. They did not model the acquisition process but instead focused on customer retention.

Our modeling approach applies an extension of the Cox proportional hazard model for multiple events. In our case, return visits to the web site by acquired customers are defined as the events. Duration time is the number of days from the previous visit to the current visit or end of study period. The model allows us to estimate a different baseline hazard function for the first and subsequent return visits of the customers by treating them as part of a separate stratum. Specifically, we split the customer visits longitudinally by visit type – first visit and subsequent visits. To estimate the retention probabilities for the first stratum – first return visit – the model uses first visit information taken from all the acquired customers – retained and not-yet-retained. For stratum two – subsequent return visits – the model uses only subsequent visit information from acquired and retained customers.

This stratification allows us to explore variation in customer retention probabilities. For example, for a website, the probability of customer retention might increase as customers become more acquainted with a site and they learn to better use
and navigate the site (Johnson et al. 2003). We examine how customer retention probabilities change for first and subsequent return visits and capture the differential impact of the various factors across these visits. Further, customer retention is also expected to vary across individuals. We incorporate heterogeneity in the retention probabilities using a gamma distribution for the ‘frailty’ parameter of the baseline hazard given its flexibility (Heckman and Singer 1984).

Our study looks at several key covariates. We define ‘customer fit’ as the accuracy with which the site’s recommendation engine reproduces the customer’s stated preferences for a set of movies. We measure switching costs to the customer using the number of movie ratings given by a customer during a visit. We also capture a customer’s interactions with the site using the average time spent rating a movie and the number of categories within the site visited by the customer. We expect customer fit to have a positive effect on customer retention, with better fit leading to higher retention. For an Internet site, we expect customers with higher switching costs to display higher retention rates. We also expect customer retention to increase as the customer’s interaction with the site increases.

Our research did not seek to improve the existing recommendation algorithm of the site. Instead, we sought to quantify the impact on retention of the idiosyncratic quality of the site’s recommendations – customer fit. Specifically, we define customer fit as the accuracy with which the algorithm captures the stated preferences of each acquired customer for a set of movies. We measure accuracy as the mean absolute
deviation (MAD) between the stated and predicted ratings for a given customer. This approach follows earlier studies (e.g., Breese et al. 1998, Ansari et al. 2000). A greater MAD reflects a lower customer fit and should negatively impact customer retention. Thus, our study provides an opportunity to explicitly examine the connection between recommendation quality – as evidenced by accuracy – and repeat visitation to the site. This gives us the ability to open an important window into the relationship between product or service quality and customer retention.

In addition to investigating the impact of Internet product or service quality on retention, our research is also relevant to managers at Internet recommendation sites. In our case, the company offers two types of products to its clients – media preference reports and the recommendation engine tool. The reports provide analysis of customer preferences and media segmentations using information the company has gathered from its Internet site. The media reports need a deep and varied database of customer ratings to be relevant and useful. Therefore, the company needs to retain its existing online customers to ensure its database and recommendation engine are constantly updated. Since the company does very little advertising – most customers find the site through search engines and links from books, movies and music websites – it is important for the company to understand how best to retain customers who visit its Internet site.

Further, the Internet recommendation site is a working module for the company’s recommendation engine tool. The web site showcases the success and
popularity of this tool. Potential clients who may want to provide a similar tool on their websites would be more willing to invest in such a tool if it was shown to be successful in increasing customer retention and stickiness on their web sites through accurate predictions about the customer’s preferences. It is vital for the company to understand the drivers of retention so it can evaluate the benefits of investing in improving the recommendation engine and other aspects of the site.

We believe our research makes an important contribution by establishing a direct and quantifiable link between customer retention and its drivers such as customer level switching costs, customer fit and customer interactions. The quantification of the response coefficients would help managers to allocate their resources to those actions which are most effective in increasing customer retention.

2.2 Modeling Framework

Our data set contains customer visit history with one or more visits for each customer. In our hazard modeling approach, an event is defined as a return visit by a customer. A duration observation is the time (in days) between two adjacent visits for a customer. Repeat customers, who visited the site more than once, have multiple duration observations. Some of the duration observations are right censored. We use the hazard rate to summarize the retention behavior of the customers and to avoid the issues that arise when computing means and medians for censored data (e.g., Dekimpe and Morrison 1991). In the biostatistics literature, similar data are commonly encountered in clinical trials for treating chronic diseases. In clinical trials, a study
subject may experience a number of ‘failures’ that correspond to repeated occurrences of an event during his/her follow-up period.

The Cox proportional hazard model uses a non-parametric baseline hazard function and allows covariates to influence the hazard rate multiplicatively (Therneau and Grambsch 2000). The Cox model is estimated using a maximum likelihood approach on the partial likelihood which integrates out the non-parametric baseline hazard function from the overall likelihood function. This approach has been used extensively in the research literature especially in engineering, biostatistics and to some extent in economics and marketing (Seetharaman 2004, Heckman and Singer 1985). Using a non-parametric hazard curve removes any restrictive assumptions about the distribution of the hazard rate in the population especially if we do not have any prior theory to support one functional form over another (Dekimpe and Morrison 1991).

Prentice, William and Peterson (1981) (PWP) proposed an extension of the Cox model for analyzing multiple events data. They showed that this model can be analyzed by a partial likelihood principle similar to a single event Cox hazard model. The model treats each event (e.g., first return visit) as an independent and different stratum. In our model, we split the event history data, a priori, longitudinally by different visits – first versus subsequent visits. This allows the estimation of a separate hazard function for each stratum – first versus subsequent return visit – using duration observations of the customers in that stratum. The covariates affecting the hazard
function for the first return visit or stratum are those observed during the customer’s previous visit. A customer can belong to more than one stratum depending on how many visits he/she made to the website.

Failure to control for unmeasured differences among individuals (usually termed frailty in epidemiological context and residual heterogeneity in social sciences) can result in misleading inferences about temporal dependence and inconsistent parameter estimates (Vanhuele et al. 1995). Recent outcomes can be similar to earlier outcomes because of a causal link (temporal dependence) or because some individuals are more prone to specific outcomes (spurious temporal dependence due to population heterogeneity). Further, if there is heterogeneity, then the characteristics of the aggregate level hazard rate will not be representative of any individual (Morrison and Schmittlein 1980). Using a gamma distribution for heterogeneity, Lancaster (1985) proved the inconsistency of the maximum likelihood estimators when heterogeneity is ignored. Those who experience a return visit will tend to be those who were more at risk for a return visit. Without controlling for heterogeneity, the estimated return probabilities of further visits will be biased upward (Omori and Johnson 1993).

One of the popular heterogeneity models in event history analysis (Clayton and Cuzick 1985) assumes that the heterogeneity follows a gamma distribution with mean one and an unknown variance. The gamma distribution (Heckman and Singer 1984) is flexible as it can take on a variety of J-shapes or unimodal shapes. Simulation results suggest that estimators for the structural parameters are very sensitive to the
assumption of a specific heterogeneity distribution and non-parametric estimation has been used as a way to resolve this problem (Heckman and Singer 1984). In epidemiological research (Clayton 1988), emphasis has been on non-parametric specification of heterogeneity effects while in econometrics (Heckman and Singer 1985) there has been some concern with the non-parametric specification. Further, there is empirical evidence (Meyer 1990) that suggests identification problems when a non-parametric specification is adopted for both the hazard and the heterogeneity distribution; model fitting tends to converge to a degenerate, zero heterogeneity solution even for simulated data with known heterogeneity characteristics. Thus, a non-parametric specification for hazard should be coupled with a parametric heterogeneity specification or vice versa.

Model Specification

Our data have multiple duration observations as some customers returned to visit the website more than once. If we pool the duration observations from all return visits together and estimate one hazard function, we would be assuming each event is similar and the underlying process is the same across subsequent events. To study how customer retention varies between the first revisit and subsequent revisits we need to allow a separate hazard function. We therefore use the PWP extension of the Cox proportional hazard model for multiple events. Given the large number of duration points, we use a continuous-time model. Combining the duration observations into a
manageable number of time intervals needed for discrete time analysis would have entailed unnecessary loss of information.

In our model, we allow for unobserved cross-sectional heterogeneity in the hazard rates (retention rates) among customers by incorporating a customer specific heterogeneity term in the hazard function. We use a gamma heterogeneity distribution for the cross section of individuals in our sample. Thus, we use a non-parametric specification for the hazard and a parametric specification for heterogeneity.

**Model Equations**

The conditional hazard function\(^3\) for customer \(i\) in stratum \(j\) with event time \(t_j\)

\[
h_{ij}(t_j | v_i) = h_{0j}(t_j | v_i) \exp(\beta_j' x_{ij}) v_i
\]

where \(t_j\) is measured from the previous event time, \(h_{0j}(\cdot | .)\) is the conditional non-parametric baseline hazard, and \(\{v_i\}\) are independent gamma variables with unit mean and unknown variance, \(\sigma^2\). They vary across the customers but are assumed to be fixed across the different strata. \(x_{ij}\) is the vector of measured explanatory variables for the \(i\)th individual in strata \(j\). \(\beta_j\) is the vector of unknown parameters associated with the explanatory variables in the \(j\)th stratum. They are assumed to be the same for all customers but vary across the different strata. The problem is to estimate \(\sigma^2, \beta_j\) and \(h_{0j}(\cdot | .)\).

---

\(^3\) The hazard function is defined as the probability that an individual experiences an event (return visit) at duration time \(t\), measured from the previous visit, conditional on having survived (not experienced an event) until then.
The overall likelihood function with duration observations $t_{ij}$, covariates $x_{ij}$ and heterogeneity variable $v_i$ for customer $i$ is

$$L(\beta, \sigma^2) = \int_0^\infty \prod_{j=1}^J \prod_{t_{ij}=1}^T \left[ h_{0j} \exp(\beta'x_{ij})v_i \right]^{d_{ij}} \exp\left(-\left(h_{0j} \exp(\beta'x_{ij})v_i\right)t_{ij}\right) g(v_i) dv_i \quad (2)$$

where $d_{ij}$ is an indicator variable for censored observation, taking a value of 0 if duration is censored, and $g(v_i)$ is the gamma heterogeneity distribution across the $N$ customers in our estimation sample.

The Cox model uses the partial likelihood function, which integrates out the non-parametric baseline hazard function from the overall likelihood function. For posterior analysis we need to re-estimate the baseline hazard function. We use the Nelson-Aalen estimator for the baseline cumulative hazard given by

$$\hat{\Lambda}(t) = \sum_{i,j,t_i = 1}^{N} \frac{\Delta N(t_i)}{\bar{Y}(t_i)} \quad (3)$$

where $\bar{Y}(t_i)$ is the number of customers who are at risk\(^4\) for a return visit at time $t_i$ and $\Delta N(t_i)$ is the number of observed return visits or events that occurred precisely at time $t_i$.

**2.3 Data and Empirical Specification**

*Data Description*

We estimate our model using customer clickstream data from an Internet recommendation site. At the time of the study, the site had six categories – movies, movies, movies, movies, movies, movies.

---

\(^4\) Customers are considered at risk for a return visit if they have experienced the previous visit. They belong to the risk set for that return visit – set of consumers who are ‘at risk’ of experiencing that return visit.
music, books, magazines, TV and radio – in which the customers could rate items and receive recommendations in return. A typical customer visiting the site, for the first time, starts with any media category – say movies – and enters his preference ratings for a set of movies. The customer’s stated preference ratings are on a scale of 1 to 10 with 1 low and 10 high. A visitor must rate at least 25 movies before he receives a list of recommended movies with predicted ratings. At this point, we deem the customer to have been “acquired”. The customer can continue to give more ratings in the movie category or visit another category within the site. The site has no other information or product to offer to the visitor. The customer pays nothing for the recommendation but must spend time and effort rating the movies before receiving the recommendation list. The site uses a patented algorithm to generate the predicted ratings. (See Appendix A for a brief overview of the algorithm.)

The information given by the customer – ratings to movies – is recorded into the database. The algorithm which generates the predictions uses this information to create a list of recommended movies with predicted ratings, using movies in its database not yet rated by the customer. The recommended list of movies is displayed to the customer in descending order of particular value. If the customer visits the site again and rates more movies, he receives additional recommendations.

In this research, we focus on the movies category which accounts for 49.9% of all observed ratings and 43.3% of all unique customers to the site. Please refer to Table 2.1 for details. The data cover a two-year period from the website’s launch in
December 1999 until February 2002. The dataset includes anonymous customer id, demographic information such as age, gender and zip code, date of visit and ratings given to each movie with a time stamp. Since we have customer visit history data, with the date and time when each rating was recorded, we can create a time to event (visit) history data set. We define a visit by its calendar date and group all customer information as though occurring on a single session on that date.

In Figure 2.1, we look at the visit history of a typical customer to the site. Point A on the line is the date of his first visit to the site when he becomes an *acquired customer*. Point B is the date of his second visit or first return visit (revisit) to the site when he becomes a *retained customer*. Points C, D, E, F and G represent his subsequent revisits to the site. Point H is the last date in the study period or the censoring date. As the customer may revisit the site after a few days, weeks or months, we assume that the customer is at risk for a subsequent revisit until point H.

Customers to the website comprise two main groups. Group 1 comprises customers who have been acquired but not yet retained. They visited the website once in the study period, with the site capturing their first visit information. Group 2 comprises customers who have been acquired and retained. They have visited the site

---

5 A customer is defined as acquired when he visits the site for the first time through the movies category and gives at least 25 ratings – the minimum number of ratings needed by the recommendation engine to generate a list of recommendations.

6 A customer is defined as retained if he revisits the site once he is acquired and gives at least 1 rating during that visit in the movies category.
more than once with the site capturing information about their first and subsequent visits. First visit information of customers from both groups is used in the estimation of retention probabilities for stratum 1 – first return visit. Subsequent visit information from customers in group 2 is used for the estimation of retention probabilities in stratum 2 – subsequent return visits.

The website acquired 6009 customers between November 1\textsuperscript{st} 2000 and February 18\textsuperscript{th} 2002\textsuperscript{7} through the movies category. We used a random sample of 65\% of these customers (3881) as our estimation sample and kept the remaining for holdout. In the estimation sample the customers had 6169 visits giving 468,815 ratings. All of the 3881 customers in the estimation sample experienced the first visit. Thus, there were 3881 duration observations used for estimating the 1\textsuperscript{st} return visit hazard function. A subset of 711 customers revisited the site. These customers contributed to the duration observations used for estimating the subsequent return visit hazard function. Please refer to Table 2.2 for details.

Table 2.3 presents some key descriptive statistics for the estimation sample. For example, the mean number of visits in the sample was 5.28. Further, the customers on average returned to the site after every 4 months (mean=126.98 days).

\footnote{7 We assume the period from December 1999 to October 31\textsuperscript{st} 2000 is sufficient for the recommendation engine to have stabilized and exclude the visitors acquired during this time.}
Model Variables

All variables are defined using information captured by the website during a customer’s visit. We define the dependent variable \( time_{vis} \) as the duration or time to revisit calculated from the previous revisit for the customer. It is the number of days until the customer experiences a revisit or is censored. For the first revisit it is calculated from the first visit of the customer. We also use an indicator variable to show if the time to event is censored.

To predict the probability of a revisit, we use information from the customer’s previous visit as the independent variables. We define \( customer_{fit} \) index as the accuracy with which the recommendation algorithm captures the stated preferences of the customer for a set of movies. We measure accuracy using the mean absolute deviation between the actual and predicted ratings for a given customer. Please refer to Appendix A for details. Larger deviations between actual and predicted ratings for the customer should lead to lower customer retention. We expect \( customer_{fit} \) index to have a negative impact on customer retention as a larger index means a lower customer fit.

We define a set of control variables to capture switching costs and customer interaction which we believe will affect customer retention. We quantify switching costs and customer interaction using proxy measures. The number of ratings, given during a visit, captures the customer’s switching costs. The amount of time spent per
rating and the number of categories visited are used for measuring customer interaction with the site. The control variables are:

\( n_{rat} \): Number of movies rated by the customer during a visit. We expect it to have a positive impact on customer retention. An increase in the number of ratings reflects an increase in switching costs and should lead to greater customer retention.

\( n_{cat} \): Number of categories including movies in which the customer gave ratings during a visit. We expect it to also have a positive impact on retention. An increase in the number of categories visited by a customer reflects greater customer interaction and should lead to greater customer retention.

\( t_{rat} \): Average time spent by the customer in rating each movie during a visit. We expect this to have a positive impact on customer retention. An increase in the time spent by the customer on the site evaluating each movie is likely to reflect a greater level of interaction and should lead to higher retention rates. Please refer to Appendix B for details.

**Model Estimation**

We estimate five models using the approach described above. *Model 1*, a null model with no covariates, estimates a single hazard function for all customers by pooling information from all the visits. *Model 2* adds the control variables \( n_{rat}, n_{cat} \) and \( t_{rat} \) to the null model. *Model 3*, which includes the customer_fit index in addition to the control variables, captures the impact of customer fit on customer retention. *Model 4* adds the cross-sectional heterogeneity variable to Model 3.
Model 5 splits the customer visit information longitudinally by visit number – first and subsequent visits. The model estimates a separate hazard function for each return visit stratum. In stratum 1, the model estimates a hazard function for the first revisit using information from the first visit of all customers when they are at risk for their first return visit. In stratum 2, the model estimates the hazard function for subsequent return visits using subsequent visit information from customers. We consider two strata – first and subsequent return visits. This is because the number of observations in subsequent visits becomes progressively smaller in our data set.

Further, in Model 5, the heterogeneity distribution is fixed and does not change with each stratum. It captures only the cross-sectional heterogeneity in the customer retention rates among all the customers in the sample. Longitudinal heterogeneity differences are captured by the baseline hazard function which varies across each return visit stratum. Model 5 performed better on BIC scores than a model where the heterogeneity distribution was allowed to vary across the return visit strata. Finally, we consider Model 4 as the benchmark pooled model against which we compare the stratified Model 5.

Models with no heterogeneity fit a hazard function by maximizing the partial log likelihood function using the Newton-Raphson algorithm (Therneau and Gramsch 2000). In the models with the heterogeneity variable, an EM algorithm is used – estimating the heterogeneity variable at the E step and the regression parameters and
the hazard at the M step. We use the SPLUS software package for estimating the different models.

The Nelson-Aalen (N-A) estimator of the baseline cumulative hazard curve (Nelson 1969) is estimated at the mean value of the variables in each stratum. The variables which enter the model, such as customer_fit index variable, are mean-centered (i.e., mean value is 0). This ensures that the N-A estimator does not incorporate the impact of the covariates, which enter the actual hazard function multiplicatively as an exponential term.

2.4 Empirical Results and Discussion

We use the Bayesian Information Criterion (BIC)\(^8\) as a goodness of fit measure for choosing the best fitting model in sample. We also compared the performance of the models in the holdout sample in predicting the hazard or survival probability for each customer.

The hazard probability for a customer gives the conditional probability of experiencing a visit (or event) during an interval \((0, t]\) between two visits of a customer. The survival probability estimates the probability of not experiencing a visit in the same interval (i.e. the duration observation is right censored). We compare this with actual observations – a single visit or no visit in an interval – to get an estimate of

\(^8\)The BIC penalizes a model for complexity (large number of parameters) with a smaller value of the BIC denoting a better fitting model. We obtained the same results for model selection using other criteria typically used in research in marketing and biostatistics such as the Akaike information criterion (AIC), AICc which corrects for any small sample bias and adjusted BIC which uses number of uncensored events instead of number of observations in the penalty term (Buckland et al, 1997; Rust et al, 1995; Volinsky and Raftery, 1999).
the prediction error. We use a holdout sample of 2128 customers with 3075 observations.

Model 5 was the best fitting model with a BIC of 30717.26, a gain of 2877.50 points over the benchmark pooled model - Model 4 (See Table 2.4 for details). In the holdout sample, Model 5 had a mean absolute error (MAE) of 0.3691 and root mean squared error (RMSE) of 0.4618 compared to Model 4 with MAE of 0.4521 and RMSE of 0.4998. Thus, our results support the use of the stratified model over the pooled model.

We expect the customer_fit index variable to have a negative coefficient. The higher the customer_fit index (greater deviation between actual and predicted ratings reflecting a lower customer fit) the lower the hazard rate or the conditional probability of a return visit (retention rate) for the customer. Further, we expect the impact of the customer_fit index on retention rate to vary between first and subsequent return visits as evidenced by different coefficient values and significance levels in the two strata. Examining Table 2.5, in stratum 1 – representing the first return visit – the customer_fit index has a significant effect on the retention rate for the first return visit with a coefficient value of -0.0705 (t = -3.16). This corresponds to a retention ratio\(^9\) of approximately 0.932 (i.e., an increase in the customer_fit index by 1 unit leads to a

\(^9\) The retention ratio is the ratio of the retention or hazard rates when the value of a variable is changed, all else equal. For customer_fit index, the retention ratio = exp (coefficient) and represents the change in the retention rate for an increase in the index by 1 unit. For variables with log transformations, it represents the change in the retention rate when the variable value is increased K times i.e., retention ratio = exp [ ln (K) * coefficient ].
decrease in the retention rate by 6.8%). In stratum 2, representing the subsequent revisits, it is not significant but has the correct sign with a coefficient of -0.0275 (t = -1.47). This corresponds to a retention ratio of approximately 0.973. Estimating a single hazard function would not have captured the differential impact of customer fit on the retention rate of newly acquired versus repeat customers.

The results suggest that customers evaluate the quality of the site’s recommendation engine during their first visit or acquisition date but not subsequently. This is relevant from a managerial perspective as the company needs to ensure that the quality of its recommendations is high during the initial contact with the customer. Thus, customer fit during the first visit significantly affects the customer’s probability of next visit, but customer fit on subsequent visits is not significantly associated with further return visits.

We expect \( n_{rat} \), the number of ratings given by the customer, to have a significant positive impact on customer retention. It captures the effort investment made by the customers and would be expected to lead to an increase in their switching costs. We expect this impact to decrease over subsequent visits as the contribution of each additional rating to the total switching cost declines. We find that in stratum 1, \( \ln(n_{rat})^{10} \) had a significant positive coefficient value of 0.3212 (t = 7.58). This corresponds to a retention ratio of 1.249 or 24.9% increase in the retention rate when

\[ ^{10} \text{The control variables enter in the log form as it resulted in a better fitting model when compared to linear or quadratic functional forms.} \]
the number of ratings is doubled. In stratum 2, the effect is reduced (coefficient 0.0831) but continued to be positive and significant ($t = 3.99$) with a retention ratio of 1.059 or 5.9% increase in the retention rate when the number of ratings is doubled. Again, using a model with a single hazard function would not have revealed this differential impact.

The results show that switching costs are relevant in the Internet world – more so than anticipated by early research (Alba et al, 1997, Bakos, 1991, Bakos, 1997). In the context of electronic marketplaces, past research expected the advantages to a site, from high switching costs of buyers, to be reduced by advances in technology, increased competition and lower search costs. But the dynamic interaction between the customers and the web site can also lead to higher switching costs from the use of proprietary customer knowledge available to the site. Customer history stored with the site through clickstream and transaction data can be used by web sites to improve customer experiences through the use of e-customization and recommendation tools (Ansari and Mela 2003). These tools are likely to work best with some minimum threshold level of customer information and improve with greater usage. A customer moving to a new site would experience switching costs as he would have to rebuild his history to get similar levels of customization and service. Thus, managers can influence switching costs on the Internet by using e-customization and recommendation tools and encouraging their customers to build an extensive history with their site.
We expect \( t_{rat} \), average time spent per rating during a visit, to have a significant positive effect on the retention rate. Time spent acts as a proxy for customer interaction with the site. Greater interaction is expected to lead to greater probability of return. In stratum 1, \( \ln(t_{rat}) \) had a coefficient value of 0.3837 (\( t = 6.63 \)). This corresponds to a retention ratio of 1.305 or 30.5% increase in the retention rate when the \( t_{rat} \) is doubled. In stratum 2, the coefficient value of 0.2313 is lower but still positive and significant (\( t = 3.73 \)) which corresponds to a retention ratio of 1.174 or 17.4% increase in the retention rate.

Similarly, \( n_{cat} \), the number of categories in which the customer gave ratings, a proxy measure for customer interaction, is expected to have a positive impact on the retention rate. In stratum 1, \( \ln(n_{cat}) \) has a coefficient value of 0.2896 (\( t = 4.18 \)) which corresponds to a retention ratio of 1.222 (a 22.2% increase in the retention rate when the number of categories is doubled). It does not have a significant impact (\( t = 1.19 \)) but is correctly signed in stratum 2 with a coefficient value of 0.0952.

The impact of \( t_{rat} \) and \( n_{cat} \) on customer retention illustrates the dynamic nature of the customer interactions with the web site. The impact of customer interactions varies across the two strata – first and subsequent return visits. Initial interactions during the first visit significantly influence the customer’s retention probabilities. The web site can manage these interactions by making the site more attractive to customers and thus, encouraging them to spend more time browsing the site. The website can also influence customer retention by encouraging its online
customers to explore more than one category. Managers can give incentives to customers during their first visit to explore other categories like books, music or TV within the website. They should concentrate their efforts during the first visit for maximum impact using, for example, cross promotions such as online banners.

Our findings support past research on determinants of customer retention and show that factors like customer fit, customer interactions and switching costs have a significant impact on customer retention in the expected direction. We add to the previous research by quantifying the impact of the above determinants. Further, this impact was shown to vary with first and subsequent return visits of the customer. The factors had a maximum impact during the first visit of the customer with the effect declining or becoming insignificant over subsequent visits. This highlights the need to manage the customer’s first impression of the site and focus marketing efforts during this visit to maximize returns.

Figure 2.2 shows the baseline hazard functions estimated for each stratum at the actual duration time points in the data set. The retention probabilities decline as the duration time (number of days from the previous visit) increases. Further, in stratum 2, the values of the retention probabilities are in general are higher than those observed in stratum 1. This reflects a higher hazard rate or conditional probability of a return visit for customers in stratum 2. Thus, retention rates, for a given duration point, change across return visits with rates being higher for stratum 2 – subsequent return visits than for stratum 1 – first return visit. Figure 2.2 demonstrates that customer
retention probabilities vary across first and subsequent return visits and underscores the need for estimating the hazard function for the two strata separately.

2.5 Managerial Implications

Customer retention has been shown to have a significant impact on a company’s performance outcomes such as share price, market value and customer lifetime value (Gupta et al. 2004). Further, results from our study show that customer fit and other determinants such as switching costs and customer interactions have a significant impact on customer retention. We can examine the impact of changes in their value (e.g., an increase in customer fit) on customer lifetime value. Thus, managers can use factors under their control, such as improving the accuracy of the recommendation engine, to influence customer and firm value.

We quantify the lifetime value of a potential customer to the site using the average number of ratings that a typical customer will contribute to the web site’s database. The duration of the customer’s life with the site is captured by the two revisit strata. Thus, customer lifetime value is the expected number of ratings, conditional on revisit, that a customer will contribute averaged over the two revisit strata. In our case, the customer lifetime value is 53.7. An increase in the customer fit by 1 unit (i.e., a decrease in the customer_fit index by 1 unit) leads to an increase in

\[ \text{Expected number of ratings per customer for stratum } j = \text{Cumulative hazard rate } \times \text{Number of ratings per customer, where the cumulative hazard rate is calculated at mean duration time and mean value of the covariates for the stratum.} \]

\[ \text{Recall that an increase in ratings adds value to the reports sold to media and research companies.} \]

11 Recall that an increase in ratings adds value to the reports sold to media and research companies.

12 Expected number of ratings per customer for stratum } j = \text{Cumulative hazard rate } \times \text{Number of ratings per customer, where the cumulative hazard rate is calculated at mean duration time and mean value of the covariates for the stratum.}
the customer lifetime value by 4.4%. Thus, customer fit has a significant impact on
customer lifetime value for the web site.

We calculated the customer fit elasticity for a similar change in customer fit. A
1 unit increase in customer fit (i.e., a 1 unit or 29% decrease in the customer_fit index
from its mean value) leads to a 5.05% increase in the retention rate, i.e., a customer fit
elasticity of 0.174. The estimate is potentially useful to managers in resource
allocation decisions across a number of marketing mix variables.

Empirical research on the impact of customer fit or product quality is thin due
to the difficulty in quantifying product value or quality (Hanssens et al. 2001). An
overall quality elasticity of 0.52 (Lambin 1976) was found for products such as soft
drinks and yogurt. The customer fit elasticity estimate from our study is therefore
about one-third of what has been found in previous research on product quality.

We also examined the performance of the two models – pooled and stratified –
in targeting customers with marketing communications. A potential marketing
intervention by the site managers is a pop-up during a customer’s visit to the web site,
encouraging them to explore other categories within the site. If customers who click
on the pop-up visit more categories, it will have a positive impact on their return
probabilities. The extent of the impact is not expected to be uniform across all
customers as it would depend on the customer specific hazard rate which is a function
of the baseline hazard rate at the customer’s duration observation, customer specific
covariates and heterogeneity value.
In our data sample, there are two groups of customers. Group 1 customers have visited the web site once and are at risk for their first return visit. Group 2 customers have returned at least once and are risk for subsequent return visits. The pooled model treats customers in groups 1 and 2 as equivalent (i.e., they have the same return visit probabilities irrespective of their return visit number) while the stratified model treats them as two different groups.

Using the parameter estimates from each model we calculate the probability of return visit for each customer. We further calculate the change in the probability of the return visit if all customers were targeted with a pop-up, encouraging them to explore an additional category within the site\textsuperscript{13}. We sort all the customers in the sample in descending order of the impact of the pop-up on their return probabilities.

The benefit and cost to the web site from the pop-up is measured in terms of the average number of movie ratings expected per customer if they returned to the web site for the next visit. If a customer clicks on the pop-up it has a positive impact on the return visit probability resulting in positive benefits to the web site. On the other hand, if the customer is lost due to annoyance with the pop-up it leads to negative benefits or costs to the web site. We calculate the average expected payoffs per customer for the

\textsuperscript{13} We assume that 10% of customers click on the pop-up ad and visit an additional category such as books or music. We further assume 10% of customers are lost due to annoyance with the pop-ups.
top 10% and optimum\textsuperscript{14} number of customers for the two models. An examination of Table 2.6 shows the average expected payoffs per customer is higher for the stratified model than for the pooled model. The optimum number of customers is much larger for the pooled model than the stratified model – resulting in greater differences in payoffs. Thus, the stratified model allows managers to sort customers better and identify more effectively those customers with higher expected payoffs from a targeted activity.

2.6 Conclusions

Using an extension of the Cox proportional hazard model for multiple events, we establish a direct and significant quantifiable link between customer retention and customer fit, as well as other determinants such as customer interactions and switching costs. Further, we show that this impact varies with the depth of repeat. Our results suggest that customer fit during the first visit significantly affects the customer’s retention probability. We also find a significant positive impact of switching costs on customer retention, indicating that switching costs continue to be relevant in the Internet world. Our findings also point to the relevance of a customer’s interaction with the site. Initial interactions, especially during the first visit, can significantly influence customer retention.

\textsuperscript{14} Optimum number of customers is defined as the number of customers when incremental cost is equal to incremental benefit.
The various factors we modeled had a maximum impact during the customer’s first visit or acquisition stage with the effects declining or becoming insignificant over subsequent visits. Our results are in line with the acquisition process perspective which states that acquisition includes, in addition to the first purchase, other non-purchase encounters that precede and follow purchase up until the time the customer makes a repeat purchase (Blattberg et al. 2001). During this acquisition period, a bonding or development stage for the customer-firm relationship, the customer forms attitudes about the firm’s product and ancillary services which affect the customer’s repurchase decision.

In the case of the Internet recommendation site, acquisition would be defined as the customer visiting the site and interacting with the site through rating items and receiving recommendations in return. Managers can impact the customer retention rates during this stage through improvements in the accuracy of their recommendation engine, increasing customer interactions by cross promoting other categories within the site and providing additional value to the customers using their information history with the site.

To make optimal resource allocation decisions among the various instruments available to improve customer retention, managers need to know the impact of each of the above variables on customer retention probabilities. We believe our research, through its estimates of the response parameters, their elasticity values and individual
level hazard rates, should help managers make these decisions more effectively and improve profitability.

We also found that customer fit, one of the dimensions of perceived quality, has a significant impact on customer lifetime value for the website and it may be in the site’s long term interest to improve the accuracy of its recommendation tool, subject to cost considerations. The estimated customer fit elasticity for the Internet website of 0.174 reflects the importance of “product quality” in a company’s marketing mix.

In operations management (Garvin 1988, Juran 1988), one of the primary dimensions of quality is customer fit. Expected future quality has been considered critical to customer satisfaction and retention as it relates to long-term relationships with customers (Czepiel and Gilmore 1987). In line with the past literature, our research shows that managing customer fit is important for a company. The company is likely to benefit from investing resources in efforts to improve its product quality which, in this case, is the accuracy of the website’s recommendation system.

Findings from our current research can be generalized to other online and offline contexts where customer retention is important. For example, web sites can increase customer loyalty by using e-customization and recommendation tools as they increase the switching costs for their customers through value additions. The sites can improve their retention rates by investing in better levels of customer fit by improving the quality of their product offering. They can also increase customer retention probabilities by increasing the customer interactions by using online promotions and
highlighting sub-sections within the site. This should encourage customers to spend more time browsing the site and also visit other sections.

There is scope for further research to better understand the various issues involved in understanding customer retention. First, it would be of interest to capture the heterogeneity in the parameters that may arise from customer specific differences in responses to the various factors outlined above. Our current model captures heterogeneity at the intercept level but assumes fixed values of the response parameters. Second, by stratifying on each revisit we do not allow for dependencies among the hazard functions in each stratum. We assume each revisit is from a different, independent process. It may be helpful to capture any such dependencies within the modeling framework. Third, we have looked at the variation in retention probabilities between first and subsequent return visits. We pooled visit information from the second return visit onwards to estimate a single hazard function for subsequent return visits. It would be of interest to explore how retention probabilities vary with depth of repeat. Fourth, given the lack of information on the firm’s acquisition process, we do not model the acquisition process. It may be helpful to check for any bias that could arise if the acquisition and retention process were not independent.
2.7 Appendices

Appendix A: Creation of Customer Fit Index Variable

The recommendation algorithm used by the site computes a customer’s preference ratings for a movie using ratings given by the customer for a set of movies. A table is created with a similarity index or coefficient for all possible pairs of movies on a scale of -1 to 1 with -1 a low and 1 a high. Movie pairs given similar ratings by a customer have a higher index than pairs with dissimilar ratings. The prediction for a movie, say Movie A, not seen by the customer, is a weighted average of the movies rated by the customer with weights reflecting how similar these movies are to Movie A. Adjustments are also made to the prediction using the average rating of Movie A by other customers in the database.

Our dataset did not include the actual list of movies recommended to the customers with their predicted ratings. We have for each visit of a customer the total list of the movies he rated. For each customer $i$, we generated a predicted rating for a movie $j$ by running the site’s algorithm using ratings from all movies in his list except movie $j$ as well as information stored in the database about the ratings given by all customers until this date.

**Step 1:** Create pairs for movie $j$ with all other movies $k$ from the list of movies rated by customer $i$. 

46
Step 2: Calculate the average similarity coefficient for each pair \((j,k)\) using every customer \(c\) in the database until this date, other than customer \(i\), who gave ratings to both movies. The average is calculated by dividing the sum of the similarity coefficients for the pair \((j, k)\) over all NC customers by \(p\) – the number of customers contributing to the sum. The average similarity coefficient is then weighted by \([(p-1)/(p+1)]\) i.e., greater weight is given to rating pairs that are more numerous. Average weighted coefficient for each pair \((j,k)\) is given by

\[
AVGCOEF(j,k) = \frac{\text{SUM}(j,k)}{p} \left( \frac{p-1}{p} \right) \left( \frac{p+1}{p} \right),
\]

where \(\text{SUM}(j,k) = \sum_{c=1}^{NC} \text{Coefficients}(j,k)_c\)

If \(AVGCOEF(j,k)>0.3\) the value is retained. If the value falls below 0.3, it is set equal to 0. Thus, the algorithm gives greater weight to rating pairs that are more similar.

Step 3: For customer \(i\), predicted rating for movie \(j\) is

\[
predrat(i,j) = \frac{\sum_{k} AVGCOEF(j,k) \times \text{rat}(i,k)}{\sum_{k} AVGCOEF(j,k)} + \frac{\sum_{c=1}^{NC} \text{rat}(c,j)}{NC} + \frac{\sum_{k} AVGCOEF(j,k) \times \left( \sum_{c} \text{rat}(c,k) \right) / NC}{\sum_{k} AVGCOEF(j,k)}
\]

, where \(\text{rat}(i,k)\) is the rating given by customer \(i\) to movie \(k\) , \(\text{rat}(c,j)\) is the rating given by customer \(c\) to movie \(j\) , \(\text{rat}(c,k)\) is the rating given by other customers \(c\) to movie \(k\), \(NC\) is the number of customers who rated movie \(j\) and \(NK\) is the number of movies \(k\) rated by the customer during a visit.
The customer fit index captures the accuracy of the recommendation engine in predicting the customer’s stated preference ratings for a movie. Our measure is the mean absolute deviation (MAD) between the actual and the predicted ratings for the set of movies rated by the customer during his visit (Breese et al. 1998, Ansari et al. 2000).

The algorithm failed to generate a prediction when there were less than two linked movies – movies rated by other customers along with the movie of interest. We believe the failure of the algorithm to give a predicted rating reflects its inability to meet the needs of customers with idiosyncratic list of rated movies. Thus, the web site’s product failed to meet the requirements of such customers. If the algorithm failed to generate a prediction for a movie $j$, we assigned the observation an absolute deviation value of 9 – reflecting the lowest customer fit value. The model results remained essentially unchanged for other values such as 4, 5, 6, 7 or 8.
Appendix B: Creation of Control Variable $t_{rat}$ – Average time spent per rating

In the data set, the time stamp information on a rating was in batches with some of the adjacent ratings having the same time stamp. We do not know the exact time when the customer gave the first rating. We only have information about when the first rating and/or adjacent ratings were processed by the computer. Hence, we calculate the visit length from the first time stamp and use adjusted number of ratings (all ratings other than those in first time stamp) to calculate average time spent per rating, $t_{rat}$ for a customer during a visit.

Some simplifying assumptions that were made include

1. We assume that all the ratings given during a calendar day belong to one visit defined by that date. If there are gaps in between two adjacent ratings greater than 120 seconds we set it equal to 120. This takes care of gaps in the ratings when the customer may have left the computer or engaged in other activities.
2. If the visit session continues over midnight than it is assumed to be part of the previous day and it is counted as one visit and not two visits and the rating date is adjusted accordingly.
3. Missing values for $t_{rat}$ for any movie by a customer were set equal to the average time spent per rating by this customer during this visit. If all values for a visit are missing – if the customer gave very few ratings which belonged to the first time stamp – then we set them equal to the average time spent per rating by the customer across
all his visits. If the customer had only one visit and all values were missing then we set them equal to the average time spent per rating across all customers.
Table 2.1: Number of Customers and Ratings for All Categories

<table>
<thead>
<tr>
<th>Categories</th>
<th>Number of Ratings</th>
<th>Percentage of Total Ratings</th>
<th>Number of Customers</th>
<th>Percentage of Total Customers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movies</td>
<td>1,067,007</td>
<td>49.91</td>
<td>12,410#</td>
<td>43.36</td>
</tr>
<tr>
<td>Books</td>
<td>550,933</td>
<td>25.77</td>
<td>12,889</td>
<td>45.03</td>
</tr>
<tr>
<td>Music</td>
<td>399,010</td>
<td>18.66</td>
<td>9,206</td>
<td>32.16</td>
</tr>
<tr>
<td>TV</td>
<td>115,653</td>
<td>5.41</td>
<td>1,989</td>
<td>6.95</td>
</tr>
<tr>
<td>Magazines</td>
<td>4,642</td>
<td>0.22</td>
<td>187</td>
<td>0.65</td>
</tr>
<tr>
<td>Radio</td>
<td>646</td>
<td>0.03</td>
<td>104</td>
<td>0.36</td>
</tr>
<tr>
<td>Total</td>
<td>2,137,891</td>
<td>100</td>
<td>28,622*</td>
<td></td>
</tr>
</tbody>
</table>

#This includes customers not unique to movies category.

*Total number of unique customers to the website.
Table 2.2: Distribution of Customers by Number of Visits

<table>
<thead>
<tr>
<th>Visit</th>
<th>Number Of Customers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3881*</td>
</tr>
<tr>
<td>2</td>
<td>711*</td>
</tr>
<tr>
<td>3 and above</td>
<td>324</td>
</tr>
</tbody>
</table>

*Number of customers with at least 2 visits – those at risk for 2nd return visit (revisit)
### Table 2.3: Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Median</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of visits</td>
<td>5.28</td>
<td>13.98</td>
<td>1</td>
<td>1</td>
<td>108</td>
</tr>
<tr>
<td>Time to revisit (days)</td>
<td>126.98</td>
<td>124.60</td>
<td>1</td>
<td>94</td>
<td>475</td>
</tr>
<tr>
<td>Number of ratings per visit</td>
<td>75.99</td>
<td>96.86</td>
<td>1</td>
<td>46</td>
<td>1119</td>
</tr>
<tr>
<td>Number of categories per visit</td>
<td>1.36</td>
<td>0.66</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Time spent per rating per visit (seconds)</td>
<td>21.79</td>
<td>15.15</td>
<td>0.76</td>
<td>16.92</td>
<td>108</td>
</tr>
<tr>
<td>Customer_fit index per visit</td>
<td>3.45</td>
<td>1.93</td>
<td>0.02</td>
<td>3.25</td>
<td>9</td>
</tr>
</tbody>
</table>
### Table 2.4: Model Comparison in Estimation Sample

<table>
<thead>
<tr>
<th>Model Number</th>
<th>Model Description</th>
<th>No. of parameters</th>
<th>LogL</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>Null Model</td>
<td>1</td>
<td>-19299.21</td>
<td>38607.15</td>
</tr>
<tr>
<td>Model 2</td>
<td>Control variables only</td>
<td>4</td>
<td>-18703.34</td>
<td>37441.59</td>
</tr>
<tr>
<td>Model 3</td>
<td>Model 2 + customer_fit index</td>
<td>5</td>
<td>-18496.18</td>
<td>37036.00</td>
</tr>
<tr>
<td>Model 4</td>
<td>Model 3 + heterogeneity</td>
<td>6</td>
<td>-16771.20</td>
<td>33594.76</td>
</tr>
<tr>
<td>Model 5</td>
<td>Model 4 with stratification (2 strata)</td>
<td>11</td>
<td>-15310.63</td>
<td>30717.26</td>
</tr>
</tbody>
</table>
Table 2.5: Parameter Estimates for Stratified Model (Model 5)

<table>
<thead>
<tr>
<th>Strata</th>
<th>Parameter</th>
<th>Coefficient</th>
<th>t - statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratum 1</td>
<td>ln(n_rad)</td>
<td>0.3212</td>
<td>7.58</td>
</tr>
<tr>
<td></td>
<td>ln(n_cat)</td>
<td>0.2896</td>
<td>4.18</td>
</tr>
<tr>
<td></td>
<td>ln(t_rat)</td>
<td>0.3837</td>
<td>6.63</td>
</tr>
<tr>
<td></td>
<td>Customer_fit index</td>
<td>-0.0705</td>
<td>-3.16</td>
</tr>
<tr>
<td>Stratum 2</td>
<td>ln(n_rad)</td>
<td>0.0831</td>
<td>3.99</td>
</tr>
<tr>
<td></td>
<td>ln(n_cat)</td>
<td>0.0952</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>ln(t_rat)</td>
<td>0.2313</td>
<td>3.73</td>
</tr>
<tr>
<td></td>
<td>Customer_fit index</td>
<td>-0.0275</td>
<td>-1.47</td>
</tr>
<tr>
<td>Heterogeneity</td>
<td>Variance of random effect</td>
<td>3.23</td>
<td>90.95</td>
</tr>
</tbody>
</table>
Table 2.6: Model Performance in Customer Targeting

<table>
<thead>
<tr>
<th></th>
<th>Pooled Model</th>
<th>Stratified Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of Customers</td>
<td>Average Expected Payoff per Customer@</td>
</tr>
<tr>
<td>Top 10% of Customers#</td>
<td>388</td>
<td>2.919</td>
</tr>
<tr>
<td>Optimum*</td>
<td>1612</td>
<td>0.764</td>
</tr>
</tbody>
</table>

# Customers ranked on the basis of maximum impact of an intervention on their hazard rate

* Optimum number of customers when incremental cost = incremental benefit

@ Measured in terms of the average number of movie ratings expected per customer in the next visit
Figure 2.1: Illustration of a Visitor’s Event History

- A: Date of 1st Visit
- B: Date of 1st Revisit
- C: Date of 2nd Revisit
- D: Date of 6th Revisit
- E: Last Date of Study
Figure 2.2: Baseline Hazard Function by Stratum
2.8 References


and Outstanding Problems,” *Statistics in Medicine*, 7, 819-841.


Forrester Research (2002). *Retail and Media North America*.


Lewis, Michael (2004), “The Influence of Loyalty Programs and Short-Term


CHAPTER 3

IMPROVING THE DIAGNOSIS AND PREDICTION OF CUSTOMER CHURN: A HETEROGENEOUS HAZARD MODELING APPROACH

Abstract

The authors develop a hazard modeling approach to predict customer churn and to study the nature of the empirical link between customer churn and factors such as customer service experience, failure recovery and payment equity. The approach uses a latent class Weibull hazard model with time-varying covariates. The model incorporates heterogeneity in both baseline hazard probabilities as well as response parameters. The authors apply the model to the churn prediction problem at a continuous service provider – a direct-to-home satellite television firm based in a South American country.

The empirical results show that the prediction of customer churn is significantly improved when heterogeneity is added to the customer churn rates and to the response parameters. Significant links are found between churn rates and variables capturing customer service experience, failure recovery efforts, and payment equity. Results also show important differences in the magnitude and significance of the response parameters across latent classes.

Keywords: Customer Churn, Hazard Models, Customer Satisfaction, Customer Heterogeneity.
3.1 Introduction

The importance of understanding customer churn has been highlighted in a number of recent studies. A 1% increase in retention rate has been shown to increase firm value on average by 5% (Gupta et al. 2004). A reduction in churn by 5 percentage points has been shown to double profits in some industries (Reichheld 1996). A critical managerial issue is therefore how to reduce churn rates, especially in industries where the annual rate is quite high (e.g., U.S. long distance telephone (30%), cell phone (30%), Internet service providers (22%), Thomas et al. 2004, Bolton 1998). In order to do this, managers need to be able to predict customer churn and establish links between customer churn and factors under their control.

One direction for improving churn prediction lies in modeling the unobserved heterogeneity in churn rates using duration or hazard models (e.g., Bolton 1998). While this is an important step, differences in churn rates across customers are unlikely to be captured completely by differences in baseline churn rates. We also expect differences to exist in how customers respond to changes in the different drivers of churn. Thus, a key goal of our research is to improve the prediction of customer churn by incorporating heterogeneity not only in a baseline churn function but also in the response parameters.

We use a latent class Weibull parametric hazard model as the baseline hazard function which allows the hazard function to be a smoothly increasing function of duration time with the heterogeneity incorporated at the segment level. The hazard rate
gives us a measure of the churn rate at any given duration time since it is the conditional probability of an event given that it has not yet occurred. In our case, an event is deemed to occur when a customer ends his/her subscription to the firm’s service.

As a method for analyzing duration data, hazard models are well established in the statistics literature. They have been shown to be well suited for analysis of duration data and superior to other methods, such as logistic and least squares regressions, in terms of stability, face validity, and predictive accuracy (e.g., Helsen and Schmittlein 1993). Of course, they also are hardly new to the marketing or CRM literature (e.g., Jain and Vilcassim 1991, Bolton 1998).

Though the importance of incorporating heterogeneity has been well established in the empirical choice modeling literature, hazard modeling papers in customer retention have so far looked only at heterogeneity in the baseline or intercept parameters. They have not incorporated heterogeneity in the response parameters of the hazard or churn function. The idea of a hazard model with heterogeneous response parameters is not new to marketing (e.g., Wedel et al. 1995) but, to the best of our knowledge, it is new to the CRM area of application.

Our research study was conducted with the cooperation of a direct-to-home satellite TV provider operating in a South American country with an emerging market economy. The firm provided its customers with digital quality TV channels showing content such as movies, sports and sitcoms through its network of satellites.
Customers subscribed to the firm’s service and paid a fixed amount for a specific content package every month. Customers were said to have churned when they stopped their subscription. The customer could stop their subscription for a number of reasons including: switching to another service provider, going without TV service, shifting residences (and possibly staying with the same service provider), or leaving the country. It is important to note that churn rates may be overstated when they are calculated using only a customer billing system as a customer database. For instance, a customer who relocates without terminating the service may be recorded as both a churned customer and then as a newly acquired customer.

One of the main issues facing this firm was how to reduce its annual customer churn rate of approximately 25%. The rate of customer churn is an important metric for managers because it is a critical input to assess customer lifetime value. To effectively plan interventions to reduce the customer churn rates, managers need to obtain both an accurate prediction of the customers’ churn rate and empirical estimates of the impact of different drivers of customer churn. At present, there is still little empirical evidence in the research literature explicitly linking customer churn to different factors (Gupta and Zeithaml 2004). Our study focuses on improving the prediction of customer churn and seeks to empirically link it to different factors that are expected to affect it.

Factors that we examine in this study include customer service experience (the customer’s evaluation of the firm’s customer service unit), failure recovery (the
customer’s evaluation of the firm’s recovery efforts when a service failure occurs) and payment equity (the customer’s perception of the fairness of the cost-benefit tradeoff). All of these factors may be expected to affect customer churn because of their impact on customer satisfaction.

An added dimension of our research is that we estimate our model using information readily available to most managers of large-scale subscription-type services. We use secondary data from a customer relationship management system. This system tracks specific transactions, interactions, and demographic data for each customer. Given our model allows us to capture the impact of time-varying covariates on the probability of churn, we can use the above information more effectively to build modeling constructs identified in the literature as impacting customer churn that have typically been measured with survey and experimental data. One goal of our research is to help managers use existing data sets more effectively, given that their systems already collect and store it on an ongoing basis.

In our study we also look at how segmentation affects the diagnosis and prediction of churn. As mentioned above, we expect differences in the churn rates among customers not just in the base levels but also in the customer’s responses to changes in the different drivers of churn. For example, in our application, the income distribution is skewed with the top 10% of the population accounting for nearly 50% of total income. Also 80% of the population is concentrated in just a few urban locations. Historically, the firm has segmented its customers *a priori* on the basis of
the demographic profiles of the customers. But given the premium nature of the service, 80% of the firm’s customers were classified as belonging to the top two socioeconomic groups in the country (23% of urban households). The skew in customer base towards the higher socioeconomic groups can limit the ability of conventional demographic segmentation to aid in churn prediction and diagnosis. We hoped to show management improvements in prediction from applying our proposed approach to the problem.

3.2 Modeling Approach

We apply existing techniques from the hazard modeling literature to predict and diagnose customer churn. In the data set, we confront the common problem of having the duration time (i.e., number of months from the time the customer subscribed to the firm’s service until he/she churned) right censored. Thus, for customers who were still active (i.e., had not churned) by the end of the study period, we do not have full information about their duration time. This makes a hazard modeling approach a natural solution as it allows us to get a measure of the customer churn rate. The hazard rate at duration time $t$ is the conditional probability of an event given that it has not occurred until time $t$. In this case, an event is deemed to occur when a customer ends his/her subscription.

The firm’s customer management software provided information that included demographic and socioeconomic variables as well as billing and transaction variables. Customer specific variables (e.g., the method of payment used by the customer)
remain fixed for the entire duration of the customer’s relationship with the firm. Variables which capture the customer’s billing and transaction history (e.g., invoiced amount) are time-varying covariates as they change every month during the customer’s relationship with the firm. The data sample and the variables used for estimating the model are described in more detail in the following section.

In our application, customer duration times are split into customer-month observations. Thus, we will be able model the impact of time varying covariates on the hazard function for each month from subscription until churn or censoring. The sample data was organized with each observation corresponding to a unique customer-time period $it$ (i.e., observation $t$ for customer $i$). For example, if customer $i$ churned at the end of three months from initial subscription then this is represented by three observations in the data where the first two contribute to the survival probability in the likelihood and the third contributes to the density. Observations from customers who did not churn before the ending time of the study period contribute only to the survival probability.

We use a latent class (heterogeneous) Weibull parametric hazard model as the baseline hazard function. The Weibull distribution gives a smoothly increasing function of duration time and the heterogeneity is incorporated at the segment level. Exploratory analysis showed the churn rate to be an increasing function of duration time and the Weibull model to be the best fitting model among several other parametric hazard models. For each customer segment, the model gives an estimate of
the baseline churn rate as well as the impact of covariates. The Weibull hazard probability (conditional probability of churn) of customer $i$ at duration time $t$ conditional on $i$ belonging to segment $j$ is given by

$$
h_j(t \mid x_{it}, i \in j) = h_{0j}(t) \exp(x_{it} \beta_j), \quad (1)
$$

where $h_{0j}(t) = \left(\frac{1}{\sigma_j}\right)^t t^{-\frac{1}{\sigma_j}} \exp(-\beta_{0j})$ is the baseline hazard function for segment $j$ with Weibull distribution parameters $\sigma_j$ and $\beta_{0j}$, $x_{it}$ are the customer and time-specific covariates and $\beta_j$ are the response parameters for segment $j$. The duration time $t$ was measured in months from the initial subscription date of the customer as the origin. The covariates were incorporated in the model with a one period lag.

The likelihood contribution of observation $it$ for customer $i$ at time period $t$ conditional on $i$ belonging to segment $j$ is given by

$$
L(it \mid i \in j) = S_j(t \mid x_{it}, i \in j)^{1-d_{it}} f_j(t \mid x_{it}, i \in j)^{d_{it}} \quad (2)
$$

where $S$ is the survival probability, $f$ is the probability density function and $d_{it}$ is an indicator variable equal to 1 if an event (churn) takes place in the time interval and equal to 0 otherwise. The unconditional likelihood contribution of customer $i$ over all the $t$ observations is given by
\[ L(i) = \sum_{j=1}^{J} L(i \mid i \in j) \Pr(j), \ j = 1, 2, \ldots J \]  \hspace{1cm} (3)

where \( J \) is the number of latent segments and \( \Pr(j) \) is the prior probability of segment \( j \), parameterized as

\[ \Pr(j) = \exp(\theta_j) / \sum_{j=1}^{J} \exp(\theta_j), \ \text{where} \ \theta_j = 0. \]  \hspace{1cm} (4)

The model parameters were estimated using maximum likelihood routines available in the LIMDEP software package. The number of latent segments is selected by estimating models with different values for \( J \) and choosing the best fitting model according to the Bayesian Information Criterion (BIC).

3.3 Data and Empirical Specification

Data Description

We study the data on customers of a direct-to-home satellite TV provider operating in a South American country. A random sample of 3000 customers was extracted from the firm’s database in September 2003 and was drawn from among those customers who had subscribed to the firm’s service during the past 12 months. A random sample of 2000 of those customers was used in the calibration sample while the remaining 1000 customers were assigned to a cross-sectional holdout sample. After cleaning the data\(^{15}\) we were left with 1870 customers in the calibration sample. The duration time for each customer was split into customer-months giving a total of

\(^{15}\) We removed customers with multiple records and illogical values on some of the covariates.
12,498 observations. The holdout sample had 931 customers with 5921 customer-month observations.

At the end of the study period, 77% of the customers in the calibration sample were active and 23% had churned. The average duration time per customer was 6.7 months (ignoring right censoring). In terms of demographics, the average age of the customers in the sample was 41 years, 34% of the customers were women, 72% of the customers were married and 67% of the customers lived in a house (and not an apartment) which in this market is an indicator of a better socioeconomic profile.

The firm’s customer management software provided demographic variables as well as billing and transaction variables for the entire duration of the customer’s relationship with the firm. For example, some of the demographic variables were age, gender and marital status, type of house, phone connectivity and payment method used (cash or credit card). Billing and transaction variables included the initial date of subscription, date of churn (if applicable), type of package subscribed, amount invoiced, amount paid, number of customer contacts made to the customer service unit for billing, technical and other service issues, number of customer contacts made to the customer recovery unit when a preliminary decision to stop a subscription was indicated to the company.

The final variables entering the model, including their functional forms (linear, log or quadratic), were selected on the basis of model fits using BIC scores. Appendix A and the following sub-section define these variables in more detail. Table 3.1 shows
summary statistics for the variables prior to any transformations. In the estimation sample these variables were scaled and standardized to ensure stability of the computation algorithm. Further, prior to log transformations, a constant was added to the variable so as to give a minimum value of 1.

Model Variables

We organized the above information into groups of variables relating to (1) customer service experience, (2) failure recovery, (3) payment equity and (4) demographic and other control variables.

Customer Service Experience

Customer service experience is the customer’s overall evaluation of the firm’s customer service based on their prior experience over the duration of the customer-firm relationship. Prior service experience has been shown to moderate positively the influence of customer satisfaction on duration (Bolton 1998) and also influence duration directly (Bolton et al. 2000). Prior experience with customer service has also been shown to influence customer attitude towards the service or evaluation of the service (Boulding et al. 1999).

Prior service experience has been measured in the extant literature using the number of transactions made by the customer in the past (Bolton et al. 2000) or the tenure of the relationship (Thomas et al. 2004). Our measure for prior experience is the variable NSERVICE, defined as the cumulative number of contacts made to the
customer service unit, by the customer, for billing, technical assistance and other service related issues.

Given the secondary data, we do not know the nature of the interactions which took place between the customer and the customer service center. The coefficient on NSERVICE could therefore be either positive or negative. The sign of the response parameter should provide us with information about the nature of prior experience between the customer and the service center. A negative (positive) sign indicates that a greater number of contacts lead to lower (higher) churn probabilities, i.e., the overall interaction or experience could be deemed to have been satisfactory (not satisfactory). We expect the sign and/or magnitude of the response parameter to be different across segments if there is heterogeneity in the nature of customers’ overall prior experience with the customer service center.

**Failure Recovery**

Failure recovery measures the customer’s satisfaction with the firm’s performance in its service recovery efforts. Service recovery refers to the actions an organization takes in response to a service failure (Gronroos 1988). Managing service recovery is an important part of the customer-firm relationship. For example, customers have been shown to be more dissatisfied by an organization’s failure to recover (compensate or respond adequately to the service failure) than by the service failure itself (Berry and Parasuraman 1991).
In addition, customers who receive a proper response to a complaint are more likely to stay, to buy new products, pay a price premium and engage in favorable word of mouth dissemination (e.g., Bowman and Narayandas 2001, Conlon and Murray 1996, Maxham 2001, Zeithaml, Berry and Parasuraman 1996). How effectively the firm responds to the failure is expected to impact whether the customer continues with the decision to quit (Keaveney 1995, Fornell and Wernerfelt 1987). In fact, companies including telephone, banks and Internet service providers (e.g., AOL) designate certain employees to field calls from customers intending to cancel. These employees are specially trained to deal with customer concerns by suggesting new price plans or services (Los Angeles Times 2005). Customer satisfaction with failure recovery efforts has been modeled, in the extant literature, using a mixed design experiment (Smith et al. 1999). Also, customer satisfaction has been shown to have a positive effect upon the duration of the customer-firm relationship (Bolton 1998). We therefore expect the customer’s evaluation of the firm’s failure recovery efforts to have an impact on churn.

Using our secondary data, we measure failure recovery with an indicator variable, DRECOVERY, which takes on a value of 1 if at least one customer contact was routed to the customer recovery unit and 0 otherwise. Given the nature of the data, we don’t know the actual customer evaluation of the firm’s recovery efforts. But the sign of the response parameter should be informative about the performance of the firm in its failure recovery efforts. If the firm is successful in its recovery efforts we
expect the customers who made the customer recovery contact to have a lower probability of churn compared to those who did not make the contact (i.e., they will have a negatively signed coefficient in the hazard model). On the other hand, if the firm is unsuccessful we expect the coefficient to have a positive sign (i.e., the churn probability is predicted to be higher than for those who did not contact the firm). We expect potential heterogeneity on this dimension to be captured through differences across segments in the sign and/or magnitude of the response parameter.

Payment Equity

Payment equity represents the customer’s evaluation of the fairness of the cost-benefit tradeoff. This has been linked to satisfaction with the firm’s service product. Customer satisfaction evaluations have been shown to affect subsequent usage of services by customers (Bolton and Lemon 1999) as well as the duration of the relationship (Bolton 1998). Thus, payment equity should affect customer churn through its influence on customer satisfaction. We represent the payment equity factor with two variables, INVOICE and DCASH. INVOICE is the cumulative amount invoiced to the customer over the duration of the customer-firm relationship to date. DCASH is an indicator variable set equal to 1 if the customer made his or her most recent payment with cash and 0 if he or she made it by way of a credit card.

We expect the coefficient on INVOICE to be negative i.e., as the cumulative amount invoiced increases we expect cumulative benefits to outweigh the cumulative costs (captured by the cumulative invoiced amount) leading to a decline in the churn
rates. Given the premium nature of the service product, we expect customers to derive benefits both from subscribing to the service (owning the product) as well as from using the product (watching the TV content). The diminishing marginal return of an additional TV channel (higher value subscription packages had more channels) is expected to be offset by the returns from subscribing (owning) the higher value package.

We expect the coefficient on DCASH to be positive and significant i.e., customers who make their payments in cash are expected to have higher churn rates than those who make their payments using credit cards. One reason for higher expected churn is the salience of the payment to the customer. Customers who make cash payments are expected to compare the benefits received from the service against the cost incurred more frequently and more intensively than those making payments via credit card. In this market, the customers who pay by cash must travel to designated collection centers every month to make the payments.

On the other hand, customers who pay by credit card see the cost of the service in an aggregated bill along with other credit card purchases and transactions. One of the features of credit cards is that they aggregate many small “losses” into one larger loss for the customer; in so doing they may reduce the total perceived value lost (Thaler 1995). Further, in a market where credit card penetration is relatively low (61% among paid TV customers) payment in cash may be a reflection of the poor credit worthiness of the customer.
Demographic and Control Variables

We also control for the effect of demographic and other variables which could have an impact on the churn probabilities of the customers. ARREAR is the amount of arrears that the customer owes to the firm. We expect the coefficient to be positive as the churn rate is expected to increase with larger arrear amounts. DPHONE is an indicator variable which takes on a value of 1 if the customer has a phone connection and 0 otherwise. We expect the coefficient to be negative. The greater the ability of the firm to contact the customer (e.g., to make an offer or address a problem), the lower the expected churn rate. The variable AGE, which captures the age of the customer at the time of the study, did not have a significant impact on the churn probability and was excluded from the final model. DFEMALE is an indicator variable which takes on a value of 1 if the customer was a woman and 0 otherwise was not predictive of churn probabilities and also was excluded. Similarly, DMARRIED an indicator variable for a married customer and DHOUSE an indicator variable for a customer living in a house (and not an apartment) were not significantly predictive of the churn probabilities. These variables also were not included in the final model (See Appendix A for definitions of the variables).

3.4 Empirical Results and Discussion

Model Selection

Exploratory analysis of the data using a Piece-wise Exponential hazard model with time intercepts showed the baseline hazard function to be a smoothly increasing
function of duration time. We therefore considered parametric hazard models which allow the baseline hazard to be an increasing function of time. We estimated the model with the following parametric hazard models – Exponential, Piece-wise Exponential with time intercepts, Loglogistic, Gompertz and Weibull. We selected the Weibull model for further analysis as it was the best fitting model and had the lowest BIC scores (see Table 3.2).

The single segment Weibull with covariates but no heterogeneity was then taken to be the benchmark model. We added latent class heterogeneity, fitting Weibull models with 2, 3 and 4 latent segments (see Table 3.3). Two-segment heterogeneity lowered the BIC score to 2013, a drop of 991 points from the benchmark model. This shows very clearly, at least for our data, that heterogeneity is very important to capture when modeling retention with historical company records. The three segment latent class Weibull model produced the lowest BIC score (1923) and was therefore selected for further analysis and discussion.

*Model Performance in the Holdout Sample*

For our model to be useful to practitioners, it is vital that it be able to provide good predictive performance on additional customer files not included in an estimation sample. We checked the performance of the selected three-segment Weibull model against the benchmark single-segment Weibull model with the cross-sectional holdout sample. The holdout sample had 931 customers with 5921 customer-month observations. The three-segment Weibull (heterogeneous) model performed better
than the single segment Weibull (homogeneous) model, giving a substantially lower log likelihood value (Table 3.4).

We also compared the performance of the two models in predicting the churn or survival probability of the customers in the holdout sample. The observed outcome for each customer in the sample was either churn or survival (still active) by the end of the study period. We computed the deviation between observed and predicted outcomes – the probability of churn or survival as predicted by each of the above models. We compared the Root Mean Squared Error (RMSE) and Mean Absolute Deviation (MAD) and found that the heterogeneous model clearly had lower RMSE and MAD than the single segment model (see Table 3.4).

The performance of the models can also be examined with lift analysis. To begin, we first calculated the top decile lift for each model. Top decile lift is defined as follows: the proportion of the 10% of customers predicted to be most likely to churn who actually churned relative to the baseline churn rate (Neslin et al. 2006). The top decile lift was 3.1 for the homogeneous model and 4.3 for the heterogeneous model. Thus, the heterogeneous model performs better than the homogeneous model in identifying the customers most likely to churn.

We then examined the prediction accuracy of the two models by comparing their cumulative lift curves against the random lift curve. The cumulative lift curve for each model shows the cumulative percentage of churners accounted for by the top x% of customers predicted as most likely to churn. The random lift curve shows the
percentage of churners expected to churn if the customers were randomly ordered. Higher lift indicates better prediction. Figure 3.1 shows the cumulative lift curve for the two models compared with the random lift curve. The heterogeneous model clearly outperforms the homogeneous model as evidenced by the higher lifts.

Parameter Estimates

Table 3.5 reports the parameter estimates for the three segment heterogeneous Weibull model, along with expected signs, where relevant. Customer service experience, NSERVICE, had a positive impact on the churn rate but the impact and significance differed across the three segments. It was significant and positive for segment 2 customers while it was positive, but not significant, for segments 1 and 3. Recall that a positive sign for the coefficient indicates that more contacts lead to higher churn probabilities (i.e., the net effect of the interactions or experience was negative). This suggests the possibility that efforts by firm personnel to address service-related issues were not successful in overcoming customer dissatisfaction for members of segment 2, but may have been for customers in segments 1 and 3.

Table 3.5 also gives the hazard ratios corresponding to each parameter estimate. The hazard ratio of 1.58 for NSERVICE for segment 2 indicates that an increase in the cumulative number of contacts by 1 unit (22.3%) leads to an increase in the hazard rate by 58%. The estimated hazard ratios are lower for segments 1 and 3.

The failure recovery variable, DRECOVERY, was significant and positive in all three segments, with the largest effect size coming in segment 2. Not surprisingly,
the result shows that customers who contacted the firm and expressed a desire to terminate the service had significantly higher subsequent churn probabilities than other customers. Though the parameter is significant in segment 3, the hazard ratio is by far the smallest of the three segments. An implication of this result is that failure recovery efforts may have been more effective in stemming churn for segment 3 customers than others. Firms could use results such as this to guide further diagnostic efforts regarding performance of failure recovery teams.\(^{16}\)

The overall results for DRECOVERY, and especially for segment 2, highlight the challenge firms’ face when trying to recover from failures in service delivery. While the firm’s efforts in segment 3 might have mitigated dissatisfaction, customers in segment 2 may have gone in the other direction as a result of the interaction with the failure recovery team. Though speculative given the limitations of our data, this would be consistent with earlier research findings that some customers can become more dissatisfied by an organization’s failure to recover than by the service failure itself (Berry and Parasuraman 1991). Again, firms could undertake further diagnostic steps to investigate this possibility.

Turning to payment equity, we find that INVOICE had a significant negative impact on churn probability, with the largest effect in segment 2. For segment 2, an

\(^{16}\) We also included an interaction term in the model for service and failure recovery, NSERVICE*DRECOVERY. Though the variable did improve overall model fit, it is only just significant for segment 1. The negative sign suggests that failure recovery may be more effective for customers who have already had extensive prior interactions with the firm in terms of customer service calls.
increase in the amount invoiced by 1 unit (a 25% increase) is associated with a decline in the hazard rate by 98% (a hazard ratio of 0.02). Recall that INVOICE is defined as the cumulative amount invoiced to date. Thus, growing levels of consumer surplus accrued by subscribers over time (i.e., revealed preference indicates that benefits exceed costs) are associated with lower churn rates. We also include a squared term in the model. This enters significantly with a positive sign for all three segments. This reflects the diminishing marginal effects of changes in cumulative invoice amounts.

The variable DCASH had a large positive impact on the churn probability, particularly for segment 2 customers. As we discuss below, a majority of the customers in segment 2 (71%) made their payments by cash. This result likely reflects the higher salience of cash as a payment mechanism as compared to credit cards. Our findings for INVOICE and DCASH confirm that payment equity is an important factor in predicting churn; our model results add to this by showing how the size of these effects vary across segments and may be associated with customer characteristics.

Among the control variables, ARREAR, which enters in log form, was significant and positive for all three segments. Thus, an increase in ARREAR is associated with an increase in the churn probability. The largest effect size was observed for segment 1 where a doubling in ARREAR is associated with an increase in the hazard rate by 160%. Interestingly, we note that customers in this segment had the lowest average values for ARREAR.
DPHONE entered the model with a negatively signed coefficient in each segment. The effect was significant for segments 1 and 2, and marginal for segment 3. The result highlights some of the challenges involved in improving customer retention in emerging markets where phone service is not universal. Customers without a phone will be much more difficult to proactively contact in order to propose new offers or address service problems.

Demographic variables for AGE, DFEMALE, DMARRIED and DHOUSE were not predictive of the churn rates and were therefore dropped from the final model. As we discuss below, the segments are quite similar in their demographic profiles. The insignificance of these terms and the similar segment profiles suggest that the firm’s \textit{a priori} demographic segmentation will be of limited diagnostic value. Some aspects of the profile of the firm’s customer base (e.g., the skew towards urban and upper income households) also may limit the explanatory power of standard demographic variables in this case.

\textit{Analysis of the Segments}

We further examined the results from the three-segment Weibull model by assigning the customers to each of the three segments based on their posterior probability of segment membership. The number of customers assigned to segment 1, 2 and 3 were 1528, 200 and 142 respectively. In segment 1, 91 customers churned (5.95%), while all customers churned in segments 2 and 3. Thus, customers in
segment 1 might be labeled non-churners whereas customers in segments 2 and 3 are churners.

Table 3.6 reports summary statistics for the model variables describing the members assigned to each of the three segments. Looking first at NSERVICE, customers in segment 2 made more calls to the customer service center (mean = 13.73) than those in the other segments (mean = 8.91 and 9.80). The mean of the failure recovery variable (DRECOVERY) also differs across segments. A greater percentage of customers in segment 2 made customer recovery calls (53%) compared to segment 1 (13%) and segment 3 (29%). Thus, customers in segment 2 may be further termed interactive churners (letting the firm know if they have a problem) whereas customers in segment 3 are passive churners (lower contact rates).

Differences are also apparent for the payment equity variables INVOICE and DCASH. Customers in segment 3 paid the lowest total invoiced amounts (mean = 475.72) compared to segment 1 (676.99) and segment 2 (699.44). Segment 1 had the lowest percentage (30%) of customers making their payments in cash compared to 76% for segment 2 and 58% for segment 3, indicating that the majority of the churners made their payments in cash.

Means for the control variables DPHONE and ARREAR also differed across segments. In segment 1, 79% of the customers in segment 1 had phone connectivity but only 65% and 70% of customers in segment 2 and 3, respectively, had phone connectivity. Customers in segment 1 had average arrears (5.78) much lower than
customers in segment 2 (15.74) and segment 3 (15.26). Again we see differences across segments in these variables and especially between churners and non-churners\textsuperscript{17}.

Turning to the demographic and control variables, average AGE across the three segments was nearly identical. Customers in segment 1 were, on average, 41 years old, segment 2 customers were 42 and segment 3 customers were 41. The percentage of customers who were women also was quite similar – 35%, 32% and 30% in segments 1, 2 and 3, respectively. There were small differences in the proportion of married customers and slight differences in type of living arrangement (DHOUSE). Broadly speaking, however, the three segments were close in demographic profile and, as noted above, demographic variables were not significant in predicting customer churn probabilities.

3.5 Conclusions

The primary objective of this research has been to develop a simple hazard modeling approach that will improve the diagnosis and prediction of customer churn with internal customer records data. A secondary objective has been to establish an empirical link between customer churn and factors which managers may be able to change. Both objectives are important for practitioners who are seeking understand

\textsuperscript{17} We note that the NSERVICE, ARREAR and INVOICE variables have been standardized. Thus, changes in these variables affect churn in the hazard model but differences in levels across customers do not.
churn in their businesses and interested in devising marketing interventions to reduce churn and increase customer lifetime values.

A promising direction for improving churn prediction lies in modeling the unobserved heterogeneity in churn rates using duration or hazard models (e.g., Bolton 1998). We also expected differences to exist in how customers respond to changes in different drivers of churn. Thus, a key goal of our research is to incorporate heterogeneity not only in a baseline churn function but also in the response parameters.

We extended the hazard model to incorporate heterogeneity in the response parameters using a latent class Weibull hazard model. We used individual-level customer data from a satellite TV provider’s customer management system. The firm operates in an emerging market country in South America and also faces issues unique to those markets. For example, only customers with very good credit worthiness can obtain and keep a credit card. Thus, method of payment could be more predictive of churn probabilities than in the U.S. market.

We estimated our model using information readily available to managers i.e., secondary data records stored internally for each customer. We use this information to build modeling constructs identified in the extant literature as impacting customer churn. These variables have typically measured with survey and experimental data. By proposing ways to capture some of these factors with secondary data, we hope our research will help managers make better use of information already available to them,
without the need to first collect survey data or conduct experiments. For example, these more costly steps might be undertaken to conduct further diagnostics but after studying the insights which could be obtained from modeling internal data.

We found that a three segment heterogeneous Weibull model provided the best fit to our data. In addition, this model outperformed a benchmark single segment homogeneous Weibull model in a holdout sample, giving lower log likelihood, RMSE and MAD values. It also had a larger top decile lift and its cumulative lift curve enveloped the single segment model. Differences in magnitude, but not sign, were also found in the segment-level response parameters pertaining to variables for customer service experience, failure recovery, and payment equity. Estimation results from our data present a convincing case for the benefits of incorporating heterogeneity into the hazard modeling approach for predicting customer churn.

We also found that conventional demographic variables (AGE, DFEMALE, DMARRIED and DHOUSE) were not predictive of the churn. The segments were also very similar in their demographic profiles. This finding is consistent with the long recognized limitations of demographic variables in predicting differences in consumer purchase behavior. The situation may be even more challenging in emerging markets given that the customer base skews toward urban and upper income households. Such challenges highlight the need to develop other approaches to predicting churn, such as taking greater advantage of the observed customer histories available within a firm’s in-house data.
There is scope for much further research in this area. First, the type of churn—voluntary or involuntary (when the firm asks the customer to leave) could have an impact as the failure times of the two types of events could be dependent. We hope to address this issue in a future research study. Second, there could be possible dependencies between the acquisition and the retention process (e.g., Thomas 2001) which could impact the churn rates. Our model currently does not incorporate the acquisition process due to a lack of specific data about the firm’s acquisition process at the customer level. If suitable data could be obtained, it would be desirable to incorporate the customer acquisition process into a unified modeling approach.
### 3.6 Appendices

**Appendix A: Variable Definition**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Customer Service Experience</strong></td>
<td></td>
</tr>
<tr>
<td>NSERVICE&lt;sub&gt;it&lt;/sub&gt;</td>
<td>Cumulative number of customer service</td>
</tr>
<tr>
<td><strong>Failure Recovery</strong></td>
<td></td>
</tr>
<tr>
<td>DRECOVERY&lt;sub&gt;it&lt;/sub&gt;</td>
<td>Indicator variable: Customer recovery</td>
</tr>
<tr>
<td><strong>Payment Equity</strong></td>
<td></td>
</tr>
<tr>
<td>INVOICE&lt;sub&gt;it&lt;/sub&gt;</td>
<td>Cumulative invoice amount</td>
</tr>
<tr>
<td>DCASH&lt;sub&gt;i&lt;/sub&gt;</td>
<td>Indicator variable: Payment by cash</td>
</tr>
<tr>
<td><strong>Demographic and Control</strong></td>
<td></td>
</tr>
<tr>
<td>ARREAR&lt;sub&gt;it&lt;/sub&gt;</td>
<td>Amount of arrears</td>
</tr>
<tr>
<td>DPHONE&lt;sub&gt;i&lt;/sub&gt;</td>
<td>Indicator variable: Phone connectivity</td>
</tr>
<tr>
<td>AGE&lt;sub&gt;i&lt;/sub&gt;</td>
<td>Age of the customer (years) in September</td>
</tr>
<tr>
<td>DFEMALE&lt;sub&gt;i&lt;/sub&gt;</td>
<td>Indicator variable: Customer is a woman</td>
</tr>
<tr>
<td>DMARRIED&lt;sub&gt;i&lt;/sub&gt;</td>
<td>Indicator variable: Customer is married</td>
</tr>
<tr>
<td>DHOUSE&lt;sub&gt;i&lt;/sub&gt;</td>
<td>Indicator variable: Customer lives in a house</td>
</tr>
</tbody>
</table>
Table 3.1: Summary Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean*</th>
<th>Standard Deviation</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer Service Experience</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSERVICE</td>
<td>9.50</td>
<td>9.36</td>
<td>6</td>
</tr>
<tr>
<td>Failure Recovery</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRECOVERY</td>
<td>0.19</td>
<td>0.39</td>
<td>0</td>
</tr>
<tr>
<td>Payment Equity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INVOICE</td>
<td>664.12</td>
<td>314.87</td>
<td>675.27</td>
</tr>
<tr>
<td>DCASH</td>
<td>0.37</td>
<td>0.48</td>
<td>0</td>
</tr>
<tr>
<td>Demographic and Control Variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARREAR</td>
<td>7.56</td>
<td>31.67</td>
<td>0</td>
</tr>
<tr>
<td>DPHONE</td>
<td>0.77</td>
<td>0.42</td>
<td>1</td>
</tr>
<tr>
<td>AGE</td>
<td>41.06</td>
<td>12.62</td>
<td>39</td>
</tr>
<tr>
<td>DFEMALE</td>
<td>0.34</td>
<td>0.47</td>
<td>0</td>
</tr>
<tr>
<td>DMarried</td>
<td>0.72</td>
<td>0.45</td>
<td>1</td>
</tr>
<tr>
<td>DHOUSE</td>
<td>0.67</td>
<td>0.47</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: * Average taken over the customers in the data sample
Table 3.2: Parametric Hazard Model Selection

<table>
<thead>
<tr>
<th>Model</th>
<th>Log Likelihood</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull</td>
<td>-1455.02</td>
<td>3004.38</td>
</tr>
<tr>
<td>Loglogistic</td>
<td>-1468.87</td>
<td>3032.08</td>
</tr>
<tr>
<td>Exponential</td>
<td>-1510.93</td>
<td>3106.76</td>
</tr>
<tr>
<td>Exponential with time intercepts</td>
<td>-1491.48</td>
<td>3171.63</td>
</tr>
<tr>
<td>Gompertz</td>
<td>-1561.60</td>
<td>3217.53</td>
</tr>
</tbody>
</table>
Table 3.3: Latent Class Weibull Model Selection

<table>
<thead>
<tr>
<th>Weibull Model</th>
<th>Log Likelihood</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null model with no covariates</td>
<td>-1758.91</td>
<td>3536.69</td>
</tr>
<tr>
<td>Single segment with covariates</td>
<td>-1455.02</td>
<td>3004.38</td>
</tr>
<tr>
<td>Two segment with covariates</td>
<td>-907.46</td>
<td>2013.01</td>
</tr>
<tr>
<td>Three segment with covariates</td>
<td>-810.56</td>
<td>1922.99</td>
</tr>
<tr>
<td>Four segment with covariates</td>
<td>-781.28</td>
<td>1968.20</td>
</tr>
</tbody>
</table>
Table 3.4: Model Performance in Holdout Sample

<table>
<thead>
<tr>
<th>Weibull Model</th>
<th>Log Likelihood</th>
<th>RMSE</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single segment (homogeneous)</td>
<td>-1197.10</td>
<td>0.3247</td>
<td>0.4214</td>
</tr>
<tr>
<td>Three segment (heterogeneous)</td>
<td>-855.68</td>
<td>0.2437</td>
<td>0.3229</td>
</tr>
</tbody>
</table>
Table 3.5: Model Parameters: Three Segment Weibull

<table>
<thead>
<tr>
<th>Variable</th>
<th>Segment 1</th>
<th>Segment 2</th>
<th>Segment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter</td>
<td>t statistic</td>
<td>Hazard Ratio</td>
</tr>
<tr>
<td>Customer Service Experience</td>
<td>NSERVICE(^a)</td>
<td>0.162</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>DRECOVERY(^b)</td>
<td>3.028</td>
<td>4.69</td>
</tr>
<tr>
<td></td>
<td>NSERVICE x DRECOVERY</td>
<td>-0.683</td>
<td>-2.09</td>
</tr>
<tr>
<td>Failure Recovery</td>
<td>INVOICE(^c)</td>
<td>-3.356</td>
<td>-10.10</td>
</tr>
<tr>
<td></td>
<td>(INVOICE)(^2)</td>
<td>0.833</td>
<td>6.02</td>
</tr>
<tr>
<td></td>
<td>DCASH(^b)</td>
<td>1.942</td>
<td>3.60</td>
</tr>
<tr>
<td>Payment Equity</td>
<td>Ln (ARREAR(^c,d))</td>
<td>1.378</td>
<td>3.45</td>
</tr>
<tr>
<td></td>
<td>DPHONE(^b)</td>
<td>-0.788</td>
<td>-3.39</td>
</tr>
<tr>
<td></td>
<td>Sigma</td>
<td>0.076</td>
<td>11.99</td>
</tr>
<tr>
<td></td>
<td>Number of Customers</td>
<td>1528</td>
<td>200</td>
</tr>
</tbody>
</table>

Note:  
\(^a\) Scaled by 10 and standardized  
\(^b\) Indicator Variable  
\(^c\) Scaled by 100 and standardized  
\(^d\) Constant value= 3.519 added to ensure minimum value is 1
### Table 3.6: Segment Level Summary Statistics of Covariates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Segment 1</th>
<th></th>
<th>Segment 2</th>
<th></th>
<th>Segment 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Standard Deviation</td>
<td>Mean&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Standard Deviation</td>
<td>Mean&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td><strong>Customer Service Experience</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSERVICE</td>
<td>8.91</td>
<td>9.12</td>
<td>13.73</td>
<td>10.29</td>
<td>9.80</td>
<td>9.05</td>
</tr>
<tr>
<td><strong>Failure Recovery</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRECOVERY</td>
<td>0.13</td>
<td>0.34</td>
<td>0.53</td>
<td>0.50</td>
<td>0.29</td>
<td>0.46</td>
</tr>
<tr>
<td><strong>Payment Equity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INVOICE</td>
<td>676.99</td>
<td>317.30</td>
<td>699.44</td>
<td>282.70</td>
<td>475.72</td>
<td>267.99</td>
</tr>
<tr>
<td>DCASH</td>
<td>0.30</td>
<td>0.46</td>
<td>0.76</td>
<td>0.43</td>
<td>0.58</td>
<td>0.49</td>
</tr>
<tr>
<td><strong>Demographic and Control Variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARREAR</td>
<td>5.78</td>
<td>29.72</td>
<td>15.74</td>
<td>38.15</td>
<td>15.26</td>
<td>38.53</td>
</tr>
<tr>
<td>DPHONE</td>
<td>0.79</td>
<td>0.41</td>
<td>0.65</td>
<td>0.48</td>
<td>0.70</td>
<td>0.46</td>
</tr>
<tr>
<td>AGE</td>
<td>40.95</td>
<td>12.44</td>
<td>41.91</td>
<td>13.21</td>
<td>41.03</td>
<td>13.74</td>
</tr>
<tr>
<td>DFEMALE</td>
<td>0.35</td>
<td>0.48</td>
<td>0.32</td>
<td>0.47</td>
<td>0.30</td>
<td>0.47</td>
</tr>
<tr>
<td>DMARRIED</td>
<td>0.72</td>
<td>0.45</td>
<td>0.76</td>
<td>0.43</td>
<td>0.67</td>
<td>0.49</td>
</tr>
<tr>
<td>DHOUSE</td>
<td>0.67</td>
<td>0.47</td>
<td>0.62</td>
<td>0.49</td>
<td>0.72</td>
<td>0.45</td>
</tr>
</tbody>
</table>

*Note: a Average taken over the customers in the segment*
Figure 3.1: Cumulative Lift Chart
3.7 References


CHAPTER 4

HAZARDS OF IGNORING INVOLUNTARY CUSTOMER CHURN

Abstract

We study the impact of ignoring involuntary churn (when a firm terminates the subscription of a customer) on the estimation and diagnosis of voluntary churn (when the customer terminates the relationship). The main event is voluntary churn and the competing event is involuntary churn. In our case, if the theoretical duration times of the two competing events are correlated then the estimates of the hazard rates and survival times of the main event can be biased.

We estimate a bivariate Weibull survival model that captures the dependency between the two event times and has been proposed in the literature as an approach to the problem of dependent competing risk. We compare this model with a benchmark model – a univariate Weibull model with only voluntary churn. We estimate the models using maximum likelihood techniques.

We find significant differences in the prediction of voluntary churn rates. The bivariate Weibull survival model does better in predicting the customers who are more likely to churn. Further, the impact of covariates on the voluntary churn rates is different across the two models. An added dimension of the study is that the key covariates that influence voluntary churn rate impact involuntary churn rate
differently. Our study highlights the need to incorporate involuntary churn when modeling the voluntary churn process.

**Keywords:** Customer churn, dependent competing risk, parametric survival model
4.1 Introduction

In many industries such as credit cards, telecommunications or direct-to-home satellite TV, there are two types of customer churn – voluntary churn and involuntary churn. Voluntary churn is defined as the event when a customer ends his subscription of the firm’s service. This could be for reasons such as leaving the industry, switching to a competitor or moving to another location. Involuntary churn, on the other hand, is defined as the event when a firm ends the customer-firm relationship. This could be for reasons such as the customer exceeds threshold risk levels, has very large arrears or makes irregular payments.

The typical practice in many industries is to ignore the involuntary churners when making predictions or analyzing voluntary churn rates. In some of these industries the rate of involuntary churn is quite large. For example, Directv US, a leading player in the direct-to-home satellite TV market, reported that 42% of their total churn rate in 2005 was involuntary.

The objective of this research is to show that ignoring involuntary churn can bias the prediction and diagnosis of voluntary churn rates. We analyze the problem within the competing risk framework where the occurrence of involuntary churn precludes the occurrence of voluntary churn for a given customer. In our case, the competing event is involuntary churn. The competing event prevents us from observing the customer’s potential duration time to voluntary churn. The two churn processes are expected to be dependent – factors idiosyncratic to the customer.
(observed as well as unobserved) may be associated with the length of both event
times.

*Dependent Competing Risk Framework*

In biostatistics as well as social sciences there is interest in the study of
duration time – the time to the occurrence of an event or alternately the probability of
occurrence of an event. Some examples of duration time are time to cessation of
smoking, time to death due to some specific disease or time to customer churn. A
competing risk is the occurrence of another event which precludes the main event of
interest from occurring. For example, it could be death from a cause other than the
disease of interest. Thus, a subject may “fail” due to one of \( k \ (k \geq 2) \) causes which are
known as competing risks.

Theoretical or potential duration time associated with each competing risk is
defined as the duration time in the hypothetical situation when other risks are
inoperative. But when all risks are operative, only the smallest of the duration time is
observed. The competing events may be dependent i.e., the theoretical duration time to
the occurrence of one event may be correlated to the theoretical duration time to the
occurrence of the competing event.

The majority of statistical methods for censored survival data assume that the
main event time, the competing event times and the censoring times are all
independent. The assumption of independence is valid in situations such as end of
study censoring or when a patient moves to another location and his exit from the
study is unrelated to his medical treatment. But there are many situations when the assumption of independence is questionable (Moeschberger and Klein 1995). One such situation is a clinical trial in which failure times from causes of secondary interest are recorded as censored observations of the failure times from the causes of primary interest (Lagakos 1979). This practice is often used in studies to deal with the problem of competing risk.

Ignoring the dependence between the competing event times or between the duration time and the censoring time can lead to biased and inconsistent estimates of the survival probability or of the cumulative hazard rates (Sannis et al. 2005, Lagakos and Williams 1978, Lagakos 1979, Moeschberger and Klein 1984, Klein and Moeschberger 1984, 1986, 1987, Slud and Byar 1988).

An approach to the problem of dependent competing risk, proposed in the biostatistical literature, is to estimate a bivariate Weibull survival model that captures the dependency between the two competing event times (Moeschberger 1974). Other approaches to tackle the problem of dependent competing risks have been proposed in the literature with no clear consensus on any one approach (Moeschberger and Klein 1995).

Given our context, we estimate the parametric bivariate Weibull survival model (Moeschberger 1974) to deal with the problem of dependent competing risk. The model provides a quantifiable measure of the dependence between the two competing events – voluntary and involuntary churn. Further, it gives estimates of the
churn rates with the flexibility for the two churn processes to have different rates. The model also allows us to study the impact of key covariates on the two churn rates. The marginal survival distributions of this model are also Weibull with unequal shape and scale constants. Exploratory work on the data sample indicated the Weibull survival function best represented each of the two churn processes – voluntary and involuntary.

4.2 Modeling Framework

As noted earlier, there are two types of churn – voluntary and involuntary. These events are mutually exclusive – occurrence of either of these events precludes us from observing the other event. We analyze the data using a dependent competing risk framework as there is no a priori reason to assume that the two competing events – voluntary and involuntary churn – are independent.

We use the bivariate Weibull survival model proposed by Moeschberger (1974). The model is an extension of the bivariate Exponential model of Marshall and Olkin (1967). We extend the model to incorporate the impact of time varying covariates along the lines used in the current literature on univariate survival models.

The bivariate Weibull survival distribution is given by

$$F(y_1, y_2) = P[Y_1 > y_1, Y_2 > y_2] = \exp\left(\max(\lambda_1 y_1, \lambda_2 y_2)\right)$$

where

$$\lambda_s = \exp(-\beta_s Z'), \quad s = 1, 2, 3$$

$$\beta_s = (\beta_{s,0}, \beta_{s,1}, \ldots, \beta_{s,n-1})$$

and

$$Z = (1, Z_1, \ldots, Z_{n-1})$$
$Y_1$ and $Y_2$ are the theoretical customer duration times for each type of event, $\lambda_s$, $c_1$ and $c_2$ are the distributional parameters, $Z$ are the customer specific covariates with some of them varying over the customer’s duration time and $\beta_s$ are the parameters that capture the impact of $Z$. A necessary and sufficient condition for $Y_1$ and $Y_2$ to be independent is that $\lambda_3 = 0$.

The marginal survival distributions of the bivariate Weibull survival distribution

\[
\overline{F}_1(y_1) = \exp\left\{-(\lambda_1 + \lambda_3)y_1^{c_1}\right\} \quad \text{and} \quad \overline{F}_2(y_2) = \exp\left\{-(\lambda_2 + \lambda_3)y_2^{c_2}\right\} \quad (2)
\]

are Weibull with unequal shape and scale constants.

The observed duration time for event $i$ is given by

\[
X_i = Y_i | Y_i = \min\{Y_1, Y_2\}, \quad i = 1, 2 \quad (3)
\]

In our context, event 1 is voluntary churn and event 2 is involuntary churn. We only observe the duration time of the event that occurs first and precludes the occurrence of the other event. The conditional distribution function $F_i(X_i)$ depends on the relative sizes of $c_1$ and $c_2$. We examine $F_i(X_i)$ for the two cases $c_1 > c_2$ and $c_2 > c_1$ with the likelihood function defined separately for the two cases. For each case, $F_i(X_i)$ has two parts defined over the regions $X_i \leq 1$ and $X_i > 1$. The reason for this complexity is the presence of the term “$\max(y_1^{c_1}, y_2^{c_2})$” in the joint survival distribution function for the two events.
The likelihood function for a case is given by

\[ L \propto \prod_{j=1}^{n_{11}} f_1(x_{ij1}) \prod_{j=1}^{n_{12}} f_1(x_{ij2}) \prod_{j=1}^{n_{21}} f_2(x_{i2j1}) \prod_{j=1}^{n_{22}} f_2(x_{i2j2}) \prod_{l=1}^{r_1} s(w_{l1}) \prod_{l=1}^{r_2} s(w_{l2}), \]  

(4)

where \( f_i(x_{ijk}) \) is the probability density of customer \( j \) experiencing event \( i \) with his observed duration time \( x \) lying in the region \( k \)\(^{18} \) and \( s(w_{jk}) \) is the survival probability of customer \( l \) with his observed right censored duration time \( w \) lying in the region \( k \). \( n_{ik} \) is the number of customers in the data sample who experienced event \( i \) with their observed duration time lying in region \( k \) and \( r_k \) is the total number of customers with right censored duration times lying in the region \( k \). See Appendix C for the exact terms that enter the likelihood equation for the two cases.

To incorporate the time varying covariates, we split the customer’s observed duration time into equal time intervals. The number of time intervals varied for each customer depending on their observed duration time. In our context, these are monthly intervals as some of the customer specific covariates change monthly.

The contribution to the likelihood of a customer who experienced an event – his conditional probability density \( f_i(x_{ijk}) \) – would be split into the contribution for each time interval. In all the intervals preceding the last interval, his contribution to the likelihood is the conditional survival probability of surviving until the end of the interval conditional on his surviving until the beginning of that interval. In the last interval, his contribution is the conditional probability density of his experiencing the

\(^{18} k = 1 \) is the region where \( X_i \leq 1 \) and \( k = 2 \) is the region where \( X_i > 1 \)
event at the end of interval conditional on surviving until the beginning of that interval. For example, the contribution to the likelihood function of a customer $j$ experiencing an event $i$ with observed duration time $x_i$ lying in region $k$ is

$$f_i(x_{ijk}) = \prod_{m=1}^{M_{ij}} S_i(v_{1,ijk})S_i(v_{2,ijk}).....S_i(v_{M_{j-1,ijk}})f_i(v_{M_{j,ijk}})$$ (5)

where $v_{m,ijk}$ are the equal time intervals for customer $j$ with $m$ taking on the values 1,2, ……$M_{j-1}, M_j$.

Similarly, the contribution to the likelihood of an active customer with his observed right censored duration time $w$ – his survival probability function $s(w_{jk})$ – would be split into the contribution for each time interval. In all the intervals his contribution to the likelihood is the conditional survival probability of surviving till the end of the interval conditional on his surviving till the beginning of that interval. For example, the contribution to the likelihood function of a customer $l$ with observed duration time $w$ lying in region $k$ is

$$S(w_{lk}) = \prod_{m=1}^{M_{lk}} S(v_{1,ik})S(v_{2,ik}).....S(v_{M_{l-1,ik}})S(v_{M_{l,ik}})$$ (6)

where $v_{m,ik}$ are the equal time intervals for customer $l$ with $m$ taking on the values 1,2, ……$M_{l-1}, M_l$.

We obtain the maximum likelihood estimates of the parameters for each case using constrained optimization techniques and select the parameters of the best fitting case.
4.3 Data and Empirical Specification

The occurrence of two types of churn in the direct-to-home satellite TV market – voluntary churn and involuntary churn – can be analyzed with the dependent competing risk framework described earlier. We estimate a bivariate Weibull survival model (main model) and compare it to the univariate Weibull survival model (benchmark model).

Data Description

We study the data on customers of a direct-to-home satellite TV provider operating in a South American country. For estimating the model, a random sample of 4500 customers was extracted from the firm’s database in September 2003 and was drawn from among those customers who had subscribed to the firm’s service during the past 12 months. After cleaning the data\footnote{We removed customers with multiple records, missing values or illogical values on some of the key covariates.} we were left with 4426 customers in the calibration sample. The duration time for each customer was split into customer-months giving a total of 28911 customer-month observations. We will call this sample data sample 1. The customers were comprised of 849 voluntary churners (19.2%), 899 involuntary churners (20.3%) and 2678 active customers (60.5%). For a cross-sectional holdout sample, a random sample of 4000 customers was extracted for the same time period. After cleaning the data sample, we had 3814 customers with
25098 customer-month observations. The holdout sample had 755 voluntary churners (19.8%), 824 involuntary churners (21.6%) and 2235 active customers (58.6%).

A sub-sample of the calibration sample (data sample 2) with only voluntary churners and active customers (3527 customers with 23307 customer-month observations) was used for estimating the benchmark univariate model. The data sample for the benchmark model represents current industry practice of filtering out the involuntary churners from their analysis. Another sub-sample of the estimation sample (data sample 3) with only involuntary churners and active customers (3577 customers and 23309 customer-month observations) was also created for model selection purposes.

The average duration time per customer in the calibration data sample 1 was 6.53 months (ignoring right censoring) with median duration time of 6 months. In the sub-sample data sample 2 the mean duration time was 6.61 months (median 7 months) while in sub-sample data sample 3 the mean duration time was 6.52 months with a median of 6 months.

The firm’s database provided customer specific billing and transaction variables as well as demographic variables for the entire duration of the customer’s relationship with the firm. For example, some of the demographic variables which we found to have an impact on the churn rates were phone connectivity and payment method used (cash or credit card). Billing and transaction variables included the initial date of subscription, date and type of churn (if applicable), type of package
subscribed, amount invoiced, amount paid, number of customer contacts made to the customer service unit for billing, technical and other service issues, number of customer contacts made to the customer recovery unit when a preliminary decision of the customer to stop the subscription was indicated to the firm.

The final variables entering the model, including their functional forms (linear, log or quadratic), were selected on the basis of model fits using BIC scores. Appendix A and the following sub-section define these variables in more detail. Table 4.1 shows summary statistics for the variables prior to any transformations. In the estimation sample these variables were scaled and standardized to ensure stability of the computation algorithm. Further, prior to log transformations, a constant was added to the variables so as to give a minimum value of 1. Also, the duration time was scaled by a factor of 10 in all the data samples. This was to ensure all regions of the bivariate Weibull distribution were represented in the data sample.

We compared the performance of the two models in terms of the diagnostics like percentage change in hazard rate and elasticity which capture the impact of different covariates on the churn rate. We also compare the prediction power of the models in a cross-sectional holdout sample looking at RMSE, MAD as well as cumulative lift curves.

The univariate survival models (Weibull, Logistic, Exponential and Exponential with time intercepts) were estimated using the LIMDEP software package. The bivariate Weibull survival model was estimated using code written for the MATLAB
software package. To validate the code and check its ability to retrieve the parameters of the underlying distribution, we generated a simulated data sample of bivariate Weibull survival data. The data sample had two competing events with the duration time of some customers right censored. Further, the data sample had a time varying covariate which was drawn from a standard normal distribution. Please refer to Appendix B for details on the algorithm for generating the simulated data set. The code for the bivariate Weibull survival model was successful in retrieving the underlying distributional parameters of the simulated data set.

Model Variables

We organized the customer specific information available from the firm’s database into groups of variables relating to (1) customer service experience, (2) failure recovery, (3) payment equity and (4) control variables. An important focus of the study is to show the difference in the response parameters and their impact on voluntary customer churn when involuntary churn is ignored. An additional dimension of the study is to highlight any differences that may exist in the response parameters of the covariates for the two types of churn. To the best of our knowledge, no study has looked at the impact of different covariates on involuntary customer churn rates.

Customer Service Experience

Customer service experience is defined as the customers’ overall evaluation of the firm’s customer service based on their prior experience over the duration of the customer-firm relationship. Prior service experience has been shown to moderate
positively the influence of customer satisfaction on duration (Bolton 1998), influence
duration directly (Bolton et al. 2000) as well as influence customer attitude towards
the service or evaluation of the service (Boulding et al. 1999).

One measure of prior service experience in the current literature is the number
of transactions made by the customer in the past (Bolton et al. 2000). Our measure for
prior experience is the variable NSERVICE, defined as the cumulative number of
contacts made to the customer service center, by the customer, for billing, technical
assistance and other service related issues.

Given the secondary data, we do not know the nature of the interactions which
took place between the customer and the customer service center. The coefficient on
NSERVICE could therefore be either positive or negative and differ in magnitude,
sign and significance across the two types of churn. The sign of the response
parameter should provide us with information about the nature of prior experience
between the customer and the service center. A negative (positive) sign indicates that a
greater number of contacts lead to lower (higher) voluntary probabilities, i.e., the
overall interaction or experience could be deemed to have been satisfactory (not
satisfactory).

Failure Recovery

Failure recovery measures the customer’s satisfaction with the firm’s
performance in its service recovery efforts. Service recovery refers to the actions an
organization takes in response to a service failure (Gronroos 1988). Managing service
recovery is an important part of the customer-firm relationship. For example, customers have been shown to be more dissatisfied by an organization’s failure to recover (compensate or respond adequately to the service failure) than by the service failure itself (Berry and Parasuraman 1991). How effectively the firm responds to the failure is expected to impact whether the customer continues with the decision to quit (Keaveney 1995, Fornell and Wernerfelt 1987, Bowman and Narayandas 2001, Conlon and Murray 1996, Maxham 2001, Zeithaml, Berry and Parasuraman 1996). We therefore expect the customer’s evaluation of the firm’s failure recovery efforts to have an impact on voluntary customer churn.

In the current literature, customer satisfaction with failure recovery efforts has been modeled using a mixed design experiment (Smith et al. 1999). Also, customer satisfaction has been shown to have a positive effect upon the duration of the customer-firm relationship (Bolton 1998). Using our secondary data, we measure failure recovery with an indicator variable, DRECOVERY, which takes on a value of 1 if at least one customer contact was routed to the customer recovery unit and 0 otherwise.

Given the nature of the data, we don’t know the actual customer evaluation of the firm’s recovery efforts. But the sign of the response parameter should be informative about the performance of the firm in its failure recovery efforts. If the firm is successful in its recovery efforts we expect the customers who made the customer recovery contact to have a lower probability of voluntary churn compared to those
who did not make the contact (i.e., they will have a negatively signed coefficient in the
hazard model). On the other hand, if the firm is unsuccessful we expect the coefficient
to have a positive sign (i.e., the voluntary churn probability is predicted to be higher
for those who contacted the firm). We expect differences across the two types of churn
in the sign and/or magnitude of the response parameter.

Payment Equity

Payment equity represents the customer’s evaluation of the fairness of the cost-
benefit tradeoff. This has been linked to satisfaction with the firm’s service product.
Customer satisfaction evaluations have been shown to affect subsequent usage of
services by customers (Bolton and Lemon 1999) as well as the duration of the
relationship (Bolton 1998). Thus, payment equity should affect voluntary customer
churn through its influence on customer satisfaction.

We represent the payment equity factor with two variables, INVOICE and
DCASH. INVOICE is the cumulative amount invoiced to the customer over the
duration of the customer-firm relationship to date. DCASH is an indicator variable set
equal to 1 if the customer made his or her most recent payment with cash and 0 if he
or she made it by way of a credit card.

We expect the coefficient on INVOICE to be negative for voluntary churn i.e.,
as the cumulative amount invoiced increases we expect cumulative benefits to
outweigh the cumulative costs (captured by the cumulative invoiced amount) leading
to a decline in the voluntary churn rates. Given the premium nature of the service
product, we expect customers to derive benefits both from subscribing to the service (owning the product) as well as from using the product (watching the TV content). The diminishing marginal return of an additional TV channel (higher value subscription packages had more channels) is expected to be offset by the returns from subscribing (owning) the higher value package.

We expect the coefficient on DCASH to be positive and significant i.e., customers who make their payments in cash are expected to have higher voluntary churn rates than those who make their payments using credit cards. One reason for higher expected voluntary churn is the salience of the payment to the customer. Customers who make cash payments are expected to compare the benefits received from the service against the cost incurred more frequently and more intensively than those making payments via credit card. In this market, the customers who pay by cash must travel to designated collection centers every month to make the payments.

On the other hand, customers who pay by credit card see the cost of the service in an aggregated bill along with other credit card purchases and transactions. One of the features of credit cards is that they aggregate many small “losses” into one larger loss for the customer; in so doing they may reduce the total perceived value lost (Thaler 1995). Further, in a market where credit card penetration is relatively low (61% among paid TV customers) payment in cash may be a reflection of the poor credit worthiness of the customer. We expect differences in the impact across the two types of churn rates.
Control Variables

We also control for the effect of variables which could have an impact on the churn probabilities of the customers. The variable ARREARS measures the amount of arrears that the customer owes to the firm. We expect the coefficient to be positive as the churn rate is expected to increase with larger arrear amounts. DPHONE is an indicator variable which takes on a value of 1 if the customer has a phone connection and 0 otherwise. We expect the coefficient to be negative for voluntary churn. The greater the ability of the firm to contact the customer (e.g., to make an offer or address a problem), the lower the expected voluntary churn rate. Again, we expect differences in the impact of the covariates on involuntary churn rates. (See Appendix A for definitions of the variables).

4.4 Empirical Results and Discussion

Model Selection

We estimated the univariate parametric survival models – Exponential, Logistic and Weibull as well as the semi-parametric survival model Exponential with time intercepts\textsuperscript{20} for the three data samples discussed in the previous section to understand the behavior of the underlying churn processes. We found the Weibull parametric survival model had the best fit with the lowest BIC values in all the three samples. (See Tables 4.2, 4.3 and 4.4 for details).

\textsuperscript{20} We used monthly time interval for the Exponential model with time intercepts – the most detailed level possible in the data sample.
We estimated the bivariate Weibull survival model for *data sample 3* (voluntary and involuntary churn) with time varying covariates as the *main model*. This model allows us to check for any dependence between the two competing events in our data sample – voluntary and involuntary churn. We also estimated the univariate Weibull model for *data sample 1* (voluntary churn only) with time varying covariates as the *benchmark model*. We estimated the models with variables entering in different functional forms (linear, logarithmic or quadratic) and report the best fitting models (lowest BIC values).

The best fitting univariate Weibull model with time varying covariates had a loglikelihood value of -787.77 and a BIC value of 1656. The best fitting bivariate Weibull model with time varying covariates (case $c_2 > c_1$) had a loglikelihood value of -1625.7 and a BIC value 3323.3. It is important to note that the goodness of fit measures (Loglikelihood and BIC) cannot be compared across the benchmark univariate Weibull model and the main bivariate Weibull model as the sample sizes are different.

*Model Parameters – Impact of covariates on different types of churn*

A study of Table 4.6 with the estimated response parameters for the main model, shows that 5 out of 7 of the parameters ($\beta_{30}, \beta_{31}, \beta_{32}, \ldots, \beta_{36}$) which influence $\lambda_3$ are significant. Computation of the derived distributional parameters $\lambda_1, \lambda_2$ and $\lambda_3$ of the bivariate Weibull survival model along with their t-statistic values shows that
the $\lambda_3$ parameter was significant (t-statistic value of 6.43)$^{21}$ with a value of 0.112. This parameter measures the dependence between the voluntary and involuntary churn process. Therefore, we cannot assume independence between voluntary and involuntary churn. Ignoring involuntary churn would lead to a bias in the estimates of voluntary churn rates as there is clear dependence between the two churn processes and we would be ignoring correlated information.

In addition, a study of the parameter values from the benchmark univariate model and the main bivariate model shows the impact of ignoring involuntary churn in the response parameters that influence voluntary churn. Tables 4.5 (univariate model) and Table 4.6 (main model) show differences in the value, sign and significance of the parameter estimates. For example, the variable NSERVICE is negative and significant (-0.078 and t-statistic value of -2.485) in the univariate model. In the main model, the combination of the response parameters ($\beta_{11}$ and $\beta_{31}$) representing the influence of NSERVICE on voluntary churn is positive and significant (0.119 and t-statistic value of 3.799) for majority of the customers$^{22}$. Similarly, the variable ln(INVOICE) is

---

$^{21}$ We estimated the standard deviation and t-statistic values for the $\lambda_i$ using their empirical distribution. The empirical distribution (10000 draws) was generated using the parameter estimates $\beta_i$ and their variance-covariance matrix. The $\lambda_i$ were calculated at the mean value of the covariates $Z$.

$^{22}$ The voluntary churn rate is defined differently for customers whose observed duration time $X_i \leq 1$ (region $k = 1$) than for customers with $X_i > 1$ (region $k = 2$). In the estimation sample 83 % of the customers who experienced voluntary churn were in the region $k = 1$. In this region for the best fitting case ($c_2 > c_1$), the voluntary churn rate is influenced by a combination of the composite parameters $\lambda_i$ and $\lambda_3$ (See Appendix C section II for more details).
positive and significant (1.637 and t-statistic value of 12.887) in the univariate model and in the main model the combination of ($\beta_{13}$ and $\beta_{33}$) representing the influence of \(\ln(\text{INVOICE})\) on voluntary churn is positive and significant (4.013 and t-statistic value of 18.753) for majority of the customers.

In the main bivariate model, the impact of the covariates on the two types of churn is different. Table 4.6 shows the differences in the value and significance of the parameter estimates. For example, the combination of the response parameters ($\beta_{11}$ and $\beta_{31}$) representing the influence of \(\text{NSERVICE}\) on voluntary churn is positive and significant (0.119 and t-statistic value of 3.79) for majority of the customers. But the response parameter $\beta_{21}$ representing the influence of \(\text{NSERVICE}\) on involuntary churn is negative and insignificant (-0.001 and t-statistic value of -0.02) for majority of the customers. Similarly, the combination of response parameters ($\beta_{15}$ and $\beta_{35}$) representing the influence of $\ln(\text{INVOICE})$ on voluntary churn is negative and significant (-0.495 and t-statistic value of -5.11) for majority of the customers. The response parameter $\beta_{25}$ representing the influence of $\ln(\text{INVOICE})$ on involuntary churn is negative and significant (-3.745 and t-statistic value of -43.21) for majority of the customers. Thus, the covariates have a differential impact on the two types of churn.

---

23 The involuntary churn rate is defined differently for customers whose observed duration time $X_2 \leq 1$ (region $k = 1$) than for customers with $X_2 > 1$ (region $k = 2$). In the estimation sample 89% of the customers who experienced involuntary churn were in the region $k = 1$. In this region for the best fitting case ($c_2 > c_1$), the involuntary churn rate is influenced by the composite parameters $\lambda_2$ (See Appendix C section II for more details).
churn. This would not have been highlighted if we had used the benchmark univariate model.

**Model Comparison – Impact of covariates on voluntary churn rates**

The impact of the covariates on voluntary churn can be examined more clearly by looking at metrics such as the percentage change in hazard rate\(^{24}\) (churn rate) or the elasticity\(^{25}\) values. Table 4.7 compares the impact on voluntary churn rates for the two models – benchmark model and main model. For example, an increase in NSERVICE – the cumulative number of contacts made by the customer to the customer service unit – leads to an increase in the churn rate of 8.06% with an elasticity value of 0.54 in the benchmark model. But in the final model, an increase in NSERVICE leads to a decrease of 1.35% in the voluntary churn rate with an elasticity value of -0.09.

The model a manager uses can lead to very different diagnostics and managerial interventions. Looking at the benchmark model, one would expect the customers’ overall service experience to have been not satisfactory leading to an increase in voluntary churn rates. But the results from the main model show that the overall customer service experience may have been positive as the voluntary churn rates actually decline.

---

\(^{24}\) Percentage change in hazard rate = Percentage change in the hazard rate when the variable value (in the functional form) changes from 0 to 1

\(^{25}\) Elasticity = Percentage change in the hazard rate / Percentage change in the original value of the variable
The impact of an increase in INVOICE – the cumulative invoiced amount is larger on voluntary churn in the benchmark univariate model (decrease in the churn rate of 80.55% with elasticity of -1.61) than in the main bivariate model (decrease of 68.43% with elasticity of -1.25). Similarly, the impact of an increase in ARREARS – the cumulative arrears owed by the customer is larger in the benchmark model (elasticity of 1.54) than in the main model (elasticity value of 0.26).

The increase in invoice amount leads to greater decrease in churn rates in the benchmark model which ignores involuntary churn than in the main model. This has implications for managers when they make decisions on the cost-benefits of a marketing intervention to persuade customers to use higher value packages with larger invoice amounts. Thus, ignoring involuntary churn and looking at the benchmark model with voluntary churners only could lead to very different results and managerial actions to bring about changes in the churn rates.

An additional aspect of the main bivariate model is that it allows us to study the differential impact of covariates on voluntary and involuntary churn rates. We find differences in the response parameters in the magnitude, sign as well as significance levels for the two churn rates (See Table 4.8 for details). For example, DRECOVERY – an indicator variable when a customer contacted the customer retention center – shows a large impact (increase) on voluntary churn rate (elasticity value of 11.865) but has a small and opposite impact on involuntary churn rate.
The impact of an increase in the invoice amount (INVOICE) leads to greater decline in involuntary churn rate (elasticity value of -1.796) than on voluntary churn rate (elasticity value of -1.247). Similarly the impact of an increase in the amount of arrears (ARREARS) is higher on involuntary churn (elasticity of 70.985) than on voluntary churn (elasticity value of 0.256). The impact of a customer having phone connectivity (DPHONE) has a significant effect only on voluntary churn rate (elasticity of -0.334).

Model performance in holdout sample

We compared the two models (benchmark versus main model) using the cross-sectional holdout sample defined earlier. We computed the deviation between the predicted probability of churn for a customer and the final outcome i.e., the customer churned. In data sample with all churners – voluntary and involuntary – the main model had a MAD (mean absolute deviation) of 0.38 and RMSE (root-mean-squared-error) of 0.55 compared to the benchmark model with MAD of 0.53 and RMSE of 1.05. Even in the data sample with only voluntary churners the main model had a MAD (mean absolute deviation) of 0.35 and RMSE (root-mean-squared-error) of 0.53 compared to the benchmark model with MAD of 0.47 and RMSE of 1.07. Thus, the main bivariate model does better than the benchmark model in predicting the customers who are more likely to churn with lower values of MAD and RMSE.

We also compared the performance of the models in terms of predicting which customers are more likely to churn using the cumulative lift curves. We examined the
prediction accuracy of the two models by comparing their *cumulative lift curves* against the *random lift curve*. The cumulative lift curve for each model shows the cumulative percentage of churners accounted for by the top x% of customers predicted as most likely to churn (Neslin et al. 2006). The random lift curve shows the percentage of churners expected to churn if the customers were randomly ordered. Higher lift indicates better prediction with greater area under the curve. Figure 4.1 shows the cumulative lift curve for the two models compared with the random lift curve. The main bivariate model performs better than the benchmark univariate model. In the all the deciles except the last one the main model has higher lifts than the benchmark model and in the last decile the lifts are equal.

4.5 Conclusions

The occurrence of two types of churn in the direct-to-home satellite TV market – *voluntary churn* and *involuntary churn* can be analyzed with the dependent competing risk framework to check the impact of ignoring involuntary churn. We estimated a bivariate Weibull survival model with data sample comprising voluntary and involuntary churn and compared it to the univariate Weibull survival model with data sample comprising voluntary churn only.

The distributional parameter $\lambda_3$ of the bivariate Weibull survival model was significant (t-statistic value of 6.43) with a value of 0.112. This parameter measures the dependence between the voluntary and involuntary churn process. Thus, ignoring
involuntary churn leads to a bias in the estimates of voluntary churn rates as there is clear dependence between the two churn processes.

The bivariate Weibull survival model allows us to study the differential impact of the covariates on voluntary and involuntary churn rates with differences in the response parameters in magnitude, sign as well as significance levels. This would not have been highlighted if we had used the benchmark univariate Weibull model which does not allow for a dependent competing risk framework.

The impact of the different covariates on voluntary churn (percentage change in hazard rates as well as elasticity values) was significantly different across the two models. Thus, ignoring involuntary churn and looking at the benchmark model with voluntary churners only could lead to very different results and managerial actions to bring about changes in the churn rates. Further, the bivariate model does better than the benchmark model in predicting the customers who are more likely to churn in a cross-sectional holdout sample.

In future research, a potential area for focus is a detailed study of the factors that influence involuntary churn using additional firm-specific data samples. A study of the dependence between the two types of churn especially factors that influence this dependence will throw greater light on this problem. In addition, a study of the association between the degree of dependence between the two churn types and the bias in the estimates of voluntary churn rates using a simulated data study would be interesting. This will give managers additional methods to identify dependence
between the two types of churn and the factors that influence it as well as the impact of ignoring the involuntary churn rates.
### Variable Description

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Customer Service Experience</strong></td>
<td></td>
</tr>
<tr>
<td>$\text{NSERVICE}_{jt}$</td>
<td>Cumulative number of customer service contacts</td>
</tr>
<tr>
<td><strong>Failure Recovery</strong></td>
<td></td>
</tr>
<tr>
<td>$\text{DRECOVERY}_{jt}$</td>
<td>Indicator variable: Customer recovery contact</td>
</tr>
<tr>
<td><strong>Payment Equity</strong></td>
<td></td>
</tr>
<tr>
<td>$\text{INVOICE}_{jt}$</td>
<td>Cumulative invoice amount</td>
</tr>
<tr>
<td>$\text{DCASH}_{j}$</td>
<td>Indicator variable: Payment by cash</td>
</tr>
<tr>
<td><strong>Control Variables</strong></td>
<td></td>
</tr>
<tr>
<td>$\text{ARREARS}_{jt}$</td>
<td>Cumulative amount of arrears</td>
</tr>
<tr>
<td>$\text{DPHONE}_{j}$</td>
<td>Indicator variable: Phone connectivity</td>
</tr>
</tbody>
</table>
Appendix B: Simulated Data Set

We generated a dependent competing risk data set with observed duration times $X_1$ and $X_2$, right censored duration times $X_3$ and a time varying covariate $Z$. We wrote the code in MATLAB software program and the main steps are described below:

1. Generate random uniform variables $U_1$, $U_2$ and $U_3$ from a standard uniform distribution.
2. Define $E_i = -\left(1 - \ln U_i\right)/\lambda_i$, $i = 1, 2, 3$ $E_i$ are i.i.d exponential variables with parameters $\lambda_i$ (Luc Devroye 1986).
3. Define $D_i = \min\left(E_1/\lambda_1, E_3/\lambda_3\right)$, $i = 1, 2$ $(D_1, D_2)$ follow a bivariate exponential distribution with parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$ (Marshall and Olkin 1967, Luc Devroye 1986)
4. Define $Y_i = D_i^{-1}$, $i = 1, 2$ $(Y_1, Y_2)$ follow a bivariate Weibull distribution with parameters $\lambda_1$, $c_1$, $\lambda_2$, $c_2$ and $\lambda_3$ (Marshall and Olkin 1967, Moeschberger 1974)
5. Generate $Y_3 = 2 * \text{Uniform}(0,1)$ $Y_3$ are randomly distributed right censored duration times which are independent of the latent duration times $Y_1$ and $Y_2$.
6. Define $X_1 = \{Y_1 \mid Y_1 \leq Y_2 \land Y_1 \leq Y_3\}$, $X_2 = \{Y_2 \mid Y_2 \leq Y_1 \land Y_2 \leq Y_3\}$ and
\[ X_3 = \{ Y_3 \mid Y_3 < Y_1 \text{ and } Y_3 < Y_2 \} \]

\( X_1 \) and \( X_2 \) are the observed event times and \( X_3 \) are the right censored survival times.

Each customer was assigned a unique customer id and an indicator variable for customer type – experienced event 1, experienced event 2 or was right censored.

7. Split the observed customer duration times into equal time episodes of length 0.1

8. Generate a time varying covariate vector \( Z \) from standard normal distribution. The covariate has a different value for each customer-time episode.
Appendix C: Conditional probability density, survival probability and hazard rate for the two cases \( c_1 > c_2 \) and \( c_2 > c_1 \)

I. Case \( c_1 > c_2 \):

**Conditional probability density** \( f_i(x_{jk}) \)

For customer \( j \) who experiences voluntary churn (event \( i = 1 \)) with observed duration time in region \( k = 1 \) (\( x_{1ji} \leq 1 \)), the conditional probability density is

\[
f_1(x_{1ji}) = \lambda_1 c_1 x_{1ji}^{c_1-1} \exp\left\{ -\left( \lambda_1 x_{1ji} + \lambda_2 + \lambda_3 \right) x_{1ji}^{c_1} \right\}
\]

For customer \( j \) who experiences voluntary churn (event \( i = 1 \)) with observed duration time in region \( k = 2 \) (\( x_{1j2} > 1 \)), the conditional probability density is

\[
f_1(x_{1j2}) = (\lambda_1 + \lambda_3) c_1 x_{1j2}^{c_1-1} \exp\left\{ -\left( \lambda_1 + \lambda_3 \right) x_{1j2}^{c_1} \right\}
\]

For customer \( j \) who experiences involuntary churn (event \( i = 2 \)) with observed duration time in region \( k = 1 \) (\( x_{2ji} \leq 1 \)), the conditional probability density is

\[
f_2(x_{2ji}) = (\lambda_2 + \lambda_3) c_2 x_{2ji}^{c_2-1} \exp\left\{ -\left( \lambda_2 x_{2ji} + \lambda_3 \right) x_{2ji}^{c_2} \right\}
\]

For customer \( j \) who experiences involuntary churn (event \( i = 2 \)) with observed duration time in region \( k = 2 \) (\( x_{2j2} > 1 \)), the conditional probability density is

\[
f_2(x_{2j2}) = \lambda_2 c_2 x_{2j2}^{c_2-1} \exp\left\{ -\left( \lambda_1 + \lambda_3 \right) x_{2j2}^{c_2} \right\}
\]

**Survival probability** \( w_{lk} \)

For customer \( l \) with observed right censored duration time in region \( k = 1 \) (\( w_{l1} \leq 1 \)), the survival probability is
$$S(w_{ij}) = \exp\left\{-\lambda_{ij} w_{ij}^{c_i} - (\lambda_j + \lambda_k) w_{ij}^{c_k}\right\}$$

For customer $i$ with observed right censored duration time in region $k = 2$ ($w_{ij} > 1$), the survival probability is

$$S(w_{ij}) = \exp\left\{- (\lambda_j + \lambda_k) w_{ij}^{c_i} - \lambda_k w_{ij}^{c_k}\right\}$$

**Hazard rate $h_i(x_{ijk})$**

For customer $j$ who experiences voluntary churn (event $i = 1$) with observed duration time in region $k = 1$ ($x_{ij1} \leq 1$), the hazard rate is

$$h_1(x_{ij1}) = -\lambda_i c_1 x_{ij1}^{c_i-1}$$

For customer $j$ who experiences voluntary churn (event $i = 1$) with observed duration time in region $k = 2$ ($x_{ij2} > 1$), the hazard rate is

$$h_1(x_{ij2}) = -(\lambda_i + \lambda_k) c_1 x_{ij2}^{c_i-1}$$

For customer $j$ who experiences involuntary churn (event $i = 2$) with observed duration time in region $k = 1$ ($x_{ij1} \leq 1$), the hazard rate is

$$h_2(x_{ij1}) = - (\lambda_j + \lambda_k) c_2 x_{ij1}^{c_2-1}$$

For customer $j$ who experiences involuntary churn (event $i = 2$) with observed duration time in region $k = 2$ ($x_{ij2} > 1$), the hazard rate is

$$h_2(x_{ij2}) = -\lambda_k c_2 x_{ij2}^{c_2-1}$$
II. Case $c_2 > c_1$ (best fitting case):

**Conditional probability density** $f_i(x_{ijk})$

For customer $j$ who experiences voluntary churn (event $i = 1$) with observed duration time in region $k = 1$ ($x_{ij1} \leq 1$), the conditional probability density is

$$f_1(x_{ij1}) = (\lambda_1 + \lambda_3)c_1x_{ij1}^{c_1-1}\exp\left\{- (\lambda_1 + \lambda_3)x_{ij1} - \lambda_2x_{ij1}^{c_2}\right\}$$

For customer $j$ who experiences voluntary churn (event $i = 1$) with observed duration time in region $k = 2$ ($x_{ij2} > 1$), the conditional probability density is

$$f_1(x_{ij2}) = \lambda_1c_1x_{ij2}^{c_1-1}\exp\left\{- \lambda_1x_{ij2} - (\lambda_2 + \lambda_3)x_{ij2}^{c_2}\right\}$$

For customer $j$ who experiences involuntary churn (event $i = 2$) with observed duration time in region $k = 1$ ($x_{ij2} \leq 1$), the conditional probability density is

$$f_2(x_{ij1}) = \lambda_2c_2x_{ij2}^{c_2-1}\exp\left\{- (\lambda_1 + \lambda_3)x_{ij1} - \lambda_2x_{ij1}^{c_2}\right\}$$

For customer $j$ who experiences involuntary churn (event $i = 2$) with observed duration time in region $k = 2$ ($x_{ij2} > 1$), the conditional probability density is

$$f_2(x_{ij2}) = (\lambda_2 + \lambda_3)c_2x_{ij2}^{c_2-1}\exp\left\{- \lambda_1x_{ij2} - (\lambda_2 + \lambda_3)x_{ij2}^{c_2}\right\}$$

**Survival probability** $w_{lk}$

For customer $l$ with observed right censored duration time in region $k = 1$ ($w_{li} \leq 1$), the survival probability is

$$S(w_{li}) = \exp\left\{- (\lambda_1 + \lambda_3)w_{li}^{c_1} - \lambda_2w_{li}^{c_2}\right\}$$
For customer $l$ with observed right censored duration time in region $k = 2$ ($w_{l1} > 1$), the survival probability is

$$S(w_{l2}) = \exp\left\{-\lambda_1 w_{l2}^c - (\lambda_2 + \lambda_3) w_{l2}^c \right\}$$

**Hazard rate** $h_i(x_{ijk})$

For customer $j$ who experiences voluntary churn (event $i = 1$) with observed duration time in region $k = 1$ ($x_{ij1} \leq 1$), the hazard rate is

$$h_1(x_{ij1}) = -(\lambda_1 + \lambda_3) c_1 x_{ij1}^{c_1 - 1}$$

For customer $j$ who experiences voluntary churn (event $i = 1$) with observed duration time in region $k = 2$ ($x_{ij2} > 1$), the hazard rate is

$$h_1(x_{ij2}) = -\lambda_1 c_1 x_{ij2}^{c_1 - 1}$$

For customer $j$ who experiences involuntary churn (event $i = 2$) with observed duration time in region $k = 1$ ($x_{2j1} \leq 1$), the hazard rate is

$$h_2(x_{2j1}) = -\lambda_2 c_2 x_{2j1}^{c_2 - 1}$$

For customer $j$ who experiences involuntary churn (event $i = 2$) with observed duration time in region $k = 2$ ($x_{2j2} > 1$), the hazard rate is

$$h_2(x_{2j2}) = -(\lambda_2 + \lambda_3) c_2 x_{2j2}^{c_2 - 1}$$
Table 4.1: Summary Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer Service Experience</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSERVICE</td>
<td>12.45</td>
<td>13.88</td>
<td>8.00</td>
</tr>
<tr>
<td>Failure Recovery</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRECOVERY</td>
<td>0.10</td>
<td>0.30</td>
<td>0.00</td>
</tr>
<tr>
<td>Payment Equity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INVOICE</td>
<td>655.55</td>
<td>301.12</td>
<td>670.42</td>
</tr>
<tr>
<td>DCASH</td>
<td>0.43</td>
<td>0.49</td>
<td>0.00</td>
</tr>
<tr>
<td>Control Variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARREARS</td>
<td>61.88</td>
<td>95.29</td>
<td>0.00</td>
</tr>
<tr>
<td>DPHONE</td>
<td>0.75</td>
<td>0.43</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: Sample size: 4426 customers
Table 4.2: Model Selection (Voluntary and Involuntary Churn Sample)

<table>
<thead>
<tr>
<th>Model</th>
<th>LogL</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull</td>
<td>-2152.63</td>
<td>4325.80</td>
</tr>
<tr>
<td>Exponential</td>
<td>-2627.53</td>
<td>5265.34</td>
</tr>
<tr>
<td>Piece-wise Exponential</td>
<td>-2556.35</td>
<td>5225.68</td>
</tr>
<tr>
<td>Logistic</td>
<td>-2212.17</td>
<td>4444.89</td>
</tr>
</tbody>
</table>
Table 4.3: Model Selection (Voluntary Churn Sample)

<table>
<thead>
<tr>
<th>Model</th>
<th>LogL</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull</td>
<td>-1483.94</td>
<td>2988.00</td>
</tr>
<tr>
<td>Exponential</td>
<td>-1706.38</td>
<td>3422.81</td>
</tr>
<tr>
<td>Piece-wise Exponential</td>
<td>-1512.94</td>
<td>3146.56</td>
</tr>
<tr>
<td>Logistic</td>
<td>-1510.16</td>
<td>3040.42</td>
</tr>
</tbody>
</table>
Table 4.4: Model Selection (Involuntary Churn Sample)

<table>
<thead>
<tr>
<th>Model</th>
<th>LogL</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull</td>
<td>-1517.46</td>
<td>3055.03</td>
</tr>
<tr>
<td>Exponential</td>
<td>-1755.50</td>
<td>3521.06</td>
</tr>
<tr>
<td>Piece-wise Exponential</td>
<td>-1550.88</td>
<td>3222.43</td>
</tr>
<tr>
<td>Logistic</td>
<td>-1525.87</td>
<td>3071.85</td>
</tr>
</tbody>
</table>
### Table 4.5: Univariate Weibull Model – Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter</th>
<th>t-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Customer Service Experience</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSERVICE</td>
<td>-0.078</td>
<td>-2.485</td>
</tr>
<tr>
<td><strong>Failure Recovery</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRECOVERY</td>
<td>-2.330</td>
<td>-17.452</td>
</tr>
<tr>
<td><strong>Payment Equity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln(INVOICE)</td>
<td>1.637</td>
<td>12.887</td>
</tr>
<tr>
<td>DCASH</td>
<td>-0.832</td>
<td>-8.915</td>
</tr>
<tr>
<td><strong>Control Variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln (ARREARS)</td>
<td>-0.482</td>
<td>-5.582</td>
</tr>
<tr>
<td>DPHONE</td>
<td>0.443</td>
<td>5.477</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.201</td>
<td>0.936</td>
</tr>
<tr>
<td>Sigma</td>
<td>0.388</td>
<td>22.119</td>
</tr>
</tbody>
</table>

Note: Values in bold were significant with t-statistic > 2
- a Scaled by 10 and standardized.
- b Indicator Variable
- c Scaled by 100, standardized and a constant value of 2.297 added to ensure minimum value is 1
- d Scaled by 10, standardized and a constant value of 2.223 added to ensure minimum value is 1
Table 4.6: Bivariate Weibull Model: Parameters Estimates

| Variable                          | Parameter 1 | Parameter 2 | Parameter 3 | t statistic | Parameter 1 | Parameter 2 | Parameter 3 | t statistic | Parameter 1 | Parameter 2 | Parameter 3 | t statistic |
|----------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| INTERCEPT                        | 2.119       | 18.88       | -0.025      | -0.24       | -0.044      | -0.18       |
| Customer Service Experience     |             |             |             |             |             |             |             |             |
| NSERVICE €                      | -0.100      | -2.76       | -0.001      | -0.02       | 0.219       | 2.61        |
| Failure Recovery                |             |             |             |             |             |             |             |             |
| Payment Equity                  |             |             |             |             |             |             |             |             |
| ln(INVOICE £)                   | -0.455      | -3.47       | 4.227       | 35.65       | 4.468       | 18.43       |
| DCASH ±                         | -0.642      | -5.07       | -0.055      | -0.83       | -0.914      | -7.19       |
| Control Variables               |             |             |             |             |             |             |             |             |
| ln(ARREARS ¥)                   | 0.338       | 2.78        | -3.745      | -43.21      | -0.833      | -4.29       |
| DPHONE ±                        | 0.642       | 6.16        | -0.055      | -0.78       | -0.002      | -0.01       |
| Distributional Parameters       |             |             |             |             |             |             |             |             |
| λ ⊗                             | 0.126       | 12.21       | 0.378       | 16.82       | 0.112       | 6.43        |
| c                               | 2.719       | 32.41       | 3.545       | 43.97       |

Note: € Scaled by 10 and standardized.
£ Scaled by 100, standardized and constant value of 2.266 added to ensure a minimum value of 1
¥ Scaled by 10, standardized and constant value of 1.877 added to ensure a minimum value of 1
± Indicator variables
⊗ Derived parameters at the mean value of the variables
Values in bold were significant with t-statistic > 2
### Table 4.7 Model Comparison – Impact of Covariates on Voluntary Churn

<table>
<thead>
<tr>
<th>Variable</th>
<th>Univariate Weibull Model</th>
<th>Bivariate Weibull Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percentage change in hazard</td>
<td>Elasticity</td>
</tr>
<tr>
<td>Customer Service Experience</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSERVICE</td>
<td>8.06</td>
<td>0.54</td>
</tr>
<tr>
<td>Failure Recovery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRECOVERY</td>
<td>927.54</td>
<td>9.28</td>
</tr>
<tr>
<td>Payment Equity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln(INVOICE)</td>
<td>-80.55</td>
<td>-1.61</td>
</tr>
<tr>
<td>DCASH</td>
<td>129.89</td>
<td>1.30</td>
</tr>
<tr>
<td>Control Variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln (ARREARS)</td>
<td>61.94</td>
<td>1.54</td>
</tr>
<tr>
<td>DPHONE</td>
<td>-35.80</td>
<td>-0.36</td>
</tr>
</tbody>
</table>

Note: Percentage change in hazard=% change in hazard when the functional form of the variable changes from 0 to 1

Elasticity= %change in hazard / % change in the value of the original variable
Table 4.8: Bivariate Weibull Model – Impact of Covariates on Voluntary and Involuntary Churn

<table>
<thead>
<tr>
<th>Variable</th>
<th>Voluntary Churn</th>
<th>Involuntary Churn</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percentage change in hazard</td>
<td>Elasticity</td>
</tr>
<tr>
<td>Customer Service Experience</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSERVICE</td>
<td>-1.35</td>
<td>-0.086</td>
</tr>
<tr>
<td>Failure Recovery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRECOVERY</td>
<td>1186.540</td>
<td>11.865</td>
</tr>
<tr>
<td>Payment Equity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln(INVOICE)</td>
<td>-68.430</td>
<td>-1.247</td>
</tr>
<tr>
<td>DCASH</td>
<td>111.780</td>
<td>1.118</td>
</tr>
<tr>
<td>Control Variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln (ARREARS)</td>
<td>13.980</td>
<td>0.256</td>
</tr>
<tr>
<td>DPHONE</td>
<td>-33.380</td>
<td>-0.334</td>
</tr>
</tbody>
</table>

Note: Percentage change in hazard= % change in hazard when the functional form of the variable changes from 0 to 1
Elasticity = %change in hazard / % change in the value of the original variable
Figure 4.1: Cumulative Lift Chart

[Graph showing cumulative lift chart with lines for Univariate Model, Bivariate Model, Random, and Perfect models. The x-axis represents the percentage of customers in descending order of churn probability, and the y-axis represents the percentage of churners.]

[Legend: ▲ Univariate Model, ■ Bivariate Model, ● Random, ○ Perfect]
4.7 References


CHAPTER 5

Conclusions

5.1 Overall Summary

The advent of customer centric organizations has spurred the need among managers for models that help them to take actions at the customer level and in the context of a dynamic customer-firm relationship. The focus of the research was accurate prediction of customer retention and customer churn and to establish empirically the impact of different factors on these rates. The research was conducted in different customer-firm relationship contexts – customer visits to an Internet recommendation web site and subscription to the service of a direct-to-home satellite television producer. The research also questioned the assumption of fixed customer retention rates across different customers and over time through a customer’s relationship with the firm.

The assumption of fixed customer retention rates to compute customer metrics such as customer lifetime value (CLV) is questionable. Intuitively, one would expect the rates to differ across customers and over time. For example, for a website, the probability of customer retention might increase as customers become more acquainted with a site and they learn to better use and navigate the site (Johnson 2003). Further, customer retention is also expected to vary across customers with some customers displaying greater stickiness or loyalty than others.
In this research, the emphasis has been on linking empirically different factors which are under the manager’s control to customer retention and customer churn. The information used to estimate the models is readily available to managers. This included clickstream data which track a customer’s onsite behavior at an Internet site or secondary data from customer relationship management software. This latter type of data set tracks customer specific transaction and demographic information at the continuous subscription product firm. We use this information to build modeling constructs identified in the extant literature as impacting customer retention and customer churn but typically measured with survey and experimental data. Customer surveys or experiments are expensive to execute and based on a customer sample from the firm’s customer base. The use of secondary data easily available to managers can provide them with less expensive options to gain insights into customer behavior and help them to use survey and other tools more effectively. We hope our research will help managers to more effectively use information which they collect on an ongoing basis to analyze customer retention and churn.

The modeling approach to analyze the data in the three research problems was based on survival modeling techniques. The modeling technique was driven by the research problem and the data available which were customer level duration time observations with some of the observations right censored. In the last research problem, the data also had the presence of dependent competing events – voluntary and involuntary customer churn. Survival models are well established in the statistics
literature. They have been shown to have better stability, face validity and predictive accuracy for analysis of duration data than other methods such as logistic and least squares regressions (Helsen and Schmittlein 1993). These models have also been utilized in the marketing and CRM literature (e.g., Jain and Vilcassim 1991, Bolton 1998).

In essay 1, we modeled the prediction of customer retention in the context of repeat visits to a website. We incorporated heterogeneity in customer retention over time by allowing customer retention probabilities to vary across first and subsequent return visits. We also showed the effects of customer fit (how well the product meets the customer’s requirement), switching costs and customer interactions on customer retention at an Internet recommendation site. Using an extension of the Cox proportional hazard model for multiple events, we established a direct and significant quantifiable link between customer retention and customer fit, as well as other determinants such as customer interactions and switching costs. Further, we showed that this impact varies with the depth of repeat. Our results suggest that customer fit during the first visit significantly affects the customer’s retention probability. We also find a significant positive impact of switching costs on customer retention, indicating that switching costs continue to be relevant in the Internet world. Our findings also point to the relevance of a customer’s interaction with the site. Initial interactions, especially during the first visit, can significantly influence customer retention.
The various factors we modeled had a maximum impact during the customer’s first visit or acquisition stage with the effects declining or becoming insignificant over subsequent visits. Our results are in line with the acquisition process perspective which states that acquisition includes, in addition to the first purchase, other non-purchase encounters that precede and follow purchase up until the time the customer makes a repeat purchase (Blattberg et al. 2001). During this acquisition period, a bonding or development stage for the customer-firm relationship, the customer forms attitudes about the firm’s product and ancillary services which affect the customer’s repurchase decision.

In the case of the Internet recommendation site, acquisition would be defined as the customer visiting the site and interacting with the site through rating items and receiving recommendations in return. Managers can impact the customer retention rates during this stage through improvements in the accuracy of their recommendation engine, increasing customer interactions by cross promoting other categories within the site and providing additional value to the customers using their information history with the site.

In essay 2, we modeled the prediction of customer churn in the context of a continuous subscription product i.e., direct-to-home satellite television. The research highlighted the importance of incorporating heterogeneity in the customer churn rates across customers with differences not only in the baseline customer churn rates but also in how customers respond to changes in different drivers of churn.
We empirically linked customer churn to customer service experience, failure recovery (how well the firm performs in its recovery efforts after a service failure) and payment equity. We found that the three segment heterogeneous Weibull model provided the best fit to our data. In addition, this model outperformed a benchmark single segment homogeneous Weibull model in a holdout sample, giving lower log likelihood, RMSE and MAD values. It also had a larger top decile lift and its cumulative lift curve enveloped the single segment model. Differences in magnitude, but not sign, were also found in the segment-level response parameters pertaining to variables for customer service experience, failure recovery, and payment equity. Estimation results from our data present a convincing case for the benefits of incorporating heterogeneity into the hazard modeling approach for predicting customer churn.

We also found that conventional demographic variables such as age, gender, marital status were not predictive of the churn. The segments were also very similar in their demographic profiles. The situation may be even more challenging in emerging markets given that the customer base skews toward urban and upper income households. Such challenges highlight the need to develop other approaches to predicting churn, such as taking greater advantage of the observed customer histories available within a firm’s in-house data.

In essay 3, we modeled the prediction of different types of customer churn for a continuous subscription product. We showed the importance of accounting for all
types of customer churn – voluntary as well as involuntary when estimating the churn probabilities. We showed that ignoring the existence of involuntary churn significantly impacts the prediction and diagnosis of voluntary churn using the dependent competing risks framework. We estimated a bivariate Weibull survival model with data sample comprising voluntary and involuntary churn and compared it to the univariate Weibull survival model with data sample comprising voluntary churn only.

The distributional parameter in the bivariate Weibull model that measures the dependence between the voluntary and involuntary churn process was significant. We found that ignoring involuntary churn leads to biased estimates of voluntary churn rates because there was clear dependence between the two churn processes. The bivariate Weibull survival model allowed us to study the differential impact of the covariates on voluntary and involuntary churn rates. The response parameters differed in magnitude, sign and significance levels. This would not have been evident if we had used the benchmark univariate Weibull model which does not allow for a dependent competing risk framework.

The impact of the different covariates on voluntary churn (percentage change in hazard rates as well as elasticity values) was significantly different across the two models. Thus, ignoring involuntary churn and looking at the benchmark model with voluntary churners only could lead to very different results and managerial actions to bring about changes in the churn rates. The bivariate model also performed better than
the benchmark model in predicting customers who are more likely to churn in a cross-sectional holdout sample.

I hope my research will help managers to better estimate customer retention and customer churn and to better understand factors that influence these rates. The empirical links between factors under a manager’s control and these rates will provide managers with the tools to make better marketing intervention decisions. There is much scope for future research in the area of improving the prediction of customer retention and customer churn. For example, there could be possible dependencies between the acquisition and the retention process which could impact the churn rates (Thomas 2001). Our research currently does not incorporate the acquisition process due to a lack of specific data about the firm’s acquisition process at the customer level. If suitable data could be obtained, it would be desirable to incorporate the customer acquisition process into a unified modeling approach. In addition, an understanding of the differences between the voluntary and involuntary churn processes would be of interest.
5.2 References


