A Churn-Strategy Alignment Model for Telecom Industry

by

Wei Yu

A Thesis submitted to
Saint Mary's University, Halifax, Nova Scotia
in Partial Fulfillment of the Requirements for
the Degree of Masters of Science in Applied Science

September, 2005, Halifax, Nova Scotia

Approved: Dr. Dawn Jutla
Co-Superviser

Approved: Dr. Shyamala Sivakumar
Co-Superviser

Approved: Dr. Sunny Marche
External Examiner

Approved: Dr. Hai Wang
Internal Examiner

Date: September 6, 2005

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-494-15261-3
Our file Notre référence
ISBN: 0-494-15261-3

Certification

Name: Wei Yu

Degree: Master of Science in Applied Science

Title of Thesis: A Churn-Strategy Alignment Model for Telecom Industry

Examining Committee:

Dr. Kevin Vessey, Dean of Graduate Studies
Saint Mary's University

Dr. David H. S. Richardson, Program Co-ordinator
Saint Mary's University

Dr. Sunny Marche, External Examiner
Dalhousie University

Dr. Dawn Jutla, Senior Co-Supervisor
Saint Mary's University

Dr. Shyamala Sivakumar, Senior Co-Supervisor
Saint Mary's University

Dr. Hai Wang, Supervisory Committee
Saint Mary's University

Date Certified: September 6, 2005

Acknowledgement

I express my appreciation to Dr. Dawn N.Jutla, and Dr. Shyamala C. Sivakumar, for their time, guidance, and valuable contribution in supervising this research.

My special appreciation goes to my husband, Andrew, and my son, Barak, for their love and support provided during my studies.

I would like to thank Teradata Center for Customer Relationship Management at Duke University, for supporting this research by providing us with the data set and encouraging comments on our research.

Abstract

A Churn-Strategy Alignment Model for Telecom Industry

By

Wei Yu

August 5, 2005

Customer churn is a costly problem in the fiercely competitive telecom industry. The thesis research proposes and investigates a novel model, the Churn-Strategy Alignment Model (CSAM), for studying alignment relationships among customer churn predictors and the competitiveness strategies identified in a modified Delta model suitable for small and medium-sized (SME) telecom companies.

Exploratory factor analysis on a large data set sourced from a telecom company allowed for aggregating low-level churn predictors into business-level constructs. Managers could use knowledge around these identified business constructs to create more informed competitiveness strategies. Goodness-of-fit results from structured equation modeling did not support the CSAM model. However, middle-stage tests were informative. For example, the discriminant validity test shows that the constructs for product, customer solution, and customer profile in our modified Delta model are distinct. Further "quality", "standard usage", and "willingness to pay" are the important high-level factors of the strategies related to them respectively.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Motivation for This Research

Churn, in this thesis, stands for customers switching to a competitor. Retaining customers is a way to control churn, and it is a core issue for Customer Relationship Management (CRM) in industries with fierce competition, such as credit card issuers, insurance companies and mobile telecommunications. Kudyba and Hoptroff (2001) viewed CRM as "a dynamic process by which firms seek to retain valued customers and attract new ones". Studies have shown that it cost companies five times the amount of money in acquiring new customers as they do in retaining the existing ones (Swift, 2001). This is true even in hyper growth areas. A survey of wireless company chief executive officers showed that customer retention was viewed a more important goal than customer acquisition (Craig and Jutla, 2001).

In the telecommunications industry, churn is the loss of subscribers switching from one carrier to another (Mozer et al., 2000). In 2000, customer churn cost the U.S. industry an estimated $10 billion with annual industry churn rate of 40%; a single company with 5 million customers and annual churn rate of 30% was estimated to lose US$870 million per year (Booz Allen & Hamilton, 2001). From November 2003, the wireless carriers were required to implement Wireless Local Number Portability (WLNP) by the Federal Communications Commission (Federal Communications Commission, 2004). WLNP allows consumers to switch from one wireless carrier to another within the same local area and without changing their phone numbers. This phone number portability removed an important switching obstacle for the subscribers. It provided more opportunity for wireless carriers to attract customers; however, at the same time, they faced the greater challenge of customer retention. Due to the ferocious competition within the telecommunication industry, churn analysis continues to be an important element in the marketing analysis undertaken by telecommunication companies in the current market.

Typically, data mining techniques are used in trying to understand the churn problem (Au et al., 2003). Data mining techniques are applied to predict whether a customer will churn by building models and learning from historical data. However, most of these techniques can only provide a result that a customer may churn or not, disclose low-level details of why they churn, but seldom disclose the high-level business reasons or links to business constructs for why customers churn. Therefore, even an accurate prediction result is of little use to telecommunication company managers, especially to the strategies of customer retention. These techniques can not illustrate which high-level factors or patterns can be interpreted from the main predictors. The goal for successful business data analysis is to both discover knowledge and apply it to decision making.

In this thesis, our research will analyze customer churn from a new perspective — linking single predictors with organizational competitiveness strategies in the mobile telecommunications industry. The benefit of this research is to provide an easy and effective way for business managers to understand customer churn, and enable these managers to more easily create strategically aligned competitiveness strategies that contribute to reducing churn.

1.2 Thesis Objectives

High-level informative patterns or factors are considered easier for business managers to understand and it is also easier for them to use this information to make strategies. Hence, the objectives of this thesis are as follows:

- Propose a customer churn analysis model, which links churn predictors with organizational competitiveness strategies using interpretable high-level patterns or factors as a bridge.
- Validate and refine the proposed model using a large business data set and exploratory factor analysis and structural equation modeling techniques.
- Interpret the experimental results and highlight any new contributions to the customer churn literature.

1.3 Organization of This Thesis

This thesis is organized as following:

- Chapter 2 and Chapter 3 are literature review focusing on the business context of customer churn and the techniques for analyzing customer churn respectively. Chapter 2 reviews the organizational strategy frameworks, customer relationship management, customer retention and customer churn, and the telecom industry evolution. Chapter 3 reviews the role of data mining in CRM, data mining methodologies and tasks, churn models, and churn predictors obtained in previous researches.

- Chapter 4 presents our proposed model-- Churn-Strategy Alignment Model (CSAM) and the methodology and techniques we designed to validate and refine the model.
- Chapter 5 presents the experiment results of model validation and refinement. It also provides the interpretations of the outcome, and some application examples of the refined model.
- Chapter 6 includes our research conclusion and future work.

Chapter 2 Business Context Review

The thesis goal is to link customer churn predictors with higher-level informative factors and organizational strategies, and thus to provide an easier way for managers to understand related high-level factors about churn and to enable them to more easily create strategically aligned competitiveness strategies that contribute to reducing churn. To clearly present this research, in this chapter, we review the business context of churn. In the following four subsections, we review organizational strategy frameworks; customer relationship management; customer retention and customer churn; and the telecom industry evolution.

2.1 Organizational Strategy Frameworks Review

According to Hax and Wilde II (2003), "Competitive Positioning" and "Resource-based View", are the two most influential organizational strategy frameworks for competitiveness proposed in the past two decades.

Competitive positioning, proposed by Michael Porter in 1980's, places the industry as the central focus of strategic attention. Based on this strategy paradigm, a successful organization is "one that appropriates monopolistic rents" (Hax and Wilde II, 2003). In other words, an organization should establish itself as the dominant competitor in an industry as a whole or in a segment of an industry. Porter's logical conclusion is that only two ways are considered to compete: "low cost" and "product differentiation" (Hax and Wilde II, 2003). Low cost calls for the economies of products. Product differentiation requires that the products created are viewed as highly valuable and unique by customers.

Many approaches can be employed to reach the differentiation goal: design of brand image, technology, features, customer service, and dealer networks (Hax and Wilde II, 2003).

The Resource-based view, proposed in 1990's, places the firm as the central focus of strategic attention. It considers the value derived from resources, capabilities, and competencies. From this point of view, what makes a firm different from others is the ability to "appropriate resources that are valuable, rare, and difficult to substitute or imitate" (Hax and Wilde II, 2003). Resource-based view contributed some factors, other than product and cost, which can also generate profitability. Examples of these factors include: management skills, information capabilities and administrative process.

Both strategy frameworks are shown to contribute towards building a successful organization. However, both of them are missing an important point: the customer. Although these classic frameworks are presented like conflicting views, they actually focus on different dimensions of organizational strategies and therefore they can richly complement each other. Consequently, firms tend to win the competition by providing standardized products, distributing through mass channels, and making limited attempts to satisfy individual final customers. All of these strategies prevent firms from having a deep insight into their customers.

Instead of focusing on how to beat competitors, Hax and Wilde II (2003) argued that the way to win is to bond with customers. They believed that customers are the "ultimate repository of all the firm's activities". They provided a new strategy framework, the Delta Model, to open new sources of strategic positioning. Delta is the Greek letter that stands for transformation and change. This framework is based on customer relationship, and uses "customer bonding" as the driving force in strategy. Customer bonding refers to the "unbreakable link, deep knowledge, and close relationship" (Hax and Wilde II, 2003) with customers. These bonds are either formed directly with customers or indirectly through complementors that a customer wishes to access.

The Delta Model identifies three possible positions: "best product", "total customer solutions", and "system lock-in" (Hax and Wilde II, 2003). The triangle in Figure 2-1 represents the three strategic positions of the Delta Model. According to the model, organizations first find themselves at one of the corners, then transform from one corner to another if necessary. In reality, organization strategies are more complicated, and they may be hybrid or fall between two corners (Hax and Wilde II, 2003; Management Sciences for Health, 2003). The organizations with "best product" strategy "attract, satisfy, and retain customer" through the "inherent characteristics of the product itself". With a "total customer solutions" strategy, an organization looks for upgrading the value of its products to the customer. It adds value either by increasing the number of related products and services (solutions) offered to each customer at a single point of delivery, or bundling/combining its products and services with support and follow-up (Management Sciences for Health, 2003). Those organizations who are reaching "system lock-in" position can be considered to play a dominant role in the market. They gain complementors' share that not only assures them of customer lock-in, but also of competitor lock-out.

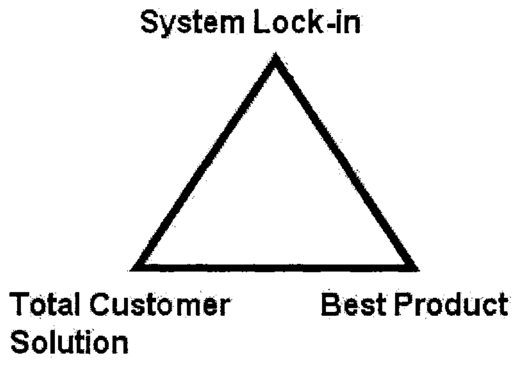


Figure 2-1 The Triangle Representation of The Delta Model (Hax and Wilde II, 2003)

2.2 Customer Relationship Management (CRM)

Understanding customers and leveraging this knowledge back to serve customers are the challenges for today's companies. Customer Relationship Management (CRM) is widely used to understand customers' needs; provide products and services that meet those needs, maximize customer satisfaction, and then obtain the revenue and profits from customers. In recent literature, there are many ways to define CRM:

- CRM is "neither a concept nor a project. Instead, it is a business strategy that aims to understand, anticipate and manage the needs of an organization's current and potential customers" (Brown, 2000, preface).
- CRM is "an enterprise approach to understanding and influencing customer behavior through meaningful communications in order to improve customer acquisition, customer retention, customer loyalty, and customer profitability" (Swift, 2001, p12).
- CRM is also viewed as "a dynamic process by which firms seek to retain valued customers and attract new ones" (Kudyba and Hoptroff, 2001, p124).

CRM turns customer information into a business decision that drives interactions with customers (Marakas, 2003, p327). Swift (2001, p40) also viewed CRM as an iterative process illustrated in Figure 2-2, which includes four major phases: Knowledge Discovery, Market Planning, Customer Interaction, and Analysis & Refinement. We will discuss each of the issues in the following paragraphs.

- Knowledge Discovery. In this phase, customer data was analyzed to identify specific market opportunities and investment strategies. It requires collecting detailed customer data from a variety of customer interactions and transactions, and transforms the data into knowledge that is usable for management and planning purposes.
- Market Planning. This phase defines specific customer strategies based on the knowledge gained. The strategies include communication plans, customer offers, delivery channels, treatment plans, and events and threshold triggers.
- Customer Interaction. This is the key action phase of executing and managing customer communications, and is the action in the plans and messages created from Knowledge Discovery and Market Planning. The outcomes of this phase generate new data for analysis & refinement. The customer interaction uses a variety of interaction channels and front office applications, such as customer care applications, sales applications, customer contact applications, and interactive applications. Through advanced technologies, these channels collect customer information, deliver marketing messages and sales opportunities, and handle service issues.

- Analysis & Refinement. This is a continuous learning phase in which data is captured and analyzed from customer interactions, and the corresponding strategies (such as price, location, communication, approaches) are refined. This refined information is then fed to the Knowledge Discovery phase as new sources.

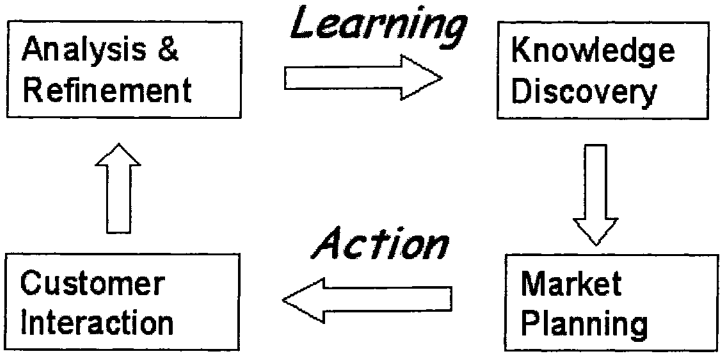


Figure 2-2 The CRM Management Process

2.3 Customer Retention and Customer Churn

Customer retention is "the ability to retain loyal and profitable customers and channels to grown the business profitably" (Swift, 2001, p42). It is one of the core issues and major objectives of CRM. Customer retention management focuses on building a deep insight into the customers, and then developing and using models to make informed decisions about which customer to attempt to retain (Swift, 2001, p78).

Contrary to customer retention, churn is viewed as the loss of customers as they switch from one organization to another (Mozer *et al.*, 2000). In the telecommunications industry, churn is broadly defined as the action that a customer's service is canceled (Lu, 2002). Churn can be either "service-provider initiated" or "customer initiated" (Lu, 2002). Service-provider initiated churn is a "competitive pull defection" (Goodwin, 2004) that the carriers terminate the service, for example, a customer's account is being closed because of payment default. Customer initiated churn is a "dissatisfaction push defection" (Goodwin, 2004) that the customers voluntarily terminate the relationship with the telecom providers; it is more complicated and the reasons behind it vary. In this thesis, we will focus on the predictors of customer initiated churn. Hence, in this thesis, churn means that customer decide to cancel the relationship with the company.

According to Gupta et al. (2003): $ChurnRate = 1 - RententionRate$

Thus customer churn and customer retention have a reverse relationship, and reducing churn rate will increase customer retention.

Recent analysis of telecom customers showed that the following are the main factors that impact customer initiated churn (see Figure 2-3)(Booz Allen & Hamilton, 2001):

- Handset related factors, such as equipment problem, and lost/stolen handset
- Cost related factors, such as an expensive price, or low usage for plan
- Coverage related factors, such as inadequate coverage, and poor reception quality.
- Competition related factors, such as cheaper price, and more flexible plans.
- Service related factors, such as billing problems, and poor customer service.

This analysis suggested that product (handset) and price (cost) related factors are the major causes of churn as they contribute to 59% of the causes of churn.

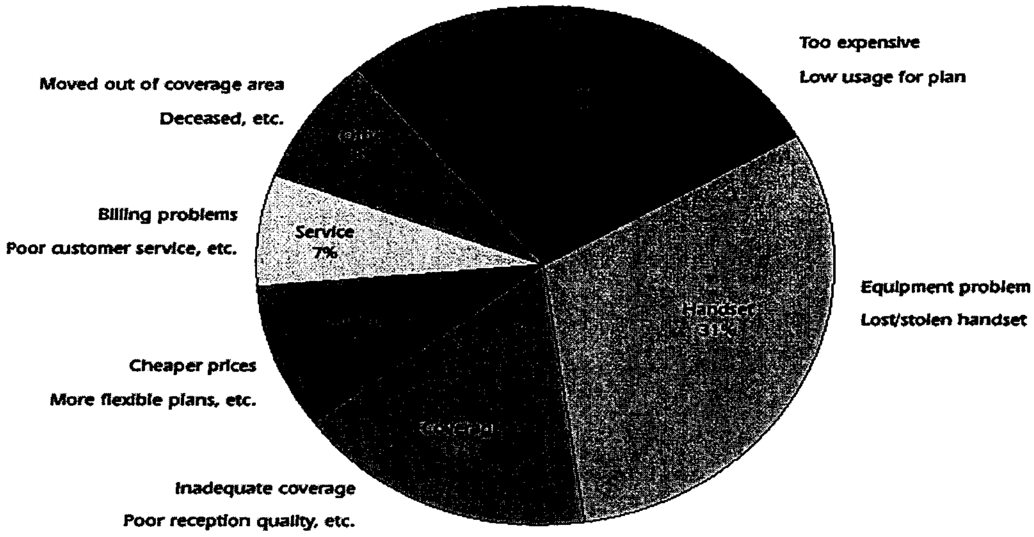


Figure 2-3 Factors That Contribute to Customer Initiated Churn

2.4 Telecom Industry Evolution

During the last five years, the wireless telecom sector has been one of the fastest-growing businesses in the economy. Serious changes to industry profitability have recently emerged (Duke Teradata, 2005). These changes include:

- Consolidation. From nearly 60 cellular companies in the 90's, only six big players are left. They account for 80% of the wireless pie; the others are now bankrupt, bought out, or struggling with heavy debts.
- Growth. The number of subscribers doubled every two years during the 90's, subscriber growth rates reduced from 50% yearly to 15%-20% by 2002.
- Competition: As an obvious result, firms engaged in a devastating price war that not only eroded revenue growth but also endangered their ability to meet their huge debts.

- Customer Strategy: The industry paradigm has arguably changed from one of making big networks and getting customers to making new services and pleasing customers.

In short, the industry has moved from an acquisition orientation to a retention orientation strategy.

A survey of wireless company chief executive officers showed that customer retention is viewed as a more important goal than customer acquisition (Craig and Jutla, 2001). Customer acquisition is defined as "acquiring the right customers, based on known or learned characteristics, which drives growth and increases margins" (Swift, 2001, p42). As we have fewer new wireless subscribers now, churn has become a major concern (Duke Teradata, 2005). In 2000, customer churn cost the U.S. industry an estimated $10 billion with annual industry churn rate of 40%; a single company with 5 million customers and annual churn rate of 30% was estimated to lose US$870 million per year (Booz Allen & Hamilton, 2001). Third quarter, 2001, statistics showed that five out of six major wireless telephone companies had more than 30% churn annually, and four out of six more than 3% churn monthly (see Figure 2-4) (Gupta *et al.*, 2003).

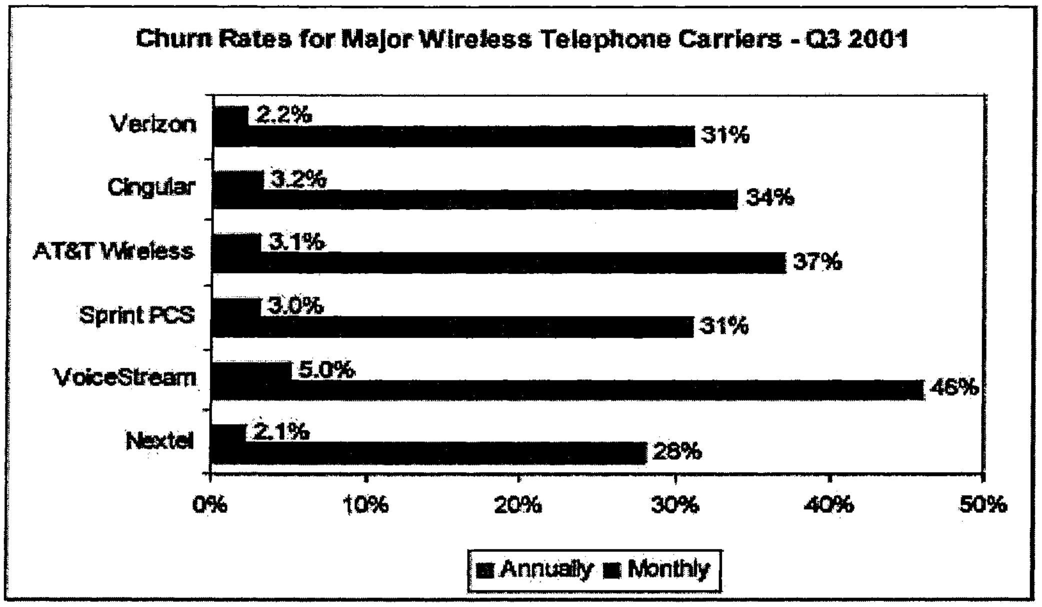


Figure 2-4 Churn in the Wireless Telephone Industry

The following are some reasons for the high level of churn in telecom industry (Duke Teradata, 2005):

- Variety of companies. There are a number of companies in the telecom industry; range from small sized local exchange carries to big sized international independent Telcos.
- Similarity of their offerings. Those companies offer similar products and services, such as similar handsets and wireless services.
- Cheap prices of handsets.
- Phone number portability. From November 2003, the wireless carriers were required to implement Wireless Local Number Portability (WLNP) by the Federal Communications Commission (Federal Communications Commission, 2004).

WLNP allows consumers to switch from one wireless carrier to another within the same local area without changing their phone numbers.

Because of these reasons, the industry's biggest marketing challenge now is to control churn rates by identifying those customers who are most likely to leave and taking appropriate steps to retain them. (Duke Teradata, 2005)

2.5 Summary

Today, organizational strategy design is based on customer relationship. CRM is where organizations manage customer relationships. Customer retention and churn control is becoming one of the most important issues in the CRM of the telecommunication industry. Both telecom industry evolution and customer relationship management require informative results of churn analysis and prediction, so that it is helpful for managers proposing market plans and transforming these plans into actions.

Chapter 3 Churn Models and Churn Predictors Review

In this chapter, we review techniques and corresponding outcomes of churn models; and churn predictors obtained from previous research. Preliminary research shows that most churn modeling techniques are drawn from data mining techniques. Hence, we first review the role of data mining in CRM and data mining methodologies and tasks. Our research focuses on the managerial propose of churn modeling, especially on their easier understandability. Hence the review of churn models here is concerned with the advantages and disadvantages of the outcomes of each model, as well as on how they influence the work in this thesis.

3.1 The Role of Data Mining (DM) in CRM

Data Mining (DM) is defined as the extraction of non-trivial, implicit, previously unknown and potentially useful patterns or knowledge from huge amount of data (Han and Kamber, 2001). DM is a multi-disciplinary field, drawing work mainly from areas including database systems, machine learning, statistics, information retrieval, and visualization (Han and Kamber, 2001; Cios *et al.*, 1998).

Knowledge Data Discovery (KDD), a synonym of DM, uses a combination of techniques including statistical analysis, neural and fuzzy logic, multidimensional analysis, data visualization, and intelligent agents. KDD can discover highly useful and informative patterns with the data that can be used to develop predictive models of behavior or consequences in a wide variety of knowledge domains (Marakas, 2003, p326).

Today, online transaction processing (OLTP) systems are nearly everywhere, which help companies collect nearly everything about the customer (Berry and Linoff, 1997). It is impossible to analyze this large amount of data by traditional data analysis technologies, such as spreadsheets and traditional statistical software tools, which can only be applied to small data sets (Hand, 1999). Increasing global and domestic competition also require that the organizations use technologies to help compete more effectively and identify new knowledge of customers that will drive new competitive strategies. More and more managers are therefore using DM to help solve their critical business problems. "One of the most recent applications of DM is in the explosive growth area of CRM" (Marakas, 2003). Kurt Thearling, a director of advanced data mining at Capital One, addressed the following typical questions that DM can help CRM in the telecom industry to answer (Thearling, 1999):

- Which customers are most likely to drop their cellular phone service?
- What is the probability that a customer will purchase at least $100 worth of merchandise from a particular mail-order catalog?
- Which prospective customers are most likely to respond to a particular offer?

The role of DM in CRM is twofold (Swift, 2001, p98):

- Convert data into information and knowledge, such that the right decisions can be made.
- Provide the mechanisms to deploy knowledge in operational systems, such that the right actions occur.

In solving churn problems, models can be built to predict the customers who are likely to switch to a competitor. These models can then be deployed in call center environments to provide guidance to operators with suggestions of approaches that are likely to help retain customers. In short, data mining techniques can help CRM in the phase of Knowledge Discovery, and the corresponding data mining results will help CRM in the phase of Market Planning (Figure 2-2).

3.2 Data Mining Methodologies and Tasks

Two basic methodologies of data mining are "hypothesis testing" and "knowledge discovery" (Berry and Linoff, 1997, p63-93). Hypothesis testing is described as a top-down method that attempts to substantiate or disprove preconceived ideas. Knowledge discovery is described as a bottom-up method that starts with the data source and tries to get it to tell something previously unknown. In CRM, hypothesis testing can be applied when end users think about possible explanations for the observed customer behavior and test these hypotheses, whereas knowledge discovery is applied when data is used to suggest new hypotheses to test.

"Knowledge discovery" can be either "directed" or "undirected". Directed knowledge discovery has a target field (such as the question "who is likely to respond to our offer?"), and can be used to recognize relationships in the data; whereas undirected knowledge discovery has no such target (such as the question "how should we define our customer segments?"), and may be used to explain those relationships once they have been found.

There are two types of DM tasks: "predictive" and "descriptive". "Predictive" tasks use current data to predict future trends in data, whereas "descriptive" tasks represent the general properties of the current data (Han and Kamber, 2001, p21).

The following three tasks are included in most books that discuss the application of data mining techniques into business: classification and prediction, association analysis, and clustering (Kudyba and Hoptroff, 2001, p23-32; Berry and Linoff, 1997, p51; Groth, 1998, p183-197). From the technique point of view, the definitions of these three tasks are as the following (Han and Kamber, 2001, p21-25):

- Classification and prediction tasks help to construct models or functions that describe and distinguish classes or concepts for future prediction. It is viewed as a predictive task.
- Association analysis helps discover the rules which imply certain association relationships, such as "occur together", among a set of objects in a database. It is viewed as a descriptive task.
- Clustering helps group or cluster the objects by the principle of "maximizing the intraclass similarity" and "minimizing the interclass similarity". It is also viewed as a descriptive task.

Both classification and clustering are used to classify and predict data; the difference between them is that classification has predetermined classes, whereas clustering does not. Both association analysis and cluster analysis are used to discover previously unknown patterns. Association analysis focuses on events that "occur together", whereas cluster analysis focuses on the partition of objects.

Customer churn prediction labels customers based on the characteristics learned from the customers lost within a period of time. Therefore, most techniques found in churn modeling are classification and prediction techniques.

The following techniques can be found for classification/prediction (Han and Kamber, 2001, p279-329; Witten and Frank, 2000, p188):

- Decision tree induction
- Bayesian classification and Bayesian belief networks
- Neural Networks
- Classification based on concepts from association rule mining
- k-Nearest Neighbor Classifiers
- Case-based reasoning
- Genetic algorithms
- Rough set
- Fuzzy set
- Linear and multiple regression
- Non linear regression
- Support Vector Machines

3.3 A Review of Churn Models

Recent researches showed two types of churn models:

- Prediction models. The outcomes can be illustrated as target customers who will churn. Through precisely predicting which customer will leave a company in a given period of time, the models also provide some features of the customer likely to churn. This helps managers retain the labeled customer.
- Customer value models. The outcomes of these models are customers' life time value, which can be interpreted as customers' contribution to the company; normally the higher the outcome, the more valuable of the customer. Then managers can focus on the valuable customers instead of all customers.

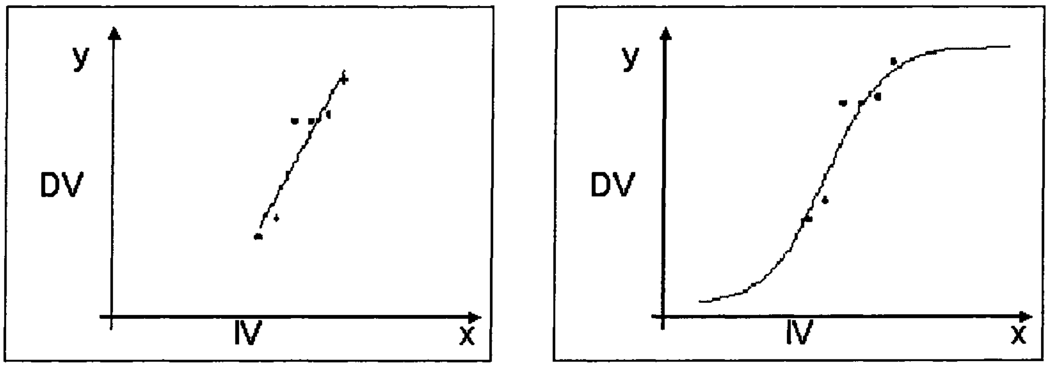
The following subsections review these two kinds of churn modeling separately.

3.3.1 Prediction Models

Most techniques for churn prediction modeling come from Data Mining techniques. Logistic regression, decision trees, and neural network are successful techniques used to build churn models (Mozer *et al.*, 2000) (Lu, 2002) (Au *et al.*, 2003) (Gupta *et al.*, 2003).

Normally, linear regression is used on continuous-valued data that have linear relationship among the variables. Logistic regression is used to predict categorical labels; it builds models that help predict the probability of some events occurring as a linear function of a set of predictor variables (Han and Kamber, 2001, p322). Figure 3-1 illustrates a simple outcome of linear regression and a simple outcome of logistic regression separately. In these two graphs, linear regression fits the relationship between the independent variable (IV) X and the dependent variable (DV) y with a straight line, whereas logistic regression fits the relationship between IV and DV with a special Sshaped curve that is mathematically constrained by the natural logarithm of the odds ratio of DV. The odds ratio of DV is calculated as $\frac{observedprobability}{1-observedprobability}$.

Howell (2002, p583) viewed Logistic Regression as "a technique for fitting a regression surface to data in which the dependent variable is a dichotomy". For a binary output variable, one can label the two values as 1 and 0. Hence, one can view an outcome regression line fit as a probability of DV. In churn analysis, the DV, churn indicator, is always labeled "1"to indicate churn, and "0" to indicate "not churn". Then the outcome may show the likelihood of the customer to churn. Moreover, the IVs used in the examination may provide the churn predictors.



(a) An example of linear regression outcome (b) An example of logistic regression outcome

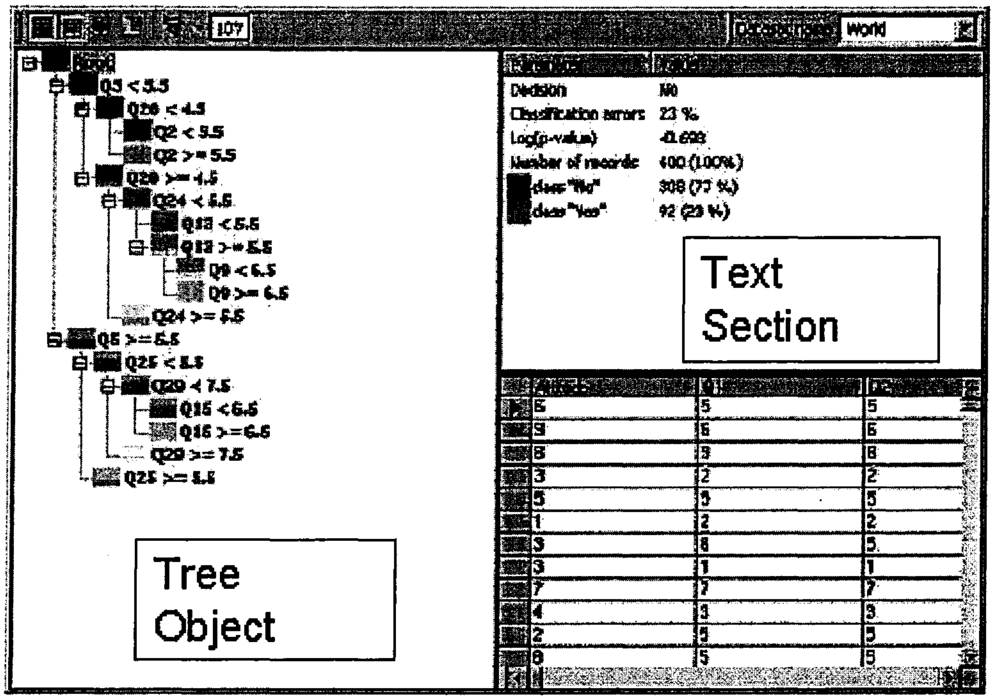
Figure 3-1 Two Simple Examples of Regression Outcome

Kantardzic (2003) said that the decision tree representation can be viewed as the most widely used logic method. He described that "a decision tree consists of nodes where attributes are tested, and the outgoing branches of a node correspond to all the possible outcomes of the test at the node." A large amount of decision tree induction algorithms can be found in machine learning and applied statistics literature. Witten and Frank (2000, p58) provide the following definition of decision tree from the algorithm point of view: “a divide-and-conquer approach to the problem of learning from a set of independent instances leads naturally to a style of representation.”

An example of a decision tree outcome report includes two parts: a tree object and a text section consisting of the measures resulting from the decision tree (Figure 3-2). Figure 3-2 shows a decision tree used to better understand the lifestyles of potential purchasers of the Discovery SUV produced by Land Rover. Marketing research manager commissioned a study of consumers' attitudes, interests, and opinions. A questionnaire was designed with 30 statements covering a variety of dimensions, including consumers' attitudes towards risk, foreign versus domestic products, product styling, spending habits, self-image, and family. The questionnaire included a final question of attitude towards purchasing the Land Rover Discovery. The respondents used a nine-point Likert scale, where a value of "1" meant that they definitely disagreed with a statement, and "9" meant that they definitely agreed. A total of 400 respondents were obtained from the mailing lists of Car and Driver, who were then interviewed at their homes by an independent surveying company.

The text section of Figure 3-2 displays the characteristics of the exploration and the confusion matrix. The total classification error is 23%, there are 92 records classified into “Yes” responses and 308 records classified into “No” responses in total for the question that tests the attitude towards purchasing the Land Rover Discovery. In the tree object part, “Q5”, “Q28”, and other labels that begins with “Q” represent the question numbers. The outcome of a decision tree can be interpreted as “if-then” rules for new labeled customers, which makes the discovered patterns easily understood. For example, a simple discovered rule may looks like

"If (the answer of question 5 is more than 5.5) And (the answer of question 25 is more than 5.5), Then the driver may want to purchase the Land Rover Discovery."



Decision	NO
Classification errors	23 %
Log(p-value)	-41.698
Number of records	400 (100%)
class "No"	308 (77 %)
class "Yes"	92 (23 %)

[illegible]	[illegible]	[illegible]
6	5	5
9	6	6
8	3	8
3	2	2
5	5	5
1	2	2
3	8	5
3	1	1
7	7	7
4	3	3
2	5	5
8	5	5

Figure 3-2 An Example of Outcome from Decision Tree

(Software : PolyAnalyst ^{TM} 4.5)

One can easily understand this simple rule and interpret it with business words. The following example is a simple rule discovered by Au *et al.* (2003):

"If District is 'Kuala Lumpur' And Payment methods is 'Cash' And Age between 36 and 44, Then Churn is True"

Domain experts may also find it meaningful since it is easy to churn for customers who pay bills by cash when compared with ones who pay by auto-pay (Au *et al.*, 2003).

Neural networks were originally used by psychologist and neurobiologist to test computational analogues of neurons. Han and Kamber (2001, p303) viewed a neural network as a set of connected input/output units where each connection has a weight associated with it. A neural network predicts the correct labels of the input variables by adjusting the weights with learning from training data set. The main disadvantage of a neural network is the difficulty in interpreting it. It is difficult for human beings to interpret the knowledge acquired by the neural networks from a set of units with weighted links (Han and Kamber, 2001, p310).

Backpropagation has proven to be a powerful neural network application. Figure 3-3 presents an illustration of Backpropagation two-step procedure. The activity from the input pattern flows forward through the network, and the error signal flows backward to adjust the weights.

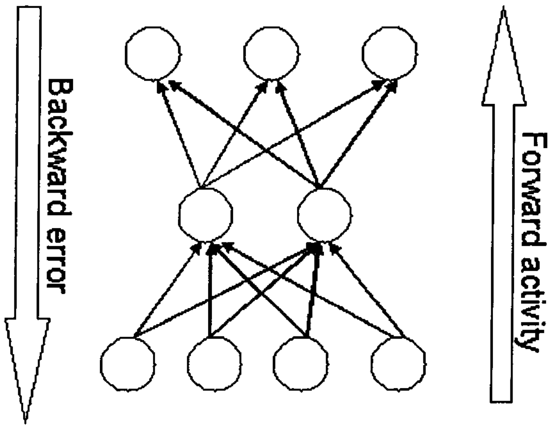


Figure 3-3 An Simple Illustration of Backpropagation Neural Networks

Table 3-1 shows the comparison of outcomes of the Logistic Regression, Decision Tree, and Neural Network models.

Table 3-1 Comparisons of the Outcomes of Logistic Regression, Decision Tree, and Neural Network Models

	Advantages of Outcome	Disadvantages of Outcome
Logistic regression	- Provide probability for prediction made (Au *et al.*, 2003) - Provide insight into relation between variables and churn (Parekkat, 2003) - A good starting point for model derivation and in understanding the model structure (Parekkat, 2003)	- Difficult to understand
Decision Trees	- Easily understand for smaller trees	- Easily confused for bigger trees
Neural networks	- Provide probability for prediction made (Au *et al.*, 2003) - Provide insight into relation between variables and the churn (Parekkat, 2003) - Effective in churn predicting (Parekkat, 2003)	- Difficult to understand

Considering the limitation of these techniques, researchers provide more effective algorithms to meet with real business world requirements. To predict the likelihood of each classification and to obtain an understandable outcome, Au *et al.* (2003) provided a new data mining algorithm, called “Data Mining by Evolutionary Learning” (DMEL) to predict the likelihood of each classification. The DMEL algorithm is based on genetic algorithms. Compared with the basic decision tree based algorithm (C4.5) which can also present understandable rules, DMEL outperformed C4.5. Compared with neural networks, they both identified more churners than C4.5. However, the advantage of DEML is that the outcome of DEML is easier for domain experts to interpret than the outcomes found by neural networks.

Considering that the customer data was non-stationary in time, Yan *et al.* (2001) provided two distinct approaches to improving the prediction results in non-stationary situations. A non-stationary situation means that the data is not static over time; i.e., the customer data changes with time. One approach was to use more historical data, for example, extending a three month window to a nine month one. The problem with using more historical data is that the training data set becomes very large. Yan *et al.* (2001) solved this problem by training one model for each shifted time window, and combined the individual model predictions with a weighted averaging formula. The training time of this approach was modest compared with that required for learning a single model over all data sets. In a situation without long data history in database, Yan *et al.* (2001) provided another approach which used new unlabeled data to improve prediction results. The purpose of using new unlabeled data was motivated by the novelty of the unlabeled data because they may be distributed quite differently than the labeled training data.

Gupta *et al.* (2003) employed a novel technique, TreeNet, which is based on one kind of decision tree -- CART (Classification and Regression Tree) (Breiman *et al.*, 1984), to build churn models for Duke/NCR Teradata 2003 Tournament. The TreeNet modeling was the winner of the tournament, and it provided much greater accuracy than the other methods tried.

In short, all of these researches focus more on the improvement of the effectiveness of the techniques and less on the understandability of the model outcomes. In this thesis, we will concentrate on the easiness of both understandability and applicability of the model outcome. We propose a novel model that can help managers easily understand customer churn, and the outcome of the model can be easily interpreted and applied in planning strategies.

3.3.2 Customer Value Models

Just predicting which customers are likely to churn is not enough for a company. With limited human resources, managers need to contact more valuable customer in advance. Calculating a customer's current value is usually based on the customer's current or recent information, such as usage, price plan, payments, collection efforts, and call center contacts (Rosset and Neumann, 2003). An example of customer value can be "the financial value of a customer to the organization", it can be calculated by "received payments minus the cost of supplying products and services to the customer" (Rosset and Neumann, 2003). Lu (2002) defined "high-value customers" as "customers with monthly average revenue of $X or more for last three months". Only customers who had received at least three months bills were considered in Lu's churn study. Rosset *et al.* (2002) employed Customer Lifetime Value (LTV) in churn analysis for telecommunication industry. LTV can be defined as "the total net income a company can expect from a customer" (Rosset *et al.*, 2002), and it is reasonable that how much is really lost due to customer churn and how much effort should be concentrate on this segment of customers.

Rosset *et al.* (2002) studied LTV modeling and its use for customer retention campaign in the telecom industry, particularly in cellular telephony. Typically, a LTV model has three components: the customer's value over time, the length of service (LOS), and a discounting factor. These three components can be either calculated/estimated separately or together. The theoretical calculations are as follows:

- The customer's value over time: $v\ (t)$ for $t>=0$, where t is time and $t=0$ is the present. In practice, the customer's future value has to be estimated from current data, using business knowledge and analytical tools.
- A length of service (LOS): it is usually described by a "survival" function $S(t)$ for $t>=0$, which describes the probability that the customer will still be active at time t. In practice, LOS has to be estimated from current and historical data as well.
- A discounting factor $D(t)$: it describes how much each $1 gained in some future time t is worth for companies right now. In practice, this function is usually given based on business knowledge. Two popular choices are:
    - Exponential decay: $D(t)= exp(-\alpha t)$ for some $\alpha>=0$ ($\alpha=0$ means no discounting)
    - Threshold function: $D(t) = I\{t<=T\}$ for some $T>0$ (where I is the indicator function).

Given these three components, the explicit formula for a customer's LTV can be written as follows:

$$LTV = \int_0^\infty S(t)v(t)D(t)dt$$

In other word, the total value is gained while customer is still active.

Customer value models are combined with prediction models when they are used to analyze churn. Researchers can either build prediction models on the selected valuable customers (Lu, 2002) or weight them more when running the prediction models (Rosset and Neumann, 2003). Both methods provide researchers the ability to concentrate on "valuable" customers, and reduce the risk of wrongly predicting premium customers.

3.4 A Review of Churn Predictors

3.4.1 Churn Predictors from Previous Researches

In the past 5 years, some researchers have studied churn predictors in the telecommunication industry. Table 3-2 to Table 3-7 summarize the predictors used by the following researches:

- Mozer *et al.* (2000) built models to predict the probability of a subscriber churning within a period of time (Table 3-2).
- Yan *et al.* (2001) built models to improve prediction of customer churn in non-stationary environments (Table 3-3).

- Lu (2002) provided some examples of reasons to identify customer initiated churn (Table 3-4).
- Au *et al.* (2003) proposed a novel data mining algorithm to predict churn accurately under different churn rates (Table 3-5).
- Rosset and Neumann (2003) calculated customer value and identify potential churn among valuable customers (Table 3-6).
- Gupta *et al.* (2003) and Scott Cardell *et al.* (2003) reported some predictors (see Table 3-7) used in 2003 churn modeling tournament (Table 3-7).

Although all of these predictors and factors contribute to the churn analysis in the telecommunication industry; the researches focused on the preciseness of churn prediction techniques, and predictors were reported as by-products of the techniques. Berry and Linoff (1997) argued that both the process of "analyzing the data" and "taking action" should be included in business data mining. Analyzing data turns data into knowledge and taking action turns knowledge into actions (Berry and Linoff, 1997). In churn prediction, without further analysis on how predictors are related to each other, how much they contribute to churn, and how they impact churn, it is challenging to tell how the predictors can be used in real management, even when we know that they are related to churn. Our research is based on these previous known predictors, and provides a deep insight into the knowledge of the relationship among the predictors.

Table 3-2 Predictors from MOZER *et al.* (2000)

Variable Name	Descriptions
Avenue through which services	Customers were how to activate

were activated	the services
Beginning dates of various services	Starting dates for services
Billing	Financial information appearing on a subscriber's bill, e.g. monthly fee, additional charges for roaming, and additional minutes beyond monthly prepaid limit.
Call quality	Service quality for a call, e.g. interference, poor coverage
Corporate capability	
Cost of roaming	Charges for roaming
Credibility / customer communications	
Credit classification	Credit rating grade
Customer classification	Customer classification, e.g., corporate vs. retail
Customer service	Quality of customer service
Dates of customer service calls	Dates of customer service calls
Handset	Handset type
Monthly charges	Monthly fee
Monthly usage	Monthly usage
Nature of customer service calls	Kinds of customer service calls
Number of abnormally terminated calls	Dropped calls which lost due to lack of coverage or available bandwidth
Number of active services of various types	
Number of calls made	Number of calls made
Number of customer service calls	Number of customer service

	calls
Pricing options	Prices for different services
Roaming/coverage	Roaming/coverage
Subscriber location	Customers' addresses
Termination dates of various services	Ending dates for services

Table 3-3 Factors or Predictors from YAN *et al.* (2001)

Variable Name	Descriptions
Customer classification	Customer classification
Customer location classification	Customer location classification
Customer credit classification	Customer credit grades
Beginning and termination dates of various services	Beginning and termination dates of various services
Monthly charges for various services	Monthly charges for various services
Monthly usage of various services	Monthly usage of various services
Monthly number of dropped calls	Monthly number of dropped calls
Monthly number of customer service calls and their classification	What kind of services customer need, and the number of that kind of customer service calls

Table 3-4 Some Examples of Reason Codes Provided by Lu (2002)

Variable Name	Descriptions
Billing problem	Financial problem
Customer expectation not met	Customer unsatisfactory
Misinformation given by sales	Misinformation given by sales
More favorable competitor's pricing plan	More favorable competitor's pricing plan

Moving and changing in business	Moving and changing in business
Unacceptable call quality	Unacceptable call quality

Table 3-5 Factors or Predictors from Au et al. (2003)

Variable Name	Descriptions
Age	Customer age
Bonus	Bonus scheme
Customer type	Customer type, e.g. government versus corporate
Dealer group	A group of dealers
District	Customer location
Gender	Female or male
Monthly charge	Monthly charge
Monthly usage	Monthly usage
Number of abnormally terminated calls	Number of abnormally terminated calls
Number of calls made	Number of calls made
Payment methods	E.g. cash vs. auto-pay
Service plan	
Subscription channel	Through which path the customer subscribe the service, e.g. dealer
Tenure	Length of service, e.g. days

Table 3-6 Factors or Predictors from Rosset and Neumann (2003)

Variable Name	Descriptions
Usage	Usage in a period of time
Price Plan	Price Plan
Payments	Payments
Collection efforts	
Call center contacts	Call center contacts

Table 3-7 Predictors from Gupta et al. (2003) and Scott Cardell et al. (2003)

Variable Name	Descriptions
Age of current handset	How long the current handset have been used
Average monthly calls (lifetime)	Average number of monthly calls in service length
Average monthly minutes (completed voice)	Average number of monthly minutes for the completed voice calls
Average monthly minutes(lifetime)	Average number of monthly minutes in the service length
Credit rating grade	Credit rating grade
CSA (condensed to 8 levels)	CART nodes CSA----Carrier Serving Area: area served by a LEC, RBOC or Telco, often using Digital Loop Carrier (DLC) technology(GOT FROM GOOGLE)
Days since last retention call	The days between today and the day making a call to retain the customer
Geographic locale or major city	Customer location
Handset price	Handset price
Length of service to date	The days between today and the starting day of service
Lifetime average minutes usage	
Recent change in monthly minutes	Recent change in monthly minutes
Number of households at address	Number of households in a particular location

Occupation	Blue/white, self
Race/origin	
Range of monthly recurring charges	Range of monthly recurring charges
Recent change in monthly minutes	Recent change in monthly minutes

3.4.2 Churn Predictors from Web Discussion

Internet increasingly becomes an important source of research. We found some extra factors that may impact churn. Table 3-8 illustrate these churn predictors from web discussion (Surfgold, 2003).

Table 3-8 Predictors from Web Discussion

Variable Name	Descriptions
Contract is up	Pay more for the same service than before
Habitual buying	Buying by the convenient instead of loyalty
Lack of decent alternative	Loyalty is forced upon people because there may not be an alternative
Move	Move
Risk minimization	Buying under someone else' advice like medicines. Someone is using that brand of product because he/she is scared of changing over to another brand.

Switching hassles	One would like to switch brands but feel the cost of switching over is way too high, and feel that the benefits are not yet big enough.
Want a new phone	Want a new phone
Wrong service(rate) plan	Wrong service(rate) plan

3.4.3 Categories of the Churn Predictors

From Table 3-2 to Table 3-8, we see that some variables have overlapped. For clear presentation, we organized these predictors based on the three apices in the Delta Model: system lock-in, customer solution, and best product. Since there is another type of predictors which describes customer characteristics and can not be classified into the above 3 categories, we provide a new category named customer profile. Customer profile classification includes the predictors of age range, income levels, education levels, occupations, residential address, tastes, demographic, psychographic, ethnic background, attitudes, and life-style etc. Table 3-9 shows the categories of the predictors.

Table 3-9 Churn Predictor Categories

Predictor Categories	**Predictors**
Customer profile	Age of current handset
	Beginning and termination dates of various services
	Billing
	Credit classification
	Customer age
	Customer classification/type
	Dates of customer service calls
	Days since last retention call
	Gender
	Handset type
	Length of service to date
	Monthly charges for various services

	Monthly usages of various services
	Number of households at address
	Occupation
	Payment methods
	Race/origin
	Recent change in monthly minutes
	Subscriber location/ location classification
	Tenure
	Want a new phone
Best product	Call quality
	Contract is up
	Handset price
	Number of abnormally terminated calls
	Price plan/ Pricing options
	Service plan
Customer solution	Bonus
	Call center contacts and their classification
	Collection efforts
	Corporate capability
	Cost of roaming
	Customer expectation not meet
	CSA----Carrier Serving Area
	Misinformation given by sales
	Number of active services of various types
	Number of customer service calls
	Roaming /coverage
	Subscription channel
System Lock-in	Dealer group
	Habitual buying
	Lack of decent alternative
	Move
	Risk minimization
	Switching hassles

3.5 Summary

Data mining techniques play an important role in CRM; they help CRM acquire knowledge of customer in the phase of Knowledge Discovery, and the corresponding data mining results will help CRM design strategies in the phase of Market Planning. Data mining prediction techniques are widely applied to predict customer churn in the telecom industry. Among those techniques, logistic regression, decision trees, and neural network are viewed as successful techniques to build churn models. However, these techniques focus more on the precision of churn prediction rather than on a deeper insight into the customers. It is opined that counting customer values provides a deeper insight of customer, but it narrows the scope in the profitability that customers could contribute to the companies. Thus this technique can not provide an overall perspective of the customers. Although these techniques have not provided a comprehensive customer view by themselves, they did provide a number of predictors that can be initially categorized into four types: customer profile, customer solution, best product, and system lock-in. The following chapters will present how we build, validate, and refine our proposed model according to these predictors.

Chapter 4 Churn-Strategy Alignment Model

Literature (Swift, 2001) indicates that more and more organizations place customer relationship in the center of their strategy planning. It also shows that successful CRM not only discovers knowledge from collected customer data, but also designs specific customer strategies based on this knowledge. It is possible to build a direct bridge between the "knowledge discovery" and "market planning". To do this, we intend to provide a knowledge discovering model for customer churn analysis. Our proposed model is based on a large customer data set collected for a telecom firm.

Chapter 3 indicates the linking of churn-related predictors to organizational strategy framework. By aligning churn-related predictors with organization strategies, one can more easily interpret churn factors from a business perspective and use this knowledge in marketing action. Our proposed model – the Churn-Strategy Alignment Model (CSAM), links single churn predictors with organizational competitiveness strategies, and groups or aggregates churn predictors into high-level business constructs. It is easier for managers to understand related high-level business constructs or factors about churn rather than individual technical low-level predictors. This chapter describes CSAM construction and the methodology employed in our research for model validation and refinement.

4.1 CSAM Construction

Chapter 3 indicates that churn predictors can be grouped under firm competitive strategies. Based on the Delta model (Hax and Wilde II, 2003) these strategies are grouped around product/service, customer solutions, and organizations themselves. For each given group of product/service, customer solutions, and organizations, high-level factors may be extracted from the individual low-level predictors. Hence grouped churn factors and activity patterns may comprise a framework relating to firm competitiveness strategy. The resultant high-level and informative churn factors are more useful to the managerial purposes of organizations, since it is easier for a manager to take action in this case as compared to using unlinked single predictors. For example, handset price (Gupta *et al.*, 2003), abnormally terminated calls (Mozer *et al.*, 2000), and roaming (Mozer *et al.*, 2000) are unlinked predictors that have been identified for churn prediction in previous research. However, in our research, handset price and abnormally terminated calls can be classified as predictors that affect product strategy while roaming impacts the customer solution strategy.

Considering that it is hard for a customer to churn from an organization that has reached a system lock-in position, we ignore the system lock-in strategy apex of the Delta model in CSAM. Because of open, non-proprietary technology standards and federal regulations in the telecommunication industry, system lock-in and competitor lock-out are also not viable strategies for many small and medium sized companies in the telecom sector.

Considering that the literature (Yan *et al.*, 2001; Au *et al.*, 2003; Rosset and Neumann, 2003) confirms that customers are different from each other and their own features may impact churn, we add a customer profile apex to CSAM. CSAM is represented in Figure 4-1. The three apices are: Customer Profile, Customer Solution, and Product.

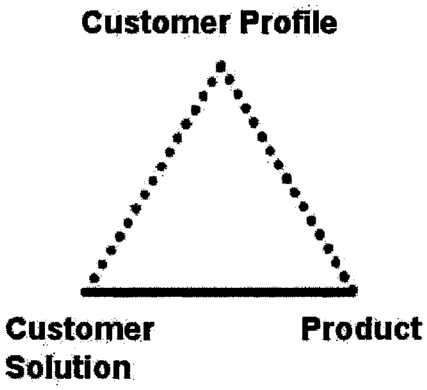


Figure 4-1 Churn-Strategy Alignment Model (CSAM)

Customer Profile is a profile of customer characteristics that are used to inform customer relationship management strategies including marketing acquisition strategies, while the Product and Customer Solution are drawn from the Delta Model (Hax and Wilde II, 2003). The dotted lines in the triangle represent the new axes of movement between customer solution strategies and customer strategies based on profile information, and between product and profile based strategies.

Customer profile includes strategies around the information about customer features such as age, demographics, education, ethnic background, geographic residence, income/financial level, life-style or taste, occupation, payment behaviors, and tenure (length of time a customer has been with a company). Customer Solution in CSAM is defined as strategies around what comprise customer's standard and extended usage of the product or service. Currently, standard usage refers to customer's receiving and making voice calls using mobile handsets. Extended usage refers to added value of other mobile services. For example, roaming and short messaging add value to the wireless service. Consequently, any added value related to the main product and service could be included in Customer Solution. Customer Solution in CSAM is similar to that of the Delta Model as Customer Solution strategies in both models imply strategies that enhance the link between the firm and customer, to the point of working with the customer to anticipate group needs, and creating new products and services to meet those needs. However the Customer Solution apex in the Delta Model may also imply customer economics focus such as those enabled by supply chain management strategies. For customer churn strategy alignment, we restrict our view to customer solutions in terms of offered products and services and to CRM strategies driven by the customer profile. Similarly, the Product apex in CSAM revolves around pricing strategies (Mozer *et al.*, 2000) and quality strategies (Au *et al.*, 2003). For example, handset price and service plan is related to pricing strategies, and blocked calls are related to quality strategy.

An extended model of CSAM with high-level churn factors and low-level churn predictors can be illustrated in Figure 4-2.

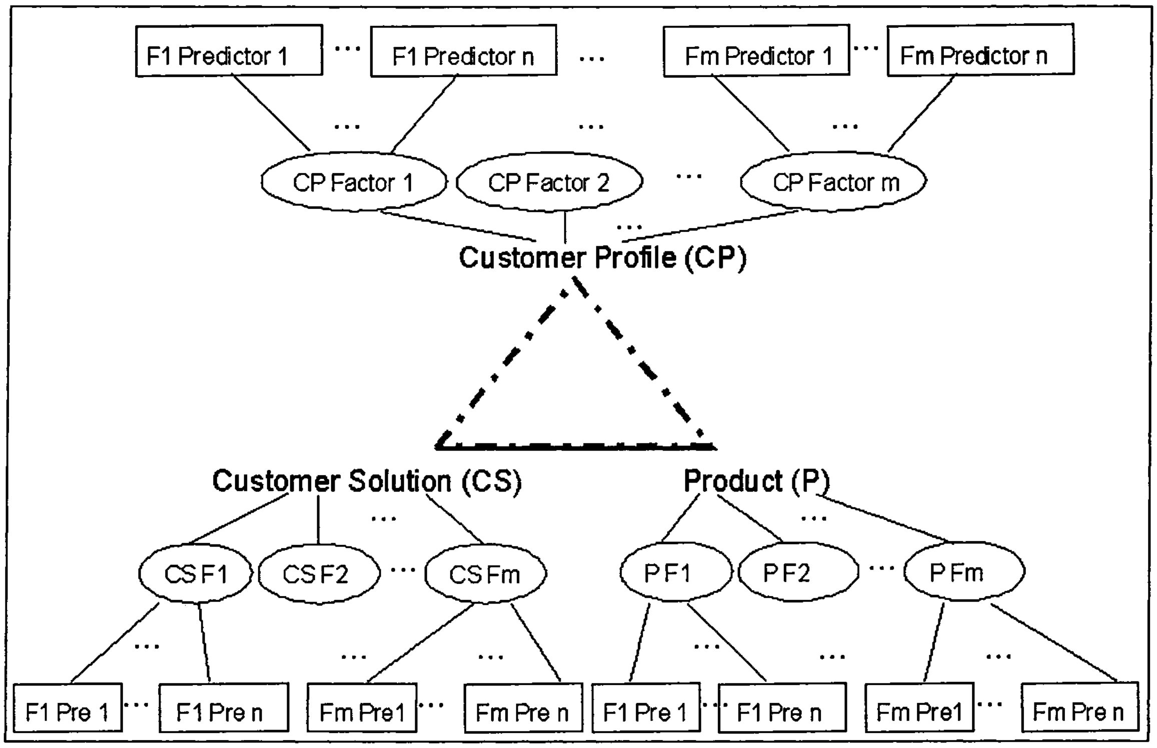


Figure 4-2 An Extended Model of CSAM with High-Level Factors and Low-Level Predictors

4.2 Methodology for Model Validation and Refinement

To validate and refine the CSAM, we apply it to a real world customer data set collected from a telecom firm. The data set was provided by Teradata Center for Customer Relationship Management at Duke University for the Churn Response Modeling Tournament in 2003. It contains 100,000 customer records obtained from a major wireless carrier in U.S.A. Only mature customers, who were with the company for at least six months, were sampled during July, September and November of 2001, and January of 2002. Each record was identified by customer ID which is from 1,000,001 to 1,100,000. A total of 172 variables are included, one for churn indication, and 171 for prediction. Our methodology and corresponding techniques focus on the process of data analysis and outcome interpretation.

For the given data set, a simple case of model validation can be illustrated as Figure 4-3, where the 3 strategy groups (Product (P), Customer Solution (CS), and Customer Profile (CP)) are viewed as 3 main churn factors respectively. Each factor is explained by a group of statistically correlated churn predictors. However, our research goal is not only to find predictors loaded on the 3 large-grain factors, but also to find high-level interpretable sub-factors under each main factor. Thus, we hypothesize that a group of interpretable high-level sub-factors can be extracted from churn predictors under each main churn factor. Therefore, there are two hypotheses to be tested:

- Statistically correlated low-level single churn predictors can be found within each group of Product, Customer Solution, and Customer Profile.
- Interpretable high-level sub-factors can be extracted from these statistically correlated low-level single churn predictors. A higher level of detail within each of the major churn factors might contribute to a more effective understanding of churn and a more nuanced explaination.

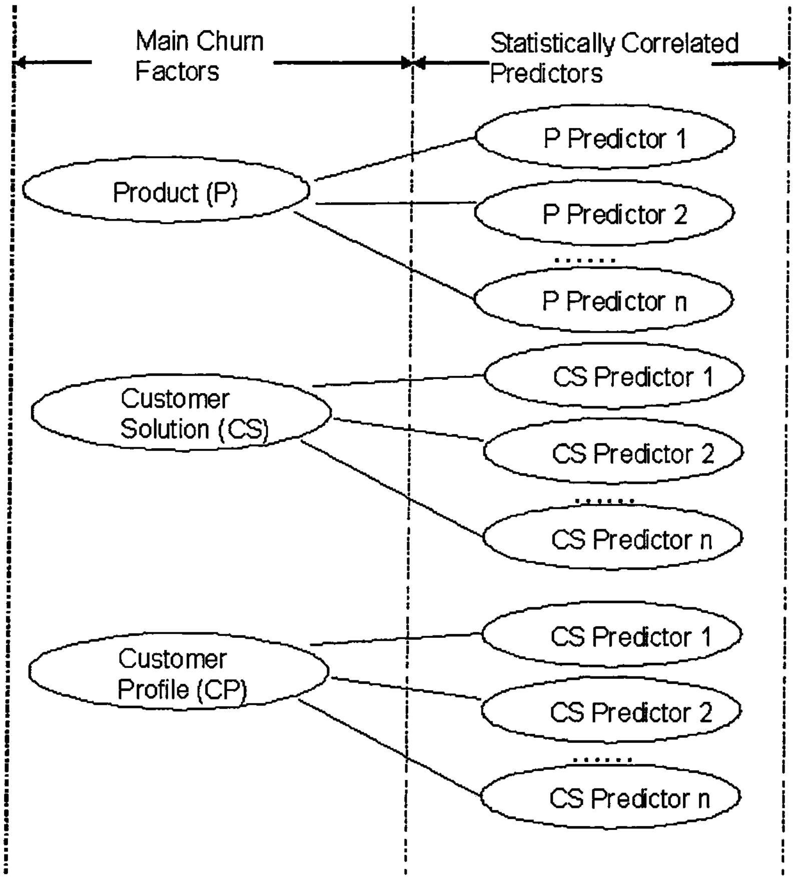


Figure 4-3 A Simple Case of Model Validation Illustration

These two hypotheses are illustrated as shown in Figure 4-4, which is the model we want to validate and refine. In this model, the 3 main churn factors are fixed; all predictors related to these main factors can be collected from the given data set. But the entire underlying sub factors are unknown and should be extracted based on the groups of particular predictors. We provide a methodology based on a hybrid of both "top-down" and "bottom-up" methods, which can be broken down into 3 steps:

1. Top-down. Based on our conceptual model of CSAM, we examine the 3 main churn factors first, and then group related churn predictors under each main factor.

The structural model can be illustrated like Figure 4-3, except that the predictors are just semantically related instead of statistically correlated.

2. Bottom-up. This step is to explore a refined model of CSAM. We examine the churn predictors first, and then fit them to intermediate groupings or aggregate sub-factors, and in turn fit these sub-factors to the 3 CSAM main factors. These sub-factors and predictors, together with our 3 main churn factors, make up a case of refinement of the CSAM model. This refined model can be used to analyze how each sub-factor and main factor impact customer churn. The models are illustrated in Figures 4-2 and 4-4.
3. Top-down-and-across. This step is to assess the obtained refined model of CSAM. It begins with the main churn factors and single predictors, and results in an assessment about how the data set fit the model. For example, how well the main churn factors are distinct; and how well each predictor contributes to the corresponding main churn factors (See Figure 4-5). Figure 4-5 is similar to Figure 4-3 except for the connections between the main factors.

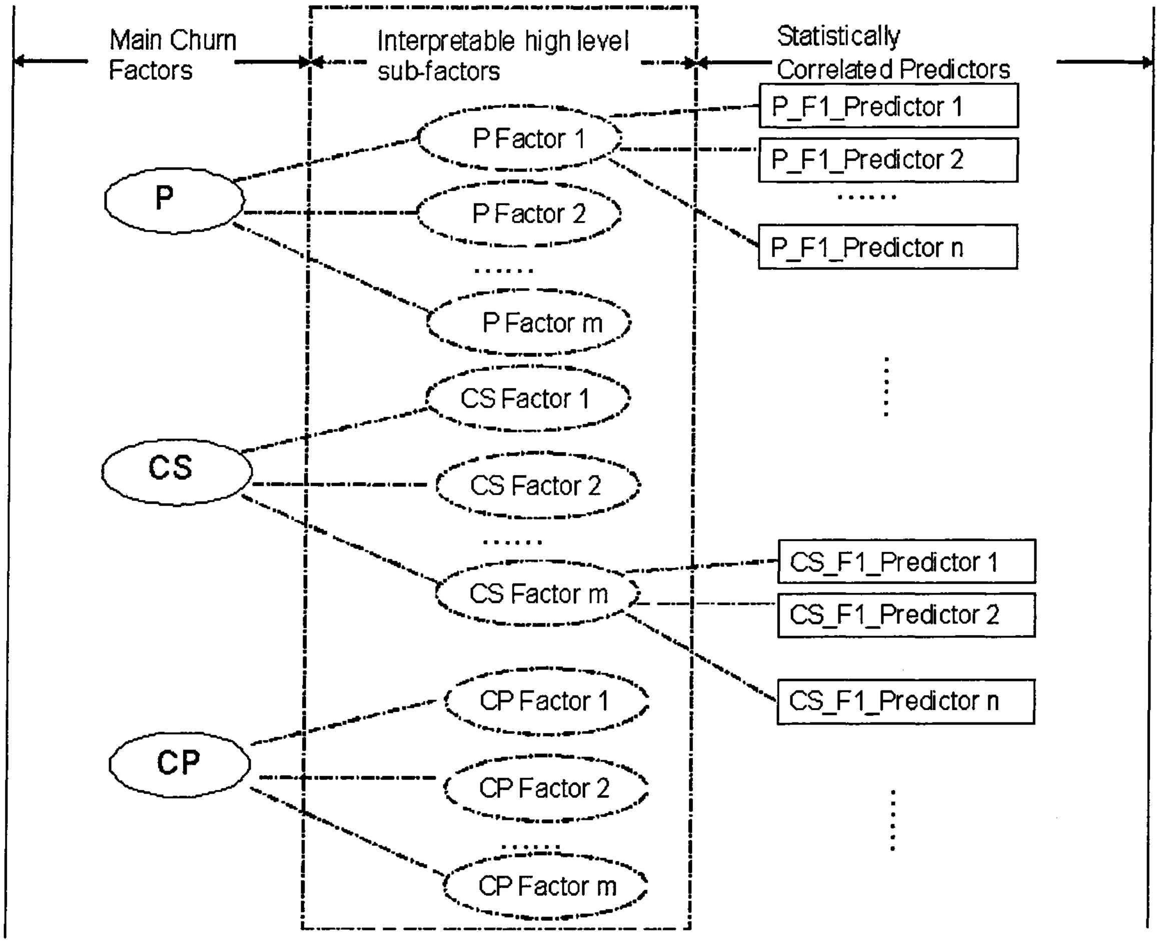


Figure 4-4 Model Validation Illustration with High-level Sub-factors

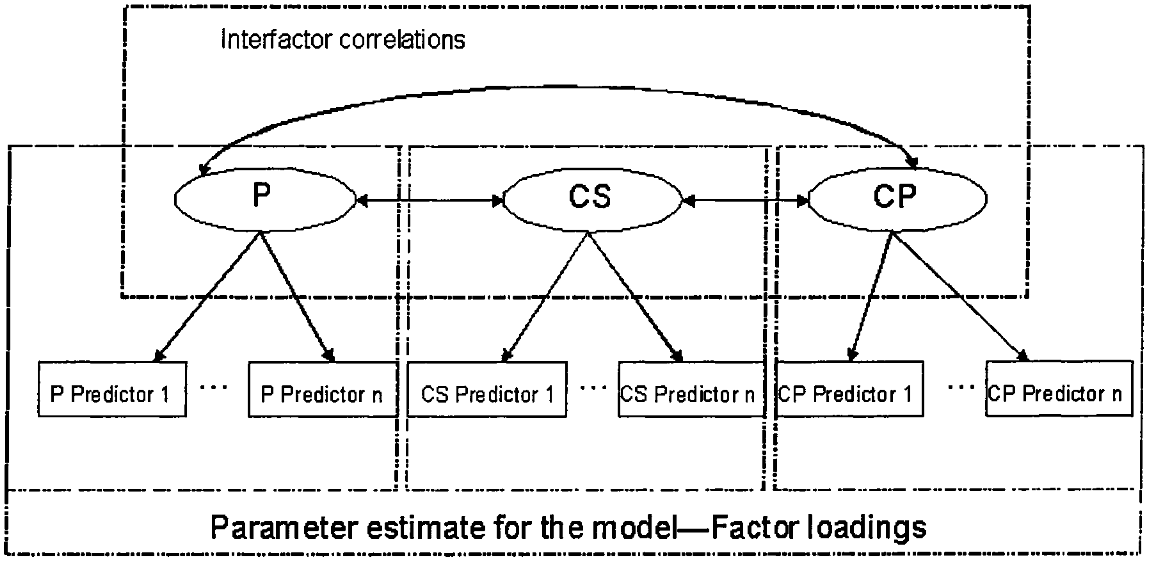


Figure 4-5 Model Assessment Illustration

Since CSAM is a novel conceptual model, there are two key measures we need to consider in the validation and refinement of this kind of model: reliability and validity. "Reliability is the extent to which an experiment, test, or any measuring procedure yields the same result on repeated trials" (Howell *et al.*, 2005). It has been defined in terms of its application to a wide range of activities. In this research, we propose to test the internal consistency reliability. Internal consistency reliability is defined as the extent to which tests or procedures assess the same characteristic, skill or quality (Howell *et al.*, 2005). Analyzing the internal consistency reliability of items dealing with a concept can disclose the extent to which items focus on the notion of a concept. In this research, internal consistency reliability tests how well the collected predictors within each group are correlated among themselves.

"Validity refers to the degree to which a study accurately reflects or assesses the specific concept that the researcher is attempting to measure" (Howell *et al.*, 2005). In this research, we intend to test construct validity of the conceptual model. "Construct validity seeks agreement between a theoretical concept and a specific measuring device or procedure" (Howell *et al.*, 2005). Construct validity can be broken down into two measures: convergent validity and discriminate validity. Convergent validity is the actual general agreement among ratings, gathered independently of one another, where measures should be theoretically related; whereas discriminate validity is the lack of a relationship among measures which theoretically should not be related (Howell *et al.*, 2005). "Internal consistency validity is generally considered a necessary but not sufficient condition for convergent validity" (Mcknight *et al.*, 2002). For this research, there is no existing theory that can tell how many predictors exists and which predictors or group of predictors can explain the factors. Thus for the validity test, we want to know how well a given data set fit the conceptual model. That is, discriminate validity is used to test how well the 3 concepts (P, CS, and CP) are empirically distinct; and convergent validity is used to test how much each observed predictor contributes to the corresponding concept, and how much each extracted factor contributes to the corresponding concept as well.

We implement our 3-step validation and refinement methodology, using the following model and data processing stages and statistical techniques in given order:

1. Model specification (top-down): select and classify churn predictors into each of the 3 strategy groups: P, CS, and CP.
2. Data preprocessing: prepare quality data for model validation and refinement.
3. Exploratory sub-factor analysis (bottom-up): get interpretable high-level sub-factors and corresponding statistically related predictors within each strategy group (see Figure 4-4).
4. Drill down analyses of the refined model: provide some application examples of the refined CSAM model to show how this refined model can help managers to understand why customers churn.
5. Structural equation modeling (Top-down-and-across): assess how distinct the 3 concepts are and how well the refined model fit the given data set (see Figure 4-5).

In the third stage, we employ techniques for Exploratory Factor Analysis (EFA) to explore the potential high-level sub factors of the given data set. EFA "seeks to uncover the underlying structure of a relatively large set of variables" (Statsoft, 2003). EFA can be conducted when there are no guiding hypotheses, and the question is simply what the underlying factors are. EFA is "the most common form of factor analysis. There is no prior theory and one uses factor loadings to intuit the factor structure of the data" (Statsoft, 2003). In this stage, without theory-based model of the sub-factors, the only way we can get the sub-factors is the data set itself. Hence, this stage is used to explore a model with sub-factors by learning from the data set.

The final stage, Structural Equation Modeling (SEM), is a general and powerful multivariate analysis technique for validating and fitting a conceptual model. It is popular in social science research. The primary reason for social science researchers adopting SEM techniques is its ability to frame and answer increasingly complex questions about the data (Kelloway, 1998). One form of SEM deals directly with how well the measures reflect their intended constructs (Kelloway, 1998). Confirmatory Factor Analysis (CFA) is an application of SEM. It tests specific hypotheses about the structure of the factor loadings and interfactor correlations. CFA is different from EFA; EFA is "guided by intuitive and ad hoc rules" (Kelloway, 1998), whereas CFA explicitly test "both overall quality the factor solution and the specific parameters composing the model" (Kelloway, 1998). In this research, we employ SEM to examine the model fit of our empirically refined model obtained from the third stage.

These five stages and techniques employed within them are briefly listed in the Table 4-1. Detailed descriptions of these five phases are provided in the following sub sections.

Table 4-1 Phases, Methodologies, and Techniques employed in model validation and refinement

Phase	Methodology	Techniques	Outcomes
1. Model Specification	Top-down	Predictor Semantic Analysis	A group of predictors for product, customer solution, and customer profile
2. Data preprocessing		Data preprocessing techniques (used either by themselves or a combination of them): - Data cleaning techniques - Data integration techniques - Data transformation techniques - Data reduction techniques	Quality data for later analyses
3. Exploratory sub-factor analysis	Bottom-up	–Internal consistency Reliability Analysis(RA) –Principal Components Analysis (PCA)	A refined model consisting of main churn factors, interpretable sub-factors, and statistically correlated churn predictors
4. Drill down analyses		Statistical methods: t-Test, One way ANOVA, Factorial ANOVA	Some applications of the refined model
5. Structural Equation Modeling	Top-down-and-across	Confirmatory Factor Analysis (CFA)	An assessment of model fit of the refined model.

4.2.1 The Model Specification

Model specification results in a rough model through a predictor semantic analysis. Here, Predictor Semantic Analysis (PSA) is defined as a method to select churn predictors from the given data set and classify them into the 3 strategy groups according to their semantic meaning and previous literature review of churn predictors (Table 3-9). To do this, we compare the meaning of each variable with predictors in Table 3-9. If a match is found between a particular variable in the data set and a predictor in Table 3-9, this variable is classified into the corresponding group based on Table 3-9. For example, we can group the variable labeled as "handset price" under product strategy since there is a predictor named handset price in the Best Product group in Table 3-9. When a variable is not found a matched predictor in Table 3-9, we classify them based on our model definition. For example, all services other than simple voice calls are considered as "added-value" service, and classified into customer solution group. Those variables describing customer features are classified into customer profile group.

After the model specification phase, we may get a model such as the one shown in Figure 4-1, where we have 3 known factors (product, customer solution, and customer profile), and all predictors related to each of these main factors are linked under the corresponding main factor.

4.2.2 Data Preprocessing

After the model specification, we need to preprocess data for model validation and refinement. Today's real world data sets are highly susceptible to noisy (wrong entries/outliers), missing (no recorded value), and inconsistent (discrepancies) data (Han and Kamber, 2001). Hence, data should be preprocessed in order to improve the quality of data, and therefore to improve the efficiency of subsequent analysis processes and to make the results more reliable. Data preprocessing techniques includes the following 4 types of techniques: data cleaning, data integration, data transformation, and data reduction (Han and Kamber, 2001). They can be used either by themselves or as a combination depending on the characteristics of the data sets. A brief description of these 4 types of techniques is as follows.

Data Cleaning Techniques

They are applied to fill in missing values, smooth out noise/identify outliers, and correct inconsistencies in a given data set.

Missing data means there is no value in a record for a given attribute. Approaches that is used for filling in missing values are as follows (Han and Kamber, 2001):

- Ignore the tuple: ignore or delete the records with missing data.
- Fill in all missing values for a given variable with a global constant
- Fill in all missing values for a given variable with the attribute mean
- Fill in all missing values for a given variable with the attribute mean for all samples belonging to the same class
- Fill in all missing values for a given variable with the most probable value of that attribute

The "ignore" approach is not effective, and can be especially poor when the percentage of missing values per attribute varies considerably. However, all "fill in" approaches bias the data. In comparison with other "fill in" methods, the last one is a popular way when "fill in" approaches are chosen. Since the data set from the telecom firm is large (thousands, tens of thousand, hundreds of thousands, or more records of customers are in the subsets of data we use), we can delete the records with significant missing data. To get more useful outcomes, we do not bias values. Therefore, in this research, missing data are treated using listwise deletion by deleting any case that had a missing value on any of the variables included in the subsequent analyses.

Noise is a random error or variance in a measured variable, and they may occur due to human or computer errors during data entry process (Han and Kamber, 2001). For example, imagine that only one customer paid more than $500 a month for his/her cell phone services while all others paid less than $200 a month. In another example, if credit classes are defined between 1 and 20; then a credit level of 40 must be an error. In the data set with these two variables, both $500 and 40 credit level are viewed as random errors. However $500 is a true value which just deviates far from others and 40 is a definite error. We distinguish these two types of errors: outliers and wrong entries. Outliers are infrequent observations in a given dataset, such as $500; whereas a wrong entry is viewed as a definite error, such as the 40 credit class level.

Wrong entries can be detected through combined computer and human inspection. Computer software can report the frequencies of all values for a given variable, and humans can look through those frequencies and compare them with the attribute documents. Wrong entries are treated as missing data.

Outliers can be detected by z score, which is a linear transformation of the original distribution of each attribute, and the equation of z is

$z = \dfrac{X-\mu}{\sigma}$, where X is any value of a given attribute, μ and σ are the mean and standard deviation of this given attribute respectively. In our research, we define that an outlier is any value whose z score is greater than 4, and an outlier is also treated as missing data.

There may be inconsistencies in some records within the data set. Inconsistencies are the discrepancies between variables' codes. For example, whenever a residential location value of "New York" and a zip code value of "21654" (Oxford) occur together in a record, there exists an inconsistency. Inconsistencies can also be detected through combined computer and human inspection. Whenever a discrepancy is found, we can either correct them or ignore this record due to whether other related attributes can be found. For example, if there is another variable, "Area code", and its corresponding value is 410, then we can make a decision that 410 is an area code related to Oxford city rather than New York city.

Data Integration Techniques

Data integration techniques integrate data from multiple sources into a coherent data set. We pay attention on the following 3 issues: schema integration, redundancy, and value conflicts (Han and Kamber, 2001).

Schema integration means successfully matching up the equivalent real-world entities from multiple data sources. For example, we can link two tables together if we know that *CustomerId* and *Cus_Id* in these two tables refer to the same real-world entity. Redundancy means that some attribute can be derived from another attribute.

Redundancy can be detected by correlation analysis, the correlation between variables A and B is measured by: $r_{A,B} = \frac{\sum(A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A\sigma_B}$, where $\overline{A}$ and $\overline{B}$ are the means of A and B respectively; σ_A and σ_B are the standard deviation of A and B respectively; and n is the number of records. If the resulting value is 0, then the two variables are independent and there is no correlation between them. If the resulting value is greater than 0, the two variable are positively correlated, and the higher $r_{A,B}$ is, the higher A is correlated with B; if the resulting value is less than 0, the two variables are negatively correlated, and the lower $r_{A,B}$ is, the higher A is correlated with B. In data analysis, the higher the absolute value of the correlation the more likely it is that variable A or B can be removed from the data set because it does not add additional information.

Value conflicts may be due to differences in representation, scaling, or encoding in different sources (Han and Kamber, 2001). For example, one source may use the

Canadian dollar whereas another source uses the US dollar as the payment unit. We should unify these semantic heterogeneities in the data integration, such as unifying the payment unit.

Data Transformation Techniques

In data analysis, the data are required to be transformed or consolidated into forms suitable for mining (Han and Kamber, 2001). Data transformation includes summary/aggregation, generalization, normalization, and attribute construction (Han and Kamber, 2001). Summary or aggregation operations can be applied to daily data to compute monthly or annual data. In generalization, raw data or low level data are replaced by higher level concepts. For example, *age* values can be mapped to *young, middle-aged, and senior*. Some data can also be scaled or normalized so as to fall in within a small specified range, such as 0.0 to 1.0. One of the popular normalization methods is min-max normalization, which maps a value *v* of a given attribute to *v'* in the range [*new_min, new_max*] by computing:

$v' = \frac{v - \min}{\max - \min}(mew\_\max - new\_\min) + new\_\min$. Here, m*in* and *max* is the original minimum and maximum values of the given attribute. This method preserves the relationship among the original values. New attributes can also be constructed and added based on the given set to help the mining process. For example, we can add a variable representing the *profit* if we have both *cost* and *revenue* variables, and construct profit by computing: *profit=revenue-cost.*

Data Reduction Techniques

Data reduction techniques are applied to find a smaller dataset which still closely maintain the integrity of the original data. This can result in a more efficient mining process, yet the same or almost the same analytical results. Data reduction can be done through either removing some attributes or reducing the number of records. Dimension reduction and numerosity reduction are two strategies for removing attributes and reducing data size respectively (Han and Kamber, 2001).

Dimension reduction reduces data size by removing irrelevant or redundant attributes. Attribute selection can be done either by domain experts or computer selection programs. In our research, we intend to select attributes by using the results of previous literature review, specifically their research results, since our selection focuses on concept matching rather than just normal attribute selection. We need to find attributes which can be viewed as churn predictors under the 3 concepts: product, customer solution, and customer profile.

Sampling is one of the most commonly used methods to reduce the size of the data set. "Random sampling is a sampling technique where we select a group of subjects (a sample) for study from a larger group (a population). Each individual is chosen entirely by chance and each member of the population has a known, but possibly non-equal, chance of being included in the sample" (Easton and McColl, 2005). The advantage of random sampling is that the likelihood of bias is reduced. In our research, we also use random sampling to reduce our sample sizes once we have a large data set. We randomly split the given data set into 2 parts, one is used for analysis, and the other is used for control. We created a program to reach this goal. The program scans each customer record only once; and each customer record has an equal chance to be randomly assigned to either the analysis part or the control part.

4.2.3 Exploratory Sub-factor Analysis

There are two methods involved in this phase: Internal Consistency Reliability Analysis (RA) and Principal Components Analysis (PCA).

Internal Consistency Reliability Analysis

Internal consistency reliability analysis is used to construct a model with causally linked variables (Mcknight *et al.*, 2002). Reliability Analysis (RA) is a statistical method to study the properties of measurement scales and the items that make them up (SPSS, 2001). One of the statistics that RA calculates is a commonly used measure-- Cronbach's coefficient *alpha* (α). It is used to estimate the reliability of a sum scale of a group of items, and provides information about the relationships between individual items in the scale. A common rule is that the predictors should have a Cronbach's alpha of .7 to judge the set reliable. In our research, within each main factor group, we viewed the collected predictors reliable with each other if their Cronbach's alpha is more than .7. We interpret this result to be the set of predictors that reflect the same concept -- the corresponding main factor.

Principal Components Analysis

In this phase, since we do not have the theory required to create a sub-factor structure model, we apply Principal Components Analysis to the CSAM model to explore an extended model with sub-factor structure. Principal Components Analysis (PCA) "uses a mathematical procedure that transforms a set of correlated response variables into a new set of uncorrelated variables that are called Principal components" (Johnson, 1998, p3). In our research, we call the extracted principal components as sub-factors. The primary objectives of a PCA are to discover the true dimensionality of the data set and to identify new meaningful underlying variables (Johnson, 1998, p96). This is also the reason we choose PCA to get the sub-factors in our research. In our research, unlinked single predictors are hypothesized to be explained by a smaller group of discriminate high-level sub-factors. Moreover, the obtained sub-factors are to be interpretable. However, not all new underlying variables are meaningful (Johnson, 1998). Furthermore, if there is non-meaningful sub-factors, then the question now is how to determine the number of sub-factors in the refined model? And, how to make the meaningful sub-factors more interpretable?

Rotation serves to make a given PCA solution more interpretable (Kline, 1998). There is more than one type of rotation option, but the choice of rotation methods depends on whether the rotated sub-factors are uncorrelated (orthogonal methods) or correlated (oblique). We employ Varimax Rotation, an orthogonal method, in our research, since it yields results which make it as easy as possible to identify each variable with a single factor (Garson, 2005).

We determine the number of sub-factors by a combination of examination of the eigenvalue and factor loading. The variances extracted by the sub-factors are called the eigenvalues (Statsoft, 2003). Sub-factors with eigenvalues greater than 1 and at least one rotated factor loadings more than .6 are retained. Whenever we find that a sub-factor contains more than one meaning, we require a re-extraction with more sub-factors. Therefore, we keep interpreting the extracted sub-factors and re-extract them till we obtain a pattern of loadings on each factor that is as diverse as possible, and each factor is more easily interpreted.

In this research, RA and PCA are conducted with SPSS for Windows version 11.5. At the end of this phase, we obtain a refined model with the 3 main factors, corresponding groups of high-level sub-factors, and corresponding groups of predictors, as the ones shown in Figure 4-2 and Figure 4-4.

4.2.4 Drill-down Analyses of The Refined CSAM Model

With the refined model obtained from last stage, we intend to provide some application examples of the resulting model to show how this refined model could help managers to analyze customer churn.

We propose to select some sub-factors from the refined model, and collect corresponding predictors from the data set and the churn indicator as well. We may also categorize some variables with reasonable rules. For example, we may categorize customer age as lower-level aged (<=18 years) middle-level aged (19-49), and senior-level aged (>=50) customers. We conduct statistical methods; t-Test, One way ANOVA, and Factorial ANOVA (SPSS, 2001; Howell, 2001) to test how this sub- factors impact churn.

4.2.5 Structural Equation Modeling (SEM)

In the SEM phase, we intend to conduct Confirmatory Factor Analysis (CFA) to assess the refined model with the given data set. CFA allows researchers to "specify and test measurement models that are more a priori" (Kline, 1998). The basic idea behind CFA is to determine whether variables are interrelated through a set of linear relationships. This is done by examining the variances and covariances of the variables. (Statsoft, 2003) CFA can be conducted when "we already know what the measures mean, and we want to test propositions". In this phase, we have an experimentally obtained model within which each observed variable is specified to load on a single sub-factor, and in turn a single main churn factor. What we are interested in here are as the follows;

- How well the 3 main factors distinct? Lower correlations means the 3 main factors not overlap in most of dimensions, and it indicates discriminant validity.
- How well each observed variable load on its specified main factors? Here, the 3 main factors are potential churn factors created by us, and the observed variables are collected from the given data set based on our CSAM conceptual model. It is meaningful to test the loading of each observed variable on the specified main factor; since a high loading represents that the observed variable contributes to the specified main factor. A result of high loadings of all observed variables on their specified main factors indicates convergent validity.

We propose to conduct CFA with LISREL 8.51. LISREL is a popular statistical package performing SEM, and it is the package of reference in most articles about structural equation modeling (Kelloway, 1998). In developing and conducting the CFA, we propose to follow Bollen and Long's (1993) description of the 5 stages of the application of SEM:

1. Model specification.
2. Model identification.
3. Estimation.
4. Testing fit.
5. Model re-specification.

Model Specification

Model specification helps clarify exactly what significant relationships exist in the proposed model. Most frequently, the structural relations that form the model are represented by a path diagram (Kelloway, 1998). In the path diagram, variables are linked either by unidirectional arrows to present causal relations, or by bidirectional arrows to present non-causal correlational relationship. Figure 4-6 is the path diagram used in our research. In this path diagram, unobserved variables like P, CS, and CP are called latent variables; they can not be measured directly, and are hypothesized to underlie the observed variables. Latent variables are usually represented by a variable name enclosed in an oval or circle. Observed variables like the collected predictors are also called indicators, and are usually represented by a variable name enclosed in a rectangle.

With the model specification, SEM applications estimate interfactor correlations, factor loadings and unique errors based on observed covariance /correlation matrix.

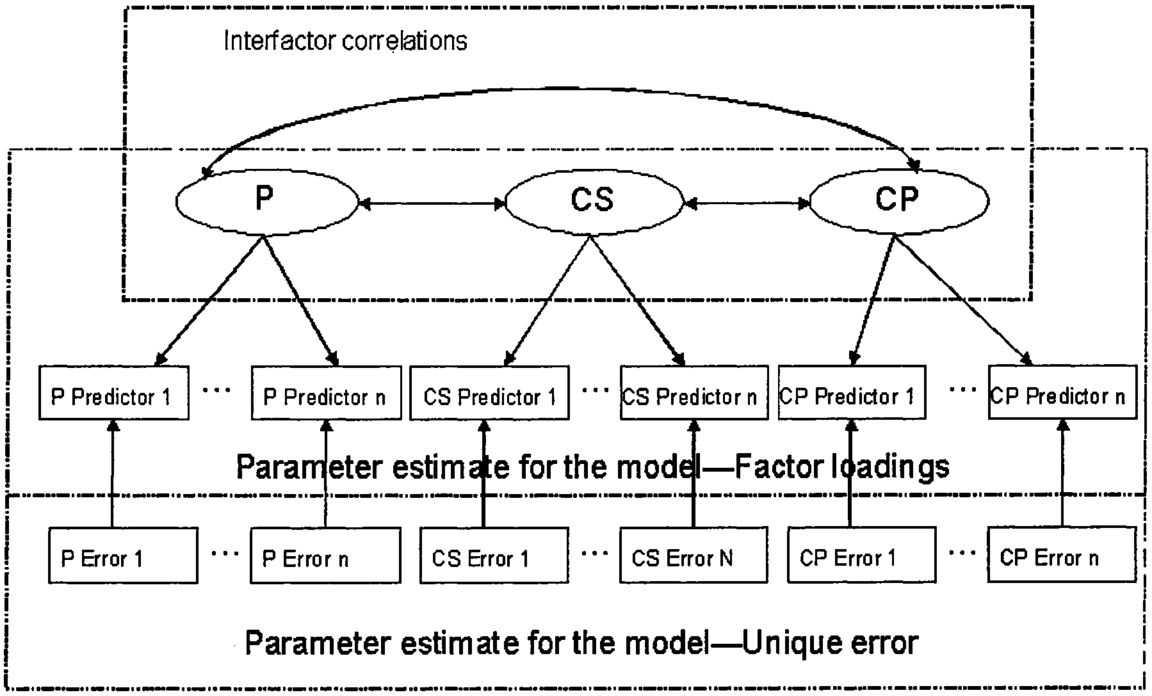


Figure 4-6 CFA Model Specification

Model Identification

Model identification is used to identify the causal relationships among latent variables and indicators. For CFA, model identification is typically dealt with defaults. That is, the latent variables are hypothesized to "cause" the observed variables. Thus the casual flow is expected to be from latent variables to the observed variables. In reality, although path diagrams can be used to represent causal flow in a system of variables, they need not imply such a causal flow. For a simple relationship illustrated in Figure 4-7, we can interpret this relationship as “X causes Y” or “a visual representation of the linear regression relationship between X and Y” (Statsoft, 2003).

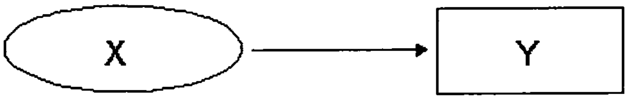


Figure 4-7 A Diagram of Relationship Between X and Y

Estimation and Testing Fit

Here, we describe estimation and testing fit together, because they relate to both the choice of estimators, and their cut values for model assessment. We use estimators (see Table 4-2) with typical cut values (Mcknight *et al.* 2002) to assess the CFA model. Moreover, discriminate validity among the latent variables is assessed to be without question if the intercorrelations are less than 0.6; convergent validity is to be assessed using the following 3 criteria (Mcknight *et al.*, 2002):

- Individual item lambda coefficient greater than 0.7
- A significant (0.05 level) t-statistic for each path
- Each path loading greater than twice its standard error

Table 4-2 Estimators and Their Cut Values in CFA

Name	Cut Value	Description
GFI	>0.9	Goodness of fit index
NFI	>0.9	Normed fit index
AGFI	>0.8	Adjusted Goodness of fit index
CFI	>0.9	Comparative fit index
RMSEA	<0.08	Root mean square of approximation

Model Re-specification

Model re-specification is model modification in case the original model does not fit the data. Models can be modified to improve the fit. (Kelloway, 1998) Theory trimming is a common approach to model improvement. It improves the model by deleting non-significant paths from the model. However, in our research, we also want to consider the indicators as groups explained by sub-factors. We intend to improve the model through deleting indicators with lower loading sub-factors. That is, if necessary, we remove indicators from the model based on how well the corresponding sub factors contribute to the latent variable. The remaining indicators relate to a higher loading sub-factor (path coefficient of sub factor >0.7).

4.3 Summary

This chapter presents our proposed model – Churn-Strategy Alignment Model (CSAM). CSAM links single churn predictors with organizational competitive strategies, which are around product, customer solution, and customer profile. CSAM provides a new vision on aligning churn predictors to strategy in the telecommunication industry, and it may assist managers in creating competitiveness strategies to reduce churn.

This chapter also provides proposed methodologies and techniques for model validation and refinement. We apply a combination of top-down and bottom-up methodologies, and propose to implement these methodologies in 5 stages:

1. Model specification (top-down): select and classify churn predictors into each of the 3 strategy groups: product, customer solution, and customer profile.

2. Data preprocessing: prepare quality data for model validation and refinement.
3. Exploratory sub-factor analysis (bottom-up): get interpretable high-level sub-factors and corresponding statistically related predictors within each strategy group.
4. Drill down analyses of the refined model: provide some application examples of the refined CSAM model to show how this refined model help managers to understand customers' churn.
5. Structure equation modeling (Top-down-and-across): assess how well the 3 concepts distinct and how well the refined model fit the given data set.

Within each phase, related techniques are reviewed, and the proposed techniques are provided.

Chapter 5 Results

In this thesis, we apply CSAM to a real world telecommunication industry data set. This chapter presents the experimental results and follows the 5 phases described in Section 4.2 of Chapter 4. This chapter provides a description of the data set; churn predictors collected in the data set; data preprocessing; the results and corresponding interpretation of Reliability Analysis (RA) and Principal Components Analysis (PCA); a refined CSAM model; drill-down analyses and examples of the application of the refined model; and the results and corresponding interpretation of Confirmatory Factor Analysis (CFA).

5.1 Introduction of The Data Set

The data set is the Churn Response Modeling Tournament Dataset sourced from Teradata Center for Customer Relationship Management at Duke University (Duke Teradata, 2003). This data set is organized into 5 data files: Calibration Data, Current Score Data, Future Score Data, Current Answers, and Future Answers. Calibration Data contains both churn indicator and potential predictors; the Current and Future Score Data contain the predictors but not churn indicators. In the tournament, participants estimated models on Calibration Data, and use these models to predict for the Current and Future Score Data.

In the model validation stage, we focus on model estimation but not on prediction. Hence, only the Calibration Data was used in this research to estimate our model.

The data set (Calibration Data) contains 100,000 customer records obtained from a major wireless carrier in U.S.A. Only mature customers, who were with the company for at least six months, were sampled during July, September and November of 2001, and January of 2002. Each record was identified by customer id which is from 1,000,001 to 1,100,000. A total of 172 variables are included, one for churn indication, and 171 for prediction. For each customer, potential predictors were calculated based on the previous four months. Churn indicator was assigned "1" (churn) or "0" (did not churn) based on whether the customer left the company during the period 31-60 days after the customer was originally sampled. The one-month treatment lag between sampling and observed churn was for the practical concern that in any application, a few weeks would be needed to score the customer and implement any proactive actions. Churners were over sampled when creating the sample data set to create a 50-50 split (the exact number is 49,562 churners and 50,438 non churners) between churners and non churners.

There are two types of variables: continuous variables and category variables. Most continuous variables are statistical values, and some were shown in the form of means and ranges, for example, "mean number of blocked (failed) data calls", and "range of number of blocked (failed) data calls". We view some continuous variables as different facets of the same predictors. For example, all of the three variables "average monthly minutes of use over the previous three months", "average monthly number of calls over the previous three months", and "average monthly revenue over the previous three months" describe the average usage over the previous three months from different perspectives. In this thesis, the ranges variables were ignored, and only one variable was considered for each predictor.

In Appendix B, a list of the 171 variables is provided.

5.2 Model Specification

Based on the work done in Chapter 4, variables were classified into three groups: product, customer solution, and customer profile. In the product group, variables related to price and quality was considered as predictors. In the solution group, we were interested in how much an added value for product/service was used by customers, for example, how many roaming calls were used by a customer. Thus variables related to the "number of calls" were considered. In the profile group, age, credit card level, ethnicity, payment patterns, and other customer characteristics were considered. In total, 6 product predictors, 16 solution predictors, and 31 profile predictors were extracted from the data set. Table 5-1 shows the predictors involved in each group.

Table 5-1 Predictors involved in Each Group

Product	Customer Solution	Customer Profile
Handset Price	Call Forwarding Calls	Account Spending Limit
Blocked or Dropped Calls	Call Waiting Calls	Active Subscribers in Household
Blocked Data Calls	Customer Care Calls	Average monthly revenue over the previous three months
Blocked Voice Calls	Completed Data calls	Average monthly revenue over the previous six months
Dropped Data Calls	Completed Voice Calls	Average monthly revenue over the life of the customer
Dropped Voice Calls	Directory Assisted Calls	Billing adjusted total revenue over the life of the customer
	DualBand	Credit Class
	Foreign Travel Dummy	Ethnicity
	Handset Web Capability	Geographic Area
	Inbound Calls Less than one Minute	Handset Refurnish Indicator

	Inbound Wireless to Wireless Calls	Monthly Revenue (Charge Amount)
	Outbound Wireless to Wireless Calls	Monthly Recurring Charge
	Received SMS Calls	Motorcycle Indicator
	Received Voice Calls	Number of Days of Current Equipment
	Roaming Calls	Number of Days Since Last Retention Call
	Three-way Calls	Number of Handsets Issued
		Number of Models Issued
		Percentage Change in Monthly Revenue vs Previous Three Month Average
		Possession Number of Credit Cards
		Revenue of Data Overage
		Revenue of Voice Overage
		Referral Numbers
		RV indicator
		Social Group
		Total calls into retention team
		Total Number of Months in Service
		Total offers accepted from retention team
		Total Overage Revenue
		Total Revenue
		Truck Indicator
		Unique Subscribers in Household

5.3 Data Preprocessing

The data set was randomly split into 2 almost equally sized sets. Experiments were conducted on one set; the other was used for control purpose. The data set used in this research contains 50,089 records. No inconsistency and wrong entry was detected. Missing data were treated using listwise deletion by deleting any case that had a missing value on any of the variables included in each test. All experiments involved in this research were conducted with SPSS for Windows version 11.0.

Outliers of continuous variables were detected by z scores. All values whose z scores are greater than 4 are treated as missing values.

All categorical variables were recoded by numbers and redefined as continuous variables. For example, handset price is defined as a category variable in the data set. This definition is based on the Teradata Center's data classification for handset price. It has 17 values: $9.99, $29.99, $39.99, $59.99, $79.99, $99.99, $119.99, $129.99, $149.99, $159.99, $179.99, $199.99, $239.99, $249.99, $299.99, $399.99, and $499.99. For this kind of variables, we recode them from 1 to the maximum number of values and redefined it as continuous variable. The recoding statement for handset price is as follows:

```
Missing Value = Missing value
Lowest  thru 10 = 1
10 thru 30       = 2
30 thru 40       =3
40 thru 60       =4
60 thru 80       =5
80 thru 100      =6
100 thru 120     =7
120 thru 130     =8
130 thru 150=9
150 thru 160=10
160 thru 180=11
180 thru 200=12
200 thru 240=13
240 thru 250=14
250 thru 300=15
300 thru 400=16
400 thru Highest =17
```

Dualband is another category variable. It has 4 values:

N = No
T = Tri-mode (analog, digital, 3G)
U = Unknown
Y = Yes

For this kind of variable, we rank the values first from the worst to the best. For example “No” dualband is the worst one, the “Yes” of dualband is better than “No” dualband, and Tri-mode is the best. “Unknown” is treated as missing values. The recoding statement for Dualband is as follows:

```
'U'=SYSMIS
'T'=1
'Y'=0.5
'N'=0
```

The recoding of other category variables can be found within Appendix A – data preprocessing syntax.

In the customer profile group, the range of values of variables is different from each other. The lowest range is from 0 to 1, whereas the largest range is from 3.75 to 16715.21. For this kind of data, we use min-max normalization to scale data so as to fall within a specified range, 0.0 to 1.0.

There are two variables whose missing values signifies a meaning: “total offers accepted from retention team”, and “total calls into retention team”. For these two variables, missing value signifies an absence of an offer from the retention team, and an absence of calls into the retention team respectively. Value “0” of “total offers accepted from retention team” signifies non-acceptance of an offer from the retention team. For these kind of variables, we split each of them into 2 variables: each was assigned “0” —

missing value, or "1"—non missing value; the other variable copy all values from the original data, but missing value means nothing. Hence, "total offers accepted from retention team" was split into 2 variables: Absence of Offer (0—absence of an offer from the retention team; 1—no absence of an offer from the retention team), and Offers Accepted from Retention Team. "Total calls into retention team" was split into 2 variables: Absence of Call to Retention Team (0-- absence of calls into the retention team, 1—no absence of calls into the retention team), and Calls to Retention Team. Thus the number of potential predictors of customer profile became 33.

5.4 Results of Exploratory Sub-factor Analysis

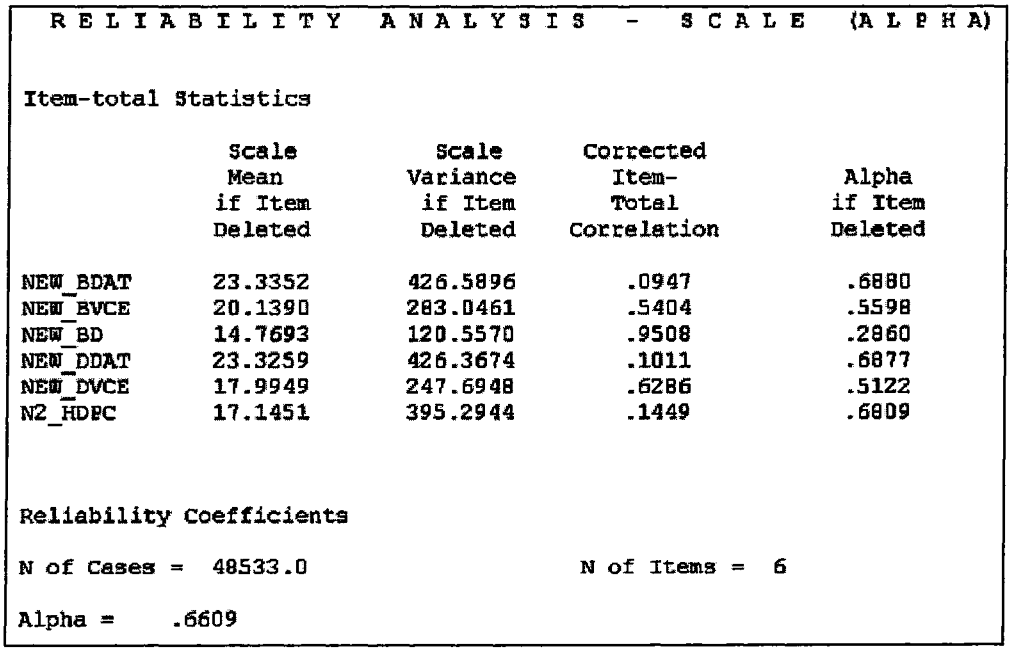
Reliability Analysis (RA) was conducted to examine the reliability of the predictors in each group. Principal Component Analysis (PCA) was conducted to extract the high-level sub-factors of each group. Both RA and PCA are conducted with SPSS for Windows 11.5. The results of each group of the product, customer solution, and customer profile will be presented in the following 3 subsections.

5.4.1 Product Group

The RA experimental result of the original 6 preprocessed predictors (see Table 5-2) is shown in Figure 5-1. This result shows that the 6 predictors were not highly correlated with each other (α=0.6609). In this situation, we need to keep removing an item or more until the remaining predictors are highly correlated with each other ($\alpha >= .7$). We used SPSS to provide the potential α value if one predictor was deleted. This value is shown on the right side of the result. From Figure 5-1, it is seen that a .6880 Alpha value can be obtained if NEW_BDAT (Blocked Data Calls) is deleted. After item NEW_BDAT was deleted (Figure 5-2), we got a new Alpha value of .6880, which is still lower than the criteria value (0.7). Hence we continue to remove an item from the group of predictors. After item NEW_DDAT was deleted, the result (α=0.7330, Figure 5-3) showed that the current group of predictors are highly correlated with each other under a product strategy grouping.

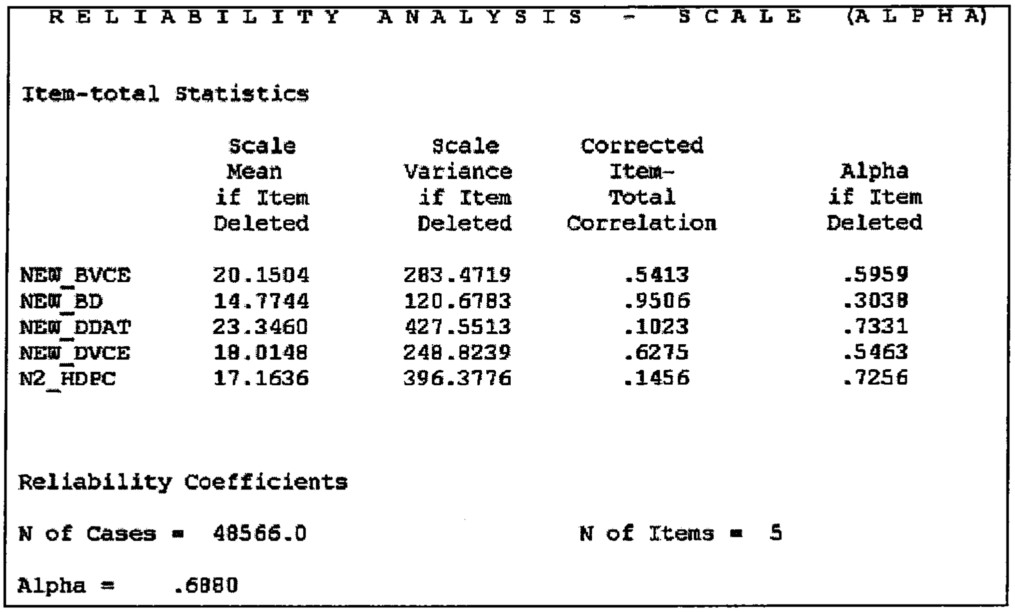
Table 5-2 Variable Names for Predictors of Product Group

Predictors	Variable name in RA & PCA
Handset Price	N2_HDPC
Blocked or Dropped Calls	NEW_BD
Blocked Data Calls	NEW_BDAT
Blocked Voice Calls	NEW_BVCE
Dropped Data Calls	NEW_DDAT
Dropped Voice Calls	NEW_DVCE



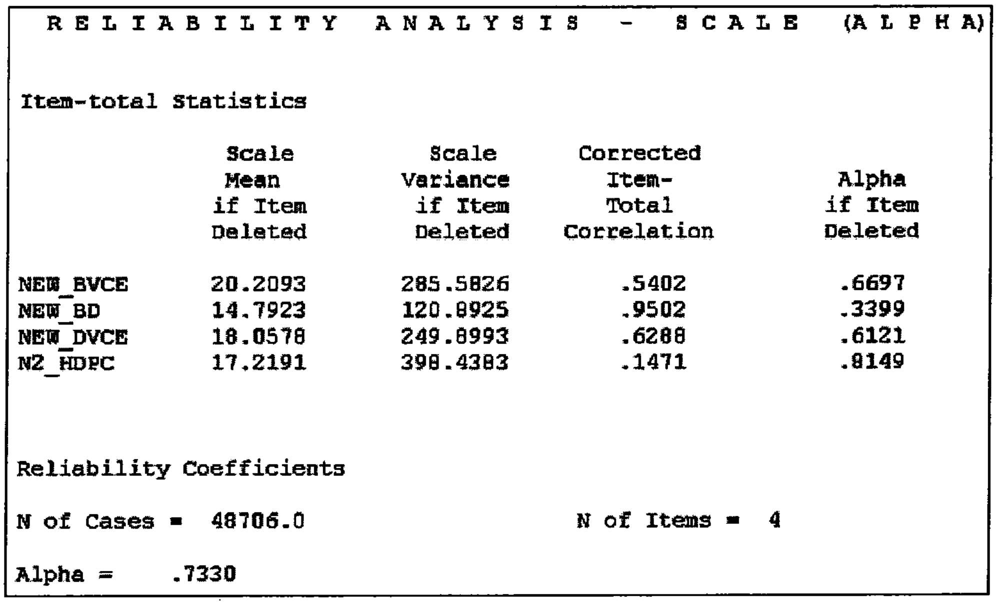
```
 R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (A L P H A)


Item-total Statistics

               Scale          Scale      Corrected
               Mean         Variance       Item-            Alpha
              if Item       if Item        Total           if Item
              Deleted       Deleted     Correlation        Deleted

NEW_BDAT      23.3352       426.5896       .0947            .6880
NEW_BVCE      20.1390       283.0461       .5404            .5598
NEW_BD        14.7693       120.5570       .9508            .2860
NEW_DDAT      23.3259       426.3674       .1011            .6877
NEW_DVCE      17.9949       247.6948       .6286            .5122
N2_HDPC       17.1451       395.2944       .1449            .6809


Reliability Coefficients

N of Cases =   48533.0                     N of Items =  6

Alpha =     .6609
```

Figure 5-1 RA Result of Original Product Group



```
  R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (A L P H A)


 Item-total Statistics

                 Scale          Scale      Corrected
                  Mean        Variance       Item-           Alpha
                 if Item       if Item       Total          if Item
                 Deleted       Deleted    Correlation       Deleted

NEW_BVCE         20.1504      283.4719        .5413           .5959
NEW_BD           14.7744      120.6783        .9506           .3038
NEW_DDAT         23.3460      427.5513        .1023           .7331
NEW_DVCE         18.0148      248.8239        .6275           .5463
N2_HDPC          17.1636      396.3776        .1456           .7256



Reliability Coefficients

N of Cases =   48566.0                     N of Items =  5

Alpha =      .6880
```

Figure 5-2 5 Predictors RA Result of Product Group



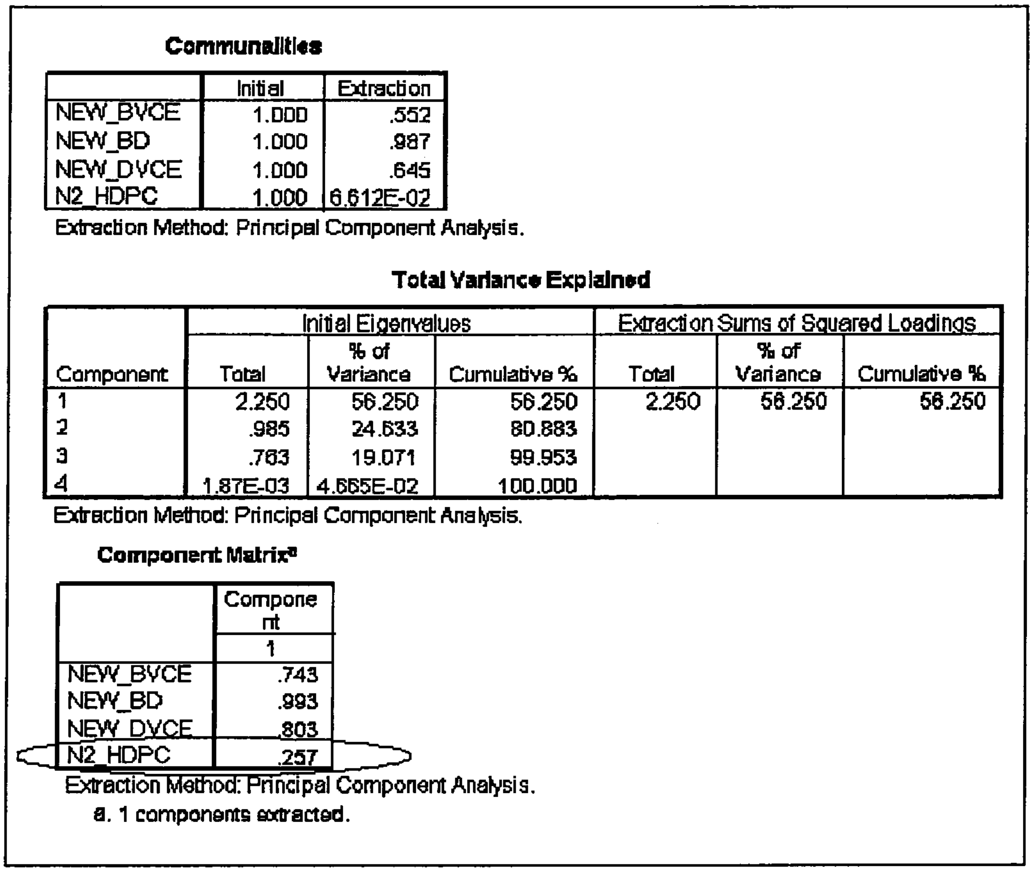
```
  R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (A L P H A)


 Item-total Statistics

                 Scale          Scale      Corrected
                  Mean        Variance       Item-           Alpha
                 if Item       if Item       Total          if Item
                 Deleted       Deleted    Correlation       Deleted

NEW_BVCE         20.2093      285.5826        .5402           .6697
NEW_BD           14.7923      120.8925        .9502           .3399
NEW_DVCE         18.0578      249.8993        .6288           .6121
N2_HDPC          17.2191      398.4383        .1471           .8149



Reliability Coefficients

N of Cases =   48706.0                     N of Items =  4

Alpha =      .7330
```

Figure 5-3 4 Predictors RA Result of Product Group

The following statement was applied to the remaining 4 predictors to examine how many sub-factors can be extracted from the product group:

```
FACTOR
 /VARIABLES new_bvce new_bd new_dvce n2_hdpc  /MISSING LISTWISE
/ANALYSIS
 new_bvce new_bd new_dvce n2_hdpc
 /PRINT INITIAL EXTRACTION ROTATION
 /CRITERIA MINEIGEN(1) ITERATE(25)
 /EXTRACTION PC
 /CRITERIA ITERATE(25)
 /ROTATION VARIMAX
 /METHOD=CORRELATION .
```

The corresponding result is shown in Figure 5-4. Since only one sub-factor was extracted with eigenvalues above 1.0, the solution cannot be rotated. We notice that N2_HDPC (Handset Price) has a lower loading (.257) in the matrix when compared with other 3 predictors. However, according to our literature review, the product group may have two sub types—price and quality. Thus we force SPSS to extract 2 sub-factors by setting a fixed number of 2 factors for SPSS solution.



Communalities

	Initial	Extraction
NEW_BVCE	1.000	.552
NEW_BD	1.000	.987
NEW_DVCE	1.000	.645
N2_HDPC	1.000	6.612E-02

Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.250	56.250	56.250	2.250	56.250	56.250
2	.985	24.633	80.883			
3	.763	19.071	99.953			
4	1.87E-03	4.665E-02	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix[a]

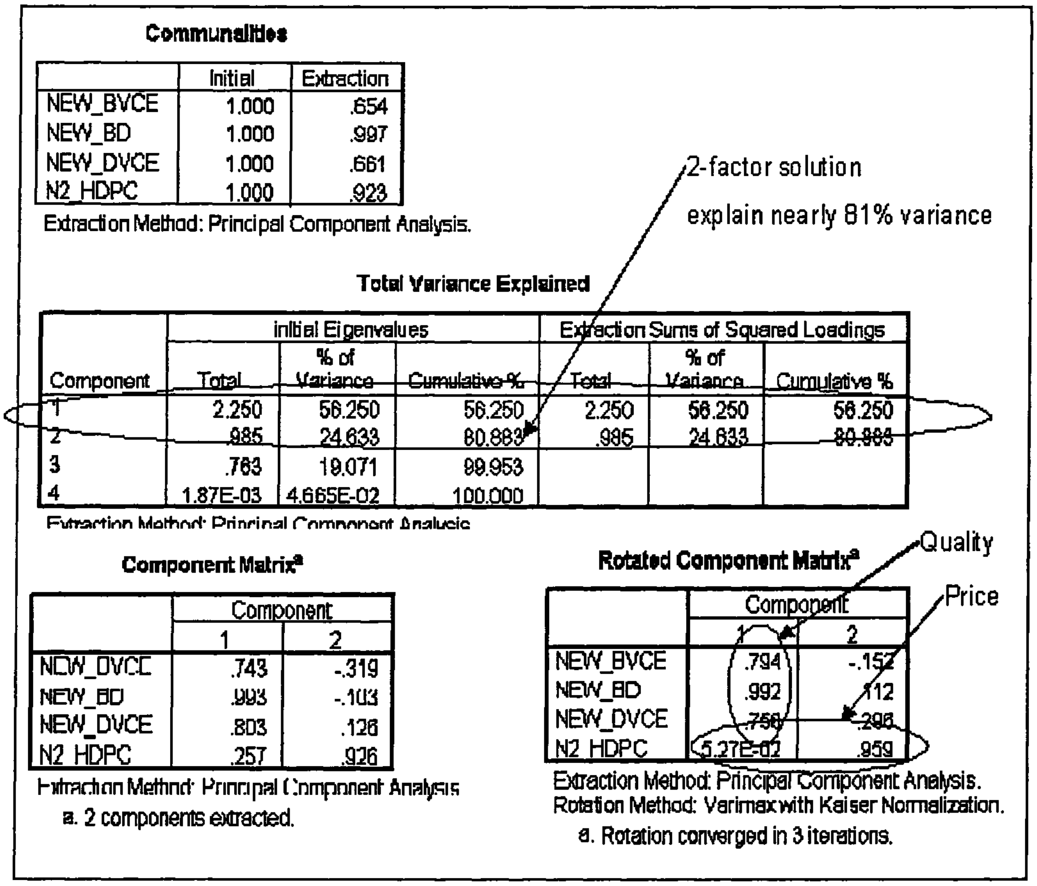
	Component
	1
NEW_BVCE	.743
NEW_BD	.993
NEW_DVCE	.803
N2_HDPC	.257

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

Figure 5-4 PCA Result of Product Group

The 2-factor solution is shown in Figure 5-5. Variable NEW_BVCE (.794), NEW_BD (.992), and NEW_DVCE (.758) load on factor 1, and N2_HDPC (.959) loads on factor 2. The first sub-factor is related to quality and the second sub-factor is related to price. These 2 sub-factors explain nearly 81% variance of the 4 observed predictors.



Communalities

	Initial	Extraction
NEW_BVCE	1.000	.654
NEW_BD	1.000	.997
NEW_DVCE	1.000	.661
N2_HDPC	1.000	.923

Extraction Method: Principal Component Analysis.

2-factor solution

explain nearly 81% variance

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.250	56.250	56.250	2.250	56.250	56.250
2	.985	24.633	80.883	.985	24.633	80.883
3	.763	19.071	99.953			
4	1.87E-03	4.665E-02	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix[a]

	Component	
	1	2
NEW_BVCE	.743	-.319
NEW_BD	.993	-.103
NEW_DVCE	.803	.126
N2_HDPC	.257	.926

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Rotated Component Matrix[a]

Quality

Price

	Component	
	1	2
NEW_BVCE	.794	-.152
NEW_BD	.992	.112
NEW_DVCE	.758	.296
N2_HDPC	5.27E-02	.959

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Figure 5-5 2-factor Solution of Product Group

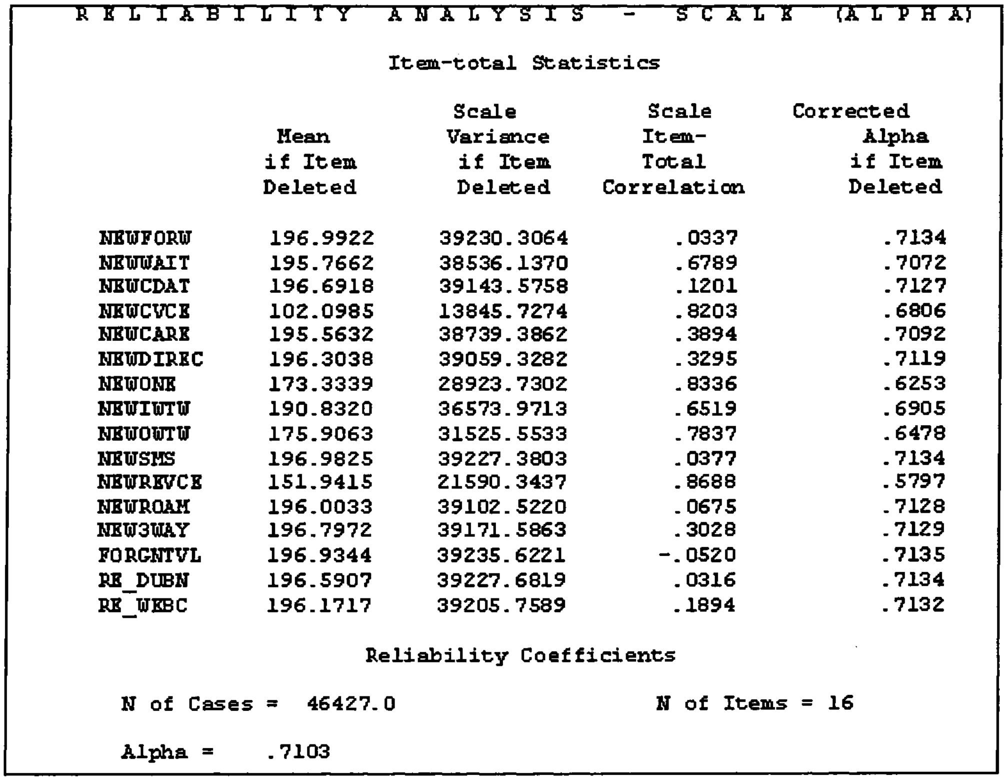
5.4.2 Customer Solution Group

The RA experimental result of the original 16 preprocessed predictors (see Table 5-3) is shown in Figure 5-6. From this Figure, it is seen that the 16 predictors are highly correlated with each other (α= .7103).

Table 5-3 Variable Names for Predictors of Customer Solution Group

Predictors	Variable name in RA & PCA
Call Forwarding Calls	NEWFORW
Call Waiting Calls	NEWWAIT
Customer Care Calls	NEWCARE
Completed Data calls	NEWCDAT
Completed Voice Calls	NEWCVCE
Directory Assisted Calls	NEWDIREC

DualBand	RE_DUBN
Foreign Travel Dummy	FORGNTVL
Handset Web Capability	RE_WEBC
Inbound Calls Less than one Minute	NEWONE
Inbound Wireless to Wireless Calls	NEWIWTW
Outbound Wireless to Wireless Calls	NEWOWTW
Received SMS Calls	NEWSMS
Received Voice Calls	NEWREVCE
Roaming Calls	NEWROAM
Three-way Calls	NEW3WAY



```
    RELIABILITY ANALYSIS - SCALE (ALPHA)

                          Item-total Statistics

                          Scale          Scale      Corrected
             Mean        Variance       Item-           Alpha
            if Item      if Item        Total         if Item
            Deleted      Deleted      Correlation     Deleted

NEWFORW     196.9922    39230.3064        .0337         .7134
NEWWAIT     195.7662    38536.1370        .6789         .7072
NEWCDAT     196.6918    39143.5758        .1201         .7127
NEWCVCE     102.0985    13845.7274        .8203         .6806
NEWCARE     195.5632    38739.3862        .3894         .7092
NEWDIREC    196.3038    39059.3282        .3295         .7119
NEWONE      173.3339    28923.7302        .8336         .6253
NEWIWTW     190.8320    36573.9713        .6519         .6905
NEWOWTW     175.9063    31525.5533        .7837         .6478
NEWSMS      196.9825    39227.3803        .0377         .7134
NEWREVCE    151.9415    21590.3437        .8688         .5797
NEWROAM     196.0033    39102.5220        .0675         .7128
NEW3WAY     196.7972    39171.5863        .3028         .7129
FORGNTVL    196.9344    39235.6221       -.0520         .7135
RE_DUBN     196.5907    39227.6819        .0316         .7134
RE_WEBC     196.1717    39205.7589        .1894         .7132

                       Reliability Coefficients

  N of Cases =  46427.0                     N of Items = 16

  Alpha =    .7103
```

Figure 5-6 Reliability Analysis Result of Customer Solution Group

The following statement was applied to the 16 predictors to examine how many sub-factors can be extracted from the customer solution group:

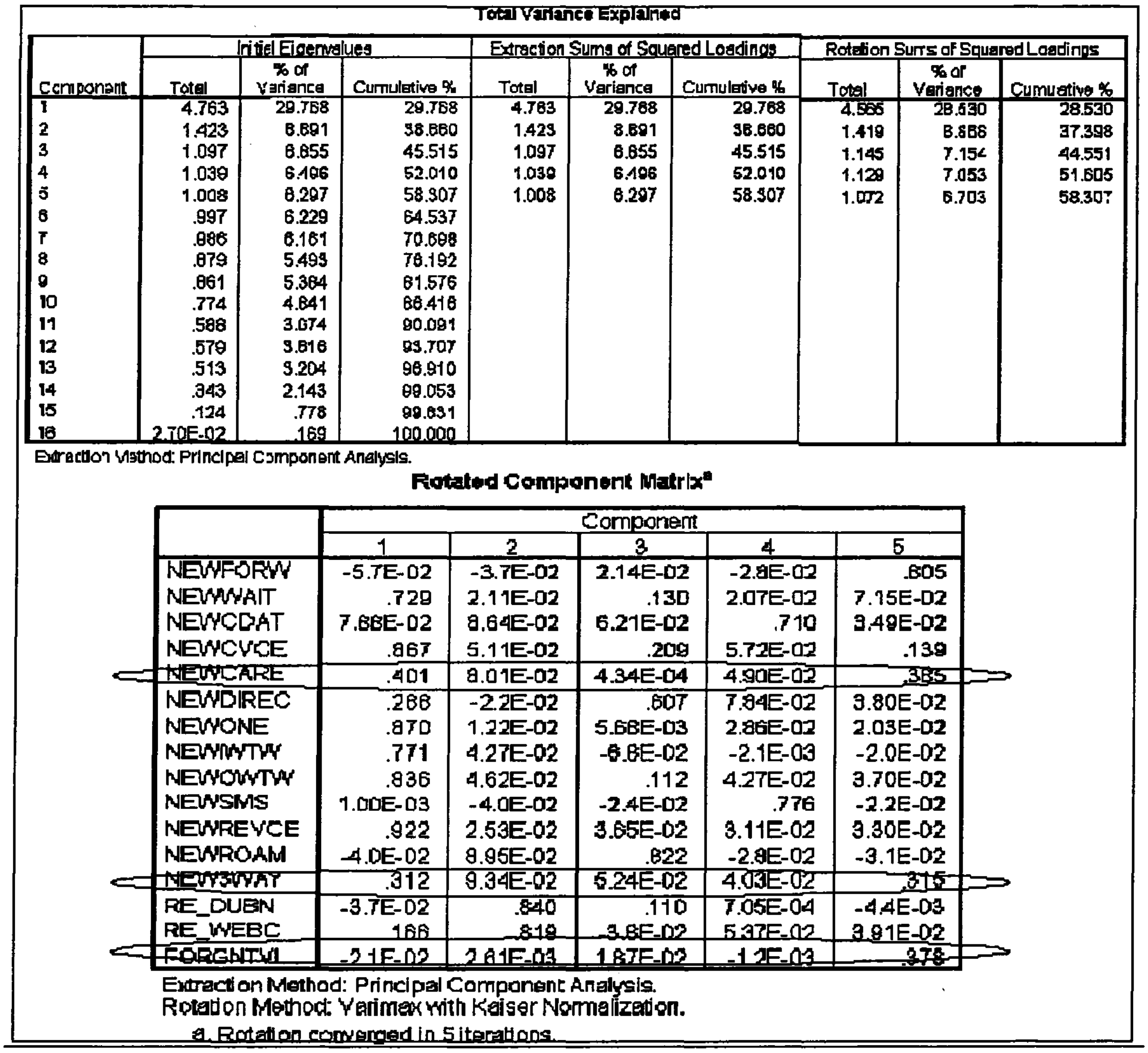
```
FACTOR
 /VARIABLES newforw newwait newcdat newcvce newcare newdirec newone newiwtw
```

```
newowtw newsms newrevce newroam new3way re_dubn re_webc forgntvl  /MISSING
LISTWISE /ANALYSIS newforw newwait newcdat newcvce newcare newdirec newone
newiwtw newowtw newsms newrevce newroam new3way re_dubn re_webc forgntvl
/PRINT INITIAL EXTRACTION ROTATION
/CRITERIA MINEIGEN(1) ITERATE(25)
/EXTRACTION PC
/CRITERIA ITERATE(25)
        /ROTATION VARIMAX
  /METHOD=CORRELATION
```

Five sub-factors with eigenvalues above 1.0 resulted (Figure 5-7). With a semantic analysis of these predictors, all of the 5 factors can be interpreted well with specific solution meaning (We will discuss these meanings in section 5.5). From the Rotated Matrix, one can find that items all loaded highly (>0.6) into different factors except 3 predictors:

- Customer Care Calls (NEWCARE , .401)
- Three-way Calls (NEW3WAY, .315)
- Foreign Travel Dummy (FORGNTVL, .378)

Further Reliability Analysis showed that dropping these 3 items improved (α= .7153) the reliability of this customer solution predictor group. Therefore, Customer Care Calls, Three-way Calls, and Foreign Travel Dummy were removed from customer solution group.



Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumuative %
1	4.763	29.768	29.768	4.763	29.768	29.768	4.565	28.530	28.530
2	1.423	8.891	38.660	1.423	8.891	38.660	1.419	8.868	37.398
3	1.097	6.855	45.515	1.097	6.855	45.515	1.145	7.154	44.551
4	1.039	6.496	52.010	1.039	6.496	52.010	1.129	7.053	51.605
5	1.008	6.297	58.307	1.008	6.297	58.307	1.072	6.703	58.307
6	.997	6.229	64.537						
7	.986	6.161	70.698						
8	.879	5.493	76.192						
9	.861	5.384	81.576						
10	.774	4.841	86.416						
11	.588	3.674	90.091						
12	.579	3.616	93.707						
13	.513	3.204	96.910						
14	.343	2.143	99.053						
15	.124	.778	99.831						
16	2.70E-02	.169	100.000						

Extraction Method: Principal Component Analysis.

Rotated Component Matrix[a]

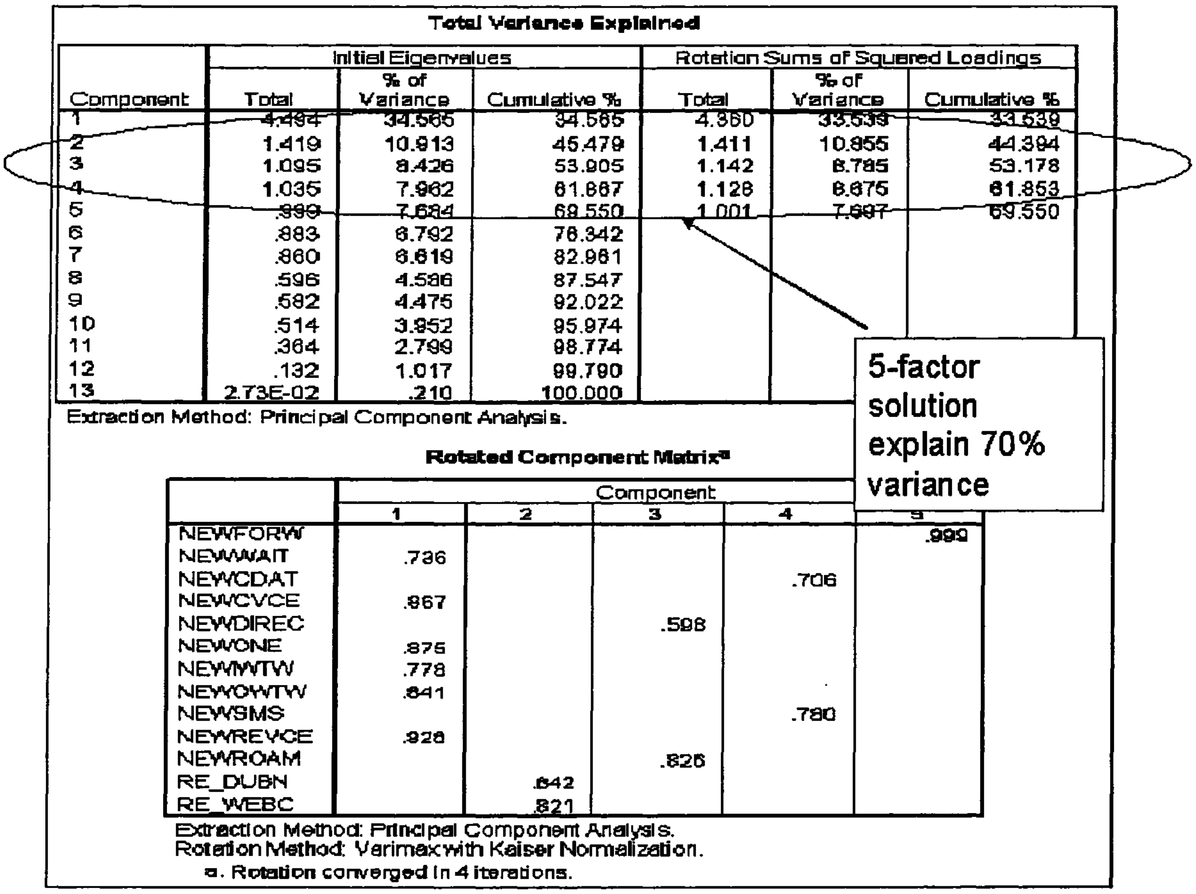
	Component				
	1	2	3	4	5
NEWFORW	-5.7E-02	-3.7E-02	2.14E-02	-2.8E-02	.605
NEWWAIT	.729	2.11E-02	.130	2.07E-02	7.15E-02
NEWCDAT	7.68E-02	8.64E-02	6.21E-02	.710	3.49E-02
NEWCVCE	.867	5.11E-02	.209	5.72E-02	.139
NEWCARE	.401	8.01E-02	4.34E-04	4.90E-02	.385
NEWDIREC	.268	-2.2E-02	.607	7.84E-02	3.80E-02
NEWONE	.870	1.22E-02	5.68E-03	2.86E-02	2.03E-02
NEWIWTW	.771	4.27E-02	-6.8E-02	-2.1E-03	-2.0E-02
NEWOWTW	.836	4.62E-02	.112	4.27E-02	3.70E-02
NEWSMS	1.00E-03	-4.0E-02	-2.4E-02	.776	-2.2E-02
NEWREVCE	.922	2.53E-02	3.65E-02	3.11E-02	3.30E-02
NEWROAM	-4.0E-02	8.95E-02	.822	-2.8E-02	-3.1E-02
NEW3WAY	.312	9.34E-02	5.24E-02	4.03E-02	.315
RE_DUBN	-3.7E-02	.840	.110	7.05E-04	-4.4E-03
RE_WEBC	.166	.819	-3.8E-02	5.37E-02	3.91E-02
FORGNTVI	-2.1E-02	2.61E-03	1.87E-02	-1.2E-03	.378

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Figure 5-7 PCA Results of Customer Solution Group

For the retained 13 predictors, PCA resulted in a 4-factor solution instead of 5-factor solution. One factor of "forwarding" was missed. Hence we have to specify a 5-factor solution when PCA was conducted on the 13 retained predictors. The 5-factor solution (Figure 5-8) showed that items loaded into each factor are the same as Figure 5-7, however, they explained more (nearly 70%) variance of the current group of predictors than that of the original group of predictors (58%).



Total Variance Explained

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.494	34.565	34.565	4.360	33.539	33.539
2	1.419	10.913	45.479	1.411	10.855	44.394
3	1.095	8.426	53.905	1.142	8.785	53.178
4	1.035	7.962	61.867	1.128	8.675	61.853
5	.999	7.684	69.550	1.001	7.697	69.550
6	.883	6.792	76.342			
7	.860	6.619	82.961			
8	.596	4.586	87.547			
9	.582	4.475	92.022			
10	.514	3.952	95.974			
11	.364	2.799	98.774			
12	.132	1.017	99.790			
13	2.73E-02	.210	100.000			

Extraction Method: Principal Component Analysis.

5-factor solution explain 70% variance

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
NEWFORW					.999
NEWWAIT	.736				
NEWCDAT				.706	
NEWCVCE	.867				
NEWDIREC			.598		
NEWONE	.875				
NEWMWTW	.778				
NEWOWTW	.841				
NEWSMS				.780	
NEWREVCE	.928				
NEWROAM			.826		
RE_DUBN		.842			
RE_WEBC		.821			

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 4 iterations.

Figure 5-8 5-Factor Solution of Customer Solution Group

5.4.3 Customer Profile Group

The RA experiment result (α= .4951, Figure 5-9) showed that these 33 predictors (Table 5-4) grouped under customer profile strategy are not statistically correlated with each other. One or more items should be dropped until we get a group of statistically reliable predictors. Based on both Alpha value in the right side of Figure 5-9 and the continuous RA results and Table 3-9, we remove a series of items until we get a group of predictors which are reliable between each other. This continuous removing procedure is presented on Table 5-5. After 7 steps, 26 predictors were statistically related (α= .7029), and were involved in the following exploratory factor analysis.

Table 5-4 Variable Names for Predictors of Customer Profile Group

Predictors	Variable name in RA & PCA
Account Spending Limit	ASLFLAG2
Active Subscribers in Household	REN_ASUB
Average monthly revenue over the previous three months	REN_AVG3
Average monthly revenue over the previous six months	REN_AVG6
Average monthly revenue over the life of the customer	REN_AVG
Billing adjusted total revenue over the life of the customer	REN_ADJ
Credit Class	RECLSCOD
Ethnicity	REETHN
Geographic Area	REN_AREA
Handset Refurnish Indicator	REN_REFL
Monthly Revenue (Charge Amount)	REN_REV
Monthly Recurring Charge	REN_MRC
Motorcycle Indicator	MTRCYCLE
Number of Days of Current Equipment	REN_EQP
Number of Days Since Last Retention Call	REN_RDAY
Number of Handsets Issued	REN_PHE
Number of Models Issued	REN_MOD
Percentage Change in Monthly Revenue vs Previous Three Month Average	REN_CHG
Possession Number of Credit Cards	CRED#
Revenue of Data Overage	REN_DATORE
Revenue of Voice Overage	REN_VCEO
Referral Numbers	REN_REFL
RV indicator	RV
Social Group	RESCGRP
Total Number of Months in Service	REN_MTH
Total Overage Revenue	REN_OVR
Total Revenue	REN_TREV
Truck Indicator	TRUCK
Unique Subscribers in Household	REN_UNT
Calls to Retention Team	REN_REN2
Absence of Call to Retention Team	RETEN1
Absence of Offer	ACCEPT1
Offers Accepted from Retention Team	REN_ACP2

```
  R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (A L P H A)

Item-total Statistics

                 Scale          Scale       Corrected
                 Mean          Variance       Item-            Alpha
                if Item         if Item       Total           if Item
                Deleted         Deleted    Correlation        Deleted

REN_ASUB         7.8585         2.5978        .0524            .4945
REN_AVG3         7.9755         2.4309        .4703            .4560
REN_AVG6         7.9414         2.4321        .4665            .4563
REN_AVG          7.9293         2.4459        .4231            .4598
REN_ADJ          7.9885         2.3891        .5313            .4470
REN_REV          7.9583         2.4415        .4911            .4573
REN_MRC          7.8476         2.5002        .2817            .4727
REN_EQP          7.9072         2.6708       -.0998            .5129
REN_RDAY         8.1872         2.6036        .1825            .4887
REN_MOD          8.0747         2.4616        .2471            .4707
REN_PHE          8.0604         2.4478        .2458            .4697
REN_CHG          7.6957         2.6542       -.0689            .4989
REN_DATO         8.1832         2.6289        .0604            .4938
REN_VCEO         8.1140         2.4659        .3344            .4660
REN_MTH          7.9446         2.4766        .2431            .4724
REN_OVR          8.1141         2.4663        .3372            .4659
REN_TREV         7.9785         2.3917        .5319            .4475
REN_UNI          8.0912         2.5912        .0528            .4948
RECLSCOD         7.3452         2.6157       -.0230            .5072
REETHN           7.6245         2.5916       -.0019            .5062
REN_AREA         7.7244         2.5426        .0210            .5076
RESCGRP          7.6069         2.5993       -.0364            .5171
REN_REFL         8.1781         2.6387       -.0017            .4968
REN_RENZ         8.1860         2.5992        .2291            .4875
REN_ACP2         8.1900         2.6176        .1953            .4907
MTRCYCLE         8.1793         2.5998        .0767            .4921
RV               8.1095         2.3813        .2158            .4699
TRUCK            7.9991         2.2684        .1834            .4781
ASLFLAG2         8.0619         2.7212       -.1714            .5542
RE_FUR           7.3353         2.6488       -.1109            .5437
CRED#            7.4815         2.3528        .0625            .5189
RETEN1           8.1718         2.5387        .1775            .4821
ACCEPT1          8.1718         2.5387        .1775            .4821
□


  R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (A L P H A)


Reliability Coefficients

N of Cases =   41473.0                     N of Items = 33

Alpha =     .4951
```

Figure 5-9 RA Result of Customer Profile Group

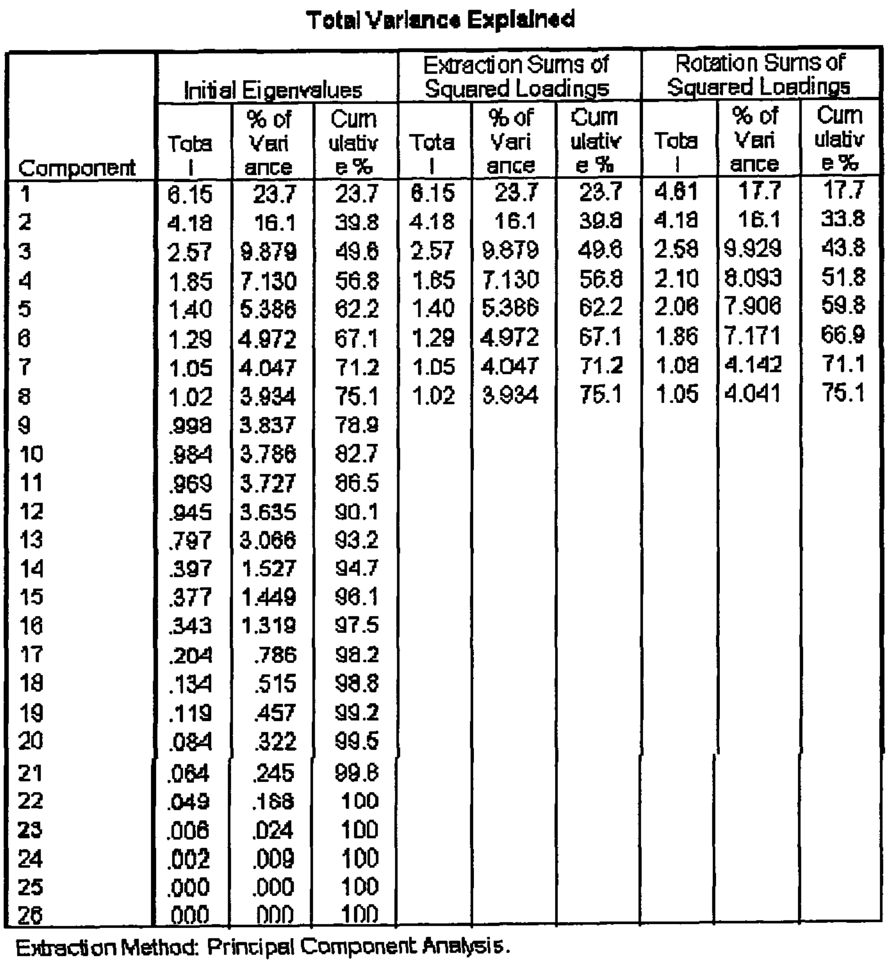
Table 5-5 Item Removing Procedure of Customer Profile Group

Steps	Dropped Variable(s)	Alpha
1		.4951
2	Account Spending Limit	.5542
3	Handset Refurnish Indicator	.6017
4	Possession Number of Credit Cards	.6295
5	Number of Days of Current Equipment, Truck Indicator	.6731
6	RV	.6971
7	Motorcycle Indicator	.7029

The following statement was applied to the 26 remaining predictors to examine how many sub-factors can be extracted from the customer profile group:

```
FACTOR
 /VARIABLES ren_asub ren_avg3 ren_avg6 ren_avg ren_adj ren_rev ren_mrc
 ren_rday ren_mod ren_phe ren_chg ren_dato ren_vceo ren_mth ren_ovr ren_trev
 ren_uni reclscod reethn ren_area rescgrp ren_ren2 ren_acp2 accept1 reten1
 ren_refl  /MISSING LISTWISE /ANALYSIS ren_asub ren_avg3 ren_avg6 ren_avg
 ren_adj ren_rev ren_mrc ren_rday ren_mod ren_phe ren_chg ren_dato ren_vceo
 ren_mth ren_ovr ren_trev ren_uni reclscod reethn ren_area rescgrp ren_ren2
 ren_acp2 accept1 reten1 ren_refl
 /PRINT INITIAL EXTRACTION ROTATION
 /FORMAT BLANK(.55)
 /CRITERIA MINEIGEN(1) ITERATE(25)
 /EXTRACTION PC
 /CRITERIA ITERATE(25)
 /ROTATION VARIMAX
 /METHOD=CORRELATION .
```

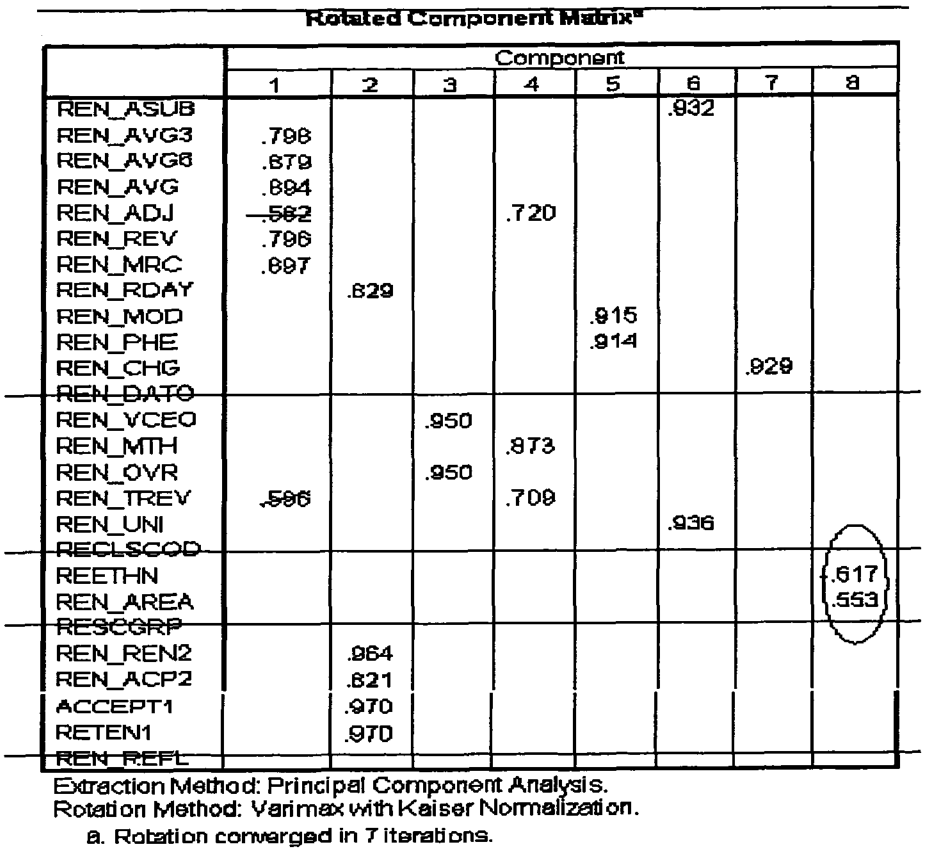
Eight factors with eigenvalues above 1.0 resulted (Figure 5-10). With a semantic analysis of these predictors, the first 7 factors can be interpreted well with specific solution meaning (we will discuss these meanings in section 5.5), but the eighth factor can not be interpreted well. From the Rotated Matrix (only showing factor loadings that are greater than .55), one can find that 20 items loaded highly together (>0.6) into the first 7 factors, and each of them can be specified to a single sub factor, but the other 6 factors have lower loading in any of the 8 factors. The 6 predictors with lower loading are Revenue of Data Overage, Credit Class, Ethnicity, Referral Numbers, Social Group, and Geographic area. We notice that Social Group and Geographic area have higher loadings within the eighth factor (.553 and .378), and they are related to the location predictor in the literature (Mozer *et al.,* 2000; Yan *et al.*, 2001; Au *et al.*, 2003). Although Ethnicity also has a higher loading (-.617), it cannot be classified into location. With the 20 high loading predictors and Social Group and Geographic area, further PCA resulted in an 8-factor solution (Figure 5-11) which has the same first 7 factors as the previous solution, however Social Group (.641) and Geographic area (.781) loaded highly together into the eighth factor. Reliability Analysis also showed that the reliability of customer profile group is improved (.7510) after the other 4 lower loading predictors (Revenue of Data Overage, Credit Class, Ethnicity, and Referral Number) were deleted. Hence there are in total 22 predictors remaining in the customer profile group.



Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.15	23.7	23.7	6.15	23.7	23.7	4.61	17.7	17.7
2	4.18	16.1	39.8	4.18	16.1	39.8	4.18	16.1	33.8
3	2.57	9.879	49.6	2.57	9.879	49.6	2.58	9.929	43.8
4	1.85	7.130	56.8	1.85	7.130	56.8	2.10	8.093	51.8
5	1.40	5.386	62.2	1.40	5.386	62.2	2.06	7.906	59.8
6	1.29	4.972	67.1	1.29	4.972	67.1	1.86	7.171	66.9
7	1.05	4.047	71.2	1.05	4.047	71.2	1.08	4.142	71.1
8	1.02	3.934	75.1	1.02	3.934	75.1	1.05	4.041	75.1
9	.998	3.837	78.9						
10	.984	3.786	82.7						
11	.969	3.727	86.5						
12	.945	3.635	90.1						
13	.797	3.066	93.2						
14	.397	1.527	94.7						
15	.377	1.449	96.1						
16	.343	1.319	97.5						
17	.204	.786	98.2						
18	.134	.515	98.8						
19	.119	.457	99.2						
20	.084	.322	99.5						
21	.064	.245	99.8						
22	.049	.188	100						
23	.006	.024	100						
24	.002	.009	100						
25	.000	.000	100						
26	.000	.000	100						

Extraction Method: Principal Component Analysis.



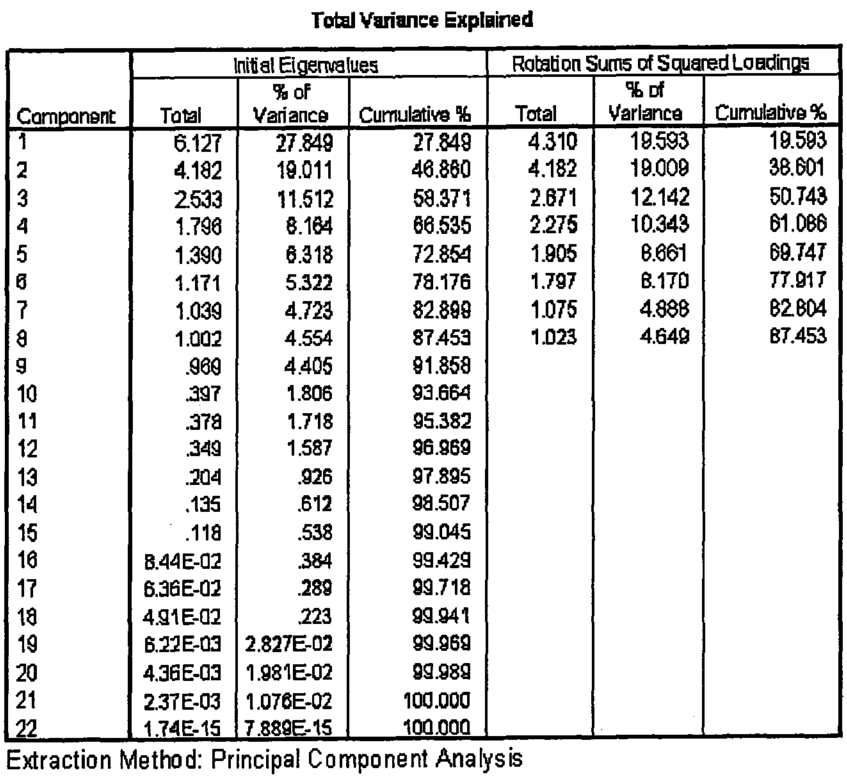
Rotated Component Matrix[a]

	Component							
	1	2	3	4	5	6	7	8
REN_ASUB						.932		
REN_AVG3	.798							
REN_AVG6	.879							
REN_AVG	.894							
REN_ADJ	~~.582~~			.720				
REN_REV	.796							
REN_MRC	.897							
REN_RDAY		.829						
REN_MOD					.915			
REN_PHE					.914			
REN_CHG							.929	
~~REN_DATO~~								
REN_VCEO			.950					
REN_MTH				.873				
REN_OVR			.950					
REN_TREV	~~.596~~			.709				
REN_UNI						.936		
~~RECLSCOD~~								
REETHN								.617
REN_AREA								.553
~~RESCGRP~~								
REN_REN2		.964						
REN_ACP2		.821						
ACCEPT1		.970						
RETEN1		.970						
~~REN_REFL~~								

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

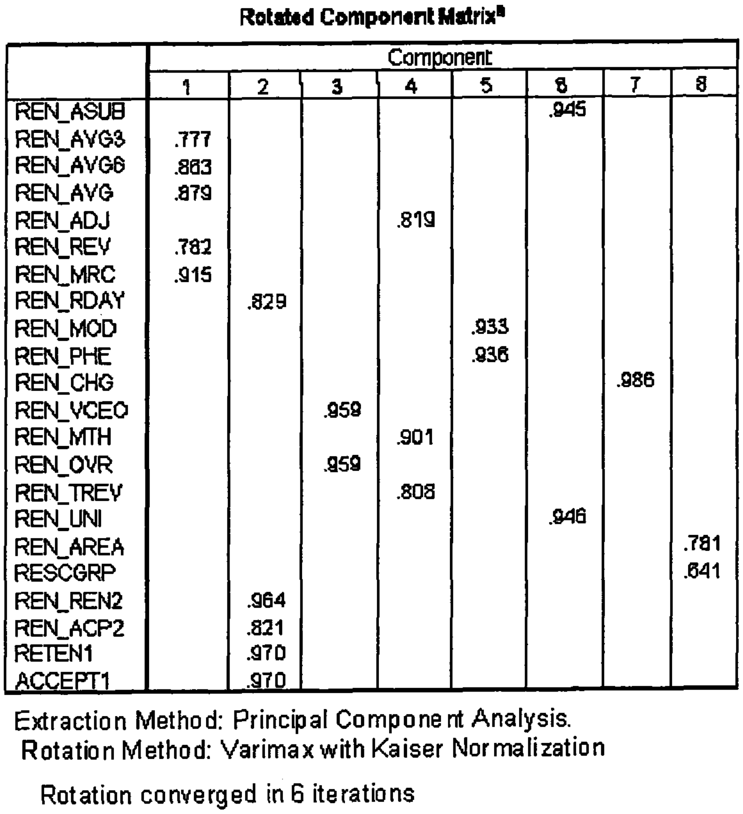
Figure 5-10 PCA Result of Customer Profile Group



Total Variance Explained

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.127	27.849	27.849	4.310	19.593	19.593
2	4.182	19.011	46.860	4.182	19.009	38.601
3	2.533	11.512	58.371	2.671	12.142	50.743
4	1.796	8.164	66.535	2.275	10.343	61.086
5	1.390	6.318	72.854	1.905	8.661	69.747
6	1.171	5.322	78.176	1.797	8.170	77.917
7	1.039	4.723	82.899	1.075	4.888	82.804
8	1.002	4.554	87.453	1.023	4.649	87.453
9	.969	4.405	91.858			
10	.397	1.806	93.664			
11	.378	1.718	95.382			
12	.349	1.587	96.969			
13	.204	.926	97.895			
14	.135	.612	98.507			
15	.118	.538	99.045			
16	8.44E-02	.384	99.429			
17	6.36E-02	.289	99.718			
18	4.91E-02	.223	99.941			
19	6.22E-03	2.827E-02	99.969			
20	4.36E-03	1.981E-02	99.989			
21	2.37E-03	1.076E-02	100.000			
22	1.74E-15	7.889E-15	100.000			

Extraction Method: Principal Component Analysis



Rotated Component Matrix[a]

	Component							
	1	2	3	4	5	6	7	8
REN_ASUB						.945		
REN_AVG3	.777							
REN_AVG6	.863							
REN_AVG	.879							
REN_ADJ				.819				
REN_REV	.782							
REN_MRC	.915							
REN_RDAY		.829						
REN_MOD					.933			
REN_PHE					.936			
REN_CHG							.986	
REN_VCEO			.959					
REN_MTH				.901				
REN_OVR			.959					
REN_TREV				.808				
REN_UNI						.946		
REN_AREA								.781
RESCGRP								.641
REN_REN2		.964						
REN_ACP2		.821						
RETEN1		.970						
ACCEPT1		.970						

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization

Rotation converged in 6 iterations

Figure 5-11 New PCA Result of Customer Profile Group

5.4.4 Summary of Results of Exploratory Sub-factor Analysis

Reliability Analysis showed that with Alpha =.7330, 4 predictors are extracted in the product group; with Alpha =.7153, 13 predictors are extracted in the solution group; and with Alpha =.7510, 22 predictors are extracted in the profile group. These predictors are statistically related to each other within each group. PCA results showed that there are 2 sub-factors in the product group, 5 sub-factors in the solution group, and 8 sub-factors in the profile group; and they explain 81%, 70%, and 87.5% of the cumulative variance of each group respectively. The result predictors and sub-factors within each group are shown in Table 5-6.

Table 5-6 Statistically Related Predictors and Sub-factors Within Each Group

	Product	Customer Solution	Customer Profile
Factor 1	Blocked or Dropped Calls Blocked Voice Calls Dropped Voice Calls	Call Waiting Calls Completed Voice Calls Inbound Calls Less than one Minute Inbound Wireless to Wireless Calls Outbound Wireless to Wireless Calls Received Voice Calls	Average monthly revenue over the previous three months Average monthly revenue over the previous six months Average monthly revenue over the life of the customer Monthly Revenue (Charge Amount) Monthly Recurring Charge
Factor 2	Handset Price	DualBand Handset Web Capability	Calls to Retention Team Absence of Call to Retention Team Number of Days Since Last Retention Call Absence of Offer

			Offers Accepted from Retention Team
Factor 3		Completed Data calls Received SMS Calls	Revenue of Voice Overage Total Overage Revenue
Factor 4		Directory Assisted Calls Roaming Calls	Billing adjusted total revenue over the life of the customer Total Number of Months in Service Total Revenue
Factor 5		Call Forwarding Calls	Number of Handsets Issued Number of Models Issued
Factor 6			Active Subscribers in Household Unique Subscribers in Household
Factor 7			Percentage Change in Monthly Revenue vs Previous Three Month Average
Factor 8			Geographic Area Social Group

5.5 The Refined CSAM

According to these statistical results (Table 5-6), a refined model with grouped predictors and main patterns and sub-factors was obtained from the data. Figure 5-12 shows the refined model. The two factors under the Product Strategy are easily identified as "quality" and "price" respectively.

Under the Customer Solution Strategy, we found that factor 1 is different from the other factors. Factor 1 shows the normal usage of the wireless, while the other factors represent the usage of capabilities such as: web, data, geographic, and call forwarding. Thus we interpret this as two patterns of usage under Customer Solution: "standard usage" and "extended usage". Standard usage refers to the normal voice usage for a mobile handset, while the extended usage refers to other capabilities. According to the special features within each factor, we label factor 2 to 5 "web capability", "data capability", "geographical capability", and "call forwarding capability" respectively.

Under Customer Profile Strategy, factor 1 represents how much a customer paid in the past; we label the factor "willingness to pay". Factor 2 shows if a customer is willing to contact the company and how many times he/she contacted; thus it was identified as "service loyalty". Factor 3 shows how much a customer paid more than the average costs; it was tagged as "extra spend". Factor 4 is related to the customer life in the company; thus it represents "tenure". Factor 5 describes how many handsets and models were changed in the past; it represents "willingness to change handset or model" which is related to customer's adoption of an innovation. Factor 6 is about the number of subscribers in a household; thus it was identified as "number of subscribers". Factor 7 shows a usage change in the past, so it was labeled "usage change". Factor 8 shows the social and geographical location of a customer, so we labeled it "resident location".

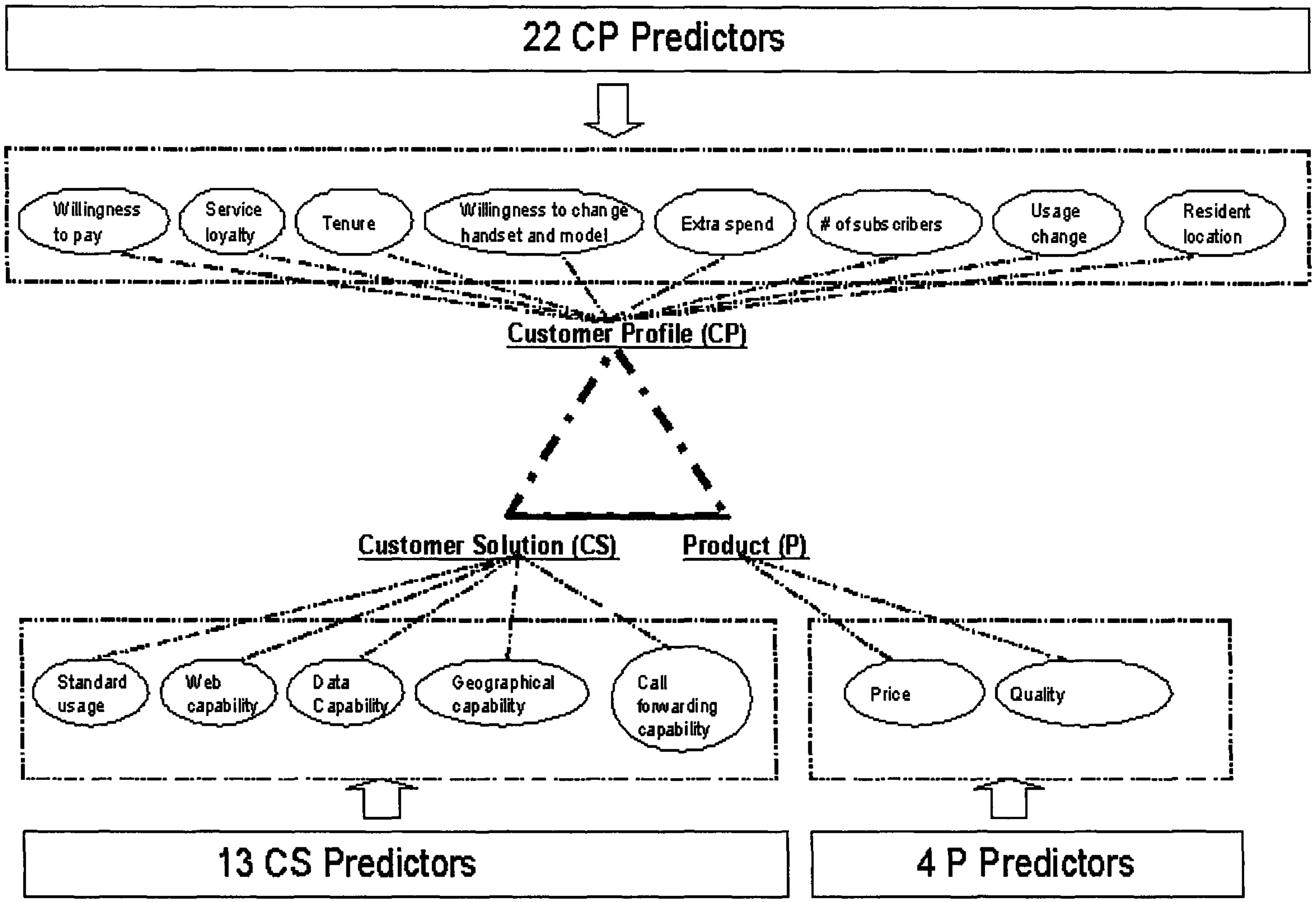


Figure 5-12 A Case of Refined CSAM Model

For the managerial purpose, it is also useful to simplify Figure 12 to the Figure 13, the refined CSAM model with only high-level interpretable sub-factors.

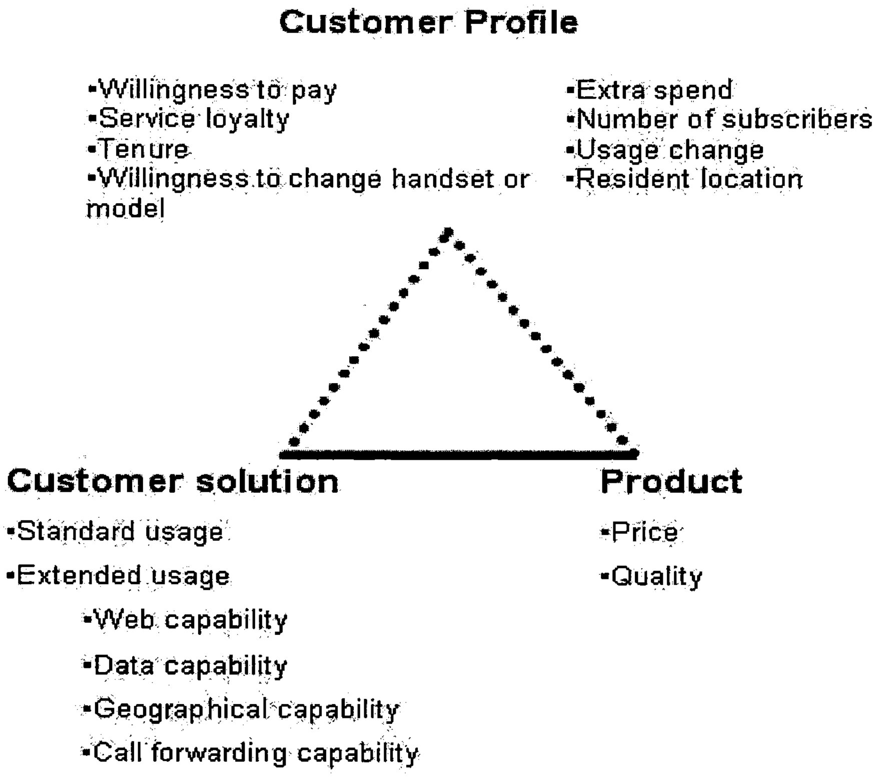


Figure 5-13 Refined CSAM with only Sub-Factors

5.6 Drill-down Analyses and Examples of The Application of Refined CSAM

Further analysis is shown for one sub-factor in each of customer solution, customer profile, and product groups in order to illustrate how drilling down and specific low level variables can be further examined after application of CSAM.

We use t-Tests, One-way Analyses of Variance (ANOVA), and factorial analyses of variance to refine examination of the identified churn factors in CSAM. This section reports test results in Table 5-7 for one example sub-factor per CSAM group.

In the product CSAM group, we use the price sub-factor as the example. The frequency of handset price predictor and the mean of churn indicator corresponding for each level of the handset price are as follows (exclude level 7 and 10, since they just have 1 customer, we ignored them here). We found that the mean of churn indicator is clearly going down from level 1 to level 9, and shift up and down from level 11 to level 17. After examine the customer number from level 11 to level 17, we found that the top two largest group (level 12 and 14) both have lower mean number of the churn indicator (.41, .29), and they have 5263 customers in total. The total customer number of the other groups from level 11 to level 17 is 221; and only in level 11 and level 13, the means of churn indicator are more than .5. Based on the observation above, we want to group some levels together. We also found that from level 1 to level 9, the means of churn indicator are more than .5 in the first 4 levels. Thus we intend to split the handset price predictor into 3 groups: level 1-4, level 5-9, and level 11-17. Since we don't want a price gap between the last two groups, and from the price point of view, level 10 ($159.99) is more near level 9 ($149.99) than level 11 ($179.99), we just set level 10 into group 2.

	Frequency	Mean of churn indicator	95% Confidence Interval for Mean	
			Lower Bound	Upper Bound
1.00	2148	.58	.56	.60
2.00	10940	.57	.56	.58
3.00	314	.56	.51	.62
4.00	4324	.52	.51	.54
5.00	4842	.50	.48	.51
6.00	3879	.48	.46	.49
7.00	1			

8.00	6726	.46	.45	.47
9.00	11015	.46	.45	.47
10.00	1			
11.00	35	.56	.39	.73
12.00	5156	.41	.40	.42
13.00	19	.58	.33	.82
14.00	107	.29	.20	.38
15.00	43	.47	.31	.62
16.00	97	.44	.34	.54
17.00	27	.30	.11	.48
Total	49674			

Based on the analysis above, we split handset prices into 3 subgroups: less than $59.99, between $59.99 and 159.99, and more than 159.99; and label them "low level price", "middle level price", and "high level price" respectively. One way ANOVA with a priori was conducted to examine the impact of handset price on churn. The result of the one way ANOVA show that for this data set there is significant difference between product subgroups on churn, *F(2, 49671) = 257.811, p < 0.01*. The results of a priori contrasts show that the mean difference is significant at 0.016 level between each pair of subgroups (*t(38131) = 18.29, p < 0.01 (2-tailed); t(9205) = 19.518, p < 0.01 (2-tailed); t(7998) = 8.247, p < 0.01 (2-tailed)*). The low level price subgroup (*M=.56, SD=.497*) have more churners than the middle level price subgroup (*M=.47, SD=.499*) and the high level price subgroup (*M=.41, SD=.492*). These results represent that customers with lower handset price are more likely to churn.

In the solution CSAM group, factorial analysis of variance was conducted to examine the impact of web capability on churn. Both main effects (Dualband and Handset Web Capability) and the interaction of web capability are significant (*F(2, 49966) = 8.63, p < 0.01; F(2, 49966) = 143.907, p < 0.01; F(2, 49966) = 12.872, p < 0.01*). Dualband has three levels; from low to high are "no dualband", "dualband", and "tri-model (analog, digital, 3G)" respectively. Handset Web Capability also have three levels; from low to high are “no web capability”, “web capability”, and “web capable mini-browser” respectively. The lowest levels in both subgroups have the most churners (*M=.55; M=.58*) respectively. The highest level in both sub-group have the fewest churners (*M=.43; M=.46*) respectively. So customers using higher capabilities are less likely to churn. Considering the web capability factor as a whole, customers using neither of the capabilities have the most churners (*M=.62*); customers using either of the highest capabilities have the lowest churners (*M=.43*).

In the profile CSAM group, one way ANOVA was conducted to examine the impact of total service months (see factor 4, Table 5-6) on churn. Although there are three predictors in this factor, only the service months contribute totally to this factor. Thus the service months were viewed representative of this factor. Similar as the operation on the handset price predictor, the service months were split into 7 subgroups: less than half year; between half year and one year; between one and one and half year; between one and half and two years; between two and three years; between three and four years, and more than four years. The difference between these subgroups are significant (*F(6,50050) = 17.367, p < 0.01*). Thus total service months have impact on churn. Five contrasts were conducted to test the difference between subgroups. Subgroup “less than half year” is significantly different from “between half year and one year” (*t(739) = -3.93, p < 0.01 (2-tailed)*), “between one and one and half year” (*t(758) = -6.52, p < 0.01 (2-tailed)*); but not significantly from “more than four years” (*t(1059) = -1.69, p = 0.091 (2-tailed)*). Subgroup “between half year and one year” is significantly different from subgroup “between one and one and half year” (*t(25997) = -8.37, p < 0.01 (2-tailed)*). Differences are not significantly large enough between the four customer profile subgroups: between one and one and half year; between one and half and two years; between two and three years; between three and four years (*t (6248) = 2.374, p < 0.025 (2- tailed)*). Customers in the company less than half year are least likely to churn (*M=.4, SD=.49*); customers between half and one year have a higher chance to churn (*M=.47, SD=.50*); customers between one and one and half year have highest chance to churn (*M=.52, SD=.50*); there is no significant difference among customers who used the service from one to four years; customers who used services for more than 4 years are less likely to churn (*M=.44, SD=.50*).

Table 5-7 Results of Drill-down Analysis

	CSAM Groups		
	Product	Customer Solution	Customer Profile
Factor	Price	Web capability	Tenure
Independent variable and subgroups	Handset Price - less than $59.99 - between $59.99 and 159.99 - more than 159.99	DualBand - no dualband - dualband - tri-model Handset Web Capability - no web capability - web capability - web capable mini-browser	Total Number of Months in Service - less than half year - between half year and one year - between one and one and half year - between one and half and two years - between two and three years - between three and four years - more than four years
Statistics	*F(2, 49671) = 257.811* *p <.01*	*F(2, 49966) = 8.63, p <.01* *F(2, 49966) = 143.907, p<.01* *F(2, 49966) = 12.872, p <.01*	*F(6,50050) = 17.367, p <.01*

5.7 Results of Structural Equation Modeling

Confirmatory Factor Analysis (CFA) is conducted with LISREL 8.51 to estimate the unknown parameters including: factor loadings and interfactor correlations among the 3 main factors and the predictors in the refined model.

Model Specification and Identification

The specification of the refined model (Figure 5-13) can be illustrated by a path diagram illustrated in Figure 5-14. This path diagram shows that we want to estimate how well the 3 concept/main factors are intercorrelated. As we discussed in Chapter 4, an intercorrelation of less than .6 can be interpreted as the two corresponding concepts are distinct. Moreover, the path loadings for each main factor can be interpreted as how each predictor contributes to the main factor. We also need to test the convergent validity by path coefficient, t-test, and standard error as discussed in Chapter 4.

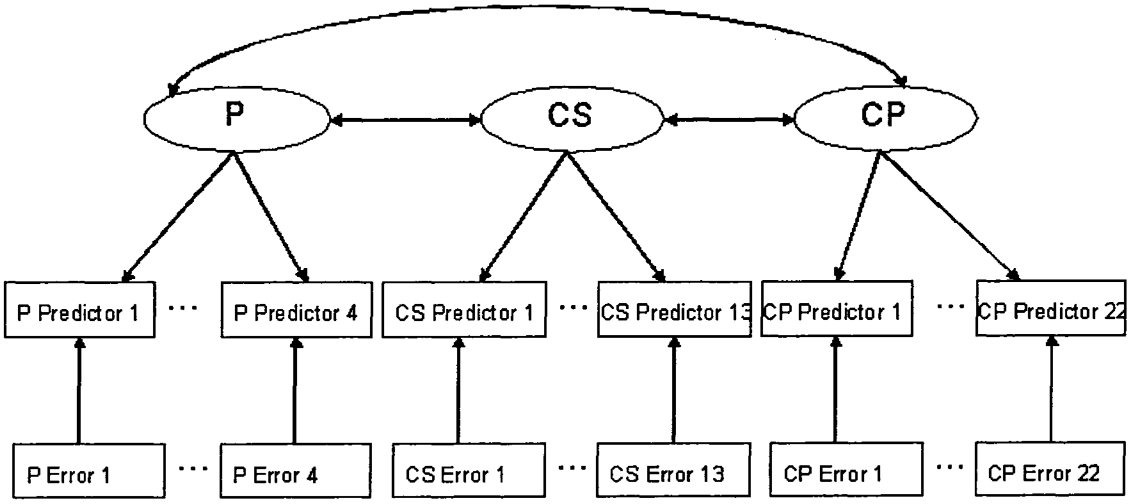


Figure 5-14 The Specification of Refined Model Illustrated by a Path Diagram

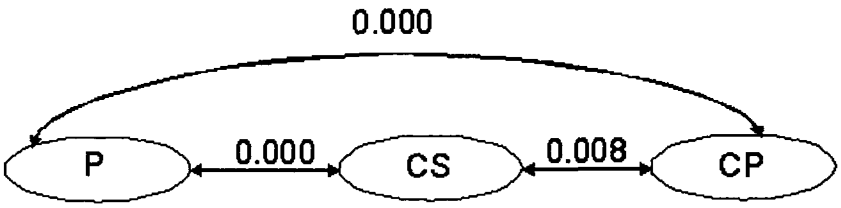
In model specification, we translate the path diagram to LISREL commands. There are 4 matrices of importance:

- Covariance matrix of observed variables. It is the input for CFA experiment.
- LX (λ) matrix. It is used to represent factor loadings. The Lambda matrix has as many rows as observed variables, and as many columns as latent variables. In Figure 5-14, all factor loadings from the latent variables to the observed variables are represented by LX matrix.
- PH (φ) matrix. It is used to represent interfactor correlations. The Phi matrix is a symmetric square matrix, with each factor mapped for a row and a column. In Figure 5-14, all interfactor correlations between the latent variables are represented by the PH matrix.
- TD (θ_δ) matrix. It is used to represent the unique factors. The Theta Epsilon matrix is a vector with as many columns as observed variables. In Figure 5-14, all errors to the observed variables are represented by the TD matrix.

Estimation and Testing Fit

The CFA result shows that the model is not a good fit for this given data set (GFI=.427, NFI=.442, AGFI=.360, CFI=.442, RMSEA=.194). The interfactor correlation and factor loadings are as follows:

- Interfactor correlations are



– Factor loadings are

```
LAMBDA-X

          Observed
          variables     P            CS           CP
        --------   --------     --------     --------
          var1      0.429          - -          - -
          var2      1.745          - -          - -
          var3      0.456          - -          - -
          var4      0.002          - -          - -
          var5        - -        0.027          - -
          var6        - -        0.635          - -
          var7        - -        0.095          - -
          var8        - -        0.756          - -
          var9        - -        0.214          - -
         var10        - -        0.956          - -
         var11        - -        0.635          - -
         var12        - -        0.647          - -
         var13        - -        0.030          - -
         var14        - -        1.008          - -
         var15        - -        0.033          - -
         var16        - -        0.009          - -
         var17        - -        0.169          - -
         var18        - -          - -        1.000
         var19        - -          - -        1.000
         var20        - -          - -        0.028
         var21        - -          - -        0.010
         var22        - -          - -        0.025
         var23        - -          - -        0.033
         var24        - -          - -       -0.026
         var25        - -          - -        0.061
         var26        - -          - -        0.009
         var27        - -          - -        0.020
         var28        - -          - -        0.059
         var29        - -          - -        0.003
         var30        - -          - -       -0.026
         var31        - -          - -        0.029
         var32        - -          - -       -0.013
         var33        - -          - -        0.019
         var34        - -          - -        0.742
         var35        - -          - -        0.699
         var36        - -          - -        0.959
         var37        - -          - -        0.007
         var38        - -          - -        0.007
```

```
var39          - -          - -       -0.031
```

The interfactor correlations are good (0.000, 0.000, and 0.008). The factor loadings for each factor have a big range (0.002~1.745, 0.009~1.008, 0.003~1.00). We can modify the model by deleting some paths from each latent variable to some observed variables to improve fit.

Model Re-specification

We modify the model by deleting some paths from the latent variables to the observed variables. Since we use sub-factors to explain the observed variables, we intend to delete those observed variables that have a lower loading (<.7) on sub-factors. That is, we employ CFA to test the contribution of the sub-factors to the main factors. All sub-factors are treated as observed variables. Since we can learn how each sub-factor contribute to the variance in the corresponding main factor from the results of PCA, we can ignore the sub factor with the smallest loading before we build the model. Then we specify the model with main factors and sub-factors as shown in Figure 5-15.

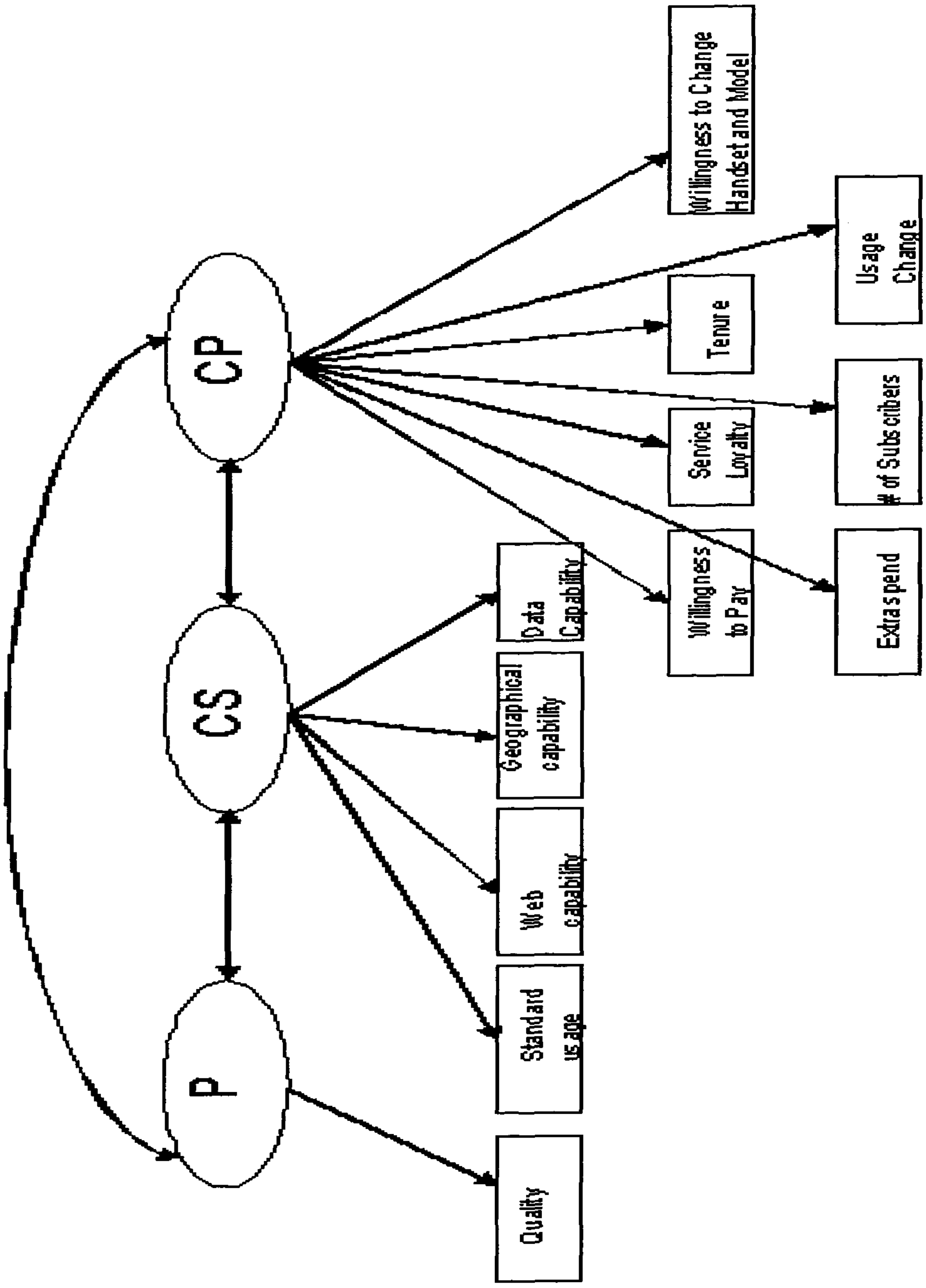


Figure 5-15 A CFA Model with Main Factors and Sub Factors

The CFA result showed that this model has 2 good fit indices (*GFI>.9, AGFI>0.8, NFI=.50, CFI=.50, RMSEA=.098*). For each main factor, only the first factor's loading is greater than .7 (1.02, .9, .8). Thus we leave all observed variables in the first sub-factor for each main factor. A modified model is illustrated in Figure 5-16. The CFA result (Figure 5-17) showed that all path t values (79.29~307.28) are significant (*p<.05*); the interfactor correlations are less than .6 (.00, .00, .51); each path loading (.10~67.05) was greater than twice its standard error (.00~.35); 8 out of 14 paths have more than .7 loadings, the other 6 are also near .7 (.44, .47, .63, .63, .65, .68). Thus the discriminate validity was demonstrated by the data, but convergent validity was not demonstrated by the data. This modified model is still not a good fit (*GFI=.73, AGFI=.62, NFI=.86, CFI=.86, RMSEA=.19*).

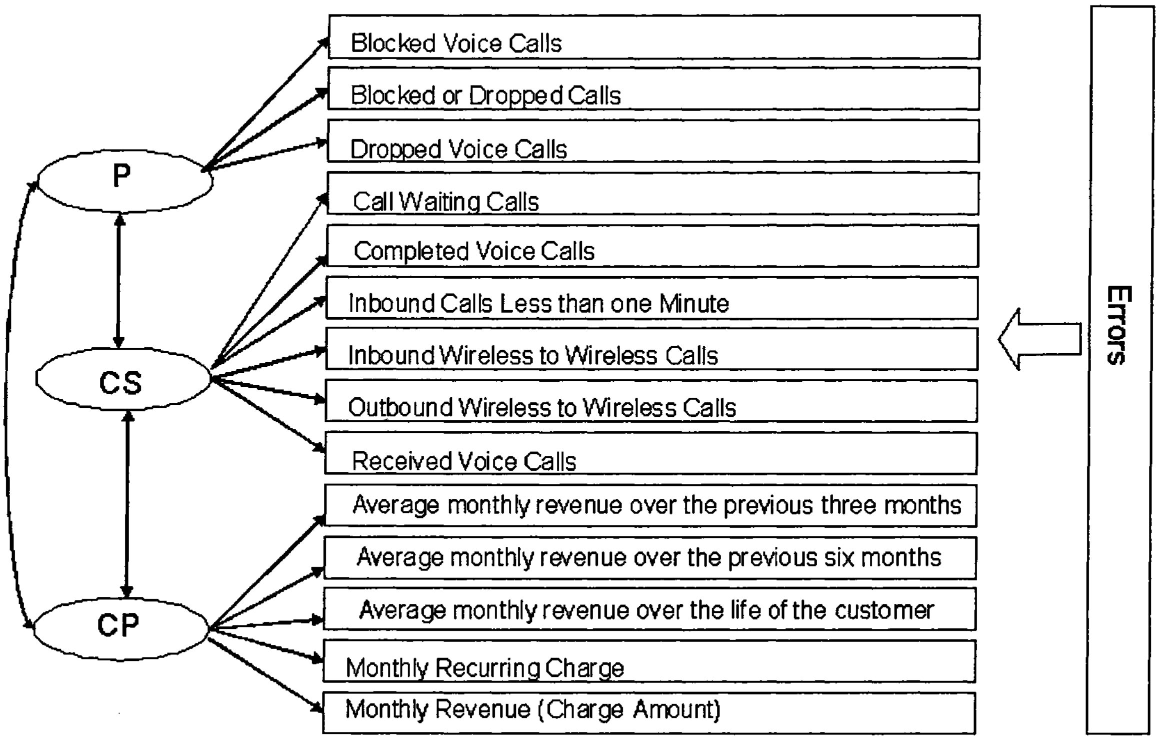

Blocked Voice Calls
Blocked or Dropped Calls
Dropped Voice Calls
Call Waiting Calls
Completed Voice Calls
Inbound Calls Less than one Minute
Inbound Wireless to Wireless Calls
Outbound Wireless to Wireless Calls
Received Voice Calls
Average monthly revenue over the previous three months
Average monthly revenue over the previous six months
Average monthly revenue over the life of the customer
Monthly Recurring Charge
Monthly Revenue (Charge Amount)
P
CS
CP
Errors

Figure 5-16 Modified Model for CFA

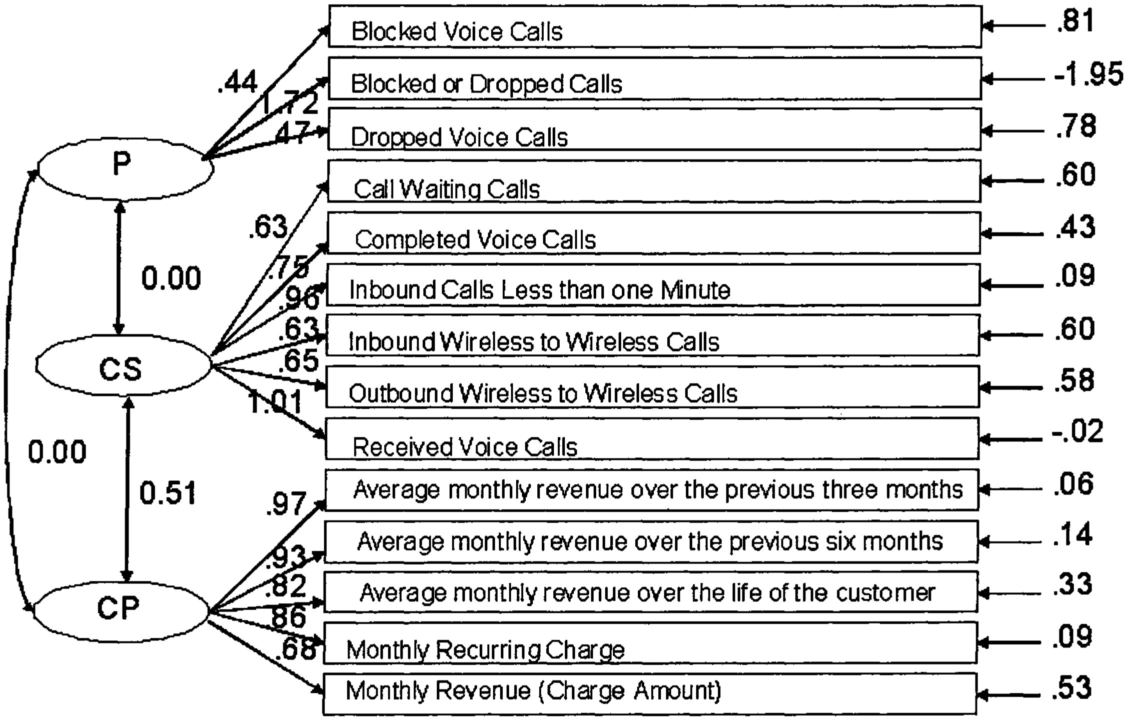

Blocked Voice Calls
Blocked or Dropped Calls
Dropped Voice Calls
Call Waiting Calls
Completed Voice Calls
Inbound Calls Less than one Minute
Inbound Wireless to Wireless Calls
Outbound Wireless to Wireless Calls
Received Voice Calls
Average monthly revenue over the previous three months
Average monthly revenue over the previous six months
Average monthly revenue over the life of the customer
Monthly Recurring Charge
Monthly Revenue (Charge Amount)
P
CS
CP
.44
1.72
.47
.63
.75
.96
.63
.65
1.01
.97
.93
.82
.86
.68
0.00
0.00
0.51
.81
-1.95
.78
.60
.43
.09
.60
.58
-.02
.06
.14
.33
.09
.53

Figure 5-17 Result of Modified Model

Although the modified model is not as good as expected, we can learn from the analysis exercise. The sub-factors: quality, standard usage, and willingness to pay contribute the most for the product, customer solution and customer profile strategies respectively. This result is easily understood by managers, and it also provides some guides for their strategy decisions. We learn that customers consider more about quality when they want a mobile service/product. Diverse capabilities of the product/service may attract customers; however, the frequently used standard usage is still the primary requirement.

When understanding a specific customer, may be the first question the mangers need to know is how much he/she is willing to pay. We can also do a drill-down analysis using the 5 predictors in our given data set which are related to the willingness to pay factor. For each predictor, we group the customers into different pay levels based on their payments using a difference of 10 dollars. For example, the range of the predictor "Average monthly revenue over the life of the customer" is from .54 to 201.66, then the customers are classified into 20 groups (lowest-10, 10-20, ..., 190-highest). The mean of churn indicator for each group of the 5 predictors are showed in Figure 5-18 (a-e).

We found that the first three groups of the mean of churn indicator within each predictor (see Figure 5-18 a-e) are more than .5, from the 4th group, the mean of churn indicator gradually moved lower or near .5, and the last several groups usually have a largely movement about the mean of the churn indicator, except for "Monthly Recurring Charge" in which the mean decreases. Here, we also use low-level, middle-level, and high-level to represent the payment levels. Based on this observation, we group the customer as shown in Table 5-8.

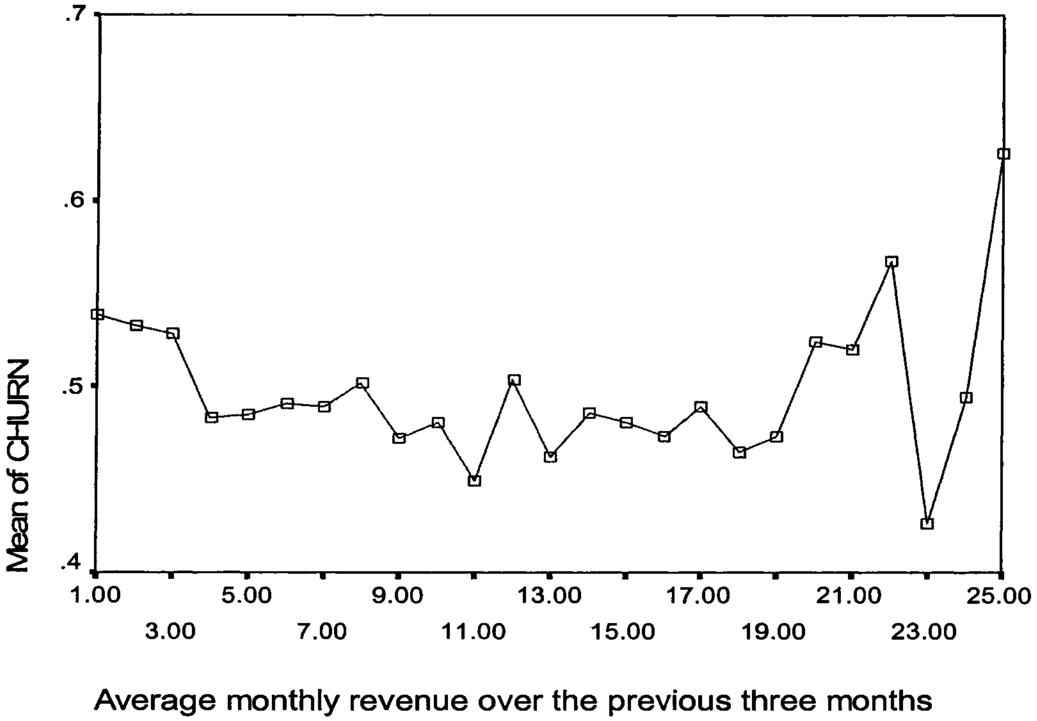


Figure 5-18 (a)

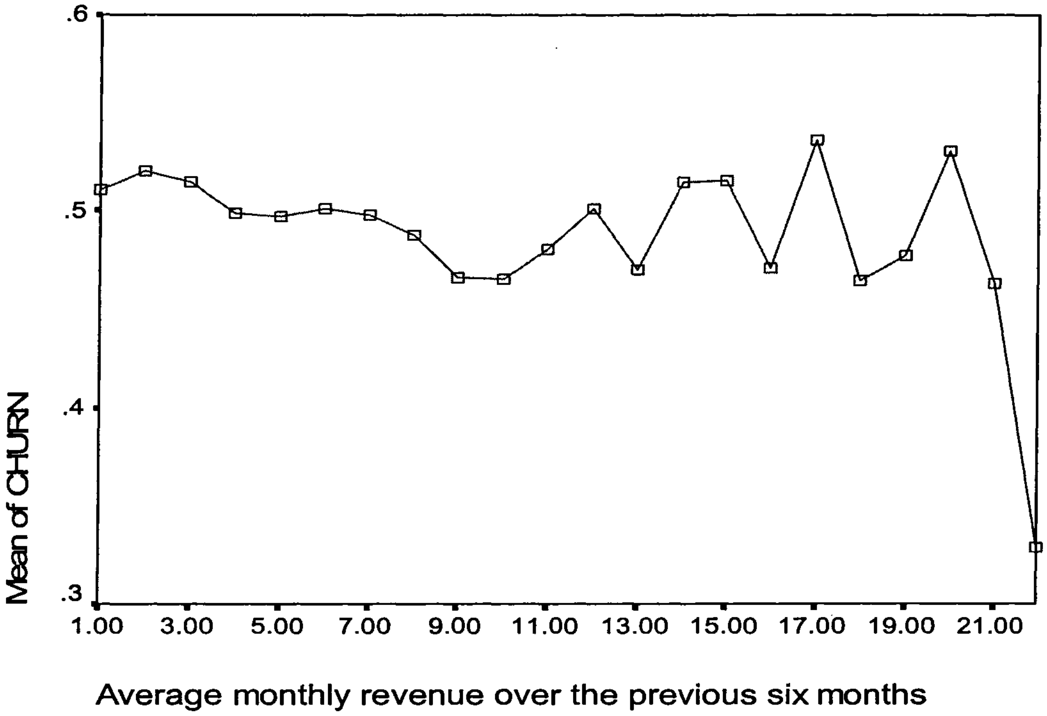


Figure 5-18 (b)

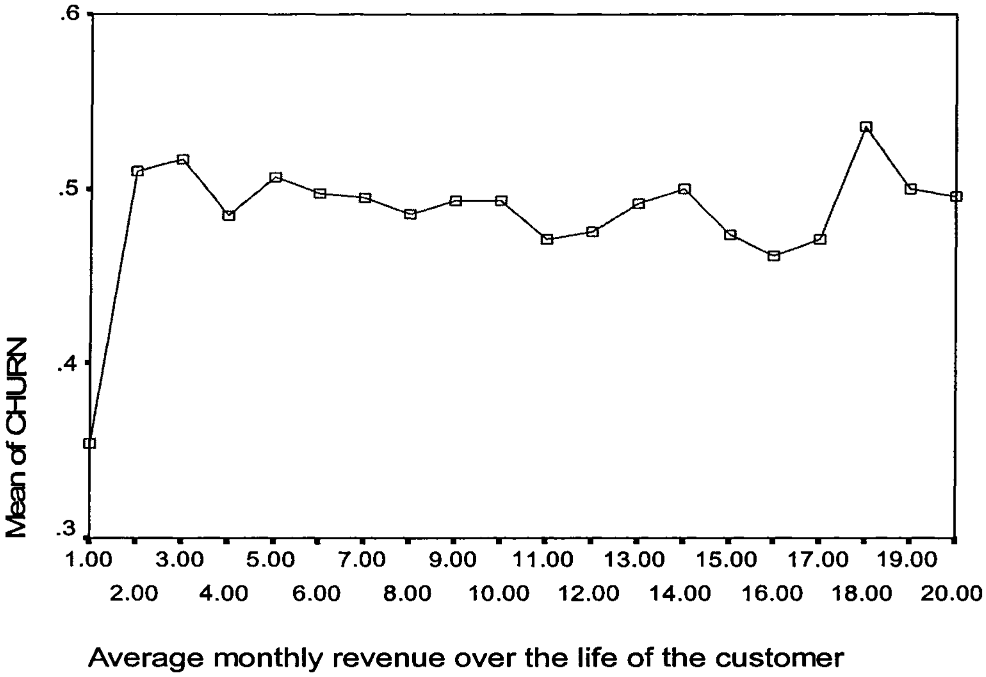


Figure 5-18 (c)

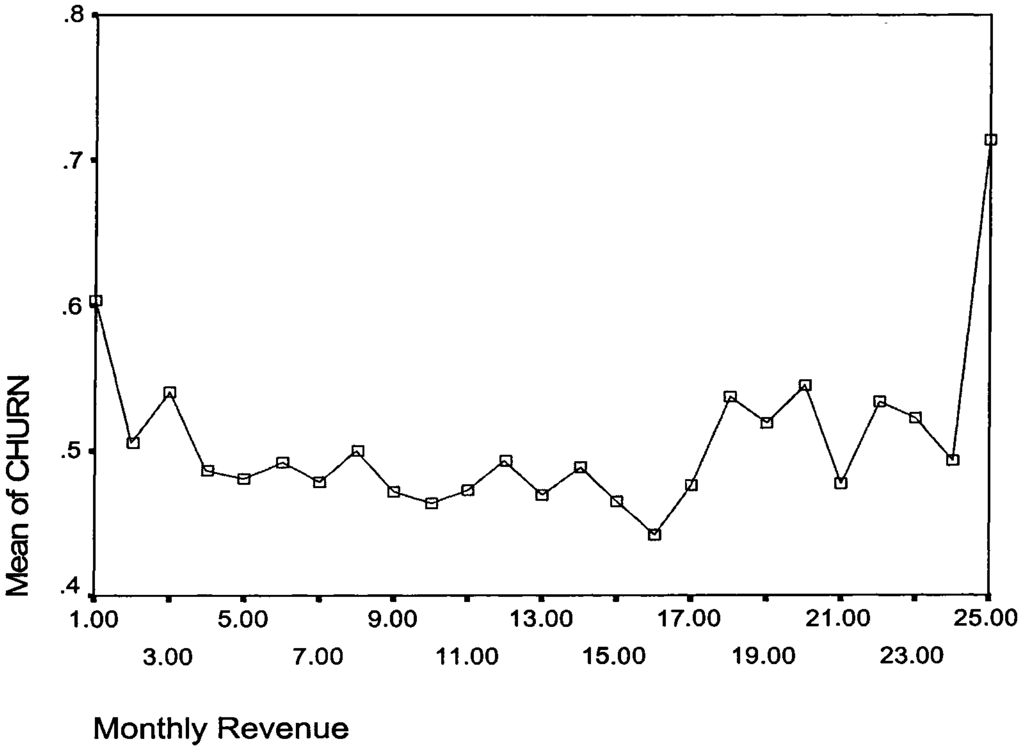


Figure 5-18 (d)

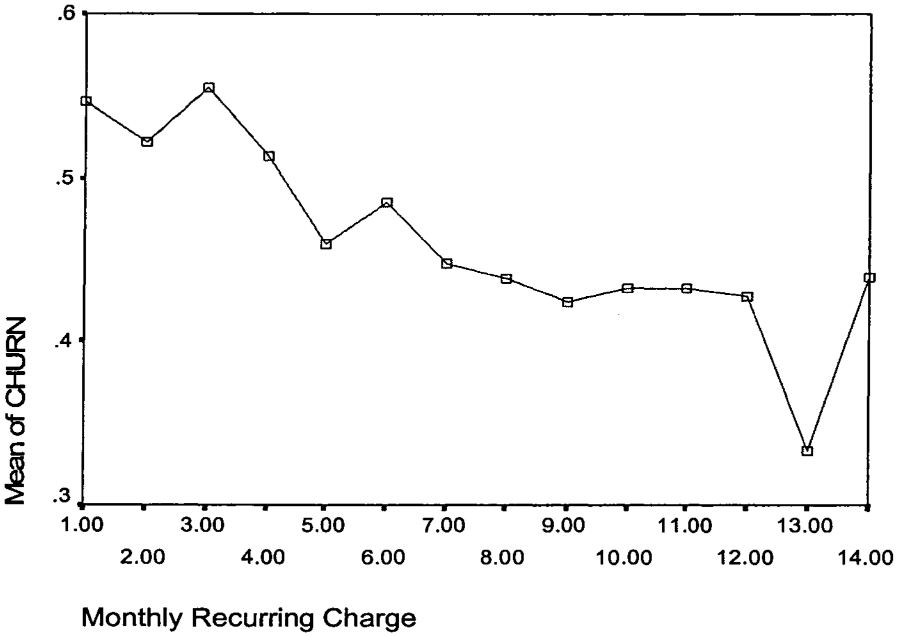


Figure 5-18 (e)

Figure 5-18 Mean Plots of each Predictors Related to "Willingness to pay" (a-e)

Table 5-8 Classification of customer for factor "willingness to pay"

Predictors	Low-level	Middle-level	High-level
Average monthly revenue over the previous three months	<=$30	$30~$190	>$190
Average monthly revenue over the previous six months	<=$30	$30~$130	>$130
Average monthly revenue over the life of the customer	<=$30	$30~$170	>$170
Monthly Revenue	<=$30	$30~$170	>$170
Monthly Recurring Charge	<=$30	>$30	

We test the difference of the mean of the churn indicator between any two groups for each predictor. The results showed that there is a significant difference between all the low-levels and the middle-levels except for the predictor "Average monthly revenue over the life of the customer". But there is no significant difference between all middle-levels and high-levels; and between all low-levels and middle-levels. These results showed that the telecommunication company may invest significant resources to reduce the churn rate of the people who are choosing the lowest margin, lowest cost products.

5.8 Summary

This chapter represented a case of CSAM validation and refinement. The exploratory factor analysis results showed that for the given data set, statistically correlated predictors can be grouped under product, customer solution and customer profile strategies. Moreover, interpretable high-level sub-factors can be extracted from the predictors to explain the corresponding churn strategy. The drill down analysis provided some applications of the refined model, and showed how managers can use the refined model to analyze churn. Confirmatory factor analysis showed that the model is not a good fit for our given data set. However, the estimates of the interfactor correlations and the factor loadings also provide new knowledge of the data. The interfactor correlations demonstrated that the 3 concepts in our conceptual model are distinct, they may reflect different facets of customer strategies. The factor loading showed a visual representation of the relationship between main churn factors and sub-factors, and main churn factors and predictors. This "not good fit" result also provides us new directions for our future work. For example, we are interested in how to refine the model for a given data set, and which variables or what kind of data should be collected for a specific churn analysis.

Chapter 6 Summary

6.1 Summary

In this thesis, we introduce a churn-strategy alignment model which links factors that cause customer churn with competitiveness strategies in the mobile telecommunications industry. We validate and refine the model by applying the model to a large industrial data set. With exploratory factor analysis, a refined model was obtained. The refined model links single churn predictors with organizational competitiveness strategies using high-level patterns or factors as a bridge to explain the individual low-level predictors. Thus, it is easier for managers to understand related informative high-level factors about churn rather than individual low-level predictors. CSAM enables business managers to more easily create strategically aligned competitiveness strategies that contribute to reducing churn.

We employ structural equation modeling to further validate the churn related strategic alignment model. Although the model did not fit very well to the given data set, we have some findings about both the conceptual model and the data set. The outcome of the structural equation modeling reveals that the 3 apices of CSAM are distinct, they can be viewed to reflect different competitive strategies. It also discloses that "quality", "standard usage", and "willingness to pay" are the important factors of the strategies related to product, customer solution, and customer profile respectively.

The CSAM provides a framework of components consisting of grouped churn predictors that allow managers to align churn reduction with these high-level patterns or factors, and in turn the higher level product, customer solution and customer profile strategies or the ways in which the firm competes. For example, managers may choose to simultaneously or step-wise focus on service loyalty strategies for churn reduction, and/or managers may focus the marketing of new products to those "early adopter" customers who are more willing to change handset and models of mobile service. We also illustrate how drill-down analyses on the complex high-level factors can further inform retention strategies. In reality, CSAM is particularly applicable to small and medium sized telecommunication companies as they are not of a size to employ system lock-in strategies.

We speculate that the model is robust in that as telecom products and services, customer solutions, and customer profiles change, the model can be easily extended while maintaining the current apex groups. For example, we have seen added capability on mobile handsets this year - camera phones are a standard industry offering in 2004. This capability would be added to the customer solution grouping on the CSAM. Customer profile is likely to be the most static CSAM group in terms of factor and variable composition.

6.2 Highlights of Findings

The following are the highlights of the findings in this research:

- The model introduces a new perspective on churn predictor analysis. Individual unrelated churn predictors can be linked to organizational competitive strategies, and they can be explained by high-level interpretable factors. These high-level factors and the higher level strategy concepts make the predictors easily understood, and enable managers to make decisions.

- Based on our given data set, "quality" and "price" are related to product strategies; "standard usage", "web capability", "data capability", "geographical capability", and "call forwarding capability" are related to customer solution strategies; "willingness to pay", "service loyalty", "extra spend", "tenure", "willingness to change handset or model", "number of subscribers", "usage change", and "resident location" are related to customer profile.
- The 3 apices in our proposed CSAM model are distinct. The strategies related to the 3 apices focus on product/service, customer solution, and customer profile respectively.
- There are visual representation between product strategy and "quality", customer solution strategy and "standard usage", and customer profile strategy and "willingness to pay". These three high-level factors are important when managers consider the customer strategies toward reducing churn.

6.3 Future Works

It would be interesting to test the fit of the CSAM model to other data sets from the mobile telecommunication industry, as well as firms in other industries with fierce competition. It would also be interesting to test in future whether the CSAM model is applicable to strategies that support m-commerce churn reduction.

Due to the “not good fit” result from the confirmatory factor analysis, it would be informative to research on how to improve the CSAM model, and how to collect high quality data to better refine the model.

References

Au, W.; Chan, K.C.C., and Yao, X., 2003. A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction. *Evolutionary Computation, IEEE Transactions.* 7(6), 532–545.

Berry, M.J.A. and Linoff, G., 1997. *Data Mining Techniques for marketing, Sales, and Customer Support.* New York; Toronto: Wiley, 454p.

Bollen, K.A. and Long, J.S., 1993. *Testing Structural Equation Models.* Beverly Hills, CA:Sage

Booz Allen and Hamilton, 2001. Winning the Customer Churn Battle in the Wireless Industry. *Insights, Information Technology Group.* 1 (1), 1-12.

Breiman L., Friedman J.H., Olshen A., and Stone C.G., 1984. *Classification and Regression Tree.* Wadsworth International Group, Belmont, CA, USA.

Brown, S.A., 2000. *Customer relationship management : a strategic imperative in the world of e-business.* Toronto ; New York : John Wiley & Sons Canada, 545p.

Cios, K.J., Pedrycz, W., and Swiniarski, R.W., 1998. *Data mining methods for knowledge discovery.* Boston : Kluwer Academic, 495p

Craig, J. and Julta, D., 2001. *e-Business readiness : a customer-focused framework.* Boston : Addison-Wesley, 437 p.

Duke Teradata, 2005. Data Description. *News and Events.* Retrieved [July 29, 2005] from http://www.teradataduke.org/ApplicationFiles/web/WebFrame.cfm?web_id=29

Duke Teradata, 2005. Industry Background. *News and Events*. Retrieved [July 29, 2005] from http://www.teradataduke.org/ApplicationFiles/web/WebFrame.cfm?web_id=29

Easton and McColl, 2005. Statistics Glossary. Retrieved [July 29,2005] from http://www.cas.lancs.ac.uk/glossary_v1.1/samp.html#

Federal Communications Commission, 2004. Retrieved [Dec 12, 2004] from http://www.fcc.gov/cgb/NumberPortability/

Garson, G.D., 2005. Factor Analysis. Retrieved [July 29,2005] from http://www2.chass.ncsu.edu/garson/pa765/factor.htm

Goodwin, R., 2004. Customer Retention Strategies - Keep the Customer Churn to a minimum. Retrieved [May 4, 2005] from http://www.crm2day.com/library/EEEFulpFEurEFbBiqQ.php

Groth, R., 1998. *Data Mining: A Hands-on Approach for Business Professionals*. Upper Saddle River, N.J.: Prentice Hall , 264p.

Gupta, S.; Kamakura, W.; Lu, J.; Mason, C., and Nelin, S., 2003. Churn Modeling Tournament. *CRM PRESENTION, INFORMS Marketing Science Conference, Maryland, June 2003* (revised version) Retrieved [Dec 20, 2004] from http://www.teradataduke.org/ApplicationFiles/web/WebWYSIWYGPage.cfm?web_page_id=82

Han, J. and Kamber, M., 2001. *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann, 550p.

Hand, D.J., 1999. Statistics and Data Mining: Intersecting Disciplines. *ACM SIGKDD Explorations Newsletter.* 1(1), 16-19

Hax and Wilde II, 2003. The Delta Model - A new Framework of Strategy. *Journal of Strategic Management Education*, 1 (1), 1-21

Howell, D.C., 2002. Statistical Methods for Psychology fifth edition. Thomson Learning, 802p.

Howell, J., Miller, P., Park, H.H., Sattler, D., Schack, T., Spery, E., Widhalm, S. and Palmquist, M., 2005. Reliability and Validity. *Writing@CSU. Colorado State University Department of English.* Retrieved [July 29, 2005] from http://writing.colostate.edu/references/research/relval/.

Johnson, D.E., 1998. *Applied multivariate methods for data analysts.* Pacific Grove, Calif. : Duxbury Press, 567 p

Kantardzic, M., 2003. *Data Mining: Concepts, Models, Methods, and Algorithms.* Hoboken, NJ: Wiley-Interscience: IEEE Press, 345p

Kelloway, E.K., 1998. *Using LISREL for Structure Equation Modeling.* Sage Publication, 147p

Kline, R.B., 1998. *Principles and practice of structural equation modeling.* New York : Guilford Press, 354 p

Kudyba, S. and Hoptroff, R., 2001. *Data Mining and Business Intelligence: A Guide to Productivity.* Hershey, PA, USA: Idea Group Publishing, 166p

Lu, J., 2002. Predicting customer churn in the telecommunications industry – An application of survival analysis modeling using SAS®. *SAS Group International 27th Annual Conference*, paper 114.

Management Sciences for Health, 2003. Business Planning to Transform Your Organization. *The Manager (Boston)*, 12 (3), 1 –30.

Marakas, G.M., 2003. *Decision Support Systems in the 21st Century (2nd Edition).* Pearson Education, Inc, 611p

Mcknight, D.H., Choudhury, V. and Kacmar, C., 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research* , 13(3), 334-359

Mozer, M.C.; Wolniewicz, R.; Grimes, D.B.; Johson, E., and Kashansky, H., 2000. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *Neural Networks, IEEE Transactions,* 11(3), 690-696.

Parekkat, A., 2003. Customer Churn - An operation comparison of logistic regression, decision tree &neural network models. VIEWS Conference, 2003

Rosset, S., Neumann, E., Eric, U., Vatnik, N., and Idan, Y. 2002. *Customer Lifetime Value Modeling and Its Use for Customer Retention Planning.* SIGKDD "02, July 23-26, 2002, Edmonton, Alberta, Canada,

Rosset, S. and Neumann, E., 2003. Integrating Customer Value Considerations into Predictive Modeling. *Third IEEE International Conference on Data Mining.*

Scott Cardell, N.; Golovnya, M. and Steinberg, D., 2003. Churn Modeling for Mobile Telecommunications: Winning the Duke/NCR Teradata Center for CRM Competition. *CRM PRESENTION, INFORMS Marketing Science Conference, Maryland, June 2003*

SPSS, 2001. SPSS for Windows 11.5 Help.

Statsoft, 2003. Retrieved [July 29, 2005] from http://www.statsoft.com/textbook/stsepath.html

Surfgold, 2003. Perils of retention rate. *Articles, Idea center.* Retrieved [May 18, 2004] from http://www.surfgold.co.kr/surfgold/readarticle.asp?article=Perils%20of%20retention%20rate.htm

Swift, R.S., 2001. *Accelerating customer relationships : using CRM and relationship technologies.* Upper Saddle River, NJ : Prentice Hall PTR, 480 p.

Thearling, K., 1999. Data Mining and CRM: Zeroing in on Your Best Customers. *DMReview.com*. Retrieved [Feb 7, 2004] from http://www.dmreview.com/master.cfm?NavID=198&EdID=1744

Witten, I.H. and Frank, E., 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufmann, 371p.

Yan, L.; MILLER, D.J.; MOZER, M.C., and WOLNIEWICZ, R., 2001. Improving prediction of customer behavior in nonstationary environments. *Proceedings of the International Joint Conference on Neural Networks.* 3, 2258-2263.

Appendix A: SPSS Syntax of Data Preprocessing

```
*************data preprocessing of product group ******************************************

DESCRIPTIVES
  VARIABLES=blck_dat blck_vce drop_blk drop_dat drop_vce hnd_pric  /SAVE
  /STATISTICS=MEAN STDDEV MIN MAX .

DO IF (ABS(zblck_da) <= 4) .
RECODE
  blck_dat
  (MISSING=SYSMIS)  (ELSE=Copy)  INTO  new_bdat .
END IF .
EXECUTE .
DO IF (ABS(zblck_vc) <= 4) .
RECODE
  blck_vce
  (MISSING=SYSMIS)  (ELSE=Copy)  INTO  new_bvce .
END IF .
EXECUTE .
DO IF (ABS(zdrop_bl) <= 4) .
RECODE
  drop_blk
  (MISSING=SYSMIS)  (ELSE=Copy)  INTO  new_bd .
END IF .
EXECUTE .
DO IF (ABS(zdrop_da) <= 4) .
RECODE
  drop_dat
  (MISSING=SYSMIS)  (ELSE=Copy)  INTO  new_ddat .
END IF .
EXECUTE .
DO IF (ABS(zdrop_vc) <= 4) .
RECODE
  drop_vce
  (MISSING=SYSMIS)  (ELSE=Copy)  INTO  new_dvce .
END IF .
EXECUTE .
DO IF (ABS(ztotmrc_) <= 4) .

RECODE
  hnd_pric
  (MISSING=SYSMIS)  (Lowest thru 10=1)  (10 thru 30=2)  (30 thru 40=3)  (40
  thru 60=4)  (60 thru 80=5)  (80 thru 100=6)  (100 thru 120=7)  (120 thru
  130=8)  (130 thru 150=9)  (150 thru 160=10)  (160 thru 180=11)  (180 thru
  200=12)  (200 thru 240=13)  (240 thru 250=14)  (250 thru 300=15) (300 thru
400=16) (400 thru Highest=17) INTO  n2_HDPC .
EXECUTE .

****************data preprocessing of customer  solution group***************
DESCRIPTIVES
  VARIABLES=callfwdv callwait comp_dat comp_vce da_mean inonemin iwylis_v
  owylis_v recv_sms recv_vce roam_mea  /SAVE
  /STATISTICS=MEAN STDDEV MIN MAX .
```

```
DO IF (ABS(zcallfwd) <= 4) .
RECODE
 callfwdv
 (MISSING=SYSMIS)  (ELSE=Copy)  INTO  newforw .
END IF .
EXECUTE .

DO IF (ABS(zcallwai) <= 4) .
RECODE
 callwait
 (MISSING=SYSMIS)  (ELSE=Copy)  INTO  newwait .
END IF .
EXECUTE .

DO IF (ABS(zcomp_da) <= 4) .
RECODE
 comp_dat
 (MISSING=SYSMIS)  (ELSE=Copy)  INTO  newcdat .
END IF .
EXECUTE .

DO IF (ABS(zcomp_vc) <= 4) .
RECODE
 comp_vce
 (MISSING=SYSMIS)  (ELSE=Copy)  INTO  newcvce .
END IF .
EXECUTE .

DO IF (ABS(zda_mean) <= 4) .
RECODE
 da_mean
 (MISSING=SYSMIS)  (ELSE=Copy)  INTO  newdirec .
END IF .
EXECUTE .

DO IF (ABS(zinonemi) <= 4) .
RECODE
 inonemin
 (MISSING=SYSMIS)  (ELSE=Copy)  INTO  newone .
END IF .
EXECUTE .

DO IF (ABS(ziwylis_) <= 4) .
RECODE
 iwylis_v
 (MISSING=SYSMIS)  (ELSE=Copy)  INTO  newiwtw .
END IF .
EXECUTE .

DO IF (ABS(zowylis_) <= 4) .
RECODE
 owylis_v
 (MISSING=SYSMIS)  (ELSE=Copy)  INTO  newowtw .
END IF .
EXECUTE .
```

```
DO IF (ABS(zrecv_sm) <= 4) .
RECODE
 recv_sms
 (MISSING=SYSMIS)  (ELSE=Copy)  INTO  newsms .
END IF .
EXECUTE .

DO IF (ABS(zrecv_vc) <= 4) .
RECODE
 recv_vce
 (MISSING=SYSMIS)  (ELSE=Copy)  INTO  newrevce .
END IF .
EXECUTE .

DO IF (ABS(zroam_me) <= 4) .
RECODE
 roam_mea
 (MISSING=SYSMIS)  (ELSE=Copy)  INTO  newroam .
END IF .
EXECUTE .

RECODE
 dualband
 ('U'=SYSMIS)  ('T'=1)  ('Y'=0.5)  ('N'=0)  INTO  RE_DUBN .
EXECUTE .
RECODE
 hnd_webc
 ('UNKW'=SYSMIS)  ('WCMB'=1)  ('WC'=0.5)  ('NA'=0)  INTO  RE_WEBC .
EXECUTE .

DESCRIPTIVES
 VARIABLES=newforw newwait newcdat newcvce newdirec newone newiwtw
newowtw
 newsms newrevce newroam re_dubn re_webc
 /STATISTICS=MEAN STDDEV MIN MAX .


*****************data preprocessing of customer profile group*************

RECODE
 asl_flag
 (MISSING=SYSMIS)  ('N'=0)  ('Y'=1)  INTO  aslflag2 .
EXECUTE .

RECODE
 crclscod
 (MISSING=SYSMIS)  ('ZY'=1)  ('ZF'=2)  ('ZA'=3)  ('Z5'=4)  ('Z4'=5)
 ('Z2'=6)  ('Z1'=7)  ('Z'=8)  ('Y'=9)  ('W'=10)  ('V1'=11)  ('U1'=12)
 ('U'=13)  ('TP'=14)  ('S'=15)  ('P1'=16)  ('O'=17)  ('L'=18)  ('M'=19)
 ('K'=20)  ('JF'=21)  ('J'=22)  ('IF'=23)  ('I'=24)  ('H'=25)  ('GY'=26)
 ('GA'=27)  ('G'=28)  ('EM'=29)  ('EF'=30)  ('EC'=31)  ('EA'=32)  ('E4'=33)
 ('E2'=34)  ('E'=35)  ('DA'=36)  ('D5'=37)  ('D4'=38)  ('D2'=39)  ('D'=40)
 ('CY'=41)  ('CC'=42)  ('CA'=43)  ('C5'=44)  ('C2'=45)  ('C'=46)  ('BA'=47)
 ('B2'=48)  ('B'=49)  ('AA'=50)  ('A3'=51)  ('A2'=52)  ('A'=53)  INTO
 crecode .
EXECUTE .
```

```
RECODE
  ethnic
  ('B'=1)  ('C'=2)  ('D'=3)  ('F'=4)  ('G'=5)  ('H'=6)  ('I'=7)  ('J'=8)
  ('M'=9)  ('N'=10)  ('O'=11)  ('P'=12)  ('R'=13)  ('S'=14)  ('U'=15)
  ('X'=16)  ('Z'=17)  INTO  RE_ETHN .
EXECUTE .

RECODE
  area
  (MISSING=SYSMIS)  ('ATLANTIC SOUTH AREA'=1)  ('TENNESSEE AREA'=19)
  ('SOUTHWEST AREA'=18)  ('SOUTH FLORIDA AREA'=17)  ('PHILADELPHIA
AREA'=16)
  ('CENTRAL/SOUTH TEXAS AREA'=3)  ('CHICAGO AREA'=4)  ('DALLAS AREA'=5)
('DC'+
 '/MARYLAND/VIRGINIA AREA'=6)  ('GREAT LAKES AREA'=7)  ('HOUSTON
AREA'=8)
  ('LOS ANGELES AREA'=9)  ('MIDWEST AREA'=10)  ('NEW ENGLAND AREA'=11)
('NEW'+
 ' YORK CITY AREA'=12)  ('NORTH FLORIDA AREA'=13)  ('NORTHWEST/ROCKY
MOUNTAIN'+
 ' AREA'=14)  ('OHIO AREA'=15)  ('CALIFORNIA NORTH AREA'=2)  INTO
newarea .
EXECUTE .

RECODE
  refurb_n
  ('R'=0)  ('N'=1)  INTO  RE_FUR .
EXECUTE .

RECODE
  retdays
  (MISSING=0)  (ELSE=Copy)  INTO  retday .
EXECUTE .

RECODE
  creditcd
  (MISSING=SYSMIS)  ('N'=0)  ('Y'=1)  INTO  cred# .
EXECUTE .


RECODE
  ref_qty
  (MISSING=0)  (Lowest thru 2=Copy)  (ELSE=3)  INTO  referal .
EXECUTE .

RECODE
  prizm_so
  ('C'=5)  ('U'=4)  ('S'=3)  ('T'=2)  ('R'=1)  (ELSE=SYSMIS)  INTO  socigrp .
EXECUTE .

RECODE
  tot_ret
  (MISSING=0)  (ELSE=1)  INTO  reten1 .
EXECUTE .
RECODE
```

```
  tot_ret
  (MISSING=0) (ELSE=Copy) INTO reten2 .
 EXECUTE .

 RECODE
  tot_acpt
  (MISSING=0) (ELSE=1) INTO accept1 .
 EXECUTE .
 RECODE
  tot_acpt
  (MISSING=0) (ELSE=Copy) INTO accept2 .
 EXECUTE .

 **************************************


 DESCRIPTIVES
  VARIABLES=aslflag2 actvsubs avg3rev avg6rev avgrev adjrev crecode re_ethn
 newarea
  RE_FUR rev_mean totmrc_m mtrcycle eqpdays retday models phones change_r
 cred#
  datovr_m vceovr_m referai rv socigrp reten1 reten2 months accept1 accept2
 ovrrev_m totrev
  truck uniqsubs
  /STATISTICS=MEAN STDDEV MIN MAX .


 DESCRIPTIVES
  VARIABLES=actvsubs avg3rev avg6rev avgrev adjrev crecode re_ethn newarea
  rev_mean totmrc_m eqpdays retday models phones change_r datovr_m vceovr_m
  socigrp months ovrrev_m totrev uniqsubs /SAVE
  /STATISTICS=MEAN STDDEV MIN MAX .


 DESCRIPTIVES
  VARIABLES=zactvsub zavg3rev zavg6rev zavgrev zadjrev zcrecode zre_ethn
  znewarea zrev_mea ztotmrc_ zeqpdays zretday zmodels zphones zchange_
  zdatovr_ zvceovr_ zsocigrp zmonths zovrrev_
  ztotrev zuniqsub
  /STATISTICS=MEAN STDDEV MIN MAX .

DO IF (ABS(zactvsub) <= 4) .
RECODE
 actvsubs
 (MISSING=SYSMIS) (ELSE=Copy) INTO nactsub .
END IF .
EXECUTE .

DO IF (ABS(zavg3rev) <= 4) .
RECODE
 avg3rev
 (MISSING=SYSMIS) (ELSE=Copy) INTO newavg3 .
END IF .
EXECUTE .
DO IF (ABS(zavg6rev) <= 4) .
RECODE
 avg6rev
 (MISSING=SYSMIS) (ELSE=Copy) INTO newavg6 .
END IF .
EXECUTE .
DO IF (ABS(zavgrev) <= 4) .
RECODE
 avgrev
 (MISSING=SYSMIS) (ELSE=Copy) INTO newavg .
END IF .
EXECUTE .

DO IF (ABS(zadjrev) <= 4) .
RECODE
 adjrev
 (MISSING=SYSMIS) (ELSE=Copy) INTO newadj .
END IF .
EXECUTE .

DO IF (ABS(zrev_mea) <= 4) .
RECODE
 rev_mean
 (MISSING=SYSMIS) (ELSE=Copy) INTO newrev .
END IF .
EXECUTE .

DO IF (ABS(ztotmrc_) <= 4) .
RECODE
 totmrc_m
 (MISSING=SYSMIS) (ELSE=Copy) INTO newtmrc .
END IF .
EXECUTE .

DO IF (ABS(zeqpdays) <= 4) .
RECODE
 eqpdays
 (MISSING=SYSMIS) (ELSE=Copy) INTO neweqpd .
END IF .
EXECUTE .

DO IF (ABS(zretday) <= 4) .
RECODE
 retday
 (MISSING=SYSMIS) (ELSE=Copy) INTO nretday .
END IF .
EXECUTE .

DO IF (ABS(zmodels) <= 4) .
RECODE
 models
 (MISSING=SYSMIS) (ELSE=Copy) INTO Nmodel .
END IF .
EXECUTE .

DO IF (ABS(zphones) <= 4) .
RECODE
 phones
 (MISSING=SYSMIS) (ELSE=Copy) INTO newphe .
```

```
END IF .
EXECUTE .

DO IF (ABS(zchange_) <= 4) .
RECODE
 change_r
 (MISSING=SYSMIS) (ELSE=Copy) INTO newchg .
END IF .
EXECUTE .

DO IF (ABS(zdatovr_) <= 4) .
RECODE
 datovr_m
 (MISSING=SYSMIS) (ELSE=Copy) INTO ndatovr .
END IF .
EXECUTE .

DO IF (ABS(zvceovr_) <= 4) .
RECODE
 vceovr_m
 (MISSING=SYSMIS) (ELSE=Copy) INTO nvceovr .
END IF .
EXECUTE .

DO IF (ABS(zmonths) <= 4) .
RECODE
 months
 (MISSING=SYSMIS) (ELSE=Copy) INTO newmth .
END IF .
EXECUTE .
DO IF (ABS(zovrrev_) <= 4) .
RECODE
 ovrrev_m
 (MISSING=SYSMIS) (ELSE=Copy) INTO newovr .
END IF .
EXECUTE .

DO IF (ABS(ztotrev) <= 4) .
RECODE
 totrev
 (MISSING=SYSMIS) (ELSE=Copy) INTO mtotrev .
END IF .
EXECUTE .


DO IF (ABS(zuniqsub) <= 4) .
RECODE
 uniqsubs
 (MISSING=SYSMIS) (ELSE=Copy) INTO newuni .
END IF .
EXECUTE .

DESCRIPTIVES
 VARIABLES=nactsub newavg3 newavg6 newavg newadj newrev newtmrc
neweqpd
 nretday nmodel newphe newchg ndatovr nvceovr newmth newovr mtotrev newuni
 /STATISTICS=MEAN STDDEV MIN MAX .

******************************************************


COMPUTE ren_asub = nactsub/4 .
EXECUTE .
COMPUTE ren_avg3 = (newavg3-1)/(243-1) .
EXECUTE .
COMPUTE ren_avg6 = (newavg6+2)/(220+2) .
EXECUTE .
COMPUTE ren_avg = (newavg-0.54)/(201.66-0.54) .
EXECUTE .

COMPUTE ren_adj = (newadj-2.7)/(4290.85-2.7) .
EXECUTE .
COMPUTE ren_rev = (newrev+6.17)/(246.62+6.17) .
EXECUTE .
COMPUTE ren_mrc = (newtmrc+6.17)/(140.48+6.17) .
EXECUTE .

COMPUTE ren_eqp = (neweqpd+5)/(1417+5) .
EXECUTE .
COMPUTE ren_rday = nretday/239 .
EXECUTE .
COMPUTE ren_mod = (nmodel-1)/(5-1) .
EXECUTE .
COMPUTE ren_phe = (newphe-1)/(6-1) .
EXECUTE .

COMPUTE ren_chg = (newchg+234.54)/(233.74+234.54) .
EXECUTE .
COMPUTE ren_dato = ndatovr/10.75 .
EXECUTE .
COMPUTE ren_vceo = nvceovr/129.94 .
EXECUTE .
COMPUTE ren_mth = (newmth-6)/(57-6) .
EXECUTE .

COMPUTE ren_ovr = newovr/131.5 .
EXECUTE .
COMPUTE ren_trev = (mtotrev-3.75)/(4410.41-3.75) .
EXECUTE .
COMPUTE ren_uni = (newuni-1)/(6-1) .
EXECUTE .

COMPUTE reclscod = (crecode-1)/(53-1) .
EXECUTE .
COMPUTE reethn = (re_ethn-1)/(17-1) .
EXECUTE .
COMPUTE REN_AREA = (newarea-1)/(19-1) .
EXECUTE .
COMPUTE rescgrp = (socigrp-1)/(5-1) .
EXECUTE .

COMPUTE ren_refl = referal/3 .
EXECUTE .
```

```
COMPUTE ren_ren2 = reten2/3 .
EXECUTE .
COMPUTE ren_acp2 = accept2/3 .
EXECUTE .
```

Appendix B: The List of 170 Variables in The Data Set

Interval Variables	Explanation
ADJMOU	Billing adjusted total minutes of use over the life of the customer
ADJQTY	Billing adjusted total number of calls over the life of the customer
ADJREV	Billing adjusted total revenue over the life of the customer
ATTEMPT_MEAN	Mean number of attempted calls
ATTEMPT_RANGE	Range of number of attempted calls
AVG3MOU	Average monthly minutes of use over the previous three months
AVG3QTY	Average monthly number of calls over the previous three months
AVG3REV	Average monthly revenue over the previous three months
AVG6MOU	Average monthly minutes of use over the previous six months
AVG6QTY	Average monthly number of calls over the previous six months
AVG6REV	Average monthly revenue over the previous six months
AVGMOU	Average monthly minutes of use over the life of the customer
AVGQTY	Average monthly number of calls over the life of the customer
AVGREV	Average monthly revenue over the life of the customer
BLCK_DAT_MEAN	Mean number of blocked (failed) data calls
BLCK_DAT_RANGE	Range of number of blocked (failed) data calls
BLCK_VCE_MEAN	Mean number of blocked (failed) voice calls
BLCK_VCE_RANGE	Range of number of blocked (failed) voice calls
CALLFWDV_MEAN	Mean number of call forwarding calls
CALLFWDV_RANGE	Range of number of call forwarding calls
CALLWAIT_MEAN	Mean number of call waiting calls
CALLWAIT_RANGE	Range of number of call waiting calls
CC_MOU_MEAN	Mean unrounded minutes of use of customer care (see CUSTCARE_MEAN) calls
CC_MOU_RANGE	Range of unrounded minutes of use of customer care calls
CCRNDMOU_MEAN	Mean rounded minutes of use of customer care calls
CCRNDMOU_RANGE	Range of rounded minutes of use of customer care calls
CHANGE_MOU	Percentage change in monthly minutes of use vs previous three month average
CHANGE_REV	Percentage change in monthly revenue vs previous three month average
COMP_DAT_MEAN	Mean number of completed data calls
COMP_DAT_RANGE	Range of number of completed data calls
COMP_VCE_MEAN	Mean number of completed voice calls
COMP_VCE_RANGE	Range of number of completed voice calls
COMPLETE_MEAN	Mean number of completed calls
COMPLETE_RANGE	Range of number of completed calls
CUSTCARE_MEAN	Mean number of customer care calls
CUSTCARE_RANGE	Range of number of customer care calls
DA_MEAN	Mean number of directory assisted calls
DA_RANGE	Range of number of directory assisted calls
DATOVR_MEAN	Mean revenue of data overage
DATOVR_RANGE	Range of revenue of data overage
DROP_BLK_MEAN	Mean number of dropped or blocked calls
DROP_BLK_RANGE	Range of number of dropped or blocked calls
DROP_DAT_MEAN	Mean number of dropped (failed) data calls
DROP_DAT_RANGE	Range of number of dropped (failed) data calls
DROP_VCE_MEAN	Mean number of dropped (failed) voice calls
DROP_VCE_RANGE	Range of number of dropped (failed) voice calls

EQPDAYS	Number of days (age) of current equipment
INONEMIN_MEAN	Mean number of inbound calls less than one minute
INONEMIN_RANGE	Range of number of inbound calls less than one minute
IWYLIS_VCE_MEAN	Mean number of inbound wireless to wireless voice calls
IWYLIS_VCE_RANGE	Range of number of inbound wireless to wireless voice calls
MONTHS	Total number of months in service
MOU_CDAT_MEAN	Mean unrounded minutes of use of completed data calls
MOU_CDAT_RANGE	Range of unrounded minutes of use of completed data calls
MOU_CVCE_MEAN	Mean unrounded minutes of use of completed voice calls
MOU_CVCE_RANGE	Range of unrounded minutes of use of completed voice calls
MOU_MEAN	Mean number of monthly minutes of use
MOU_OPKD_MEAN	Mean unrounded minutes of use of off-peak data calls
MOU_OPKD_RANGE	Range of unrounded minutes of use of off-peak data calls
MOU_OPKV_MEAN	Mean unrounded minutes of use of off-peak voice calls
MOU_OPKV_RANGE	Range of unrounded minutes of use of off-peak voice calls
MOU_PEAD_MEAN	Mean unrounded minutes of use of peak data calls
MOU_PEAD_RANGE	Range of unrounded minutes of use of peak data calls
MOU_PEAV_MEAN	Mean unrounded minutes of use of peak voice calls
MOU_PEAV_RANGE	Range of unrounded minutes of use of peak voice calls
MOU_RANGE	Range of number of minutes of use
MOU_RVCE_MEAN	Mean unrounded minutes of use of received voice calls
MOU_RVCE_RANGE	Range of unrounded minutes of use of received voice calls
MOUIWYLISV_MEAN	Mean unrounded minutes of use of inbound wireless to wireless voice calls
MOUIWYLISV_RANGE	Range of unrounded minutes of use of inbound wireless to wireless voice calls
MOUOWYLISV_MEAN	Mean unrounded minutes of use of outbound wireless to wireless voice calls
MOUOWYLISV_RANGE	Range of unrounded minutes of use of outbound wireless to wireless voice calls
OWYLIS_VCE_MEAN	Mean number of outbound wireless to wireless voice calls
OWYLIS_VCE_RANGE	Range of number of outbound wireless to wireless voice calls
OPK_DAT_MEAN	Mean number of off-peak data calls
OPK_DAT_RANGE	Range of number of off-peak data calls
OPK_VCE_MEAN	Mean number of off-peak voice calls
OPK_VCE_RANGE	Range of number of off-peak voice calls
OVRMOU_MEAN	Mean overage minutes of use
OVRMOU_RANGE	Range of overage minutes of use
OVRREV_MEAN	Mean overage revenue
OVRREV_RANGE	Range of overage revenue
PEAK_DAT_MEAN	Mean number of peak data calls
PEAK_DAT_RANGE	Range of number of peak data calls
PEAK_VCE_MEAN	Mean number of inbound and outbound peak voice calls
PEAK_VCE_RANGE	Range of number of inbound and outbound peak voice calls
PLCD_DAT_MEAN	Mean number of attempted data calls placed
PLCD_DAT_RANGE	Range of number of attempted data calls placed
PLCD_VCE_MEAN	Mean number of attempted voice,calls placed
PLCD_VCE_RANGE	Range of number of attempted voice calls placed
RECV_SMS_MEAN	Mean number of received SMS calls
RECV_SMS_RANGE	Range of number of received SMS calls
RECV_VCE_MEAN	Mean number of received voice calls
RECV_VCE_RANGE	Range of number of received voice calls
RETDAYS	Number of days since last retention call
REV_MEAN	Mean monthly revenue (charge amount)

REV_RANGE	Range of revenue (charge amount)
RMCALLS	Total number of roaming calls
RMMOU	Total minutes of use of roaming calls
RMREV	Total revenue of roaming calls
ROAM_MEAN	Mean number of roaming calls
ROAM_RANGE	Range of number of roaming calls
THREEWAY_MEAN	Mean number of three way calls
THREEWAY_RANGE	Range of number of three way calls
TOTCALLS	Total number of calls over the life of the customer
TOTMOU	Total minutes of use over the life of the customer
TOTMRC_MEAN	Mean total monthly recurring charge
TOTMRC_RANGE	Range of total monthly recurring charge
TOTREV	Total revenue
UNAN_DAT_MEAN	Mean number of unanswered data calls
UNAN_DAT_RANGE	Range of number of unanswered data calls
UNAN_VCE_MEAN	Mean number of unanswered voice calls
UNAN_VCE_RANGE	Range of number of unanswered voice calls
VCEOVR_MEAN	Mean revenue of voice overage
VCEOVR_RANGE	Range of revenue of voice overage

Category Variables	**Explanation**
ACTVSUBS	Number of active subscribers in household
ADULTS	Number of adults in household
AGE1	Age of first household member
AGE2	Age of second household member
AREA	Geographic area
ASL_FLAG	Account spending limit
CAR_BUY	New or used car buyer
CARTYPE	Dominant vehicle lifestyle
CHILDREN	Children present in household
CRCLSCOD	Credit class code
CREDITCD	Credit card indicator
CRTCOUNT	Adjustments made to credit rating of individual
CSA	Communications local service area
DIV_TYPE	Division type code
DUALBAND	Dualband
DWLLSIZE	Dwelling size
DWLLTYPE	Dwelling unit type
EDUC1	Education of first household member
ETHNIC	Ethnicity roll-up code
FORGNTVL	Foreign travel dummy variable
HND_PRICE	Current handset price
HHSTATIN	Premier household status indicator
HND_WEBCAP	Handset web capability
INCOME	Estimated income
INFOBASE	InfoBase match
KID0_2	Child 0 - 2 years of age in household
KID3_5	Child 3 - 5 years of age in household
KID6_10	Child 6 - 10 years of age in household
KID11_15	Child 11 - 15 years of age in household
KID16_17	Child 16 - 17 years of age in household
LAST_SWAP	Date of last phone swap
LOR	Length of residence

MAILFLAG	DMA: Do not mail flag
MAILORDR	Mail order buyer
MAILRESP	Mail responder
MARITAL	Marital status
MODELS	Number of models issued
MTRCYCLE	Motorcycle indicator
NEW_CELL	New cell phone user
NUMBCARS	Known number of vehicles
OCCU1	Occupation of first household member
OWNRENT	Home owner/renter status
PCOWNER	PC owner dummy variable
PHONES	Number of handsets issued
PRE_HND_PRICE	Previous handset price
PRIZM_SOCIAL_ONE	Social group letter only
PROPTYPE	Property type detail
REF_QTY	Total number of referrals
REFURB_NEW	Handset: refurbished or new
RV	RV indicator
SOLFLAG	Infobase no phone solicitation flag
TOT_ACPT	Total offers accepted from retention team
TOT_RET	Total calls into retention team
TRUCK	Truck indicator
UNIQSUBS	Number of unique subscribers in the household
WRKWOMAN	Working woman in household