THE APPLICATION OF DATA MINING TECHNIQUES IN QUALITY AND RELIABILITY PREDICTION AND IMPROVEMENT

By

Yanan Liang

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

April, 2009, Halifax, Nova Scotia

Approved: Dr. Muhong Wang
Supervisor

Approved: Dr. Qigang Gao
Examiner

Date: March 27, 2009

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: 978-0-494-50623-3
Our file Notre référence
ISBN: 978-0-494-50623-3

Certification

The Application of Data Mining Techniques in
Quality and Reliability Prediction and Improvement

by

Yanan Liang

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

March 27, 2009, Halifax, Nova Scotia

Examining Committee:

Approved: Dr. Qigang Gao, External Examiner
Department of Computer Science, Dalhousie University

Approved: Dr. Muhong Wang, Senior Supervisor
Department of Finance, Computing & Information Systems and Management Science

Approved: Dr. Hai Wang, Supervisory Committee Member
Department of Finance, Computing & Information Systems and Management Science

Approved: Dr. Pawan Lingras, Supervisory Committee Member
Department of Mathematics and Computing Science

Approved: Dr. Pawan Lingras, Program Coordinator

Approved: Dr. Tony Charles, Dean of Graduate Studies Rep

Table of Contents

List of Tables

List of Tables

List of Figures

Abstract

THE APPLICATION OF DATA MINING TECHNIQUES IN QUALITY AND RELIABILITY PREDICTION AND IMPROVEMENT

By Yanan Liang

The telecommunications industry is confronted with growing customer expectations and demands for better quality at a lower cost, both of which have added more challenges to the task of quality and reliability prediction and improvement. This research is motivated by a need for the knowledge to support managerial decision making, as one of the wireless devices manufacturers wants to ensure the quality of the product. This research aims to examine these existing business problems and to develop a classification and prediction model, which will give the manufacturer access to information from a range of corporate databases, deemed essential to the manufacturing, activation, and Return Material Authorization databases. The methodology in this thesis is based on a data mining approach, which focuses on the application domain. The results are expected to identify the influential factors that cause the deactivation of the devices, mainly from a product quality point of view.

April 1, 2009

Acknowledgements

First, I would like to thank Dr. Pawan Lingras and Dr. Hai Wang for their great help and support and thoughtful advice during my studies at Saint Mary's University. I am fortunate to have had the opportunity to study with these professors, who show great dedication to their teaching careers and who are very successful in their research fields. Their spirit, perseverance, and personality inspire me every day. I cannot thank them enough.

Also, I want to thank Dr. David Richardson, former Dean of Science and program coordinator, whose help in the early stages of my University life was invaluable. Dr. Richardson is a great person who thinks only of the good of the students. I also want to thank Dr. Qigang Gao, from Dalhousie University, for his valuable suggestions.

In addition, I would like to express my appreciation to the Natural Sciences and Engineering Research Council of Canada (NSERC), and Co-operative Education at Saint Mary's University for the financial support I received, and for the help I received obtaining valuable Co-op experiences. I appreciate this help.

I am, as ever, especially indebted to my parents for their love and support throughout my life. I also wish to thank my husband, who shared his love and experiences with me.

I love my grandparents, and I miss them everyday, I will never forget their love and care in my childhood. Moreover, my sincere thanks go to my family and friends in China and my friends Xiongxin Ren, George Li, and Hui Zhou. My life has been so enjoyable because of all of them.

Chapter 1 Introduction

Nowadays, the telecommunications industry is facing dramatic changes due to the influence of information technology and global competition. As the telecommunications market grows increasingly competitive, customer expectation is increasing as well. Today's customers can exercise their purchase power by choosing from many carriers or product providers to satisfy their communication needs. Accordingly, the major task facing the telecommunications industry is how to provide better services and better quality at a lower cost. In this chapter, the strategic overview of an organization in the telecommunications industry is presented, the mandate of its quality assurance (QA) group is reviewed, and the main concerns of the QA group are discussed. It was these concerns that motivated the research presented in this thesis.

1.1 Strategic Overview of Organization

The research presented in this thesis is conducted for an organization that is a designer, manufacturer, and marketer of wireless communication products, such as mobile phones and the accessories. The organization has one of the largest market shares in the telecommunications market in terms of device shipments. The market share of the company has steadily increased in the past decade. The organization is continually developing its successful story: the revenue is growing, the subscriber base is doubling, a number of advanced devices have been launched, and the carrier partners are extended over many networks globally. This growth has been driven by a number of factors, including new product introduction, global carrier launches, the development of attractive end-user pricing plans by the carrier partners, and so forth.

Therefore, the overall performance of the organization is remarkable. One of the key factors is its partnership with a large number of mobile operators. The organization now has the majority of the largest carriers in the world signed on as partners. Its global carrier partnerships have been critical to its success; they have helped the organization to build and expand its large subscriber base and to increase revenue. Another key factor in the organization's success is its more stable Average Selling Prices (ASP). The targeting of the multiple levels of customers has helped the products to become one of the most preferred wireless products on the market, and has enabled the organization to maintain more stable ASP for its devices. The growing popularity of the wireless communication devices indicates that customers and carriers are willing to support the stable ASP of the devices, if the feature set is right and the devices are executed well.

While the organization has been remarkably successful, there are some limitations that have slowed the growth of the organization.

i. Low presence in certain regions and market segments. The organization's market share is not even. It now needs to put increasing efforts into addressing and benefiting from the growing demand in the emerging markets and other regions the organization has not reached yet.

ii. Limited manufacturing and R&D capabilities. Currently, the organization is trying to increase its manufacturing capacity so that manufacturing capacity and facilities are doubled. In addition, the organization is working with outsourcing partners to provide both site redundancy and to increase manufacturing flexibility. R&D investment generally reflects an organization's willingness to forego current operations or profit to improve future performance or returns and its ability to conduct research and development. The R&D teams in this organization need to continually innovate and develop future generations of products and technologies at this time.

iii. Engineering design and user interface (UI) limitations. In order to improve the design and UI, the organization needs to seek ways to include better features, improved design, and clearer UI in the devices.

To continue performing strongly, to overcome its limitations, and to capitalize on its current share of the telecommunications market, the organization has set up many goals:

- ➢ Focus on the execution of the business strategy to ensure that the opportunities ahead will be seized.
- ➢ Continue to provide the products, services and support to extend the success of the carrier partners globally.
- ➢ Continue to grow the subscriber bases globally.
- ➢ Develop and launch products for new market segments; continue the geographic

expansion of the business into markets the organization has not reached yet.

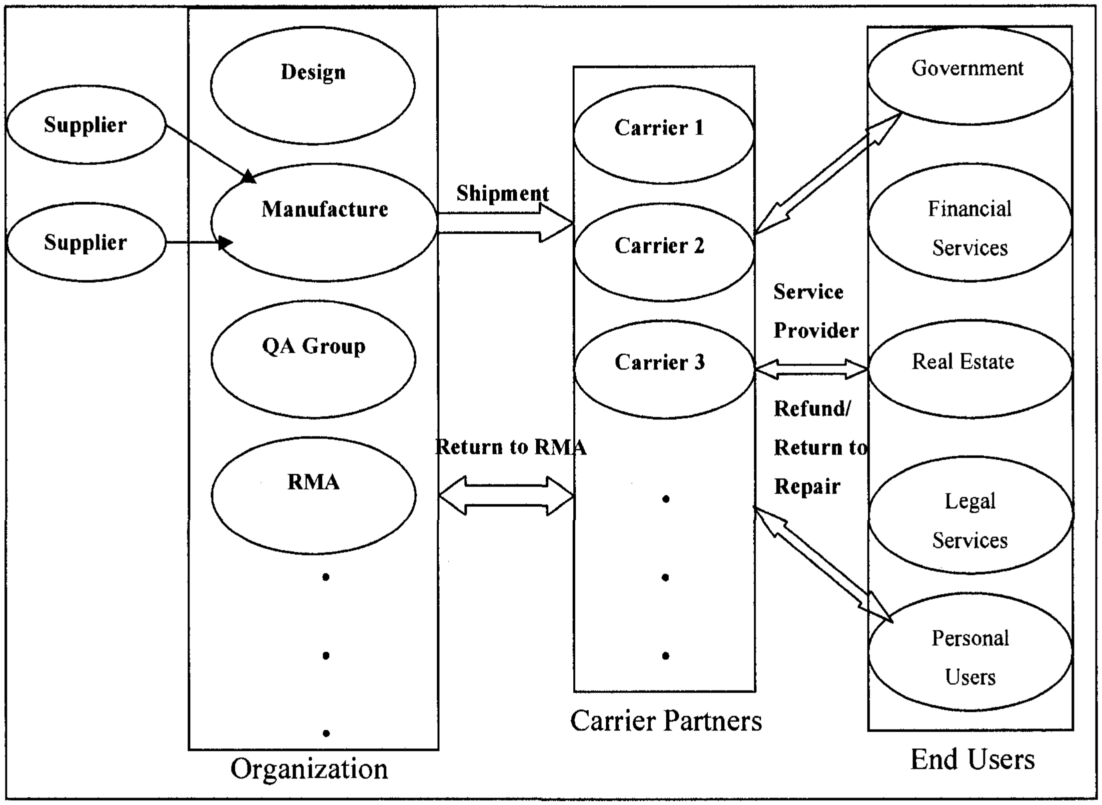


Figure 1.1 Product Distribution Diagram

Figure 1.1 shows how the wireless devices are manufactured and distributed to end users. Some components of the devices are produced by many suppliers, and the devices are manufactured and assembled in several manufacturing locations. The organization has turned its business from a direct sales model to an indirect one, selling the products primarily through the carrier partners. Therefore, the devices are first shipped to the carriers. The carriers sell the devices and provide services to the end-users. The target subscribers are not only professional and enterprise customers, but also personal phone users. Customers may return defective, failed and/or damaged devices, via the carriers, to the company for examination and repair. Return Material Authorization (RMA) is the service channel for customers and carriers who return the unsatisfactory products. After RMA repair, devices are returned to customers through the carriers.

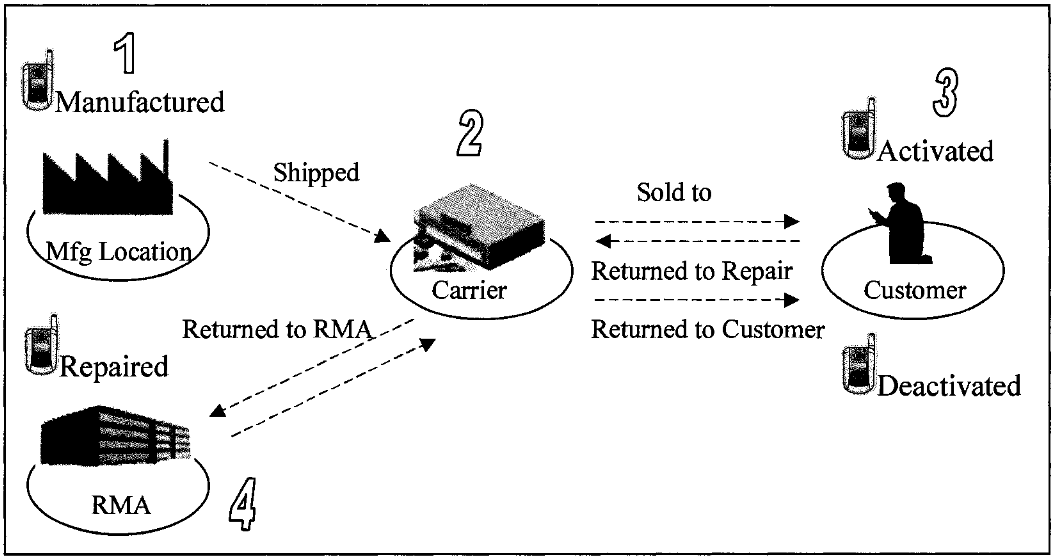


Figure 1.2 Distribution Channel of S.P.C.S

The process of Single Point of Contact Sale is illustrated in Figure 1.2. The devices are manufactured in the manufacturing locations (Stage 1). At this stage, the manufacturing information (such as the serial number, manufacturing date, manufacturing location, product type, and so on) is recorded for each of the devices and entered into the database. The devices are then shipped to carriers (Stage 2). Once a device is sold to and activated by a customer, the device's carrier, region, and activation date are recorded (Stage 3). When a customer stops using a device, it is referred to as a deactivation. The device's deactivation date can be traced and found in the database as well. Most of the devices are never sent back for repair. For those devices, the process ends at stage 3. Some devices are returned for repair due to various failure modes or customer abuse. Those devices are returned to the carriers first, who send them to RMA for repair. In the RMA process, the RMA information (such as RMA issue date, repair date, and failure mode description) is recorded for each returned device (Stage 4). Most of the repaired devices will be returned to customers after the repair and will continue to be used. This process is called Single Point Contact Sale.

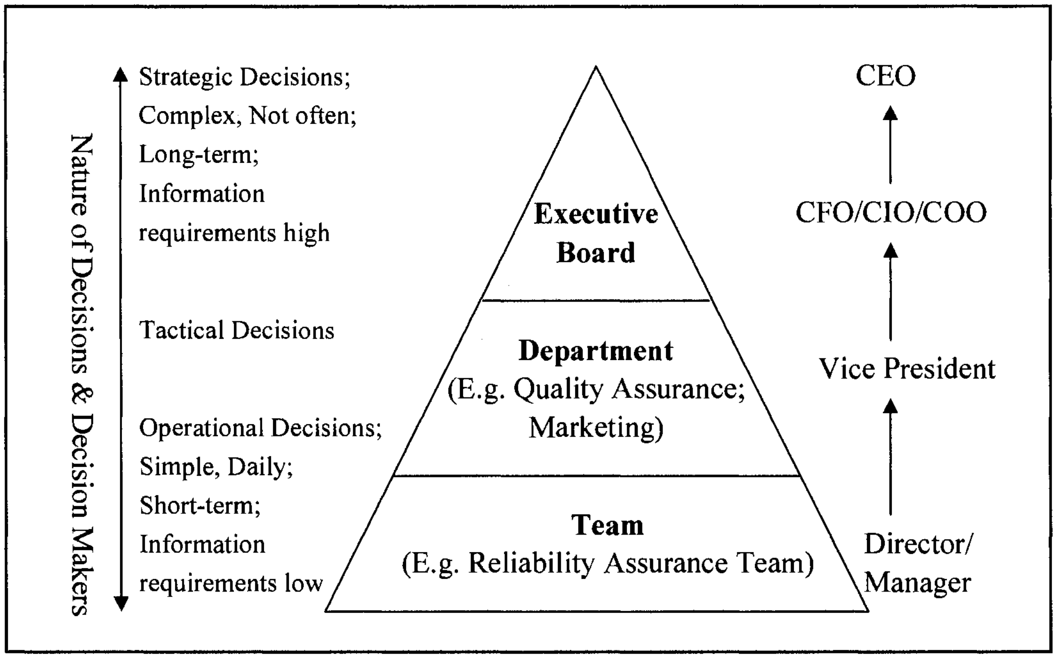

Nature of Decisions & Decision Makers
Strategic Decisions; Complex, Not often; Long-term; Information requirements high
Tactical Decisions
Operational Decisions; Simple, Daily; Short-term; Information requirements low
Executive Board
Department (E.g. Quality Assurance; Marketing)
Team (E.g. Reliability Assurance Team)
CEO
CFO/CIO/COO
Vice President
Director/ Manager

Figure 1.3 Organization Structure vs. Nature of Decisions

The above discussion is a strategic overview of the organization. The organization's structure is like a pyramid, as shown in Figure 1.3. The organization consists of many departments, such as manufacturing, engineering design, quality assurance, marketing, accounting, software development, and so on. Inside each department, there are many teams. The executive board makes strategic decisions, which provide guidelines on the direction and scope of the organization over the long-term. Operational and tactical decisions are made at the department level. The operational decision is short-term decision with immediate impact, while the tactical decision is medium range decision with more significant influence than the operational ones. The operational and tactical decisions provide the basis for making successful strategic decisions and for planning future operations.

All the data within the organization are stored in a central repository, and some local servers and databases exist in each department. The data can be accessed and analyzed to suit what information managers need. All the teams and departments work cooperatively, sharing data and information in certain degrees in order to make better decisions. The decision making process usually combines human judgment and information or knowledge. Any missing link in the information chain will have consequences, depending on the nature of the decisions. It should be understood that in the global market, products move so quickly these days from concept to production to market. A company needs to keep a vigorous flow of information in order to avoid the top executives making mistakes or falling behind. There is also a need for top executives to know where they are heading, why they are making decisions and where they need to focus. Therefore, it is very important to keep the information flowing within the organization.

1.2 Introduction to the QA Group

The task of the Quality Assurance (QA) group is to provide adequate confidence in the product's ability to satisfy given requirements. The goal of the QA group is to ensure that products fulfill or exceed customers' expectations. Quality assurance covers many activities, including failure testing, statistical process control, and total quality control. Many organizations use statistical process control to bring their products to Six Sigma levels of quality, which means that the likelihood of an unexpected failure is confined to six standard deviations of the normal distribution. This probability is less than four one-millionths. The failure testing result is used to drive engineering and manufacturing process improvements, while total quality control is the most necessary inspection control in all cases.

Quality control methodology is developed to control and improve the product of a manufacturing process. The objective of quality control is to ensure that the variables that measure a product's quality fall into ranges that are acceptable to prospective customers. Quality control methods may fall into one of three categories: (1) monitoring techniques, designed to track the level of quality variables and to detect undesirable shifts in product quality; (2) troubleshooting techniques, used to locate the cause of undesirable changes in product quality; (3) screening techniques, designed to remove defective or poor quality products entering the process as raw materials and to perform the same job for finished products before shipment to a customer. It is often said that quality control is 10% statistics and 90% engineering and common sense. The quality control methods can tell you when, but not why trouble occurs. Finding the cause of poor product quality and correcting the situation requires knowledge of the process and problem-solving abilities.

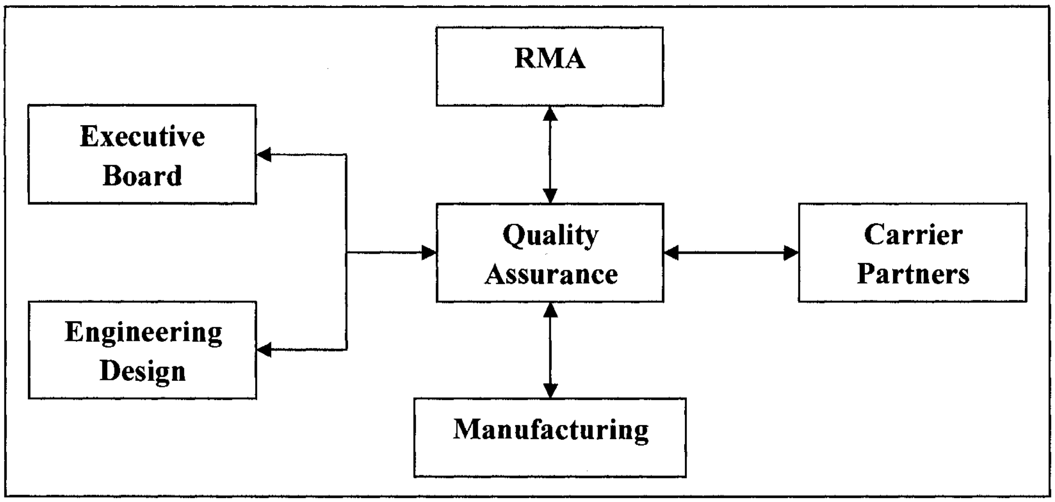


Figure 1.4 The QA Group's Information Flow Chain

Figure 1.4 shows the position of the QA group in the information flow chain in the organization. The manufacturing department provides the QA group with production information, such as production quantity, production lines, and so on. The QA group gives the manufacturing department feedback on manufacturing quality, on supplier quality and on scrap cost, and so on. Return Material Authorization (RMA) provides the QA group with repair information on the devices, such as repair date, failure mode descriptions, and so on. The QA group also gives feedback to RMA on RMA issues, mainly on some top failure modes, on a regular basis. The VP of the QA group reports to the executive board, thus supporting the board's strategic decision making, while the executive board provides directions, guidelines, and suggestions in order to ensure QA's role in reaching organizational goals. The QA group also makes recommendations to the engineering design group if many failure modes or defects are caused by a design. Engineering design provides the updates of new devices and new features to the QA group. QA also receives information from carriers regarding the activation, deactivation, and quality of the devices. QA summarizes the quality issues and provides some confidence to the carriers about product quality.

The QA group plays a very important role in the organization's success. As the manufacturing capacity, outsourcing partners, and supplier relationships expand, QA needs to ensure that the quality of the products and components from all the parties is consistent and satisfactory. If the quality of the products or components falls short, further investigation will be necessary. As a result, QA makes sure that the overall quality of the product will be improved. The second reason why the QA group is so important is because it can get feedback from end-users and make recommendations on the engineering design. New product design and development is often a crucial factor in the survival of a company. In the telecommunications market, organizations must continually revise the design and the range of products due to continuous technological change as well as to growing competition and the changing preferences of the customers. The QA group's recommendations on the weakness of a product design will help to drive the engineering design to the next level.

Most importantly, because the organization is continually providing the products and support to extend the success of their carrier partners globally and continually launching the products with the carriers, it is vitally important to retain good relationships with the carriers. A significant amount of the company's revenue derives from device sales and service fees from carriers. Therefore, the carrier partnership is critical to the increasing sales revenue and profit. The company's history of reliability has earned it the trust and confidence of both carrier partners and end users. When the organization grows and expands its subscriber base, it is essential that it continues to maintain the high quality of its products and support. The carriers also want to see increased average revenue and lower customer churn through carrying their products. The QA group is one of the channels connecting the organization to the carriers. On the one hand, the QA group is providing confidence to the carriers about product quality. On the other hand, if the carriers have any concerns about a drop in sales or product quality, the QA group is responsible for discovering if the drop in sales is caused by product quality issues or other reasons. Moreover, if the quality of the product is guaranteed, the carriers are willing to support the price of the devices, thus supporting the Average Selling Prices (ASP) strategy of the organization. So the carrier analysis by the QA group is necessary and very important.

The QA database team develops and maintains the QA database on a regular basis. The QA database is a major data resource for the QA group. Each month, the QA database team imports the manufacturing and activation data into the QA database from the manufacturing database and the relay database from different servers within the company. The central repository in the organization houses all the sources of data and provides the QA database with the RMA and shipment data. All the data in the QA database is stored in relational tables. The QA database structure is shown in Figure 1.5. The statistics team has a number of data requests. They use statistical tools to analyze the data, and then report to the VP of the QA group to support the tactical and operational decision making process.

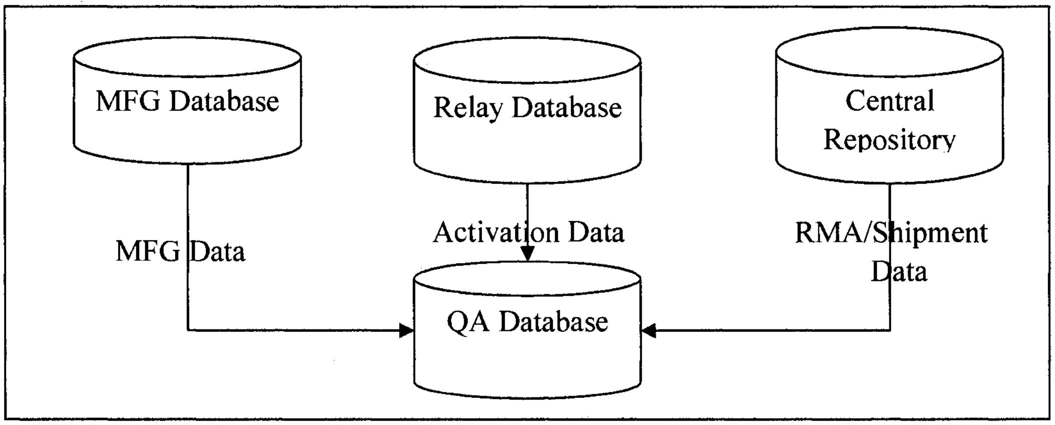


Figure 1.5 The QA Database Structure

Presently, some major concerns are described by the decision makers in the QA group as follows:

1. Deactivation

When a device is deactivated, it means that the device is no longer being used. In this research, we assume that customers use only one device during the time they are using the wireless services. The deactivation of a device not only means the device is no longer being used, but also means the customer has terminated the customer relationship with the service provider. A significant proportion of the company's revenue is generated from device sales and service fees from carriers. A steady stream of revenue is generated as long as a device is activated and being used. There are many reasons that can lead to the deactivation of devices:

- Customer behavior. A customer may stop using a device due to financial and/or personal reasons or other reasons beyond his/her control (e.g., the device may be lost or stolen). The customer behavior is obviously related to the former case but not to the latter. The prediction of customer behavior is particularly difficult. However, if customer information, such as age, gender, occupation, as well as contract information, billing information, and device usage data, is available, then it is possible to estimate the patterns of a customer's behavior. Because the QA group does not have access to customer information data, contract and billing information, and usage data, those factors are not included in this research.
- Carrier and geographical differences. Some carriers may provide poor services but charge high service fees, or customers may relocate to other locations. In situations like this, customers can exercise their purchase power to switch to

other carriers to satisfy their communication needs. In this research, we are able to identify from which carriers and regions the devices are more likely to be deactivated.

- ➢ Advanced technology and/or new products launched, especially those from competitors. The duration of a device being used can be affected when a new product with advanced technology hits the market; it may cause customers to stop using the current device, thus cutting down the useful life of the device. Currently, we do not have information in this regard; therefore, this factor is not included in this research.
- ➢ Failure mode and/or customer abuse. Sometimes a device does not function properly; therefore, it is deactivated due to failure modes or customer abuse. Although most of the failure modes and customer abuses can be repaired, the customer will still be very likely to deactivate the device if the quality issues appeared frequently. It is one of QA's mandates to minimize the impact of quality issues on deactivation through quality control and statistical analysis.

Overall, the prediction of deactivation can be complicated. There are many factors that may affect the deactivation, some of which are not measurable, or information may not be available. But the QA group has access to RMA data, so they will be able to analyze the quality issues and reduce the deactivation of devices through the quality control, thus maximizing the carriers' and the customers' satisfaction and the company's revenue. Currently, the Pareto charts are used to represent the numeric results in the carrier report, including, for example, the quantity of deactivation. However, there is no systematic method for predicting deactivation; thus, there is no specific requirement in collecting relevant information in this regard. One of the major concerns in this research is to identify influential factors that may cause deactivation. The analysis will be based mainly on the factors recorded in the manufacturing data, activation data as well as RMA data.

2. Age

Age is measured as how many days a device has been in use, which is also referred to as time in use. Age is an important measurement of customer loyalty. The longer a customer uses the device, the more revenue will be generated from the service fees. As mentioned early, in this research we assume that customers use only one device during the time they are using wireless services. If some of the customers can afford the newer models, they will buy new models but stay with the same wireless services. The situation like this is hard to trace, and it might be a rare case. Both carriers and manufacturer like to increase the time in use. It is observed that some of the devices deactivated with long ages, while many of devices deactivated with very short ages. The average age in the carrier report only indicates the numeric result, but it does not show the root cause of short-age devices. Like deactivation, many factors may affect the age. Customer behavior is again an important factor, and quality issues are also major factors. This research aims to find out why some of the devices have short ages, to develop a method for identifying the influential factors on age based on the data available (activation and RMA data), and to show how time in use changes by carrier, location, and so forth.

3. Time Delay between Manufacturing and Activation

We do not think time delay has much impact on deactivation. But both manufacturer and carriers want to reduce the time difference so that they can speed up the activation of the devices and maximize profit. It will be interesting to see how time delay changes by carrier and location in the time delay analysis.

4. Failure Mode Association and Classification

Failure mode is an important factor affecting deactivation and age. The statistical method assumes each failure mode is independent; it ignores the association between failure modes. Thus, only the top few failure modes are being analyzed individually. However, it is beneficial to discover the relationship between failure modes. This information will help to improve the engineering design and cut the exam time in RMA. The research aims to identify the combination of failure modes with higher frequency and to verify if they directly cause the deactivation of the devices. It is also useful to estimate the approximate time to fail for each failure mode in order to determine the warranty period.

The use of database and statistical techniques are well established in the QA group. Both the data analysis tools and the statistical software return rather good results, but those data analysis and reporting tools have some limitations, which restrict the power of the knowledge discovered from the data: (1) The size of the data sets are huge; the existing tools lack the ability to handle such big data sets. (2) The data analysis tools are not directly connected to the database; various data requests and data transformations have to be completed manually. These tasks are time consuming, and they might not be precise because of the increased chance for human error. Additionally, there is a data security issue. (3) Other methods, such as Pareto chart, aim to monitor the numeric data, but they ignore the relationship between the target attribute and the input attributes, as well as the correlations between input attributes.

1.3 Introduction to Data Mining

With the growth of the organization and the success of its products, it is necessary that quality control and quality assurance become more sophisticated in order to meet customer expectations and the information needs of the organization. Making informed decisions about product quality requires accurate data and measurements as well as powerful data analysis tools.

Telecommunications operators have to manage one of the most complex systems in all industries. They not only manage the junction of voice and data networks, but also the activation and deactivation of millions of devices worldwide each day. Therefore, the organizations in the telecommunications industry generate a large amount of data. Those data include call detail data, network data, customer data, and product data. The organization now realizes it is important to take advantage of those data, so there is a demand for powerful data analysis tools to handle such big data sets.

Data mining (DM) is a natural solution because it is a process of extracting knowledge hidden in large volumes of data. Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and databases. The patterns, associations, or relationships among all this data can provide information. And information can be converted into knowledge about historical patterns and future trends.

Decisions require a solid foundation of data and analyses to be acceptable and defensible. Figure 1.6 shows the transformation of data to information, knowledge, and decisions, with rapidly increasing value as it moves from the bottom of a decision pyramid to the top. Data by themselves have little value; once they are processed, analyzed, and organized, they become information. Once the information is processed by the user to develop and enhance the understanding of a situation or a problem, it transforms into knowledge. This knowledge can reduce the risk of making decisions with undesirable consequences. One approach is to start at the top of the decision pyramid and travel down the pyramid to determine what subset of data and processing is essential for decision support on a specific set of decisions. This approach can result in significant cost savings in data collection and analyses because it prioritizes decision support needs. In this thesis, we follow the same approach by first defining the business needs and decisions to be made and then determining the data that is relevant to the problems from the QA database.

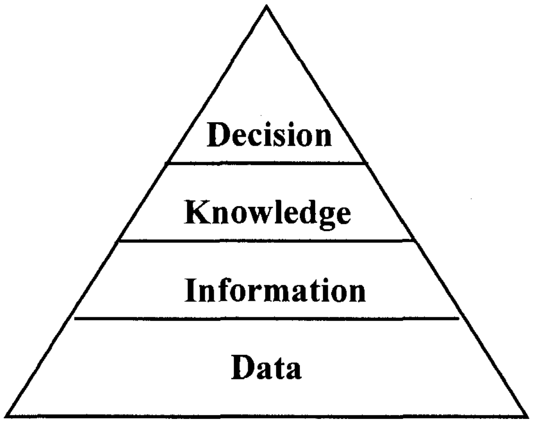


Figure 1.6 Data-Decision Hierarchy Pyramid

The utilization of data mining can enhance the ability to discover some hidden knowledge that might significantly improve various organizational strategic and operational decisions. The telecommunications industry has been one of the earliest adopters of data mining technology, mostly because of the amount and quality of the data that it collects. Data mining is now applied in many areas in manufacturing, engineering design, fault detection, quality assurance, and decision support systems. These areas have many successful data mining applications.

Therefore, there is a need to establish a data mining model and integrate it into the decision support system, not only because of the potential of data mining to overcome the problems with current data analysis tools, but also because, in the long run, it will create a repeatable and reproducible knowledge resource, and it will serve as a foundation of future decision making or prediction. Additionally, there is no conflict between data mining and statistics; the QA group can still take advantage of the existing statistical tools because they do not overlap. The main differences of these two methodologies are:

- Statistics assumes a pattern, and the algorithms attempt to prove it; DM describes a pattern which the algorithms find.
- DM processes data that is usually given as a large database or a large flat file; statistics are often applied to small and clean data sets.
- The objective of DM is to find patterns, knowledge, and valuable new information in data; through statistical analysis, data is processed according to a defined objective of analysis.
- Statistics consider data variation, but this is not considered in DM.
- In DM, residual data is useful, and it is processed; in statistics, it is removed from the original data set.

Although data mining might be the solution, telecommunications data pose some interesting issues for data mining as well. The first is scale; telecommunication databases may contain millions of records. The second issue is that the raw data is often not suitable for data mining. For example, call detail data is time-series data that represent individual events. Before this data can be effectively mined, a useful "summary" must first be identified. Thus, data mining requires more attention and effort in the data preprocessing stage. Rarity, or an unbalanced distribution of the target attribute, is also an issue because many data mining applications in the telecommunications industry involve predicting rare events, such as return for repair. In addition, some built-in strong patterns are involved in the manufacturing process; thus, there is a risk that the knowledge discovered by data mining will not be interesting but just common sense.

1.4 Motivation and Objectives

The concerns described by the decision makers in the QA group, the limitations of the current data analysis tools, and the interesting characteristics of the telecommunications data are the motivations of this thesis research, while many publications are focusing on creating data mining algorithms for manufacturing and quality assurance, this thesis research focuses on the description of business problems and the utilization of the data mining process in the context of enterprise application.

This thesis research aims to put all the major business problems and concerns of the quality assurance group together as one system, the Deactivation Analysis System. In this thesis, we utilize statistical analysis and data mining techniques to build the system. This approach demonstrates the uniqueness of this thesis: no data mining techniques have ever been used by the QA group. Hopefully, this thesis will also serve as a guideline of how to initialize a data mining project and how to take advantage of data mining techniques to solve certain problems within a QA group.

Due to the limited access to certain data, this thesis research is only performed on the product level data, which is distinguished from the prediction of customer behavior on the business level. That is, this thesis develops a Deactivation Analysis System through a data mining approach, which identifies the influential factors to deactivation and age mainly from a product quality point of view. Thus, it will support the decision making process and give the decision maker access to the information summarized from manufacturing, activation and deactivation, and RMA data.

1.5 Thesis Outline

The thesis is organized as follows: Chapter 1 sets out the objectives of the thesis research, introducing the organization and describing its business problems and requirements. Chapter 2 reviews the related concepts, methods, and approaches and their applications in the industries. Chapter 3 presents the methodologies that will be used in our data mining model. Chapter 4 presents the data preprocessing and preliminary analysis. Chapter 5 describes details of the experimental design and evaluates the results. Chapter 6 concludes the research, makes recommendations to the organization, and initiates future work.

Chapter 2 Background Review

In the information age, most organizations, such as financial institutions, retailers, and the telecommunications industry, generate large volumes of data from a variety of sources. In the telecommunications industry, the data is overwhelming. The industry keeps track of billions of phone calls, as well as customer and product information. It is impossible to analyze it all manually. Unlike traditional data analysis methods, such as statistical analysis, which only works well on small and "clean" data sets, data mining might be a solution to the problem of handling the large and messy data sets in the telecommunications industry. As a result, data mining has received a lot of attention in the research community lately. It is a broad area that integrates statistics, machine learning, and artificial intelligence. In this chapter, we first review some popular data mining techniques and their algorithms from a theoretical point of view, and then the data mining applications in the telecommunications industry and quality assurance field. A review of related work serves as a foundation of the study in this thesis research.

2.1 Decision Tree

Predicting future outcomes and identifying influential factors are often the main goals of data analysis and data mining. A decision tree is a flow-chart-like structure where each node denotes a test on an attribute, each branch represents an outcome of the test, and each tree leaf represents a class or a class distribution (Han & Kamber, 2001). The decision tree model is one of the most popular data mining methods because it returns interpretable rules and logic statements that enable more intelligent decision making. A decision tree normally predicts discrete outcomes; and is referred to as a regression tree if the outcome is a continuous variable.

Decision tree learning is generally best suited to problems with instances that are represented by attribute-value pairs. That is, instances are described by a fixed set of attributes, and each attribute takes on a small number of disjoint possible values or real values. A decision tree also works well on a target function that has discrete output values; for instance, the simplest case exists when there are only two possible classes (Boolean classification) or more than two possible output values (Hamilton et al., 2003).

Decision trees have several advantages compared with other data mining methods:

- Simple to understand and interpret. Decision makers or managers are able to understand decision tree models with only a little data mining knowledge.
- Have value even with small data set. A decision tree can be as small as just a few nodes and branches to solve small problems, such as whether or not we should play golf today. A decision tree can grow huge depending on the complexity of the problem and the associated attributes. A smaller tree is normally easier to understand and provides meaningful classification rules. Bigger trees can be too complicated and generate too many decision rules.

- Can be combined with other decision techniques. When we talk about decisions or actions, it is usually associated with costs. For example, a cost-sensitive learning algorithm can be added to the decision tree, which takes cost into consideration and improves the classification accuracy (Wu, 2005).
- Able to handle both numerical and categorical data. Some techniques are specialized in analyzing data sets that have only one type of variable. But a decision tree can handle the mixed data sets.
- Possible to validate a model using statistical tests. Statistical tests make it possible to account for the reliability of the model.

Creating a decision tree is an important technique in machine learning, and it is used extensively in data mining. The results generated from decision trees are easy to read; they describe the trends in the underlying relationships in a data set and may be used for classification and prediction tasks. The technique has been used successfully in many different areas, such as medical diagnosis, plant classification, and target marketing strategies. There are some major algorithms for the induction of decision trees, including C4.5, CART, and CHAID, which are also among the most influential data mining algorithms in the research community. In the following sections, we will examine each algorithm closely.

2.1.1 C4.5 Algorithm

C4.5 is an algorithm developed by Ross Quinlan in 1993 to generate decision trees. It is an extension of an earlier version called ID3, also developed by Quinlan. C4.5 offers additional features, such as:

- ✧ Using information gain ratio as an option in the testing measures
- ✧ Handling training data with missing attribute values
- ✧ Handling continuous attributes
- ✧ Avoiding over-fitting the data
- ✧ Reduced error pruning; added rule post-pruning (Hamilton et al., 2003)

C4.5 builds decision trees from a set of training data using the concept of information entropy. C4.5 examines the normalized information gain (difference in entropy), which is also called gain ratio, that results from choosing an attribute for splitting the data. Gain ratio is a modification of the information gain that reduces its bias on high-branch attributes (Witten & Frank, n.d.). Suppose S represents a data set and A is an attribute, Values A is the set of all possible values for attribute A, and S_v is the subset of S for which attribute A has value v. Intrinsic information measures how much information we need to tell which branch an instance belongs to. It is defined as:

$$IntrinsicInfo(S,A) = -\sum_{i=1} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \quad (2.1)$$

(Wu, 2005).

The information gain is defined as:

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2.2}$$

(Wu, 2005).

Gain Ratio normalizes information gain by:

$$GainRatio(S,A) = \frac{Gain(S,A)}{IntrinsicInfo(S,A)} \tag{2.3}$$

(Wu, 2005).

Gain ratio takes the number and size of branches into account when choosing an attribute; thus, it overcomes the bias of the information gain measure, which favors attributes with many values over those with few values (Wu, 2005). The attribute with the highest normalized information gain is the one used to make the decision. The algorithm then recurs on the smaller lists.

C4.5 handles missing data by assigning a probability to each possible value of the related attribute. The assigned probability is estimated by the observed frequency of the attribute value among the instances at the tree node. C4.5 handles continuous-valued attributes by dynamically defining a best cut point of the attribute. After sorting the instances according to the attribute values, C4.5 discretizes values with midpoints as candidate thresholds; the value of the threshold that maximizes information gain must always lie at the boundaries, so the best cut point is the one that maximizes information gain (Wu, 2005).

C4.5's pruning strategy is to avoid over-fitting of the data by removing the branches reflecting noise, thus improving accuracy. The divide and conquer algorithm partitions the data until every leaf contains cases of a single class or until further partitioning is impossible because two cases have the same values for each attribute but belong to different classes. Consequently, if there are no conflicting cases, the decision tree will correctly classify all training cases. This so-called over-fitting is generally thought to lead to a loss of predictive accuracy in most applications (Quinlan, 1986). Over-fitting can be avoided by including a stopping criterion that prevents some sets of training cases from being subdivided or by removing some of the structure of the decision tree after it has been produced. Most authors agree that the latter is preferable because it allows potential interactions among attributes to be explored before deciding whether the result is worth keeping (Kohavi & Quinlan, 1999). Post-pruning is adopted in C4.5 by using sub-tree replacement (replace a sub-tree by a single leaf) and sub-tree raising (replace a sub-tree by its most frequently used branch) rather than pre-pruning. Pruning of nodes continues until further pruning decreases the accuracy of the tree.

2.1.2 CART Algorithm

CART builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). The classic CART algorithm was first introduced by Breiman et al. (1984). The CART decision tree is a binary recursive partitioning procedure capable of processing continuous and nominal attributes both as targets and predictors, treating ordinal inputs as interval attributes. The term "binary" implies that each group of instances, represented by a "node" in a decision tree, can only be split into two groups. Thus, each node can be split into two child nodes, in which case the original node is called a parent node. The term "recursive" refers to the fact that the binary partitioning process can be applied over and over again. Thus, each parent node can give rise to two child nodes and, in turn, each of these child nodes may themselves be split, forming additional children. The term "partitioning" refers to the fact that the data set is split into sections or partitioned (Lewis, 2000).

The differences between C4.5 and CART are in the tree structure, the splitting criteria, the pruning method, and the way missing values are handled. CART constructs trees that have only binary splits. This restriction simplifies the splitting criterion. CART uses the Gini diversity index as a splitting criterion (Kohavi & Quinlan, 1999). Let RF (C_j, S) denote the relative frequency of cases in S that belong to class C_j. The Gini index is defined as

$$I_{gini}(s) = 1 - \sum_{j=1}^{x} RF(C_j, S)^2 \tag{2.4}$$

(Kohavi & Quinlan, 1999).

CART's pruning strategy is different from C4.5 as well. CART prunes trees using the cost-complexity model, in which whole parameters are estimated by cross-validation

(Kohavi & Quinlan, 1999). First, it grows the largest tree, called the "maximal" tree. And then it prunes away the "weakest links," which are the nodes that add least to the overall accuracy of the tree. CART determines a pruning sequence for every node all the way back to the root node, which is the exact order in which each node should be removed.

Cost Complexity = Resubstitution Misclassification Cost + $\beta \times$ Number of terminal nodes (2.5)

(Yohannes & Webb, 1999).

In Formula 2.5, β represents penalty per additional terminal node. If β=0, then cost complexity attains its minimum for the largest possible tree. On the other hand, as β increases and is sufficiently large, a tree with one terminal node (the root node) will have the lowest cost complexity. Then the CART takes a test data set and drops it down the largest tree in the sequence and measures its predictive accuracy. Therefore, the CART procedure requires test data to guide tree evaluation, and the predictive accuracy is evaluated by the mean error rate.

Unlike C4.5, CART does not penalize the splitting criterion during the tree construction if examples have unknown values for the attribute used in the split. The criterion uses only those instances for which the value is known. Also CART finds several surrogate splits that can be used instead of the original split. As CART implies, it also supports regression trees. Regression trees are somewhat simpler than classification trees because the growing and pruning criteria used in CART are the same. The regression tree structure is similar to a classification tree, except that each leaf predicts a real number. The estimation criteria used is the mean squared error.

2.1.3 CHAID Algorithm

CHAID is a type of decision tree technique. It was published in 1980 by Gordon V. Kass. It can be used for prediction or for detection of interaction between variables. CHAID stands for Chi-Squared Automatic Interaction Detector. The Chi-squared goodness of fit test is used to identify significant predictors and to merge predictor categories that do not differ in their prediction of the dependent variable. For CHAID, the inputs are either nominal or ordinal. Many software packages accept interval inputs and automatically group the values into ranges before growing the tree. The CHAID algorithm is like sequential cross-tabulation. For each predictor:

1. Cross tabulate the m categories of the predictor with the k categories of the dependent variable.
2. Find the pair of categories of the predictor whose 2xk sub-table is least significantly different on a Chi-squared test and merge these two categories.
3. If the Chi-squared test statistics is not "significant" according to a preset critical value, repeat this merging process for the selected predictor until no non-significant chi-square is found for a sub-table.
4. Pick the predictor variable whose chi-square is largest and split the sample into m <= l subsets, where l is the number of categories resulting from the merging process on that predictor.

5. Continue splitting until no "significant" chi-square results (Wilkinson, 1992).

The CHAID algorithm saves some computer time, but it is not guaranteed to find the splits that predict best at a given step; only a search of all possible category subsets can do that. For some software, CHAID is limited to categorical predictors, so it cannot be used for quantitative or mixed categorical quantitative models. And because it uses multi-way splits by default, it needs rather large sample sizes to work effectively. With small sample sizes, the respondent groups can quickly become too small for a reliable analysis. Nevertheless, CHAID is an effective tool for searching heuristically through rather large tables quickly (Safavian & Landgrebe, 1991).

2.1.4 Summary and Applications of Decision Tree Algorithms

In summary, the comparison of these three decision tree algorithms in terms of tree structure, splitting criterion and pruning reveals that they are different in a few important ways. Each has relative advantages and disadvantages, depending on the data mining purpose and the data structure.

CART is a data-exploration and prediction algorithm, which summarizes a classification and regression tree. The CHAID modeling is essentially a "stepwise" statistical method. CHAID grows non-binary trees through a relatively simple algorithm that is particularly well suited for the analysis of larger data sets, and it has been particularly popular in marketing research (Tang et al., 2005). CHAID is similar to

CART, but it differs in choosing a split node. It depends on a Chi-squared test used in contingency tables to determine which categorical predictor is farthest from independence with the prediction values. It appears that CHAID is most useful for identifying major data trends and data analysis, whereas CART is more suitable for prediction. In other words, CHAID should be used when the goal is to describe or understand the relationship between a response variable and set of explanatory variables, whereas CART is better suited for creating a model that has high prediction accuracy of new cases. The C4.5 algorithm is mainly about entropy calculation, which examines the attributes to add at the next level of the tree using an entropy calculation and chooses the attribute that minimizes the entropy. In the experimental design and result evaluations, we will compare and consider the differences between the algorithms and splitting methods in order to find optimal solutions for each model and each business problem.

Due to the nature of decision trees, a wide range of fields have deployed successful applications of decision trees. Early users of decision trees tended to be in information-intensive industries, such as financial services and direct mail marketing. Decision trees have some successful application areas, including campaign management, customer segmentation, and target marketing. The following are some examples of applications of three decision tree algorithms.

Rosset et al. (1999) presents a two-stage rules-based fraud detection system which first involves generating rules using a modified C4.5 algorithm. Next, it involves sorting rules based on accuracy of customer level rules, and selecting rules based on coverage of fraud of customer rules and difference between behavioral level rules. In the first stage, the authors build a rule generator based on a modification of the C4.5 algorithm. The relevant changes in the algorithm are concentrated in three areas, splitting criterion, stopping rule, and pruning significance tests. The splitting criterion is used to select the "best" greedy split in each stage during tree construction. It is based on calculating the "information content" of each of the suggested splits with regard to the class distribution and choosing the one with the highest content. The stopping rule dictates the size of groups can be accepted as "leaves" in the tree. The goal of using a stopping rule is to prevent the system from creating rules representing small samples with no statistical generalization ability. For both of these areas the key to working on bi-level data is that the "size of groups" concept has to be defined with respect to the level at which the attribute being split belongs. So, when splitting on a customer-level attribute, the amount of customers of each class found in each "leaf" is counted. The idea of the necessary changes for the pruning significance tests is similar. The modified C4.5 algorithm and the rule selection generate many "meaningful" rules and the results indicate that this route is promising.

In order to show the benefits of data mining in health care management, Bach and Cosic (2007) show a way to raise awareness of women in terms of contraceptive methods they use or do not use by using a decision tree. In the application of data mining in explaining women's choice of contraceptive methods, the authors use CHAID algorithm. CHAID is different from other decision tree algorithms because it follows pre-pruning method, that is, it tries to stop the branching before the over fitting occurs. CHAID uses Chi-squared test to make a decision about merging the fields that do not create statistically significant differences in the values of a target field. The tree grows until there are no more splits that lead to statistically significant differences in classification.

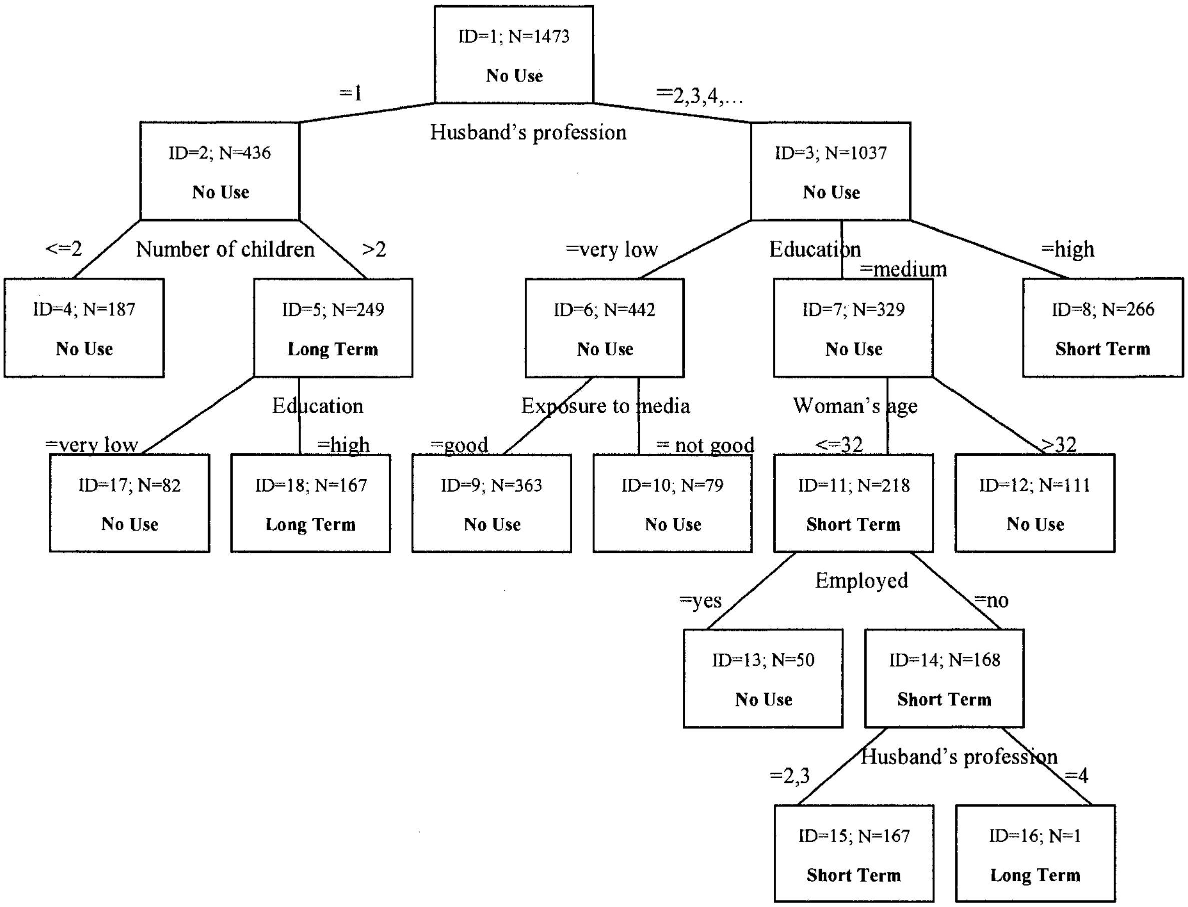


Figure 2.1 CHAID Decision Tree for the Choice of Contraceptive Method

In the “choice of contraceptive method” in Figure 2.1, CHAID does the first split on “husband’s profession”, which is obviously the most important. Another thing that can be relevant is the number of “ever born children,” meaning that significant changes in percentages of contraceptive methods used occurs after second child, when long term method prevails. Third significant variable, “woman’s education,” has a significant influence when level of education is increasing. “Women’s age” also has an important role in this case, since, until the age of 32, women mostly use short term methods and after that “non usage” prevails. Further on, CHAID divides a population of women younger than 32 according to their “status of employment,” where working women mostly do not use contraception and unemployed women mostly use short term methods.

According to these results, the authors come out with some conclusions and recommendations, such as the health care management can raise the awareness among women through a campaign for contraceptive products. A marketing strategy should focus on short term contraceptive products, and a promotion of products should target the population of women whose education is very low, whose husbands do not work on hierarchically low positions, and whose exposure to the media is not good. Meanwhile, because of the instability of the decision tree algorithms slight variations in the data can result in different attribute selection at each branch node. Therefore, very different set of rules can be produced based on the slightly different data set used.

In Timofeev's Master's thesis (2004), the Boston Housing is used as an example for regression trees. The data set includes 13 independent variables, and one response variable – value of house. At the upper levels of the tree there are more significant variables, and less significant variables are at the bottom of the tree. For Boston Housing data set, "average number of rooms" is most significant, since it is located in the root node of the tree, and then come variable "percent lower status" and "weighted distance to employment center." The result also shows the optimization (pruning) of the tree is very important for regression trees, since the maximum tree is too big. There are several methods for tree pruning; "cross-validation" is often used, and another one is adjusting the parameter of minimum number of observations. By increasing the parameter, the size of the tree will decrease and vice versa.

2.2 Association Rules

In data mining, association rule learners are used to discover elements that co-occur frequently within a data set. Questions such as "if a customer buys product A, how likely is he to buy product B?" and "What products will a customer buy if he buys product C and D?" are usually answered by association rules, so association rules are normally used in market basket analysis. In practice, other fields benefit from association rules as well, such as:

- ➢ Common combination of services and products from banking services that customers like to purchase together.
- ➢ Commonly associated options (call waiting, call display) help determine how to structure product bundles that maximize revenue.

- Certain combinations of conditions can indicate the high risk of various complications from medical patient histories.

The Apriori algorithm (Agrawal & Srikant, 1994) is a great achievement in the history of mining association rules. It is by far the most well-known association rule algorithm. The Apriori pruning principle is if there is any item set that is infrequent; its superset should not be generated or tested (Mining frequent patterns, n.d.). So the fundamental differences between Apriori and the other association rule mining algorithms (e.g., AIS, SETM, and AprioriHybrid Algorithms) are the way of generating candidate item sets and the selection of candidate item sets for counting. Usually, the other algorithms generate too many small candidate item sets. The Apriori algorithm addresses this important issue. The Apriori generates the candidate item sets by joining the large item sets of the previous pass and deleting those subsets that are small in the previous pass without considering the transactions in the database. By only considering large item sets of the previous pass, the number of candidate item sets is significantly reduced (Dunham et al. 2001).

```
$L_1$ = {large 1-itemsets};
For ( k=2; $L_{k-1} \neq \emptyset$ ; k++) do again
    $C_k$ = apriort-gen($L_{k-1}$); // New candidates
    For all transactions $t \in D$ do again
        $C_t$ = subset($C_k$, t); //Candidates contained in t
        For all candidates $c \in C_t$ do
```

```
c.count++;
end
$L_k = \{ c \in C_k \mid c.count \geq minsup \}$
End
Answer = $\bigcup_k L_k$;
```

Figure 2.2 Apriori Algorithm

MacDougall (n.d.) describes a case study in which the association node in the SAS Enterprise Miner is used to examine relationships between the demographic characteristics, political party affiliation, and media influences of respondents in a survey of US voters. Although the data set is small in this study, only about 1800 observations, the large number of interview questions covers a wealth of detail about each respondent's demographic and social background as well as political views. The author suggests that lift is very useful because it measures the extent to which the rule improves the ability to predict the right-hand side. Therefore, the author combines the lift value and other measures (such as confidence, support, and frequency) with her judgment, thus coming up with some meaningful and interesting rules. For example, the association rules for independents with higher lift show that independents tend to be younger and less educated and tend not to vote. They support a mixture of issues including death penalty, gays in the military, and tax cuts. The study demonstrates that generating association rules can be a useful starting point for exploring unfamiliar data; the results are better than a decision tree and clustering in terms of solving these kinds of problems. The study shows the effectiveness of the data mining process from exploration, data transformation, to association analysis.

2.3 Data Mining Applications in the Telecommunications Industry and Quality Assurance group

In recent years, the telecommunications market has grown significantly due to the rising population of wearable and affordable communication devices. The competition in the telecommunications market is intense: carriers and product providers have become much more competitive in pricing and cost, the bundling of services and features, and the quality of products and services. So the telecommunications market is facing growing consumer demand, growing competition, and growing optimization. There are some key functions that enhance the competitive advantages. (1) Understanding the needs of their business (business intelligence). (2) Managing actions based on those needs (business management). (3) Running day-to-day operations effectively (business operations). Business intelligence includes customer churn (loss of customers) prediction, market segmentation, modeling, and so on. Data mining is playing an increasingly important role in business intelligence, as it helps businesses to understand their needs and to meet the information requirements for business management and business operations.

The telecommunications industry generates and stores a tremendous amount of data. These data include call detail data, network data, and customer data. At the same time, these data offer a huge potential as sources of new knowledge. The need to handle such large volumes of data and to extract knowledge from the data has become an issue, and data mining is a natural solution for transforming the data into useful knowledge. Knowledge is the most valuable asset of an organization; the extracted knowledge can be used to model, classify, and make predictions for numerous applications. And normally decisions are made based on a combination of judgment and knowledge from various domains. The idea of extracting knowledge from manufacturing, business, or medical data is not new. Traditionally, it was the responsibility of analysts, who generally used statistical techniques; however, analysts have increasingly turned to data mining, due to its ability to handle large volumes of data. Thus, the telecommunications industry was an early adopter of data mining technology.

There are two major data mining applications in the telecommunications industry that are well explored and developed, one is fraud detection, and another is churn modeling. Fraud is a serious problem for telecommunications companies, leading to billions of dollars in lost revenue each year. The most common method for identifying fraud is to build a profile of customer's calling behavior and to compare recent activities against this behavior in order to classify if the customer is insolvent or solvent. Misclassification costs for fraud are generally high. When building a classifier to identify fraud, one should ideally know the relative cost of letting a fraudulent call go through versus the cost of blocking a call from a legitimate customer. Customer churn, also known as customer turnover or customer attrition, is a business term used to describe the loss of clients or customers. Customer churn is a significant problem because of the associated loss of revenue and the high cost of attracting new customers. To better prevent customer churn, companies need to enhance customer satisfaction, and to group customers by some patterns and then promote the product accordingly. To predict customer churn is essentially to predict customer behavior which will support efficient direct marketing campaigns and quality cross-selling services.

This is an example of data mining application in fraud detection. Aiming to reveal the behavioral patterns from customer profiles, service usage, and financial transaction data, Daskalaki et al. (2003) build a decision support system to handle customer insolvency for a large telecommunications company. The authors apply three classification algorithms (discriminant analysis, decision tree, and neural networks) on the same set of data. Using a case-by-case comparison, it appears that the decision tree classifier performs the best in terms of maximum prediction accuracy and minimum error rate. When considering the misclassification rate of positive examples, however, none of the algorithms are considered satisfactory for business use. The result shows that the application of a combined algorithm increases the confidence of the classification prediction by reducing significantly the false positive rate. The result also demonstrates that the proposed architecture can be the core component of a decision support system that provides advice on future customer insolvencies.

In addition to the widely used data mining in the telecommunications industry, an exponential growth of data mining applications in quality assurance and improvement has been observed. The reasons for this growth may be that large volumes of data are generated during manufacturing and RMA and those small improvements can have a significant impact in the industry. There is a strong potential that data mining can be used to improve quality control; the key is to accurately determine types of failure modes and defects.

The drop test is a common method for systematically determining the reliability of portable electronic products under actual usage conditions used by quality assurance groups. The test mainly improves the engineering design, decreases design cycles, and determines the best design parameters. Zhou et al. (2001) apply the C4.5 algorithm to a drop test analysis of electronic goods. To meet the needs of quickly identifying the root causes of defects in very complex manufacturing processes, Chen et al. (2004) first generate association rules for defect detection in semiconductor manufacturing by applying a typical algorithm of mining association rules, the correlation between combinations of machines and the result of defect is defined. Then an integrated processing procedure (a Root Cause Machine Identifier) is proposed to discover the root cause of the problem. And then the results of experiments are evaluated and proved to be accurate and efficient in real manufacturing cases. Al-Salim and Abdoli (2005) present a two-stage methodology for finding the clusters of quality problems for quality improvement in industrial firms. The first step is to determine the related quality problems using association rules based on their likelihood to appear together. This step effectively reduces the number of possible combinations of quality problems. In the second step, an optimization model is employed for selecting the minimum cost of quality improvement systems. Results show that this methodology is beneficial and attractive in making the quality improvement process more efficient and in providing support to managerial decisions. The results also suggest that the association rules technique works well in classifying quality problems that occur together. And the cluster (segmentation) analysis can be an alternative to the association rules technique to be implemented for quality problems grouping in the future.

2.4 Data Mining Commercial Packages Review

Nowadays, the data mining projects that succeed should provide not only the high accuracy and performance of the data mining algorithm, but also some functions listed below:

- Ability to handle large data sets without interrupting the databases. Because the calculation behind data mining is so complicated, it might cause the database system to crash.
- It should be easy to integrate the software within the existing system, including hardware requirements, software, databases, and overall integration.
- The data mining system should be able to transfer the knowledge into an understandable format. Sometimes data mining results need expert knowledge to interpret them. Many decision makers lack this knowledge; therefore, easy to understand data mining results are required to make timely decisions.
- Automation: the data mining system should be able to generate reports automatically.

- The system needs to be flexible, scalable, and user-friendly.

Some commercial products we have investigated and evaluated are listed below:

- Oracle Product: Oracle Data Miner is the graphical user interface for Oracle Data Mining. It covers all the data mining algorithms that can help to complete tasks in various industries. In addition, the newly released edition includes the Oracle Data Miner PL/SQL Code Generator. The code generator generates PL/SQL code for all the data mining activities, including data preparation, data transformation, and model operations. It adds more flexibility for Oracle PL/SQL users. The main advantage of this software is that it is easy to integrate into the database system. With Oracle Data Miner, the data never leaves the database; all data movement is eliminated. Therefore, it enhances the security of the database (Oracle, 2008).
- SAS product: With SAS Enterprise Miner, organizations are able to build a model based on real-world data collected from a variety of sources. They can use the model to produce patterns in the information that can support decision making and predict new business opportunities. SAS is one of the most recognized data analysis programs, and it has been used in many areas, such as telecommunications and manufacturing. In the telecommunications market, the SAS data mining solution helps organizations to attract and retain profitable customers consistently and to build customer profiles in order to make better decisions that minimize churn while maximizing profits. SAS delivers an integrated approach to support customer retention, customer segmentation, and campaign management. SAS also helps manufacturers to improve supply chain management, manufacturing

processes, and customer relationship management. SAS provides a complete set of statistical analysis tools and includes a business intelligence reporting system as well, which puts everything in a user-friendly GUI. It also provides a framework for data mining called SEMMA methodology. Overall, SAS is capable of satisfying an organization's information needs from various perspectives by turning data into intelligence (SAS, 2008).

- SPSS product: SPSS Predictive Analytics Solutions and Data Mining Solutions which help telecommunications operators to understand what their most desirable customers believe is the most attractive offer that will lead to greater customer retention. The package includes Analytical Customer Relationship Management (CRM), Fraud Detection, Marketing and Sales Analysis, and Segmentation Management. The patterns uncovered using data mining help organizations make better and timelier decisions. SPSS data mining solutions and services have enabled hundreds of organizations to achieve remarkable results in many areas, such as boosting sales and reducing marketing costs, improving the response rate of direct mail campaigns, and so on. With the visual data mining interface, the end users can quickly interact with the data and begin discovering patterns that can be used to support the decision making process (SPSS, 2008).
- StatSoft Ltd product: Statistica Data Miner has the functionalities to help a telecommunications company to prevent client churn and market segmentation and to identify fraud and other irregularities in the use of telecommunications services. In addition, Statistica provides a full range of statistical tools for data analysis

(StatSoft, 2008).

These are the related products offered by different companies. Usually, the companies advertise the characteristics of their commercial products, but not the details of their techniques.

2.5 Summary

Data mining in various forms is becoming a major component of business operations. Almost every business process today involves some forms of data mining. Even though data mining has become critical to businesses, most of the academic research in data mining is conducted on mostly publicly available data sources. This is mainly due to two reasons: (1) The difficulty academic researchers face in getting access to large, new, and interesting sources of data. (2) The limited access researches have to domain experts who can provide a practical perspective on existing problems and provide a new set of research problems. There is limited interaction between industry practitioners and academic researchers working on related problems in similar domains. This research thesis, however, is based on real-world problems, identified by decision makers from a telecommunications company. To the best of our knowledge, there is no existing research published on the exact same topic as this research thesis. Thus, this thesis breaks new ground in this field. The uniqueness of the data set, the business problems, and the methodologies we proposed in this thesis will be discussed in detail in the later chapters.

Chapter 3 System Architecture and Methodology

In Chapter 2, we reviewed some popular data mining methodologies in the research community and their applications in the telecommunications market and in quality assurance. In this thesis, both statistical analysis and data mining techniques are used as part of the knowledge discovery process. Therefore, in this chapter, we introduce the business problems in the QA group, present the proposed system architecture, and discuss in detail the methodologies we employ in this thesis research. Data mining models, such as decision trees, association rules, and other statistic methods provided in SAS Enterprise Miner, have been selected as the most suitable methods for our data and business problems.

3.1 System Architecture

As we briefly discussed in Chapter 1, the mission of the Quality Assurance group is to maintain the highest level of customer satisfaction through continuous improvements in quality in compliance with ISO 9001, which is one of the standards in the ISO 9000 family of quality management systems. Quality Assurance, or QA for short, refers to a planned and systematic process, which provides confidence in a product's effectiveness. The QA process covers activities including quality control, failure testing, and statistical control. Quality control is involved in developing systems to ensure products are designed and produced to meet or exceed customer requirements. These systems are often developed in conjunction with other business and engineering disciplines using a cross-functional approach. Failure testing is a valuable process to perform on consumer products, which usually involves operating a product, often under stressful conditions (e.g., an increasing vibration) until a failure occurs, thus exposes many unanticipated weaknesses in the product, and the data is used to drive engineering and manufacturing process improvements. Traditional statistical process controls in manufacturing operations usually proceed by randomly sampling and testing a fraction of the output. It aims to bring the organization to the Six Sigma level of quality.

As Figure 1.4 shows, the QA group plays a very important role in the information chain, and the operation of the QA group is to help support the organization in reaching its goals. The QA group provides feedback on quality issues to manufacturing and engineering design departments in order to drive manufacturing and engineering process improvements. As the organization continually introduces new products and expands production capacities, the QA group needs to make recommendations on engineering design to ensure that the quality of the products from various manufacturing processes are consistent and satisfactory.

The worldwide carrier partnership is a key factor in the company's success. If the quality of the products is assured, the carriers will become likely to carry the products and will support stable average selling prices. A committed relationship with the carriers benefits the organization the most by increasing profit and growing subscriber bases. The QA group is responsible for showing the confidence of the carriers about the product quality. Besides the daily operations of the QA group, including reliability and failure testing, statistical control, and quality control, the QA group communicates with some major carriers on a regular basis. Reports are sent to the carriers regularly, giving updated information on key components such as activation, deactivation, age, time delay, and top failure modes analysis. These components are also major concerns of decision makers in the QA group.

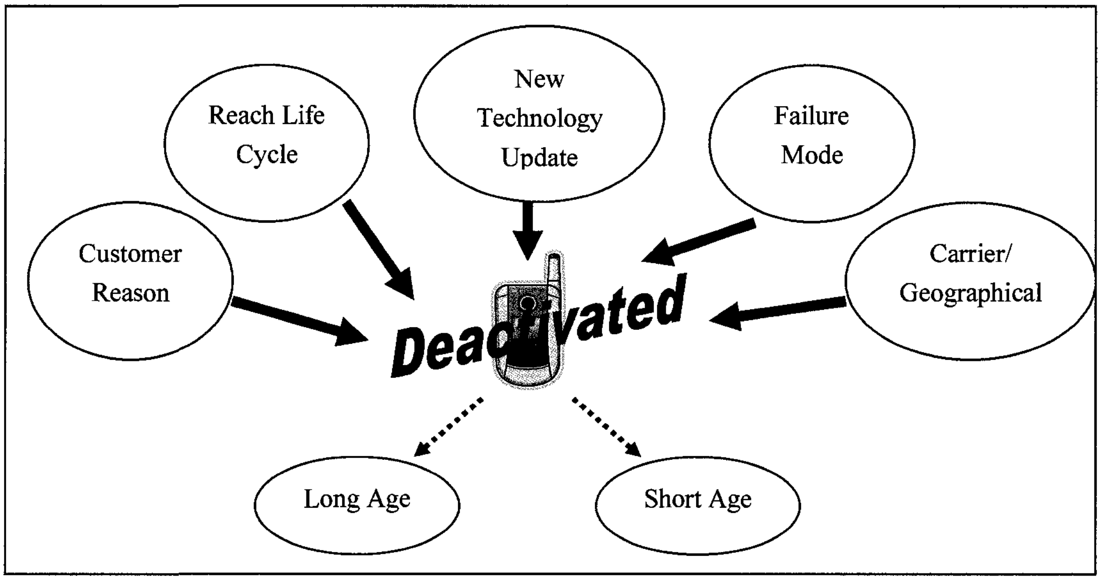


Figure 3.1 Deactivation Influential Factors Identification

As Figure 3.1 shows, the deactivation of a device may be caused by many factors. For example, customer may stop using the device for personal reasons. Unfortunately, there is no customer information available to analyze this factor; therefore, the factor has to be left out. Failure mode is related to product quality. If a device is not working properly, a customer will very likely stop using it. Both the carriers and the organization like to see lower deactivation rates. They are very interested to find out which factors have the influences on the deactivation of devices the most, whether they are carrier, location, or failure modes. Because age is a measure of how long the customer sticks with this device. There are number of devices deactivated with a very short age. It is very interesting to find out which factors influence age and how time in use changes by carrier, location, and so forth.

In addition to the factors discussed above, the organization is also interested in discovering which failure modes have a higher likelihood of appearing together. Furthermore, the organization wants to find out which carrier or location has shorter time delays. Time delay is measured by the time difference between a device's manufacturing date and its activation date. From an organizational point of view, companies like to see shorter time delay; any factor that speeds up a product's entry into the market will maximize profit.

Currently, data is accumulated very quickly, as many devices are activated and deactivated. As tons of thousands of devices are being sold, activated, and deactivated daily, data is accumulated quickly on the product. The management realizes that the data may provide valuable information and knowledge about the product. The database and statistical techniques used to perform quality control and statistical control currently are efficient in maintaining daily operations and in providing sufficient information to the QA group, but the existing tools cannot handle such large data sets. The statistical analysis tools currently used in the QA group work well when a data set is “small,” “clean,” and “ready.” With objectives clearly predefined, these statistical analyses are more focused on monitoring numeric criteria in the data, but they overlook the interrelationship of variables. Data mining can be a possible solution to a “large,” “noisy,” and “incomplete” data set because its ability of handling such data set. Most importantly, data mining is able to discover previously unknown patterns or knowledge to enhance the organization’s competitive advantage. Unfortunately, there is no data mining project currently initialized in the QA group. While other departments do attempt to predict customer churn, they do not use a data mining approach. Therefore, this thesis research, which seeks to pioneer a new data mining approach, aims to integrate the key concerns and components into one system and to resolve a key business concern using a data mining approach.

The solution system architecture is presented in Figure 3.2, which is proposed for the deactivation prediction. There are six analytical components, which are listed in the right column in Figure 3.2. The first component is the Deactivation Analysis, which aims to identify influential factors to deactivation. The analysis will mainly be based on the data sets from carrier, region, and the summarized information on RMA data. The second component is the Age Analysis, which utilizes the methods of decision tree, regression tree, and statistical analysis to determine the factors causing the short life-cycle of a device. The purpose of the third component, Failure Mode Association Analysis, is to discover which failure modes and their combinations are more likely to occur. The fourth component is the Failure Mode Classification Analysis; this analysis aims to identify the failure modes or failure mode combinations that directly cause deactivation. The fifth component is the Time to Fail Analysis. The purpose of this component is to compare the time to fail with the current warranty period, in order to determine the efficiency of the warranty. And the last component is to present how time delay changes by factors.

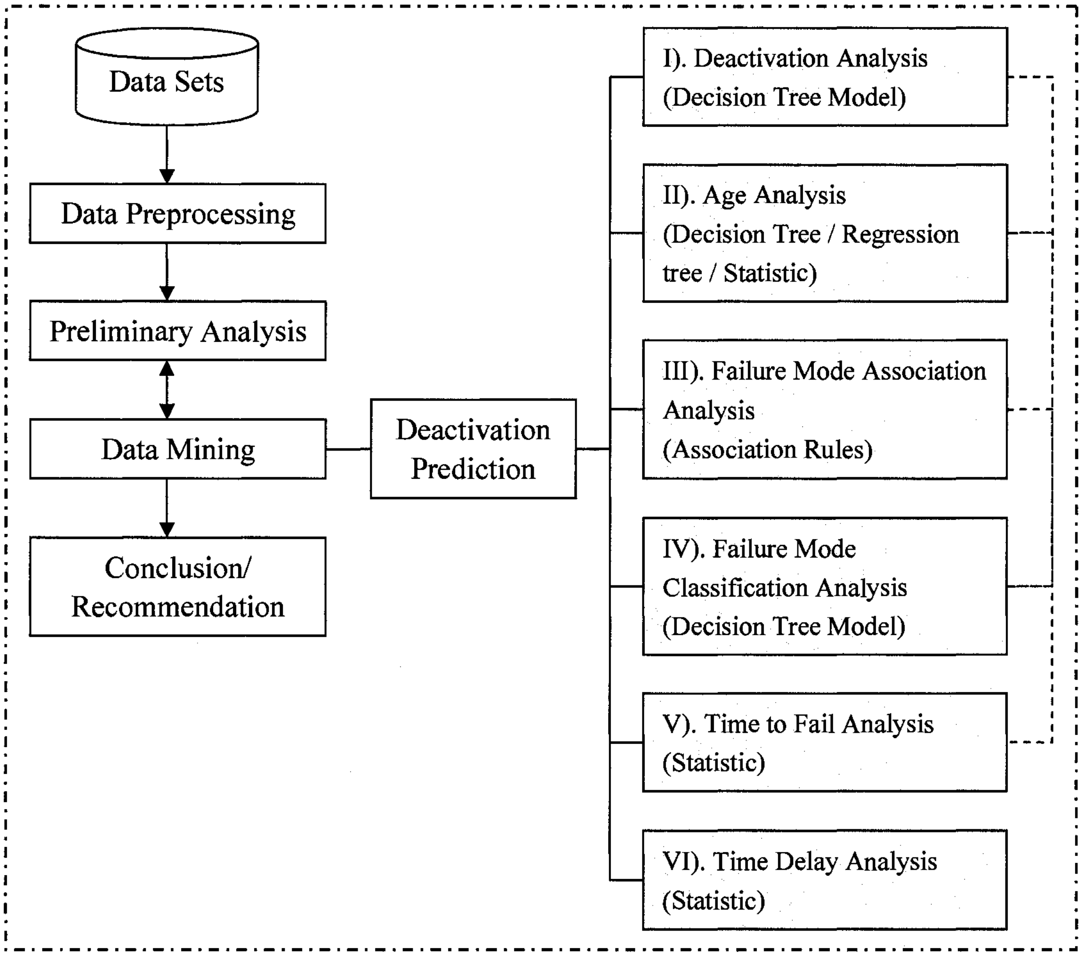

Data Sets
Data Preprocessing
Preliminary Analysis
Data Mining
Conclusion/ Recommendation
Deactivation Prediction
I). Deactivation Analysis (Decision Tree Model)
II). Age Analysis (Decision Tree / Regression tree / Statistic)
III). Failure Mode Association Analysis (Association Rules)
IV). Failure Mode Classification Analysis (Decision Tree Model)
V). Time to Fail Analysis (Statistic)
VI). Time Delay Analysis (Statistic)

Figure 3.2 The System Architecture of Deactivation Prediction

A major challenge faced in this study is to mine the highly specialized data set. Due to the features in the data sets from real-world practice, the variables (attributes) contained in the data sets are relatively limited in comparison with the number of devices (records). There are a dozen variables, while there are millions of records. It makes the data set "long" in one dimension and "thin" in the other. There are also some strong built-in patterns in the data, which will be discussed in the next chapter. Consequently, the mining results may be affected by the data structure to some degree. Nevertheless, the research generates meaningful results that coincide with common sense and with known knowledge in the business. The analytical procedure may offer some guidelines to management in the QA group for further data collection and analysis. In this chapter, data mining models used in the solution system are discussed. In chapter 4, the detailed data introduction and data preprocessing will be discussed.

3.2 Decision Tree

3.2.1 Decision Tree Method

The decision tree model is capable of classifying observations based on the values of nominal or binary targets. As shown in Figure 3.3, a tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. One rule is applied after another, resulting in a hierarchy of segments within segments. The hierarchy is called a tree, and each segment is called a node. The original segment contains the entire data set and is called the root node of the tree. A node with all its successors forms a branch of the node that created it. The final nodes are called leaves. For each leaf, a decision is made and applied to all observations in the leaf.

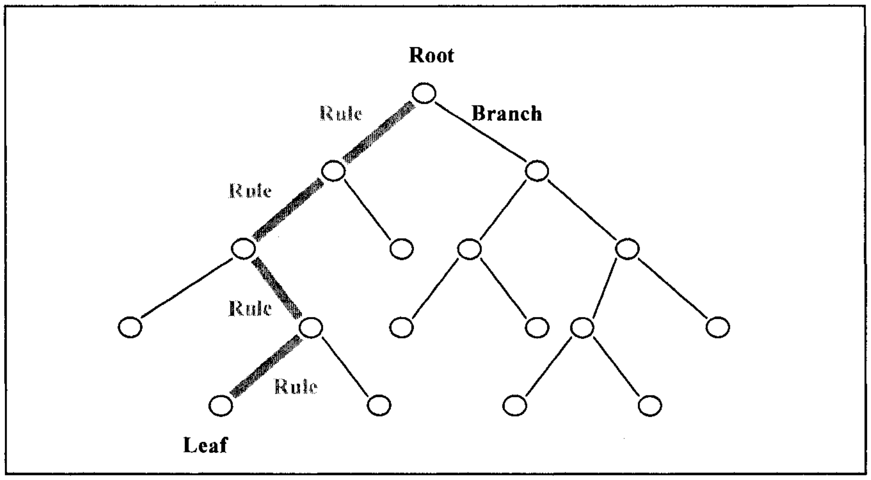


Figure 3.3 Decision Tree Diagram

There are three algorithms provided in SAS Enterprise Miner for generating decision tree models. They are CHAID, CART, and C4.5. There are three splitting criteria, Chi-Squared test, Gini Reduction and Entropy Reduction, which are used in each of the three algorithms respectively as their splitting methods.

1) Chi-Squared Test

Pearson's Chi-Square is used to assess two types of comparison: tests of goodness of fit and tests of independence. A test of goodness of fit establishes whether or not an observed frequency distribution differs from a theoretical distribution (Pearson's chi-square test, n.d.).

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{3.1}$$

(Pearson's chi-square test, n.d.), where O_i = an observed frequency; E_i = an expected (theoretical) frequency, asserted by the null hypothesis; n = the number of possible outcomes of each event.

2) Entropy Reduction

Entropy, a measure from information theory, characterizes the purity, or homogeneity, of an arbitrary collection of examples (Hamilton et al., 2003). So, if there is a group of samples, we want to classify them into two classes, Yes or No. If the group of samples is all classified as Yes, then this is a very "pure" group. And if the group is fifty-fifty Yes and No, then the group is very impure. The higher the entropy value of a group, the more impure it is.

For a random variable X with n outcomes $\{X_i : i = 1, \ldots, n\}$, information entropy, a measure of uncertainty and denoted by H(X), is defined as

$$H(X) := -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \tag{3.2}$$

(Entropy, n.d.), where $p(x_i)$ is the probability mass function of outcome x_i. For example, if p_1 is the proportion of the group that is Yes and p_2 is the proportion that is No. If p_1 and p_2 equal 0.5, then the entropy is at its maximum of 1. If either p_1 or p_2 equal 1, then the group is completely pure and the entropy is at its minimum of 0.

Ideally, we want all pure groups. When we split a group of samples, we want the remaining groups to be “purer,” so the entropy must be reduced. In determining which attribute to use to split the samples, the attribute that reduces the entropy the most will be used. The reduction of entropy is called information gain. If we want to split group S based on attribute A, then the information gain is:

$$InformationGain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{3.3}$$

(Mous, 2005). The formula aggregates over the domain of attribute A. The information gain depends not only on the entropy of a new node, but also on how many samples there are in that new node.

3) Gini Reduction

The Gini coefficient is a measure of statistical dispersion. It is defined as a ratio with values between 0 and 1. A low Gini coefficient indicates a more equal distribution, while a high Gini coefficient indicates a more unequal distribution. 0 corresponds to prefect equality and 1 corresponds to perfect inequality. The Gini index is the Gini coefficient expressed as a percentage. The Gini coefficient G is a summary statistic of the Lorenz curve. If the data is ordered by the increasing size of individuals, G is given by

$$G = \frac{\sum_{i=1}^{n} (2i - n - 1) x_i'}{n^2 \mu} \tag{3.4}$$

(Damgaard, n.d.).

Generally, the best split is the one that does the best job of separating the data into groups where a single class predominates in each group. So the best split is one that increases the purity of the subsets by the greatest amount, and a good split also creates nodes of similar size or at least does not create very small nodes. In other words, the splitting method decides which input variable returns the highest purity and lowest variance in child nodes.

3.2.2 Regression Tree Method

1) F-test

Rather than testing each β individually, the global test that encompasses all β's is the following overall hypothesis (Testing Utility of Model, 2004):

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

H_a: at least one $\beta_j \neq 0$

The test statistic to test this hypothesis is called F-statistic and is calculated as:

$$F = \frac{(SS_{yy} - SSE)/k}{MSE} = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} \tag{3.5}$$

Where

Numerator d.f. = k

Denominator d.f = n – (k+1)

(Testing Utility of Model, 2004). The F-statistic is the ratio of the explained variability (as reflected by R^2) to the unexplained variability (as reflected by 1- R^2), each divided by the corresponding degrees of freedom. And the larger the F-statistic, the more useful the model is. If we fail to reject H_0, it means that there is no evidence that any of the predictors are linearly associated to the response. If we reject H_0, it means that at least one of the predictors is linearly associated to the response, but all we know is that one of the predictors is associated to y, but we do not know which ones.

SAS computes the F-statistic as:

$$F = \frac{Model \quad Mean \quad Square}{Error \quad Mean \quad Square} \tag{3.6}$$

(Testing Utility of Model, 2004). The degrees of freedom are: Numerator d.f. = Model d.f. = k; Denominator d.f. = Error d.f. = $n - (k+1)$.

2) Variance Reduction

In mathematics, variance reduction is a procedure used to increase the precision of the estimates that can be obtained for a given number of iterations. Every random variable that the simulation outputs is associated with a variance that limits the precision of the simulation results. In order to make a simulation statistically efficient, that is, to obtain greater precision and smaller confidence intervals for the output random variable of interest, variance reduction techniques can be used (Variance Reduction, n.d.). Therefore, in data mining, the variance reduction splitting criterion is applied to grow the branches in order to minimize the total variance.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{3.7}$$

(Variance Reduction, n.d.), where $\bar{x}$ is the population mean.

3.2.3 Important Performance Measures

Classification accuracy is an important measurement of a model's performance. As Table 3.1 shows, a confusion matrix (Confusion Matrix, n.d.) is used for evaluating the classification model. Out of the four possible outcomes, the true positive and the true negative are correct classifications. A false positive, however, is a situation where the outcome is incorrectly predicted as "yes" when it is in fact "no," while a false negative is a situation where the outcome is incorrectly predicted as "no" when it is in fact "yes." In the confusion matrix, each element represents the number of test instances for which the actual class is the row and the predicted class is the column; good results correspond to large numbers down the main diagonal and small, ideally zero, off-diagonal elements.

		Predicted Class		
		yes	no	
Actual Class	yes	True Positive	False Negative	← Type II error
	no	False Positive	True Negative	
		↑ Type I error		

Table 3.1 Confusion Matrix

We want introduce two terms, "misclassification rate" and "sensitivity," which are also two major evaluation criteria we will use in our data mining result evaluation in Chapter 5. Misclassification rate indicates the percentage of false negatives in the model. Sensitivity indicates the percentage of true positives captured by the model. An optimum model would be the one that has the highest sensitivity values and the lowest misclassification rate.

$$Sensitivity = \frac{number \quad of \quad True \quad Positives}{number \quad of \quad True \quad Positives + number \quad of \quad False \quad Negatives} \tag{3.8}$$

$$Misclassification = \frac{number \quad of \quad False \quad Negatives}{number \quad of \quad True \quad Positives + number \quad of \quad False \quad Positives} \tag{3.9}$$

In statistics, the terms Type I error (false positive) and Type II error (false negative) are used to describe possible errors made in a statistical decision process. In addition, decision rules generated from decision trees are also an important measurement. To generate rules, one must trace each path in the decision tree from root node to leaf node, recording the test outcomes as antecedents and using the leaf-node classification as the consequent.

Overall, as stated in Section 3.1 and Chapter 1, the objectives of the deactivation analysis (Component 1), the age analysis (Component 2), and the failure mode classification (Component 4) have been, respectively, the development of models for predicting the deactivation of the devices, distinguishing the long and short ages of the devices, and classifying the devices on the failure mode data. Therefore, the data mining task for these problems is defined as a classification problem, since the ultimate goal is to classify each device, or each returned device, as deactivated or still in use and as deactivated with longer life or shorter life.

Because service fees are an important source of revenue, the company wants to detect deactivated devices and devices with short life as well as determine the root causes of deactivation and short age. According to these purposes, deactivation prediction and analysis makes sense in business terms if it can be of use for the QA group and for the company.

The major advantage of the decision tree over other modeling techniques is that it produces a model that may represent interpretable rules or logic statements. The explanation capability that exists for trees producing axis parallel decision surfaces is an important feature. Besides, classification can be performed without complicated computations, and the technique can be used for both continuous and categorical variables. Furthermore, decision tree model results provide clear information on the importance of significant factors for prediction or classification.

There are many successful applications of the decision tree model in the telecommunications industry and in the QA environment, as shown in Chapter 2. And according to our data mining goals and the features of the decision tree model, we choose to use the decision tree model to perform the classifications and predictions in our system.

3.3 Association Rules

Association discovery is the identification of items that occur together in a given event or record. It aims to extract interesting correlations and frequent patterns, associations, or casual structures among sets of items in transaction databases or other data repositories. The association rules are expressed as "if item A is part of an event, then item B is also part of the event X percent of the time." Some software allows for the sequence discovery, which goes one step further than association discovery by taking into account the ordering of the relationships among items. For example, rule A => B implies that event B occurs after event A occurs.

There are two important basic measures for association rules: support and confidence. All the association rules generated from association rules mining need to satisfy the predefined minimum support and confidence from a given database (Kotsiantis & Kanellopoulos, 2006). Since the database is large, usually thresholds of support and confidence aim to drop those rules that are not very interesting or useful. The level of support is how frequently the combination occurs in the database. Support of an association rule is defined as the percentage of records that contain $A \cup B$ to the total number of records in the database.

$$\text{Support (A=>B)} = P(A \cup B) = \frac{count(A \cup B)}{count(Total)} \qquad (3.10)$$

Confidence of an association rule is defined as the percentage of transactions that contain $A \cup B$ to the total number of records that contain A. Confidence is a measure of the strength of association rules. Suppose the confidence of the association rule A=>B is 80%. This means that 80% of the transactions that contain A also contain B. So the confidence can be considered the conditional probability P(B|A).

Confidence (A=>B) = P(B|A) = $\frac{count(A \cup B)}{count(A)}$ (3.11)

(Dunham et al., 2001).

There is another important measurement called lift, which is equal to the confidence divided by the expected confidence, where expected confidence is equal to the number of consequent transactions divided by the total number of transactions (i.e., the confidence the rule would have if the occurrence of condition and decision values are statistically independent). Lift is a factor by which the likelihood of consequence increases given an antecedent. The lift is larger than 1 if the association between A and B is due to more than just chance. Lift corresponds to a positive correlation between the events A and B.

Theoretically, if the number of item sets grows too quickly, the system can run out of disk or memory resources. To avoid this problem, the support level can be set to a higher number in order to reduce the item sets to a more manageable number. The maximum number of items in an association determines the maximum size of the item set to be considered. Normally, we use the evaluation criteria to evaluate how strong the association rules are, but there are some common problems with association rules mining: the algorithm may generate an extremely large number of association rules and not all of the discovered strong association rules (i.e., passing the minimum support and minimum confidence thresholds) may be interesting enough to present. Therefore, a combination of expert knowledge as well as personal judgment is necessary in order to evaluate the association rules.

3.4 SAS Enterprise Miner

All the models for this study are developed by SAS Enterprise Miner, version 4.3. The framework for data mining is the SEMMA methodology, as defined by the SAS Institute. SEMMA (SAS Institute, 2004) is simply an acronym for "Sample, Explore, Modify, Model, and Assess." This logical superstructure provides users with a comprehensive method by which individual data mining project can be developed and maintained. Enterprise Miner is the first and only data mining solution that addresses the entire data mining process, through an intuitive point-and-click graphical user interface. However, not all data mining projects need to follow each step of the SEMMA methodology; the tools in Enterprise Miner software give users the freedom to deviate from the process to meet their needs, but the methodology does give users a scientific, structured way of conceptualizing, creating, and evaluating data mining projects.

As Figure 3.4 shows, the nodes of Enterprise Miner are organized into the categories of the SEMMA data mining process. Beginning with a statistically representative sample of your data, this methodology makes it easier to apply exploratory statistical and visualization techniques in order to select and transform the most significant predictive variables, to model the variables to predict outcomes, and to confirm a model's accuracy. Furthermore, the node is a key concept in Enterprise Miner; most of the time we interact with the program by dragging and dropping, right-clicking, or double-clicking the node that corresponds to a particular task, and we view the nodes by clicking the Tools tab at the bottom of the left window pane.

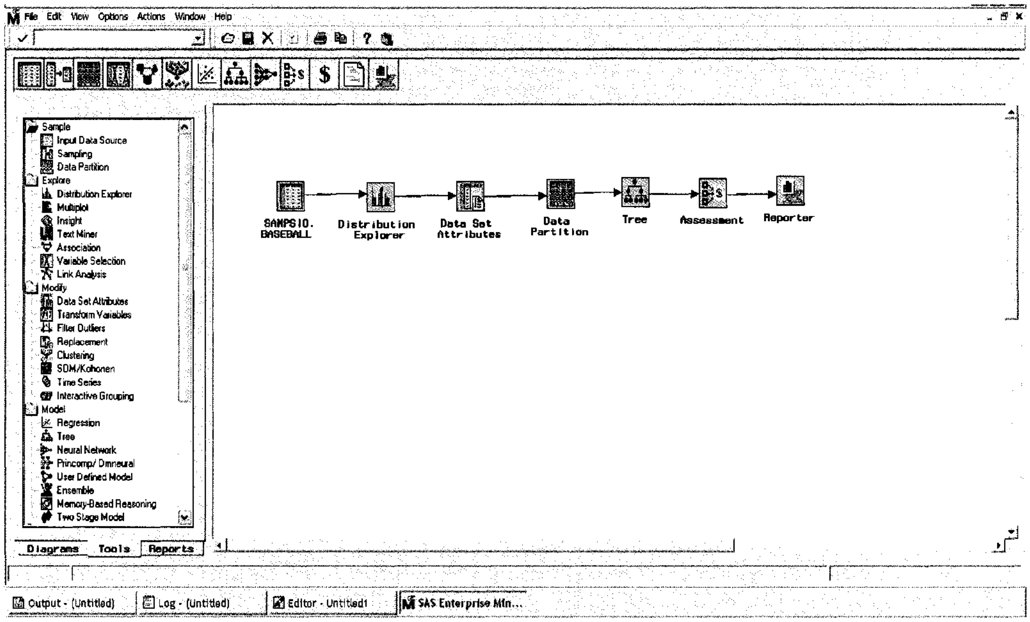


Figure 3.4 SAS Enterprise Miner GUI

3.5 Other Methodologies

Oracle PL/SQL is Oracle Corporation's proprietary procedural extension to the SQL database language used in the Oracle database (PL/SQL, n.d.). PL/SQL is a normal programming language that includes all the features of most other programming languages; in addition, it has the ability to integrate easily with SQL. The data manipulation is faster in PL/SQL than in SQL or Java or other programming languages within an Oracle database. And it is easier to use than other programming languages as well. The key strength of PL/SQL is its tight integration with the Oracle database. Because all the data of the QA group accumulated is stored in the Oracle database in the QA group, PL/SQL has become a major tool for data retrieval and data manipulation. In this thesis research, we apply Oracle PL/SQL to retrieve and integrate the data; it proves that PL/SQL is an efficient tool that works especially well with Oracle databases. The detailed PL/SQL queries used to retrieve, derive, and integrate data will be introduced in the next chapter.

Microsoft Access is a relational database management system from Microsoft that combines a relational database engine with a GUI and software development tools. It can be used to build simple applications. Access normally is used by small businesses, by departments within large corporations, and by individual programmers to handle the creation and manipulation of data. One of the benefits of Access is its relative compatibility with SQL. SQL queries may be viewed and edited as normal SQL statements. In addition, Access allows for relatively quick development because of the friendly GUI design tools and the high level of integration between the GUI design and data objects. All database tables, queries, forms, and reports are stored in the database. Access can be applied to small projects, but the program scales poorly to larger projects with large data sets or many users. In this research, after the data is retrieved from the Oracle database, we integrate those data into a few different data files and import those data files into an Access database. The experiment shows that Access is a proficient tool to handle comparably smaller data sets. It covers all the basic functionalities of SQL.

Therefore, it is a useful tool, especially to perform the preliminary analysis in this thesis research.

3.6 Summary

The methodologies reviewed in this chapter might be the best candidates to serve our goals and to solve the business problems in this thesis research. In next chapter, we will discuss the data preprocessing and preliminary analysis in detail. Data preprocessing is one step before we apply SAS Enterprise Miner to generate data mining models, and it is a very important step. Oracle PL/SQL is the major tool we used to complete the data preprocessing part, and we will perform preliminary analysis with the help of MS Excel and Access in the next chapter.

Chapter 4 Data Preparation and Preliminary Analysis

The real world data is noisy, inconsistent, and large. Therefore, the data is needed to be preprocessed in order to improve its quality. Data preparation includes data cleaning, data transformation, and data integration, which is discussed in this chapter. The preliminary analysis explores large volumes of data graphically, uncovers patterns and trends, and reveals extreme values in the data sets. With the analysis, we expect to explore the relationships between the attributes, to pre-analyze the potential influential factors, and to explore the data structures. We will discuss the findings from our preliminary analysis in this chapter as well.

4.1 Data Preparation

As Figure 1.5 shows, the data in the QA database is imported from three major sources. The originally selected data, or attributes, from the QA database can be categorized into three categories as follows:

1. Manufacturing related attributes: mfg_date and mfg_loc. These attributes contain information on the manufacturing date and manufacturing location of devices. Each device has a unique serial number, which is the primary key for many data tables in the database. The unique serial number joins tables, so it helps in obtaining additional information for each device across the tables.
2. Activation related attributes: act_date, Region, Carrier, Age, and time_delay.

3. RMA related attributes: had_rma, num_fms, num_returns, and due_to_rma. These attributes contain information on the Return Material Authorization.

The attributes are classified into one of the three categories according to their nature. Due to the limited access to information outside the QA database, we do not have customer related data, only product related data.

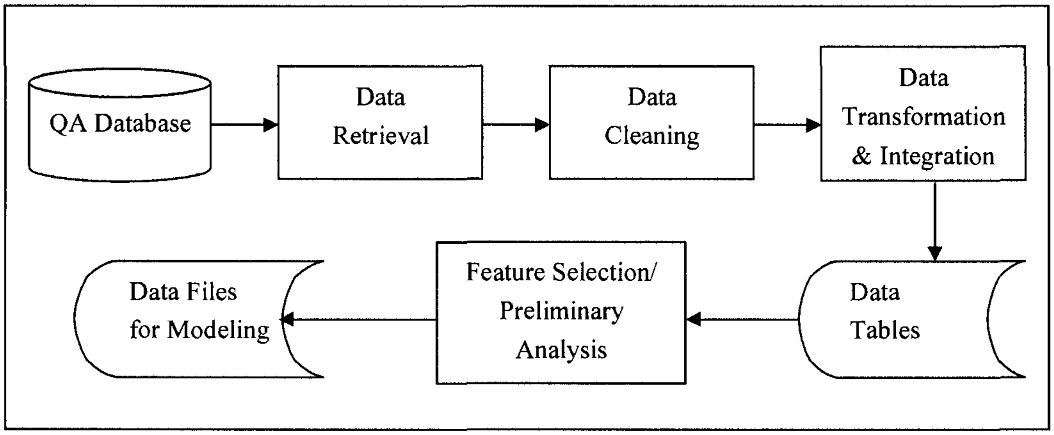

QA Database
Data Retrieval
Data Cleaning
Data Transformation & Integration
Data Tables
Feature Selection/ Preliminary Analysis
Data Files for Modeling

Figure 4.1 Data Preprocessing Design

Although the information stored in the QA database is well organized, there are missing data and inconsistent data. Data cleaning aims to fill in the missing data and to remove the inconsistent data. The process of data transformation and integration is generally used to generate derived attributes and to integrate the data. Oracle PL/SQL is a primary tool to retrieve the data from databases and to generate the derived attributes. Preliminary analysis and feature selection help us to better understand the data by identifying the trends and outliers and by removing irrelevant or redundant features.

In the original database, some attributes have missing data because not all data was collected. For example, the repair date of a device may not have been entered or not entered properly into the database because of human error. When data is imported into the QA database, a data checking procedure is applied that filters out records with missing data or fills in the missing data if necessary. If any records are missing the date, or if the missing date causes an odd derived attribute (such as age is negative number), we simply delete those records. If the attribute represents a time period measured in days, then the missing data is replaced with -1 because 0 is already used for certain situations. Sometimes "NA" is placed to represent the undetermined status of the devices.

The data sets we chose to analyze in this thesis are only for one product. The product is an old model, which was very popular when it was introduced to the market, and sold millions of units. Therefore, there are enough records available for data mining analysis. Because it is already out of production, the production life cycle of this product has been reached. When we scan the QA database, combined with expert knowledge from this department, we select some data that we think is most relevant to our data mining purposes and apply the Oracle PL/SQL to retrieve, transform, and integrate the selected data into three tables. The data tables in the QA database are relational, which are efficient for retrieval, updating, and maintenance of the data. Derived attributes are calculated fields not in the original data. They work better than the original data in certain circumstances, especially for the time series data. Due to the large number of activated devices and the small number of attributes, the tables are very "thin" and "long." The following are the introductions of each table, with notes on how to calculate each of the derived attributes, some assumptions and PL/SQL queries, and also a brief discussion of why we chose this data set to do data mining.

Attribute	Distinct Value	Definition	Measurement
1. mfg_date	5	When is the device manufactured?	Semiannually.
2. act_date	6	When is the device activated?	Semiannually.
3. Region	3	Where is the device activated?	Each value represents a region.
4. Carrier	5	By which carrier is the device activated?	Each value represents a carrier; all the other carriers are grouped into Others.
5. mfg_loc	2	Manufacturing location.	Each value represents a mfg location.
6. Age	Continuous Attribute	How many days is the device being used?	Days.
7. time_delay	Continuous Attribute	How many days in between the device's mfg_date and act_date?	Days.
8. had_rma	2 (Binary)	Has the device ever been returned for repair?	Value: 1, 0; 1 means yes, this device has been returned to repair at least once; 0 means no, this device has never been returned to repair.
9. due_to_rma	2 (Binary)	Is the device deactivated because of the failures (returned for repair)?	Value: 1, 0; 1 means yes, this device was deactivated because of the failure modes; 0 means no, the device was deactivated due to other reasons or the device was returned to repair but still in use.

10. num_fms	5	How many failure modes are found with this device?	Value: 1, 2, 3, 4, >=5; Number of fms larger than or equal to 5 are all grouped into >=5 category.
11.num_returns	3	How many times has this device been returned for repair?	Value: 1, 2, >=3; 1 means this device only had been returned once, (but we might find multiple failures). Number of returns larger than or equal to 3 are all grouped into >=3 category.
12.Deactivated	2 (Binary)	Is the device still in use or has it been deactivated?	Value: 1, 0; 1 means yes, this device has already been deactivated; 0 means no, this device is still in use.

Table 4.1 The Data Table for Deactivation Prediction, Age and Time Delay Analysis

The first data set, which contains 12 attributes, is used to perform deactivation analysis, age analysis, and time delay analysis (Table 4.1).

1. mfg_date: This attribute represents the time when a device was manufactured. It has 5 distinguished values, which are converted from the original record. For example, if a device was manufactured on March 20, in Year 2, then its mfg_date = YEAR2-A, that is, it was created on the first half of Year 2. The data set consists of devices manufactured in two and a half years.
2. act_date: This attribute represents the time when a device was activated (put in use), which has 6 distinguished values. Similar to mfg_date, for example, if a device is sold and activated on October 7, in Year 3, then act_date = YEAR3-B, which

represents the second half of Year 3. By definition, the value of act_date is always later than the value of mfg_date.

3. Age = deact_date – act_date. This attribute is a derived variable. It measures how many days a device has been used, where deact_date is the date when the device was last used by the customer. For this data analysis, June 30 of Year 4 is the cut-off date for all the records. As a result, if a device is still in use, then its deact_date is June 30, Year 4; otherwise, it is whenever the device was stopped being used.
4. time_delay = act_date – mfg_date. This attribute is a derived variable, which measures the number of days from the time a device was manufactured to the time the device was activated.
5. had_rma is a variable determined by its rma_issue_date. If a device has never been returned for repair, then its had_rma = 0; otherwise, had_rma = 1, which means the device was returned for repair at least once.
6. due_to_rma is a derived attribute, which indicates whether a deactivation was caused by a returned repair. If this is the case, then a deactivated device would have a deact_date occurred earlier than its last rma_issue_date. The value of due_to_rma is calculated by the PL/SQL query:

   WHEN deact_date < MAX (rma_issue_date) OVER (PARTITION BY serial_num) THEN 1, ELSE 0.
7. num_fms and num_returns are calculated by counting distinct failure mode descriptions and distinct RMA issue dates respectively.

8. Deactivated is determined by

WHEN ('Jun-30-Year4' - deact_date) > 60 then 1 ELSE 0

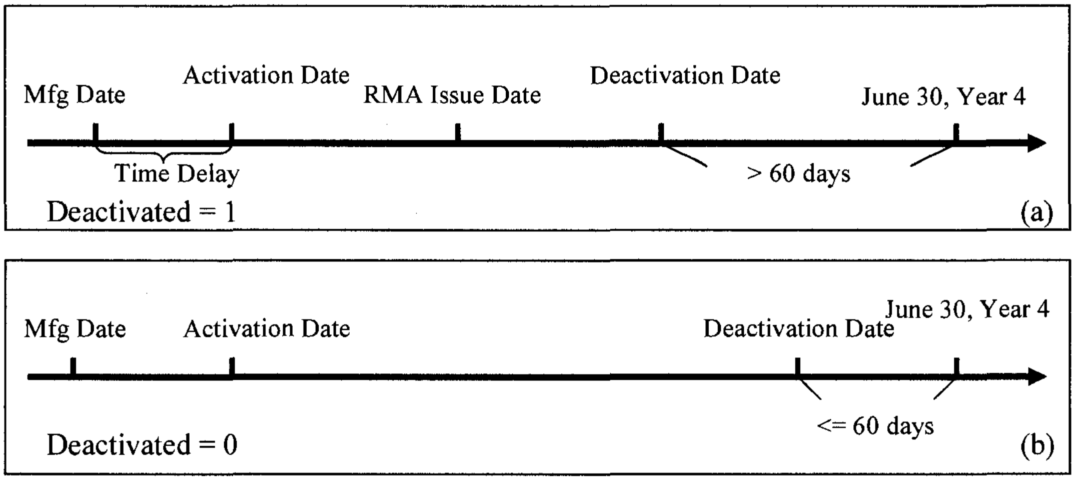


Figure 4.2 Time Line Illustration

Basically, the deactivation date is compared to June 30, Year 4, the cut-off date. If the deactivation date is more than 60 days before June 30, Year 4, we are confident that the device is deactivated. If the deactivation date is less than 60 days before June 30, Year 4, it might have been the case that the customers were traveling and that they simply stopped using the device for a period of time with the intention of continuing to use the device after they came back; or perhaps the device needed to be repaired, in which case the customer would have continued to use it after it was returned. 60 days is a reasonable amount of time to distinguish between deactivation and a temporarily discontinued device. We consider the devices deactivated within 60 days before June 30, Year 4 are not deactivated but temporarily discontinued. So when the deactivation date appears as June 30, Year 4, this means the device is still being used, but we simply use June 30, Year 4 as its deactivation date. The situations are illustrated in Figure 4.2.

As Figure 3.1 shows, carrier and geography could be one of the reasons causing the deactivation. Thus, we create Table 4.1 from the QA database, which we believe is the most relevant, and aim to determine the influential factors to deactivation from a carrier and geography perspective. Also, some devices were deactivated with short age or with long age. We want to determine the influential factors on age and how age distribution changes by carrier and region. We are also interested in discovering how time delay changes by those factors. All the RMA related attributes in Table 4.1 are derived variables. Only 10% of all the devices have been returned to RMA, and some of the devices were returned more than once. We believe the derived attributes work better. We also create two separate data tables that contain the detailed information on failure modes for different data mining purposes. In the experiment design on this data table, we simply want to discover if had_rma, due_to_rma, num_fms and num_returns have a correlation with the deactivation or age. We will exam the failure mode data in detail later.

Unlike Table 4.1, in which each record is for one device, in Table 4.2, each record is for one failure mode. For example, if one device returned to RMA had 3 different failures, 3 records were entered in Table 4.2. If this device was returned more than once, and if the second time 2 failures were found with the device, even if the second failures were the same as the first, those two later records would be entered in Table 4.2 as well.

Attribute	Distinct Value	Definition	Measurement
1. mfg_loc	2	Manufacturing location.	Each value represents a mfg location.
2. Carrier	5	By which carrier is the device activated?	Each value represents a carrier, all the other carriers grouped into Others.
3. time_to_fail	Continuous Attribute	Time between the activation date and the failure mode appear date.	Days.
4. fm_index	165	The failure mode (FM) description.	Each value represents a particular failure mode description.
5. cat_num	37	FM category.	Each value represents a failure mode category.
6. fail_cat	3	Failure category (Warranty failure or Non-warranty failure or Risk).	Each value represents a failure mode category.
7. g_cat	6	Failure category (customer abuse or functional failure or no fault found and so on).	Each value represents a failure mode category.
8. Deactivated	2 (binary)	Is this device still in use or has it been deactivated?	Value: 1, 0; 1 means yes, this device has already been deactivated; 0 means no, this device is still in use.

Table 4.2 The Data Table for Time to Fail Analysis

The variable time_to_fail, where time_to_fail = rma_issue_date – act_date, measures how many days lie between when the device was activated and when the failure mode appeared. In addition, fm_index is the specific failure mode description (e.g., battery connector broken/deformed). cat_num is the failure mode category. For example, “keypad not functioning,” “keypad cracked,” and “keypad worn out” all belong to one failure mode category, “keypad.”

Time to fail is an important factor in determining a warranty period. We want to pick some important failure modes from Table 4.2 in order to test whether or not the warranty period offered by the manufacturer is efficient. This way, we can make a practical recommendation to the manufacturer concerning warranty periods.

Serial_Num	FM1	FM2	...	FM165	Deactivated
1	0	0	...	0	0
2	1	0	...	0	1
3	0	1	...	0	0
4	0	0	...	1	1
5	0	1	...	0	1
6	1	0	...	0	0

Table 4.3 The Data Table for Failure Mode Association and Classification Analysis

1. Table 4.3 shows the format of the data set used for Failure Mode Analysis, which is also a cross-tabulation. The Serial_Num column contains the unique serial number for each device. FM1, FM2...FM165 represent the failure mode descriptions.
2. In the intersection of failure mode and serial number, 1 means this device had this particular failure mode, while 0 means this device never had this particular failure mode. How many times the failure mode occurred with this device is not considered; only the presence of the failure mode is relevant.
3. This table only contains the devices that have been returned to RMA.

4. "Deactivated" represents whether or not this device is already deactivated or still in use.
5. There are 165 different FMs in this table.

We create Table 4.3 mainly for two purposes. First, we want to discover the associations of failure modes. We want to find out which failure modes are more likely to occur together. This information will improve the engineering design and the RMA process and reduce the RMA exam time. Second, we want to discover which failure modes or combinations will directly cause deactivation. As Figure 3.1 shows, failure modes can be a reason for deactivation. The RMA related attributes in Table 4.1 are derived attributes. Thus, we want to investigate further which specific failure modes will cause deactivation.

4.2 Preliminary Analysis

The primary objective of the preliminary analysis is to examine each factor according to its distribution and its relationship with the target variable. The second purpose is to examine the interrelationships between paired attributes. The analysis aims to break the complicated problem into some manageable pieces in order to examine closely the potential influential factors and their combinations with the target variable. Additionally, the analysis can be used to discover trends and outliers, as well as serve as a manual feature selection for data mining. MS Access, Excel, and Pareto charts are the major tools used for the preliminary analysis.

4.2.1 Preliminary Analysis on Deactivation Prediction

The purpose of this study is to identify the factors that will possibly influence the deactivation of the devices. The devices are deactivated due to many reasons: the product reaches its physical life cycle; technology updated; user abuse; or the quality of the services and products (from carrier and product provider) and so forth. These factors can be broken down into many sub-factors. Some of the factors are related to customer behavior and unexpected circumstances, which certainly adds difficulty and uncertainty to the prediction. Here are the findings from the variables listed in Table 4.1.

1. *mfg_date*: The manufacturing quantity is normally distributed. There is no surprise that the devices created early are more likely to be deactivated, and the devices created later are more likely to be still in use, as shown in Figures 4.3 and 4.4.

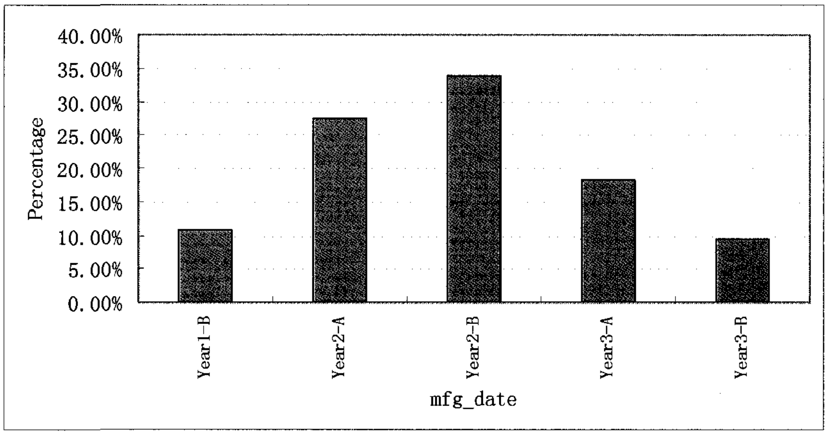


Figure 4.3 The Distribution of mfg_date

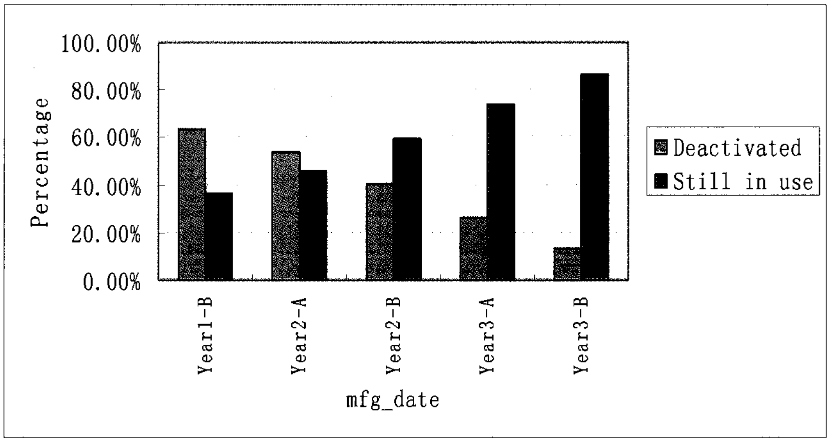


Figure 4.4 Percentage of Deactivation vs. mfg_date

The product was first produced at manufacturing location B, and then the production level decreased rapidly at this location. The production was mainly carried out at location A.

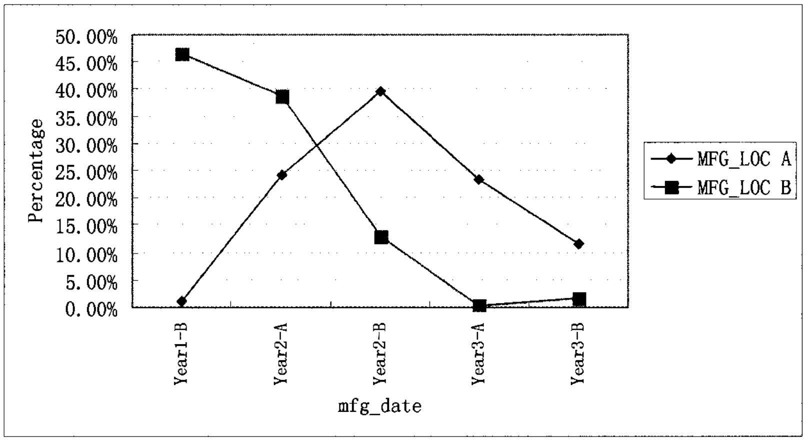


Figure 4.5 Relationships between mfg_date and mfg_loc

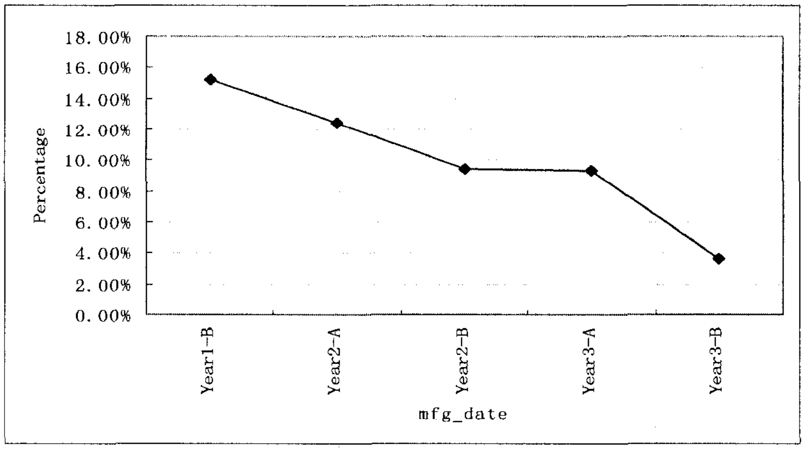


Figure 4.6 mfg_date vs. had_rma Rate

As shown in Figure 4.6, the rate of had_rma decreases significantly vs. mfg_date. 16% of the devices manufactured in YEAR1-B have been returned for repair, while only 4% of those manufactured in YEAR3-B have been returned. This makes sense because the longer a device is being used, the more likely some quality issues may occur. It is also possible that the manufacturing technology gets more mature and that the increasing quality of the production reduces the return rate.

2. *act_date:* Like mfg_date, the devices activated in early years are more likely to be deactivated now; the devices activated later are more likely to still be in use.

3. *Region:* The percentages of the market share of each carrier in the three different regions are not even. Carriers A and B are focused in Region 1, while Carriers C and D are focused in Region 2, and the other carriers are focused in Region 3. Therefore, Region and Carrier are highly correlated.

4. *Carrier:*

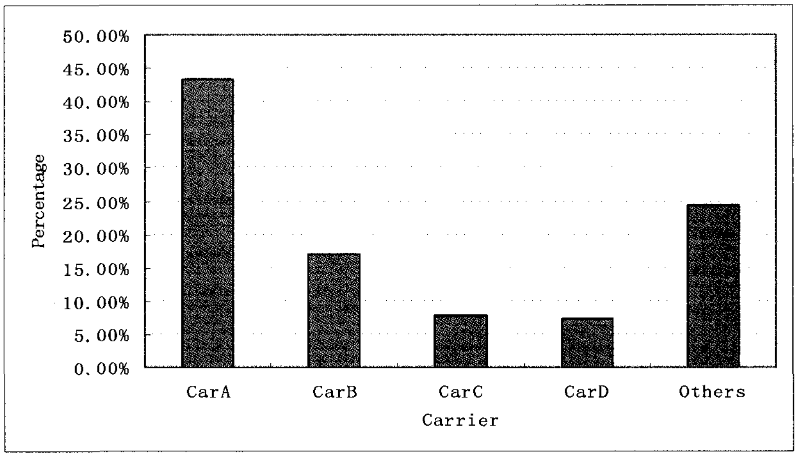


Figure 4.7 Carrier Distribution Diagram

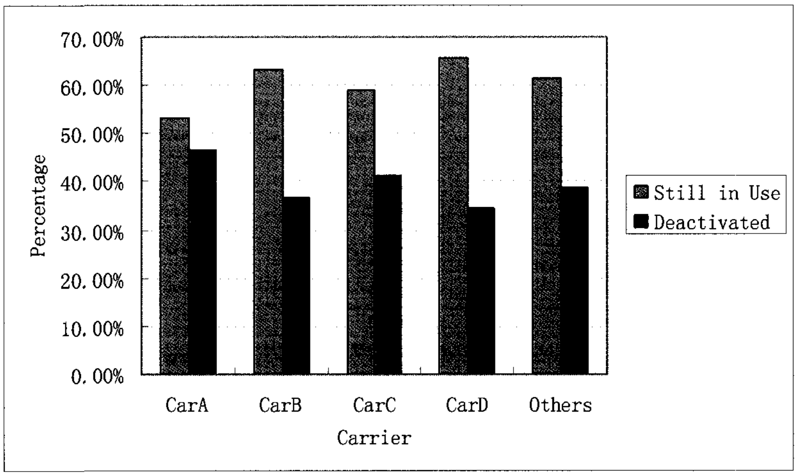


Figure 4.8 Carriers vs. Percentage of Deactivation and Still in Use

There are no significant differences in deactivation rates among the carriers. Higher percentage of deactivation does not necessarily indicate a quality issue; it might be caused by region, product launch date, or the other product line carried by the carrier. However, the average time delay for each carrier is significantly different. Carrier, then, may be a factor in determining how fast a product gets activated after it is produced.

5. *mfg_loc:* The product was manufactured in two locations. mfg_loc is highly correlated with mfg_date, as shown in Figure 4.5. We do not expect mfg_loc will be an important factor in predicting age or deactivation.

6. *Age:* This attribute is only calculated for devices that are already deactivated. If a device is still in use, then its age is indeterminate.

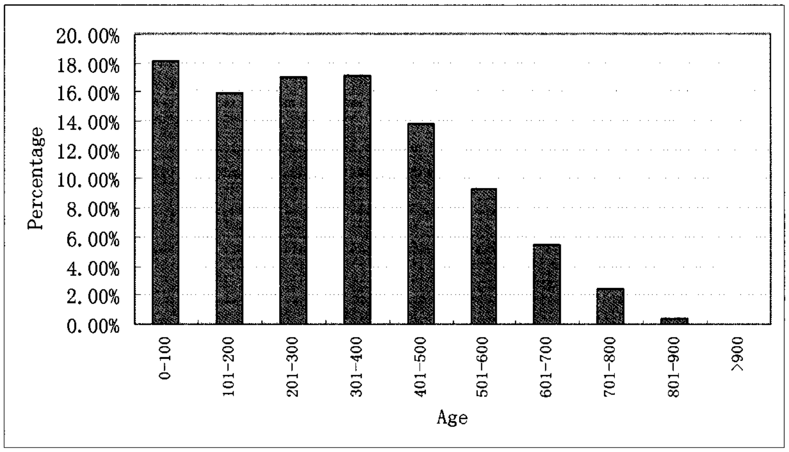


Figure 4.9 Age Distribution (Deactivated Devices)

For all the deactivated devices, the majority have an age between 0 to 400 days. Some of the devices, however, have an age between 400-700 days. There are very few devices used longer than 700 days. Thus, we have two extreme cases: the devices that were deactivated after a very short period of time and devices that were deactivated after a very long time. In the age analysis, we aim to determine which factors influence age, thereby discovering the root cause for short age values. Carrier, region, and failure mode might be the influential factors to age.

7. *Other Attributes:* 10% of all the devices were returned to RMA for repair. Among the devices returned to RMA, 67% are deactivated. Among those returned and deactivated devices, 70% are deactivated due to RMA, the rest are deactivated due to other reasons. The figures are illustrated in Figure 4.10.

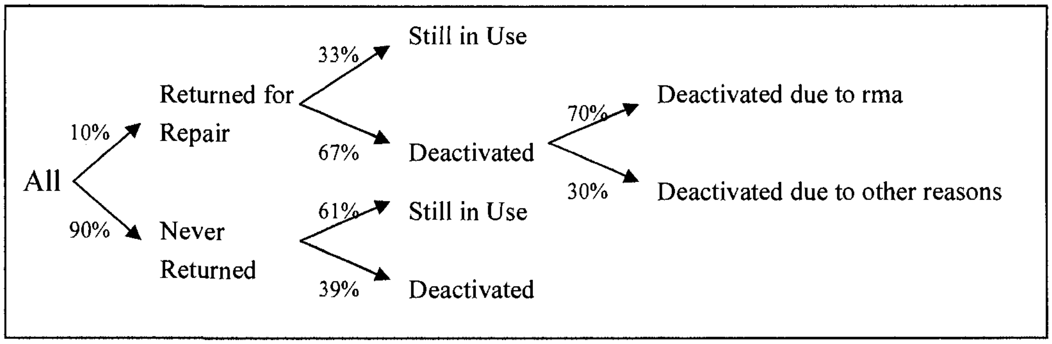


Figure 4.10 Tree Illustration of Data Structure

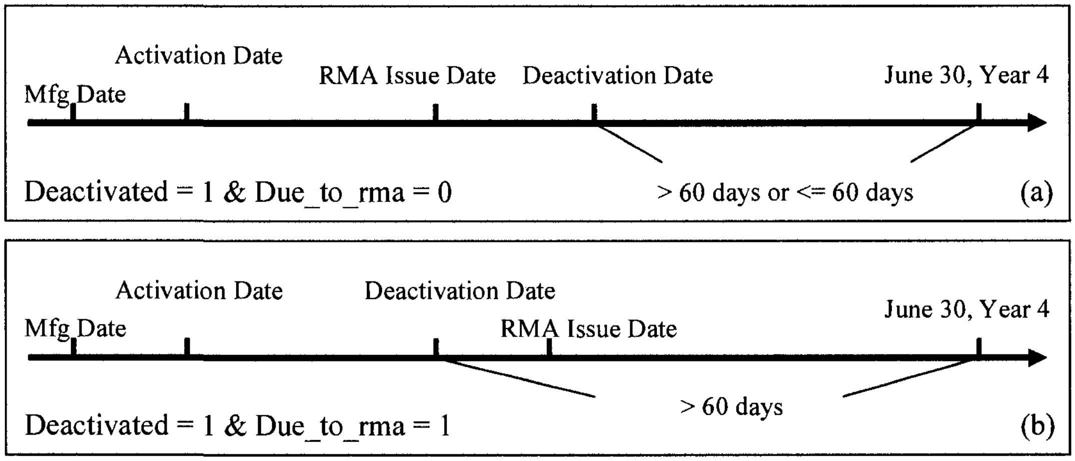


Figure 4.11 Time Line Illustration of due_to_rma

due_to_rma is defined as whether or not the device is deactivated because of the failures (return for repair) or other reasons. due_to_rma is determined by WHEN deact_date < MAX (rma_issue_date) OVER (PARTITION BY serial_num) THEN 1, ELSE 0. 1 means yes, the device was deactivated because of RMA; 0 means no, the device was deactivated due to other reasons. A device may be returned to repair more than once. If the deact_date date is before the latest rma_issue_date, then the device is no longer being used, possibly because of RMA.

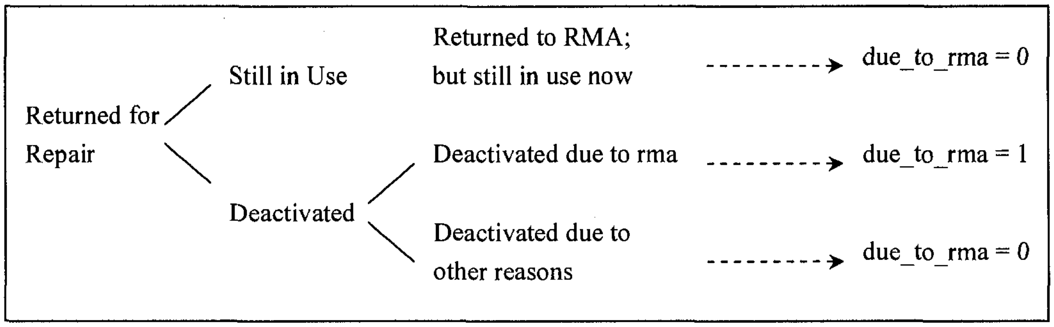


Figure 4.12 Tree Illustration of due_to_rma

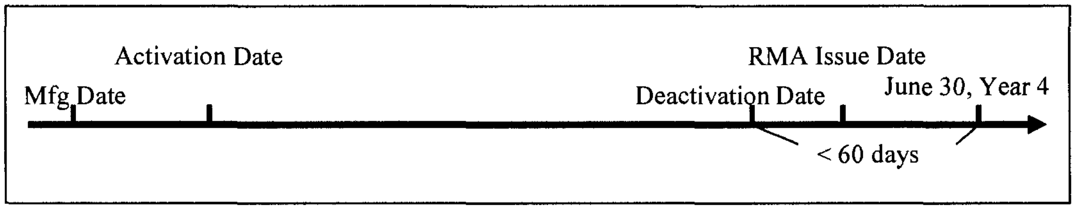


Figure 4.13 Time Line Illustration of due_to_rma (Exception)

due_to_rma is only valid for those devices that returned for repair, since only those devices that were returned to RMA have the RMA issue date. We assume that there are only three possible outcomes as shown in Figure 4.12. If the device returned to RMA, and deactivated because of RMA (Deactivation date < rma_issue_date), then due_to_rma = 1 (the device is deactivated due to RMA); if the device returned to RMA, and deactivated possibly due to other reasons (Deactivation date > rma_issue_date), then due_to_rma = 0 (the device is deactivated due to other reasons). Ideally, if the device returned to RMA, but still in use now, due_to_rma for those devices should be assigned as "NA" instead of "0." But we did not give enough thought on this when we integrated the data. And also if due_to_rma = NA for those devices, due_to_rma will no longer have capability to predict the deactivation of the devices. Therefore, if the devices returned to RMA but still in use now, then due_to_rma = 0 as well.

However, there is another possibility, as shown in Figure 4.13. There may be devices that were returned to RMA but the deactivation date is less than 60 days earlier than June 30, Year 4, therefore we cannot determine if the devices will be continued to use or deactivated after the repair. Fortunately, there are only about 2000 units belonging to this situation, so we simply ignore and delete them. due_to_rma explains whether or not the devices were deactivated due to RMA or due to other reasons. We will also want to discover, in later sections, which failure modes will directly cause the deactivation of devices.

The rate of deactivation due to RMA is increasing, as shown in Figure 4.14. In the early years, about 50% of the deactivated devices were deactivated because of RMA. In recent years, over 90% of the deactivated devices were deactivated because of RMA. This statistic shows that the devices produced in later years were more likely to be deactivated because of the RMA rather than other reasons. This situation makes sense because, in recent years, technology is updated so fast and customers have more choices. Thus, customers can easily switch to other products if any quality issues appear.

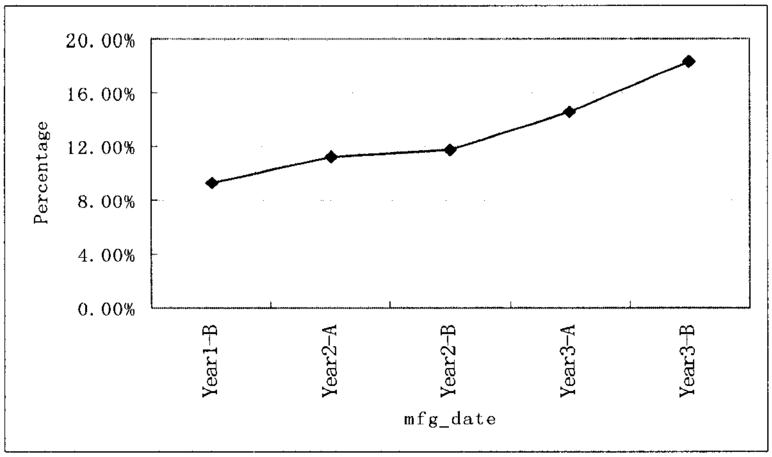


Figure 4.14 mfg_date vs. due_to_rma Rate (Deactivated Units)

As Table 4.4 shows, the devices deactivated due to RMA are more likely to have shorter ages. So we expect that RMA related factors will have a considerable effect on age.

due_to_rma	Average Age
0	317.22
1	212.80

Table 4.4 Average Age by due_to_rma

8. *Deactivated:* Overall, about 40% of devices in the market are deactivated, and 60% of the devices are still in use.

In summary, deactivation analysis and prediction can be very complicated; many factors influence deactivation. As noted in the single factor analysis discussed above, carrier and the RMA related attributes might be the influential factors to deactivation and age. Some of the factors are derived from other factors, and some factors have strong correlations with others. The knowledge discovered from single factor analysis is quite straightforward, and it is not very “interesting” because the manufacturer may already be familiar with the findings. The “long” and “thin” data set may add additional challenges to the data mining analysis, which will be discussed in the next chapter. It has been proven that the data mining methodology is capable of handling very large data sets. We expect that the data mining results will match the results from our preliminary analysis. In addition, data mining will identify the factors or the combination of the factors that will influence deactivation. We also expect data mining will discover some business rules. We expect to come to the conclusion that data mining can be an alternative solution for such business problems.

4.2.2 Preliminary Analysis on Time to Fail

The knowledge discovered from time to fail study can be useful in determining warranty policy. In this study, we use Table 4.2 as the data source.

There are in total 165 different failure modes, which can be further grouped into 37 categories. The failure modes that occurred more frequently are more important, and we are more interested in analyzing those top failure modes. It is useful to compare the time to fail to the warranty period in order to determine and recommend a better and more efficient warranty period for certain top-ranking failure modes.

4.2.3 Preliminary Analysis on Failure Mode Association and Classification

We believe that failure mode and/or customer abuse play important roles in deactivation and age prediction. We use Table 4.3 to perform this analysis. The columns represent 165 failure modes, which are exactly the same 165 distinct values of the variable fm_index as represented in Table 4.2. In this analysis, we aim to discover the association rules operating between failure modes, which are expressed like "if FM1 occurred, then FM2 also occurred on the same device." So we are looking for the failure modes that occurred together more frequently. In addition, we want to use classification to discover which failure modes or failure mode combinations will directly cause the deactivation of devices.

Number of FMs	Percentage
1	30.24%
2	19.87%
3	16.01%
4	17.77%
5+	16.11%

Table 4.5 Percentage of Num_FM

Most of the devices only have one type of failure mode. The devices that have 2, 3, 4 and 5+ failure modes are almost equally important. Table 4.5 shows that it is necessary to include up to 5 items in an association analysis. Therefore, it is necessary to perform 3-way, 4-way, and 5-way association analyses on the data in order to discover up to 5 FMs in an association. The association rules discovered from this study will indicate the combination of failure modes that occurred more frequently. This information will be very helpful in improving engineering design and the RMA process. But we also have to be prepared for the possibility that the association rules generated may not be very "interesting." Since the 165 failure modes are not evenly distributed, the association rules will still be emphasized on the top failure modes.

In summary, the data structure, the built-in patterns, and the limited attributes certainly add some challenges to the data mining. But there is a lot of useful information in the data, as we can tell from the preliminary analysis. We will apply the selected data mining methods on the data and evaluate the data mining results in the next chapter.

Chapter 5 Data Mining Results and Interpretations

A typical data mining process involves the phases of problem identification, data preparation, data exploration, modeling, and evaluation. The previous chapters addressed the problem identification, data preparation, and data exploration phases. In this chapter, appropriate data mining methods, including decision trees, regression trees and association are utilized on the data sets. Each business problem is investigated in multiple experiments. The experiment designs are carried out by SAS Windows 9.1 Enterprise Miner, Release 4.3. The purpose of the studies, the data sets and the model setting for each experiment design will be discussed in the following sections. The results will be validated and interpreted respectively. We expect to come up with the optimal solutions for each business problem.

5.1 Deactivation Classification and Prediction

The objectives of this part of the study are to identify the devices that have a high likelihood of being deactivated and the factors that influence deactivation. We use Table 4.1 as the data resource. Before applying data mining, we have to separate the data into two subsets: one for those units with RMA (returned for repair, had_rma = 1), the other for those units without RMA (never returned for repair, had_rma = 0). The data set with RMA consists of 9 variables, including the target variable Deactivated but excluding had_rma, Age, and time_delay. We will separate age and time delay and analyze them individually. The data set without RMA has 6 variables, including the target variable Deactivated but excluding the RMA related variables (num_fms, num_returns, due_to_rma). There are millions of records in Table 4.1: 10% of the records had RMA, while the rest never had RMA. There are two reasons why we separate the records into two data sets. First, the complete data set is too large to handle. We want to use a relatively smaller data set to perform the data mining methods comparison. Second, as the number of input variables to a model increases, there is an exponential increase in the data required to densely populate the model space. If the modeling space becomes too sparsely populated, the ability to fit a model to real world data is hampered. And for those units never returned to RMA, the RMA related variables will be redundant.

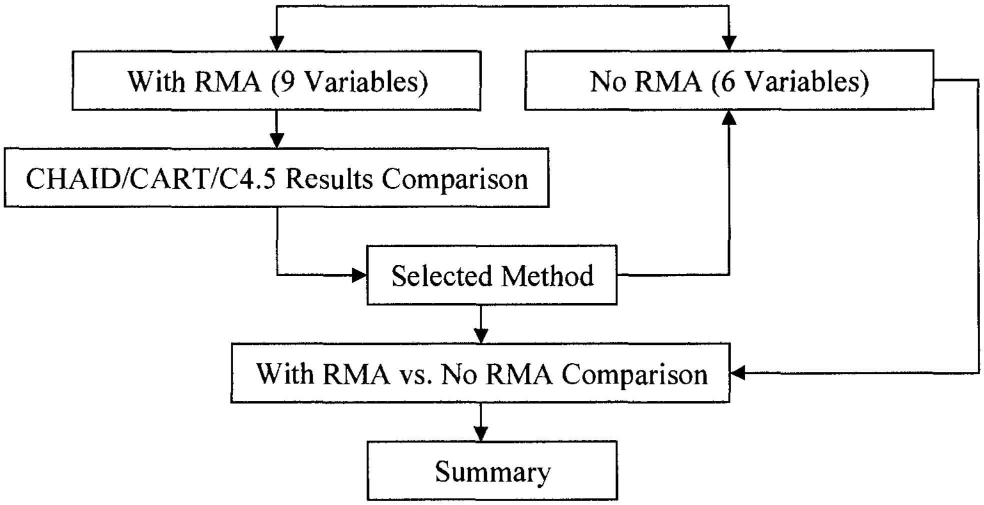

With RMA (9 Variables)
No RMA (6 Variables)
CHAID/CART/C4.5 Results Comparison
Selected Method
With RMA vs. No RMA Comparison
Summary

Figure 5.1 Deactivation Classification and Prediction Experiment Design

Figure 5.1 shows the procedure used in the deactivation analysis. At first, each of the three methods is employed on the data set with RMA to build a binary split decision tree. The results of the three models are then compared, and the method that provides the best fit model will be selected. Subsequently, the selected method will be used on the data set without RMA. Finally, the results obtained from data with and without RMA will be compared. This experiment design provides a computational advantage of saving time because only 10% of the records have RMA information.

To approximate those three methods when using the tree node, we follow the instructions of SAS Enterprise Miner and apply the following model settings for each method.

	CHAID	CART	C4.5
Splitting Criterion	Chi-Square	Gini Reduction	Entropy Reduction
Minimum number of observations in a leaf	202	202	202
Observations required for a split search	2024	2024	2024
Maximum number of branches from a node	6	2	6
Maximum depth of tree	8	8	8
Splitting rules saved in each node	5	5	5
Surrogate rules saved in each node	0	5	0
Model assessment measure	Total leaf impurity (Gini Index)	Total leaf impurity (Gini index)	Automatic
Observations sufficient for split search	32000	1000	32000
Maximum tries in an exhaustive split search	0	5000	0
P-value adjustment	Apply Kass after choosing number of branches	NA	NA

Table 5.1 Model Settings for Binary Split Decision Trees

To perform the analysis, the split of 80% for training, 10% for validation and 10% for testing is used. To compare the three data mining methods, we evaluate the results from misclassification rate, sensitivity, important variables, and decision rules. In order to visualize the results for the three developed models, the following tables provide the confusion matrixes (validation data) generated by SAS Enterprise Miner for validation of the data and for the model comparisons:

		Output		
		0	1	Total
Target	0	5342	1218	6560
	1	2506	11183	13689
	Total	7848	12401	20249

Table 5.2 Confusion Matrix - CHAID

		Output		
		0	1	Total
Target	0	5626	934	6560
	1	2773	10916	13689
	Total	8399	11850	20249

Table 5.3 Confusion Matrix - CART

		Output		
		0	1	Total
Target	0	5479	1081	6560
	1	2594	11095	13689
	Total	8073	12176	20249

Table 5.4 Confusion Matrix - C4.5

1. Misclassification Rate

The three methods are comparable on misclassification rate. For all the data sets, C4.5 shows the lowest misclassification rate.

	CHAID	CART	C4.5
Training Data	18.035%	18.046%	17.993%
Validation Data	18.391%	18.307%	18.149%
Testing Data	17.823%	17.798%	17.784%

Table 5.5 Misclassification Rate Comparison Table

2. Sensitivity

Sensitivity is a statistical measure of how well a binary classification test correctly identifies a condition. It measures the proportion of actual positives that are correctly identified as such. In our study, a sensitivity of 100% means that the test recognizes all deactivated devices as such. The sensitivities for each method are calculated in Table 5.6, according to the confusion matrixes and Equation 3.8.

Method	Sensitivity Calculation
CHAID	= 11183/13689 = **81.69%**
CART	= 10916/13689 = **79.74%**
C4.5	= 11095/13689 = **81.05%**

Table 5.6 Sensitivity Comparison Table

As Table 5.6 shows, CHAID has the highest sensitivity rate. So there is an 81.69% chance that CHAID correctly predicts the deactivated devices, which is highest among the three methods.

In statistical analysis, there are two types of error in model estimation, type I and type II errors which are universally accepted. They are often referred to as false positives and false negatives, respectively. The Type I errors and Type II errors are calculated in Table 5.7, according to the confusion matrixes, and Equation 3.10 and 3.11.

Methods	Type I error	Type II error
CHAID	= 1218/6560 = **18.57%**	= 2506/13689 = **18.31%**
CART	= 934/6560 = **14.24%**	= 2773/13689 = **20.26%**
C4.5	= 1081/6560 = **16.48%**	= 2594/13689 = **18.95%**

Table 5.7 Type I and Type II Errors Comparison Table

A Type I error is usually interpreted as a false alarm or insufficient specificity; which is the error of rejecting the null hypothesis when the hypothesis is actually true. A Type I error is also referred as a false positive rate, which is the proportion of negative instances that were erroneously reported as being positive. A Type II error can be similarly interpreted as an oversight, a lapse in attention, or inadequate sensitivity; it is the error of failing to reject the null hypothesis when the alternative hypothesis is actually true. It is also referred to as the false negative rate, which is the proportion of positive instances that were erroneously reported as negative. In this case, a Type I error means that the devices are actually still in use, but the model has classified them as deactivated. A Type II error here means that the devices are deactivated, but the model has classified them as still in use. From the comparison shown in Table 5.7, the decision tree model generated by CART tends to be more likely to classify a deactivated device as still in use, thus underestimating deactivation rate.

3. Important Variables

CHAID and C4.5 pick the same important attributes in the same order according to the importance value of the attributes given by the SAS Enterprise Miner decision tree result, as Table 5.8 shows (Black bar represents the importance value of the variable on training data; gray bar represents the importance value of the variable on validation data).

Variable	Training	Validation	Importance
due_to_rma	1	1	
act_date	0.319	0.315	
Carrier	0.122	0.143	
mfg_date	0.108	0.125	
num_returns	0.067	0.051	
num_fms	0.065	0.064	

Table 5.8 Important Variables Selected by SAS Enterprise Miner – CHAID

due_to_rma seems to be the most important variable; it is ranked as such by all three methods. Refer to Figure 4.13, if due_to_rma = 1, there is a 100% certainty the device is deactivated; if due_to_rma = 0, there are two possibilities, either the devices are deactivated due to other reasons or the devices are still in use. That is the way we set up due_to_rma this attribute. Because of the 100% certainty if due_to_rma = 1, the devices are deactivated, the model automatically picks it as the most important attribute because of its high accuracy in terms of classify the deactivation.

As the preliminary analysis in Chapter 4 shows, the devices are more likely to be in use if they were created and activated at a later date, while the devices are more likely to be deactivated if they were created and activated at an early date. The models pick mfg_date and act_date, as important variables, which is consistent with the preliminary analysis and common sense. We also expect that carrier, num_returns, and num_fms will have some impact on deactivation. As a result, the three models are almost the same in picking the important variables.

4. Significant Decision Rules

The decision rules generated by the three methods are quite comparable. Each rule gives a certain set of conditions and the number of instances that satisfies those conditions. Each rule also gives the percentage of deactivated devices, as well as the active ones under the given conditions. Because we are more interested in deactivation, the decision rules with a higher percentage of deactivation are considered as significant rules.

CHAID found 8 significant rules; CART found 5; and C4.5 found 9 rules. We ranked the rules according to the percentage of deactivation (Table 5.9).

	CHAID	CART	C4.5
1.	IF NUM_FMS = 2 AND ACT_DATE = YEAR1-B AND DUE_TO_RMA = 0 THEN 1 : 69.4% 0 : 30.6%	IF ACT_DATE = YEAR1-B AND NUM_FMS = 2, 3 AND NUM_RETURNS = 1 AND MFG_DATE = YEAR1-B AND CARRIER = CARA,CARB AND DUE_TO_RMA = 0 THEN 1 : 70.7% 0 : 29.3%	IF ACT_DATE = YEAR1-B AND DUE_TO_RMA = 0 THEN 1 : 60.8% 0 : 39.2%
2.	IF NUM_FMS = 3, 4 AND ACT_DATE = YEAR1-B AND DUE_TO_RMA = 0 THEN 1 : 64.6% 0 : 35.4%	IF ACT_DATE = YEAR2-A, YEAR2-B AND NUM_FMS = 2, 3 AND NUM_RETURNS = 1 AND MFG_DATE = YEAR1-B AND CARRIER = CARA,CARB AND DUE_TO_RMA = 0 THEN 1 : 62.4% 0 : 37.6%	IF CARRIER = CARB AND ACT_DATE = YEAR2-A AND DUE_TO_RMA = 0 THEN 1 : 58.5% 0 : 41.5%
3.	IF MFG_DATE = YEAR1-B AND NUM_RETURNS = 1 AND CARRIER = CARA AND ACT_DATE = YEAR2-A AND DUE_TO_RMA = 0 THEN 1 : 59.4% 0 : 40.6%	IF NUM_FMS = 5+, 1, 4 AND NUM_RETURNS = 1 AND MFG_DATE = YEAR1-B AND ACT_DATE = YEAR1-B, YEAR2-A, YEAR2-B AND CARRIER = CARA,CARB AND DUE_TO_RMA = 0 THEN 1 : 59.0% 0 : 41.0%	IF NUM_RETURNS = 1 AND CARRIER = CARA AND ACT_DATE = YEAR2-A AND DUE_TO_RMA = 0 THEN 1 : 56.8% 0 : 43.2%
4.	IF CARRIER = CARB AND ACT_DATE = YEAR2-A AND DUE_TO_RMA = 0 THEN 1 : 58.5% 0 : 41.5%	IF NUM_FMS = 2, 3, 4 AND ACT_DATE = YEAR2-A AND MFG_DATE = YEAR2-A AND CARRIER = CARA,CARB AND DUE_TO_RMA = 0 THEN 1 : 57.0% 0 : 43.0%	IF MFG_DATE = YEAR1-B AND CARRIER = CARA AND ACT_DATE = YEAR2-B AND DUE_TO_RMA = 0 THEN 1 : 56.1% 0 : 43.9%
5.	IF NUM_FMS = 1 AND ACT_DATE = YEAR1-B AND DUE_TO_RMA = 0 THEN 1 : 57.9% 0 : 42.1%	IF NUM_FMS = 2, 3 AND CARRIER = CARB AND ACT_DATE = YEAR2-B AND MFG_DATE = YEAR2-A AND DUE_TO_RMA = 0 THEN	IF CARRIER = CARB AND MFG_DATE = YEAR2-A AND ACT_DATE = YEAR3-A AND DUE_TO_RMA = 0 THEN

		1 : 54.1% 0 : 45.9%	1 : 55.7% 0 : 44.3%
6.	IF NUM_FMS = 2, 3, 4 AND MFG_DATE = YEAR2-A AND NUM_RETURNS = 1 AND CARRIER = CARA AND ACT_DATE = YEAR2-A AND DUE_TO_RMA = 0 THEN 1 : 56.7% 0 : 43.3%		IF NUM_FMS = 2 AND MFG_DATE = YEAR1-B, YEAR2-A AND CARRIER = CARB AND ACT_DATE = YEAR2-B AND DUE_TO_RMA = 0 THEN 1 : 54.9% 0 : 45.1%
7.	IF MFG_DATE = YEAR1-B AND CARRIER = CARA AND ACT_DATE = YEAR2-B AND DUE_TO_RMA = 0 THEN 1 : 56.1% 0 : 43.9%		IF NUM_RETURNS = 3+, 2 AND CARRIER = CARB AND MFG_DATE = YEAR2-B AND ACT_DATE = YEAR3-A AND DUE_TO_RMA = 0 THEN 1 : 53.6% 0 : 46.4%
8.	IF CARRIER = CARB AND MFG_DATE = YEAR1-B, YEAR2-A AND ACT_DATE = YEAR3-A AND DUE_TO_RMA = 0 THEN 1 : 55.4% 0 : 44.6%		IF NUM_RETURNS = 2 AND MFG_DATE = YEAR2-B AND CARRIER = CARB AND ACT_DATE = YEAR2-B AND DUE_TO_RMA = 0 THEN 1 : 53.6% 0 : 46.4%
9.			IF NUM_FMS = 3 AND MFG_DATE = YEAR1-B AND CARRIER = CARB AND ACT_DATE = YEAR2-B AND DUE_TO_RMA = 0 THEN 1 : 53.2% 0 : 46.8%

Table 5.9 Significant Decision Rules Generated by Three Models

From the top two rules each method generated, we can see 1) devices that were created in YEAR1-B, activated in either YEAR1-B or YEAR2-A, and had failures which did not directly cause the deactivation (due_to_rma = 0), are more likely deactivated till now. 2) devices that were created in YEAR1-B, activated in YEAR1-B or YEAR2-A, activated under Carrier A or B, and had failures which did not directly cause the deactivation (due_to_rma = 0) - no matter how many failures were found with the devices - those devices are likely deactivated till now(more than 50% chance). Therefore, the decision rules generated from the three methods are similar and comprehensive. In fact, the decision rules generated match the preliminary analysis and common knowledge.

In comparing the three methods used on the data set with RMA, C4.5 performs a little bit better than the other two methods because of its lower misclassification rate. Therefore, we will apply C4.5 on the data set without RMA in order to identify the influential factors on deactivation. The data set with no RMA is very large, but has only 6 attributes.

The misclassification rate on the data with no RMA is 34.382%. As Table 5.10 shows that sensitivity is equal to 43.28%, which means the test only recognizes 43.28% of all deactivated devices as such.

	Frequency	Output		
		0	1	Total
Target	0	86494	22081	108575
	1	38735	29561	68296
	Total	125229	51642	176871

Table 5.10 Confusion Matrix (No RMA)

Rule 1:
IF MFG_DATE EQUALS YEAR1-B
AND ACT_DATE IS ONE OF: YEAR1-B YEAR2-A
THEN
1 : 63.2%
0 : 36.8%

Rule 2:
IF MFG_DATE EQUALS YEAR1-B
AND ACT_DATE EQUALS YEAR3-A
THEN
1 : 55.8%
0 : 44.2%

Because of the higher misclassification rate and the lower sensitivity rate, the model is not very reliable, possibly because of the data structure, which is long and thin. Only two rules are significant. It seems that time factors (mfg_date and act_date) are the most important predictors of deactivation, which is common knowledge; when the devices were created and activated early; they are more likely to be deactivated till now. As well, it seems like carriers and regions are not important in predicting deactivation.

Through these experiments, we have come to some conclusions about deactivation prediction.

- Decision tree model can be a solution for such problems. Three methods, CHAID, CART and C4.5, are comparable in terms of misclassification rate, important variables, and decision rules. The important variables and decision rules generated match the preliminary analysis and common knowledge. The results demonstrate that the data mining process and methods are efficient in solving certain problems. If more data, such as customer information, device usage, and contract information are available, we expect that a decision tree should generate better results that can directly benefit the organization. Currently, the decision rules generated are common knowledge, but these results verify that the model is suitable for certain business problems.

- Carrier and Region are not significant factors. They do not have a big impact on deactivation prediction. But RMA related attributes are. Further analysis of which failure modes are directly related to deactivation will be necessary and useful.
- The separation of the data into two data sets, with RMA and without RMA, is necessary to reduce the misclassification rate. This separation also shows, in any case, that effective feature selection and redundant attribute elimination will be crucial in improving the classification accuracy.
- The misclassification rate on the data with RMA is 17.993%, while the misclassification rate on the data without RMA is 34.382%. If the devices had RMA, they are more likely to be deactivated till now. However, if the devices were never returned to RMA, the reasons for their deactivation appear random and hard to predict. It makes sense, then, that the misclassification rate on the data with RMA is lower. From these results, we can say that the RMA related attributes are

significant in predicting deactivation. However, we will need to further analyze which failure modes are directly related to deactivation. Also, because the data set contains millions of record, but has only 12 attributes. There are many cases have the same values for each attribute but belong to different classes. For example, two devices might have exactly the same values for each attribute, they were manufactured in the same manufacturing location, they were carried by the same carrier, they all returned to RMA, but one of them is deactivated, one of them is still in use. The model cannot classify these devices that have exactly the same values for each attribute, there are 50/50 percent chance either it is deactivated or still in use, so the misclassification rate is increased due to there are too many devices in the data set are like that. There are two possible solutions for that: one is to use sampling, while the other is to include more attributes when the data is available.

5.2 Age Analysis

Age is measured as how many days the device has been used; it is also identified as time in use. Age represents how long a device is used by a customer, which reflects a customer's loyalty to the product. Because age is only valid on the deactivated devices, so in this research, we do not include age as a factor in predicting deactivation. The age may or may not have correlation with deactivation. Some of the devices were deactivated with only short ages, while some devices were deactivated with longer ages; the reasons for that are multiple. Both the manufacturer and carriers like to see customers keep and use the device longer. In this section, we expect to identify the influential factors that will distinguish the devices with short age and long age and to find out how time in use changes by the factors through using a combination of data mining and graphical tools. We proposed three experiment designs, as Figure 5.2 shows, by using three approaches.

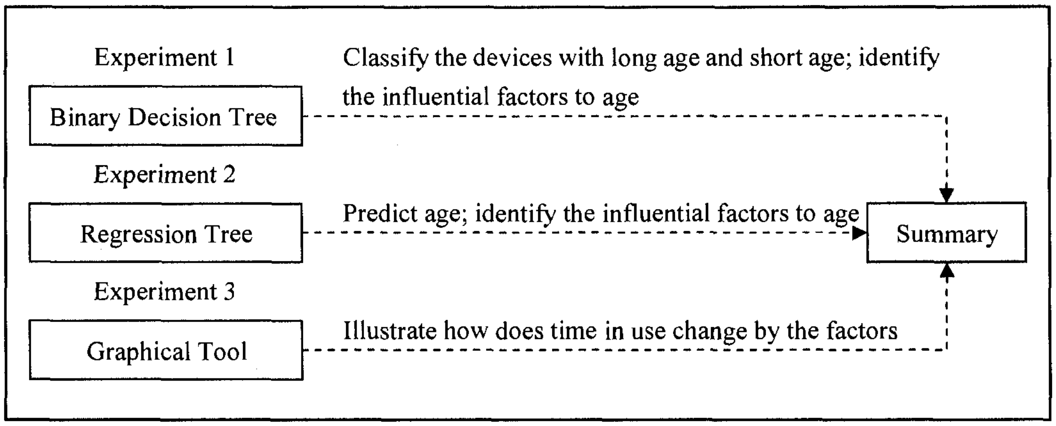

Experiment 1
Binary Decision Tree
Classify the devices with long age and short age; identify the influential factors to age
Experiment 2
Regression Tree
Predict age; identify the influential factors to age
Summary
Experiment 3
Graphical Tool
Illustrate how does time in use change by the factors

Figure 5.2 Age Analysis Experiment Design

5.2.1 Experiment 1: Binary Decision Tree

When we only look at the devices that are already deactivated, their age most likely fell between 0-400 days. Some of the devices have ages between 400-700 days, but there are very few devices that were used longer than 700 days. We want to compare two extreme cases: devices that were deactivated with short ages and devices that were deactivated with long ages. In the experiment design for the age analysis, therefore, the deactivated devices are divided into two groups, one with Age <= 100 days and the other with Age >= 700 days. The two groups represent the two extreme cases. It is interesting to see whether our data mining model can identify different sets of influential factors.

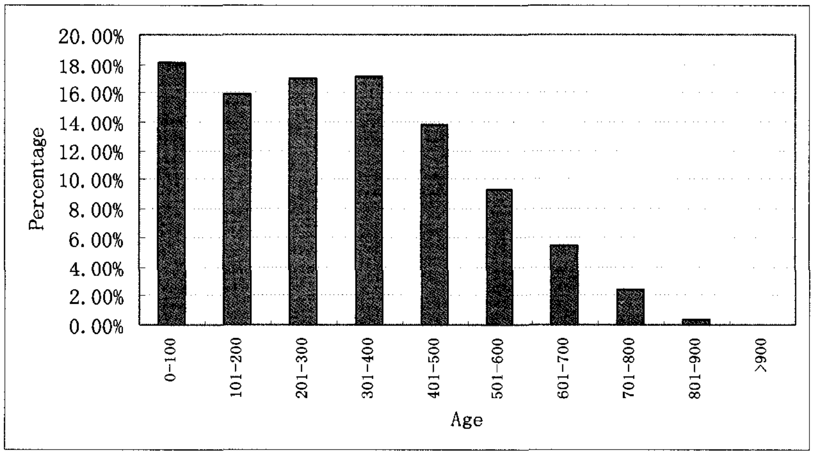


Figure 5.3 Age Breakdown for Deactivated Units

We choose to use a decision tree (C4.5) because we aim to classify the devices with short ages and long ages and to identify the factors that influence the age. This is a typical classification problem. We use the data from Table 4.1, but only include the deactivated devices (deactivation = 1). All the attributes listed in Table 4.1 are included, except "Deactivated." Age is the target variable, which has two values, 100 and 900, as we set age = 100 for the units that have age <=100 days, while we set age = 900 for the units that have age >700 days. So the model becomes a binary decision tree.

The misclassification rate is 7.179%, which is low. This low rate shows the high accuracy rate of the decision tree model. Table 5.11 shows the important variables the model picked. And there are a total of 20 decision rules generated. 14 rules are significant in classifying the short age, while 6 rules are significant in classifying the long age as shown in Table 5.12.

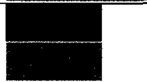
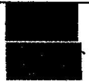
Variable	Training	Validation	Importance
Act_date	1	1	
Due_to_rma	0.342	0.340	
Mfg_date	0.282	0.301	
Time_delay	0.255	0.231	
Carrier	0.091	0.076	
Had_rma	0.068	0.067	
Region	0.067	0.071	
Mfg_loc	0.053	0.045	

Table 5.11 Important Variables Selected by SAS Enterprise Miner (Age – C4.5)

Short Age (age <=100 days)	Long Age (age > 700 days)
Act_date = YEAR2-B, YEAR3-A, YEAR3-B,YEAR4-A	Due_to_rma = 0; mfg_date = YEAR1-B, YEAR2-A; act_date = YEAR2-A
Due_to_rma = 1; act_date = YEAR1-B, YEAR2-A	Due_to_rma = 0; act_date = YEAR1-B
Mfg_date = YEAR2-A; act_date = YEAR2-A; time_delay = 50 days	Due_to_rma = 0; had_rma = 0, 1; time_delay = 100 days; mfg_date = YEAR1-B; act_date = YEAR2-A
Region = 2, 3; time_delay = 150 days; Act_date = YEAR1-B; mfg_date = YEAR2-A	Due_to_rma = 0; region = 1, 2; time_delay < 100 days; mfg_date = Y2004; act_date = YEAR2-A
Mfg_date = YEAR1-B; Due_to_rma = 0; time_delay > 150 days	

Table 5.12 Significant Decision Rules (Age – C4.5)

Since age = deact_date – act_date; therefore, the act_date date is correlated with the target variable, and it is ranked as the most important variable. It seems that in any activation period, there are a number of devices deactivated within 100 days.

due_to_rma is still an important factor. Even the devices that had RMA, as long as they did not deactivate due to RMA, generally have longer ages. But if the devices had failures that caused deactivation, they generally have shorter ages. In addition, time delay might have significant impact on age as well.

5.2.2 Experiment 2: Regression Tree

In previous experiments, we used a typical decision tree model to solve a classification-type problem. Regression-type problems are generally those where one attempts to predict the values of a continuous variable from one or more continuous and/or categorical predictor variables. In 5.2.1, when we designed an experiment that grouped the age variable into short age and long age, it became a typical binary decision tree problem. Because age is a continuous variable in Table 4.1, in this experiment, we want to apply a regression tree, which predicts a continuous variable and to discover the influential factors to age.

CHAID and CART are all capable of generating regression tree models. CHAID uses the F-test as a splitting criterion for interval targets, while CART uses variance reduction. All the other settings are the same as those for a binary decision tree. When we compare CHAID and CART, they are very similar in terms of accuracy rate, important variables and rules. CART performs little bit better than CHAID, so we use CART for modeling in this experiment.

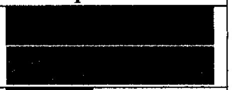
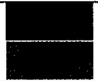
Variable	Training	Validation	Importance
Act_date	1	1	
Due_to_rma	0.317	0.314	
Time_delay	0.165	0.162	
Carrier	0.152	0.152	
Mfg_date	0.138	0.142	
Num_returns	0.093	0.094	
Num_fms	0.071	0.070	
Region	0.053	0.068	
Mfg_loc	0.027	0.024	

Table 5.13 Important Variables Selected by SAS Enterprise Miner (Age – CART)

Table 5.13 shows the important variables selected by the regression tree model using the CART method. In comparison, Table 5.11 shows the important variables selected by the binary decision tree model that used C4.5. The important variables that the two methods picked are very similar. According to the results, act_date, due_to_rma, time_delay, carrier, and mfg_date are the most important variables in predicting age. In the regression model, num_returns and num_fms seem to have some impact on age prediction; however, in the binary decision tree model, it seems as if those two variables do not have any effect on distinguishing long and short age. In addition, had_rma seems not to have any impact in the regression model, as well as in the binary decision tree model. Region and mfg_loc have a small impact in both models, but it is not significant.

The regression tree model returns a typical tree model. In a classification tree, the terminal nodes are assigned to a specific class according to the class assignment rule. In regression trees, however, there are no classes to which terminal nodes are as assigned. Instead, for each of the terminal nodes produced by the CART regression, summary statistics of the dependent variable are computed. There were 93 rules found with the regression tree model. We rank the rules according to the average age given by each rule.

Short Age	Long Age
IF NUM_FMS = 3, 2 AND CARRIER = CARA, CARB AND DUE_TO_RMA = 1 AND ACT_DATE = YEAR3-A THEN AVE : 86.1974 SD : 91.3543	IF ACT_DATE = YEAR1-B AND DUE_TO_RMA = 0 THEN AVE : 494.811 SD : 231.319
IF DUE_TO_RMA = 1 AND MFG_DATE = YEAR1-B, YEAR2-A, YEAR3-B AND ACT_DATE = YEAR3-B THEN AVE : 57.7822 SD : 55.8714	IF TIME_DELAY < 77.5 AND MFG_DATE = YEAR1-B AND ACT_DATE = YEAR2-A AND DUE_TO_RMA = 0 THEN AVE : 474.616 SD : 215.277
IF 61.5 <= TIME_DELAY AND DUE_TO_RMA = 0 AND MFG_DATE = YEAR1-B, YEAR2-A, YEAR3-B AND ACT_DATE = YEAR3-B THEN AVE : 74.8881 SD : 60.4802	IF REGION = 2, 3 AND 77.5 <= TIME_DELAY < 107.5 AND MFG_DATE = YEAR1-B AND ACT_DATE = YEAR2-A AND DUE_TO_RMA = 0 THEN AVE : 439.513 SD : 237.99
IF NUM_FMS = 5+, 3, 2, 4 AND TIME_DELAY < 129.5 AND MFG_DATE = YEAR2-B, YEAR3-A AND ACT_DATE = YEAR3-B THEN AVE : 98.0885	IF 131.5 <= TIME_DELAY < 169.5 AND MFG_DATE = YEAR1-B AND ACT_DATE = YEAR2-A AND DUE_TO_RMA = 0 THEN AVE : 417.542

SD : 77.9821	SD : 201.137
IF 358.5 <= TIME_DELAY AND MFG_DATE = YEAR2-B, YEAR3-A AND ACT_DATE = YEAR3-B THEN AVE : 68.7984 SD : 60.695	IF REGION = 3 AND CARRIER = OTHERS AND MFG_DATE = YEAR2-A AND ACT_DATE = YEAR2-A AND DUE_TO_RMA = 0 THEN AVE : 439.444 SD : 203.942
IF CARRIER = CARA, CARD, CARB AND 129.5 <= TIME_DELAY < 358.5 AND MFG_DATE = YEAR2-B, YEAR3-A AND ACT_DATE = YEAR3-B THEN AVE : 113.905 SD : 74.188	IF MFG_LOC = A, B AND REGION = 1 AND 77.5 <= TIME_DELAY < 107.5 AND MFG_DATE = YEAR1-B AND ACT_DATE = YEAR2-A AND DUE_TO_RMA = 0 THEN AVE : 463.831 SD : 208.668
IF CARRIER = OTHERS, CARC AND 129.5 <= TIME_DELAY < 358.5 AND MFG_DATE = YEAR2-B, YEAR3-A AND ACT_DATE = YEAR3-B THEN AVE : 92.6358 SD : 71.0565	IF CARRIER = CARA, CARD, OTHERS AND MFG_LOC = B AND TIME_DELAY < 97.5 AND REGION = 1 AND MFG_DATE = YEAR2-A AND ACT_DATE = YEAR2-A AND DUE_TO_RMA = 0 THEN AVE : 440.56 SD : 199.853

Table 5.14 Significant Decision Rules (Age – CART)

If a device is created and activated early, YEAR1-B or YEAR2-A, and if it is not sent back for repair or is still used after a repair (due_to_rma = 0), then it is more likely to have a long age. A device tends to have a short age regardless of when it was created and activated if its due_to_rma = 1, that is, if it had some failure modes that directly caused the deactivation.

In conclusion, the regression tree performs fairly well for age prediction. Currently, the statistical analysts use statistical data analysis tools to predict age for the QA group. Compared with the statistical method, a regression tree takes all the factors into consideration, but it is a little bit more difficult to interpret. The failures that cause the deactivation of devices normally also cause short age, and the ages of those devices are around 100 days. Therefore, further investigation of which failures directly cause deactivation is necessary. Sometimes the devices really had failures or customer abuse, or sometimes the customers simply deactivated and returned the devices, but RMA could not find any failures with these returned devices. Since a number of devices are deactivated after very short period of time, customer information would be helpful to predict age as well. We expect that the regression tree model could be helpful in identifying which group of customers intended to hold the device for only a short time if we had access to customer data.

5.2.3 Experiment 3: How does time in use change by factors?

In the previous age analysis, we used decision tree and regression tree to discover the influential factors to age. In this section, we want to perform a single factor analysis using graphical tools, to verify the findings from data mining results, and to illustrate how time in use changes by the factors.

We do not use quantity to do the comparison because the quantities of devices are not evenly distributed among carriers, regions, manufacturing locations, and other factors. For another example, if 1000 infants were born in January 2008 in both Canada and

China. But China has a much larger population than Canada; the birth rate in Canada would actually be much higher than that in China in January 2008. In our example, for instance, say Carrier A carries 10,000 units and Carrier B carries 5,000 units, and say Carrier A has 1,000 units deactivated within 100 days and Carrier B has 800 units deactivated within 100 days. It may look like Carrier A has more devices with short ages, but when we look at the percentage of deactivated devices within 100 days (rate for A = 1000/10000 = 0.1, rate for B = 800/5000 = 0.16), Carrier B has a higher percentage of deactivation. Therefore, we use percentage to do the comparison and illustrate how age changes by factors. Only deactivated units are included in the analysis because only their ages are valid and can be determined.

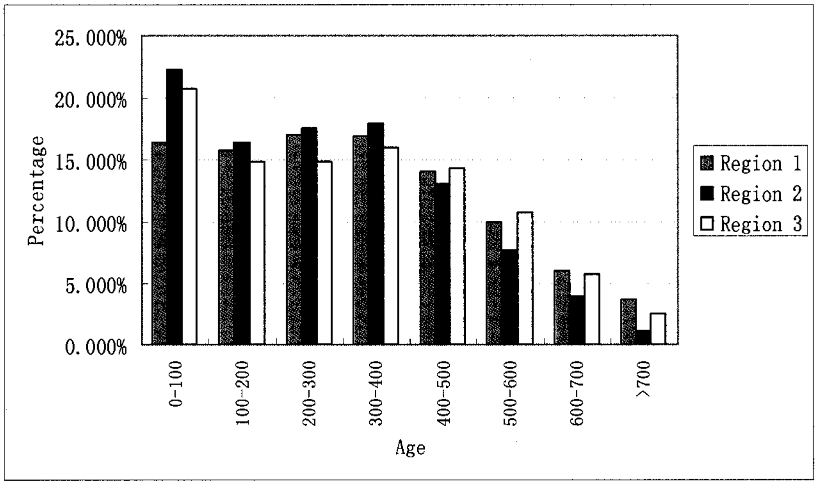


Figure 5.4 Age Change by Region

Looking at Figure 5.4, there is no significant difference among three regions. Moreover, Region 2 has a higher percentage of devices with shorter ages compared with those in the other two regions.

Figure 5.5 shows that there is no significant difference among carriers. Carriers B, C and Others have a higher percentage of the devices with shorter ages but a lower percentage of the devices with longer ages. Carriers A and D show a similar trend, which is opposite to the trend shown by carriers B, C, and Others.

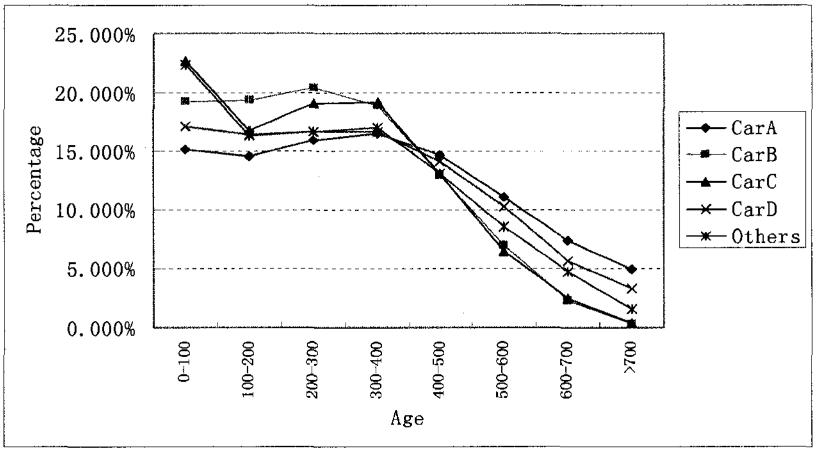


Figure 5.5 Age Change by Carrier

In Figure 5.6, we can see that two manufacturing locations have significantly different trends. Manufacturing location B has a normal distribution; thus, the percentage of devices with short age is almost identical with the percentage of devices with long age. But manufacturing location A has a higher percentage of devices with short age.

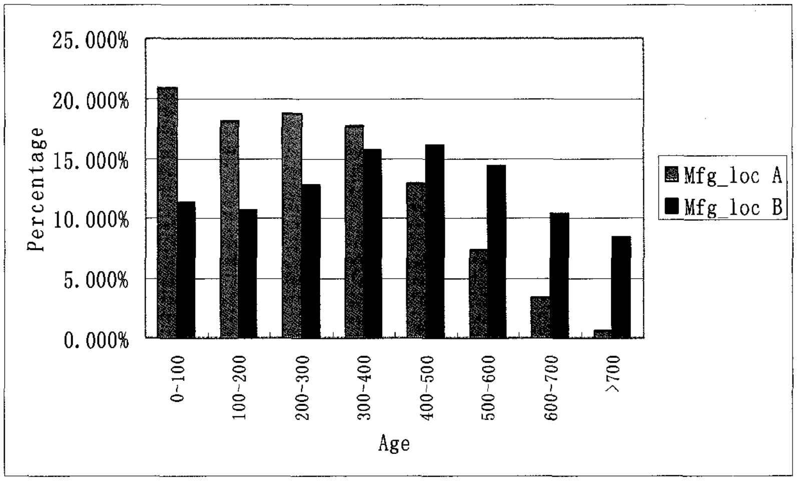


Figure 5.6 Age Change by mfg_loc

From Figure 5.7 shows that those devices which deactivations were caused by RMA normally have shorter ages.

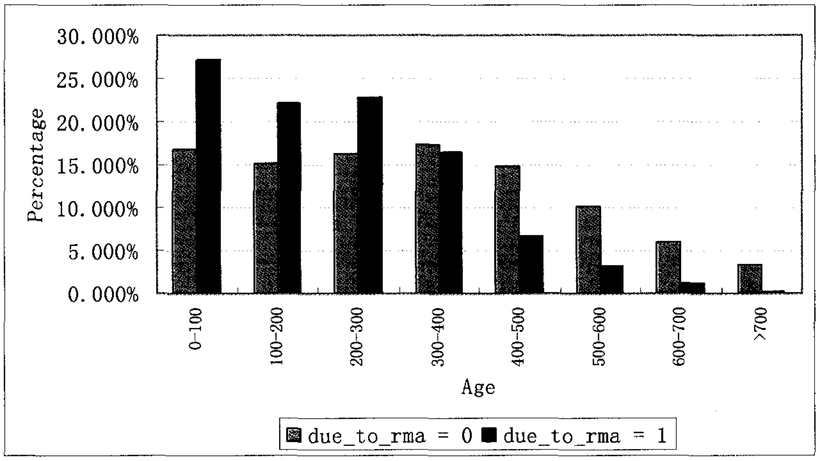


Figure 5.7 Age Change by due_to_rma

Looking at Figure 5.8, the devices with longer time delay are more likely to have shorter ages.

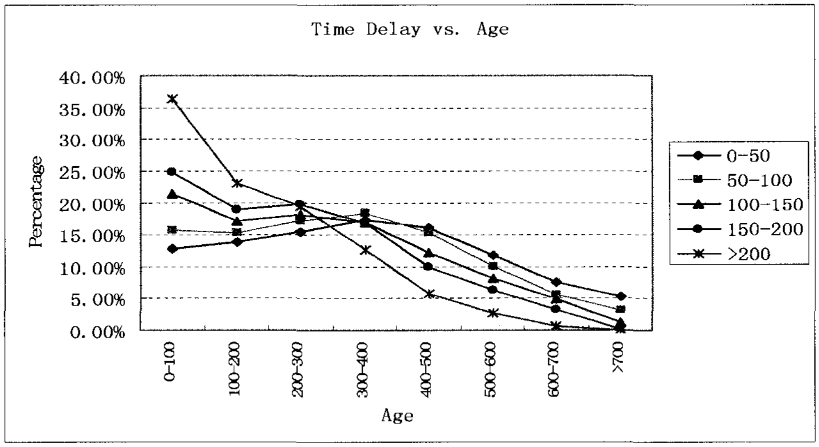


Figure 5.8 Age Change by time_delay

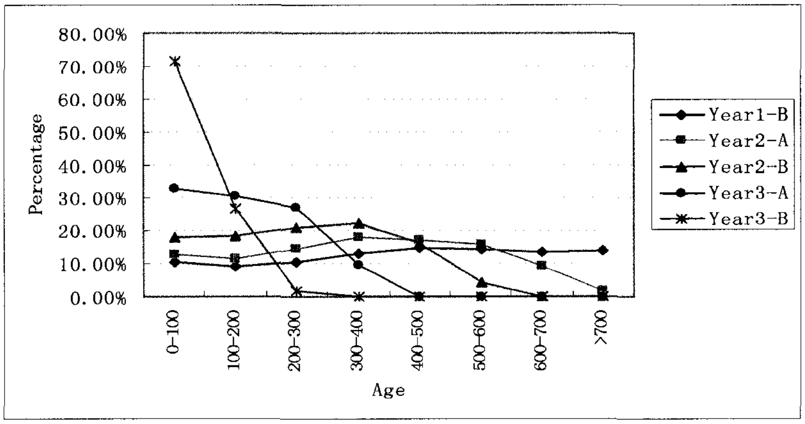


Figure 5.9 Age Change by mfg_date

Figures 5.9 and 5.10 reveal that the devices created in YEAR3-B, and/or activated in YEAR3-B and YEAR4-A, are more likely to be deactivated within 100 days. There might be two reasons for this: first, the devices were not created and activated for that long; in fact, there are not many devices used longer than 100 or 200 days. Second, when this type of product first entered the market, it had a large and stable market share. As time went by, a number of different types of products entered into the market, giving customers more choices. These new choices could cause customers to deactivate these older devices.

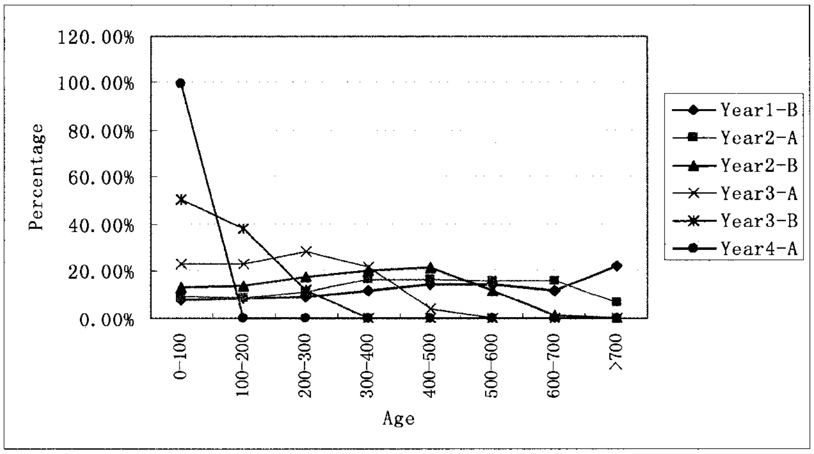


Figure 5.10 Age Change by act_date

As Figure 5.11 shows, the devices that never had RMA tend to have longer ages, while devices that had RMA tend to have shorter ages. However, the difference is not that significant.

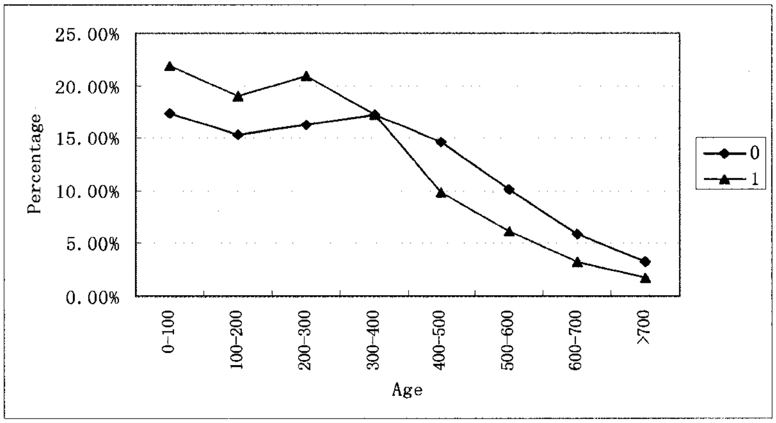


Figure 5.11 Age Change by had_rma

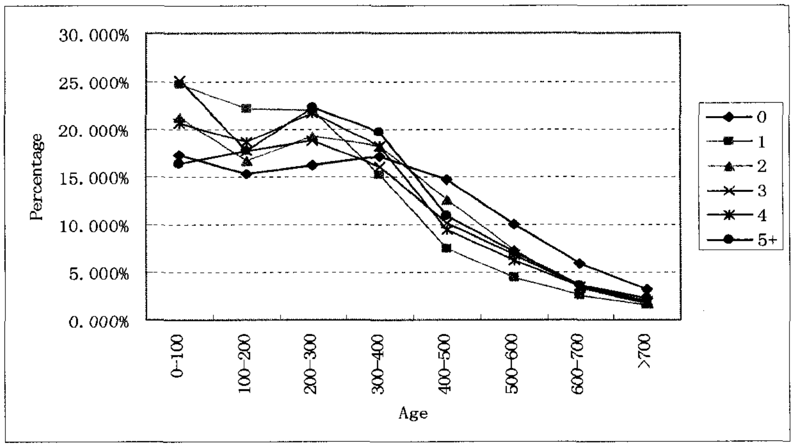


Figure 5.12 Age Change by num_fms

Because there is no significant difference between the devices returned to RMA and the devices never returned to RMA, there are no significant differences among the number of failure modes the devices had.

In this experiment, we not only used diagrams to present how time in use changes by factors, but also verified the data mining results from two previous experiment designs, namely, that due_to_rma and time_delay are the two most important influential factors for age. Other factors have some differences and some impacts on age, but they are not that significant. Further study on the failures related to deactivation and time delay analysis will be performed in the next sections.

5.3 Time Delay Analysis

Time delay is a measurement of how many days elapsed between a device's manufacturing date and activation date: time_delay = act_date – mfg_date. Both manufacturers and carriers want the time delay to be short in order to speed up the product's entry into the market and to maximize profit. In practice, a manufacturer produces a new batch of devices after receiving an order from a carrier. The carrier decides how many units to order according to market conditions. Therefore, it is in the best interest of both parties to reduce time delay. It is believed that this attribute is not correlated to RMA or deactivation. However, it is interesting from the manufacturer's perspective whether or not time delay varies among different carriers and different regions. Therefore, the purpose of this analysis is to explore the relationship of time delay to carriers and regions.

Because only Carrier and Region are related to the time delay prediction, data mining methods might not be applicable. We tried a regression tree model and also a decision tree model by grouping the target variable time_delay into equal sized categories such as 0-50 and 50-100. The models showed a high misclassification rate and did not generate meaningful rules. Therefore, we abandoned the data mining methods in time delay analysis, but used the graphical method instead.

Figure 5.13 shows that most devices were activated within 100 days of being manufactured. There were also many devices activated a long time after the manufacturing date. To further the analysis, we break the results down to see which carrier or region has the longer and shorter time delay.

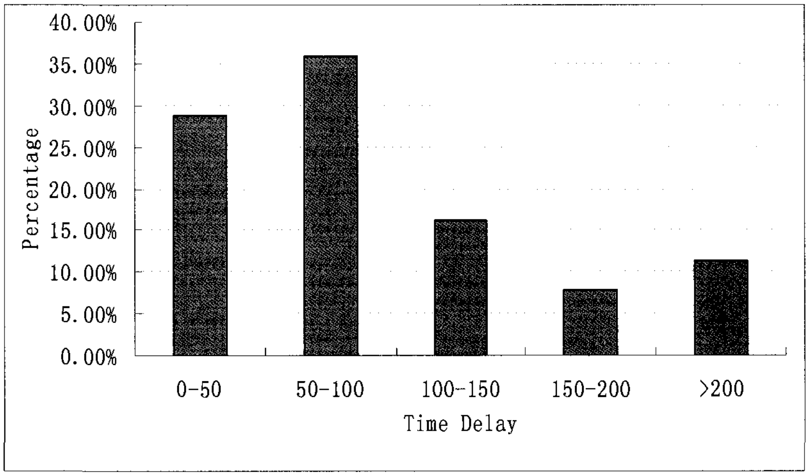


Figure 5.13 Time Delay Distribution

In Figure 5.14 one can see that, Region 2 had a larger percentage of devices with longer time delay, but it had a smaller percentage of the devices with shorter time delay. Region 1, however, is the complete opposite. Therefore, in Region 1, devices entered into the market and into use faster than in the other two regions.

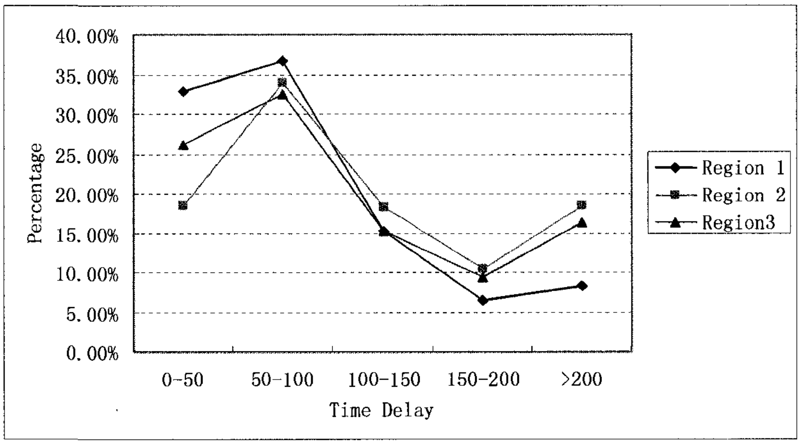


Figure 5.14 Time Delay Change by Region

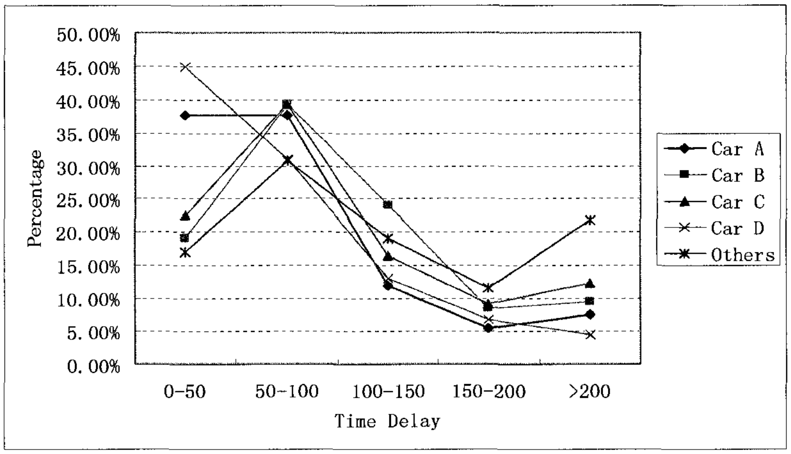


Figure 5.15 Time Delay Change by Carrier

Figure 5.15 reveals that Carriers A and D had shorter time delays, while the carrier marked as Others had a higher percentage of devices with longer time delay but a lower percentage of devices with shorter time delay.

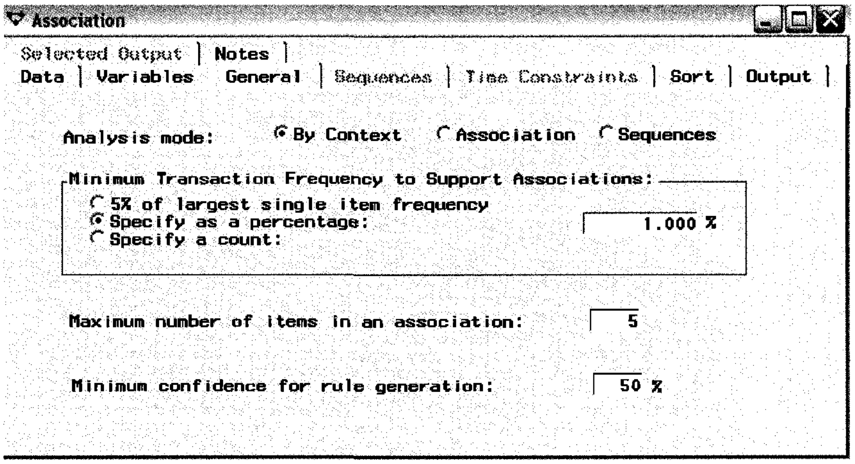
In conclusion, both region and carrier have impacts on time delay. The region where the manufacturer is located is more likely to have short time delay. Bigger carriers tend to have shorter time delay. There are multiple reasons for this situation: bigger carriers have a better understanding of their market, so they are more likely to put out the right order at the right time, and bigger carriers are more recognized. The smaller carriers, which we grouped into Others, have a comparably longer time delay because of the same two reasons. Therefore, a good market plan and manufacturing plan will be the key to cutting time delays.

5.4 Failure Mode Association Rule Mining and Classification

In the previous analysis, it was found that failure modes were very important in predicting age and deactivation. In this section, we focus on analyzing the correlation among failure modes, that is, which failure modes, or combinations of failure modes are most likely to occur. The results of this analysis will provide useful information to the manufacturing department, engineering designers, and RMA, which is one of the mandates of the managers in the QA group. Therefore, we propose two experiment designs in this section: 1) Failure mode association rule mining: to identify the failure modes that are more likely to occur together in a given device. Through the experiment, we aim to identify the combinations of failure modes that occur more frequently. The knowledge discovered from this experiment will improve the engineering design and the RMA process. 2) Classification. We want to identify which failure modes and failure mode combinations directly relate to deactivation.

5.4.1 Failure Mode Association Rule Mining

Association discovery is the identification of items that occur together in a given event or record. The data set is represented in Table 4.3. Among all the records, only 10% of devices were returned to RMA. So the data set in Table 4.3 is only 10% of all data records. In Chapter 4, we revealed how we generated this table and data structure. Figure 5.16 shows the model settings for the association model.


Association
Selected Output | Notes |
Data | Variables General | Sequences | Time Constraints | Sort | Output |

Analysis mode: ● By Context ○ Association ○ Sequences

Minimum Transaction Frequency to Support Associations:

○ 5% of largest single item frequency
● Specify as a percentage: 1.000 %
○ Specify a count:

Maximum number of items in an association: 5

Minimum confidence for rule generation: 50 %

Figure 5.16 Association Model Settings

A total of 572 association rules are generated. 54 rules contain 2 items in each rule, 194 rules contain 3 items, 230 rules contain 4 items, and 94 rules contain 5 items. All the rules generated are significant and meet the minimum requirements for support and confidence. The association rules contain a total of 30 distinct FMs, which are also the top 30 failure modes of all 165 FMs in terms of frequency. The other failure modes are not significant compared to the top 30. If a failure mode occurred only once, that meant it had a very small chance of occurring. These low probability failures are irrelevant because, even if the organization loses one customer because of this failure mode, it will not hurt the business. Also, the association rules have to meet the minimum support and confidence; it makes sense that the association rules only include the top failure modes. The detailed discussion on association rules follows.

(1) 2-way Relation Association Rules

We rank the association rules according to the confidence factor, as well as support level. There are two very significant 2-way relation association rules.

	Relations	Lift	Support (%)	Confidence (%)	Rule
1	2	1.40	46.91%	95.68%	N57Ke ➔ N143Ho
2	2	1.34	43.08%	91.45%	N80LC ➔ N143Ho

Table 5.15 Significant Association Rules (2-way Relation) – 1

The first rule means that if a device's keypad is not functioning, then 95.68% of the time, it also has housing cosmetic damage. The second rule means that if a device has its LCD scratched, then 91.45% of the time it will also have housing cosmetic damage. Because their transaction counts are large, their support levels are also high. Therefore, those two rules are most significant. But these strong relationships tend to be obvious. We need to dig deeper into the rules to find clues to some hidden relationships. Normally, the association rules with confidence >= 80% are considered as significant. Table 5.16 shows some 2-way relation association rules that have confidence levels of more than 80%, but lower support levels.

	Relations	Lift	Support (%)	Confidence (%)	Rule
1	2	1.36	19.89%	92.67%	N36Ke ➔ N143Ho
2	2	1.44	14.03%	98.61%	N23Ke ➔ N143Ho
3	2	1.91	13.35%	93.83%	N23Ke ➔ N57Ke
4	2	1.82	12.23%	85.95%	N23Ke ➔ N80LC
5	2	1.39	4.56%	95.21%	N81Ba ➔ N143Ho
6	2	1.83	4.30%	89.83%	N81Ba ➔ N57Ke

Table 5.16 Significant Association Rules (2-way Relation) – 2

If the devices have "keypad scratched," "keypad broken/deformed," or "battery connector broken/deformed" failure modes, they are very likely to have housing cosmetic damage as well. If the devices have "keypad broken/deformed" or "battery connector" failure modes, then their keypads are more likely to be not functioning. In addition, "keypad broken/deformed" is very likely to lead to "LCD scratched". In addition, there are some rules with higher confidence (>=80%) but lower support, as Table 5.18 shows.

Left Hand Side	Right Hand Side
N124Internal Microphone N85Antenna N39SIM Card N101Shield N86LCD N102Headset Mic N146Headset Jack N148Headset Jack N88USB/Serial Connector	N143Housing: Cosmetic Damage
N124Internal Microphone N39SIM Card N85Antenna N101Shield N102Headset Mic N146Headset Jack	N57Keypad: Not Functioning
N85Antenna N146Headset Jack N88USB/Serial Connector	N80LCD: Scratched

Table 5.17 Significant Association Rules (2-way Relation) – 3

From the 2-way relation association rules generated, we can tell that, normally, failure modes and customer abuse will lead to some cosmetic and surface damages. If the device does not function properly, such as in "keypad not functioning" or "keypad broken/deformed" failure modes, it has to be returned to RMA for repair, and then RMA finds some cosmetic damage on the device at the same time. If "Housing: Cosmetic Damage" is on the left-hand side and the failures previously listed on the right hand side, it means that if a device has cosmetic damage, then it also has those failures. This is a rare case because, most of the time, customers do not return the devices for repair just because of some cosmetic damages. Only when the devices do not work properly, will customers return the devices for repair. The failure modes that lead to keypad not functioning are either customer abuses that physically break the keypad or other keypad related functional failures. Similar to housing cosmetic damage, it is not very likely that devices are returned for repair just because of a scratched LCD. The device has to have some functional failures in order to be returned, at which point the scratch will be found and repaired together with the functional failures.

(2) 3-way Relation Association Rules

As discussed in the preliminary analysis, 3-way, 4-way, and 5-way relation association rules are equally important as 2-way relations association rules, since there are a number of devices that have more than 2 failures and since the percentage of those devices is almost equal to the devices that have 2 failures. In fact, the 3-way, 4-way, and 5-way relation association rules are very similar to the 2-way relations rules.

	Relations	Lift	Support (%)	Confidence (%)	Rule
1	3	1.42	29.47%	97.18%	N80LC & N57Ke ➔ N143Ho
2	3	1.35	15.55%	92.30%	N80LC & N36Ke ➔ N143Ho
3	3	1.43	14.45%	98.01%	N57Ke & N36Ke ➔ N143Ho

Table 5.18 Significant Association Rules (3-way Relation) – 1

Again, we rank the 3-way relation association rules by confidence and support. Table 5.18 shows the top three 3-way relation association rules. The left hand side items are the combinations of LCD scratched, keypad not functioning and keypad scratched. And they all lead to housing cosmetic damages.

Because there are so many 3-way association rules generated, the rules have to be scanned and evaluated manually. The 3-way relation association rules tend to repeat the same items as 2-way relation association rules. The left hand side items are mostly the combinations of failure modes seen in 2-way association rules, and the right hand side is either "housing cosmetic damage" or "keypad not functioning" or "LCD scratched."

(3) 4-way & 5-way Relation Association Rules

Because the level of support is how frequently the combination occurs in the database, the numbers of transactions that contain 4 and 5 items are smaller, and so the support level for those rules is lower as well. Therefore, the 4-way and 5-way association rules have comparably lower support levels. Similar to the 3-way relation association rules, on the left hand side of the 4-way and 5-way association rules, there are mostly the same combinations of the functional failures and customer abuse. Additionally, the failures on the right hand side are most likely the combinations of "housing cosmetic damage," "keypad not functioning," and "LCD scratched."

In conclusion, an association rules mining is a powerful tool for discovering associations within a large set of items. The association rules discovered represent knowledge about the failure modes that are occurring together more frequently. The rules show that, normally, the top functional failures and customer abuse will lead to some cosmetic and surface damages. Therefore, the rules, especially the "one-way" rules, are efficiently describing situations where one failure mode causes another failure mode and not vice versa. This potentially allows one to identify the failure modes that are the source of devices being returned for repair.

Association rules mining is sufficient to solve certain problems. Because of the limitations of the data, the association rules discovered are the failure modes that occurred frequently during a device's lifetime. In the data set, one device is one record, but a device might be returned more than once. In future research, we may consider using one return as one record. So, instead of discovering the association of failure modes in a device's life time, we will be able to discover the association of failure modes for each return for each device. This further research could benefit the RMA process by cutting down the exam time.

5.4.2 Deactivation Classification

In this experiment, we aim to determine which failure modes and failure mode combinations will directly relate to deactivation. This is a typical classification problem: to classify devices as deactivated or still in use.

We use CART to generate the binary decision tree. Because the quantity of 165 failure modes ranges from over one hundred thousand to 1, when we input the data, the data set attribute node automatically rejected some of the failure modes with small quantities. There are about 65 FMs included in the decision tree model, which are top failures in terms of quantity. Because we are more interested in deactivation (1) than still in use (0), the rules with a higher percentage of deactivation (1) than activation (0) are considered as significant rules. Also, because this data set contains all the devices returned to RMA, the devices returned to RMA are more likely to be deactivated. Therefore, most of the decision rules have a higher percentage of deactivation. We consider the decision rules with a deactivation rate of more than 75% as very significant rules. There are 10 significant rules and ranked according to the percentage of deactivation (Table 5.19).

	FM1	FM2	FM3	1:0
1	N38Liquid Damage			98.3%:1.7%
2	N87BER Beyond Repair			90.6%:9.4%
3	N123Device Error			84.0%:16.0%
4	N123Device Error	N84Housing Silver Bezel		83.0%:17.0%
5	N145No Fault Found			81.8%:18.2%
6	N57Keypad	N143Housing Cosmetic Damage		80.9%:19.1%
7	N123Device Error	N143Housing Cosmetic Damage		80.5%:19.5%
8	N84Housing Silver Bezel	N143Housing Cosmetic Damage		79.1%:20.9%
9	N57Keypad	N143Housing Cosmetic Damage	N81Battery Connector	78.1%:21.9%
10	N80LCD	N143Housing Cosmetic Damage		77.9%:22.1%

Table 5.19 Significant Decision Rules

Therefore, the devices that contain either one failure or failure mode combination in Table 5.19 are more likely to be deactivated. The decision rules generated from the decision tree model are very similar to the association rules from the last section. Some combinations of failure modes (such as keypad not functioning and housing cosmetic damage; keypad not functioning, housing cosmetic damage and battery connector broken/deformed) not only occurred more frequently, but they also relate directly to deactivation. While cosmetic or surface damages alone do not cause deactivation, they are usually combined with some functional failures; however, the functional failures are more likely to cause deactivation. In addition, we find that there are a number of devices that are deactivated, but with no fault found. Part of the reason for this finding is the refund policy in North America and in some other places: customers can return devices and get a full refund within a period of time without giving any reasons and with no questions asked. In fact, most of the time, these returned devices are in good condition. However, these devices still have to go to RMA to be examined, but there are no faults found and no cosmetic damages. For these devices, there are no quality issues, but if the customer data are available, it will be interesting to find out which group of customers tends to return the devices within the refund period with no fault found with the devices.

There are some failures alone that will very likely cause deactivation, such as "BER Beyond Repair," "Liquid Damage," and "Device Error." In these cases, customer abuse or functional failure are critical, since they will directly lead to the devices not functioning, and they are hard or even impossible to repair. These failures are also hard to predict.

Overall, the decision tree model is suitable for such a business problem, and the decision rules generated are reasonable and will hopefully benefit the RMA and manufacturing process.

5.5 Time to Fail Analysis

This experiment is designed to compare the time to fail with the warranty period for some particular failure modes. Another purpose is to find out how time to fail changes by the failure modes. The knowledge discovered will help the manufacturer to better decide and improve the warranty period.

To analyze the time to fail, we apply the statistical and graphical tools on Table 4.2. Because time to fail is a continuous variable, ideally, if we want to predict the time to fail by using data mining methods, a regression tree would be a good choice. However, we do not have enough predictive variables available to predict time to fail, since we assume time to fail is independent of carriers, regions, and manufacturing locations. Therefore, we abandon the data mining method in this experiment and apply some simple statistical and graphical methods instead. We want to pick some important failures that were discovered in the association rule mining results and decision rules from section 5.4.

1. Keypad: Not Functioning

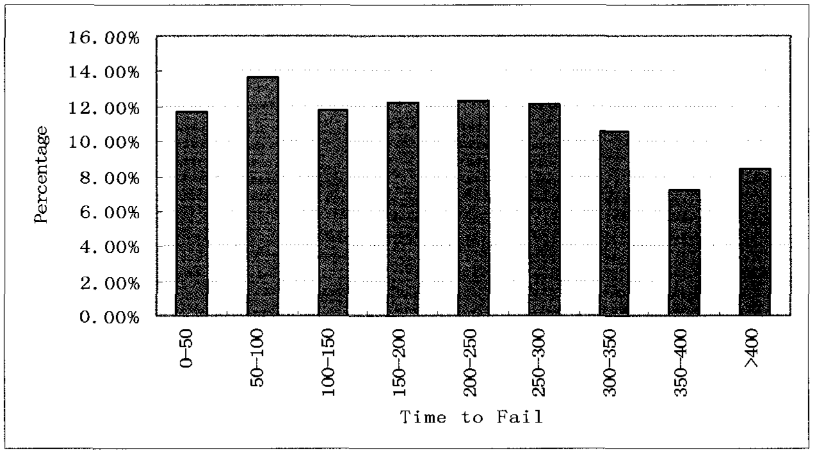


Figure 5.17 Time to Fail Distribution Diagram – Keypad Not Functioning

"Keypad not functioning" is an important functional failure. It usually directly causes the device to fail or to be deactivated. As Figure 5.17 shows, keypad not functioning can occur at any time after a device is activated. The average time to fail for this failure is 214 days. Currently, the warranty for "keypad not functioning" is 365 days. The average time to fail for those devices that failed within the warranty period is 179 days, while the average time to fail for those devices that failed after the warranty is 457 days.

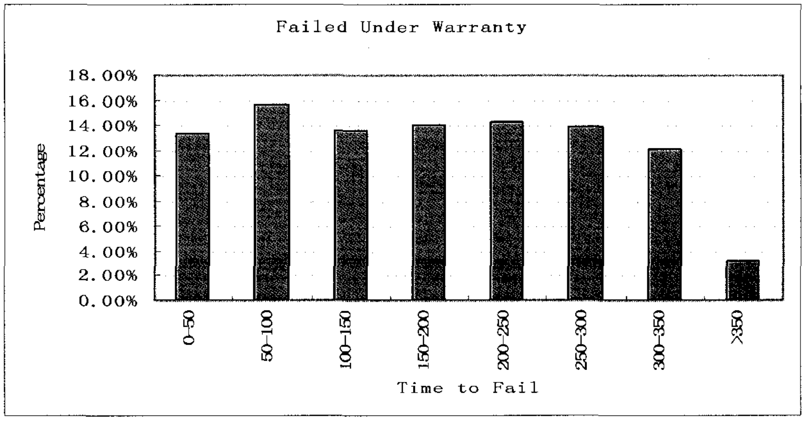


Figure 5.18 Failed Under Warranty – Keypad Not Functioning

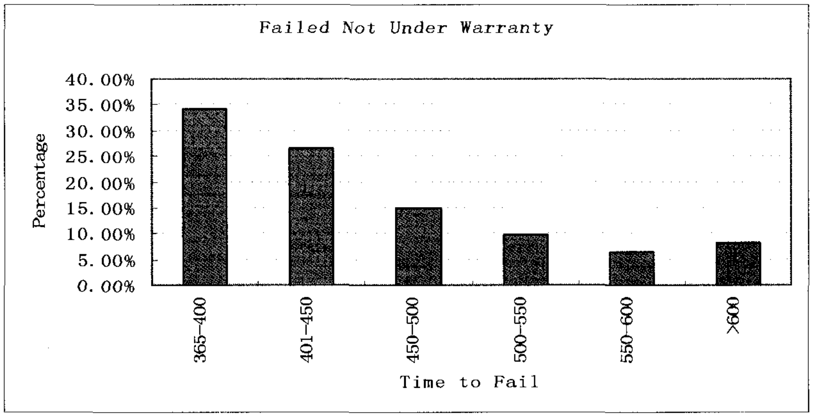


Figure 5.19 Failed Not Under Warranty – Keypad Not Functioning

We think that the organization could consider extending the warranty for the keypad to a lifetime warranty. Although we do not know the cost of repairing keypad, this cost needs to be weighed against the cost of losing the customer because when the keypad is not functioning, the device is very likely to be deactivated. Therefore, a comparison of the repair cost for the keypad versus the cost of losing a customer due to this failure would reveal whether or not the warranty should be extended. Moreover, only 10% of devices returned to repair and there is comparably small number of devices that have the keypad not functioning failure after the warranty period.

2. Device Level (Software): Device Error

“Device error” is another important functional failure; it is normally ranked as the second most important functional failure. Unlike “keypad not functioning,” the distribution of time to fail of device error is normally distributed. The average time to fail for device error is 236 days. The current warranty period for device error is 365 days. Same as "keypad not functioning," if the cost of repair for this failure is less than the cost of losing the customer because of this failure, we think the organization should consider extending the warranty to retain the customer.

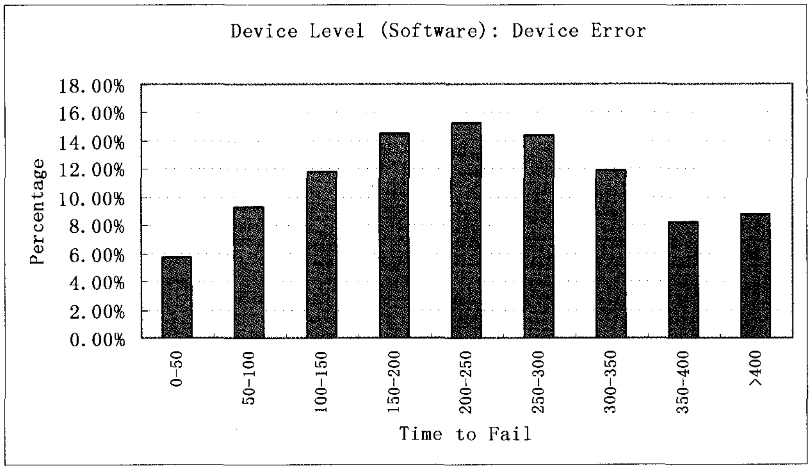


Figure 5.20 Time to Fail Distribution Diagram – Device Error

Determining a warranty period is an interesting topic. Many organizations, in many different industries, consider warranties as a strategy to maximize customer satisfaction and to maintain good customer relation. Competition is intense, especially in the telecommunications industry, and devices have comparably short life. Thus, the extension of a warranty might be a solution for maintaining the current customers and attracting new ones.

Chapter 6 Conclusion and Future Work

6.1 Conclusion

This research studied real world business problems in the telecommunications industry, specifically in a quality assurance department. This research proposed some experiment designs and aimed to find the optimal solutions for each business problem, mainly through a data mining approach. The research explored and demonstrated the potential for applying data mining to quality and reliability prediction and improvement.

Deactivation analysis and prediction is a core subject in this thesis research. But deactivation prediction is a broad subject involving many influential factors. In addition to the attributes included in our data sets, information on a customer's age, occupation, gender, device usage, and the customer's contract is also relevant and useful in deactivation prediction. Our models and results would be significantly enhanced should this information be available. So the deactivation analysis in this research was based on two components: one is general information, such as carrier and geographical information, and another is RMA related information, such as failure mode and repair information. Furthermore, age and time delay analyses were performed in this thesis.

The data, the data mining methods, and other data analysis methods were defined and applied to serve the objectives of this research. In the deactivation analysis, the experiments were designed to identify the influential factors on deactivation. The decision tree models built by CHAID, CART and C4.5 were compared and evaluated. Those three decision tree algorithms are very comparable. The results not only showed what is merely common sense and inconsequential, but they also showed that RMA related information is an important predictor to deactivation, while carrier and region are not significant in predicting deactivation. The results also showed that effective data preparation and feature selection are the keys to improving the accuracy of classification and prediction. The experiments also demonstrated that the decision tree model is a potential solution for certain classification and prediction problems. If we have access to customer information, contract information, and device usage data, we could expect the decision tree model to return better results that would directly benefit the organization.

In the age analysis, the experiments were designed to analyze the influential factors on age and how age changes by those factors. The devices were deactivated with either a long age, a short age, or an average age. We used three methods in three experiments: a binary decision tree, a regression tree, and graphical tools. The experiments identified the factors to distinguish long and short age, and illustrated how time in use changes by those factors. The results from the three experiments were consistent and showed that time delay and RMA related information are two important influential factors. The charts and the diagrams effectively represented how time in use changes by the factors, and they are easier to understand than the data mining results.

In the time delay analysis, we chose not to use a data mining approach, because there were only two attributes we believed were relevant to time delay and because it was safe to assume that time delay is irrelevant to deactivation. So we chose to use graphical tools instead. The results showed that significant differences of time delay exist between both carriers and regions.

The previous analyses proved that RMA related information (in other words, the failure modes found with the devices) is very important in determining the deactivation of devices and the age of devices. The failure mode analysis in this thesis was composed of three parts: 1) Association rules discovery, which discovers the failure modes that are more likely to occur together during the life time of the devices in order to improve the engineering design, manufacturing, and RMA process. 2) Deactivation classification, used to find which failure modes or failure mode combinations are closely related to the deactivation of the devices. 3) Time to fail analysis, which studies time to fail in order to better determine the warranty period.

The association rule mining was proved to be a powerful tool for discovering the associations of failure modes that occurred more frequently in a given device. It was observed that cosmetic damages are usually associated with critical functional failures or customer abuse. The decision tree model was efficient in identifying the failure modes or the failure mode combinations that directly cause deactivation. The time to fail analysis showed that the simple statistical method is capable of comparing the time to fail with the existing warranty period.

This research is a first step in exploring how useful data mining could be to the QA group in this organization. More and more organizations have realized the advantages of data mining, so they are now involving in data mining applications and development. The impact of data mining application is potentially great; the use of data mining techniques, combined with statistical data analysis tools, will result in huge commercial gains, particularly in large organizations.

6.2 Future Work

The research was conducted on real world cases in the telecommunications industry; it is an effort to apply data mining techniques, and to demonstrate the potential values of data mining applications. The groundwork laid by this effort could be used in the future development of a system. A complete and comprehensive system needs to be developed in order to put all the modules into one system that will keep track of different components. In the future, it would be reasonable to connect the system to the database and make data preprocessing a standard procedure. This procedure would help keep the data secure, reduce human error, and retain the consistency and reliability of the data. In order to make it more realistic, future work could also include some elaborate interfaces with more interactive functions. A reporting system would also be necessary to present the results in a comprehensible format.

In any future work, it would be beneficial to include customer information data, contract information, and device usage data. So the access to the data (e.g., customer data, device usage data and contract information data) will be appropriate, and with that data available, we expect the data mining will return better results, which will directly benefit the organization. In the future, some alternative data mining methods can be considered, such as clustering, which is the assignment of objects into group so that objects from the same cluster are more similar to each other than objects from different clusters. Clustering is very useful in term of discover the data structure. Some business problems and solutions in this thesis research can be expanded and specified. For instance, in the future, instead of discovering the association of failure modes in a device's life time, the association of failure modes for each return for each device would benefit the RMA process by cutting down the RMA exam time. The time to fail analysis can be expanded to all the major failure modes. The system will be able to recommend whether or not the current warranty period is efficient, thus reducing the cost or attracting more customers. In addition, customer complaints are an important data resource; they are direct feedback from customers regarding product quality issues. Semantic data mining would be extremely helpful in order to mine important information from customer complaints, which would save time from manually scanning and reading all the customer complaints.

Finally, documentation of the development of data mining applications and systems will be useful in the future in order to create reproducible and repeatable data mining results. Because data mining is still a new technology in the department and in the organization, encouraging more and more people to train on and to take advantage of data mining will maximize the value of this technology.

References

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile,* 12-15 September 1994 (pp. 487-499). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Al-Salim, B., & Abdoli, M. (2005). Data mining for decision support of the quality improvement process. In *Proceedings of the 11th Americas Conference on Information Systems, Omaha, USA,* 11-14 August 2005 (pp. 1462 – 1469).

Bach, M. P., & Cosic, D. (2007). Data mining usage in health care management: literature survey and decision tree application. *Medicinski Glasnik,* 5(1). Retrieved from http://www.ljkzedo.com.ba/M8_10.pdf

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* New York: Chapman & Hall.

Chen, W. C., Tseng, S. S., & Wang, C. Y. (2004). A novel manufacturing defect detection method using data mining approach. In *Proceedings of the 17th International Conference on Innovations in Applied Artificial Intelligence, Ottawa, Canada,* 17-20 May 2004 (pp. 77-86). New York, USA: Springer Verlag Inc.

Confusion matrix. (n.d.). Retrieve February 11, 2008, from http://en.wikipedia.org/wiki/Confusion_matrix

Damgaard, C. (n.d.). *Gini coefficient.* Retrieved January 29, 2008, from http://mathworld.wolfram.com/GiniCoefficient.html

Daskalaki, S., Kopanas, I., Goudara, M., & Avouris, N. (2003). Data mining for decision support on customer insolvency in telecommunications business. *European Journal of Operational Research,* 145(2), 239-255.

Dunham, M. H., Xiao, Y., Gruenwald, L., & Hossain, Z. (2001). *A survey of association rules.* Retrieved January 5, 2008, from http://rodin.cs.uh.edu/~ceick/6340/grue-assoc.pdf

Entropy (information theory). (n.d.). Retrieved November 29, 2007, from http://en.wikipedia.org/wiki/Information_entropy

Feldman, D., & Gross, S. (2003). Mortgage default: classification trees analysis. *The Journal of Real Estate Finance and Economics,* 30(4), 369-396.

Hamilton, H., Gurak, E., Findlater, L., & Olive, W. (2003). *Knowledge discovery in databases: C4.5 tutorial.* Retrieved October 25, 2007, from University of Regina, Department of Computer Science Web site: http://www2.cs.uregina.ca/~dbd/cs831/index.html

Han, J., & Kamber, M. (2001). *Data mining: concepts and techniques.* San Francisco: Morgan Kaufmann.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics,* 29(2), 119-127.

Kohavi, R., & Quinlan, J. R. (1999). *Decision tree discovery.* Retrieved October 19, 2007, from http://ai.stanford.edu/~ronnyk/treesHB.pdf

Kotsiantis, S., & Kanellopoulos, D. (2006). *Association rules mining: a recent overview.* Retrieved November 12, 2007, from University of Patras, Department of Mathematics Web site: http://www.math.upatras.gr/~esdlab/en/members/kotsiantis/association%20rules %20kotsiantis.pdf

Lewis, R. J. (2000). *An introduction to classification and regression tree (CART) analysis.* Retrieved January 19, 2008, from http://www.saem.org/download/lewis1.pdf

MacDougall, M. (n.d.). *Shopping for voters: using association rules to discover relationships in election survey data.* Retrieved December 2, 2007, from http://www2.sas.com/proceedings/sugi28/122-28.pdf

Mining frequent patterns. (n.d.). Retrieved March 19, 2008, from www.cefns.nau.edu/~dl259/teaching/cs445/LectureNotes/05_1.ppt

Mous, L. (2005). *Predicting bankruptcy with discriminant analysis and decision tree using financial ratios.* (Bachelor's Thesis, Erasmus University, 2005). Retrieved from www.tbm.tudelft.nl/live/pagina.jsp?id=bff0eff5-fe86-47b9-8464-ff2a04478b5c &lang=en&binary=/doc/lonneke-bachelor.pdf

Oracle. (2008). *Oracle data miner.* Retrieved from http://www.oracle.com/technology/products/bi/odm/odminer.html

Pearson's chi-square test. (n.d.). Retrieved November 19, 2007, from http://en.wikipedia.org/wiki/Pearson%27s_chi-square_test

PL/SQL. (n.d.) Retrieved December 10, 2007, from http://en.wikipedia.org/wiki/PL_SQL

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning,* 1(1), 81-106. doi: 10.1007/BF00116251

Quinlan, J. R. (1993). *C4.5 programs for machine learning.* San Francisco, USA: Morgan Kaufmann Publishers Inc.

Rosset, S., Murad, U., Neumann, E., Idan, Y., & Pinkas, G. (1999). Discovery of fraud rules for telecommunications – challenges and solutions. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, USA,* 15-18 August 1999 (pp. 409-413). New York, USA: ACM.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man and Cybernetics,* 21(3), 660-674. doi: 10.1109/21.97458

SAS. (2008). *Data mining with SAS Enterprise Miner.* Retrieved from http://www.sas.com/technologies/analytics/datamining/miner/

SAS Institute, Inc. (2004). *Getting started with Enterprise Miner 4.3.* Retrieved from http://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_91/em_gs_7281.pdf

SPSS. (2008). *Data mining improves decision making.* Retrieved from http://www.spss.com/data_mining/index.htm

StatSoft. (2008). *Statistica data miner & customized solutions.* Retrieved from http://www.statsoft.com/products/dataminer.html

Tang, T., Zheng, G., Huang, Y., Shu, G., & Wang, P. (2005). A comparative study of medical data classification methods based on decision tree and system reconstruction analysis. *IEMS,* 4(1), 102-108. Retrieved from http://ie.kaist.ac.kr/iems/contents/vol4no1/4-1-10.pdf

Testing utility of model – F-test. (2004). Retrieved June 9, 2008, from http://www.public.iastate.edu/~alicia/stat328/Multiple%20regression%20-%20F%20test.pdf

Timofeev, R. (2004). *Classification and regression trees (CART) theory and application.* Retrieved March 12, 2008, from http://edoc.hu-berlin.de/master/timofeev-roman-2004-12-20/PDF/timofeev.pdf

Variance reduction. (n.d.). Retrieved December 10, 2007, from http://en.wikipedia.org/wiki/Variance_reduction

Wilkinson, L. (1992). *Tree structured data analysis: AID, CHAID and CART.* Retrieved February 1, 2008, from http://bus.utk.edu/stat/DataMining/Tree%20Structured%20Data%20Analysis%20(SPSS).pdf

Witten, I. H., & Frank, E. (n.d.). *Data mining: the practice.* Retrieved January 15, 2008, from http://www.informatik.uni-freiburg.de/~ml/teaching/ss07/dmPractical/slides/ChapterAlgos1-2x3.pdf

Wu, W. (2005). *An integrated CRM data mining method for predicting best next offer.* Unpublished master's thesis, Dalhousie University, Halifax, Canada.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., et al. (2007). Top 10 algorithms in data mining. *Knowledge and Information Systems,* 14, 1-37. doi: 10.1007/s10115-007-0114-2

Yohannes, Y., & Webb, P. (1999). *Classification and regression trees, CART.* Retrieved March 18, 2008, from www.ifpri.org/pubs/microcom/micro3.pdf

Zhou, C., Nelson, P. C., Xiao, W., Tirpak, T. M., & Lane, S. A. (2001). An intelligent data mining system for drop test analysis of electronic products. *IEEE Transactions on Electronics Packaging Manufacturing,* 24(3), 222-231. doi: 10.1109/6104.956808