April 2010

Predicting Policyholder Behavior and Benefit Utilization: An Analysis on Long-Term Care Insurance

Ashleigh Rita Smeal
Worcester Polytechnic Institute

Heather Lee Standring
Worcester Polytechnic Institute

Jie Bai
Worcester Polytechnic Institute

Xinyi Zhang
Worcester Polytechnic Institute

Predicting Policyholder Behavior and Benefit Utilization

An Analysis on Long-Term Care Insurance

A Major Qualifying Project, submitted to the faculty of
Worcester Polytechnic Institute in partial fulfillment of the requirements for the
Degree of Bachelor of Science

Submitted by:

Jie Bai

Ashleigh Smeal

Heather Standring

Xinyi Zhang

Submitted to:

Project Advisors:
Prof. Jon P. Abraham
Prof. Helen G. Vassallo

Project Liaison:
Don Charsky, Ability Resources, Inc.

Spring 2010

Abstract

In order to better serve their customers, a project to create a methodology for identifying variables that could indicate future long-term care insurance usage was commissioned by Ability Resources, Inc. As a basis for constructing a predictive model, tools such as SAS and Excel were implemented. A k-means clustering algorithm in SAS was utilized to group policyholders with similar characteristics, and a performance evaluation was executed in Excel. Together, these processes created a tool that determined the impact each characteristic had on policyholder benefit utilization. The validity of the process was assessed by applying it to supplemental data generated by the team. After several trials, the Variable Identification Procedure proved accurate.

Authorship

This project was completed through a combined effort from all group members. Each individual contributed equally throughout the project. Tasks, including researching methods, executing procedures, analyzing data, and writing the report were divided amongst the members.

Acknowledgements

This MQP group has many individuals and organizations to thank for aiding in the completion of this project.

We would like to acknowledge the contributions of the following:

Don Charsky, our project sponsor and CEO of AbilityRe, for conceptualizing this project, providing direction, and guidance along the way.

Professor Abraham and Professor Vassallo, our advisors, both provided critical input throughout the project. Their expertise, patience, and feedback are greatly appreciated.

Jaclyn Dempsey, of AbilityRe, provided the group with a multitude of data and explanations that were critical in framing our understanding of the topic.

Fred Yosua, President of Ability Insurance Company, is thanked for his help in calibrating the methods developed by offering insights based on experience in the field.

Jim DuEst, Director of Database Management and Development, and Mike Noce of AbilityRe, contributed to our project by offering background and insight on long-term care insurance.

Mike Batty and Chris Stehno, Deloitte employees, offered key insights into predictive modeling techniques and served as a liason between the team and a supplemental data aggregator.

Lance Rubeck, of Equifax Marketing Services, educated the team on the potential value added to a study by using supplemental data.

Professor Vermes and Professor Weekes, of the WPI Mathematical Sciences Department, suggested possible methods for creating a surface.

Criselda Toto, of the WPI Mathematical Sciences Department, provided explanations and guidance in using SAS.

Table of Contents

Table of Figures

Table of Tables

Executive Summary

Ability Resources, Inc (AbilityRe) is a reinsurer located in Framingham, Massachusetts. As a part of their business model, the company has purchased a block of Long-Term Care Insurance policies. Through an analysis of customer data from this block of policies, AbilityRe recognized an opportunity to improve their services. As a result, this project was commissioned to evaluate and understand the behavior of policyholders and to determine a method to predict future benefit usage.

The goal of this project was to identify policyholder variables that may indicate or aid in the prediction of future spending and usage. To meet this goal, the following objectives were outlined: establish a data set of policyholder information that includes the combination of records held by AbilityRe along with supplemental data, outline a clustering methodology that would group policyholders based on common characteristics, and perform an evaluation to determine the impact each variable had on differentiating policyholders. However, as a result of HIPPA regulations, the focus of this project was redirected from identifying variables in a data set to developing and testing a specific procedure, which could be used to identify variables in the future if a data set were available.

The methodology that was developed by the group is called the Variable Identification Procedure and consists of three steps. The first step is to cluster policyholders based on one characteristic. This was completed in SAS, which applies a k-means clustering methodology to the input data. Clustering is performed by dividing the observations into k clusters based on the closest mean, then recalculating the k centroid values. This step helped to determine which variables may be useful in differentiating policyholders. The next step occurred simultaneously and consisted of each individual in the data set being given a score from zero to one hundred. To

do this, a subset of policyholders was scored by AbilityRe, then this information was used to define a least squares plane. The remaining policyholders were scored by extrapolation from this plane using the financial ratio and years owning the policy before going on claim. Finally, the scoring and clustering were evaluated using a macro in Microsoft Excel, which calculated the difference between the line of the average score and the cumulative average score of each cluster.

Several simulated supplemental data sets were generated and tested to prove the accuracy of the Variable Identification Procedure. Each data set was designed differently to test the capacity of the process at differing levels of variable randomness or predictive capability. After applying the Variable Identification Procedure in four trials, the method proved it could successfully distinguish between random and predictive variables. This method, when applied to policyholder supplemental data, will allow AbilityRe to better understand the behavior and benefit usage of its customers.

Chapter 1: Introduction

As the national economy continually develops throughout the 21st century, it is now more important than ever that businesses align their products with consumer needs. By improving operating methodologies and the company's awareness of external factors, the dual issue of satisfying customers, while maintaining profits, can be addressed. It is through an understanding of the consumers' behavioral tendencies, lifestyle choices, and individual characteristics that a company can offer a valuable product or service. This information can help to streamline company activities by focusing on only those products that will be mutually beneficial for the company and the consumer. Insurance is one industry that is improving its services in this way.

The insurance industry provides risk management products in a wide breadth of areas to protect the consumer against loss. Within insurance companies, predictive modeling is becoming a more common practice as a way to improve processes and products. Such techniques are allowing businesses to "innovate, become more efficient, make more accurate and consistent decisions, and grow profitably."[1] A predictive modeling methodology uses algorithms to estimate unknowns and allows for the combination of recorded policyholder data with supplemental data from a variety of sources. Identifying these macroeconomic trends requires skill and understanding in the field of mathematics. Such techniques can lead to proactive business practices that increase efficiency by minimizing cost and maximizing consumer value. Although the most apparent benefit is the cost savings, the effects of predictive modeling have also been shown to improve the insured's and insurer's experiences.[2]

---

[1] Mike Batty et al. (2009)."Bringing Predictive Models to Life." Print.
[2] Mike Batty et al. (2009)."Bringing Predictive Models to Life." Print.

In general, insurers fall into one of two types: Property and Casualty, or Health and Life. Although predictive modeling has seen extensive usage in Property and Casualty, it is now being applied within the Health and Life sector as well. One form of Health and Life Insurance is Long-term care insurance (LTCI). This form of coverage provides some financial security to lessen the impact that the cost of long-term care can have. Long-term care refers to the assortment of services available to assist individuals unable to care properly for themselves, while allowing them to maintain some level of independence. Long-term care may be required as a result of mental or physical impairment. In 2003, the chance that an individual over the age of 65 would meet the eligibility requirements for needing long-term care assistance at some point in their lifetime was 68%.[3]

Ability Resources, Inc. (AbilityRe) is a reinsurer that specializes in LTCI policies. Through observation of policyholder satisfaction and their usage of the policy, AbilityRe identified an opportunity to improve its services. They recognized the importance of understanding the insured's needs in order to provide a more valuable product. It was the goal of this project, in collaboration with AbilityRe, to identify policyholder variables that may indicate or aid in the prediction of future spending and usage. Several objectives were outlined in order to meet this goal. First, a data set of policyholder information was established. This included the combination of records held by AbilityRe along with supplemental data. Second, a clustering methodology outlined by the group would group policyholders based on common characteristics. Finally, an evaluation was performed to determine the impact each variable had on policy usage.

---

[3] Family Caregiver Alliance. (2005). *Selected Long-Term Care Statistics.* Retrieved on March 27, 2010 from http://www.caregiver.org/caregiver/jsp/content_node.jsp?nodeid=440.

Improving the methods for analyzing and predicting policyholder usage will aid AbilityRe in maintaining and increasing the positive impact their products and services offer.

Chapter 2: Background

For long-term care insurance companies, providing more valuable services and products starts with understanding the behavior of the insureds. It is the goal of this project to identify behavioral patterns or characteristics that would aid in the prediction of future claims and benefit usage. In order to achieve this goal, a basic understanding of long-term care insurance must first be achieved. This section provides an overview of long-term care, long-term care insurance policies, the conditions that must be met for a claim to be initiated, and an introduction to the mission of Ability Reasources, Inc.

2.1. Long-Term Care Overview

Approximately nine million people over the age of 65 will require some form of assistance due to a disability or chronic illness this year.[4] In order to facilitate this increasing need, a variety of services exists to help people of all ages with tasks that can usually be done without assistance. Activities such as dressing, bathing, using the bathroom, and eating are a few of the everyday events that may be difficult for people requiring long-term care.[5] Long-term care is the assortment of services that work to support those needing this type of assistance over an extended period of time. This can include living in a nursing home, living in a community or assisted living facility, or having home care.[6] The goal of these services and long-term care is to allow the people requiring them to maintain some independence and functionality.[7]

---

[4] U.S. Department of Health and Human Services. (March 2009). *Long Term Care.* Retrieved on November 5, 2009 from http://www.medicare.gov/longTermCare/static/home.asp
[5] U.S. Department of Health and Human Services. (October 2008). *Understanding LTC Basics*. Retrieved on November 5, 2009 from http://www.longtermcare.gov/LTC/Main_Site/Understanding_Long_Term_Care/Basics/Basics.aspx
[6] U.S. Department of Health and Human Services. (March 2009).
[7] U.S. Department of Health and Human Services. (October 2008).

2.2. Long-Term Care Insurance

The need for forms of long-term care (LTC) to medically assist the aging population has always been present in society. However, a recent increase in the demand for services in this sector has become apparent. Several factors could be contributing to the increased usage of LTC including longevity of the population and family structure.[8] Long-term care insurance (LTCI) is a form of coverage that will aid in alleviating some of the out-of-pocket financial burden LTC providers or facilities can have on the elderly individuals in need. Additionally, LTCI provides a unique solution to the problems raised by Medicare and Medicaid.

Medicare coverage is intended for short-term care, such as that which is required after hospitalization. Medicaid becomes available only after all other personal assets have been depleted.[9] In fact, long-term care insurance was first seen on the market in the second half of the twentieth century and was created by insurance companies to offer coverage in situations where Medicare and Medicaid were not applicable.[10] This form of insurance is relatively new to employee compensation packages. In fact, the first group long-term care insurance contract was written in 1987. However, by 2003 approximately 13% of all full-time workers in the public and 19% of workers in private establishments were offered this benefit.[11] As the costs for all forms of long-term care are rapidly increasing, experts agree that it is becoming more important for individuals to invest in some form of LTCI.

Similar to other forms of insurance, long-term care insurance requires premiums to be paid by the policyholder on a regular basis. These premiums can be very costly depending on the

---

[8] Long Term Care Insurance Tree. (2009). *What are ADLs?* Retrieved October 8, 2009, from http://www.longtermcareinsurancetree.com/ltc-basics/what-are-adls.html.
[9] W. Konrad. (2009, June 26). Getting Insurance for One's Frailest Years. *The New York Times*.
[10] Pfuntner, J., & Dietz, E. (2004, January 28). *Long-term Care Insurance Gains Prominence*. Retrieved October 8, 2009, from United States Bureau of Labor Statistics: http://www.bls.gov/opub/cwc/cm20040123ar01p1.htm.
[11] Pfuntner & Dietz. (2004).

type of coverage that is elected to be included in the plan and the age at which to policy is purchased. Some common plan provisions include the maximum daily benefit, types of care covered, and length of coverage. One component of the LTCI plan is the maximum daily benefit, which is the amount of coverage that will be paid daily once the policyholder is on claim status. In 2009, the average cost of care for one day in a semi-private room at a nursing home was $183.25 and $46.22 an hour for in-home assistance from a nurse.[12] This average cost can see significant increases in metropolitan areas, resulting in a substantial bill for even one day of care. The maximum daily benefit would be used to pay a portion of the long-term care-giving bill. One way to ensure the maximum daily benefit remains a useful amount is to include an inflation protection option within the policy. Insurance riders such as this one can be purchased by the policyholder and allow amendments to the coverage provided in the plan to be made over time.[13]

Another stipulation of a LTCI policy is the types of coverage that would be provided. Some plans may provide coverage for a specific type of care such as home health care attendants, assisted living facilities, or nursing homes, while other insurance policies provide coverage for all types of care. A final example of a plan specification is the duration of benefits. Some plans may provide a limited number of years of benefit payments once a policyholder goes on claim while others offer lifetime benefits.[14] The cost of plan premiums can vary greatly depending on the specific coverage forms provided by the plan.

Benefits are distributed to policyholders once an illness or disability has been recognized and treated for a certain amount of time, this is known as an elimination period. After this time, a

---

[12] Genworth Financial. (2009, April). *Genworth 2009 Cost of Care Survey*. Retrieved October 8, 2009, from Genworth Financial:http://www.genworth.com/content/etc/medialib/genworth_v2/pdf/ltc_cost_of_care.Par.8024.File.dat/cost_of_care.pdf.
[13] America's Health Insurance Plans. (2004). *Guide to Long-Term Care Insurance.* Retrieved October 8, 2009, from http://www.ahip.org/content/default.aspx?docid=21018.
[14] W. Konrad. (2009).

claim can be filed with the insurance provider. Long-term care insurance policies cover a broad spectrum of services when an individual is no longer able to perform activities of daily living (ADLs) or is cognitively impaired.[15] Providing health care for individuals in these situations can become very costly and long-term care insurance helps to offset the financial burden incurred.

2.2.1. Activities of Daily Living

Indicators such as medical record and current health conditions are unable to provide a complete assessment of the functionality and self-sufficiency of an individual. There was a need for a more comprehensive way to determine the daily capabilities of a person. As a result, researchers developed the activities of daily living (ADLs), which analyze quality of life as well as the ability for an individual to live safely and independently. The ADLs are a list of actions that are considered the basics of independent self-care.[16] The Katz Activities of Daily Living Scale is the most commonly used measure of ADLs although there are over forty-three similar scales in use. In the Katz scale, ADLs are defined to be bathing, dressing, toileting, transferring, continence, and feeding.[17] Although difficulty completing activities of daily living may be seen in all age groups, it is predominantly recognized in the elderly population especially those 85 years of age or older.[18]

An increasing number of private long-term care insurance providers as well as public long-term care programs, such as Medicaid, rely on ADLs to determine benefit eligibility. On average, difficulty or inability to complete any two of the activities would render an individual

---

[15] America's Health Insurance Plans. (2004).
[16] Long Term Care Insurance Tree. (2009).
[17] J. M. Wiener, R. J. Hanley, R. Clark, & J. F. Van Nostrand. (1990). *Measuring the Activities of Daily Living: Comparisons Across National Surveys.* United States Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation Office of Social Services Policy.
[18] J. M. Wiener, R. J. Hanley, R. Clark, & J. F. Van Nostrand. (1990).

eligible for benefits.[19] Testing an individual's ability to complete these activities is a reliable method to assess their ability or disability status because it measures not only their physical capabilities, but also their cognitive function. A second advantage of using this scale is that the inability to complete certain ADLs is indicative of the services required, which aids in determining the most effective form of care for an individual.

2.2.2. Cognitive Impairment

Long-term care insurance policies not only cover physical disabilities, but also some mental diseases may be covered as well. Cognitive impairment, sometimes known as cognitive dysfunction, is defined as abnormally poor mental function, which may be exhibited by symptoms such as confusion, forgetfulness, or difficulty concentrating. One phrase that has been used to describe this type of impairment is 'brain fog', because "it can feel like a cloud that reduces your visibility or clarity of mind."[20] Although, cognitive dysfunction was rarely diagnosed in the past, it has recently become a more documented handicap. With this recognition, a clear distinction has been made between actions of individuals with fatigue and depression, and those that have more complicated mental impairments. Research has indicated that this is a progressive disease, meaning that individuals exhibiting early onset symptoms are likely to have a more marked dysfunction in the future.

Several medical and physical conditions can be attributed to the cause of cognitive dysfunction. The extensive list includes heavy metal poisoning, menopause, and sleep disorders. Additionally, there are many types of cognitive impairment; about 100 types have already been

---

[19] Long Term Care Insurance Tree. (2009).

[20] Lawrence Wilson. (2008). Brain Fog. The Center for Development. Retrieved on October 23, 2009 from http://www.drlwilson.com/Articles/brain_fog.htm.

identified. Likewise, the range of symptoms is extensive, with 4,035 symptoms already known and being studied.[21] Several methods for identifying and diagnosing mental impairment are utilized. Most commonly, a checklist is employed, such as the one seen in Table 1.

Treatments for conditions that have been diagnosed typically involve correcting any underlying medical condition, and memory and focus exercises. As one of the side effects of mental impairment is slowed performance of an individual, patient progress or improvement may take a significant amount of time.

1. Appearance of complaints or objective evidence of memory problems.
2. Traditionally normal daily living skills are deteriorating.
3. Thinking ability, other than memory, is not normal.
4. Increased levels of depression.

Table 1 How Doctors Diagnose Mild Cognitive Impairment[22]

Some long-term care policies include triggers for both ADLs and cognitive impairment. Triggers are conditions, specified by the insurance company that must be present before the policy is eligible to be activated. Under a cognitive impairment trigger, coverage starts when the policyholder has been certified to require substantial supervision to protect from threats to personal health and safety.[23]

2.2.3. Benefit Utilization

Long-term care insurance benefits can be utilized in a variety of ways, some of which are more commonly known than others. For example, nursing home care is frequently used in the United States and provides those in need with medical attention, therapy, and nurses at all hours

---

[21] Health Grades Inc. (2009). *Cognitive Impairment*. Retrieved on October 09, 2009 from http://www.wrongdiagnosis.com/sym/cognitive_impairment.htm.
[22] John Morley. (2008). *Managing Cognitive Dysfunction*. Retrieved on October 09, 2009 from http://www.thedoctorwillseeyounow.com/articles/senior_living/cogdys_6/.
[23] ElderLawNet, Inc.(2008). *Long-Term Care Insurance*. Retrieved on October 09, 2009 from http://www.elderlawanswers.com/elder_info/elder_article.asp?id=2595.

of the day. However, this type of service is quite expensive and tends to be unaffordable for many. Other types of benefit utilizations include providing nurses, certified nursing assistants, physical, occupational, and respiratory therapists, and home health aides or homemakers. Two types of care can be distinguished, informal and formal care. Informal care can generally be administered or delivered to the policyholder's home by family or friends. Formal care is typically provided in settings such as a home, adult day services center, assisted living facilities, nursing home, hospice facility, or some combination of these.[24]

Long-term care services may also be received in a continuing care retirement community. This type of setting usually provides housing, services, and various levels of long-term care when needed, all in one location and to the level required to meet the needs of the residents.

Long-term care policies may provide benefits by offering a fixed daily amount of money or through reimbursement of the cost of care up to a daily maximum. Additionally, most policies include the option to name a proxy to act on the policyholder's behalf in the case that the policyholder has lost the ability to file claims.

2.2.4. Insurance Riders

Suppose that every individual who bought an insurance policy was able to create his or her own policy. This would result in hundreds of policies with a wide range of benefits and eligibility constraints. In order to keep the number of policy types for each insurer low, while still allowing consumers to customize their policies, insurers have products that are known as insurance riders. Essentially, riders are amendments that can be appended for an additional cost

---

[24] Metropolitan Life Insurance Company. (2009). *The Essentials of Long-Term Care Insurance*. Retrieved on October 09, 2009 from www.metlife.com/.../long-term-care essentials/mmi-long-term-care-insurance-essentials.pdf.

to a consumer's base policy. These products offer a wide range of services, from protecting against inflation to allowing couples to combine their individual benefits.[25]

The riders offered on a policy are dependent on the insurer selling the policy; however, there are several riders common in the industry. While the riders mentioned in this section are typical, they may vary slightly from one insurer to another. One of the most well known long-term care riders is the inflation rider. Given that premiums are lower and more affordable now than they will be in the future, consumers are urged to buy long-term care policies many years before they are expected to use their benefits. Because of this, inflation plays a crucial role in the benefits that a policyholder will receive once on claim.[26] The inflation rider allows policy benefits to increase at a certain rate, usually compounded around 5%, making it possible for the benefits to maintain their value with the increasing cost of long-term care.

Another rider that is common among long-term care insurers is the restoration of benefits rider, which, depending on the insurer, may also be built into a policy. Policyholders with this rider are rewarded for recovering after using a portion of their benefits. If a policyholder goes on claim, then recovers, and goes off claim for a certain period of time without needing long-term care assistance, the benefits that they used will be restored. [27] This means that their policy value can be brought back to the initial value.

Another rider in long-term care insurance policies is the shared care rider, which can be added if a couple has two separate long-term care insurance policies. While this is seen frequently, it is also typical for some insurance companies to sell shared policies as well, which

---

[25] J. Brown, & A. Goolsbee. (June 2002). Does the Internet Make Markets More Competitive? Evidence from the Life Insurance Industry. *The Journal of Political Economy*. 110, 3, 481.
[26] M. Cohen, J. Miller, & M. Weinrobe. (August 2002). Inflation Protection and Long-Term Care Insurance: Finding the Gold Standard of Adequacy. Retrieved on October 8, 2009 from http://assets.aarp.org/rgcenter/health/2002_09_inflation.pdf.
[27] P. Shelton. (2003). Long-Term Care: Your Financial Planning Guide. Kensington Publishing Corp.: New York, NY. 53-54

provide benefits similar to that of the shared care rider. Either way, this option allows couples to withdraw their benefits from one combined pool. Thus, if one of the persons requires more benefits than was expected, the partner's benefits can be used. Additionally, if one partner dies, his remaining benefits can be added to the benefits of the living partner.[28] While this rider is good in the case that one of the partners needs more coverage than expected, it can be problematic if one partner uses all of the benefits, essentially draining both policies.

Finally, the return of premium benefit rider is a product that generally appears among long-term care insurer benefits. When a policyholder whose contract includes this rider passes away, the premiums that he paid over his lifetime, less the claims made, will be returned to a beneficiary designated by the policyholder.[29] In addition, all of the premiums returned to the beneficiary are paid out tax-free. The additional cost for this rider is significant when compared to the other riders; however, it varies widely among insurers.

2.3. Ability Resources, Inc.

Ability Resources, Inc. (AbilityRe) was founded in 2007 and is located in Framingham, Massachusetts. This company includes reinsurers and insurance services, which provide strategic solutions to insurers in a difficult market. AbilityRe strives to insure quality and professionalism in delivering upon policyholder obligations. The complete company structure can be seen in Figure 1.[30]

---

[28] The Prudential Insurance Company of America. (September 2008). Long Term Care Product Guide. Retrieved on October 8, 2009 from http://www.nfn.crumplifeinsurance.com/BISYSdocs/ltc/LTC%20EVOLUTION%20Product%20Guide.pdf.
[29] S. K. Davidson. (March 2006). US Patent No. 20060059020A1. Washington D.C.: US Patent and Trademark Office.
[30] Ability Resources, Inc. *Company Profile*. Retrieved October 2009, from Ability Resources, Inc.: http://www.abilityresources.com/.

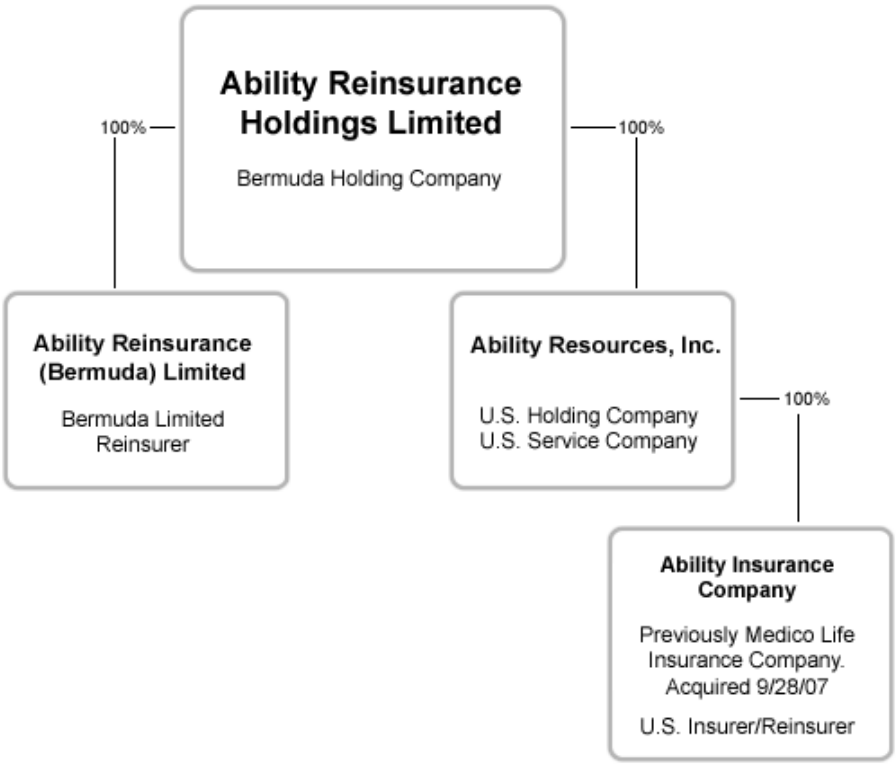


Figure 1 AbilityRe Company Structure[31]

2.3.1. Claims Process

Although many insurers have similar processes for paying out a claim, it is important to understand an individual insurer's claims process in order to grasp what happens on both the company and policyholder level. In addition, in order to suggest possible adjustments and improvements it is necessary to research the current methods used. AbilityRe has a claims process currently in place.

The first necessary step for a claim to be paid to the policyholder is for the claims request form to be filled out. This form is to be completed by either the policyholder, or a caregiver if necessary, and sent to AbilityRe. In addition to asking for the policyholder's basic information such as name and policy number, the form inquires the level of assistance that is necessary for

[31] Ability Resources, Inc.

the policyholder to complete the six activities of daily living and the type of care that is required, such as nursing home or home health care.

After submitting the form, the current process at AbilityRe requires that the policyholder receive a visit from a nurse in order to ensure that the person is eligible to receive benefits. Due to the fact that AbilityRe acquires policyholders that can reside anywhere in the country, it may be difficult to ensure that nurses are available to visit each policyholder. To help with this, a system has been established that allows insurers to find a nurse trained in long-term care in areas throughout the country. For AbilityRe, this means that when a policyholder makes a claim, the company requests a nurse in the area where the claimant lives. Once a nurse is assigned to the claimant, the nurse will be provided with the person's benefits and coverage plan. This allows the nurse to become familiar with the policyholder's coverage before the visit. The nurse will then observe the policyholder at home, determining whether or not the eligibility requirements set forth in the contract are met. If it is determined that the policyholder is eligible to begin receiving claims, the claimant or caregiver indicates to the company the form in which the benefits are to be paid out.

Chapter 3: Methodology

The primary goal of this project was to work with Ability Resources, Inc. (AbilityRe) to identify variables in long-term care insurance policyholder data that may indicate or aid in the prediction of future spending and usage. This project focused on several main objectives. One was to gather and review the policyholder information maintained by AbilityRe. A second objective was to supplement the existing data with information that could be obtained from external sources. The extent to which additional information was gathered was limited by public accessibility, fees, and legal restrictions associated with obtaining such data. Once a comprehensive data set was compiled, a type of trend analysis known as k-means clustering was performed using several different policyholder characteristics. Following this, an evaluation of the results obtained from the clustering was done to reveal which characteristics proved to be the most effective possible predictor variables. As a result, the team developed predictions of future claim amounts and policy usage. Finally, a profile explaining the behavior of policyholders in regards to their action paths and motivations was drafted. The overall methodology for this project is represented by the flow chart in Figure 2.

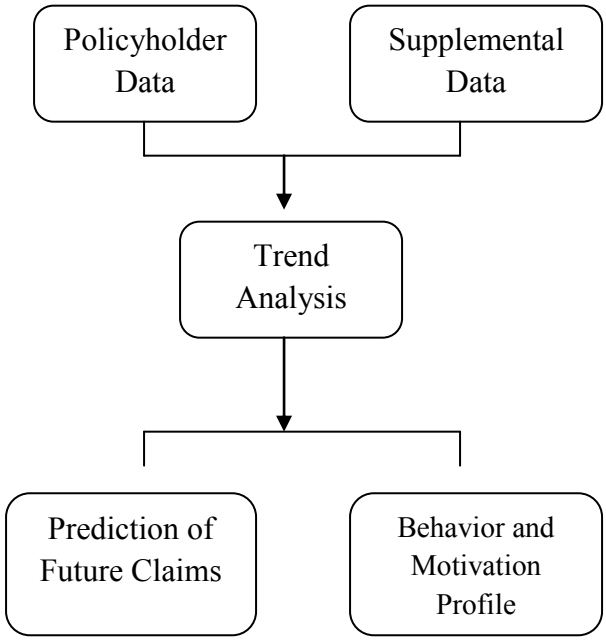


Figure 2 Methodology Flow Chart

The methodology was executed over the course of twenty-eight weeks, starting with the gathering of information, the data analysis, and then the creation of the deliverables. The timeline of this project can be seen in Appendix A: Project Timeline.

3.1. Data Set Organization

The first objective of this project consisted of gathering the information on policyholders of long-term care insurance stored by AbilityRe. Once this data was collected, the team worked on data set organization. This process occurred in two phases. First, the facts and records provided were reconciled and checked for accuracy. Second, the information was summarized and pictorially described through graphs and charts. Both of these steps helped the group to understand the data that was provided and recognize any gaps in necessary information.

Upon receiving the data from AbilityRe, the team worked to ensure its accuracy and usability by performing some basic checks. The data was quite extensive and contained a wide breadth of information on each policyholder. Every category for which information was provided

was considered a variable by the team. Some variables were information directly provided by the policyholder to the insurance company, which are referred to as non-calculated fields. Others were fields that have been calculated by the insurance company from information on file. The team began by working to understand the data dictionary that explains the many variables provided. Next, computations were performed to ensure that the calculated fields have the correct information listed based on the facts provided in the non-calculated fields. Other checks may have been done as well to ensure the file was completely reconciled and that it was ready to be used in an analysis.

In the second phase, the data was summarized by creating graphs, showing pictorially the trends in the data. The creation of charts using one or more variables increased the focus on these areas and revealed the mean, median, and mode of the distribution of policies about these variables. An example of a graph that would help to describe the data is the distribution of long-term care insurance policies based on the purchase age, which can be seen in Figure 3.

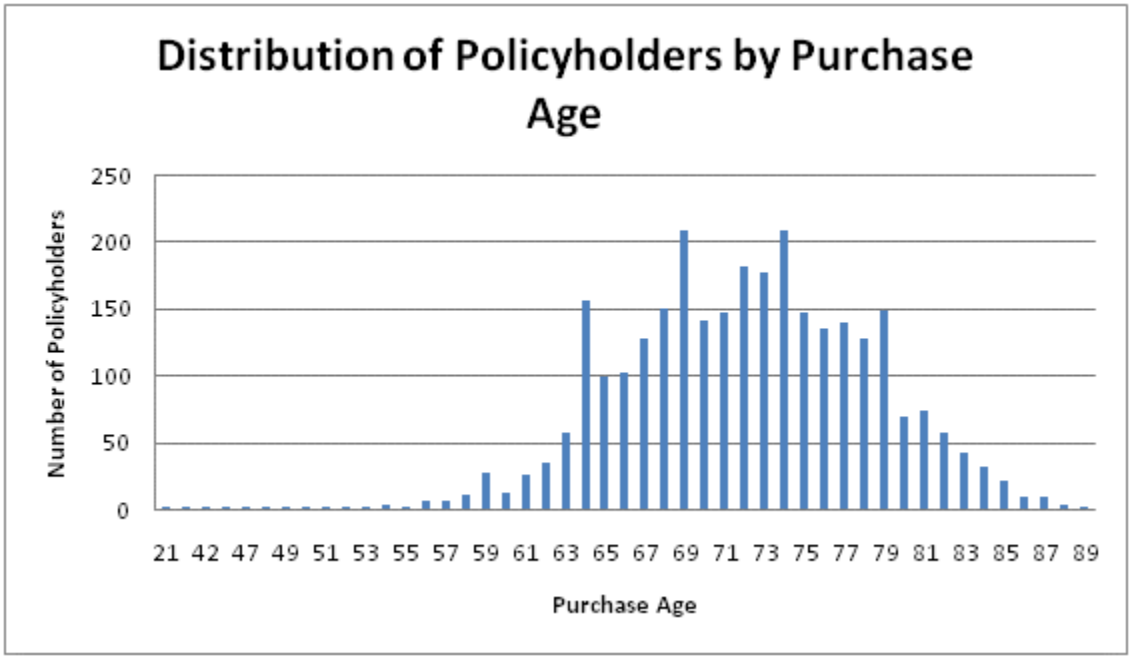


Figure 3 Distribution of Policies by Purchase Age

Graphs and charts helped the team to understand the distribution of the data provided and revealed any trends that needed to be further investigated. Once the data set organization was complete, the team worked to collect supplemental information to be used in the analysis

3.2. Supplemental Data Sources

Supplemental data are any information that can be appended to existing policyholder files on an individual basis to create a more complete profile of each long-term care insurance user. For the purposes of this project, AbilityRe was able to provide all the non-personally identifiable fields kept on record for each on-claim insured to the project group. Since the goal of this project was to identify variables in long-term care insurance policyholder data that may indicate or aid in the prediction of future spending and usage, expanding the variables that could be evaluated was necessary.

Several options are available in identifying sources of supplemental data. The project group initially identified two viable options: the government's census data and marking data aggregators. The United States Census website is able to provide a wealth of information such as average family size, household income, and number of bedrooms in a household.[32] However, these facts are given on an aggregate basis using zip code. In the group's block of long-term care insurance policyholders, the majority of the policies were purchased in a couple of states. Thus, using data on a zip code instead of an individualized level would provide little distinction between the insureds, making it difficult to identify realistic trends in spending and policy usage. Conversely, data aggregators collect information on an individual basis primarily for marketing purposes. Most companies that specialize in the collection of such data establish a cost structure

[32] *U.S. Census Bureau, Housing and Household Economic Statistics Division*. (2009). *U.S. Census Bureau.*

that has a fixed price for all the information available on one person. Applying this price to the number of records in AbilityRe's on-claim policyholder file would result in a substantial cost.

In the winter 2009 edition of *Contingencies* magazine, an article entitled "Bringing Predictive Models to Life" discussed gathering supplemental data for use in insurance predictive modeling. The authors noted that using supplemental data in these types of models was critical because it greatly increased the segmentation power.[33] To gain insight into how this was done, the group contacted two of the article's authors, Mike Batty and Chris Stehno of Deloitte.

The Deloitte consultants were able to explain many of the intricacies of using marketing data that had not been considered by the group. First, since this data is collected for the purposes of marketing and sales, historical data on an individual basis is neither stored nor available by the aggregators. In most cases, data is refreshed every six months. For this project, the group required individual variables at the time the policyholder went on-claim. This meant that the group had to limit the AbilityRe data set to only those individuals who had gone on-claim within the last two years. Beyond this point, it can be assumed that the marking data will no longer create an accurate description of the policyholder the day they went on-claim. Second, they discussed the value added by creating synthetic variables, which are those that are not directly provided but could be calculated using the information given. Finally, Batty and Stehno suggested establishing a univariate review with AbilityRe to assure that all of the variables collected and used in this project would meet the company's legal and compliance rules.

Seeing opportunities for mutual learning and gain, Batty and Stehno worked with the group to obtain a corporate discount from the marketing data aggregator they had worked with in the

---

[33] Mike Batty et al. (2009)."Bringing Predictive Models to Life." Print.

past, Equifax. Equifax had the potential to be able to provide a wealth of marketing data on an individual basis using the key identifier of name and home address, which typically yields a 95% accuracy rate. The list of on-claim policyholders and their information needed to be sent to Equifax through AbilityRe so that the group would not be exposed to any personally identifiable data in the process.

Due to time constraints on the project and the complications that can arise from working with third parties, the group established a contingency plan. Without obtaining external supplemental data, the group was still able to test the accuracy of the process outlined. First, the team divided into two groups: one that created pseudo supplemental data and one that tested the effectiveness of the procedure designed. The pseudo data was generated in such a way that only some of the variables created resulted in viable clusters, which indicates that the variable would be useful in predicting policyholder benefit usage.

3.3. Policyholder Scoring

Calculating a score for each policyholder was a challenging but crucial part of the analysis. This allowed each policyholder to be rated on a scale from one to one humdred based on how well they used the benefits that they were able to receive from their policy. Throughout the project period, several different methods were suggested for determining a ranking for the individuals. Collaboration between team members and AbilityRe representatives was crucial to the creation of the final scoring method used.

The first part that was necessary to determine was which variables were needed for the calculation of an individual's score. After several discussions with AbilityRe representatives, the team decided that the calculation required the incorporation of two variables into the score for each policyholder. These variables were the amount of time a policyholder owned the policy

before going on claim for the first time and a calculated financial ratio. The financial ratio was used to relate the dollar amount of premiums paid by a policyholder to the amount of claims that were paid out to them from AbilityRe.

In order to calculate the financial ratio, the team began with a simple ratio of the amount of claims paid out to a policyholder over the amount of premiums that were paid to AbilityRe by the policyholder. The amount of premiums paid by the policyholder was to be computed using a series of calculations; essentially, the amount of time the policyholder owned the policy was to be multiplied by the amount that the policyholder paid on a regular basis. It was sufficient to assume that the policyholder would not pay any more premiums to AbilityRe because once going on claim, a policyholder stops paying premiums.

The calculations for determining the amount of claims paid out to the policyholder were a bit more rigorous than those for the premiums. The team felt that it was necessary to look at both the claims previously paid out and the amount of claims that would likely need to be paid out in the future. In order to calculate the amount of claims paid out to the policyholder, the length of time spent on claim was to be multiplied by the policyholder's benefits. On the other hand, calculating the projected future claims usage was not as straightforward as the other parts of the ratio. This involved determining the average time spent on claim by an individual and subtracting that from the time that each policyholder had already spent on claim. This result was then to be multiplied by the policyholder's benefits. The projected benefits were to be added to the claims previously paid to the policyholder. Finally, the total claims to be received by the policyholder were to be divided by the amount of premiums paid out. The team's initial plan for calculating this ratio can be seen in Appendix B: Proposed Scoring Method.

After proposing this method to AbilityRe, the team was informed that many of the values that were to be calculated in the initial method were already on record with AbilityRe. Upon learning this, the team asked to be provided with these values. Not only did this simplify the calculations for the financial ratios, it also resulted in values that are more accurate.

Upon receiving all of the data for the financial ratio, the team noticed that there were values that had not previously been considered. For example, in addition to the premiums paid by the policyholder, AbilityRe also has a refund value for some policyholders. Policyholders who received refunds have the Return on Premium rider added to their policy, meaning that they receive a percentage of their premiums back if they do not go on claim within a certain number of years after purchasing it. In order to account for this, the refund value was subtracted from the premiums paid for those policyholders whose refund value was not zero. Additionally, AbilityRe also had on record an Active Life Reserve (ALR) and Disabled Life Reserve (DLR) for most policyholders. AbilityRe team members explained that the DLR is the amount that a policyholder who is currently on claim is expected to use for the current claim. On the other hand, the ALR is calculated for each policyholder and is the amount that AbilityRe expects to pay out in claims to that policyholder in the future. Together, these two numbers made a projected reserve for each policyholder.

In order to analyze the impact that each variable played in the calculation of the ratio, the team decided to calculate the financial ratio in several different ways. The first method added the ALR and DLR for each policyholder with the claims previously paid out. Essentially, this was the same as the initial idea to use a projected reserve for future claims usage. Because there were some policyholders for whom there was no record for ALR and DLR, the team decided to calculate the projected reserve as was intended by the original plan in the second calculation of

the financial ratio. The final method for calculating the financial ratio did not include any projected reserve. All equations used for calculating the different financial ratios may be found in Appendix C: Formulas for Calculating Financial Ratios. Ultimately, it was decided that the financial ratio calculated with the reserves provided by AbilityRe would be used.

The computation of the amount of time a policyholder owned their policy before going on claim for the first time required a much simpler calculation than that of the financial ratio. This was calculated by determining the number of decimal years between the time that an individual purchased a policy and the first reported claim date. This can be done in Excel through the use of the DATEDIF function.

In order to obtain a score that took both of these values into account, the team had to ask help from AbilityRe team members. The team determined that the best way to do this was to select a minimal number of policyholders that were different from one another so that it would be possible to distinguish which policyholders should be given which scores. To begin this process, all policyholders were plotted on a grid which was then divided into nine buckets based off of the two values. The length of time the policyholder owned the policy before going on claim was broken down into three sections: 0 to less than 5 years, 5 to less than 15 years, and 15 or more years. The financial ratio was also broken down in to three sections: values from 0 to less than 0.25, values from 0.25 to less than 2, and values higher than 2. The combination of these two variables placed each policyholder into one bucket. The group then chose one person from the center of each bucket. The team then sent only the corresponding ID numbers to AbilityRe and asked them to assign scores to the nine individuals.

Once the nine scores were received, a method needed to be determined that would allow for the assignment of scores to the other 2917 policyholders. The group decided that the best way

to do this would be to combine the nine points chosen with their corresponding scores. This provided the group with nine points in a three dimensional graph. Once this was done, the group would employ a method to fit a surface to the nine points. By obtaining a surface, it was possible to determine the height of the surface for any corresponding point on the grid; thus, a score could be calculated for any combination of financial ratio and number of years before going on claim. Several methods for fitting a surface were proposed to the group by different faculty members in the WPI Mathematics Department. Each method was considered and discussed by the group to determine which were most feasible and effective for the scores given.

The first method that the group utilized was the use of biquadratic functions. This allowed for a surface, which touched each of the nine points, to be created. The biquadratic function used by the group can be seen below.

$$z = a + bx + cy + dx^2 + exy + fy^2 + gx^2y + hxy^2 + ix^2y^2.$$

In the given equation, z represents the score corresponding with some financial ratio, x, and number of years before going on claim, y. By solving for the coefficients in this equation, the group obtained the equation for a surface that fit through the nine points. To solve for these coefficients, the group first substituted the nine scores and corresponding financial ratios and number of years before going on claim into the above equation, in order to attain nine different biquadratic functions. Once the group had nine equations with nine unknowns, it was possible to solve for each of the unknowns using matrices. The group placed all of the values into three matrices as shown here:

$$\begin{bmatrix} 1 & x_1y_1 & \cdots & x_1^2y_1^2 \\ 1 & \ddots & \ddots & x_2^2y_2^2 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & x_9y_9 & \cdots & x_9^2y_9^2 \end{bmatrix} \begin{bmatrix} A \\ B \\ \vdots \\ I \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_9 \end{bmatrix}$$

After these matrices were set up, the group calculated the inverse of the matrix of x and y values and multiplied it by the matrix of z values, resulting in a matrix with the values for the nine coefficients. After substituting these coefficients into the biquadratic function, the group was provided with an equation that allowed for the calculation of a score for every individual in the set of policyholders. The predominate flaw of this approach is the effect outliers have on the surface. In the case of an outlier being present in the data set, the overall shape is greatly augmented to accommodate this point, which leads to distortion of the surface.

Another method that the group employed to determine a score for each individual was calculating a least squares plane fit to the nine points. Unlike the previously discussed method, this surface would not touch each of the nine points that the group had. Instead, it would fit a surface that minimized the sum of squared errors between the surface and the each of the given points. To begin this method, the group used the equation, $= A + Bx + Cy$, for the surface. In order to minimize the sum of the squared differences between the surface and the given scores, Π, the group took the partial derivatives with respect to each coefficient and set them equal to zero.

$$\pi = \sum [z_i - (A + Bx_i + Cy_i)]^2$$

After solving for each of the partial derivatives, the group attained three equations with three unknowns. Once again, the group employed the use of matrices to solve for the three unknowns. The three matrices utilized by the group can be seen below.

$$\begin{bmatrix} n & \sum x_i & \sum y_i \\ \sum x_i & \sum x^2 & \sum xy \\ \sum y_i & \sum xy & \sum y^2 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} \sum z \\ \sum xz \\ \sum yz \end{bmatrix}$$

By solving for each of the unknowns A, B, and C, the group obtained the equation for the least squares surface fit to the nine points given, making it possible to solve for the scores of all of the policyholders in the data set.

3.4. K- Means Clustering

K-means clustering is a commonly-used partitional clustering method. It is one of the simplest and most efficient ways to analyze and categorize data. It was developed by J.MacQueen in 1967 and then refined by J.A.Hartigan and M.A.Wong around 1975. The concept of k-means clustering is simple and intuitive. The clustering procedure is done by minimizing the sum of squared distances between observations and their corresponding cluster centroid. The following formula depicts the mathematical representation of the distance between each point and its centroid:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and its centroid c_j.[34]

The first step of the clustering mechanism is defining "k" centroids, one for each cluster. The points for centroids can be chosen randomly; however, different locations have an impact on the effectiveness of the algorithm. For this reason, it is better to place the centroids as far from each other as possible. The second step is calculating the distances between each centroid and every observation. Once the distances are calculated, each observation is grouped with its nearest centroid. At this point, the "k" clusters are initially formed. Knowing the contents of each cluster,

---

[34] "K-Means Clustering". A Tutorial on Clustering Algorithms. Retrieved on April 11, 2010 from http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

new centroids are computed based on the observations in each cluster. After new centroids are decided, the observations are regrouped using the same method. This loop is repeated until a stage is reached where "new" centroids are the same as "old" ones. Once this happens, the final clustering result is obtained. The following is a diagram of the k-means clustering process.

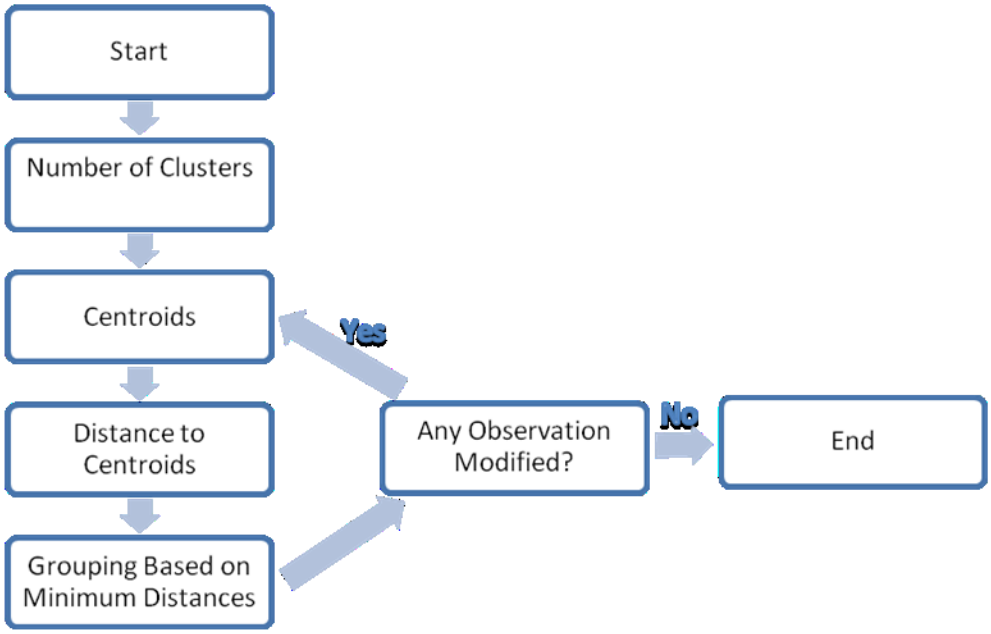


Figure 4 K-Means Clustering Technique

Several statistical software packages can perform a clustering method that would be useful for this project. For example, Microsoft Excel and Visual Basic have the necessary functionality. However, due to its computing capacity, the group determined that SAS 9.1 would be the best software to utilize. SAS is a highly effective program for data mining and manipulation; additionally, it also includes many specialized functions.

SAS, which stands for Statistical Analysis Software, is an integrated system of software products introduced by the SAS Institute Inc., which provides programmers the capability to perform data entry, retrieval, management, mining, warehousing and much more.

Since the team had very little prior training in SAS software, understanding the program was initially challenging. The team attempted several approaches to acquiring the necessary understanding of SAS due to the complexity of the software. Although online searching did provide some learning materials and related coding for clustering, the majority of the findings did not meet the project's needs. This was because the scripts were based on ideal scenarios or included macro programming, which increased the complexity of the problem.

Fortunately, the Mathematical Sciences Department at WPI offers several courses in statistics, which employ SAS in a laboratory setting. Criselda Santos Toto, a teaching assistant within the department, often oversees the SAS lab. She was willing to offer instruction for this project. During the two SAS learning sessions attended by both the group and Toto, basic SAS commands were reviewed and necessary codes were written and run. Toto also showed the group the SAS online documentation, which provides a list of all the commands in SAS. It was in this online documentation that the FASTCLUS procedure was discovered. This procedure performs a disjoint cluster analysis based on the Euclidean distances between quantitative data with at least 100 observations. The FASTCLUS procedure can perform k-means clustering using the least-square criterion or perform more precise clustering by using the least p^{th} power clustering criterion. [35]

After the two SAS training sessions, the group had established and written customized coding for the project under the instruction of Toto. The complete customized coding can be seen in Appendix D: SAS Customized Code. To test the new coding, a small sample data set was extracted from the original AbilityRe data set and clustered. The outcome was optimistic because

---

[35] SAS Institute Inc. (2003). *The FASTCLUS Procedure.*

the data was successfully clustered into groups. However, some details required modification in applying the code to the full data set due to the large number of observations in the file.

3.5. Evaluating Scoring Method

Once the data points were grouped into separate clusters through the k-means clustering algorithm, it was important to analyze them so that their meaning could be understood. The score given to each policyholder made it possible to study the types of individuals within each cluster. By taking the average score of all of the policyholders within a cluster, it was possible to determine if the variables used in clustering were good defining characteristics for policyholder behavior. If the variables chosen were effective in defining policyholder behavior, one would expect the average scores of the clusters to vary and range from high to low. Additionally, the score given to policyholders made it possible to compare individuals in different clusters.

In order to determine which variables best define future policyholder behavior, it was necessary to compare the ranges of average scores between clustering sets. Individual graphs of the clustering set scores were utilized to make these comparisons. This made it possible to break the process down into three steps. First, a control was established in order to evaluate each clustering set on the same level. The control was then plotted against each clustering set. Next, a numerical difference between the clustering sets and the control was calculated. Finally, the differences calculated between controls and clustering sets were compared to one another.

The first step in plotting the control versus the clustering sets was to calculate the average score over all policyholders. If policyholders were randomly clustered, one would expect that the average score of each cluster would be equal to the average score over all policyholders. Therefore, a straight line on a graph, when plotting cumulative score of clusters versus individual clusters, would depict the cumulative average score over all clusters. This was used as the

control against which all clustering comparisons were made. To plot the clustering sets in individual graphs, the clusters within each set were first sorted in ascending order by average score. The cumulative score was then plotted against the clusters. If the variables chosen have an impact on the use of benefits, one would expect a somewhat exponential curve. An example of this process can be seen in Figure 5. In this example, there were nine clusters and the average score over all policyholders was 60. The straight line depicts the control used whereas the curved red line represents a clustering set that used some variables A, B, and C. The area in between the curves shaded in green is the calculated difference that was used to compare clustering sets to one another.

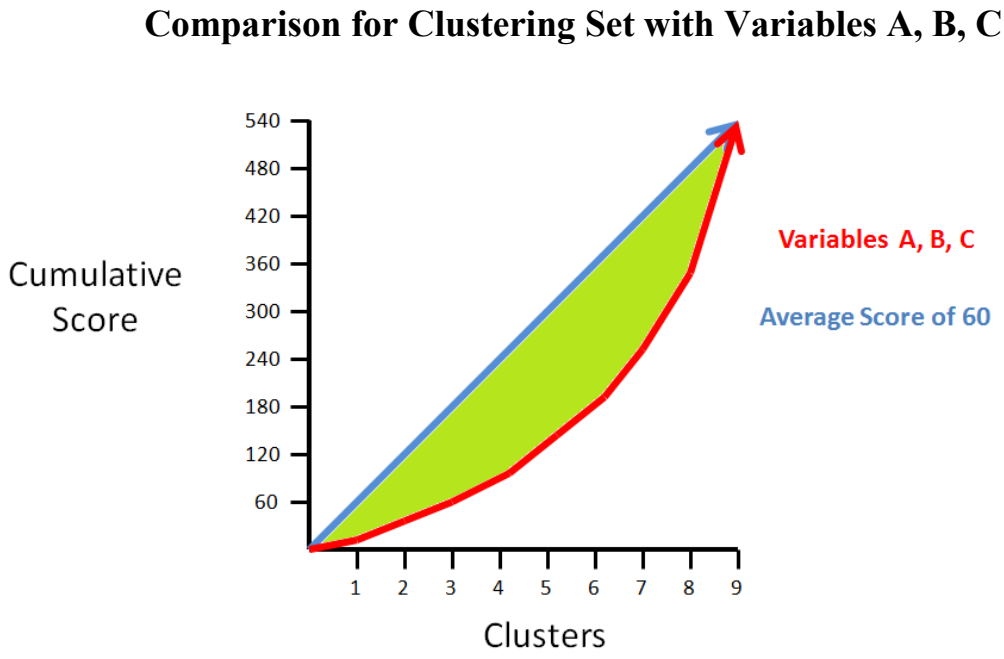


Figure 5 Clustering Set Versus Control

To calculate the area between the curves, it was necessary to break the graph into two separate pieces. The first area that was calculated was that under the straight line. Because it is a straight line, the area can be calculated by using the simple formula for the area of a triangle seen below:

$$Area = \frac{1}{2} * (Base * Height)$$

Next, it was necessary to calculate the area under the curve made by the clustering set. To perform this calculation, the area was broken down into a series of trapezoids. This made it possible to use the simple equation for the area of a trapezoid found below:

$$Area = \frac{1}{2} * Height * (Base\ 1 + Base\ 2)$$

Once the area of each trapezoid was calculated, their areas were summed together to determine the total area under the curve. This area was then subtracted from the area under the straight line, resulting in the area between the two curves. An image of the breakdown of the calculated areas can be seen in Figure 6.

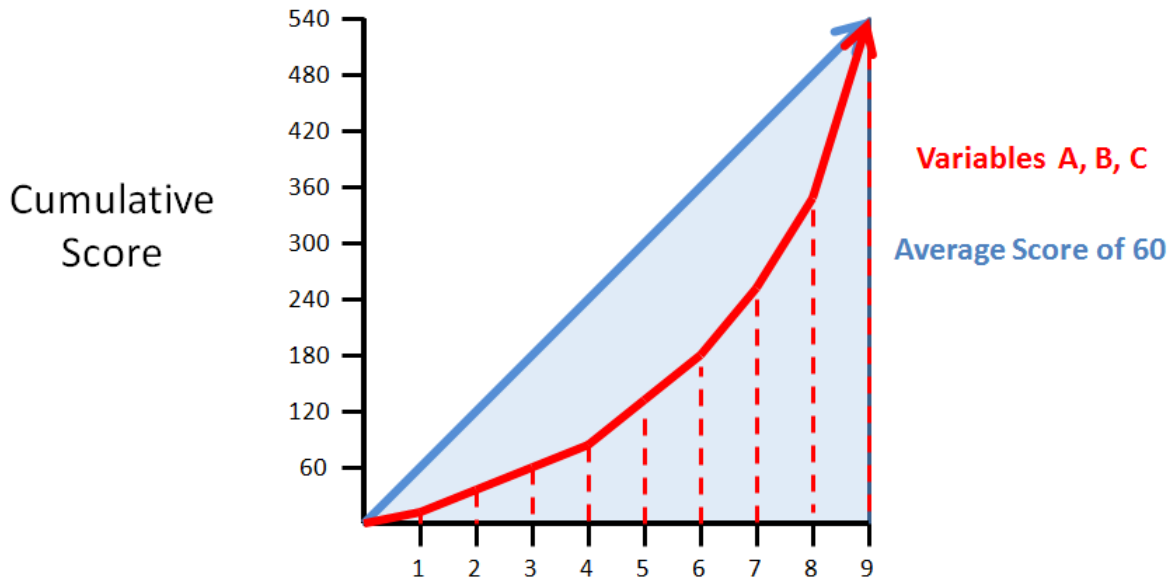


Figure 6 Calculating Area Between Curves

The last step in the process for evaluating the scoring for the clustering sets was to compare the calculated values for the areas between curves to one another. Because the steepness of the clustering set curve is dependent on the range of average scores in the set, a broader range

would result in a shallower curve. This means that a larger area between curves suggests a better group of predictor variables. Figure 7 shows a side-by-side comparison of two sample clustering sets. The graph on the left shows a clustering set created with some variables A, B, and C. On the right, the graph depicts a clustering set using variables D, E, and F. In this case, the clustering set with variables A, B, and C has a larger area between curves, indicating that variables A, B, and C are a better group of predictor variables than D, E, and F.

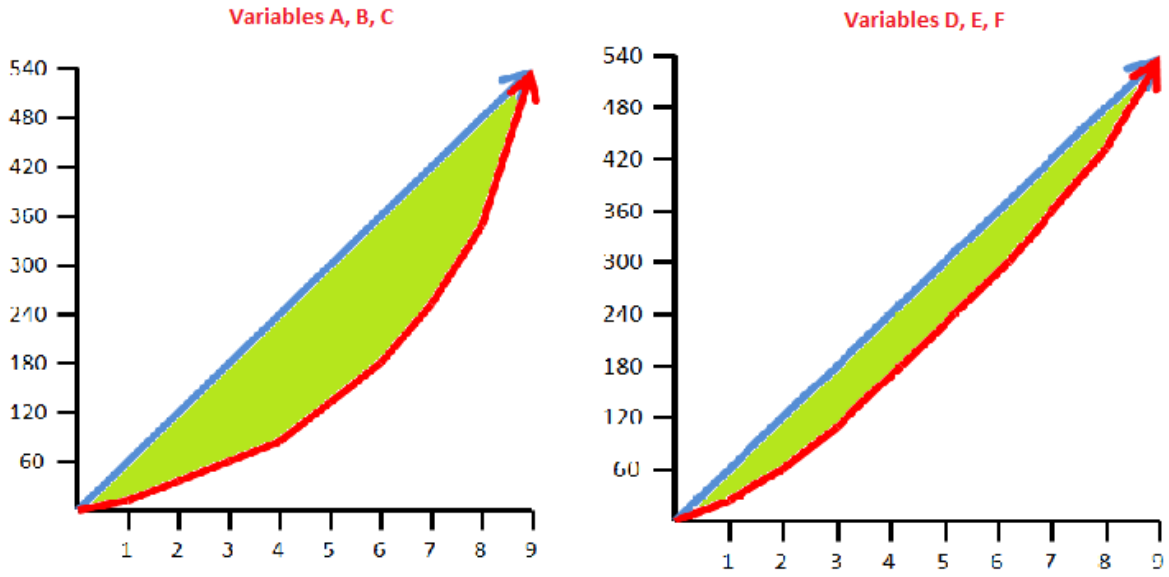


Figure 7 Comparison of Two Sample Clustering Sets

The calculated areas between curves can also be compared to the area that would result from an ideal clustering variable. This provided a more meaningful numeric representation because it allowed the group to designate a percentage relative to the ideal area that an individual variable achieved. A variable that clusters perfectly would put each individual in their own cluster, resulting in maximum differentiation between policyholders. This area was calculated and despite it being ideal, the group realized that it may not have been a realistic application. A more realistic, ideal area was obtained by lining up policyholders in ascending score order and

dividing them into equal clusters. This was a more practical area calculation because the clustering technique utilized by the group never resulted in policyholders being placed in individual groups, but always in ten clusters.

Chapter 4: Results and Discussion

After outlining a procedure to identify policyholder variables that would indicate and aid in the prediction of policyholder benefit usage, the assumptions made can now be validated or refuted. This chapter presents the results of calibrating the scoring method, identifying a range to gauge variable significance, and implementing the procedure on several simulated data sets.

4.1 Calibrating Procedures

As a way to identify potentially predictive variables, the group established the Variable Identification Procedure, which can be seen in Figure 8.

Figure 8 Variable Identification Procedure

The first step is to cluster policyholders based on one characteristic. This was completed through the use of the FASTCLUS program in SAS. This function utilizes the k-means clustering algorithm to group policyholders with similar characteristics into clusters. By doing this, the group was able to determine which variables may be useful in differentiating policyholders. Independently of step one, each individual in the data set was given a score from zero to one hundred. Scores were determined in several ways using the techniques outlined in the methodology chapter of this report. Finally, the group evaluated the scoring by testing the level of differentiation in the data the variable caused and comparing that value to the separation that could be the result of randomized clustering. A valuable predictive variable would separate the

policyholders into distinct clusters and the average score would vary amongst clusters. The evaluation step was carried out using a macro in Excel, which would calculate the difference between the line of the average score and the cumulative average score of each cluster. As a result of the shape of the graph that was produced, the group called this area the "banana area".

Several components of the Variable Identification Procedure had to be calibrated before being applied to the policyholder data from AbilityRe. In particular, the scoring method needed to be aligned with the view AbilityRe has on policyholder usage, and the impact of the value of the banana area required adjustment to account for the effect of randomness. The group adjusted both of these areas and the procedure that was followed in each case is described in the remainder of this section.

4.1.1 Policyholder Scoring

As was discussed in the methodology chapter of this report, the scoring method was calibrated by collaborating with AbilityRe team members. First, the group asked AbilityRe employees to assign a score to a subset of nine policyholders. The team selected one policyholder from each bucket based on the two quantitative variables: financial ratio and the amount of time a policyholder owned the policy before going on claim. AbilityRe team members were unaware of how the nine policyholders were selected and the two values that the team calculated to determine the bucket that each policyholder belonged to. This allowed the policyholders' scores to be based off the intuition of AbilityRe team members, rather than the variables that the team thought to be useful. The scores were given to the policyholders in the sample set by reviewing information the reinsurer had on file about the individual. Thus, all scores are from the perspective of the insurance company. This means that in general a score of zero would be a policyholder that has cost the insurance company a lot of money to insure, but

has paid premiums for a short amount of time. Conversely, a score near one hundred would be a policyholder that paid premiums for a significant amount of time, but used minimal benefits.

Once the scores were obtained, they were linked to the policyholder's financial ratio and the amount of time that the policy was owned before going on claim. These three variables combined to create a three dimensional representation, which allowed for the creation of a surface. Several mathematical approaches were implemented in an attempt to construct a representative surface. After constructing the surface, the scoring method could be applied to the remaining 2917 policyholders.

Upon reviewing the scores that were being extrapolated from the surface, the group realized that having more points would allow for a more accurate surface to be created. As a result, the group asked AbilityRe team members to score additional policyholders. In this request, the group included three policyholders in each of the buckets based on financial ratio and the amount of time a policyholder owned the policy before going on claim, as well as twenty policyholders that have the most extreme combination of those variables. Obtaining multiple points from each bucket helped to ensure that the scoring method would create an accurate surface for that area. Additionally, the extreme points represent the edges of the surface, so having these policyholders scores guaranteed that the points were captured and portrayed appropriately. The complete list of scores generated by AbilityRe team members for the subset of policyholders can be seen in Appendix E: Policyholder Scores from AbilityRe.

4.1.2 Reasonableness Range

After establishing a methodology for calculating a number that represented the impact that each variable had on the clustering of policyholders, a technique that would determine the level of significance at which that number lay was needed. A variable that had no impact on the

clustering of policyholders would randomly group them into any cluster. With this thought process in mind, a range of banana areas was calculated using randomly generated clusters. This made it possible to determine a range of values that could be considered insignificant if obtained by the clustering performed on policyholders.

In order to establish the range of banana areas considered insignificant, the 2,926 policyholders were randomly placed into one of ten clusters. A macro was written in Excel that allowed this process to be repeated for 10,000 cluster sets. Once the 10,000 cluster sets were obtained, they were placed into a macro that calculated the banana area for each cluster set. This macro was a version of that used for the policyholder clustering, but it contained modifications that made it possible to handle the larger number of cluster sets. This macro provided the group with a list of banana areas obtained from running the 10,000 cluster sets. With this data, it was decided that the middle 9,500 values would be used as the range of insignificant values. Removing the 250 values on either side of the range made it possible to eliminate any outliers in the set of values. Once these values were removed, the range of values that was obtained was 1,078,048 – 2,902,578. After finding this range, the group determined that any banana area values in excess of 2,902,578 could indicate a possible predictor variable that would be useful in determining future claims because they were an improvement to what could be reasonably expected by randomized clustering.

4.2 Evaluating Technique

It was the original intent of the project group to apply the Variable Identification Procedure to a supplemental data set, which would include variables, such as marketing data and policyholder information, on each individual policyholder in the AbilityRe data set. The intended outcome was to identify specific variables that indicated or could aid in the prediction of future

benefit usage. However, due to concerns over privacy regulations outlined in HIPPA, AbilityRe was unable to provide the information Equifax required to generate the supplemental data. Consequently, the focus of this project was redirected from identifying variables in a data set to developing and testing a specific procedure, which could be used to identify variables in the future if a data set were available. To test the functionality and accuracy of the Variable Identification Procedure, the group divided into two teams. The first team generated several simulated data sets that modeled the kinds of information, which might be included in supplemental data. The other team clustered, scored, and evaluated the variables that were included. This method created a blind process ensuring that the results were genuine and not crafted by known expectations. The results of the trials are detailed in the remainder of this section.

4.2.1 Simulated Data Trial One

The goal of the simulated data sets was to test the Variable Identification Procedure on information that resembled supplemental data to see if this methodology would yield useful results. Each data set was designed differently to test the capacity of the process at differing levels of randomness or predictive capability. This allowed the group to see if the procedure could distinguish useful variables amongst data in which no trends were present. The spectrum of possible trials can be seen in Figure 9. The first simulated trial set is on level two of the spectrum.

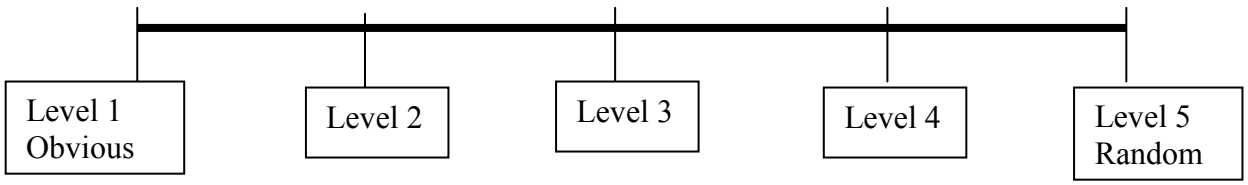


Figure 9 Trial Spectrum

4.2.1.1 Creating Data Set

The first trial data set was a level two in the spectrum of possible data sets, which indicates that although the clustering was not obvious without the use of SAS and Excel it was fairly easy to determine relationships with these programs. This data set was established with the intent of creating three variables that would cluster well, indicating that they were predictor variables, and six additional variables that would be generated randomly, indicating random variables. Additionally, it was a further objective of the first team, that the three predictor variables, when clustered in tandem, would produce an even larger banana area than any one of these variables could create alone. To create the data set the first team began by constructing a three dimensional space. From this space, ten points were chosen that maximized the distance between all points. Each point has coordinates made up of only ternary values. The three dimensional space can be seen in Figure 10.

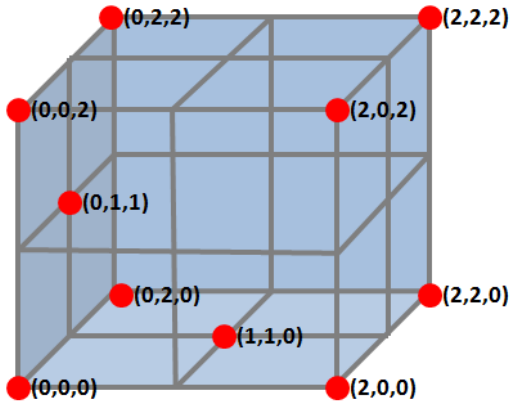


Figure 10 Three Dimensional Space of Centroids

These ten points served as the centroids for the ten clusters that could be created if the three predictor variables were clustered together. The centroid coordinates were transformed into a value between zero and one hundred. Next, additional points were generated around each

centroid by applying a normal distribution model. The standard deviation was selected so that the points would most likely not fall into the range of a neighboring cluster.

At this point ten clusters had been created and there were three variables that could be used for clustering. Six additional variables were then randomly generated using values from zero to one hundred. This resulted in a total of nine variables in the data set. All values in the data set had randomly generated values; however, the three predictor variables were randomly generated for a much smaller range of numbers. A table was created which had the random values that were associated with each cluster.

The next step in the process was to randomly assign policyholders into ten clusters. To do this, the policyholders were sorted according to their score and then broken into ten approximately equal clusters. This resulted in ten clusters that had different average scores Once policyholders had been assigned to a cluster, a VLOOKUP was performed in Excel that assigned each individual the nine random numbers from the previously created table.

Simultaneously, in a separate table, potential categorical and numeric variables were listed as possible characteristics of policyholders that would be included in supplemental data. The variables that were chosen include: occupation, house value, income, number of magazine subscriptions, number of children, university, driving record, and proximity to medical facility. Once these categories were picked, values in each category needed to be assigned for each number from zero to one hundred, so that they could be mapped to the random values generated for each policyholder. The categorical variables contained ten possibilities each, so that ten distinct clusters could be identified. However, for the numeric variables, linear equation were used to transform each number from zero to one hundred to a value that would be meaningful for

that variable. For example, the linear transformation for house value was 50000 + (2000 * value), where value represents a number from zero to one hundred.

After the variables were created, they could be mapped to the table of policyholders. This resulted in the creation of a data set with both categorical and numeric variables, of which three were predictor variables and six were the result of random number generation.

The second group was not aware that only three variables would generate banana areas that would be considered significant, nor were they previously informed that a linear relationship was used and would be helpful in decoding the relationship. To further blind the study, the order of the variables was rearranged so that the first three variables tested would not all necessarily be predictor variables. Thus, the second team would have to determine a method for decoding the variable's relationship, test each variable individually, and test the variables in combination to uncover which variables in this trial would aid in the prediction of future claim usage. In this trial the three predictor variables that the first team chose to use were: occupation, income, and proximity to nearest medical facility. Each of these variables was intended to appear significant on its own, but would result in even greater predictive capacity when combined.

4.2.1.2 Clustering Data Set

After receiving the first trial of simulated data generated by the first team, the second team started to apply the evaluating mechanism. This uses SAS as the fundamental tool to cluster the data, then applies an Excel macro to compare the banana areas for the scores associated with each variable, and finally identifies the variables with excessively large banana areas. Variables that meet this criteria may be characteristics that in the real world could influence the behaviors of policyholders.. This procedure, if successful, could be a very powerful tool to discover hidden trends in policyholder data.

As one would expect, some of the variables to be tested were quantitative, while others were categorical. In addition to the real variables, age at purchase and gender, contained in the AbilityRe data set, nine other variables were provided by the first team.

The first step was to cluster variables individually. The FASTCLUS function in SAS was able to perform this operation. Since the function required the number of clusters, the team decided to cluster the data in ten clusters. However, some variables, for example the number of children, for which the distribution indicated less than ten clusters clearly, parameters in the FASTCLUS function were modified to better fit the data.

Quantitative variables were imported into SAS and the clustering results were exported as an Excel spreadsheet. Next, the clustering results were inserted into the macro to calculate the banana area, and the average scores of each cluster were put in ascending order to compute the area difference for this specific quantitative variable.

For the non-numeric variables, the team decided to cluster the data based on categories. For instance, the car colors of the policyholders that have cars were black, blue, green, purple, red, silver, white and yellow, while no car was another category. In this case, clusters for car colors already existed. To compare this variable to others, however, an area difference was still required. Average scores for each car color were calculated and inserted into the macro in ascending order, and the banana areas were produced.

For the simulated data trial one, the banana areas for each variables were listed in ascending order as can be seen in Table 2 below. The ideal banana area that could be achieved would group policyholders in the clusters that they were assigned to by the first tea. Thus, all variables can be described as having achieved a percentage of this ideal area.

Variable	Area Difference	Percentage
Score	33,033,873.30	100.00%
Income	16,354,280.50	49.51%
Proximity to Nearest Medical Facilities	14,808,338.40	44.83%
Occupation	6,671,199.50	20.20%
Age at Purchase	3,563,869.70	10.79%
Number of Magazines Subscribed	2,386,487.80	7.22%
Driving Record	1,982,522.90	6.00%
Number of Children	1,858,677.90	5.63%
Gender	1,811,997.00	5.49%
University	1,751,585.20	5.30%
House Value	1,527,073.40	4.62%

Table 2 Simulation Trial One Banana Areas for Individual Variables

As previously stated, the banana area for policyholder score was the perfect or largest banana area that could be obtained. Additionally, as a result of random trials, it was found that any banana areas in excess of 2,902,578 could indicate a possible predictor variable that would be useful in determining future claims. Therefore, for simulated data trial one, it was quite obvious that the income, proximity to nearest medical facility, occupation and age at purchase were the variables that could be used to predict policyholders' behaviors. Because the effect of a a variable may be magnified when combined with another, the variables with excessively large banana areas were combined in different ways and re-clustered. In this first trial, as the results for individual variables indicated, variable income, proximity to nearest medical facility, and

occupation were selected to be re-clustered. Although clustering age at purchase resulted in a higher than random area, it was omitted from the re-clustering because the team decided that it would only re-cluster with three variables. This allowed for four re-clustering combinations, and would serve as the standard number of variables to be re-clustered in future data sets.

The approach to cluster multiple variables was quite similar to that performed for single variables. However, it was necessary to standardize all variables to the same scale in order to avoid the possibility that one variable would overwhelm the others. For example, the difference between income and proximity to a medical facility is very large so in comparing these two, it is necessary to find a scale that correctly represents a difference in values. To standardize the variables, the second team adopted several approaches to put all three variables into the zero to one hundred scale.

For quantitative variables, the group tried two approaches. The first approach was that the minimum of the data set was subtracted from the value of each variable and then divided by the difference of the maximum and minimum. The formula for this approach is shown below:

$$Standardized\ Data = \frac{Original\ Data - Minimum}{Maximum - Minimum}$$

This approach would stretch the lowest point of the data set to zero and the highest point to one hundred, with all other points falling into the zero to one hundred range. However, the disadvantage of this approach was the differences between each data point were also stretched out. In some cases, this could weaken or even eliminate the data pattern, which may influence the results and conclusions obtained from the analysis.

The second approach for quantitative variables was first to calculate a multiplier that would bring the maximum number in the data set to one hundred, and then multiply each of the data points by this multiplier. The formula for this approach can be seen below:

$$Standardized\ Data = Original\ Data * \frac{100}{Maximum}$$

Although this approach does not bring the lowest point of the data set to zero, the distances between each data point were better preserved. Therefore, the standardized values could better indicate and describe the relation among the data.

For categorical variables, the group originally attempted to assign numbers to each category based on the characteristics of the category. For example, in trial one, availability and capability to utilize policies were considered to be the factors that most affected the value assigned to each occupation. Although this makes sense logically, due to the time constraints of this project, the team discovered that it was difficult to determine the order of occupations.

Two other methods were adopted as substitutions for the first approach. In the first method, the zero to one hundred range was divided into ten buckets. Each bucket was assigned a range of ten values, and then each range was randomly assigned an occupation. Therefore, policyholders who worked in the area of education might receive a 15 while those who worked in a factory might receive a 75.

The second approach was that the occupations were first placed in ascending order by their average score, which was obtained during the single variable clustering process. Next, they were assigned a value that was calculated using the same function in the second approach for quantitative variables which is shown below:

$$Standardized\ Data = Original\ Data * \frac{100}{Maximum}$$

The values assigned to each occupation using this method can be seen in the chart below:

Government	83.56894717
Engineering	85.46648641
Unemployed	90.81815945
Health Care	91.69441374
Retail	91.95797854
Construction	92.22069734
Factory	92.28517496
Education	97.16702504
Sales	99.99854373
Manufacturing	100

Table 3 Simulation Trial One Occupation Values

After the three variables were all normalized to the zero to one hundred scale, the variables were then combined in four ways to be re-clustered. The three variables, income, proximity and occupation were clustered together first, then income and proximity were combined, then proximity and occupation, and finally income and occupation were clustered. The banana areas for each of the four combinations are shown in Table 4.

	Approach Q1&C1		Approach Q2&C1		Approach Q2&C2	
	Area Difference	rank	Area Difference	rank	Area Difference	rank
I+P+O	24,529,332.0	2nd	29,964,566.9	MAX	24,864,935.0	MAX
I+P	24,669,767.9	MAX	24,672,584.8	2nd	24,596,835.8	2nd
I+O	16,423,934.3	3rd	17,405,748.3	4th	17,583,121.6	3rd
P+O	16,262,527.6	4th	20,580,889.5	3rd	17,571,705.9	4th
Comparison	I+P+O<I+P		I+P+O MAX		I+P+O MAX	
Conclusion	I>P>O		P>I>O		I>P>O	

Table 4 Simulation Trial One Multiple Variable Clustering Banana Areas

Note: I stands for income, P stands for Proximity to nearest medical facility, and O stands for occupation. Additionally, the notation for approaches followed indicates the formula utilized (i.e. Approach Q1 & C1 indicates that the first methods for both categorical and quantitative variables were used).

As demonstrated in the chart above, using the second approaches for both quantitative and categorical variables produced the results that one would expect because the order obtained when clustering single variables was maintained. For example, when proximity and occupation, the variables that produced smaller banana areas individually, were clustered together the resulting banana area was the smallest of the four joint banana areas produced. Additionally, when all three variables were combined, it resulted in a larger banana area than any two variables combined.

Although this clustering data set worked well with the simulated data trials, there were some restrictions and limitations in the methods used. The FASTCLUS function in SAS requires that the user input the number of clusters. In this case, the second team decided to group the data

into ten clusters, which may be fewer or more than actually needed. Additionally, the Excel macro used to calculate the banana areas also has the restrictions being applied to ten clusters. It would not be difficult to modify parameters for either the FASTCLUS function or the macro; however, it would still require that a number of clusters be input. Moreover, if these parameters were to be changed in the simulated data sets, it could result in a different outcome than was reached in this trial.

4.2.2 Simulated Data Trial Two

For the second data trial that was performed, the first team intended to create a data set that would fall at level three on the trial spectrum. This meant that the data would be more random than the first data trial that was created, but the policyholder characteristics would still be able to be identified through the use of the Variable Identification Procedure.

4.2.2.1 Creating Data Set

In the first data trial, the generated data for each cluster had a very small probability of overlapping another cluster. In order to make the data more random than the first trial, the first team decided to increase this probability to a level that would more likely result in overlapping data. This was accomplished by utilizing the same three-dimensional space and centroid points that had been determined in the first trial. To increase the likelihood that clusters would overlap one another, the generated values were further from each cluster centroid. This was accomplished by doubling the standard deviation of the normal distribution model that was used in the first trial.

Once the new values were obtained, the team was able to use the same table of supplemental data that was created in the first trial to assign values to each policyholder. Variables different from those in the first trial were chosen as those that would cluster the data

correctly. The variables chosen for this trial were: driving record, occupation, and number of children.

4.2.2.2 Clustering Data Set

The same methods used for trial one were utilized to test the simulated data provided by the first team in trial two. The banana areas found for each variable are listed in descending order below:

Variable	Area Difference	Percentage of Ideal
Driving Record	12,709,377.70	38.47%
Occupation	10,454,527.00	31.65%
Number of Children	3,691,521.10	11.17%
Age at Purchase	3,563,869.70	10.79%
Income	2,797,409.20	8.47%
Number of Magazines Subscribed	2,430,896.90	7.36%
Proximity to Nearest Medical Facilities	2,262,752.50	6.85%
Car Color	2,082,562.20	6.30%
Gender	1,811,997.00	5.49%
House Value	1,444,132.50	4.37%
University	1,441,180.60	4.36%

Table 5 Simulation Trial Two Banana Areas for Individual Variables

In this trial, it was obvious that the variables Driving Record, Occupation, Number of Children and Age at Purchase led to excess large banana areas. However, those for Driving Record and Occupation were much larger than the other two variables. Because re-clustering just

two variables would only result in one grouping, the second team decided to also include the top third value to assess the effect of multiple variables. Once again, age at purchase was not included in the re-clustering because only the top three variables were chosen.

As it was decided in trial one that the second categorical and second quantitative approaches produced the best results, they were once again used to convert all of the variables to a zero to one hundred scale. The results of the re-clustering of the three variables in the table below.

Variable	Area Difference	Percentage of Ideal	Rank
D+O+C	20,333,712.10	61.55%	MAX
D+O	17,954,436.50	54.35%	2nd
D+C	13,572,181.50	41.09%	3rd
O+C	5,749,776.50	17.41%	4th

Table 6 Simulation Trial Two Multiple Variable Clustering Banana Areas

Note: Here D stands for driving records, O stands for occupation, C stands for the number of children, and the third column denotes the ranking of clusterings.

Similarly to the first trial, the combination of all three variables produced the largest banana area of all trials. In this trial, it is worthy to note the effects which re-clustering had on the values that were obtained. For example, in re-clustering occupation with number of children, the banana area that is achieved is smaller than that obtained when clustering occupation on its own. This shows that when these two variables were clustered in tandem, a negative effect was achieved.

In creating the simulated trial two data, it was the first team's goal to generate data that would cluster better when all three variables were combined. In addition, it was also the first team's intention to add randomness into the data that would make it more difficult for trends to be found in clustering. Although clustering two of the variables together resulted in a smaller banana area than would be expected, the Variable Identification Procedure was able to identify three variables that, when clustered in tandem, resulted in the largest banana area of any prior clusterings.

4.2.3 Simulated Data Trial Three

The third simulated data trial was intended to fall at level four on the Trial Spectrum. This would result in data that was even more random than that crated in the second data trial. The first team anticipated that the amount of randomness introduced to the data in this trial would result in values that could not as easily be clustered. However, the Variable Identification Procedure was able to distinguish those characteristics that were meaningful.

4.2.3.1 Creating Data Set

In order to create data that was more random than the second trial, the first team decided to include some policyholders with randomly generated data in each cluster. First, the standard deviation of the normal distribution used to generate points around each centroid was changed back to its initial value that was used in the first trial. This was done to generate values for the three variables that were chosen to be the predictor variables for this trial: number of magazine subscriptions, university, and house value. Once again, the six other policyholder characteristics were assigned random values from zero to one hundred for each policyholder. To introduce the random policyholders into the data set, approximately one-third of the policyholders in each cluster were reassigned random values from zero to one hundred for the three clustering

variables. Thus, one-third of each cluster would be comprised of policyholders with completely random data. Mapping the policyholder values to the appropriate values for each characteristic was executed in the same manner as the previous two trials.

4.2.3.2 Clustering Data Set

The same methods were employed to test the third simulated data set as trial two. The banana areas calculated for each variable are listed in descending order below:

Variable	Area Difference	Percentage
House Value	**15,579,262.90**	**47.16%**
University	**14,179,993.10**	**42.93%**
Number of Magazines Subscribed	**5,523,017.90**	**16.72%**
Age at Purchase	**3,563,869.70**	**10.79%**
Driving Record	**2,564,857.90**	**7.76%**
Occupation	**2,375,491.30**	**7.19%**
Car Color	**2,240,043.20**	**6.78%**
Income	**2,126,796.60**	**6.44%**
Gender	**1,811,997.00**	**5.49%**
Proximity to Nearest Medical Facilities	**1,680,271.00**	**5.09%**
Number of Children	**1,630,645.10**	**4.94%**

Table 7 Simulation Trial Three Banana Areas for Individual Variables

After studying these values, it was obvious that the variables House Value, University, Number of Magazines and Age at Purchase led to excess large banana areas, with those of House Value and University much larger than the other two. Similarly to trial two, the top three variables were used to provide enough comparisons of joint clusterings for analysis.

Additionally, the second categorical and quantitative approaches were once again used to standardize the data. The banana areas of each of the combinations are shown in the table below:

Multiple Variables	Area Difference	Percentage of Ideal	Rank
H+U+M	**18,618,784.30**	**56.36%**	**3rd**
H+U	**23,047,352.30**	**69.77%**	**MAX**
U+M	**18,800,059.50**	**56.91%**	**2nd**
H+M	**15,889,072.80**	**48.10%**	**4th**

Table 8 Simulation Trial Three Multiple Variable Clustering Banana Areas

Note: H stands for house value, U stands for university, M stands for number of magazines subscriptions.

The results produced from this trial were different from those produced in earlier trials. In this trial, the combination of all three variables generated a smaller banana area than two of the combinations of just two variables. However, each of the banana areas created from the re-clusterings were larger than any single variable on its own. Thus, it is still possible to see the effects of joint clustering.

The data generated in this trial had a higher amount of randomness than any of the previous trials. In creating the data set, the first team could not be sure of the extent of the effect that adding random data would have in clustering. As seen in this trial, the Variable Identification Procedure was still able to identify three predictor variables in a data set in which one-third of the data was completely random. Although the three variables, when clustered together, did not result in the largest banana area, this is likely a result of the random data that was included in the data set and could not have been foreseen.

4.2.4 Simulated Data Trial Four

The fourth simulated data trial fell at level five on the trial spectrum. This meant that the data set produced was completely random. Therefore, for this trial, it was the objective of the first team to produce a data set which would generate no meaningful results.

4.2.4.1 Creating Data Set

For this trial, the first team did not use the same procedure as was used in the first three trials to generate values for each characteristic of a policyholder. Instead of using a three dimensional space to create the clusters, the first team randomly assigned values from zero to one hundred for each of the nine variables for a policyholder. This resulted in each policyholder having a random value from zero to one hundred for each of the nine characteristics. After this was complete, each characteristic was mapped to the value that it corresponded to in the table created in the first trial.

4.2.4.2 Clustering Data Set

The fourth set of simulated data provided by the first team was clustered in SAS and evaluated using the Excel macro. The following chart shows the banana areas for each variable in descending order:

Variables	Area Difference	Percentage
Age at Purchase	**3,563,869.70**	**10.79%**
Driving Record	**2,206,793.20**	**6.68%**
Gender	**1,811,997.00**	**5.49%**
Income	**1,772,786.70**	**5.37%**
Car Color	**1,752,009.90**	**5.30%**
House Value	**1,636,892.60**	**4.96%**
University	**1,613,729.80**	**4.89%**
Number of Children	**1,474,980.90**	**4.47%**
Proximity to Nearest Medical Facilities	**1,470,726.40**	**4.45%**
Number of Magazines Subscribed	**1,236,246.10**	**3.74%**
Occupation	**1,055,745.50**	**3.20%**

Table 9 Simulation Trial Four Banana Areas for Individual Variables

As can be seen in this table, clustering the variables resulted in banana areas that were quite close to that generated by the random clustering. The largest banana area the variables have was relatively small compared to previous trials. In this trial, no re-clustering was executed because only one variable had a banana area in excess of randomly generated data. This meant that even if re-clustering were to be performed, only one variable would be included, resulting in the same outcome that was achieved in the clustering of individual variables.

When the first team created this data set, it was their intention that the Variable Identification Procedure would not find any useful trends in the data. As can be seen from the clustering that was executed by the second team, the results obtained in this trial were exactly as the group intended.

Chapter 5: Recommendations and Conclusions

The greatest challenge of this project was to identify appropriate mathematical approaches to solve a behavioral problem. This section will present recommendations for changes that could be made to the mathematical approaches used, which may increase the accuracy and efficiency of the Variable Identification Procedure.

5.1 Scoring Method

One area of this project that could be improved in the future is the scoring method. The scores were meant to be used as a measure of how well the policyholder used the long-term care insurance policy. In an effort to identify the best way to model the surface associated with policyholder scores, the team tested several approaches before determining that the least squares regression plane yielded the most accurate results. However, in each of these methods the underlying factors that were used, being the financial ratio and number of years a policyholder owned a policy before going on claim, remained constant. It is possible that these factors rather than the alternative methods lacked accuracy. To test this hypothesis both of the factors should be reevaluated. Additionally, new factors could be considered to determine how well policyholders are using benefits.

5.2 Clustering

The group identified several potential revisions to the clustering methodology, which may improve the accuracy of the Variable Identification Procedure. These modifications include identifying an appropriate k for the k-means clustering technique, determining the appropriateness of using the distance between policyholders, and considering the effects of allowing policyholders to be present in multiple clusters.

Through working with the k-means clustering function in SAS, it became clear that adjusting the number of clusters that were to be created, k, did have some effect on determining a variable's predictive capability. While some variables would naturally fit into a fewer number of clusters, by programming SAS to divide the policyholders into ten clusters the group may have been over differentiating the data. Further work should be done to identify an appropriate number of clusters to be formed for each variable being tested using the k-means clustering technique.

The k-means clustering technique calculates multiple distances to determine the assignment of policyholders into clusters. Although this method was able to achieve the purpose of this project, it is possible that using density instead of distance will lead to clusters that are more accurate. In the future, other clustering techniques should be trialed which utilize the computation of a density rather than distance so that the best method may be identified.

Finally, the team did not allow a policyholder to be placed in multiple clusters during the assignment of clusters. This discrete approach may have over pigeonholed the policyholders, especially in the situation where an individual exhibits an equal amount of two distinct characteristics. Allowing policyholders to fit into more than one cluster may have an effect on the identification of predictor variables, so this approach should be trialed and evaluated for its accuracy.

5.3 Supplemental Data

The original goal of this project was to identify policyholder characteristics that may be able to indicate future benefit usage. To achieve this goal, the team created the Variable Identification Procedure, which would be able to identify any variables that may predict future benefit usage. The variables to be tested were to be drawn from both information AbilityRe had

recorded on each policyholder, as well as supplemental data, which could be obtained from a data aggregator. However, as a result of privacy concerns, supplemental data was not obtained during the course of this project. The acquisition of supplemental data in the future would allow the original goal of this project to be met, which would provide insight that could help AbilityRe offer services that are more valuable to the long-term care insurance policyholders.

5.4 Project Conclusion

This project experience offered a unique opportunity to apply mathematical methodologies in the evaluation of the impact that behavioral characteristics have on benefit usage. Although the team was unable to identify specific predictor variables as a result of privacy concerns, the utility of the Variable Identification Procedure will be applicable at any time if such information is to become available.

Additional Material on Possible Policyholder Behavior Patterns

Throughout this project, the team worked to study the behavior of the claimants, both when they purchased the long-term care insurance policy, and when they made claims against it. One way to look at the behavior and the policyholders is through the algorithms and mathematics used in predictive modeling. A second approach to understanding the activities of the policyholder is using the concept of frames to understand behavioral patterns.

This chapter presents and discusses the organizational and individual behaviors that have the potential to impact the identification of predictor variables, which would aid in the identification of future claim amounts. These characteristics include the amount and impact of policyholder interaction, the position taken by the insurer relative to policyholders making claims and receiving benefits, and the impact on both the insurer and the policyholder of utilizing supplemental data in determining variables that may indicate future spending amounts. An area that is particularly important in this analysis is an insurance company's dual goal of satisfying its customers while maintaining and maximizing profits. In an ideal scenario, these goals would be consonant; however, it is interesting to investigate the conditions that affect their mutual achievability.

6.1 Policyholder Interaction

The second step of the Variable Identification Procedure outlined by the group involved scoring the policyholders. Scores were determined using a mathematically interpolated surface based on a calculated financial ratio and the number of years an individual had the policy before going on-claim. To calibrate the scoring method, AbilityRe scored a subset of policyholders. It is important to note that doing this resulted in the scores being from the insurer's perspective. For example, an individual who has paid premiums for a substantial period of time, but has received

few benefits from the policy would be considered to have a "good" use of the policy and as a result be awarded a high score. Obviously, the policyholder has given the insurance company profits in the form of premiums while inflicting minimal costs. However, it is interesting to consider how the perception of what is considered a good use of the policy may change when considered from the policyholder's perspective.

Premiums paid for long-term care insurance policies guarantee a means of indemnity in the event long-term care is needed by the policyholder. From the policyholder perspective, it would be a good use of the policy to eventually claim the benefits for which premiums had been paid, so as to recoup the cost of the policy. Extending this notion leads to the conclusion that the best use of a policy would be paying premiums for a relatively short period of time before going on-claim, thus paying only a fraction of the amount that will be received in benefits. This is a plausible scenario because in many long-term care insurance policies the individual ceases premium payments once on-claim. This type of behavior would be given a low score in the method used by AbilityRe because a small amount of profits would have been generated as compared to the high costs that would have been incurred.

In addition to the apparent difference in defining a good use of a policy between the viewpoint of the insurance company and the policyholder, an individual given a high score by the AbilityRe scoring method still may not represent an ideal individual to insure. It is possible that individuals who held the policy for several years before going on-claim had required long-term care insurance for some time, but failed to initiate the claim. The only way to test this assumption, and uncover the potential reasoning behind it, is through a comprehensive behavioral survey of all policyholders in the AbilityRe block of insurance policies. This form of

analysis was beyond the means and scope of the project; however, several explanations for such behavior can be hypothesized.

Behaviors can be explained by reframing the situation, so that the view of the issue changes. This technique helps in identifying strategies and possibilities that will be effective in understanding and addressing the behavior. In the case of policyholders choosing whether or not to initiate a claim for long-term care insurance, two frames of reference seem applicable: the human resource frame and the symbolic frame.[36] The human resource frame focuses on needs and skills that should be addressed. Conversely, the symbolic frame focuses on meaning and the importance of creating new ways and symbols.

The human resource frame can be applied to the behavior of not initiating a claim. One of the assumptions of this frame is that "organizations exist to serve human needs rather than the reverse." Under this frame, one explanation for the policyholder not initiating the claim would be that the policyholder did not know the appropriate procedure, or first step in filing a claim. The human resource frame puts the responsibility of explaining and ensuring the understanding of the claims process in the hands of the insurance company. Additionally, this frame considers Maslow's hierarchy of needs which distinguishes five levels of needs that are satisfied in order. The levels are physiological, safety, belongingness and love, esteem, and self-actualization. In the case of a policyholder going on-claim in long-term care insurance, the benefits being received may aid in satisfying the "prepotent" needs, physiological and safety. As was stated in the Background Chapter of this report, individuals are eligible for benefits if they are no longer able to complete the activities of daily living or have cognitive impairment. In either situation individuals are unable to satisfy the two most basic needs as defined by Maslow. Recognizing

[36] L. Bolman and T. Deal. (2003). *Reframing Organizations*.

the situation has been framed in this way by the policyholder will allow the insurer to respond appropriately and work proactively to offer more valuable services. In this case, more literature or personal conversations describing the details of the claim filing process would need to be made available to all individuals owning a long-term care insurance policy. This information should be available at the time of policy purchase to ensure that claims are filed as they are needed and not delayed as a result of miscommunication. This is advantageous for the insurance company because a policyholder who waits to go on-claim may develop an intensified condition, which may have been preventable had proper care been taken, resulting in an increased net amount of money paid in benefits.

The symbolic frame can also be applied to the scenario of an individual failing to initiate the claims process. One of the main ideas of this frame is, "what is most important is not what happens but what it means."[37] In this frame the rationale for an individual not initiating a claims process would be the result of the policyholder attempting to avoid what it means to be on claim. For an individual in need of long-term care, admitting the need for benefits to pay for the care may be difficult because it is in essence admitting a need for help and a resignation to a loss of personal independence. By recognizing that the problem is being framed in this way, creative solutions can be implemented to combat the reaction and encourage obtaining benefits. The symbolic frame suggests that the solution lies in replacing the meaning that is lost. In this case, it would be imperative to highlight the financial benefits that will result from allowing the insurance company to cover some of the costs associated with long-term care. Additionally, the care should be described so that the policyholder recognizes that it is intended to help maintain

[37] L. Bolman and T. Deal. (2003).

independence by allowing the individual to continue with the tasks that are manageable while receiving support for those that are more challenging.

6.2 Insurer Perspective

Although the definition of good policy usage may be different between the insurance companies and those who are insured, which leads to different goals, it is possible that the objectives are mutually achievable. In order to accomplish this, a change needs to occur in the basic assumptions that motivate an insurance company's actions. The book *Intentional Revolutions* outlines seven methods of influence which if applied properly can aid in dynamically changing an environment. Ultimately, change would result in achievement of both goals, adding value to the customer while increasing company revenues.

The first method of influence is persuasive communication, which is "the art of presenting a proposal or suggestion so as to maximize the probability that it will be accepted."[38] This method is already being applied within the claims process of an insurance policy. Long-term care insurance providers, such as AbilityRe, outline a procedure that can be followed in the event that a claim is required. However, since a possible policyholder behavior is to ignore the need to initiate a claim immediately, it is clear that persuasive communication alone is not enough. To increase the compliance, an insurance company could mandate that a claim is filed within a specified time period following the trigger event; this method of influence is known as coercion. A drawback of this method is that policyholders may feel needless pressure to file a claim for minor events that may not be eligible for benefits, which would overwork the system.

---

[38]E. Nevis. et al. (1996). *Intentional Revolutions*.

Additionally, if a policyholder feels too constrained within this model he/she may opt to cancel their policy and purchase a new policy from a competitor firm creating a loss of profits.

A third method of influence is through role modeling, which requires the insurance company to act in the way they wish to be treated so that this behavior can be emulated by the policyholders. In the case of ensuring claims are made appropriately and promptly, the insurer should handle claims in this manner as well. For example, a good behavior would be dispatching nurses to the claimants in a timely fashion to assess the level of care needed and review benefit options once the claim is initiated. This type of reaction would show that the insurance company takes the claims process seriously and that prompt action is considered a priority. These actions would be desirable if emulated by the policyholder. Although this method is subtle it is very effective in bringing about change.

Participation is another highly effective influence method. Participation involves actively engaging the stakeholders, such as policyholders, to understand their needs, concerns, and perspectives in an effort to create acceptance and greater compliance with the finalized plan. Central to the concept of participation is the idea of sharing power, especially the power to change or influence a decision.[39] Policyholders represent a large and often dispersed population; as such the logistics of a participatory influence method are complex. One way this method could be implemented is through a survey of policyholders to gain a better understanding of the behaviors surrounding decision making. This type of investigation would reveal any barriers and the factors that influence the claims process. Outcomes of the survey can be used to revise the current process. To maximize the impact, the influence method of structural rearrangement could

[39] E. Nevis. et al. (1996). *Intentional Revolutions*.

be applied as well. This technique involves making it more likely that tasks are executed successfully. Ideally, once a new process is established as a result of participation, structural rearrangement can ensure that the resources are reallocated and adjusted as necessary. If an outcome in the survey suggested that policyholders did not understand the process of initiating a claim, a structural arrangement that might follow as a result would be to increase the number of customer service representatives available, or possibly arrange for presentations to be given to policyholders preemptively. Both of these services would add value to the insurance policy, leading to increased customer satisfaction and potentially increasing the profits of the company.

A sixth method of influence is expectation. Expectation is a more implicit method than persuasive communication or participation because it is a subtle way of eliciting a behavioral change.[40] The concept behind expectancy is that a self-fulfilling prophecy is brought about. "The predictor makes some assumptions about the target of the prediction and then acts in such a way as to make the predictions come true." If the insurance company wants the policyholders to maximize the value of their plan and use their benefits appropriately, the chances of this occurring will increase if the company believes such action is possible and acts on that belief. The insurer must adopt the mindset that individuals are not intentionally attempting to misuse policy benefits or planning the timing of the policy purchase so that the cost in premiums is lower than the benefits that will be received. Both of these examples are of negative expectancies. Instead, the insurance company must adjust its expectancy to the idea that individuals purchase long-term care insurance policies to protect against financial losses that may be incurred if such care is necessary in the future. This would be an example of a positive expectancy. These revised outlooks cannot be imposed, but must be internalized by insurance

[40]E. Nevis. et al. (1996). *Intentional Revolutions*.

company leadership to ensure they are fully invested and ready to modify the ways in which they interact with policyholders. By acting on the positive expectancies, the desired behaviors will be elicited from the insured. Although this method is often criticized for being overly optimistic, several studies have shown its lasting effect.

A final method of influence is extrinsic rewards. This environmental change focuses on rewarding positive behavior and ignoring negative behavior. Extrinsic rewards can take many forms ranging from verbal praise to monetary payments. The latter incentive may be more effective in the case of insurance companies and their interaction with policyholders. To counteract the behavior of failing to initiate a claim, deductions or stipends may be offered to those individuals who file the claim at the onset of a condition. If medical attention is received at this point, it is possible that the overall benefit costs may be reduced or better managed, thus adding value to the service for the policyholder and decreasing costs for the insurer. Conversely, if the claim is delayed and the condition worsens, insurer profits would decline and even the claimant might suffer. Through working with the AbilityRe dataset it was clear that an individual may utilize a long-term care insurance policy and go on-claim several times. Thus, policyholders can learn through experience the behavior that is considered good and therefore rewarded by the insurer.

The impact of the change can be maximized by combining the methods of influence in a way that highlights each method's strengths while diminishing its weaknesses. The greatest challenge in initiating this change is having the result internalized by all stakeholders and suppressing the reflex to operate under the formerly used practices.[41] However, the result of

[41] E. Nevis. *Intentional Revolutions*.

changing the behavior will be substantial because achieving the consonant goals of the insurance company and the policyholder will satisfy both parties.

6.3 Supplemental Data Application

The Variable Identification Procedure, which was developed by this project group, is intended to be applied to supplemental data, so that predictor variables may be identified. As a result of HIPPA regulations, the project group was unable to test the methodology by applying it to supplemental data that could have been collected on an individual policyholder basis. However, some policyholder information, such as age and gender, was available to be used by the group as it was stored by AbilityRe. This data provided a good basis to start the variable analysis, but more information about each policyholder is necessary before determining predictor variables. The increasing availability and applications of supplemental data within the insurance industry warrant a discussion of its positive and negative effects.

An article from The Economist discussed this contemporary issue, which has been rapidly gaining public attention and concern. The pandemic generation, collection, and consumption of data has had transformative effects on business, society, and culture. Additionally, there are several methods for best internalizing and understanding all that is available and new regulatory concerns that have arisen as a result of its wide usage. Since the amount of information and its usage in predictive modeling is unlikely to slow in the near future, understanding the best practices is critical.

The amount of digital information available is growing at an increasingly rapid rate, and although the technology used to generate, maintain, and aggregate the data has improved, the amount of data available has already exceeded the available storage capacity. This phenomenon,

where technology is producing more information then can be feasibly stored or used, is known as "big data".[42]

As a result of the growing information trend, businesses such as insurance companies are looking for ways to analyze the data and identify trends, which would aid in predicting the future needs of customers. Predictive modeling can lead to proactive business practices that increase efficiency by minimizing cost and maximizing consumer value. Identifying these macroeconomic trends requires skill and understanding in the field of mathematics; however, these models are not always perfect predictors of the real word, so human judgment and monitoring are still necessary.

Sales data is the most valuable type of information for a company looking to implement predictive modeling practices to improve its business. This form of business intelligence was once exclusive to only the largest of corporations, but has become more common as a result of the decreasing costs of the necessary technology. Data mining tools, such as the ability to forecast and correlate data, result in more targeted marketing and a better understanding of the customer's needs. To supplement the information that has already been gathered on a customer by the company and is stored within corporate databases and records, supplemental data is often gathered on customers. This information could include variables such as occupation or family status. As a result, an increasing number of business decisions are based on mathematical algorithms as opposed to individual intuition. This statement holds true for AbilityRe, because in sponsoring this project, the company was hoping to develop a method and identify policyholder variables that would aid in predicting future claims amounts. The method that was developed, the

[42] "Data, Data Everywhere." (2010). Print.

Variable Identification Procedure, uses mathematics rather than intuition alone to assess variables predictive ability.

One negative associated to supplemental data is that due to the vast quantities, it is often difficult to conceptualize the data that is available. This task involves taking the "inhuman scale of the information and the need to present it at the very human scale of what the eye can see."[43] Text or numeric data in large amounts are difficult to understand completely, but, often, when presented visually can be interpreted in a fraction of the time. Data visualization allows users to more fully understand the problem, ultimately resulting in the potential for more complete and creative solutions.

In using supplemental data, there are an array of ethical and legal issues that need to be considered. First is the issue of privacy of personal information, which individuals would like to preserve and companies would like to exploit. This issue arose in the form of HIPPA regulations as the team attempted to obtain supplemental data on policyholders. This act protects policyholders of long-term care insurance because it falls within the broadly protected category of health insurance. Additionally, this act is meant to give greater control over the availability and usage of medical records to patients. However, it also prohibits holders of the information to exchange the data with marketing companies, such as supplemental data providers, without explicit patient permission.[44] As a result, the predictive modeling capabilities within the health insurance domain and more specifically in long-term care insurance are limited, unless provisions for obtaining supplemental data are made. A second concern is that information

---

[43] "Data, Data Everywhere." (2010). Print.

[44]"Understanding Health Privacy". *Health Information Privacy*. U.S. Department of Health and Human Services. Retrieved on April 7 2010 from <http://www.hhs.gov/ocr/privacy/hipaa/understanding/index.html>.

security must be made a priority for corporations so that their systems and networks are protected from breaches.[45] Supplemental data can be purchased from data aggregating companies, so not only is the information that is being stored personally identifiable, but it also represents an investment as it was purchased for a fee. Having such records to use in a predictive modeling scenario may give one company a competitive advantage over peers, so ensuring that the information is kept secure is critical. A third concern relates to the power of computer algorithms. While contributing greatly to understanding, mathematics should not be thought to completely replace human intuition and reaction. Techniques such as data clustering, as used in the group's Variable Identification Procedure, may not be prepared to accurately interpret information, and there is the possibility that these computerized methods will cluster or categorize individuals together which should have been separated. Additionally, these methods may make generalizations that could be avoided if experts in the field analyzed the data. This leads to another issue, which is that of storing digital records. While some believe information should be retained, others argue that the information becomes obsolete and should be refreshed regularly. Outdated information may lead to inaccurate conclusions and predictions, but constantly refreshing supplemental data could result in a significant financial investment for the company. Finally, it is imperative that the integrity of the data be maintained, which requires companywide international cooperation.[46] Supplemental data that is intended to be used in predictive modeling to improve a company's products or services need to remain free of errors so that the conclusions made accurately represent the consumers.

[45] "Data, Data Everywhere." (2010). Print.
[46] "Data, Data Everywhere." (2010). Print.

Predictive modeling and computer algorithms can be utilized to simplify, condense, and interpret supplemental data available making it more “digestible for humans”. Supplemental data has the potential to reveal trends which could improve the services and products provided by a company, therefore increasing consumer satisfaction and potentially increasing profits. However, many factors must be taken into consideration while working with supplemental data to ensure that the analyses executed are purposeful and ethical.

References

Ability Resources, Inc. *Company Profile*. Retrieved October 2009, from Ability Resources, Inc. http://www.abilityresources.com/.

America's Health Insurance Plans. (2004). *Guide to Long-Term Care Insurance.* Retrieved October 8, 2009, from http://www.ahip.org/content/default.aspx?docid=21018.

Batty, Mike, James Guszcza, Alice Kroll, and Chris Stehno. "Bringing Predictive Models to Life." *Contingencies* Winter 2009: 4-14. Print.

Bolman, Lee, Terrence E. Deal. *Reframing Organizations*. San Francisco, CA: John Wiley & Sons, Inc, 2003. Print

Brown, J. & Goolsbee, A. (June 2002). Does the Internet Make Markets More Competitive? Evidence from the Life Insurance Industry. *The Journal of Political Economy*. 110, 3, 481.

Cohen, M., Miller, J. & Weinrobe, M. (August 2002). *Inflation Protection and Long-Term Care Insurance: Finding the Gold Standard of Adequacy*. Retrieved on October 8, 2009 from http://assets.aarp.org/rgcenter/health/2002_09_inflation.pdf.

"Data, Data Everywhere." *The Economist* 27 February 2010: 3-18. Print.

Davidson, S.K. (March 2006). US Patent No. 20060059020A1. Washington D.C.: US Patent and Trademark Office.

ElderLawNet, Inc.(2008). *Long-Term Care Insurance*. Retrieved on October 09, 2009 from http://www.elderlawanswers.com/elder_info/elder_article.asp?id=2595.

Family Caregiver Alliance. (2005). *Selected Long-Term Care Statistics.* Retrieved on March 27, 2010 from http://www.caregiver.org/caregiver/jsp/content_node.jsp?nodeid=440.

Genworth Financial. (April 2009). *Genworth 2009 Cost of Care Survey*. Retrieved October 8, 2009, from Genworth Financial:http://www.genworth.com/content/etc/medialib/genworth_v2/pdf/ltc_cost_of_care.Par.8024.File.dat/cost_of_care.pdf.

Health Grades Inc. (2009). *Cognitive Impairment*. Retrieved on October 09, 2009 from http://www.wrongdiagnosis.com/sym/cognitive_impairment.htm.

"K-Means Clustering". A Tutorial on Clustering Algorithms. Retrieved on April 11, 2010 from http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html.

Konrad, W. (2009, June 26). Getting Insurance for One's Frailest Years. *The New York Times*.

Long Term Care Insurance Tree. (2009). *What are ADLs?* Retrieved October 8, 2009, from

http://www.longtermcareinsurancetree.com/ltc-basics/what-are-adls.html.

Metropolitan Life Insurance Company. (2009). *The Essentials of Long-Term Care Insurance*. Retrieved on October 09, 2009 from www.metlife.com/.../long-term-care-essentials/mmi-long-term-care-insurance-essentials.pdf.

Morley, John. (2008). *Managing Cognitive Dysfunction*. Retrieved on October 09, 2009 from http://www.thedoctorwillseeyounow.com/articles/senior_living/cogdys_6/.

Nevis, Edwin, Joan Lancourt, Helen Vassallo. *Intentional Revolutions*. SanFrancisco, CA: Jossey-Bass Inc, 1996. Print.

Pfuntner, J., Dietz &, E. (2004, January 28). *Long-term Care Insurance Gains Prominence*. Retrieved October 8, 2009, from United States Bureau of Labor Statistics: http://www.bls.gov/opub/cwc/cm20040123ar01p1.htm.

The Prudential Insurance Company of America. (September 2008). Long Term Care Product Guide. Retrieved on October 8, 2009 from http://www.nfn.crumplifeinsurance.com /BISYSdocs/ltc/LTC%20EVOLUTION%20Product%20Guide.pdf.

SAS Institute Inc. *The FASTCLUS Procedure.* 2003. Retrieved on Feb. 22, 2010 from http://support.sas.com/onlinedoc/912/getDoc/ statug.hlp/fastclus_sect1.htm.

Shelton, P. (2003). Long-Term Care: Your Financial Planning Guide. Kensington Publishing Corp.: New York, NY.

"Understanding Health Privacy". *Health Information Privacy*. U.S. Department of Health and Human Services. Web. Retrieved on April 7 2010 from http://www.hhs.gov/ocr/privacy/hipaa/understanding/index.html.

U.S. Census Bureau, Housing and Household Economic Statistics Division. *U.S. Census Bureau*, 2009. Web. 3 December 2009.

Wiener, J. M., Hanley, R. J., Clark, R., & Van Nostrand, J. F. (1990). *Measuring the Activities of Daily Living: Comparisons Across National Surveys.* United States Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation Office of Social Services Policy.

Wilson, Lawrence. (2008). *Brain Fog*. The Center for Development. Retrieved on October 09, 2009 from http://www.drlwilson.com/Articles/brain_fog.htm.

Glossary

Activities of daily living (ADLs)- a list of actions that are considered the basics of independent self-care, which serve as a metric for determining an insured's ability level and helps in the determination of the necessity of going on claim for long-term care insurance

Cognitive impairment- abnormally poor or low mental function

Elimination period- an established amount of time between the onset of an illness or disability and disbursement of claim payments from insurance company

Insurance riders- amendments that can be purchased by the insured from the insurance company at any time to allow changes to the coverage provided in insurance policy

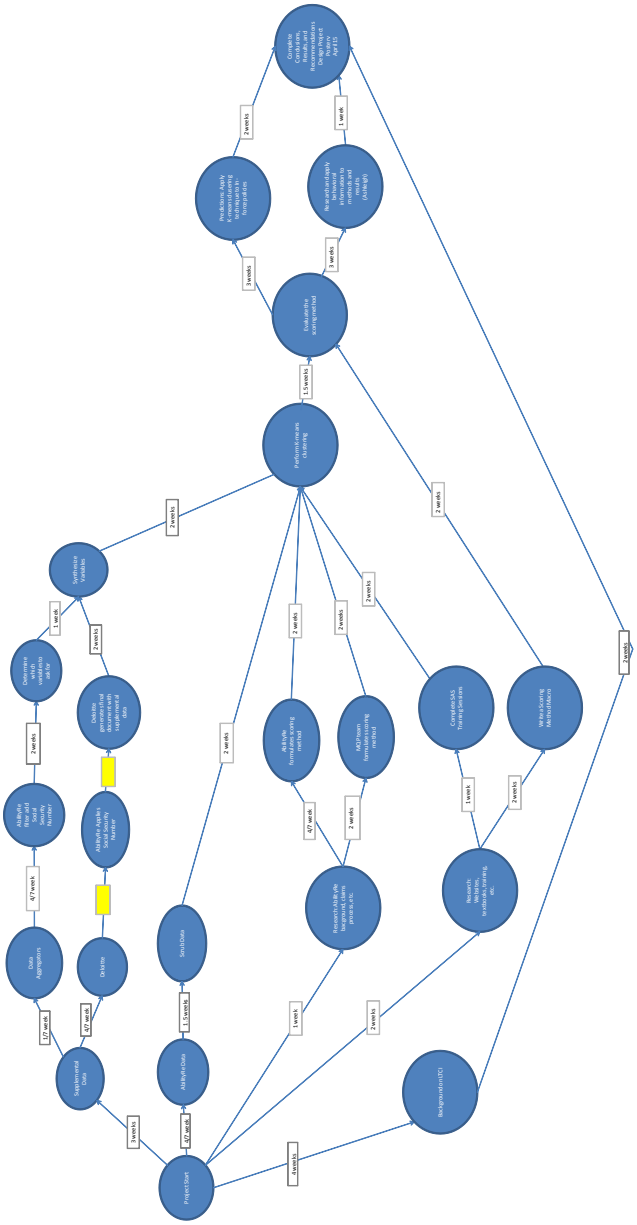
Long-term care (LTC)- the assortment of services that work to support individuals needing medical assistance over an extended period of time

Long-term care insurance (LTCI) - a form of insurance coverage that alleviates a portion of the out-of-pocket financial burden LTC providers or facilities can have on the elderly individuals in need

Maximum daily benefit- is the amount of coverage that will be paid daily by the insurance company to the insured once the policyholder is on claim status

Premium- the cost paid regularly by the insured for the insurance policy coverage that is received

Appendix A: Project Timeline



Appendix B: Proposed Scoring Method

Proposed Scoring Method- Financial Perspective

Ratio= How much policy holder used / How much policy holder paid in premiums

1 -> .66= Good Use
.65 -> .33 = Avg Use
.32 -> 0 = Poor Use

Legend:
Green- Numbers to be used in final ratio
Red-data from policy holder data files
Orange- calculated fields

1. How much the policy holder used
   a. Total Current Benefit Used
      i. For first claim incident- Calculate Benefit Amount Used
      [(Close_date)-(report_date)] x (sum[(nh_daily_benefit_amount)+ (hhc_daily_benefit_amount)+ (alternative_care_benefit_amount)])= Current Benefit Used Claim 1

      ii. For each subsequent claim incident for a single policy holder
      * repeat step i.

      iii. Sum Calculated Benefit Amount Used for all claim incidents
      Current Benefit Used Claim 1 + Current Benefit Used Claim 2+...= Total Current Benefit Used

   b. Projected Future Benefit Use
      i. Repeat steps a) i.-iii. For all deceased on claim policy holders

      ii. Sum all of the deceased on claim policy holders Total Current Benefits Used
      sum(Total Current Benefits Used)= Deceased Total Current Benefits Used

      iii. Average
      Deceased Total Current Benefits Used/ Number of deceasred on claim policy holders= Average time receiving benefits

      iv.Projection
      [Average Time Receiving Benefits- (sum (close_date)-(report_date))] x (sum[(nh_daily_benefit_amount)+ (hhc_daily_benefit_amount)+ (alternative_care_benefit_amount)])= Projected Future Benefit Use

   c. Sum of Total Current Benefits Used and Projected Future Benefit Use
   Total Current Benefit Used + Projected Future Benefit Use = How Much the Poilcy Holder Used

2. How much the policy holder paid in premiums
   a. Date the policy holder went on claim (report_date) / "-"Date the policy holder bought the policy = How long the policy holder has been paying premiums

   b. How long the policy holder has been paying premiums / Premium Period (payment_days_quantity) "=" Number of Premium Periods

   c. Number of Premium Periods X Premium Amount (charge_amount) "=" How much the policy holder paid in premiums

Appendix C: Formulas for Calculating Financial Ratios

1. Calculated Financial Ratio Including Reserve =

$$\frac{\mathbf{DLR + ALR + Benefits}}{\mathbf{Preminums - Refund}}$$

2. Calculated Financial Ratio Without Reserve=

$$\frac{\mathbf{Benefits}}{\mathbf{Preminums - Refund}}$$

3. Calculated Financial Ratio with Projected Reserve=

$$\frac{\mathbf{Projected\ Reserve + Benefits}}{\mathbf{Preminums - Refund}}$$

   a. Where Projected Reserve=

$$(\mathbf{Average\ Claim\ Length\ for\ Dead - Current\ Claim\ Time})\frac{\mathbf{Benefits}}{\mathbf{Total\ time\ on\ Claim}}$$

   b. Where Average Claim Length for Dead=

$$\frac{\mathbf{\sum Claim\ Lengths\ for\ Dead}}{\mathbf{Number\ of\ Claims\ from\ Dead\ Policyholders}}$$

Appendix D: SAS Customized Code

The customized FASTCLUS coding in SAS that was written by the group with the help of Toto is as follows:

```
libname project 'E:\WPI\COURSES\09\MQP\AbilityRe\CLUSTERING';
proc print data=project.policydata;
run;
quit;
```

-The command print will show the data that will be clustered.

```
proc contents data=project.policydata;
run;
quit;
```

-This step is intended to show all the variable names that could be used for clustering.

```
proc fastclus data=project.policydata maxclusters=2 OUT=project.out_set1 list
OUTITER OUTSEED=temp;
var Age_at_Purchase;
id Unique_Identifier;
run;
```

-This is the main procedure performing k-means clustering. In the test, the group took the age at purchase of the policy holders as the variable and unique identifier as the observation IDs to run the clustering procedure. In this case, the only data that was clustered is the age at which the policyholders purchased their policies, and the clustered data file is named as project.out_set1.

```
data project.policydata2;
set project.policydata;
keep Policy_Number;
run;
```

-This step creates a new data file "policy.policydata2" and edits the column attributes for future use.

```
proc print data=project.policydata2;
run;
quit;
```

-The print command shows the new data file "policy.policydata2" that was just created.

```
proc sort data=project.out_set1;
by CLUSTER;
run;
```

-This step sorts observations in the order of clusters to which they belong.

```
symbol v=dot;
proc gplot data=project.out_set1;
plot Avg_Family_Size*Unique_Identifier=CLUSTER;
```

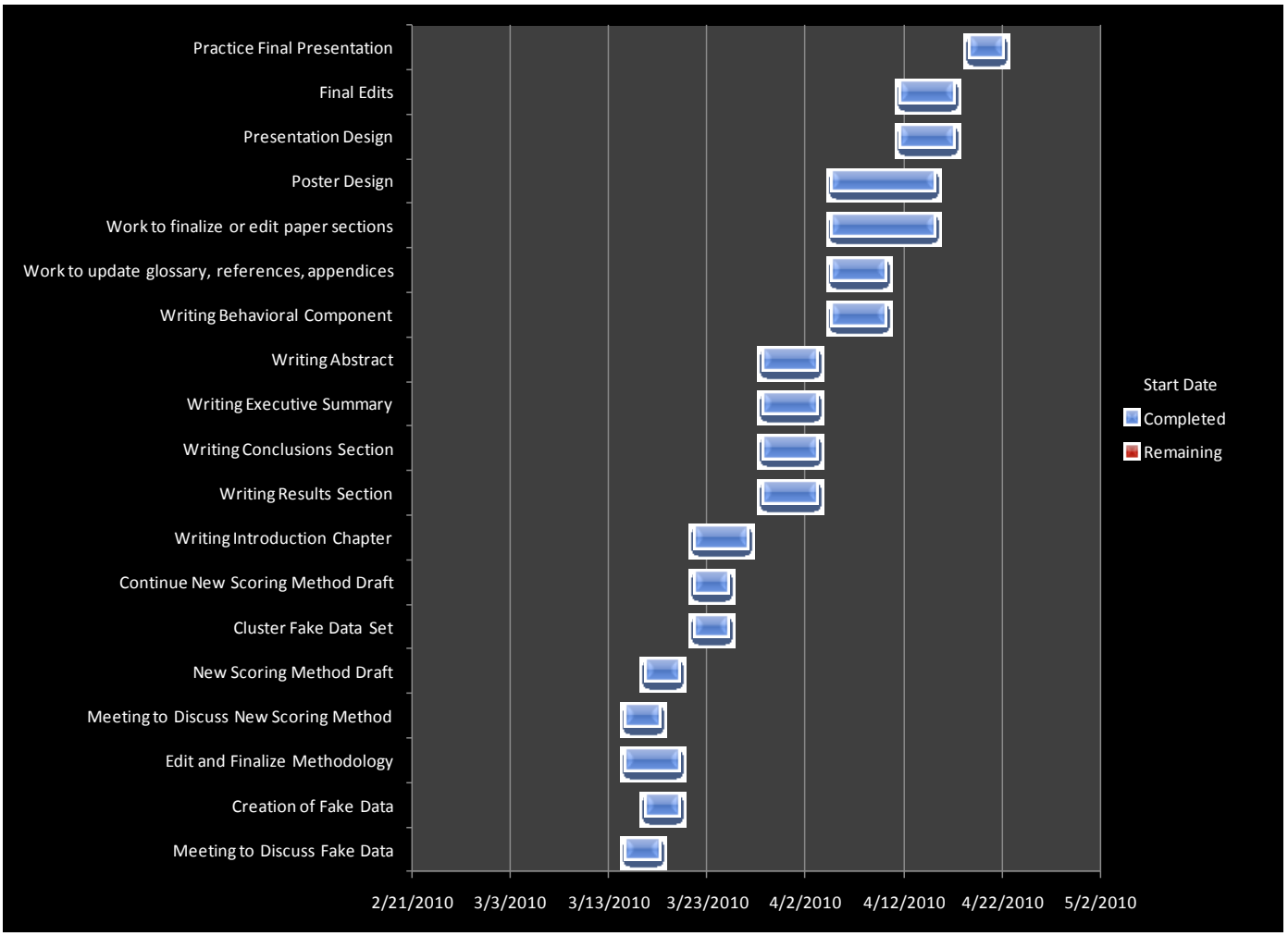
```
run;
```
-The distribution of the clustering results is plotted in this procedure. The observations can be plotted in different patterns and using different colors, providing a clearer picture of the results for further analysis. In this test, the observations are plotted as dots.
```
proc print data=project.out_set1;
run;
```
-The clustered data file project.out_set1 is showed by the print command here.
```
quit;
```

Appendix E: Policyholder Scores from AbilityRe

Edge Points			
Unique ID	Financial Ratio	Number of Years Before Claim	Ability Re Score
25063	0	3.61369863	
9173	0.246516029	2.934246575	80
109655	1.003475243	3.934246575	55
126170	4.648232397	3.101369863	12
117399	13.55911541	5.909589041	
139665	45.40809832	7.164383562	1
445056	315.0883308	9.336986301	15
140369	42.55239704	13.42191781	4
140000	22.44959508	14.95616438	18
139744	13.56838874	19.17808219	75
140248	7.825045221	20.55068493	4
57221	7.181361952	21.2	32
445124	5.591829127	21.01917808	28
58935	4.587817021	23.05479452	45
140155	3.488403851	22.12876712	43
5049	2.50610903	31.45205479	46
71925	2.630641987	27.45205479	50
3423	0.490514478	33.11506849	75
140115	0.225551303	32.48767123	85
25416	0	32.33424658	83

Bucket Sampling			
Unique ID	Financial Ratio	Number of YearsBefore Claim	Ability Re Score
98966	0	2.769863014	91
51197	0.157068063	4.350684932	86
13259	0	4.230136986	89
10816	0.215395254	11.85753425	97
87704	0.083633976	5.904109589	99
139047	0.135263551	12.95068493	100
127308	0.197601325	19.64657534	93
129075	0.087107707	17.90136986	
96373	0.225196116	22.2109589	84
12327	1.158822129	17.90410959	90
55522	0.473134038	19.8	92
87816	1.464531423	23.54520548	81
28123	0.543824497	7.854794521	87
34860	0.302416675	13.10410959	
54143	1.974904967	5.731506849	37
135710	0.899856253	4.920547945	92
100079	1.54640615	4.975342466	47
10956	0.490259111	4.780821918	82
140130	4.925918367	4.635616438	40
445024	17.59076203	4.931506849	6
41542	8.195270735	4.501369863	25
314	3.540802883	5.904109589	33
2733	6.616956618	12.38630137	7
11014	3.436488089	9.731506849	35
20578	3.780774895	19.50136986	70
30430	4.960206305	17.87945205	9
58935	4.587817021	23.05479452	45

Appendix F: Gantt Chart for End of Project



Appendix G: Excel Macro to Calculate Banana Areas

This macro is to be used with the supplemental spreadsheet entitled “Score Evaluation Template.xlsm”.

```
Sub readData()

Dim i As Integer
Dim numSets As Integer

'Number of cluster sets that have been included in the data sheet
'This should be altered to fit the number of clusterings that have been performed
numSets = 1000

For i = 1 To numSets
   ActiveSheet.Range("A2").Select
   Worksheets("DATA").Range("A2:A2927").Copy
   ActiveSheet.Paste
   ActiveSheet.Range("B2").Select
   Worksheets("DATA").Range("B2:B2927").Copy
   ActiveSheet.Paste
   Worksheets("DATA").Range(Worksheets("DATA").Range("C2").Offset(0, (i - 1)),
Worksheets("DATA").Range("C2").Offset(0, (i - 1)).End(xlDown)).Copy
   ActiveSheet.Range("C2").Select
   ActiveSheet.Paste

   Range("A1:C1").Select
   ActiveWorkbook.Worksheets("Variable " & i).Sort.SortFields.Clear
   ActiveWorkbook.Worksheets("Variable " & i).Sort.SortFields.Add Key:=Range( _
      "C2:C2927"), SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:= _
      xlSortNormal
   ActiveWorkbook.Worksheets("Variable " & i).Sort.SortFields.Add Key:=Range( _
      "B2:B2927"), SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:= _
      xlSortNormal
   With ActiveWorkbook.Worksheets("Variable " & i).Sort
      .SetRange Range("A1:C2927")
      .Header = xlYes
      .MatchCase = False
      .Orientation = xlTopToBottom
      .SortMethod = xlPinYin
      .Apply
   End With
   Range("E2:H2").Select
   ActiveWorkbook.Worksheets("Variable " & i).Sort.SortFields.Clear
```

```
    ActiveWorkbook.Worksheets("Variable " & i).Sort.SortFields.Add Key:=Range( _
        "F3:F12"), SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:= _
        xlSortNormal
    With ActiveWorkbook.Worksheets("Variable " & i).Sort
        .SetRange Range("E2:H12")
        .Header = xlYes
        .MatchCase = False
        .Orientation = xlTopToBottom
        .SortMethod = xlPinYin
        .Apply
    End With

    ActiveSheet.Copy After:=Sheets(i + 2)
    Sheets("Variable " & i & " (2)").Select
    ActiveSheet.Name = "Variable " & (i + 1)
    Range("A2:C2927").Select
    Selection.ClearContents

    Sheets("Variable " & i).Select
    Range("A1:N2927").Select
    Selection.Copy
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    ActiveSheet.Range("N20").Copy
    ActiveWorkbook.Sheets("Differences").Select
    Range("B1").Select
    ActiveCell.Offset(i, 0).PasteSpecial xlPasteValues

    Sheets("Variable " & (i + 1)).Select

Next i

End Sub
```