Data Mining in the Insurance Industry

Solving Business Problems using SAS® Enterprise Miner™ Software

The Power to Know™

Table of Contents

Content for “Data Mining in the Insurance Industry” was provided by Sanford Gayle, Senior Systems Analyst, SAS Institute Inc.

List of Figures

Abstract

Data mining can be defined as the process of selecting, exploring, and modeling large amounts of data to uncover previously unknown patterns. In the insurance industry, data mining can help firms gain business advantage. For example, by applying data mining techniques, companies can fully exploit data about customers' buying patterns and behavior and gain a greater understanding of customer motivations to help reduce fraud, anticipate resource demand, increase acquisition, and curb customer attrition.

This paper discusses how insurance companies can benefit by using modern data mining methodologies and thereby reduce costs, increase profits, retain current customers, acquire new customers, and develop new products.

Changes in Information Technology

As in other sectors of the economy, the insurance industry has experienced many changes in information technology over the years. Advances in hardware, software, and networks have offered benefits, such as reduced costs and time of data processing and increased potential for profit, as well as new challenges particularly in the area of increased competition.

Technological innovations, such as data mining and data warehousing, have greatly reduced the cost of storing, accessing, and processing data. Business questions that were previously impossible, impractical, or unprofitable to address due to the lack of data or the lack of processing capabilities can now be answered using data mining solutions. For example, a common business question is, “How can insurance firms retain their best customers?” Through data mining technology, insurance firms can tailor rates and services to meet the customer’s needs, and, over time, more accurately correlate rates to the customer behaviors that increase exposure.

Modern data mining technologies also offer more accurate and better performing models that are generated in less time than that with previous technologies. Graphical user interfaces (GUIs) enable more complex models, more granularity of customer and product markets, and more sophisticated comparisons across methodologies. By generating better, extensively tested models in less time than was previously possible, insurance firms can more accurately address issues such as moral hazard in underwriting and the adverse selection in marketing.

Changes in U.S. Legislation

Recent federal legislation in the United States has cleared the path for changes in the way U.S. insurance firms can operate and compete in the United States and internationally. Although they have their roots in the Depression of the 1930s, the legislative changes offer modern-day opportunities and challenges for those insurance firms that employ enabling technologies such as data mining to compete in the growing global economy of the 21st century.

The Financial Services Modernization Act of 1999

November 12, 1999, U.S. President Clinton signed into law the Financial Services Modernization Act,[1] which effectively repealed Depression-era financial legislation by enabling insurance companies, banks, and securities firms to affiliate with one another. Prior to that signing, the United States was one of only two major world economies with legislation prohibiting insurance companies, banks, and securities from offering each other’s products and services.[2] The U.S. prohibitions were based on Depression-era judgements made about the causes of the Stock Market crash of 1929 and the ensuing economic woes. Those judgements led to the Glass-Steagall Act of 1933 and, 23 years later, to the Bank Holding Company Act of 1956.[3]

Erosion of the Glass-Steagall Act of 1933

During the past 60 plus years of Glass-Steagall, the financial services industry in the United States has seen the steady erosion of the effects of the legislation through increased mergers, acquisitions, and distribution agreements among insurance companies, banks, and securities firms.

An early reconsideration of Glass-Steagall included Senator Glass himself, who in 1935 joined other senators to weaken significantly parts of the act. Despite Senator Glass’ efforts and many other legislative attempts over the years, Glass-Steagall remained intact legislatively, if not functionally (Bentson 1990, p. 222).

During the 1980s and 1990s, the erosion of Glass-Steagall accelerated due to a rapid increase in financial sector mergers and acquisitions in the United States and internationally. For example, since 1980 there have been over 7,000 bank mergers in the United States. From December of 1997 to mid-year 1998 alone, the five largest mergers or acquisitions

[1] The Financial Services Modernization Act of 1999 (S.900) or simply, “the Modernization Act,” also is known as the “Gramm-Leach-Bliley Act of 1999 (Thomas 1999).”

[2] Japan was the other country with such prohibitions, which were enacted as a part of post-World War II reconstruction of Japan’s economy and which were based on the U.S. judgements about the causes of the Depression (Bentson 1990, p. 2).

[3] The Glass-Steagall Act of 1933, also known as “The Bank Act of 1933,” separated commercial banking and investment firms. The Bank Holding Company Act of 1956 separated banking and insurance (House Committee on Banking and Financial Services 1997).

approved or announced involved approximately $500 billion of bank assets (Francis 1998).

Perhaps the most important single merger to set the stage for the demise of Glass-Steagall was the April 1998 megamerger of Citicorp and the Travelers Group, which resulted in Citigroup. That merger—the biggest of its kind—created a combined equity valuation of about $145 billion (Wettlaufer 1998). Given the growing weight of such mergers, acquisitions, and distribution agreements, the repeal of Glass-Steagall was seen by many in the industry as inevitable.

Opportunities and Challenges for Insurance Firms

U.S. government officials have hailed the Modernization Act as "the most significant overhaul of our financial services industry laws in more than a generation (Watts 1999)," and "the most important legislative changes to the structure of the U.S. financial system since the 1930s (Clinton 1999)." By most accounts, these are reasonable statements of the historical and legislative significance of the act, but the question remains, "What opportunities and challenges is the repeal of Glass-Steagall likely to bring to the insurance sector of the financial services industry in the next few years and beyond?"

Although a range of predictions exists from business as usual in the near term (two to five years) to "one stop-stop shopping" at financial institutions (Pellegrini 1998), most observers of the insurance industry see opportunities and challenges occurring as a result of the following:

- Further consolidation.
- Changes in distribution methods.
- Increased competition.
- Demutualization.
- Redomestication.

Opportunities for growth and stability exist for insurance firms that are able to provide customers with new, innovative products and services. The challenges are the result of operating in a highly competitive and dynamic business environment.

Further Consolidation

With the repeal of Glass-Steagall, the consolidation of the financial services industry so prevalent in the 1980s and 1990s is likely to continue for the near future resulting in the formation of several major players along with numerous, smaller niche players. Through the use of holding companies and subsidiaries, financial firms of all sizes will be able to offer a myriad of products and services.

Both the large, full-service companies and the smaller specialized firms will need to analyze their target markets carefully and conduct sales and marketing campaigns that provide the best return on investment.

Changes in Distribution Methods

Changes in the way insurance is distributed began prior to the enactment of the Modernization Act. For example, for some time banks and insurance firms have signed distribution agreements enabling insurance products to be offered along with bank loans. Some insurance firms have taken the distribution agreement approach a step further by creating separate distribution companies that can take advantage of creative distribution channels such as the Internet. These two trends—the growth of distribution agreements and the creation of separate distribution companies—are likely to continue.

Insurance companies that succeed at changing their distribution methods may find some degree of protection for acquisition if the banks find the distribution agreements to be just as lucrative yet simpler than providing the underwriting themselves. To do so, insurance firms that survive and maintain some degree of autonomy will need sophisticated data warehousing, data mining, and reporting software that can incorporate new features such as web-enablement to provide up-to-date, dynamic business intelligence.

Increased Competition

As insurance companies, banks, and securities firms continue to form new corporations, holding companies, and distribution agreements, the trend toward competition in the financial services industry should continue if not increase significantly. Competition for customers will lead to innovation and diversity in financial products and services. The impact on the consumers could be significant. In fact, the U.S. Treasury Department estimates annual savings to consumers of $15 billion as firms compete for business in a fluid market comprised of educated, computer literate consumers (Watts 1999).

To survive in the new world of financial services, insurance firms must be able to match their product and service offerings with the sophistication and variations in their marketplace. That kind of business intelligence requires a data mining system that can sift enterprise-wide data enabling insurance firms to:

- Analyze the profiles and preferences of existing customers.
- Predict customer buying habits.
- Focus sales and marketing campaigns on prospects who have a high probability of becoming customers.

[4]In order to engage in the new financial activities, all banks affiliated under a holding company or through operating subsidiaries are prohibited from expanding into insurance underwriting and securities activities until they demonstrate satisfactory or better ratings under the federal government's Community Reinvestment Act (Thomas 1999).

[5]An important provision of the Modernization Act ensures that banks entering the insurance business will be subject to state insurance regulations. If banks do choose to underwrite insurance, they may do it only within a state-licensed and state-regulated insurance company affiliate owned by a bank or a bank holding company (Thomas 1999).

Demutualization

The increase in demutualization among life insurance companies that preceded the signing of the Modernization Act is likely to continue for the near future. The bull market of the years prior to the signing increased the market capitalization of many U.S. banks to unseen levels. That increased capitalization has brought them a degree of power not shared by most mutuals in the insurance sector. By going stock, insurance firms might reap some of the same benefits. The risk, of course, of demutualization is of opening up the firm to acquisition. Rather than take that risk, some mutuals will instead opt to form mutual holding companies, which will provide a measure of protection while enabling access to equity markets.

Redomestication

The Modernization Act enables mutual life insurance companies to relocate their business operations if the State where they are located does not enact provisions for the formation of mutual holding companies. Relocating or "redomestication" to another state with such provisions will enable mutuals to form holding companies, which can issue stock to raise capital.

Redomestication does not come without costs. Obviously, there is the physical move and all that it entails including, perhaps, the cost and disruption of moving an IT shop. Less obvious may be the costs associated with analyzing and selling in new markets with different demographics. To succeed, mutuals that relocate must have a flexible, scalable, and comprehensive data mining system that can analyze such massive and ongoing changes in business practices.

Using Data Mining in the Insurance Industry

Data mining methodology often can improve upon traditional statistical approaches to solving business solutions. For example, linear regression may be used to solve a problem because insurance industry regulators require easily interpretable models and model parameters. Data mining often can improve existing models by finding additional, important variables, by identifying interaction terms, and by detecting nonlinear relationships. Models that predict relationships and behaviors more accurately lead to greater profits and reduced costs.

Specifically, data mining can help insurance firms in business practices such as:

- Establishing rates.
- Acquiring new customers.
- Retaining customers.
- Developing new product lines.
- Creating geographic exposure reports.
- Detecting fraudulent claims.
- Performing sophisticated campaign management.
- Estimating outstanding claim provisions.
- Assisting regulators understand the firm's rates and models.
- Coordinating actuarial and marketing departments.

Establishing Rates

An important problem in actuarial science concerns rate setting or the pricing of each policy. The goal is to set rates that reflect the risk level of the policyholder by establishing the "break even" rate (premium) for the policy. The lower the risk, the lower the rate.

Identify Risk Factors that Predict Profits, Claims, and Loses

The critical question in rate making is the following: "What are the risk factors or variables that are important for predicting the likelihood of claims and the size of a claim?" For example, in the automobile insurance industry, a significant, positive correlation exists between the likelihood of a claim and the policyholder's closeness to a large urban area. Actuaries might use this knowledge to specify automobile rates by postal codes for a given policyholder profile. As a result, a 30-year old male having one ticket in the past three years is likely to pay a higher rate if he lived and drove in a large urban area. Similarly, in the health insurance industry smokers typically pay higher health premiums. Some insurance companies make further distinctions between perceived health risks of certain behaviors such as the insured's use of tobacco products.

Although many risk factors that affect rates are obvious, subtle and non-intuitive relationships can exist among variables that are difficult if not impossible to identify without applying more sophisticated analyses. Modern data mining models can more accurately predict risk, therefore insurance companies can set rates more accurately, which in turn results in lower costs and greater profits.

Improve Predictive Accuracy - Segment Databases

To improve predictive accuracy, databases can be segmented into more homogeneous groups. Then the data of each group can be explored, analyzed, and modeled. Depending on the business question, segmentation can be done using variables associated with risk factors, profits, or behaviors. Segments based on these types of variables often provide sharp contrasts, which can be interpreted more easily. As a result, actuaries can more accurately predict the likelihood of a claim and the amount of the claim.

For example, one insurance company found that a segment of the 18- to 20-year old male drivers had a noticeably lower

accident rate than the entire group of 18- to 20-year old males. What variable did this subgroup share that could explain the difference? Investigation of the data revealed that the members of the lower risk subgroup drove cars that were significantly older than the average and that the drivers of the older cars spent time customizing their "vintage autos." As a result, members of the subgroup were likely to be more cautious driving their customized automobiles than others in their age group.

Acquiring New Customers

Another important business problem that is related to rate making is the acquisition of new customers. Although traditional approaches involve attempts to increase the customer base by simply expanding the efforts of the sales department, sales efforts that are guided by more quantitative data mining approaches can lead to more focused and more successful results.

Focusing Marketing Strategy to Reach Targeted Group

A traditional sales approach is to increase the number of policyholders by simply targeting those who meet certain policy constraints (illustrated in Figure 1). A drawback to this approach is that much of the marketing effort may yield little return. At some point, sales become more difficult and greater marketing budgets lead to lower and lower returns.

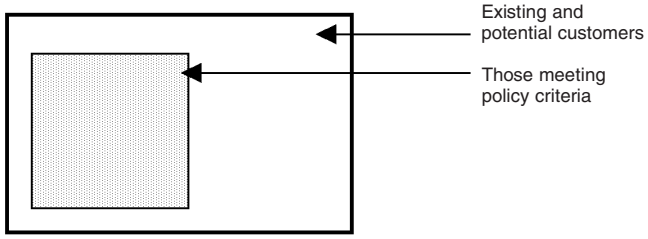

Figure 1: Marketing to All Those Meeting Policy Criteria

Increasing a Marketing Campaign's Return on Investment

In contrast to the traditional sales approach, data mining strategies enable analysts to refine the marketing focus. For example, the focus could be refined by maximizing the lifetime value of policyholders, that is, the profits expected from policyholders over an extended period of time. Thus as Figure 2 illustrates, the crucial marketing question becomes, "Who of those meeting the criteria are most likely to actually purchase a policy?"

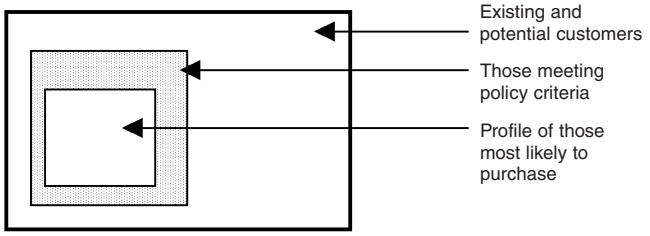

Figure 2 : Marketing to Those Most Likely to Purchase

Because only the segmented group of those likely to purchase is targeted, the return per unit of marketing effort is greater.

Can even better results be obtained? In other words, as more data are collected, can better models be developed and can marketing efforts be focused further? To sharpen the focus, analysts in the insurance industry can utilize advanced data mining techniques that combine segmentations to group (or profile) the high lifetime-value customer and produce predictive models to identify those in this group who are likely to respond.

For example, perhaps, the first group for the marketing campaign is made up of those who meet the policy criteria, are likely to purchase, and are likely to remain loyal by not switching to another company (as illustrated in Figure 3). Segmenting the universe of potential customers to focus on specific groups can make marketing campaigns more efficient and further increase the return per unit of marketing effort.

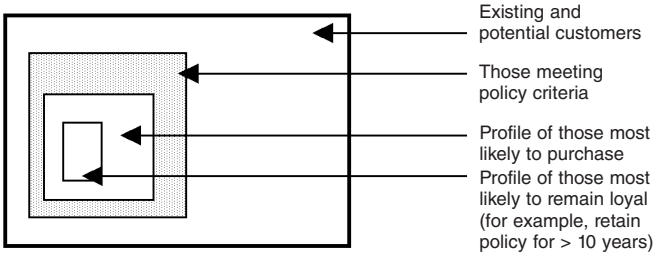

Figure 3 : Marketing to Those Most Likely to Retain Policies

Companies can increase response rates and profitability by first targeting those prospects that have characteristics similar to those of the high lifetime-value customers. Moreover, additional data mining work could be performed to identify the best time of day, the best season, and best media for marketing to the targeted group.

Retaining Customers

The problem of retaining current customers is related to the problem of focusing the marketing campaigns to those most likely to purchase.

Offer Bundled Packages

Experience shows that a customer holding two policies with the same company is much more likely to renew than is a customer holding a single policy. Similarly, a customer holding three policies is less likely to switch than a customer holding less than three. By offering "quantity discounts" and selling bundled packages to customers, such as home and auto policies, a firm adds value and thereby increases customer loyalty, reducing the likelihood the customer will switch to a rival firm.

Analyze at the Customer Level

Successfully retaining customers requires analyzing data at the most appropriate level, the customer level, instead of across aggregated collections of customers. The insurance industry has long been a leader in analyzing profitability at the customer level by answering data-driven questions such as, "What policies are customers most likely to purchased as bundles?" Data warehousing, improved computer technology, and data mining are the enabling technologies to answer such questions.

Using a data mining technique called *association analysis,* insurance firms can more accurately select which policies and services to offer to which customers. With this technique insurance companies can:

- Segment the customer database to create customer profiles.
- Conduct rate and claim analyses on a single customer segment for a single product. For example, companies can perform an in-depth analysis of a potential new product for a particular customer segment.
- Analyze customer segments for multiple products using group processing and multiple target variables. For example, how profitable are bundles of policies (auto, home, and life) for certain customer segments of interest?
- Perform sequential (over time) market basket analyses on customer segments. For example, what percentage of new policyholders of auto insurance also purchase a life insurance policy within five years?

Aim Retention Campaigns at Those Most Likely to Switch Firms

Database segmentation and more advanced modeling techniques enable analysts to more accurately choose whom to target for retention campaigns. Current policyholders that are likely to switch can be identified through predictive modeling. A logistic regression model is a traditional approach, and those policyholders who have larger predicted probabilities of switching are the target group.

Identifying the target group may be improved by modeling the behavior of policyholders. By including nonlinear terms and more interaction terms, neural network models can generate more accurate data on the probability of policyholders switching.

Additionally, decision tree models may provide more accurate identification by dividing (segmenting) the policyholders into more homogeneous groups. Greater accuracy in identifying the target group can reduce costs and has the potential for greatly improving the results of a retention campaign.

Developing New Product Lines

Markets change over time, and so do the products sought by consumers. It is critical that the firm identifies and monitors the changing needs of the insurance prospect and adjusts the offered policies so that insurance agents can sell them. In turn, insurance firms can continue to realize a profit.

Identifying Profitable Customer Profiles

Insurance firms can increase profitability by identifying the most lucrative customer segments and then prioritize marketing campaigns accordingly. For the insurance industry, there may be no such thing as an unacceptable consumer profile. For example, Lloyd's of London has developed a reputation for insuring virtually anything for anyone with the provision that Lloyd's brokers can set the rates. In one unusual example, food critic and gourmet, Egon Ronay, insured his taste buds through Lloyd's for £250,000 (Lloyd's of London 1999).

Problems with profitability can occur if firms do not offer the "right" policy or the "right" rate to the "right" customer segment at the "right" time. For example, the most profitable customer segment may be the more risky customers, which may command higher rates. Or perhaps, it is the customer segment that is more likely to purchase a bundle of insurance policies. Another possibility is that the most profitable segment is the lifetime customer segment that over time adds more product and policies. Each of these different customer segments requires different analyses to measure the expected profitability.

Adjusting to Market Changes

To adjust to market changes, insurance companies need to know what types of new policies will be profitable. For example, actuaries conceivably could set rates for a life insurance policy for private citizens traveling on the Space Shuttle. The marketing department could develop a customer profile, but would the expected sales revenue justify the effort? Perhaps not, because the job of selling such a policy to a large enough segment of the population may be nearly impossible.

Clearly, the question of what products and policies to offer is closely related to the problems of acquisition, cross-selling, and the retention of customers. Data mining techniques can enable insurance companies to use the answer to one business question as input to another question. Analyses will seek to answer various questions related to customer segments, the introduction of new products, and profitability. For example, the analysts might ask:

- "After performing various analyses on potential products for various customer segments, which products are most profitable?"
- "Which have the biggest market?"
- "Which appear to be easiest to market?"

Insurance firms can now utilize all of their available information to better develop new products and marketing campaigns.

Prioritizing the Introduction of New Products

Once new products are identified as potentially the most profitable, they can be prioritized for introduction to the market. Depending upon the goals of the firm, the new products can be prioritized based on expected profit, expected number of new customers, and/or the expected speed of acceptance. Data mining technology can be used to identify consumer groups, to model their behaviors, to calculate expected profits from marketing campaigns, and then to compare potential strategies for action.

Creating Geographic Exposure Reports

Insurance firms also can augment their business and demographic databases with socio-geographic data, which is also referred to as *spatial attribute data* or *latitude/longitude data.* The reason for augmenting existing data with socio-geographic data is that the social profile, including geographic location of potential customers, can be an important risk factor in the rate making model. For example, driving conditions and the likelihood of accidents and auto thefts vary across geographic regions. Differences in risk factors indicate differences in likelihood of claims, expected claim amounts, and, ultimately, in rates.

Including purely geographic data in a *data warehouse*[6] enables the insurance firm to create digital maps. Business analysts can overlay (or "map") the data then assess and monitor exposure by geographic region. Such data processing and data mapping capabilities are not merely for the purpose of plotting the geographic data for display. Instead, the data also can be included in rate making and other analytical models. If an area of over-exposure is identified, then the risk can be mitigated, possibly by rate adjustment or by re-insurance.

Detecting Fraudulent Claims

Obviously, fraudulent claims are an ever-present problem for insurance firms, and techniques for identifying and mitigating fraud are critical for the long-term success of insurance firms. Quite often, successful fraud detection analyses such as those from a data mining project can provide a very high return on investment.

Better Fraud Detection Results in Lower Cost of Claims

Fraud is a thriving industry. For example, the United States spends more than $1 trillion each year on health care. Of that amount, over $100 billion is estimated to be lost annually to fraud (Goldmann 2000). Add in $20 billion in property-casualty fraud, and an "insurance fraud industry" would be in the top 25 of the Fortune 500. The shear magnitude of the fraud problem implies that the firms that are better able to detect fraudulent claims also are in a position to offer more competitively priced products, to reduce costs, and to maintain long-term profitability.

Just Random Chance or Is There a Pattern?

Fraudulent claims are typically not the biggest claims, because perpetrators are well aware that the big claims are scrutinized more rigorously than are average claims. Perpetrators of fraud use more subtle approaches. As a result, in searching for fraudulent claims, analysts must look for unusual associations, anomalies, or outlying patterns in the data. Specific analytical techniques adept at finding such subtleties are market basket analysis, cluster analysis, and predictive modeling.

By comparing the expected with the actual, large deviations can be found that can be more thoroughly investigated. For example, Empire Blue Cross and Blue Shield of New York used a database to compare the number of bronchoscopies reportedly performed by otolaryngologists. Newly reported numbers were compared to the existing numbers gathered from the insurer's 4.1 million members. The number of bronchoscopies from one otolaryngologist was well above average, and further research confirmed that the physician was submitting false bills. As a result, the doctor received a ten-month jail sentence, was forced to surrender his medical license, and was required to pay Empire $586,000 in restitution. Empire estimates it saved $4 million in 1997 alone by using data mining for fraud detection (Rao 1998).

[6]Accessing, aggregating, and transforming data are primary functions of data warehousing. For more information on data warehousing, see the "Recommended Reading" section of this paper.

Providing Reinsurance

Reinsurance occurs when part of the insurer's risk is assumed by other companies in return for part of the premium fee paid by the insured. By spreading the risk, reinsurance enables an individual company to take on clients whose coverage would be too great a burden for one insurer to carry alone. Although reinsurance reduces the risk of large claim amounts for an insurance firm, reinsurance also reduces the revenue of the firm. The goal, then, is to find the appropriate level of reinsurance. Too little reinsurance and the firm is exposed to too much risk. Too much reinsurance and the firm gives away revenue for little in return.

In addition, part of the risk/revenue equation of reinsurance is the need for the firm seeking reinsurance to be aware of the credit risk exposure; that is, the counter-party to the contract—the firm providing the reinsurance—should be able to perform as contracted on claims.

Limitations of Traditional Methods of Analysis

Using traditional methods can lead to policies being reinsured when in fact their risk of a claim during the reinsurance period is minimal. For example, for most classes of general insurance, the distribution of claim amount is markedly skew, having a long tail on the right. If an insurer receives a large number of policies for a particular book of business, the total claim payment amount might be expected to be approximately normal since it is the sum of a large number of individual claims.

If, on the other hand, the available data are limited, then the tail probabilities of the loss distribution play a far more significant role. In such situations the confidence level associated with the insurer's predictions and estimates are reduced, and this diminished level of confidence represents the primary motivation for reinsurance.

In such situations, an insurer will often obtain reinsurance on potential losses exceeding some established amount.

Advantages of Data Mining for Reinsurance

Data mining technology is commonly used for segmentation clarity. In the case of reinsurance, a group of paid claims would be used to model the expected claims experience of another group of policies. With more granular segmentation, analysts can expect higher levels of confidence in the model's outcome. The selection of policies for reinsurance can be based upon the model of experienced risk and not just the traditional "long tail of a book of business."

Exploration of Claims Distribution

For a typical book of business, an insurer might adopt the log normal distribution as the underlying loss distribution. But for situations in which the data are not warehoused properly or for situations in which a large variance and extreme claim amounts are common, the Pareto distribution might represent a more appropriate distribution due to the heavier tails or larger tail probabilities. Indeed, for an insurer or reinsurer to use the log normal distribution for rating when the Pareto distribution is the true distribution would likely prove to be an expensive blunder, which illustrates the importance of having the right tool to identify and estimate the underlying loss distribution.

Role of a Modeler

The modeler can examine the body of claims history looking for the most granular segmentation that will enable differentiation of risk across a book of business. Once established, that experience can be used to identify the high-risk policies and earmark them as reinsurance candidates.

Greater accuracy in modeling indicates better predictions. Firms can avoid sharing premiums unnecessarily by reinsuring only when it is necessary to reinsure. Firms can avoid losses from not reinsuring when, in fact, the policies were risky and should have been reinsured. Avoiding both problems increases profits by reducing costs.

When the techniques of predictive modeling have been applied, the underlying loss distribution is estimated for the most at-risk policies. From there, the question of what to reinsure can be answered. The policies that are reinsured are the ones most likely to result in a claim, yielding a higher level of risk avoidance for the premium dollars given to other insurers.

Performing Sophisticated Campaign Management

Developing a customer relationship has a long-standing tradition in business. Small firms and many retailers are able to relate to their customers individually. However, as organizations grow larger, marketing departments often begin to think in terms of product development instead of customer relationship development and maintenance. It is not unusual for the sales and marketing units to focus on how fast the firm can bring a mass-appeal product to market rather than how they might better serve the needs of the individual customer.

Ultimately, the difficulty is that as markets become saturated, the effectiveness of mass marketing slows or halts completely. Fortunately, advanced data mining technology enables insurance companies to return their marketing focus to the aspects of individual customer loyalty. Creative, data-driven, scientific marketing strategies are now paving the way back to the customer relationship management of simpler, efficient economies, while on a much grander, more comprehensive scale.

A Customer-Centric Focus

Many leading insurance companies are making an effort to move away from the product-oriented architectures of the past and toward a customer-centric focus to better serve their customers. Data mining technology can be utilized to better understand customers needs and desires. Analysis of marketing campaigns provides in-depth feedback and serves as the foundation of future campaign development.

Marketing – Another Frontier of Automation

To know your customers in detail—their needs, desires, and responses—is now possible through the power of data mining methodologies and supporting computer technologies. Campaign management solutions consist of graphical tools that enable marketers to analyze customer data, analyze previous marketing campaigns, design new campaigns, assess the new campaigns prior to implementation, monitor the campaigns as they are presented, and evaluate the effectiveness of the completed campaigns.

Customer-centric marketing is accomplished by integrating data mining and campaign management. The integration sets up a cyclical relationship in that data mining analysts can develop and test customer behavior as required by marketers. Then marketers can use the models to predict customer behavior. By assessing predicted customer behavior, marketing professionals can further refine the marketing campaigns.

An Overall Solution

Campaign management tools should support the entire direct marketing lifecycle—Analysis, Planning, Execution, and Evaluation—not just one or two phases of the lifecycle. Unfortunately, many campaign management products contain separate tools embedded in the system for querying, updating, analyzing, and extracting data. Often, these systems require significant customizations to accommodate the promotional campaign strategies of individual enterprises. The best solution is one that integrates the major functional areas of data access, data warehousing, data mining, as well as campaign management to provide an efficient, cohesive whole.[7] If the tools are not explicitly designed to work together, problems of portability and throughput can arise. In most cases, a workable piecemeal solution may be far from a good overall solution.

Effective Campaign Management

The integration of data access, data warehousing, data mining, and campaign management technologies enables marketing professionals to utilize pre-established data mining models within the context of their campaign management system. Marketers are able to select from a list of such models and apply the model to selected target subsets identified using the campaign management system. The scoring code is typically executed on the selected subset within the data mining product, and the *scored file*[8] is then returned to the marketing professional, enabling marketing to further refine their target-marketing campaigns. This form of integration is commonly referred to as "dynamic scoring" because it reflects the real-time execution of the scoring code.

Estimating Outstanding Claims Provision

The settlement of claims is often subject to delay. For example, in liability insurance, the severity (magnitude) of a claim may not be known until years after the claim is reported. In cases of employer's liability, there is even a delay involved in reporting the claim. Such delays may result in non-normal distribution of claims, specifically, skewed distributions and long-tailed distribution across time and across classes of business.

Still the business of the firm must continue, and an estimate of the claim severity is often used until the actual value of the settled claim is available. The estimate can depend on the following:

- Severity of the claim.
- Likely amount of time before settlement.
- Effects of financial variables such as inflation and interest rates.
- Effects of changing social mores. For example, the tobacco industry has been greatly effected by the changing views toward smoking.

Predicting Actual Settlement Values

A claims provision, which is necessary for continued business operations, is developed by estimating insurance claims. The accuracy of the claims provision is important, because the funds set aside for paying claims typically cannot be invested in long-term, higher yielding assets. If the claims provision is too small, the firm may experience financial problems. Conversely, if the claims provision is too large, the firm may become unprofitable. Thus, the more accurate the estimation, the more opportunity for profit. The analysis of the distribution of claims across customers, across geography, and across time lead to better estimates of the claims provision.

Data mining technology can be utilized to establish the distribution of claims and the pattern of past claim run-offs. The data are analyzed and modeled, and when a predictive model is developed, the current outstanding claims are

[7] SAS software integrates well with campaign management systems as the data mining and data warehousing solution. SAS software typically provides analytical tools for assessing the success of promotional campaigns. In addition, SAS partners with several campaign management software vendors to offer the most sophisticated campaign management solutions on the market.

[8] *Scored file* refers to a data set in which values for a target variable have been predicted by a model. Scored data sets consist of a set of posterior probabilities for each level of a (non-interval level) variable.

scored. Specifically, the model parameters and the claims data are used to predict the magnitude of the actual settlement value of the outstanding claims. This estimate of the actual settlement value can be used to develop a claims provision.

Updating the Predictive Model

The estimate of the claims provision generated from a predictive model is based on the assumption that the future will be much like the past. If the model is not updated, then over time, the assumption becomes that the future will be much like the distant past. However, as more data become available, the predictive data mining model can be updated, and the assumption becomes that the future will be much like the recent past.

Data mining technology enables insurance analysts to compare models and to assess them based on their performance. When the newly updated model outperforms the old model, it is time to switch to the new model. Given the new technologies, analysts can now monitor predictive models and update as needed.

Assisting Regulators

Rate changes require review from regulators. However, the statistical analyses employed by quantitative experts to obtain those rates may be based on highly non-linear models and other arcane analytics. Explaining those models and results to some audiences can be a challenge. Fortunately, through the use of graphical representations that are accompanied by the human language rules and express the analytics in text form, highly sophisticated analyses can be made simpler and easier to understand by general audiences. For example, by augmenting graphical representations such as tree models with logic rules expressed as readable text, analysts can assist regulators and others in the insurance industry to see the statistical bases for rate changes and other practices.

Coordinating Actuarial and Marketing Departments

Coordinating the efforts of the actuarial and marketing departments can enhance revenues and lead to greater profits. Traditionally, the actuarial or rate-setting problem has been separated from the targeted-marketing problem, and the actuarial and marketing departments traditionally operated relatively independently of each other. In fact, they are interdependent; the actions of one affect the other.

Recognition of this interdependency by management can lead to cooperation. Coordination of efforts can be achieved by strategic use of data mining techniques. The marketing department can utilize the findings of the actuarial department, and the results of marketing campaigns can become data for additional actuarial research.

Implementing Data Mining Projects

Much has been written about the best way to implement data mining projects. Many authoritative books and shorter works that are written by IT experts cover the topic in detail.[9] One message found in many of these works is that implementing a data mining project must take into account real-world, practical challenges. A data-centric approach is especially effective and can be divided into the following functional areas:

- Access the data.
- Warehouse the data.
- Analyze the data.
- Report the results of the analyses.
- Exploit the results for business advantage.

Accessing the Data

Reliable, accurate data is a prerequisite for data mining. A complete data access strategy should include the following key elements:

- Access to any or all types of data sources.
- Access to data sources regardless of the platform on which they reside.
- Preservation of the source data through the use of security routines.
- An easy-to-use, consistent GUI that, while not requiring an extensive knowledge of each data type, does provide the flexibility to meet the specific needs.
- Integration with the existing technology rather than access routines that require retooling of hardware and/or software or extensive, additional learning by users.

A properly designed and implemented data warehouse can help accomplish these key elements of a data access strategy.

[9] See "Recommended Reading" section in this paper.

Warehousing the Data

A data warehouse enables insurance industry researchers to easily access data that might be stored in various *data tables*[10] across a wide variety of platforms. Through data warehouses, research analysts can merge and aggregate data into subject areas. However, prior to analysis, data that contain errors, missing values, or other problems needs to be *cleaned.*[11] One approach to cleaning data is to simply delete cases that contain missing values. However, the results are biased because the deleted data may have otherwise been an important part of various relationships. Major data cleaning tasks such as making variable names consistent, imputing missing values, identifying errors, correcting errors, and detecting outliers can be performed relatively easily using data mining technology.

Good data warehousing tools can vastly improve the productivity of the data mining team. Important results are obtained faster and often at much lower cost. SAS/Warehouse Administrator™ software provides a visual environment for managing data warehouses. Using the Warehouse Administrator, analysts specify metadata, which defines data sources, data stores, code libraries, and other warehouse resources. The Warehouse Administrator then uses this metadata to generate or retrieve the code, which extracts, transforms, and loads the data into the data warehouse and *data marts.*[12]

Analyzing Data Using the SEMMA Methodology

Even after data are merged and aggregated into subject areas, highly complex relationships often are not easily seen by visually inspecting the raw data or by applying basic statistical analyses. Actual patterns may appear random or go undetected. Additionally, linear models may poorly characterize nonlinear relationships. Modern data mining technology can overcome these limitations by employing approaches such as the following:

- Sophisticated GUI data exploration and plotting tools to better display relationships among variables.
- Variable selection methodologies to identify the most important variables to include in models.
- Advanced modeling techniques such as linear models with interactions.
- Nonlinear neural networks, and tree models.

[10]The terms *data table* and *data set* are synonymous.

[11]The terms *cleaned, transformed,* and *scrubbed* are synonymous. *Scrubbed data* and similar terms refer to data that are prepared for analysis by correcting errors such as missing values, inconsistent variable names, and inconsequential outliers before being analyzed. For more information on preparing data for mining see the "Recommended Reading" section in this paper.

[12]*Data marts* are tables that provide easy access to a subset of the data in a data warehouse. Typically, data marts store information of interest to a particular department or individual. They can also be used to store the results of ad hoc queries or cross-subject analyses.

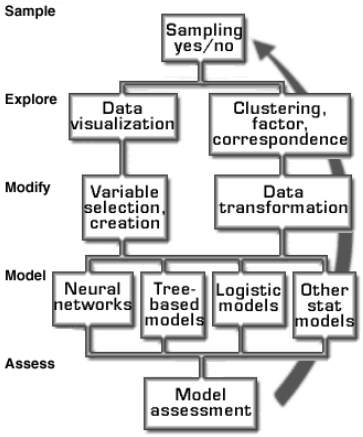

Figure 4 : The SEMMA Analysis Cycle

- Assessment techniques to assist analysts in selecting the best performing model based on profit and loss criteria.

Once accessed, the data can be explored using GUIs that utilize sophisticated data mining algorithms. For example, subsetting data can reveal important relationships for marketing campaigns. Disaggregating along region and firm might reveal costly anomalies of operations. Drilling down into the data might reveal missed profit opportunities.

The actual data analyses for data mining projects involve selecting, exploring, and modeling large amounts of data to uncover hidden information that can then be used for business advantage. The answer to one question often leads to new and more specific questions. Hence, data mining is an iterative process, and the data mining methodology should incorporate this iterative, exploratory approach.

To provide a predictable yet flexible path for the data mining analysis to follow, SAS has developed a data mining analysis cycle known by the acronym SEMMA. This acronym stands for the five steps of an analysis that are ordinarily a part of a data mining project. Those five steps are:

- Sample.
- Explore.
- Modify.
- Model.
- Assess.

Figure 4 illustrates the tasks of a data mining project and maps those tasks to the five steps of the SEMMA methodology.

Beginning with a representative sample, the SEMMA analysis cycle guides analysts through the process of exploring data using visual and statistical techniques, transforming data to uncover the most significant predictive variables, modeling the variables to predict outcomes, and assessing the model

by testing it with new data. Thus, the SEMMA analysis cycle is a modern extension of the scientific method.

Sample

The first step in a data mining analysis methodology is to create one or more data tables by sampling data from the data warehouse. The samples should be big enough to contain the significant information, yet small enough to process quickly. This approach enables the most cost-effective performance by using a reliable, statistically representative sample of the entire database. Mining a representative sample instead of the whole volume drastically reduces the processing time required to get crucial business information.

If general patterns appear in the data as a whole, these will be traceable in a representative sample. If a niche is so tiny that it is not represented in a sample and yet so important that it influences the big picture, it can be discovered using summary methods.[13]

Explore

After sampling the data, the next step is to explore them visually or numerically for inherent trends or groupings. Exploration helps refine the discovery process. If visual exploration does not reveal clear trends, analysts can explore the data through statistical techniques including factor analysis, correspondence analysis, and clustering. For example, new parents are often more acutely aware of their need for life insurance but may be seeking the most insurance for the least amount of money. This group may be more likely to respond to direct mailings for term insurance.

Modify

Modifying the data refers to creating, selecting, and transforming one or more variables to focus the model selection process in a particular direction or to augment the data for clarity or consistence.

Based on the discoveries in the exploration phase, analysts may need to modify the data to include information such as the grouping of customers and significant subgroups, or to introduce new variables such as a ratio obtained by comparing two previously defined variables. Analysts may also need to look for outliers and reduce the number of variables to narrow them down to the most significant ones. In addition, because data mining is a dynamic, iterative process, there often is a need to modify data when the previously mined data change in some way.

Model

Creating a data model involves using the data mining software to search automatically for a combination of data that reliably predicts a desired outcome.

After the data have been accessed and modified, analysts can use data modeling techniques to construct models that explain patterns in the data. Modeling techniques in data mining include neural networks, tree-based models, logistic models, and other statistical models such as time series analysis and survival analysis.[14]

Each type of model has particular strengths, and is appropriate within specific data mining situations depending on the data. For example, neural networks are good at combining information from many predictors without over-fitting and therefore work well when many of the predictors are partially redundant.[15]

Assess

The next step in data mining is to assess the model to determine how well it performs. A common means of assessing a model is to apply it to a portion of the data that was set aside during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model.

Similarly, analysts can test the model against known data. For example, if one knows which customers in a file had high retention rates and the model predicts retention, analysts can check to see whether the model selects these customers accurately. In addition, practical applications of the model, such as partial mailings in a direct mail campaign, help prove its validity.

Iteration

Although assessing the data models is the last step in the SEMMA methodology, assessing the effectiveness of data models is often not the final step in an actual implementation of SEMMA. Because SEMMA is a cycle, the internal steps are often performed iteratively as needed within a particular data mining project.

Reporting the Results

Enterprise reporting capabilities are essential to making data useful by enabling users to create and publish reports from data warehouses and other information sources. Key features of a reporting system should include:

- Complete system integration.
- Simplified warehouse reporting.
- Ease of use through graphical user interfaces.

[13]For more information about sampling, see SAS Institute Inc., "Data Mining and the Case for Sampling," 1998.

[14]See "Recommended Reading" section in this paper.

[15]For more information on neural networks, see Sarle 1997.

• Rapid distribution of reports.
• Web enablement.

Modern reporting tools such as those found in SAS® Enterprise Miner™ enable business users to create, publish and print richly formatted reports from information stored in their data warehouse. Through easy-to-use interfaces, users have the ability to create graphs, tables, charts and text within a single report from their desktops.

Exploiting the Results for Business Advantage

The new information obtained from data mining can be incorporated into an executive information or online analytical processing and reporting system, and then disseminated as needed throughout the organization. The firm's decision makers can use the data mining results to answer important business-related questions such as, "How can we increase the ROI of our marketing campaigns?" for strategic planning and action. By exploiting data mining results in this manner, firms can better prepare for long-term growth and improve their opportunities for long-term prosperity.

Summary

The key to gaining a competitive advantage in the insurance industry is found in recognizing that customer databases, if properly managed, analyzed, and exploited, are unique, valuable corporate assets. Insurance firms can unlock the intelligence contained in their customer databases through modern data mining technology. Data mining uses predictive modeling, database segmentation, market basket analysis, and combinations thereof to more quickly answer crucial business questions with greater accuracy. New products can be developed and marketing strategies can be implemented enabling the insurance firm to transform a wealth of information into a wealth of predictability, stability, and profits.

References

Bentson, George J. (1990), *The Separation of Commercial and Investment Banking. The Glass-Steagall Act Revisited and Reconsidered,* Oxford: Oxford University Press.

Clinton, William J. (1999), "Statement by the President," November 12, 1999, http://www.pub.whitehouse.gov/uri-res/I2R?urn:pdi://oma.eop.gov.us/1999/11/15/6.text.1 (accessed 21 Feb. 2000).

Francis, David (1998), "Throwing Stones at Glass-Steagall," http://www.csmonitor.com/durable/1998/05/11/p58s1.htm (accessed 23 Feb. 2000).

Goldmann, Peter (2000), "Malcolm Sparrow: Fight Fraud by Understanding Its Complexities," http://www.fraudreport.com/article.cfm?id=52&month=04&year=2000 (accessed 29 Mar. 2000).

House Committee on Banking and Financial Services (1997), "Financial Services Competition Act of 1997," http://www.house.gov/banking/hr10rep.htm (accessed 24 Feb. 2000).

Lloyd's of London (1999), "Unusual Risks," http://www.lloyd's.com/heritage/unusualtext.htm (accessed 9 Dec. 1999).

Pellegrini, Frank (1998), "The New Citigroup: One-Stop Shopping," http://www.time.com/time/daily/0,2960,10851,00.html (accessed 21 Feb. 2000).

Rao, Srikumar S. (1998), "Diaper-Beer Syndrome," Forbes, April 6, 1998. http://www.forbes.com/forbes/98/0406/6107128a.htm (accessed 9 Dec. 1999).

Sarle, W.S., ed. (1997), "Neural Network FAQ," periodic posting to the Usenet newsgroup comp.ai.neural-nets, ftp://ftp.sas.com/ pub/neural/FAQ.html (accessed 9 Dec. 1999).

SAS Institute Inc. (1998), SAS Institute Best Practices Paper, "Data Mining and the Case for Sampling," Cary, N.C.: SAS Institute Inc.

Thomas, Legislative Information on the Internet (1999), "S.900: Gramm-Leach-Bliley Act (Enrolled Bill (Sent to President))." http://thomas.loc.gov/cgibin/query/D?c106:4:./ temp/~c106VzxndU:: (accessed 28 Feb. 2000).

Watts, J.C., Chairman, House Republican Conference, "Financial Services Modernization: H.R. 10, The Financial Services Act of 1999," June 29, 1999, http://143.231.67.32/IssueFocus/IssueBriefs/IB106/19990629fin.htm (accessed 22 Feb. 2000).

Wettlaufer, Dale, "Universal Banking," April 6, 1998, http://www.scribe.fool.com/LunchNews/1998/LunchNews980406.htm (accessed 21 Feb. 2000).

Recommended Reading

Insurance

Hogg, R.V. and Klugman, S.A. (1984), Loss Distributions, New York: John Wiley & Sons, Inc.

Hossack, I.B. et al., eds. (1999), Introductory Statistics with Applications in General Insurance, Cambridge University Press.

Data Mining

Berry, M.J.A. and Linoff, G. (1997), Data Mining Techniques, New York: John Wiley & Sons, Inc.

SAS Institute Inc. (1997), SAS Institute White Paper, "Business Intelligence Systems and Data Mining," Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1998), A SAS Best Practices Paper, "Data Mining and the Case for Sampling: Solving Business Problems Using SAS® Enterprise Miner™ Software," Cary, NC: SAS Institute Inc.

Weiss, S.M. and Indurkhya, N. (1998), Predictive Data Mining: A Practical Guide, San Francisco, California: Morgan Kaufmann Publishers, Inc.

Data Warehousing

Inmon, W.H. (1993), Building the Data Warehouse, New York: John Wiley & Sons, Inc.

SAS Institute Inc. (1995), SAS Institute White Paper, "Building a SAS® Data Warehouse," Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1996), SAS Institute White Paper, "SAS' Rapid Warehousing Methodology," Cary, NC: SAS Institute Inc.

Singh, H. (1998), Data Warehousing Concepts, Technologies, Implementations, and Management, Upper Saddle River, New Jersey: Prentice-Hall, Inc.

SAS Enterprise Miner Software

SAS Institute Inc. (2000), SAS Institute White Paper, "Finding the Solution to Data Mining: A Map of the Features and Components of SAS® Enterprise Miner™ Software Version 3," Cary, NC: SAS Institute Inc.

Statistics

Hays, W.L. (1981), Statistics. New York: Holt, Rinehart and Winston.

Hildebrand, D.K. and Ott, R.L. (1996), Basic Statistical Ideas for Managers, New York: Duxbury Press.

SAS World Headquarters
SAS Campus Drive
Cary, NC 27513 USA
Tel: (919) 677 8000
Fax: (919) 677 4444
U.S. & Canada sales:
(800) 727 0025

SAS Europe, Middle East & Africa
PO Box 10 53 40
Neuenheimer Landstr. 28-30
D-69043 Heidelberg, Germany
Tel: (49) 6221 4160
Fax: (49) 6221 474850

SAS Asia/Pacific & Latin America
SAS Campus Drive
Cary, NC 27513 USA
Tel: (919) 677 8000
Fax: (919) 677 8144

www.sas.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies. Copyright © 2000, SAS Institute Inc. All rights reserved.
38113US.1000