

## Mining the Most from Credit and Non-Credit Data

**O**VER THE PAST SEVERAL YEARS, it has become common for personal-lines insurers to use credit scores to more accurately price and underwrite auto and homeowners insurance. This novel use of credit scoring is one of the most successful innovations in personal lines insurance in recent decades. According to a recent study by [Conning & Company](#), approximately 90 percent of the top 100 personal lines insurance companies use credit information for pricing, underwriting, or both. Approximately half of these companies adopted credit scoring in 1998 or after.

What started out as a strategy for innovative insurance companies to gain market share has therefore become a matter of survival in today's more competitive marketplace. Today, an insurer that doesn't use credit is likely being selected against by its competitors.

In this way, the very success of credit scoring has diminished its ability to give insurers a competitive edge in the marketplace. Insurers that want to regain or maintain their competitive edge are therefore led to ask: What is the next logical step beyond credit scoring?

We have performed a number of large-scale insurance data mining projects that have involved both credit and non-credit information. The results of these projects are somewhat surprising. Our experience leads us to believe that, no matter how sophisticated an insurer's pricing or underwriting system, the use of credit usually increases the actuarial fairness of insurance rates.

This particular result isn't what's surprising. The great majority of insurers wouldn't spend large amounts of time and money to implement a mediocre pricing/underwriting system. What is surprising is that highly effective scoring models can be built without using credit information at all. The key is to custom-design pre-

dictive models that take maximum advantage of a company's internal data sources, as well as external, non-credit sources of information. As we will discuss, such "non-credit" models have several advantages over commonly used credit scores.

By the same token, it's possible to use both credit and non-credit to build "mixed" scoring

models. Our experience leads us to believe that nearly any pricing/underwriting system or scoring model can be enhanced through the use of credit information. But there's no need for insurers to restrict themselves to credit in-

formation in their scoring models. The key to building the most effective scoring models possible is to take full advantage of all available sources of information.

Credit scoring is best viewed from a "data mining" perspective. It should become clear in the course of this article that credit scores are really a special case of the infinite variety of scoring algorithms made possible

by recent advances in computing power and data mining techniques. At the same time, the data mining perspective allows us to address a question that's often raised in the regulatory debates over credit: Does credit information really help explain insurance losses, even after all other available information has been taken into account?

### Unpacking the Black Box

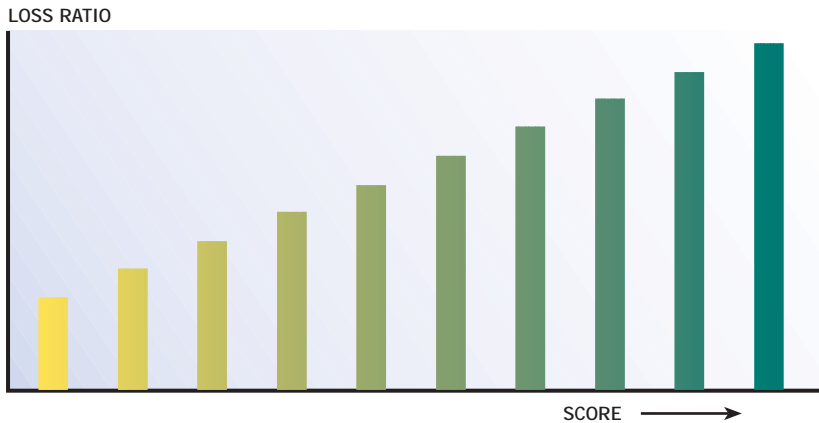
The raw building blocks—the "atoms"—from which credit scores are built are individuals' raw credit reports, col-



**JAMES GUSZCZA** IS A MANAGER, **CHENG-SHENG PETER WU** IS A DIRECTOR IN THE ADVANCED QUANTITATIVE SERVICES PRACTICE OF DELOITTE & TOUCHE'S ACTUARIAL AND INSURANCE CONSULTING GROUP. THEY WORK IN DELOITTE'S LOS ANGELES OFFICE.

FIGURE 1

## The Lift Curve



To create a lift curve:

- ▶ Policies are sorted from lowest to highest by model score
- ▶ Policies are then grouped into 10 equal-sized deciles
- ▶ "Lift" is measured as the resulting loss-ratio differential

lected and maintained by such credit bureaus as Equifax, Trans Union, and Experian. Credit reports contain a wealth of information that can be grouped into four classifications:

- ▶ General information
- ▶ Trade line information
- ▶ Inquiries
- ▶ Public records and collections

The raw fields on these reports can be aggregated in various ways to create a plethora of predictive variables. Examples of credit predictive variables include:

- ▶ Number of trades
- ▶ Months since oldest trade
- ▶ Amount past due
- ▶ Number of trades 60 or more days past due
- ▶ Ratio of amounts currently due to available credit
- ▶ Number of inquiries
- ▶ Number of collections
- ▶ Number of lawsuits

Many of these variables are individually predictive of insurance profitability. The idea of credit scoring is to take a collection of credit variables and derive from them a composite variable—a score—that is more predictive of insurance profitability than any of the credit variables taken alone. A typical credit score used

in personal lines insurance today might be composed of 10 to 30 individual credit variables.

Various kinds of statistical techniques can be used to build credit scores from raw credit variables. The most widely understood such technique is multiple regression analysis. Multiple regression can be enhanced or replaced by such advanced techniques as principal components analysis, generalized linear models (GLMs), clustering, classification and regression trees (CART), multiple adaptive regression splines (MARS), and neural networks.

The result of the statistical analysis is an equation or scoring engine. The scoring engine takes a collection of credit variables and produces a number that estimates the policy's future loss ratio. This number can be calibrated in various ways to produce a final credit score. For example, the commonly used FICO credit score by Fair, Isaac, the analytical consulting firm, has a range of 500 to 1000.

The business power of the scoring engine is measured by a lift curve. To create a lift curve, all the policies in the analysis database are sorted from best to worst, based on their credit score. Next, the policies are divided into 10 equal-sized

groups, called deciles. (We could choose five, 10, 20, or even 100 groups depending on the desired degree of resolution.) The "best" decile contains the 10 percent of the policies with the best credit score. For example, the scores in the best decile might range from 750 to 850. Similarly, the "worst" decile contains the 10 percent of the policies with scores ranging from, say, 500 to 578.

The predictive power of the model is measured by computing the overall loss-ratio relativity of each decile. For example, the best decile produced by a given credit model might have a loss ratio 40 percent lower than average; the worst might have a loss ratio 60 percent higher than average. This difference is the model's "lift."

Commonly used business actions based on this lift analysis would be to place policies into different price tiers; target new business growth for the better deciles; not renew policies in the worst decile; and raise the rates of the policies in the second-to-worst decile. These are just examples. A well-constructed scoring model is, in a sense, the ultimate actuarial precision tool: It can help sort profitable from unprofitable policies, even within traditional rating and underwriting categories.

Aside from the specifics of the predictive variables, nothing about this process is unique to credit information. Each step—variable selection, the use of a variety of statistical techniques, lift curve analysis, and model calibration—is entirely general and can be applied to any set of predictive variables. In fact, we regularly employ this process to build models using sets of non-credit variables, as well as sets containing both credit and non-credit variables. From the data mining perspective, credit scores are just the beginning.

Although many insurance industry credit scores are available, competitive pressures have led many insurers to develop their own proprietary, customized scoring engines. These same competitive pressures, as well as growing uncertainty surrounding the regulation of credit,

are motivating insurers to look beyond credit to develop next-generation predictive models. Before turning to this issue, we will discuss an issue that arises in regulatory debates over credit.

### Our Large-Scale Data Mining Experience

Skeptics of credit scoring have questioned whether credit information is truly predictive of insurance losses. Might it be possible that credit is little more than a proxy for other kinds of rating and underwriting information available to insurers? For example, the oft-observed correlation of poor credit with high insurance losses might occur partially because youthful drivers tend to have poor credit. If this were the case, the high insurance losses would be partially “explained” by driver age, not credit.

We’re aware of only two publicly available studies relating to this debate. The first was published by Tillinghast, and was associated with the 1997 National Association of Insurance Commissioners credit study. Tillinghast analyzed the loss ratios of nine insurance companies by average credit score. Based on this univariate study, Tillinghast concluded that poor credit scores are associated with relatively high loss ratios. While helpful, this study leaves open the possibility that credit could be a proxy for other available information.

The second study was published by [James Monaghan](#) in the 1999 Casualty Actuarial Society Forum. Monaghan performed a number of two-way analyses of credit score and such traditional insurance rating variables as driving record, driver age, marital status, and territory. In no case did the predictive power of credit vanish in the presence of another variable. Although Monaghan’s presentation is an advance over Tillinghast’s univariate study, it’s limited to two-way analyses. One would still like to see credit and a wide variety of non-credit variables all analyzed together, in a true multivariate fashion. Our insurance data mining work has allowed us to do precisely this.

We’ve completed a number of large-

scale data mining projects involving both credit and non-credit variables. The goal of these projects has been to build mixed credit/non-credit predictive models. As with commonly used credit scores like FICO, a variety of credit variables are used. But our scores aren’t restricted to credit variables. Indeed, any socially acceptable variable that might be useful for predicting insurance losses is fair game.

Before building predictive models, we create literally hundreds of predictive variables from driver, vehicle, coverage, agent, billing, and other information. These variables are all generated from data available from internal insurance company systems. In addition, we create external predictive variables from ZIP code-level weather, census, and demographic data sources. Credit variables and credit scores are also classified as external predictive variables.

This is the richest context imaginable in which to analyze the relationship of

FIGURE 2

## Examples of Internal and External Data Sources

- ▶ Motor Vehicle Reports
- ▶ Credit Reports
- ▶ Geographic/Demographic Information
- ▶ Enhanced Census Information
- ▶ Weather Data
- ▶ Billing/Payment History
- ▶ Agency Information
- ▶ Historical Claims Data
- ▶ Policyholder/Driver Information
- ▶ Vehicle Information

1/3  
Reden and Anders  
Page 63

Credit scoring has given insurers a powerful tool, useful for multiple tasks: more accurate pricing, more automated and objective underwriting, and more focused retention efforts.

credit and insurance losses. In a typical data mining project we have at our disposal hundreds of non-credit predictive variables and tens of thousands of data points. Because our projects are true multivariate studies, conducted using rigorous statistical methodologies on large quantities of data, we feel that our experience might add value to the debate over credit.

We've found that credit scores and credit variables are among the strongest variables in our models. These credit variables are associated with some of the most statistically robust parameters in our models. Furthermore, credit variables typically show up early in our variable selection procedures.

Finally, we've found that removing the credit variables from a model does some-

what dampen the loss ratio "lift" the model produces. This further indicates that credit variables aren't mere proxies for other variables in the models.

In short, our large-scale multivariate modeling work has convinced us that the relationship between credit scores and loss ratio remains strong even after many other variables have been taken into account. In other words, credit variables do play an important role in explaining insurance losses and profitability, even in a model that contains a variety of other variables.

### **Beyond Credit Scoring**

In recent years, many insurers have incorporated credit scores into their underwriting, pricing, and tiering plans. This has created more of a level playing field among insurers, and has therefore somewhat dampened the competitive advantage of credit scoring. At the same time, some states have restricted or even banned the use of credit in pricing and underwriting insurance.

Finally, little or no credit information is available for certain insureds. All of this presents a challenge: How can insurers enhance or replace the credit scoring engines that they've woven into their underwriting, pricing, and tiering schemes? Based on our experience, we feel the answer is clear: by building effective multivariate non-credit scores, mixed credit/non-credit scores, or both.

While we've seen that including credit scores and credit variables in a multivariate predictive model does increase the predictive power of that model, we've found that non-credit scores can still perform extremely well. Indeed, we discovered that the predictive power of non-credit models can be as good as, if not better than, that of typical credit scoring models.

Of course, eliminating credit variables means that such non-credit models must be carefully designed and constructed. As we've discussed, the overall process of building a non-credit score is much the same as that of building a credit score. We call this "data mining" methodology. But

1/3  
winkelvoss  
Page 64

regardless of how one chooses to name it, the major hallmarks of this process are:

- The use of a large dataset. For example, all of a company's policies from a five-year time period
- The creation of a large number of predictive variables
- The use of traditional as well as cutting-edge statistical techniques to build a scoring engine
- The use of lift analysis to judge the predictive power of the model and to help create business rules

Data mining is now possible because of the availability of cheap computing power and recent advances in statistical theory and machine learning. The power of data mining transcends the use of credit information, and unleashes considerable power to transform masses of raw data into profitable business actions. For this reason, we feel that insurance data mining is here to stay.

Because the power of data mining transcends the predictive power of credit variables, we see no reason why insurers' scoring models should be limited to credit data. Just as we can build credit-only models and non-credit models, we can build mixed credit/non-credit predictive models. Advantages of such models:

- The use of additional (non-credit) variables adds further predictive power to a model.
- Many of the non-credit variables leverage the insurance company's internal data sources.
- Models that don't depend entirely on credit information might be more acceptable to regulators.
- Non-credit variables could be used to update a policy's score each policy term.
- Models with non-credit variables can make accurate predictions even for policies with little or no available credit history.
- Credit variables can be tempered or removed entirely in cases where the use of credit is restricted.

To summarize, we see two alternatives that can help insurers go beyond credit scores. One option is to build models that contain the optimal mix of credit and non-credit variables; the other option is

to use the optimal mix of non-credit variables only.

Credit scoring has given insurers a powerful tool, useful for multiple tasks: more accurate pricing, more automated and objective underwriting, and more focused retention efforts. This power is due both to the inherent predictive value of credit information, and also to the power of data mining. Insurance data mining goes beyond the predictive power of credit information, and can leverage nearly any kind of internally or externally available information.

In the future, insurers will continue to search for new data sources that will give their scoring algorithms a competitive edge. The result will be an ongoing diminishing of the information asymmetry between insurers and policyholders, and a more efficient market for insurance.

Ultimately, insurance is about making the best use of all available information to manage risk. To stay competitive in the

ever more efficient market for insurance, insurers must continually maintain and develop cutting-edge predictive modeling strategies. ●

#### References

"[Insurance Scoring](#) in Personal Automobile Insurance—Breaking the Silence," Conning Report, Conning & Company (2001).

"Credit Reports and Insurance Underwriting," NAIC White Papers, National Association of Insurance Commissioners (1997).

Monaghan, J. E., "The Impact of Personal [Credit History](#) on Loss Performance in Personal Lines," CAS Forum, Casualty Actuarial Society (2000).

Wu, C. S. P. and Guszczka, J. C., "Does [Credit Score](#) Really Explain Insurance Losses? Multivariate Analysis from a Data Mining Point of View," CAS Forum, Casualty Actuarial Society, forthcoming. ●

1/3  
Actex  
Page 65