Demystifying Analytics
Proven Analytical Techniques and Best Practices for Insurers
About the Author

Stuart Rose is Global Insurance Marketing Director at SAS, a market-leading business analytics software vendor. He is responsible for thought leadership and marketing content for applying analytics within the insurance industry. Rose began his career as an actuary and now has more than 25 years of experience in the insurance industry.

He is a regular contributor to insurance publications and the Analytic Insurer blog. Rose frequently speaks at insurance conferences and is co-author of the book Executive’s Guide to Solvency II. He holds a BSc in mathematical studies from Sheffield University.
Introduction

Analytics is fast becoming a mainstream practice in the insurance industry. According to a recent survey of 165 insurance executives, 72 percent consider analytics the biggest industry game changer through 2015.

Many insurance companies are just beginning to take steps toward becoming an “analytic insurer” – one that embeds analytics into daily operations to make better decisions that reduce costs, improve pricing and more. And those organizations with more advanced analytics capabilities are actively seeking to build on previous successes and grow their analytics capabilities. It’s no wonder, given the increasing volumes of data being produced through enterprise business systems, online interactions, social media and other channels. Turning all of it into useful information is a challenge for most organizations.

What about your business? To what degree are you operationalizing analytics? Where do you see opportunities to grow your capabilities and improve your business?

This white paper will highlight some of the best practices and techniques that business analysts, data scientists and other domain experts can use to turn their data into valuable insights.

Data Preparation

Of all the steps in the analytical life cycle, gathering data often takes the most time. Data can reside in multiple transactional systems, such as policy administration solutions, claims management systems and agency applications. Database architects spend a significant amount of time running queries from various data sources and joining tables, all the while competing for system resources with other users.

At the same time, analysts who use this data must have a clear goal in mind when preparing the data for analysis. For example, they need to consider the objective of their models, define them accordingly, and understand what data will actually be available at execution time.

SAS recommends several data preparation techniques that can help you process the growing volume and variety of data that insurance companies collect, generate and access externally.

Data Warehousing and Analytical Data Tables

Data warehousing is the process of bringing together disparate data from across an organization to support decision making. A data mart is a subset of a data warehouse that is usually oriented toward the needs of a specific line of business or department. Data marts improve response time by allowing users to have access to the specific type of data they need to analyze most often.

But no data is perfect. And that’s why analysts need core data preparation tools for file importing and appending, merging and dropping variables, addressing missing values, filtering outliers and developing segmentation rules. Together, these tools allow analysts to achieve the quality of data optimally suited for analytics that resolves specific business problems.

Binning

There are times when converting a numeric variable into a categorical one is desirable. Binning is one way analysts can divide the range of a numeric variable into a set number of subranges (or bins) and replace each value with its bin number. Binning can also be a powerful way of dealing with variable skewed distributions, as the outliers are all placed together in one bin. For example, your actuarial department can use binning with the ratemaking process to determine the bands for pricing models. Some of the standard binning definitions are weights by records, weights by premium, or more commonly, method weight by exposure.

Exploratory Data Analysis

Insurance companies, regardless of their type or size, are constantly generating data every minute of every day. Everyone - including executives, departmental decision makers, underwriters, claims adjusters and call center workers – hopes to learn things from this data to help them make better decisions, take smarter actions and operate more efficiently.

But what’s the best way to turn data into insight? The science of extracting insight from data is constantly evolving. But regardless of how much data you have, one of the best ways to discern what information is important – and what isn’t – is through exploratory data analysis. Let’s take a look at some of the most valuable ways you can do this.
Line Graphs

Line graphs are most often used to track changes or trends over time. Line charts are also useful when comparing multiple items over the same time period. The stacking lines are used to compare the trend or individual values for several variables. For example, a line graph could be used to show the variation in losses accrued over a defined period (for instance, 14 months, as shown in Figure 1).

Bar and Pie Charts

Bar charts are most commonly used for comparing the quantities of different categories or groups. Values of a category are represented using the bars, and they can be configured with either vertical or horizontal bars, with the length or height of each bar representing the value. For example, a pie chart could be used to easily visualize the percentage of an insurer's gross written premium or revenue for each line of business.

Another form of a bar chart is called the progressive bar chart, or waterfall chart. A waterfall chart shows how the initial value of a measure increases or decreases during a series of operations or transactions. Figure 2 illustrates how you can use a bar chart to compare the number and type of insurance certifications your organization has by region.
Scatter and Bubble Plots

A scatter plot (or X-Y plot) is a two-dimensional plot that shows the joint variation of two data items. In a scatter plot, each marker (which is typically a symbol such as a dot, square or plus sign) represents an observation. The marker position indicates the value for each observation. Scatter plots are useful for examining the relationship (or correlations) between X and Y variables. Variables are said to be correlated if they have a dependency on, or are somehow influenced by, each other.

As shown in Figure 3, a bubble plot is a variation of a scatter plot in which the markers are replaced with bubbles. In a bubble plot, each bubble represents an observation. The location of the bubble represents the value for two measured axes; the size of the bubble represents the value for a third measure. These plots are useful for data sets with dozens to hundreds of values or when the values differ by several orders of magnitude.

Correlation Matrix

A correlation matrix, as shown in Figure 4, enables you to quickly identify which variables are related. It also shows how strong the relationships are between variables. Darker boxes indicate a stronger correlation; lighter boxes indicate a weaker correlation.

Correlation matrices can be used as an effective first-pass variable selection tool. Data variables with a strong relationship to the target can be kept, while other variables can be eliminated. An insurance company can use a correlation matrix to determine, for instance, which variables have an influence on loss history so people can more accurately price insurance products.
Clustering

Clustering looks for similarities within data and creates groups of data points based on these similarities (see Figure 5). Sometimes patterns that are not apparent in a large volume of data can be found within simpler clusters. Insurance companies often use clustering for customer segmentation to improve marketing campaign performance. By dividing a large customer base into smaller, more homogeneous groups, they can treat each customer in a group in a similar fashion in terms of messages, offers, pricing and more, and be confident in their relevancy.

Traditional Analytical Techniques

All insurance companies have one thing in common: Key insights, relationships and answers needed to run their businesses better are buried somewhere in corporate data. For example, you can analyze your data to answer pressing questions such as:

- Which customers will purchase what products and when?
- Which customers are leaving and what can be done to retain them?
- How should insurance rates be set to ensure profitability?
- How can you predict which claims are likely to result in litigation or subrogation?

To get answers to these kinds of complex questions, create effective strategies and gain an edge in today's competitive market, you need advanced analytics solutions that reveal hidden patterns and insights. Let's take a look at some of the most valuable, traditional analytical techniques.

Regression Models

Regression models are one of the most commonly used predictive modeling techniques. The simplest form is a linear regression model that has a single variable or input. In reality, multiple factors influence outcomes, so regression models often have more than one input; this is called a multiple regression model. But it's important to note that multiple regression models do not work well with a large number of inputs. So selecting just the right variables is the most important aspect of any analytical modeling effort.

Another type of regression model is a logistic regression. Logistic regressions are frequently used when the variable has a binary response, such as when asking the question, “Was the auto accident reported to the police?”

Generalized Linear Modeling

As a growing number of insurance companies have adopted the use of generalized linear models (GLMs) for ratemaking, the insurance industry has undergone a significant transformation. By identifying additional characteristics that can segment
existing rating cells into smaller cells with different rates, an insurer can use this more granular structure to compete with other insurers by using lower prices – and still continue to charge an appropriate premium. In addition, GLMs provide statistical diagnostics that help analysts select only significant variables and validate model assumptions. Today, GLMs are widely recognized as the industry standard.

Decision Trees
Decision trees, also known as classification or regression trees, allow you to analyze problems or predict behavior that’s influenced by multiple factors. A decision tree is a hierarchical collection of business rules that describe how to divide a large collection of records into successively smaller groups of records. Decision trees present influencing factors incrementally as a series of one-cause, one-effect relationships that are easier to understand than more complex, multiple-variable techniques.

Decision trees are popular with business users with a limited background in advanced analytics since they can easily build and interpret the results of this powerful tool. This technique is also employed by advanced analytical users (e.g., to analyze many parameters and further fine-tune analyses). Figure 6 illustrates a decision tree for gross profit data.

Forecasting
Generating forecasts (illustrated in Figure 7) is a crucial step for decision making and strategic and tactical planning. Forecasting is the process of analyzing conditions or events that take place over time to anticipate what is likely to happen in the future. Time series forecasting uses historical data to project future outcomes. With forecasting, you can identify previously unseen trends and anticipate fluctuations so you can plan ahead more effectively.
Emerging Analytical Techniques

The digital age has brought with it a huge increase in the amount of data available to companies. However, for most insurers, dealing with big data is not the problem. Insurers are more concerned about the variety, veracity and velocity than the volume of data being created in today’s digital environment. Most importantly, the vast majority of new data sources are semistructured and unstructured, which require new analytical techniques to generate insights. Let’s take a closer look at some of the most important, emerging analytical techniques used to analyze big data.

Link Analysis

From a technical perspective, link analysis converts data into a set of interconnected, linked objects that can be visualized as a network of effects. Insurers are successfully using link analysis to detect and prevent organized claims fraud by going beyond transaction and account views to analyze all related activities and relationships at a network dimension. Using a network visualization interface, as shown in Figure 8, insurance claims investigators can identify linkages among seemingly unrelated claims.

Text Analytics

Text analytics is the use of computer software to annotate and extract information from documents and electronic text sources and analyze that information for business purposes. Insurance companies are beginning to explore the vast power of text analytics by analyzing and gaining insights from unstructured data, such as medical records, emails, underwriter notes, call center logs, videos, even Twitter and blogs.

An example of a simple textual analytics technique is a word cloud visual, as shown in Figure 9. This technique can be used to display high- or low-frequency words, with the size of the word representing its frequency within a body of text.

Figure 8: An example of a hierarchical link analysis network showing the connection between multiple claims.

Figure 9: An example of a word cloud.
Figure 10: An example of text analytics exploring term relationships with interactive concept linking visualization.

Another textual visualization technique that can be used for semistructured or unstructured data is the network diagram (see Figure 10). Network diagrams view relationships in terms of nodes (representing individual actors within the network) and ties (which represent relationships between the individuals, such as friendship, kinship, organizations, business relationships, etc.). These networks are often depicted in a diagram where nodes are represented as points and ties are represented as lines.

**Model Management**

Model management and deployment are critical steps in the analytics life cycle. They involve validating a model, selecting a champion model, and publishing or deploying your models (in batch or real time). Model performance is also evaluated and monitored constantly in order to decide whether current models should be retired or upgraded, or if new ones are needed. Let’s explore some important best practices for model management.

**Model Validation and Comparison**

With advances in modeling software and computing performance, it’s both easy and feasible to build multiple candidate models for use in analyses. By experimenting with different binning definitions, analysts can efficiently select alternate input variables, model techniques and build multiple models. But before using these models, insurance companies need to evaluate them to determine which ones provide the greatest benefit. As shown in Figure 11, models can be compared across a number of metrics based on their statistical merit using measures like lift, receiver operating curve (ROC) and Kolmogorov-Smirnov (K-S) charts.

**Model Deployment**

Analytical models are at the heart of critical business decisions - for example, when you need to find new opportunities or manage uncertainty and risks. But their real business value is only realized when they are put into production, either in batch or real time, for use by decision makers on a day-to-day basis. It’s only when analytical models can be used in a call center environment, by an underwriter assessing risk, or a special investigative unit (SIU) determining fraud that insurers truly operate based on smarter decisions.

The challenge is that many insurers spend at least six months converting a predictive model into a programming language that can be used by traditional transaction systems (e.g., a claims management solution). In today’s environment, customer
For insurance companies, greater regulatory scrutiny increases the need for model transparency. However, keeping records of the entire modeling process – from data preparation through model performance – can be time consuming and unproductive.

As a best practice, employ model governance tools that enable you to automate model governance – complete with documentation to create qualitative notes, summaries and detailed explanations. Having an efficient model governance approval process can also save you time and money, as it can help you avoid audits and regulatory fines.

How Can SAS® Help?

Analytics is no longer a luxury. It has become fundamental to the success and growth of insurance organizations worldwide because it supports better decision making about customers, prices, offers, risks and more. The key is turning the massive amounts of data your company has collected – and continues to collect – into timely, trusted insights and competitive advantage.

As a best practice, employ model governance tools that enable you to automate model governance – complete with documentation to create qualitative notes, summaries and detailed explanations. Having an efficient model governance approval process can also save you time and money, as it can help you avoid audits and regulatory fines.
With the SAS Analytics framework, you can deploy solutions such as:

- **SAS Visual Analytics**. SAS Visual Analytics provides a complete platform for analytics visualization, enabling you to identify patterns and relationships in data that weren’t initially evident. Interactive, self-service BI and reporting capabilities are combined with out-of-the-box advanced analytics so everyone can discover insights from any size and type of data, including text.

- **SAS Enterprise Miner™**. This solution is designed for people who need to analyze increasing volumes of data to identify and solve critical business or research issues, and support well-informed decision making. SAS Enterprise Miner streamlines data mining and analytical processes to create accurate, predictive and descriptive analytical models using vast amounts of data.

- **SAS Model Manager**. SAS Model Manager provides a comprehensive framework to support full lifecycle management and model governance. It makes it easier to manage modeling processes and quickly put the best models into production. Performance monitoring and alerts automate the model updating process so you can avoid model degradation and ensure that models are always performing at their highest levels.

Learn More

Organizations are always looking for that “X factor” - something to clearly differentiate them from competitors. For insurance companies, the X factor is often hidden in mountains of data. But data alone is worthless; it must be prepared, analyzed and visualized in order to get valuable insights. And with the emergence of big data, the possibilities for breakthrough insights are increasing dramatically. One thing is clear: Insurance companies that invest in analytical capabilities and apply best practices, as described in this paper, will identify their X factor much faster than their competitors. At the same time, they will be able to make faster, smarter decisions across the entire business.

To learn more about how SAS can help you apply analytics best practices to address your insurance business challenges, contact your SAS sales representative or visit us online at [sas.com/en_us/industry/insurance.html](http://sas.com/en_us/industry/insurance.html).