Paper 087-2007

Critical Success Factors for Rate Making with SAS® Enterprise Miner™

Hrushikesh Khopkar, SAS Institute Inc., Cary, NC
Jay King, SAS Institute Inc., Cary, NC

ABSTRACT

Rate making and risk-based pricing give users a definite competitive advantage in the general insurance business. This advantage is all the more compelling if the company is an industry leader. This paper attempts to tie the theory of rate making to technology and business definitions and describes how to set a benchmark for validation and customer handover to ensure a successful rate-making project.

The objective of this paper is to reduce the number of iterations involved in the development of a rate card, or formula, and to discuss available options. To underwriters, this paper provides a road map and proven best practices. To project managers, it offers a guide to scope commitment, dependency tracking, and progress tracking. To analytical developers, this paper describes the business definitions and thresholds that impact the acceptance of results. This paper presents the life-cycle of a rate-making project and highlights how the advances in theory and technology can be usefully leveraged to price insurance products tactically.

THE CHALLENGES OF RATE MAKING

Rate making is the process by which a company, using a wide assortment of data, attempts to predict both the likelihood that a claim will be made and the total amount of claims in a policy period in order to price products accordingly. To do this, insurers must quantify the risks they are willing to assume and the premiums they will charge to assume them, taking into account the overall goals of the company and, perhaps, government regulations. The imperatives in rate making are as follows:

1. **The need to set the price of the product before knowing the cost of the product:** Pricing insurance products is unlike pricing manufactured goods. In the latter case, we know the costs before we go to market. In the former, an insurer has no way of knowing what the realized cost per policy will be because it depends on whether the policy buyer and others in that financial year will file claims and, if so, how many claims will be filed and how large these claims will be.

2. **The need to compete on price but still ensure overall solvency:** Insurance is becoming increasingly complex and the key to survival is to ensure that the loss ratio of an insurance company achieves the desired target. Competitors can emerge with innovative rate structures, targeting particular low-risk "good customers" by offering attractive rates. Actuaries have the challenge of accurately forecasting the risk of people filing claims—which requires a powerful solution that can take into account numerous factors.

3. **The need to balance strategic, tactical, operational, personnel, and project management priorities:** A pricing plan that is too complex is difficult to operationalize and maintain in IT, marketing, and financial environments. On the other hand, a pricing plan that is too simple does not give the advantage of differential pricing. Pricing that is too aggressive is detrimental to solvency and pricing that is too high might cause a loss in market share. Therefore, it is imperative that a rate making initiative balance strategic, operational, and staffing needs. Rate making requires the involvement of analytically skilled insurance and managerial personnel and IT professionals with knowledge of the data sources involved.

THE CONCEPTUAL SOLUTION FOR PRICING INSURANCE PRODUCTS

The solution to the rate-making challenge must be based on policy characteristics that are submitted at the time of underwriting (for example, characteristics of the policy owner, other individuals covered, property covered, and so on). The only way to infer the loss propensity with respect to these characteristics is by historical analysis of the link between the policy characteristics and the corresponding losses. Conceptually, the solution must

1. Charge a price to every prospective customer, so that the pool of collections in a term is sufficient to service the overall claim amount (in that term) filed by prospective customers.

2. Charge a relatively higher price to prospects who are either more likely to file claims than others or who are likely to file larger claims than others.

AN EXAMPLE SOLUTION

Table 1 is an example of rating factors that were developed using the optimal binning and generalized linear modeling methods that are discussed later in this paper.

POLICY CHARACTERISTICS	RATING FACTOR BUCKET CORRESPONDING TO POLICY CHARACTERISTICS	PARAMETERS OF FREQUENCY PREDICTION EQUATION	PARAMETERS OF SEVERITY PREDICTION EQUATION
Intercept		**-2.69**	**10.27**
Makemodel1, Makemodel2, Makemodel3	MAKEMODEL_BIN 2	-0.14	-1.16
Makemodel4, Makemodel5, Makemodel6, Makemodel7	MAKEMODEL_BIN 3	0.51	-0.54
Makemodel8, Makemodel9	MAKEMODEL_BIN 4	0.34	-0.96
Makemodel10, Makemodel11, Makemodel12	MAKEMODEL_BIN 5	-0.30	-0.83
Makemodel13, Makemodel14	MAKEMODEL_BIN 6	0.22	-0.99
Makemodel15, Makemodel16, Makemodel17	MAKEMODEL_BIN 7	0.27	-1.12
Makemodel18, Makemodel19, Makemodel20	MAKEMODEL_BIN 8	0.41	-0.94
Makemodel1345, Makemodel1346, Makemodel13465	MAKEMODEL_BIN 9	0.68	-0.39
Makemodel3965, Makemodel842, Makemodel745	MAKEMODEL_BIN 10	0.43	-0.77
Makemodel24, Makemodel825, Makemodel548	MAKEMODEL_BIN 11	0.51	-0.64
Makemodel1344, Makemodel174, Makemodel8365	MAKEMODEL_BIN 12	0.48	-0.75
Makemodel34154, Makemodel43678, Others	MAKEMODEL_BIN 13	0.22	-0.26
Makemodel7777	MAKEMODEL_BIN 14	0.00	0.00
City11, City34, City234	CITY_BIN 2	0.00	-0.11
City234, City234, City342	CITY_BIN 3	0.11	-0.05
City1234, City6345, City634	CITY_BIN 4	0.22	-0.01
City435, City234, City2346	CITY_BIN 5	0.20	-0.36
City778, City789, City-567	CITY_BIN 6	-0.06	-0.04
City21, City6346, City61346	CITY_BIN 7	0.31	0.17
City2341, City16516, City6164	CITY_BIN 8	0.12	0.00
City45525, City79679, City1124	CITY_BIN 9	0.10	-0.03
City235, City613, City6134	CITY_BIN 10	-0.10	0.08
City6421, City,7413, City136, others	CITY_BIN 11	0.06	0.06
City847, City832, City7175	CITY_BIN 12	0.00	0.00
NCB_PERCENT =0 or NCB_PERCENT = Missing	NCB_BIN 2	0.82	-0.09
NCB_PERCENT > 0	NCB_BIN 3	1.58	-0.77
NCB_PERCENT = -55 or NCB_PERCENT = -35 or NCB_PERCENT = -25	NCB_BIN 4	0.37	-0.08
NCB_PERCENT = -45 or NCB_PERCENT = -20	NCB_BIN 5	0.57	-0.08
NCB_PERCENT = -65	NCB_BIN 6	0.00	0.00
VEHICLE_AGE =0	VAGE_BIN 2	0.46	0.30
VEHICLE_AGE =1	VAGE_BIN 3	0.47	0.25
VEHICLE_AGE >1 AND VEHICLE_AGE <=3	VAGE_BIN 4	0.38	0.19
VEHICLE_AGE =4	VAGE_BIN 5	0.25	0.08
VEHICLE_AGE >=5	VAGE_BIN 6	0.00	0.00

Table 1: Example Rating Factors

Given the rating card in Table 1 and the following information, the estimated accident claim cost for a policy is calculated in Table 2:

car of make-model = 'Makemodel1'
city = 'City11'
no-claim-bonus = NCB_PERCENT =0 %
vehicle age = VEHICLE_AGE =0

	Makemodel	City	NCB_Percent	VEHICLE_AGE	
	'Makemodel1'	'City11'	'0%'	'0'	
Predicted frequency of accident claim per insured vehicle-year	=exp((-2.69) + (-0.14) + (-0.00) + (0.82) + (0.46))				=0.21 claims per insured vehicle-year
Predicted severity of accident claim per claim	=exp((10.27) + (-1.16) + (-0.11) + (-0.09) + (0.30))				= $9,997 per claim

Table 2: Example Rating Model

Thus, the

Predicted Pure Premium per policy = 0.21 * $9,997 = $2,099 per insured vehicle-year

and, assuming that the

Insurer Declared Value (IDV) = (sum insured where vehicle / driver characteristics as listed above) = $201,741

then,

Rate as % of IDV = (Predicted Pure Premium) / (sum insured) = $2,099 / $201,741 = 1.04%.

Note that this paper does not focus heavily on liability insurance.

CRITICAL SUCCESS FACTORS IN RATE MAKING

Rate making is a series of choices. There are a number of alternative definitions, techniques, and workarounds at each step. The authors consider these steps as critical success factors in rate making:

1. **Agreeing on the scope, actuarial definitions, and their database representation**

   As trivial as it may sound, obtaining agreement among business and field personnel, the actuarial team, and the database team about the connotation and data representation/extraction of entities forms the most critical success factor. Unambiguous definitions of what constitutes an exposure, how a peril is to be identified in the claim data, how to summarize the loss cost per policy (excluding the Loss Adjustment Expense), what constitutes a completed term, and what are the merges required to obtain the unified view of policy and claim data form the bedrock of underwriting and rate making. The length of data history that is available, the competitive position of the company in the market, the stage of the business cycle, the stability of portfolio composition, and, finally, the credibility of the data quality should be checked for sufficiency.

2. **Agreeing on the desired structure of a rate card and business model and processes**

   Rates are very closely related to the business model and processes. The preferences and expectations for the end result of the rating exercise are extremely critical.

   Theoretically, the factors, factor levels, and factor weights that affect the number of claims differ from those that affect the severity of claims. However non-actuarial field personnel often find this information very arcane and difficult to assimilate and communicate. Additionally, different geographies might have different prerogatives about whether the level of factors should be the same for predicting frequency and severity.

The number of rating factor levels that are feasible for the business to operate is also of concern. High granularity of rating classes is recommended in a highly competitive market. This will enable the insurer to offer the best price to each prospective customer and still acquire a large number of customers. However, it makes communicating and operationalizing these classes more cumbersome. Capping the number of rating classes in a factor affects the ability to implement the model.

Another decision that affects the complexity of the rating model is the inclusion of complex, statistically significant rating factors. For example, a pure premium differential between males and females might differ by age. Including such an interaction effect might improve predictive accuracy, but it makes model interpretation and implementation more complex. However, a number of governments are prohibiting the use of certain statistically significant factors, such as age, gender, and marital status, for political reasons. Such prohibitions force significant adjustments to the rate-making process and result in one population subsidizing the rates of another.

Business preferences about compulsory inclusion and exclusion and caps should be noted early in the exercise. Compulsory business rules that cap the number of factors force certain factors to be included or excluded, and directives to take the existing underwriting rules and processes into account should be considered.

Certain economies and organizations prefer to express the rate as a percentage of the IDV against the pure premium in currency terms. The IDV typically has a depreciation schedule, and it is important to have firm preferences about this aspect before beginning the rating exercise.

Another issue is the choice about monitoring model performance and determining when to update or recalibrate the model. Typically, the length of the policy term of a particular product is set as the frequency for updating the rates. However, it is ideal to perform an analysis every quarter with incremental data to validate the efficacy of the champion model.

Organizations often prefer to have a rate card in a traditional table format rather than a scoring equation. It's important to have a consensus among the stakeholders on this issue. Rate discovery is an analytical exercise that is carried out offline. Identifying how the new rates will be implemented in the field system is critical to successful implementation.

Finally, an articulate and measurable strategic objective is needed. For example, a 20% reduction in Loss Ratio (LR), a change of product terms or target marketing, or a withdrawal of the product keeps the rating exercise on the desired track and serves as a validation criterion.

3. Defining the rate-making computing environment

Defining the computing environment that is used in the rate-making process, though not analytical, is also important for the successful implementation of a rate-making solution.

The critical tools needed for producing the frequency and severity statistical models are: SAS Enterprise Miner Server software, SAS Enterprise Miner Client software, SAS Enterprise Miner Tree Desktop Application, SAS® Enterprise Guide®, and the Genmod Extension node for SAS Enterprise Miner. The Genmod Extension node is distributed through a SAS consulting engagement and is not generally available as production-supported code from SAS. Readers who want to learn more about developing SAS Enterprise Miner extension nodes should enroll in the SAS training class, *Extending SAS Enterprise Miner 5* (support.sas.com/training/us/crs/pmex.html). As with all software installation projects, acquiring the tools that you need requires the cooperation of the IT department as well as business users and actuaries to ensure that the software is loaded on the necessary platform(s) for the work that each group will perform.

Granting appropriate permissions on data and project areas, creating users on the appropriate platforms, and creating users in metadata are essential to provide relevant resource access. Though this requirement may sound trivial, projects are often needlessly delayed because end users of the tools are not granted the appropriate permissions to use the software, access the data, or access the appropriate resources on development servers.

The importance of determining an appropriate backup policy for SAS Enterprise Miner development work, production reports, data, and metadata goes without saying. Backups are a necessity for any project that is critical to the business.

Many of the processes and modeling techniques that are performed in a rate-making project are memory- and disk-intensive. Monitoring disk and memory usage is, therefore, critical in order to anticipate and mitigate any problems that may arise.

4. Ensuring the adequacy and appropriate granularity of the data

A rating exercise is based on the premise that policy characteristics are correlated with the claim loss experience. Extracting and aggregating representative historical data is the first contact point in which theory is put into practice. Mature organizations typically have a data warehouse in place that can serve as a source of summarized information. In other cases, data must be extracted from operational systems for policy and claim management. To obtain the data, perform the following tasks:

Step1: Identify the policies in scope (based on business requirements).

Step2: Aggregate the transaction-level data in the policy management system to a policy-level granularity.

Step3: Identify the claims transactions.

Step4: For the claims identified in Step 3, aggregate the claim transactions to a claim level along with claim-related flags.

Step5: Join the exposure data with the claims data. This gives you a unified view of exposure and claims and is popularly referred to as the *severity table*.

Step6: Summarize the severity table to the policy level, aggregate the claim amount, and generate a claim count per policy per peril. This table structure is popularly called the *frequency table*.

Step7: It is critical to check the quality of the data extraction. Items such as the number of policies that are written between the defined inception date and expiration dates, the number of claims, the missing values, and the illegal values should be checked.

It is not uncommon to iterate through these steps. Often information that is learned in Steps 5 through 7 leads to insights that change the definitions that were determined in Step 1.

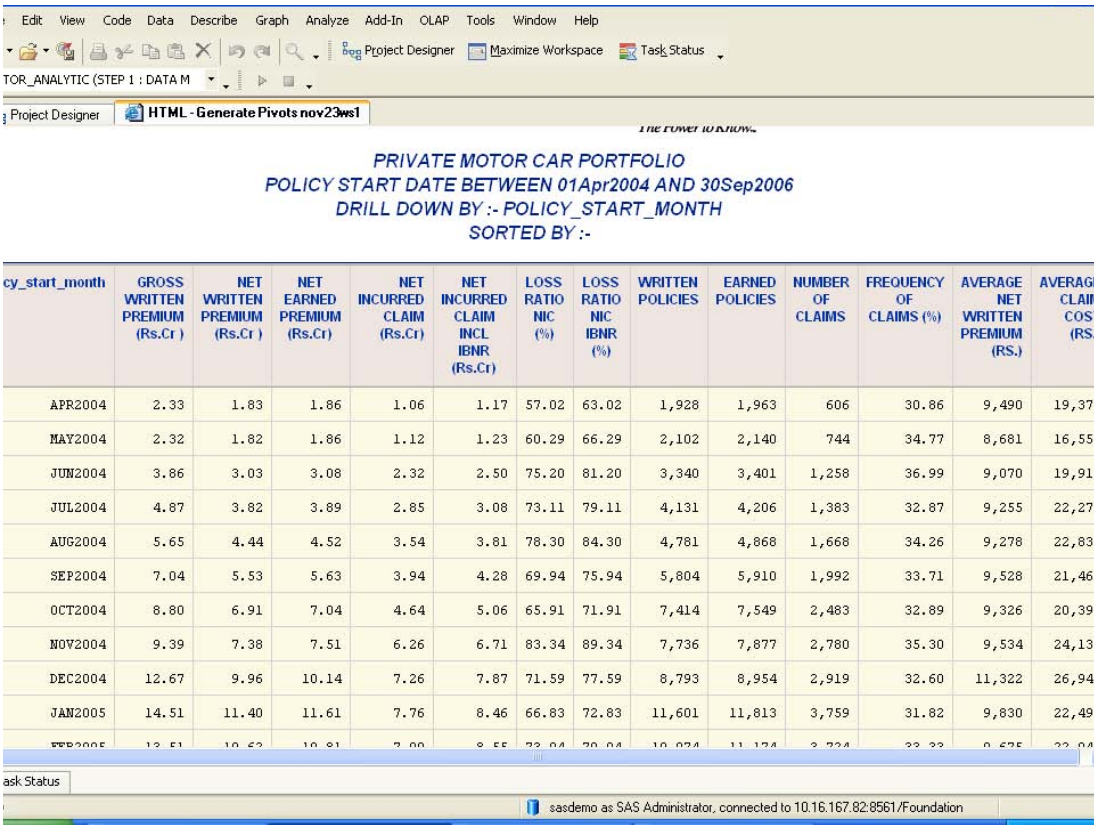
The source data is typically available either in a relational format in the operational systems or in a dimensional schema in the organization's warehouse. SAS® Data Integration Studio (formerly known as SAS® ETL Studio) enables the extraction, joining, and aggregation of raw data into the severity and frequency tables. It also provides the ability to schedule all such jobs.

5. Determining baseline portfolio loss metrics and generating underwriting reports

Given the availability of the tables from the previous section, the next major step is generating summarization reports. At the portfolio level, one would typically report metrics such as actual claim cost, total business (sum insured), loss ratio (claim cost/premium earned), actual claim cost as a percentage of sum insured, number of claims, and number of policies. Additionally, these metrics are also reported by candidate rating factors.

Executives often have a day-to-day view of operations and, hence, may be getting varying impressions about the company's portfolio. Standardized reports give consistent insight into the state of the portfolio and underwriting metrics. It is critical that everyone agrees to the loss ratios and premium amounts that are shown in the raw data before embarking on a predictive exercise.

SAS can generate underwriting reports of varying sophistication. A simple PROC MEANS can produce a static report through SAS Enterprise Guide, for example (see Figure 1). Ideally, such a process is parameterized and deployed as a stored process (see Figure 2) that could be executed with either SAS Enterprise Guide or from within Microsoft Excel using the SAS® Add-In for Microsoft Office. Optionally, one can deploy interactive reports using the SAS® Enterprise BI Server or SAS Enterprise Guide (see Figure 3).



PRIVATE MOTOR CAR PORTFOLIO
POLICY START DATE BETWEEN 01Apr2004 AND 30Sep2006
DRILL DOWN BY :- POLICY_START_MONTH
SORTED BY :-

cy_start_month	GROSS WRITTEN PREMIUM (Rs.Cr)	NET WRITTEN PREMIUM (Rs.Cr)	NET EARNED PREMIUM (Rs.Cr)	NET INCURRED CLAIM (Rs.Cr)	NET INCURRED CLAIM INCL IBNR (Rs.Cr)	LOSS RATIO NIC (%)	LOSS RATIO NIC IBNR (%)	WRITTEN POLICIES	EARNED POLICIES	NUMBER OF CLAIMS	FREQUENCY OF CLAIMS (%)	AVERAGE NET WRITTEN PREMIUM (RS.)	AVERAG CLAI COS (RS
APR2004	2.33	1.83	1.86	1.06	1.17	57.02	63.02	1,928	1,963	606	30.86	9,490	19,37
MAY2004	2.32	1.82	1.86	1.12	1.23	60.29	66.29	2,102	2,140	744	34.77	8,681	16,55
JUN2004	3.86	3.03	3.08	2.32	2.50	75.20	81.20	3,340	3,401	1,258	36.99	9,070	19,91
JUL2004	4.87	3.82	3.89	2.85	3.08	73.11	79.11	4,131	4,206	1,383	32.87	9,255	22,27
AUG2004	5.65	4.44	4.52	3.54	3.81	78.30	84.30	4,781	4,868	1,668	34.26	9,278	22,83
SEP2004	7.04	5.53	5.63	3.94	4.28	69.94	75.94	5,804	5,910	1,992	33.71	9,528	21,46
OCT2004	8.80	6.91	7.04	4.64	5.06	65.91	71.91	7,414	7,549	2,483	32.89	9,326	20,39
NOV2004	9.39	7.38	7.51	6.26	6.71	83.34	89.34	7,736	7,877	2,780	35.30	9,534	24,13
DEC2004	12.67	9.96	10.14	7.26	7.87	71.59	77.59	8,793	8,954	2,919	32.60	11,322	26,94
JAN2005	14.51	11.40	11.61	7.76	8.46	66.83	72.83	11,601	11,813	3,759	31.82	9,830	22,49

Figure 1: Static HTML Report Generated in SAS Enterprise Guide Showing PROC MEANS Output of Various Factors by Policy Start Month

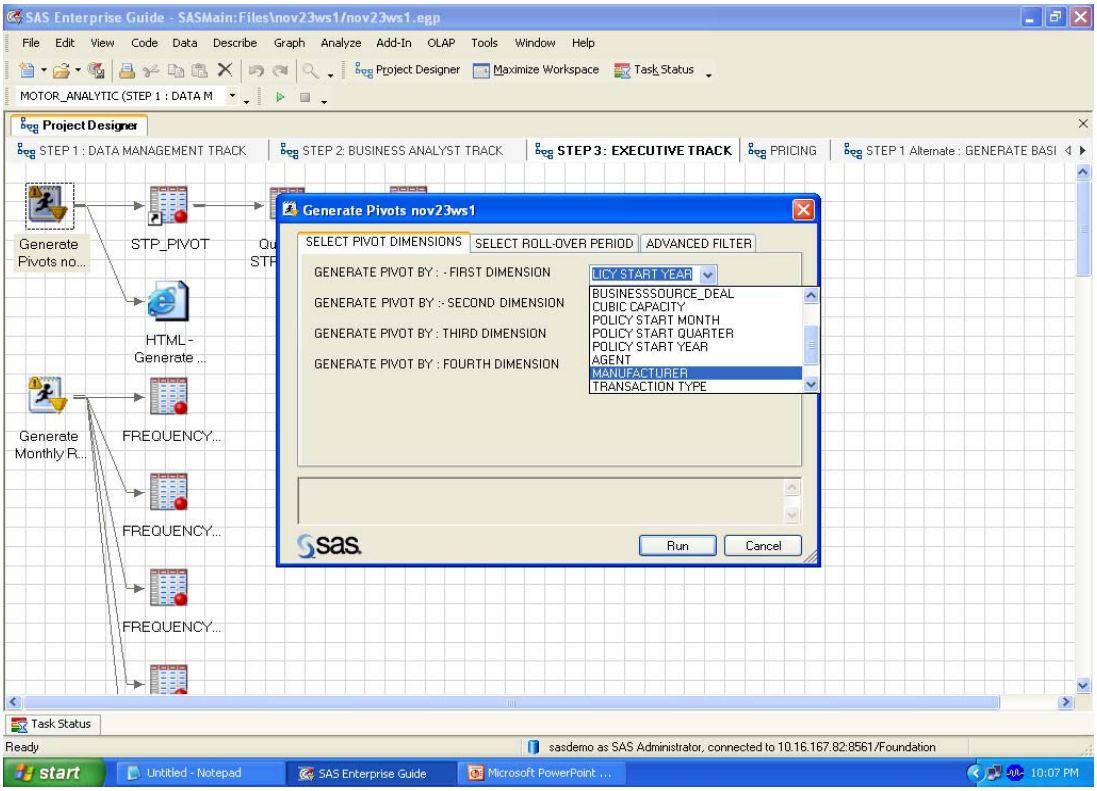


Figure 2: A SAS Stored Process in SAS Enterprise Guide that Generates Pivot Tables

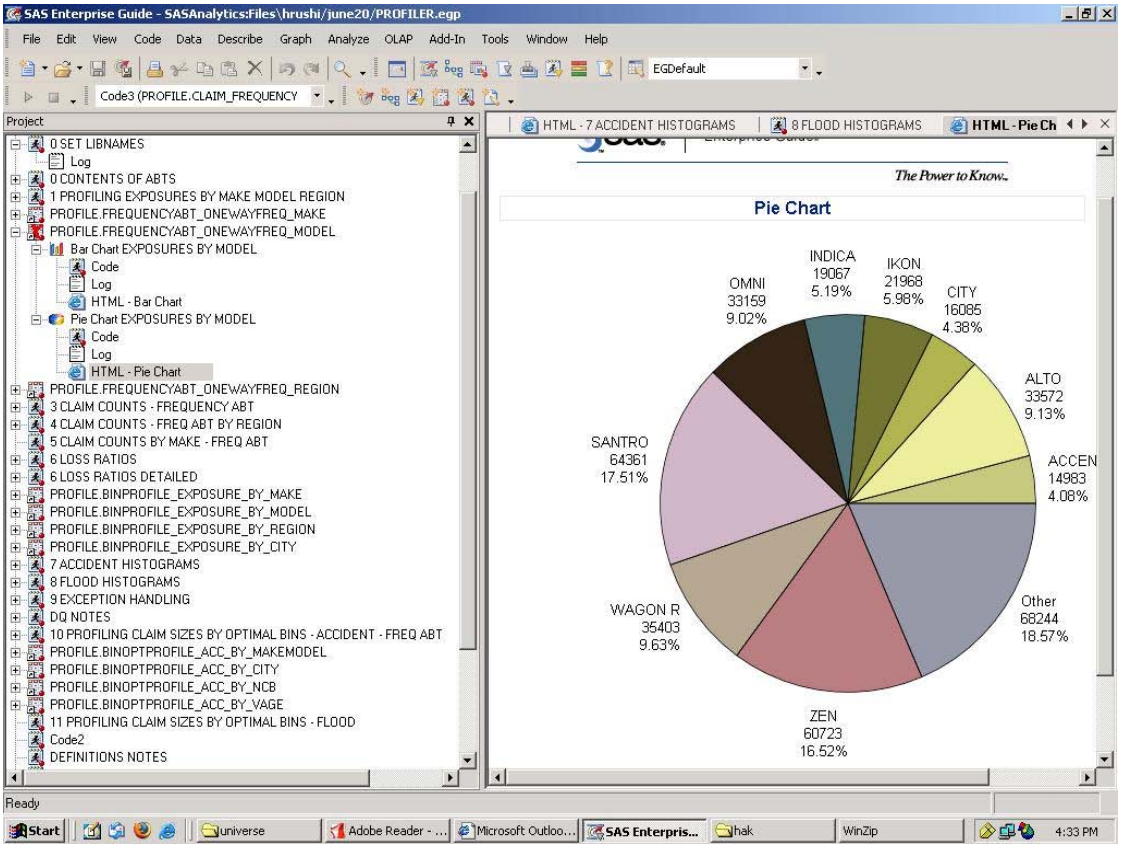


Figure 3: A Pie Chart Generated in SAS Enterprise Guide that Shows Exposure by Vehicle Model and is Drillable to Rating Factor Details

6. Performing one-way analysis to confirm business expectations and form initial hypotheses

Rate relativities refers to the difference in rates between classes of insured. Traditionally, this was achieved by one-way frequencies of the number of claims against the rating factors (univariate approach). Table 3 shows a simple example of such tables. Though we now have the power of multivariate techniques to identify the exact relativities, one-way analysis still provides the most intuitive starting point for discussion and hypothesis formation.

	Claims				Exposures	
Age of	Age of Driver			Age of	Age of Driver	
Vehicle	16-21	22+		Vehicle	16-21	22+
0-3	20	35		0-3	80	700
4-8	50	25		4-8	250	1000
9+	300	5		9+	2000	500
	Actual Frequency				Frequency Relativity	
Age of	Age of Driver			Age of	Age of Driver	
Vehicle	16-21	22+		Vehicle	16-21	22+
0-3	0.25	0.05		0-3	1	0.2
4-8	0.2	0.025		4-8	0.8	0.1
9+	0.15	0.01		9+	0.6	0.04

Table 3: One-Way Frequency Tables

7. Determining optimal binning of explanatory factors

The core agenda of risk-based pricing is identifying rating classes, or bins, based on loss experience. Risk-based pricing identifies risky members of a factor such that all of them are in one group and the rule is predictive and repeatable in nature. The number of levels of a rating variable implies the number of prices in the rating plan and also the granularity of risk segmentation. Setting the number and members of each level, or bin, is a critical and hotly-debated exercise. It is a critical issue not just from the standpoint of risk management but also from the standpoint of IT system management, marketing operations management, and organizational culture.

Identifying factor levels can be arbitrary, judgmental, or optimality driven. Decision trees are transparent, intuitive, non-parametric, and robust to influential values, outliers, and missing values; therefore, they can be used to find optimal bins, or factor levels. This forms the core "mining" work where one is looking for a pattern in the data to explain the phenomenon that is being analyzed. The factor levels, or optimal bins, for the frequency of claims can be generated by setting the number of claims (an ordinal variable) as the target variable while the optimal bins for the severity of claims can be generated by setting the size of the claims as the target variable. Figure 4 shows a simple example of a SAS Enterprise Miner process flow that uses decision trees to find optimal bins for vehicle model and vehicle age for claim severity, and Figure 5 shows an example of the decision tree output.

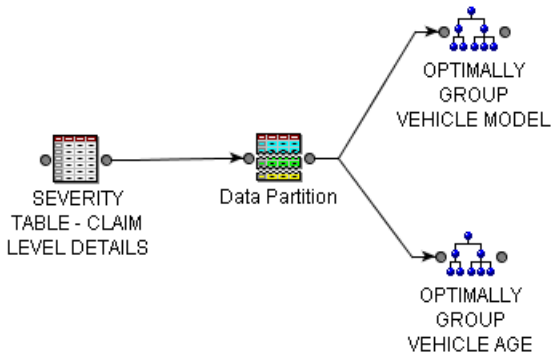


Figure 4: Example SAS Enterprise Miner Process Flow Using Decision Trees to Find Optimal Classes

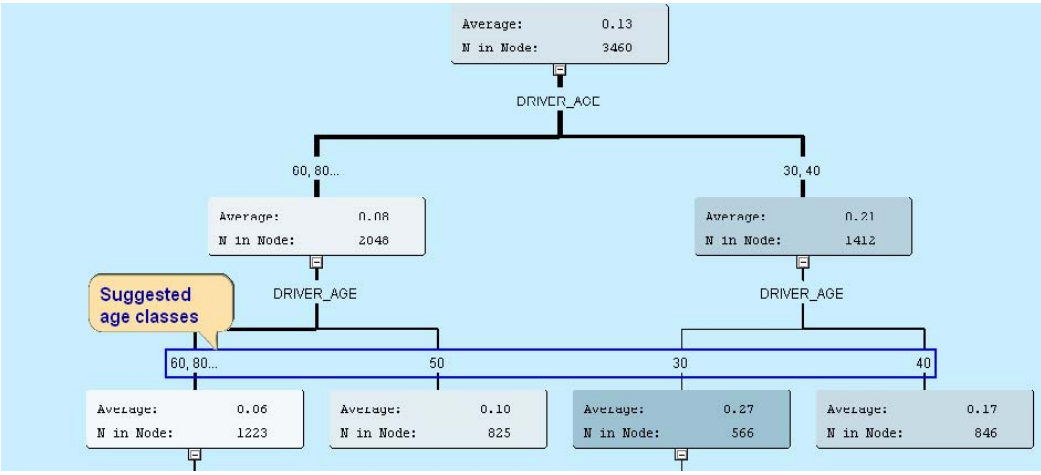


Figure 5: Example Decision Tree Rules

SAS Enterprise Miner also provides an interactive interface for decision trees, which further enhances the business utility of the results by providing an option to manually segment the population. This allows the analyst to develop rating factors, which, although they are not "optimal" bins, are based on business rules or personal judgment. Figure 6 shows a tree that is being built interactively. Though suboptimal, these factor levels might have greater acceptability in the business community.

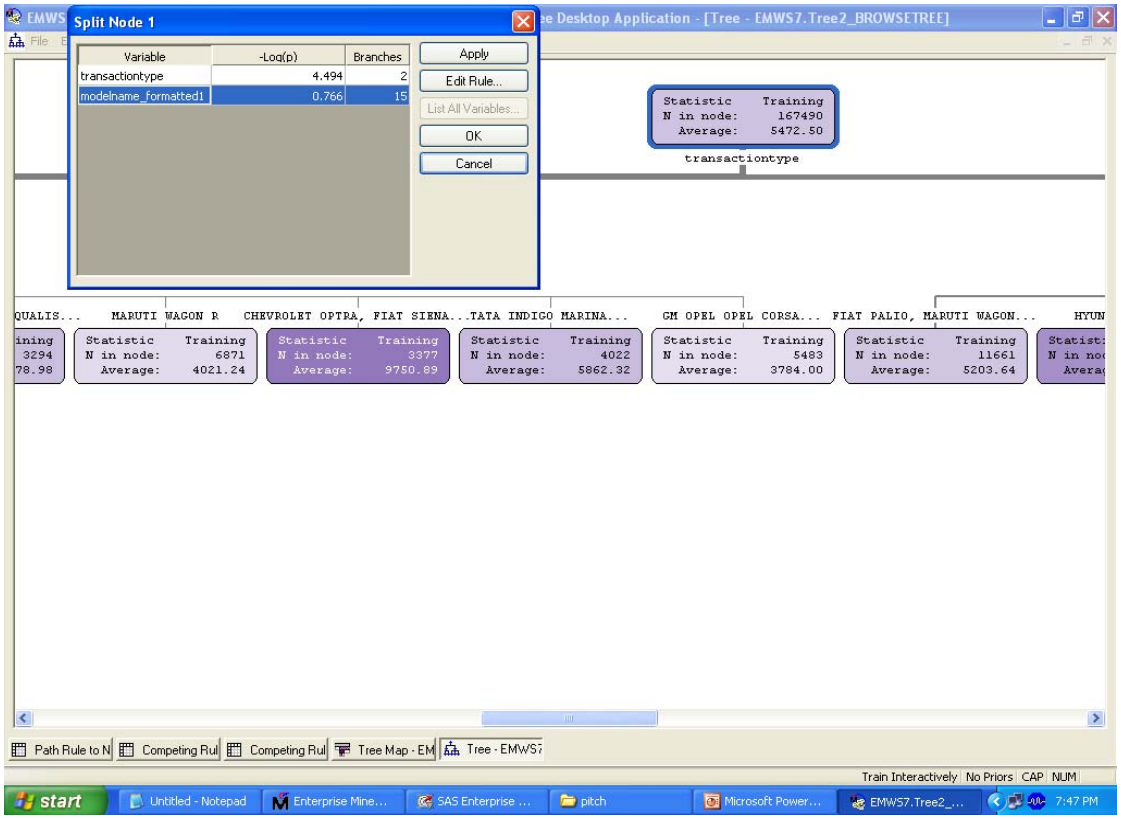


Figure 6: SAS Enterprise Miner Tree Desktop Application Helps Develop Decision Trees Interactively

In addition to decisions about the number of levels for factors and possible business overrides, several other decisions must also be made during this process. What should the level of significance for factor splits be? What splitting criterion should be used? Should there be a minimum number of members in a node (policies or claims) to create a split? If so, what should that number be?

Once all decisions have been made, the Segment Profiler node in SAS Enterprise Miner (see Figure 7) can then juxtapose the distribution of a factor in a segment (outer circle) with that of the overall population (inner circle). It gives visual proof and justification for why one risk segment should be priced differently than the other.

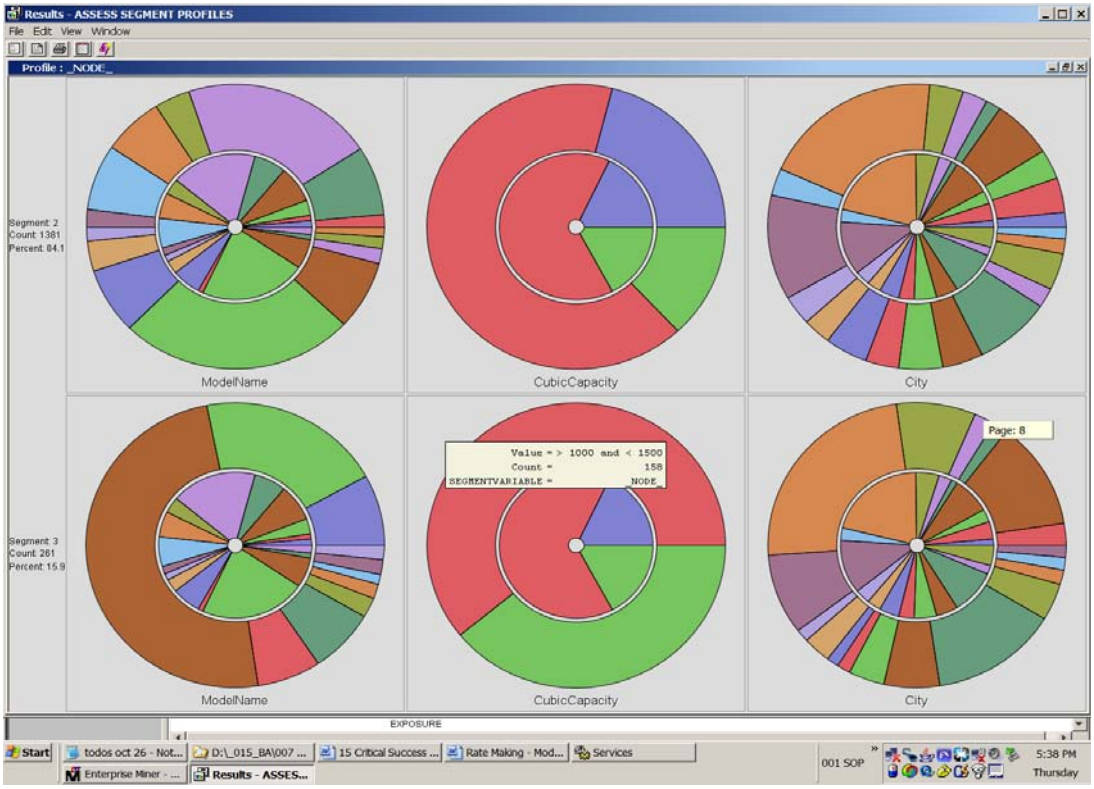


Figure 7: SAS Enterprise Miner Segment Profiler Node Output

8. Choosing the distribution of the target variables and the modelling technique to be adopted

At this point, the modeler is faced with a design decision—determine the most appropriate and practical technique and equation specification.

If one-way analysis is considered the "first generation" technique for rate making, then the "second generation" technique is based on identifying the parametric distribution of the loss amount and the number of claims and a multivariate relation to explain the distribution by variations in the policy characteristics. There have been significant advances in modeling loss, which typically turn out to be skewed and non-normal (though this is still in the exponential family of distributions). For example, Figure 8 shows an example of the distribution of claim severity. The SAS/STAT® GENMOD procedure, which stands for generalized linear modeling, provides a rich set of capabilities for fitting such distributions.

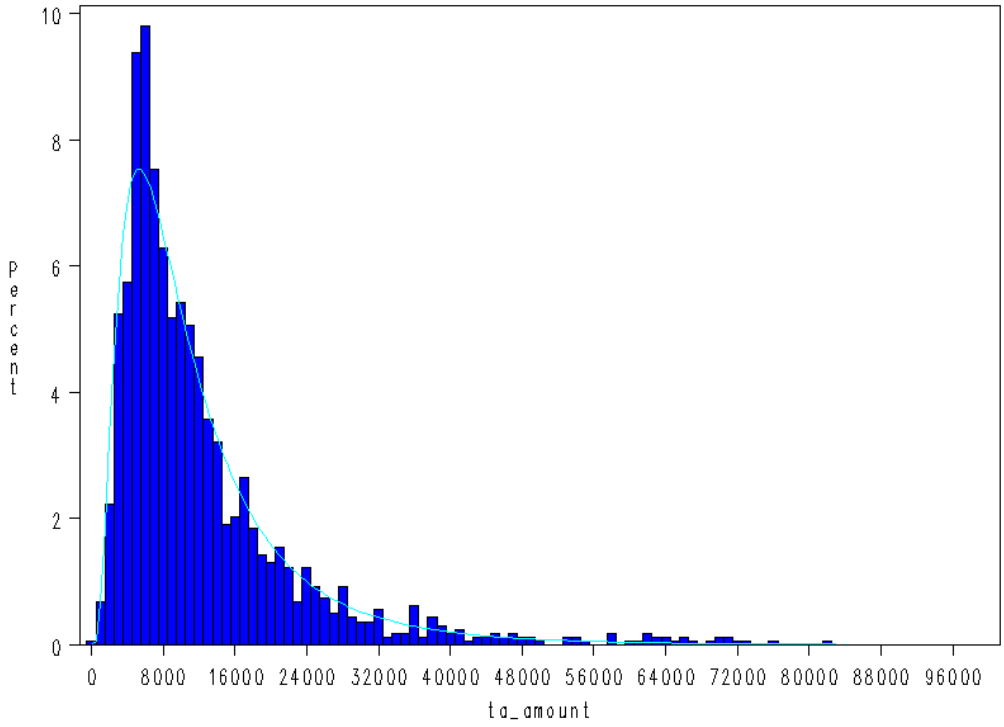


Figure 8: Claim Severity Plot

Within the second generation paradigm, there are a host of insurance-specific distributions and adaptations. A common motivation for fitting Zero-Inflated Poisson (ZIP) models is that count data often exhibits more zeros than would be expected under standard Poisson model assumptions (see Figure 9). It is one common source or form of overdispersion. The Poisson, with only one mean, or rate parameter, is often unable to adapt to situations where there is a larger fraction of zero counts. These techniques allow for modeling pure premium (compound distribution) directly.

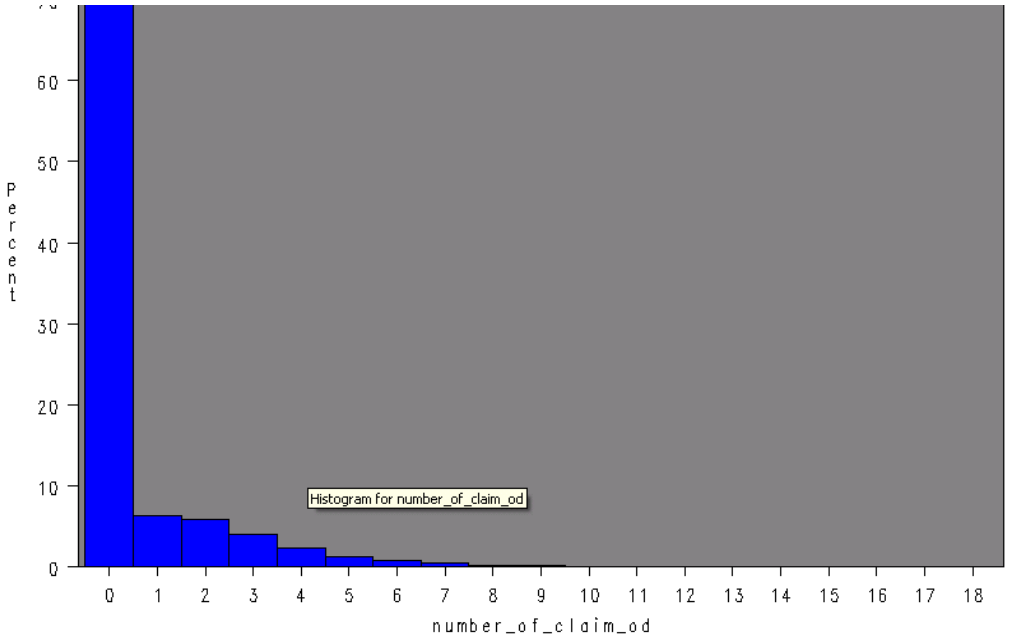


Figure 9: Histogram of Claim Counts

Some practitioners recommend the simple non-parametric approach of using decision trees to explain pure premium directly (that is, pure premium as the target and all of the candidate variables as input). This forms an easy-to-understand and easy-to-implement solution but does not provide the relativities of different factors and their levels. The authors recommend that factors be binned one at a time and the binned levels be specified into a generalized linear estimation model.

Other modern data mining techniques, such as neural networks, are not typically considered for use because they lack transparency. Although such techniques might provide excellent results, the need to create rating systems that are easily understood by management, the field, customers, and government regulators typically excludes them from consideration.

9. Specifying regression and parameters

a. Fitting a generalized linear model for severity and frequency with SAS PROC GENMOD

SAS Enterprise Miner provides a process flow diagram with a drag-and-drop approach to the modeling process. Modeling with trees and regressions from the standard tool set, or with the Generalized Linear Modeling (GENMOD) extension node (distributed through a SAS consulting engagement and not generally available as production-supported code from SAS), can be done within this flow diagram. The GENMOD extension node provides a graphical user interface to the power of SAS/STAT and the GENMOD procedure and presents the results in graphical output. The process flow in Figure 10 presents a typical rating model where the severity model predicts the mean claim severity for a prospect and the frequency model predicts the mean number of claims per exposure for a prospect.

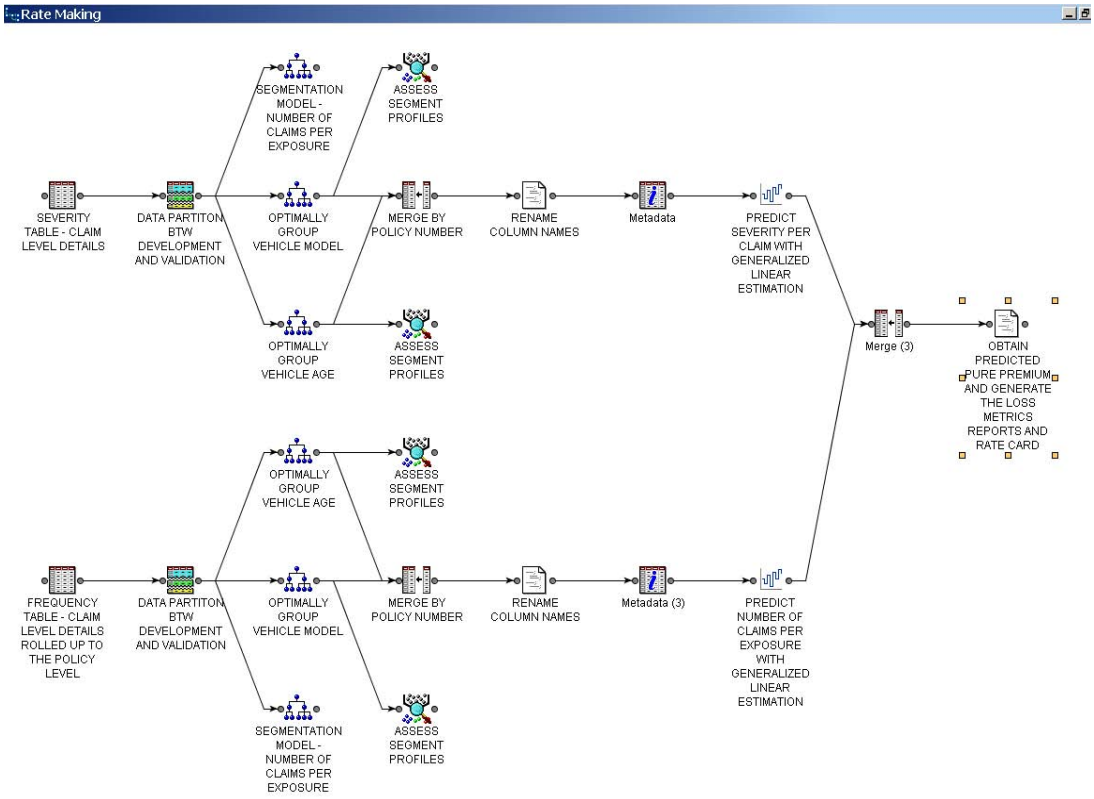


Figure 10: Example SAS Enterprise Miner Rate-Making Process Flow

At this point, it is important to specify the distribution, link function, significance levels, and confidence levels. Screen shots of the user interface for these settings are shown in Figure 11.

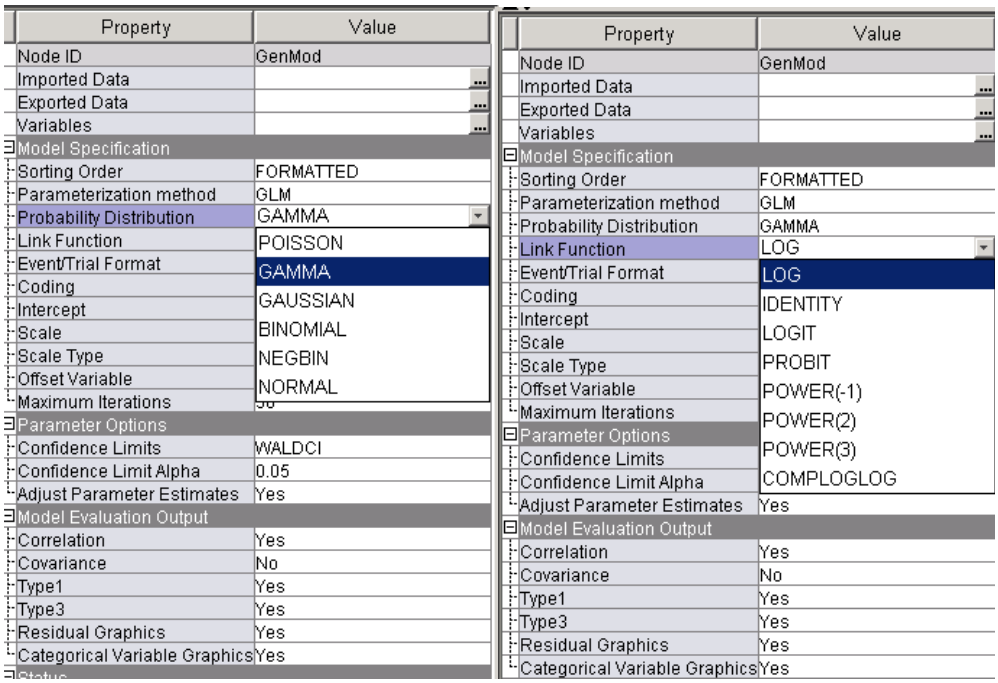


Figure 11: GENMOD Extension Node Properties Sheet

b. Using zero-inflated Poisson regression models with SAS PROC NLMIXED

As stated earlier, a common motivation for fitting ZIP models is that count data often exhibits more zeros than would be expected under standard Poisson model assumptions. It is one common source or form of overdispersion. The Poisson model, with only one mean, or rate parameter, is often unable to adapt to situations where there is a larger fraction of zero counts.

ZIP models can also be viewed as a particular type of finite mixture models, a two-component Poisson model for counts where a standard Poisson probability mass function (PMF) is combined with a discrete probability mass point at zero to handle 'extra' zeros (that is, extra relative to a standard Poisson). Engineers and physicists call that point mass a *Dirac delta function* (Wikipedia 2007). The ZIP PMF is:

$$f(y) = p0*F_1(y) + (1 - p0)*F_2(y), \qquad y = 0,1,2,3\ldots$$

where $F_1(y)$ is a Dirac delta distribution and $F_2(y)$ is a Poisson distribution. Another way to look at this is that y is distributed as a weighted average of two Poisson distributions, one of which is a degenerate distribution that has all of the probability lumped at zero. As additional probability or mass is added in the spike at zero, the same amount of probability or mass must be taken away or subtracted from the second nondegenerate Poisson PMF. Zeros have two sources, the standard Poisson and an additional Poisson source:

$$Y \sim p0 + (1 - p0)* F_2(Y=0) \qquad \text{for } y=0, \text{ and}$$

$$Y \sim \quad (1 - p0)* F_2(Y=y) \qquad \text{for } y=1,2,3\ldots$$

The Poisson distribution is as follows:

$$y = \frac{e^{-\lambda}\lambda^{y}}{y!}$$ or `exp(-lamda)*lambda**y/fact(y)`

Here, $F_2(Y=0)$ is as follows:

```
Y ~ p0 + (1 – p0)*exp(-lamda)*lambda**y/fact(y)       substituting (y=0),
Y ~ p0 + (1 – p0)*exp(-lamda)*lambda**0/fact(0)       hence
Y ~ p0 + (1 – p0)*exp(-lamda)                         for y=0
Y ~      (1 – p0)*exp(-lamda)*lambda**y/fact(y)       for y=1,2,3…
```

Combining both together a little more compactly, you have the following:

```
Y ~ (y=0)*p0 + (1 – p0)* exp(-lamda)*lambda**y/fact(y)   for y=1,2,3…
```

In the SAS NLMIXED procedure one can use the following code to model the ZIP probability:

```
Prob =  (y=0)*p0 + (1 – p0)* exp(-lamda)*lambda**y/fact(y);
```

Substituting using one of the many predefined probability distribution functions achieves the same result:

```
Prob = (y=0)*p0 + (1 – p0)* pdf('poisson',  y, lambda) ;
```

Here, p0 directly represents a zero-inflation factor. Occasionally, one will see in the literature a model formulation that is reversed, where the probability p0 is for the positive portion of the model. A handy mnemonic, or memory aid, is to remember that in Lambert's (1992) original formulation, she was interested in estimating the probability of manufacturing defect-free silicon chips for computers and how many manufacturing defects were on each chip containing defects.

Currently, no extension node uses the NLMIXED procedure. This procedure can be easily called using a SAS Code node in SAS Enterprise Miner. Code nodes allow users to integrate any licensed SAS procedures, or customized DATA steps, into the process flow diagram. Once the standard parameter settings for running a procedure and the standard output necessary for evaluation are determined, an extension node can be easily developed using SAS best practices, which can be learned in the SAS training class, *Extending SAS Enterprise Miner 5* (support.sas.com/training/us/crs/pmex.html).

10. Validating results by statistical parameters

Validating modeling results in order to determine model adequacy cannot be underestimated. The GENMOD extension node provides the standard output of the GENMOD procedure, including goodness of fit tests, analysis of parameter estimates, Type 1 and Type 3 analyses, correlation matrices, and covariance matrices (including heat maps for quick graphical analysis). As a fully integrated extension node, the GENMOD node also generates standard SAS Enterprise Miner fit statistics and graphics for lift and gains charts.

Residual graphics can also be generated by the GENMOD node: Plots of Cooks' D versus the predicted values; standardized residuals versus predicted values (see Figure 12, for an example); observed values versus predicted values; and so on. Graphics that are not generated by default can be created by using the plot wizard within SAS Enterprise Miner.

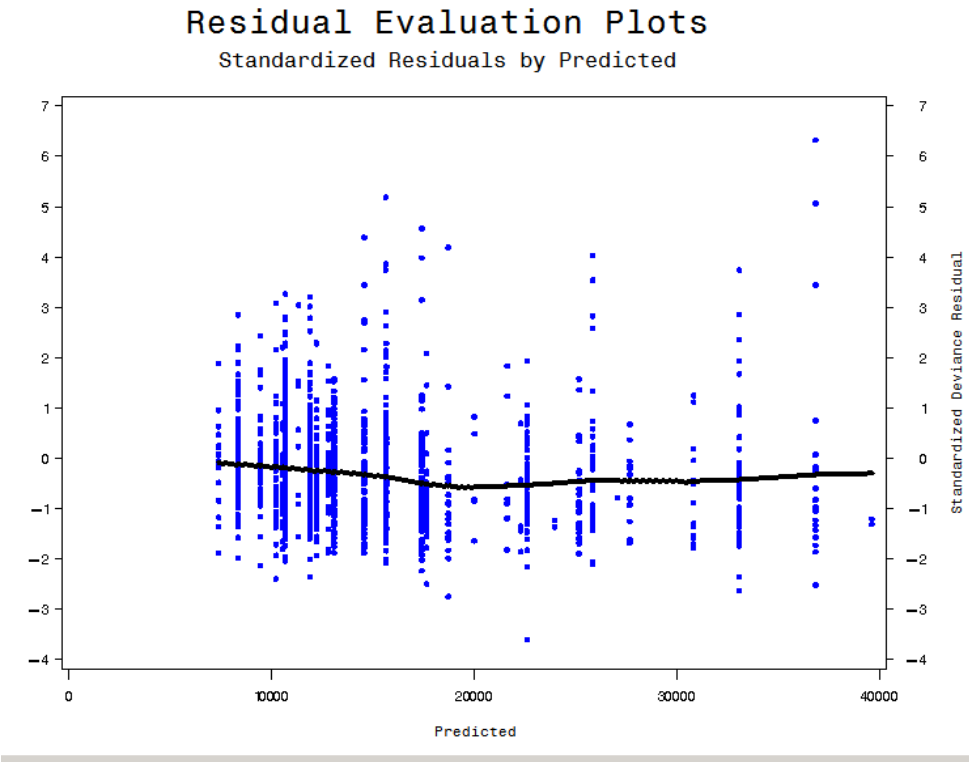


Figure 12: Example of Residual Graphics from the GENMOD Extension Node

The GENMOD extension node also presents the factor relativities graphically (see Figure 13).

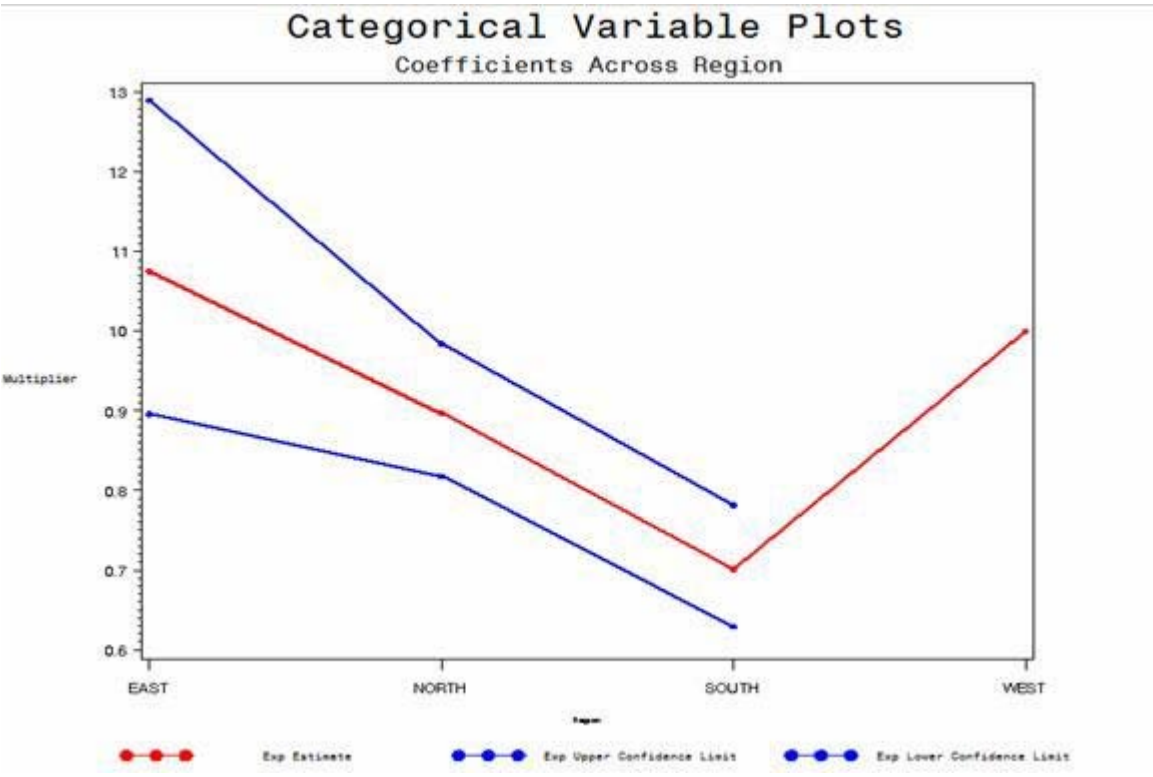


Figure 13: Example of Factor Relativities Plot

11. Validating results by business parameters and business implications

The solvency of the business line should be confirmed by checking to see if the amount collected by using the statistically fitted pricing plan indeed covers the claim cost of that line of business. This can be achieved by

1. Scoring all the policies in the portfolio with the frequency and severity model to determine pure premium, and
2. Checking to see if the premium thus collected would actually be equal to the claim cost on that portfolio.

This activity is aided by the automatic generation of scoring code in SAS Enterprise Miner. All activities in the process flow generate SAS scoring code, including the GENMOD extension node. The scoring code is SAS DATA step code that can be used to score new data. It should be noted that the standard SAS Enterprise Miner nodes also produce C and JAVA score code and, in some instances, PMML code (see the SAS Enterprise Miner documentation for the complete list of nodes that generate PMML). The GENMOD extension node only produces SAS code, however.

Almost everyone in an organization has an appreciation for and judgment of differential pricing. The frequency and severity relativities should be well-received by users and management for the rate-making recommendation to be acceptable. The relativities derived from generalized linear models are multivariate in nature and empirically valid. The base level is chosen arbitrarily and can be changed proportionally.

Inevitably a new “challenger” model will be compared against the current “champion” model. Comparing the challenger and champion rating plan should be done in order to determine if the challenger model is better and whether it can achieve the desired strategic objective stated at the beginning of the exercise.

12. Documenting, managing versions, and transferring knowledge

Rate-making initiatives bear huge returns in the medium to long run; however, longevity and repeatability of the rate-making project hinges on the softer aspects of project management. The importance of documentation and version management cannot be neglected, especially when operating in a heavily regulated market.

SAS Enterprise Miner provides the user with the ability to store projects, such as the one in Figure 10, in model packages or XML format. *Model packages* store all information about the model, including all logs and output. The XML format is a more compact way of sharing projects that store only the project layout, and provides a starting point for the next project or for sharing knowledge. SAS has also launched SAS® Model Manager to address the important needs of organizations that must provide audit trails of the model-building process and version control of implemented models.

13. Determining the acceptability of the rating through scenario analysis, alternative pricing plans, and business overrides

The overriding objective of the rate-making exercise is to ensure solvency, differential pricing, and underwriting profitability. Though there is only one statistically optimal solution for a given regression, there are a number of different settings for running a regression and also a number of thresholds for optimality. Business executives will only have faith in the recommendations of the actuary after trying out different settings and thresholds for the various factors. With SAS Enterprise Miner, the user has the ability to override the optimal segmentation in the decision tree. This proves to be a vital element in the overall acceptability of the rating recommendation across the organization.

14. Putting the model into an operational system

To close the loop, the rating logic or rules discovered in the mining exercise must be operationalized in the underwriting system to score prospective customers. There are different levels of sophistication for scoring from which an organization may choose. It is critical to choose an appropriate mix of sophistication and operational ease when deploying the models. SAS Enterprise Miner provides the capability to score within its own environment, within Base SAS® (for example, in batch or called as a stored process), or in a non-SAS operational environment by providing the JAVA or C code for scoring the corresponding model.

The complexity of the model, the nature of the operational system, the frequency of update, and the availability of skills involved have a significant bearing on the implementation. Both SAS Data Integration Studio and SAS Enterprise Guide have add-in tools that can read SAS Enterprise Miner model packages and automatically deploy the SAS scoring code, thereby dramatically reducing the time to implementation.

15. **Driving organization culture and communication**

Finally, because the price of the product impacts customer retention, customer acquisition, sales bonuses, reserve policies, cross-sell policies, channel management policies, and a number of other key performance indicators, it is immensely important to communicate, involve, and coordinate with all stakeholders. A successful rate-making initiative needs buy-in from all involved parties.

CONCLUSION

Risk-based pricing gives a definite competitive advantage. Rate making is a strategic area, but it's also a technical and cross-functional subject. It's important that it is not seen as an arcane research project. Everyone needs to understand its importance, and the entire organization should leverage it for competitive advantage. Given the right business definitions and organizational commitment, SAS can ensure the success and returns on the initiative.

REFERENCES

Lambert, D. 1992. Zero-inflated Poisson Regression with an Application to Defects in Manufacturing. *Technometrics*. 34:1–14.

ACKNOWLEDGMENTS

The authors thank Dinkar Sathe, Sambhaji Shedale, David West, Wayne Thompson, and Matt Flynn for their contributions to this paper.

RECOMMENDED READING

Karp, Andrew. 2002. "Advanced Tips and Techniques with PROC MEANS". Proceedings of the Twenty-seventh Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc. Available at www2.sas.com/proceedings/sugi27/p018-27.pdf.

Murphy, Karl, Michael Brockman, and Peter Lee. 2000. "Using Generalized Linear Models to Build Dynamic Pricing Systems for Personal Lines Insurance". Casualty Actuarial Society Winter Forum. Casualty Actuarial Society; Landover, MD. Available at www.casact.org/pubs/forum/00wforum/00wf107.pdf.

Wikipedia authors. 2007. Dirac Delta Function. Wikipedia The Free Encyclopedia (accessed March 14, 2007). Available at en.wikipedia.org/wiki/Dirac_delta_function.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:

Hrushikesh Khopkar
SAS Institute Inc.
E-mail: hrushikesh.khopkar@sas.com
hrushikesh.khopkar@gmail.com

Jay King
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
Work Phone: 919-531-1161
E-mail: jay.king@sas.com