IBM Research Report RC-21483

Probabilistic Estimation Based Data Mining for Discovering Insurance Risks

C. Apte, E. Grossman, E. Pednault, B. Rosen, F. Tipu, and B. White
T.J. Watson Research Center
IBM Research Division
Yorktown Heights, NY 10598

September 13, 1999

Abstract

The UPA (Underwriting Profitability Analysis) application embodies a new approach to mining Property & Casualty (P&C) insurance policy and claims data for the purpose of constructing predictive models for insurance risks. UPA utilizes the ProbE (Probabilistic Estimation) predictive modeling data mining kernel to discover risk characterization rules by analyzing large and noisy insurance data sets. Each rule defines a distinct risk group and its level of risk. To satisfy regulatory constraints, the risk groups are mutually exclusive and exhaustive. The rules generated by ProbE are statistically rigorous, interpretable, and credible from an actuarial standpoint. Our approach to modeling insurance risks and the implementation of that approach have been validated in an actual engagement with a P&C firm. The benefit assessment of the results suggest that this methodology provides significant value to the P&C insurance risk management process.

Keywords
rule-based predictive modeling, insurance risk modeling

1 Introduction

The Property & Casualty (P&C) insurance business deals with the insuring of tangible assets, e.g. cars, boats, homes, etc. The insuring company evaluates the risk of the asset being insured taking into account characteristics of the asset as well as the owner of the asset. Based on the level of risk, the company charges a certain fixed, regular premium to the insured. Actuarial analysis of policy and claims data plays a major role in the analysis, identification, and pricing of P&C risks.

Actuaries develop insurance risk models by segmenting large populations of policies into predictively accurate risk groups, each with its own distinct risk characteristics. A well-known segment is male drivers under age 25 who drive sports cars. Examples of risk characteristics include mean claim rate, mean claim severity amount, pure premium (i.e., claim rate times severity), and loss ratio (i.e., pure premium over premium charged). Pure premium is perhaps the most important risk characteristic because it represents the minimum amount that policyholders in a risk group must be charged in order to cover the claims generated by that risk group. Actual premiums charged are ultimately determined based on the pure premiums of each risk group, as well as on the cost structure of the insurance company, its marketing strategy, competitive factors, etc.

Ideally, insurance companies would like to develop risk models based on the entire universe of potential policies in order to maximize the accuracy of their risk assessments. Although no insurer possesses complete information, many insurers, particularly ones operating across large territories, have access to vast quantities of information given their very sizable *books of business*. A book of business corresponds to either a type of policy or to the set of policies of that type in a territory, depending on context. It is common for such firms to have millions of policies in each of their major regions, with many years of accumulated claims data. The actuarial departments of insurance companies make use of this data to develop risk models for the markets served by their companies. The availability of large quantities of insurance data represents both an opportunity and a challenge for data mining.

The first and perhaps most fundamental challenge for standard data mining techniques is that pure premium is the product of two other risk characteristics: claim frequency and claim severity. Claim frequency is the average rate at which individual policyholders from a risk group file for claims. Frequency is usually expressed as the number of claims filed per policy per unit time (i.e., quarterly, annually, etc.); however, it can also be expressed as a percentage by multiplying by 100. For example, a frequency of 25% means that the average number of claims filed in a given unit of time is 0.25 times the number of policies. This is not to say that 25% of policyholders file claims; only about 19.5% will file one claim in the given time period and an unlucky 2.6% will file two or more claims. Thus, the 25% refers to a rate, not a probability. Claim severity is more straightforward. It is simply the average dollar amount per claim.

If one were forced to use standard data mining algorithms, such as CHAID [1], CART [2], C4.5 [3], or SPRINT [4], one might try to view frequency modeling as a classification problem and severity modeling as a regression problem. However, further examination suggests that these modeling tasks are unlike standard classification or regression problems. Viewing frequency prediction as a classification problem is misleading. It is certainly not the case that every individual policyholder will file a claim with either 100% certainty or 0% certainty. In actuality, every individual has the potential to file claims, it is just that some do so at much higher rates than others. The predictive modeling task is therefore to discover and describe groups of policyholders, each with its own unique filing rate, rather than attempt to discover groups that are "classified" as either always filing claims or never filing claims.

From the point of view of standard data mining algorithms, severity prediction appears to be very much a regression problem, given that the data fields that correspond to this variable have continuous values across a wide range. However, the distributional characteristics of claim amounts are quite different from the traditional Gaussian (i.e., least-squares optimality) assumption incorporated into most regression modeling systems. Insurance actuaries have long recognized that the severity distribution is often highly skewed with long thick tails. Reliance on the Gaussian assumption for modeling individual claims can lead to suboptimal results, which is a well-known problem from the point of view of robust estimation [5].

A more fundamental obstacle for standard data mining algorithms is that specialized, domain-specific equations must be used for estimating frequency and severity. Equations for estimating frequency must incorporate variables that reflect the earned exposure of each data record; i.e., the amount of time that the corresponding policy is actually in force during the stated time interval. Equations for estimating severity must take into account claim status; i.e., whether a claim is fully settled or still open. Some types of claims can take several years to settle, most notably bodily injury claims. To obtain reliable risk models, all claims must be considered when estimating frequency, but only those claims that are fully settled should be used when estimating severity. Standard data mining algorithms are typically not equipped to make these distinctions, nor are they equipped to perform the necessary calculations.

A further complication for standard data mining algorithms is that insurance actuaries demand statistical rigor and tight confidence bounds on the risk parameters that are obtained; i.e., the risk groups must be *actuarially credible*. Actuarial credibility, which is discussed in subsequent sections, is a further requirement that standard data mining algorithms are ill-equipped to handle because they are not designed to perform the necessary calculations to ensure that only actuarially credible risk groups are identified.

The above challenges have motivated our own research [6, 7] and have lead to the development of the IBM ProbE$^{\mathbf{TM}}$ (Probabilistic Estimation) predictive modeling kernel. This C++ kernel embodies several innovations that address the challenges posed by insurance data. The algorithms are able to construct rigorous rule-based models of insurance risk, where each rule represents a risk group. The algorithms differ from standard data mining algorithms in that the domain-specific calculations necessary for modeling insurance risks are not only integrated into the algorithms, they are in fact used to help guide the search for risk groups.

The IBM UPA$^{\mathbf{TM}}$ (Underwriting Profitability Analysis) application [8] is built around ProbE and provides the infrastructure for using ProbE to construct rule-based risk models. UPA was designed with input from marketing, underwriting, and actuarial end-users. The graphical user interface is tailored to the insurance industry for enhanced ease of use. Innovative features such as sensitivity analysis help in evaluating the business impact of rules. An iterative modeling paradigm permits discovered rules to be edited and the edited rules to be used as seeds for further data mining. In a joint development project with a P&C company, the UPA solution amply demonstrated the value that a discovery-driven approach can bring to the actuarial analysis of insurance data.

2 Statistical Modeling of Insurance Risks

Some of the features incorporated into ProbE were strongly influenced by the mathematical rigor with which actuaries approach the problem of modeling insurance risks. Actuarial science is based on the construction and analysis of statistical models that describe the process by which claims are filed by policyholders (see, for example, [9]). Different types of insurance often require the use of different statistical models. The statistical models that are incorporated into the current version of ProbE are geared toward property and casualty insurance in general, and automobile insurance in particular.

For any type of insurance, the choice of statistical model is strongly influenced by the fundamental nature of the claims process. For property and casualty insurance, the claims process consists of claims being filed by policyholders at varying points in time and for varying amounts. In the normal course of events, wherein claims are not the result of natural disasters or other widespread catastrophes, loss events that result in claims (i.e., accidents, fire, theft, etc.) tend to be randomly distributed in time with no significant pattern to the occurrence of those events from the point of view of insurance risk. Policyholders can also file multiple claims for the same type of loss over the life of a policy. As illustrated in Figure 1, these properties are the defining characteristics of Poisson random processes [9]. ProbE thus uses Poisson processes to model claim filings.

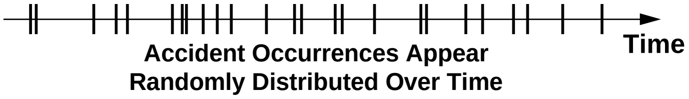

Figure 1: Event Stream of a Poisson Random Process

In addition to modeling the distribution of claims over time, actuaries must also model the amounts of those claims. In actuarial science, claim amounts for property and casualty insurance are modeled as probability distributions. Two kinds of distributions are usually considered: those for the amounts of individual claims, and those for the aggregate amounts of groups of claims.

The distributions incorporated into the current version of ProbE were selected based on an examination of actual historical automobile claims data. The claim amounts were found to have a highly skewed distribution. Most claims were small in value relative to the maximum amounts covered by the policies, but a significant proportion of large claims were also present. When the claim amounts were log transformed, the skewness virtually disappeared and the resulting distribution was found to be highly Gaussian in shape. As illustrated in Figure 2, these properties are the defining characteristics of log-normal distributions.

Aggregate loss distributions are not used by ProbE during data mining, but only for post-mining analysis purposes when estimating the aggregate parameters of each risk group. There is no closed-form solution for the aggregate loss distribution given that the individual loss distribution is log-normal (a sum of log-normal random variables is not itself log-normal). An approximation must therefore be made. In ProbE, the central limit theorem is invoked and the normal (i.e., Gaussian) distribution is used to model aggregate losses.

Because different distributions are used to model individual versus aggregate losses, different statistical procedures are employed for estimating the parameters of those distributions. For the log-normal distributions used to model individual losses, the relevant statistical parameters are the means and standard deviations of the natural logarithms of individual claim amounts. For the normal distributions used to model aggregate losses, the means and standard deviations of the (raw) claim amounts are the parameters that need to be estimated.

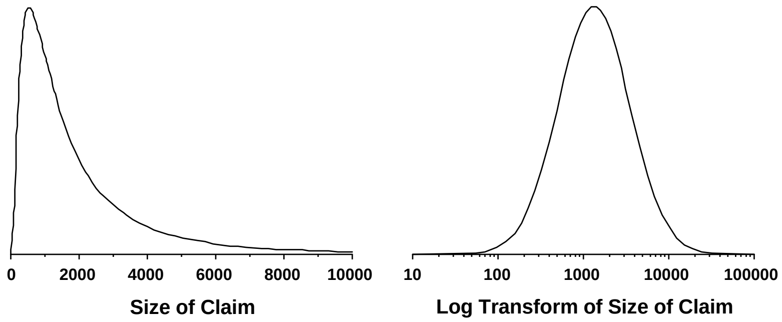

Figure 2: Density Functions for a Log-Normal Random Variable

3 Top Down Identification of Risk Groups

The traditional method used by actuaries to construct risk models involves first segmenting the overall population of policyholders into a collection of risk groups based on a set of factors, such as age, gender, driving distance to place of employment, etc. The risk parameters of each group are then estimated from historical policy and claims data. Ideally, the resulting risk groups should be homogeneous with respect to risk; i.e., further subdividing the risk groups by introducing additional factors should yield substantially the same risk parameters. Actuaries typically employ a combination of intuition, guesswork, and trial-and-error hypothesis testing to identify suitable factors. The human effort involved is often quite high and good risk models can take several years to develop and refine.

ProbE replaces manual exploration of potential risk factors with automated search. Risk groups are identified in a top-down fashion by a method similar to those employed in classification and regression tree algorithms [2, 1, 4, 3]. Starting with an overall population of policyholders, ProbE recursively divides the policyholders into risk groups by identifying a sequence of factors that produce the greatest increase in homogeneity within the subgroups that are produced. The process is continued until each of the resulting risk groups is either declared to be homogeneous or is too small to be further subdivided from the point of view of actuarial credibility.

One of the key differences between ProbE as embodied in the UPA and other classification and regression tree algorithms is that splitting factors are selected based on statistical models of insurance risks. In the case of UPA, a joint Poisson/log-normal model is used to enable the simultaneous modeling of frequency and severity and, hence, pure premium. The joint Poisson/log-normal model explicitly takes into account insurance-specific variables, such as earned exposure and claim status. In addition, it provides feedback to ProbE's search engine on the degree of actuarial credibility of each proposed risk group so that only those splitting factors that yield actuarially credible risk groups are considered for further exploration. By explicitly taking these aspects of the problem into account, ProbE is able to overcome the major barriers that cause standard data mining algorithms to be suboptimal for this application.

4 The Joint Poisson/Log-Normal Model

The optimization criterion used to identify splitting factors is based on the principles of maximum likelihood estimation. Specifically, the negative log-likelihood of each data record is calculated assuming a joint Poisson/log-normal statistical model, and these negative log likelihoods are then summed to yield the numerical criterion that is to be optimized. Minimizing this negative log-likelihood criterion causes splitting factors to be selected that maximize the likelihood of the observed data given the joint Poisson/log-normal models of each of the resulting risk groups.

Historical data for each policy is divided into distinct time intervals for the purpose of data mining, with one data record constructed per policy per time interval. Time-varying risk characteristics are then assumed to remain constant within each time interval; that is, for all intents and purposes their values are assumed to change only from one time interval to the next. The choice of time scale is dictated by the extent to which this assumption is appropriate given the type of insurance being considered and the business practices of the insurer. For convenience, quarterly intervals will be assumed to help make the discussion below more concrete, but it should be noted that monthly or yearly intervals are also possible

Assuming that data is divided into quarterly intervals, most data records will span entire quarters, but some will not. In particular, data records that span less than a full quarter must be created for policies that were initiated or terminated mid-quarter, or that experienced mid-quarter changes in their risk characteristics. In the case of the latter, policy-quarters must be divided into shorter time intervals so that separate data records are created for each change in the risk characteristics of a policy. This subdivision must be performed in order to maintain the assumption that risk characteristics remain constant within the time intervals represented by each data record. In particular, subdivision must occur is when claims are filed under a policy in a given quarter because the filing of a claim can itself be an indicator of future risk (i.e., the more claims one files, the more likely one is to file future claims). The actual time period covered by a database record is the earned exposure of that record.

Figure 3 depicts the database records that are constructed as a result of subdivision. In this figure, Q0, Q1, Q2, etc., represent the ending days of a sequence of quarters. T0 represents the day on which a particular policy came into force, while T1 represents the day the first claim was filed under that policy. Though not illustrated, T2, T3, T4, etc., would represent the days on which subsequent claims were filed. For data mining purposes, the policy claims data is divided into a sequence of database records with earned exposures t1, t2, t3, etc. As illustrated, new policies typically come into force in the middle of quarters. Thus, the earned exposure for the first quarter of a policy's existence (e.g., t1) is generally less than a full quarter. The earned exposures for subsequent quarters, on the other hand, correspond to full quarters (e.g., t2, t3, and t4) until such time that a claim is filed, the risk characteristics change mid-quarter, or the policy is terminated. When a claim is filed or the risk characteristic changes, the data for that quarter is divided into two or more records. The earned exposure for the first database record (e.g., t5) indicates the point in the quarter at which the claim was filed. The earned exposure for the second record (e.g., t6) indicates the time remaining in the quarter, assuming only one claim is filed in the quarter as illustrated in the diagram. If two or more claims are filed in the quarter, then three or more database records are constructed: one record for each claim and one record for the remainder of the quarter (assuming that the policy has not been terminated). Likewise for other changes in risk characteristics, such as adding or removing drivers, cars, etc., from the policy.

For Poisson random processes, the time between claim events follows an exponential distribution. Moreover, no matter at what point one starts observing the process, the time to the next claim event has

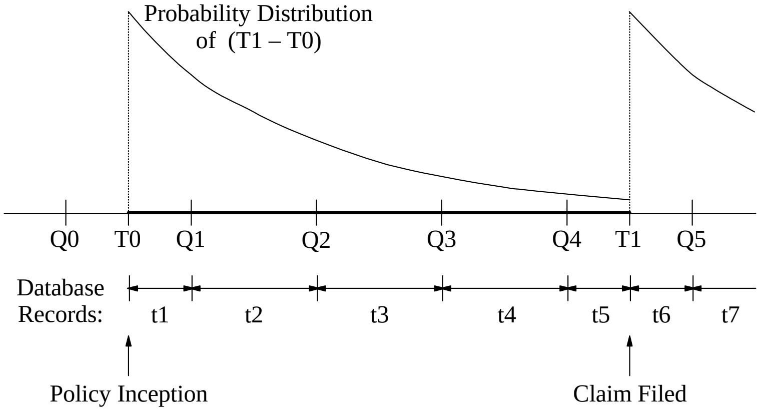

Figure 3: The Effect of Subdividing Policy-Quarters in which Claims are Filed.

the same exponential distribution as the time between claim events. For the example shown in Figure 3, the probability density function for the time between the policy inception and the first claim being filed is given by

$$f(\mathrm{T1} - \mathrm{T0}) = \lambda e^{-\lambda(\mathrm{T1}-\mathrm{T0})} = \lambda e^{-\lambda(\mathrm{t1}+\mathrm{t2}+\mathrm{t3}+\mathrm{t4}+\mathrm{t5})}\ , \tag{1}$$

where λ is the claim frequency of the risk group. More generally, from the additivity properties of Poisson random processes, it can be shown that the probability density for the time T (i.e., the total earned exposure) between $k + l$ claim filings (where k is the number of settled claims and l is the number of open claims) is given by

$$f(T \mid k + l) = \lambda^{k+l} e^{-\lambda T}\ . \tag{2}$$

The maximum likelihood estimate used by ProbE for the frequency parameter λ is thus the same one that is typically used by actuaries for estimating frequency:

$$\hat{\lambda} = \frac{k + l}{T} = \frac{\text{Total Number of Claims}}{\text{Total Earned Exposure}}\ . \tag{3}$$

In the case of claim amounts, the joint probability density function for the severities $s_1, \ldots, s_k$ of k settled claims is given by:

$$f(s_1, \ldots, s_k) = \frac{1}{\prod_{i=1}^{k} \sqrt{2\pi}\,\sigma_{\log}\, s_i} \cdot e^{-\dfrac{\sum_{i=1}^{k}(\log(s_i) - \mu_{\log})^2}{2\sigma_{\log}^2}}\ . \tag{4}$$

The estimates of the mean log severity $\mu_{\log}$ and the variance of the log severity $\sigma_{\log}$ are likewise the ones typically used for log-normal distributions:

$$\hat{\mu}_{\log} = \frac{1}{k}\sum_{i=1}^{k} \log(s_i) \tag{5}$$

and

$$\hat{\sigma}^2_{\log} = \frac{1}{k-1}\sum_{i=1}^{k}(\log(s_i) - \hat{\mu}_{\log})^2 \ . \tag{6}$$

Equations 5 and 6 are used during training to estimate the parameters of the severity distribution for individual claims. These estimators presume that the individual severity distributions are log-normal. The usual unbiased estimators for the mean and variance of severity are used after data mining has been completed to estimate the parameters of the aggregate severity distribution:

$$\hat{\mu} = \frac{1}{k}\sum_{i=1}^{k} s_i \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{k-1}\sum_{i=1}^{k}(s_i - \hat{\mu})^2 \ . \tag{8}$$

Only fully settled claims are considered when applying Equations 5-8. The severity fields of unsettled claims are often used to record reserve amounts; i.e., the money that insurers hold aside to cover pending claims. Reserve amounts are not actual losses and therefore are not used to develop models for predicting actual losses.

As mentioned earlier, negative log-likelihoods are calculated for each database record in a risk group based on Equations 2 and 4. The nonconstant terms in the negative log-likelihoods are then summed and used as the criterion for selecting splitting factors in the top-down identification of risk groups. The constant terms do not contribute to the selection of splitting factors and, hence, are omitted to reduce the amount of computation.

With constant terms removed, the negative log-likelihood score for the ith database record is:

$$\xi_i = \begin{cases} \lambda t_i & \text{for non-claim records} \\ \lambda t_i + \log(\frac{\sigma_{\log}}{\lambda}) & \text{for open claim records} \\ \lambda t_i + \log(\frac{\sigma_{\log}}{\lambda}) + \frac{(\log(s_i) - \mu_{\log})^2}{2\sigma^2_{\log}} & \text{for settled claim records,} \end{cases} \tag{9}$$

where t_i is the earned exposure for the ith record. Note that the Poisson portion of the model contributes an amount $\lambda t_i + \log(1/\lambda)$ to the score of each claim record and an amount λt_i to the score of each non-claim record. The sum of these values equals the negative logarithm of Equation 2. The log-normal portion of the model contributes nothing to the scores of non-claim records, and an amount $\log(\sigma_{\log}) + (\log(s_i) - \mu_{\log})^2/(2\sigma^2_{\log})$ to the score of each settled claim record. The sum of these values equals the negative logarithm of Equation 4 with constant terms (i.e., $\sum_{i=1}^{k}\log(\sqrt{2\pi}\, s_i)$) removed. In the case of open claim records, an expected value estimate of the log-normal score is constructed based on the scores of the settled claim records. After dropping constant terms from this expected value estimate, open claim records contribute an amount $\log(\sigma_{\log})$ to the log-normal portions of their scores.

If the database records for a risk group contain k settled claims and l open claims, then the sum of the above scores is given by:

$$\xi = \lambda\left(\sum_{i=1}^{N} t_i\right) + (k+l)\log\left(\frac{\sigma_{\log}}{\lambda}\right) + \left(\frac{1}{2\sigma^2_{\log}}\right)\sum_{i=1}^{k}(\log(s_i) - \mu_{\log})^2 \ . \tag{10}$$

In the above equation, N is the total number of database records for the risk group, the first k of which are assumed for convenience to be settled claim records. Equation 10 is then summed over all risk groups to yield the overall score of the risk model. The top-down procedure described in the previous section identifies risk groups by minimizing the overall score in a stepwise fashion, where each step involves dividing a larger risk group into two smaller risk groups so as to reduce the value of the overall score to the maximum extent possible.

From the point of view of data mining technology, the important thing to note about the above equations is that insurance-specific quantities such as earned exposure and claim status enter into both the equations for estimating model parameters and the equations for selecting splitting factors. Earned exposure effectively plays the role of a weighting factor, while claim status plays the role of a correction factor that adjusts for missing data in one of the two data fields to be predicted (i.e., the settled claim amount given that a claim was filed). Equation 10 essentially replaces the entropy calculations used in many standard tree-based data mining algorithms. It should be noted that entropy is, in fact, a special case of negative log-likelihood. Its calculation need not be restricted to categorical or Gaussian (least-squares) distributions. The development of the joint Poisson/log-normal model presented above illustrates the general methodology one can employ to customize the splitting criteria of tree-based data mining algorithms to take into account data characteristics that are peculiar to specific applications.

5 Actuarial Credibility

ProbE's top-down modeling procedure is constrained to produce risk groups that are actuarially credible. In actuarial science, credibility [9] has to do with the accuracy of the estimated risk parameters (in this case, frequency, severity, and ultimately pure premium). Accuracy is measured in terms of statistical confidence intervals; that is, how far can the estimated risk parameters deviate from their true values and with what probability. A fully credible estimate is an estimate that has a sufficiently small confidence interval. In particular, estimated parameter values X must be within a certain factor r of their true (i.e. expected) values $E[X]$ with probability at least p:

$$P\left\{\left|\frac{X - E[X]}{E[X]}\right| \leq r\right\} \geq p \; . \tag{11}$$

Typical choices of r and p used by actuaries are $r = 0.05$ and $p = 0.9$. In other words, X must be within 5% of $E[X]$ with 90% confidence.

To ensure that actuarially credible risk groups are constructed, ProbE permits a maximum fractional standard error to be imposed on the estimated pure premiums of each risk group. In the process of subdividing larger risk groups into smaller risk groups, ProbE only considers splitting factors that yield smaller risk groups that obey this constraint. Specifically, each risk group must satisfy the following inequality:

$$\frac{\sqrt{Var[X]}}{E[X]} \leq r' \; , \tag{12}$$

where X is the pure premium estimate of the risk group, $E[X]$ is the expected value of the pure premium, $Var[X]$ is the variance of the pure premium estimate, and r' is the maximum allowed fraction standard error. If a splitting factors that satisfies Equation 12 cannot be found for a given risk group, that risk group is declared to be too small to be subdivided and no further refinement of the risk group is performed. Actuarial credibility is ensured by the fact that, for any pair of values of p and r in Equation 11, there exists a corresponding value of r' for Equation 12 such that

$$P\left\{\left|\frac{X - E[X]}{E[X]}\right| \leq r\right\} \geq p \quad \text{if and only if} \quad \frac{\sqrt{Var[X]}}{E[X]} \leq r' \; . \tag{13}$$

In particular, if X is approximately Gaussian and $p = 0.9$, then the corresponding value for r' as a function of r is

$$r' = \frac{r}{1.645} \; . \tag{14}$$

For a 5% maximum error with 90% confidence, the corresponding value of r' would thus be 3.04%.

When applying the above credibility constraint, the mean and variance of the pure premium estimate are approximated by their empirical estimates. Thus, the fractional standard error for pure premium is approximated by

$$\frac{\sqrt{Var[X]}}{E[X]} \approx \sqrt{\frac{1}{k+l} + \frac{1}{k}\left(\frac{\hat{\sigma}^2}{\hat{\mu}^2}\right)} \ . \qquad (15)$$

Note that this fractional standard error varies as a function of the statistical properties of each risk group. The determination of when a risk group is too small to be subdivided is thus context-dependent. The ability to impose a context-dependent actuarial credibility constraint on the top-down process by which risk groups are constructed is another important feature of ProbE that distinguishes it from all other tree-based modeling methods, such as CHAID [1], CART [2], C4.5 [3], or SPRINT [4].

Equation 15 can also be used to obtain a rough estimate of the amount of data needed to justify a given number of risk groups. In general, the standard deviation of claim severity tends to be at least as large as the mean claim severity; hence, $\hat{\sigma}^2/\hat{\mu}^2 \geq 1$ in most cases. To achieve a 5% maximum error with 90% confidence, a risk group must therefore cover at least 2,164 claim records, or about 108,200 quarterly records given that the average quarterly claim rate for automobile insurance tends to be about 2%. Multiply 108,200 by the number of risk groups and it becomes quite evident that a very large number of quarterly data records must be considered in order to achieve fully credible results.

6 Predictive Accuracy

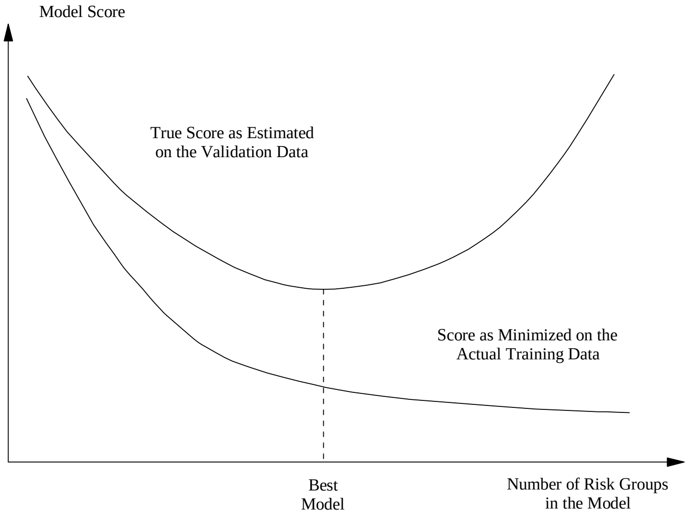

Figure 4: Mining for the Best Model

To further ensure the predictive accuracy of the risk models that are produced, ProbE also incorporates a method for avoiding overfitting. Overfitting occurs when the best model relative to the training data tends to perform significantly worse when applied to new data. In the case of the UPA, model performance is measured using the negative log-likelihood score defined by Equation 10, wherein lower scores indicate better performance. Risk groups are identified by searching for splitting factors that minimize this score with respect to the training data. However, the score can be made arbitrarily small simply by introducing enough splitting factors. As more splitting factors are introduced, a point of overfitting is reached where the value of the score as estimated on the training data no longer reflects the value that would be obtained on new data. Adding splitting factors beyond this point would simply make the model worse.

Overfitting mathematically corresponds to a situation in which the score as estimated on the training data substantially underestimates the expected value of the score that would be obtained if the true statistical properties of the data were already known. Results from statistical learning theory (see, for example, [10]) demonstrate that, although there is always some probability that underestimation will occur for a given model, both the probability and the degree of underestimation are increased by the fact that we explicitly search for the model that minimizes the estimated score. This search biases the difference between the estimated score and the expected value of the score toward the maximum difference among competing models.

To avoid overfitting, the available training data is randomly divided into two subsets: one that is used for actual training (i.e., estimation of parameters and selection of splitting factors); the other that is used for validation purposes to estimate the true performance of the model. As splitting factors are introduced by minimizing the score on the actual training data, a sequence of risk models is constructed in which each successive model contains more risk groups than its predecessors. The true score of each model is then estimated by evaluating Equation 10 on the validation data for each risk group in the model and summing the results. The model that minimizes this unbiased estimate of the true score is selected as the most accurate risk model given the available data. As illustrated in Figure 4, the introduction of each successive splitting factor simultaneously increases the number of risk groups and decreases the score of the risk model on the actual training data. The most suitable risk model is the one with the smallest score on the validation data.

7 The UPA Solution

The UPA solution consists of the UPA application and a methodology for processing P&C policy and claims data using the application. The UPA application is a client-server Java-based application. On the server side, the ProbE C++ data mining kernel is used for actual execution of mining tasks. The client-server implementation is multithreaded and a process scheduling subsystem on the server manages and synchronizes requests for ProbE runs that may flow in from any of the clients. Results of mining are available in various graphical and tabular formats, some of which may require a business analyst to interpret while others can be directly interpreted by a business decision maker.

In preparation for mining, company policy and claims data may be combined with exogenous data, such as demographics, and stored as a set of records. Each record is essentially a snapshot of a policy during an interval of time, including any claim information. Trend information is captured in a set of derived fields. The application is geared to predict pure premium, which is the product of claim frequency and claim severity. Though not explicitly present in the raw data, it is readily computed once mean frequency and mean severity have been estimated.

The user has control over three distinct phases in the mining process.

1. *Training* is the process in which the application discovers the statistically significant subpopulations that exist in the data.

2. *Calibration* is the process in which the application applies a second data set to the rules discovered in the training phase, and calibrates the statistics associated with each rule, such as claim rate, claim amount, and pure premium.

3. *Evaluation* permits a user to evaluate the rules on yet another data set to confirm the actuarial credibility of the calibrated rules.

The training, calibration, and test data sets are constructed so as to be disjoint (i.e., they have no records in common). This is necessary to ensure the statistical reliability of the rules and subsequent analysis. Both the calibration and test data sets are obtained by randomly sampling the entire data set that was constructed for analysis. The claim rates and severities measured on the calibration and test data sets therefore reflect the rates and severities of the entire data set. The training data set, on the other hand, is a stratified random sample in which the proportion of claim to non-claim records is greater than in the entire data set.

The distinction between training and calibration does not generally exist in other data mining algorithms. The distinction was made in the UPA in order to satisfy the demand for actuarial credibility while simultaneously keeping the computational requirements to a reasonable level. Because ProbE makes many passes over the training data in the process of identifying risk groups, it is desirable from a computational standpoint to keep the size of the training set to a minimum. However, actuarial credibility demands large quantities of data. Instead of attempting to achieve full credibility on a large amount of training data, the compromise made in the UPA solution is to achieve a weaker level of credibility on a smaller amount of training data, but to then use a much larger quantity of calibration data to re-estimate the model parameters of each risk group identified during training in order to achieve fully credible parameter estimates. The fractional standard error of pure premium that needs to be achieved on the training data in order to achieve a desired fractional standard error of pure premium on the calibration data is given by the following equation:

$$r'_{\text{training}} = r'_{\text{calibration}} \sqrt{\frac{\text{Number of Claim Records in the Calibration Set}}{\text{Number of Claim Records in the Training Set}}} \; . \tag{16}$$

Another compromise made in the UPA solution is to stratify the training data by randomly excluding a large percentage of non-claim records. Stratification can dramatically reduce the size of the training set; for example, in the case of quarterly automobile data, 90% of the quarterly non-claim records can be removed with minimal impact on the predictive accuracy of the resulting risk groups. Stratification is justified by the fact that its effect is to nonlinearly rescale all estimated claim frequencies, and this nonlinear rescaling can be accounted for in Equations 3 and 10 by linearly scaling the earned exposures of the remaining non-claim records in inverse proportion to the fraction of non-claim records that remain. Thus, if 90% of non-claim records were removed, with 10% remaining, then the effect of stratification can be mathematically compensated by dividing the earned exposures of all remaining non-claim records by 0.10. The earned exposures of the claim records, on the other hand, would remain the same. By scaling the earned exposures in this fashion, the resulting estimated claim frequencies given by Equation 3 and the resulting model scores given by Equation 10 would then be the same, to within estimation error, as those obtained without stratification. Any differences in these values, and hence any differences in the choice of splitting factors, would be entirely due to the sampling noise introduced through stratification. Note that stratification has no effect on the determination of actuarial credibility because Equation 15 depends only on the claim records that are present.

Mining runs produce risk models that are represented as collections of rules. A typical rule is illustrated below:

```
RULE #22
IF
Field "VANTILCK" "Vehicle Antilock Break Discount?"
= "Antilock Brake"
Field "VEHTYPE" "Type of Vehicle"
= "Truck"
THEN
claim rate        0.0115561
```

```
mean severity     5516.84
std dev severity  11619.9
pure premium      63.753
loss ratio        0.688204
608 training claims out of 53221 training points
```

Several statistics are reported for each rule, including claim rate, mean severity, standard deviation of the severity, pure premium (i.e., claim rate times severity), and loss ratio (i.e., pure premium over premium charged). Two additional statistics that are reported for each rule are the number of total examples that match the rule and the total number of those examples that are claim-related. In the case of the rule illustrated here, 53,221 examples matched the rule, out of which 608 had incurred claims.

The risk models produced by ProbE can be used as the basis for establishing new price structures for the premiums charged to policyholders. In addition, the models can be analyzed to uncover nuggets; i.e., previously unknown risk factors that, if incorporated into existing price scenarios, could improve overall profitability.

8 Uncovering nuggets in the rules

The first step in uncovering nuggets begins with the lift charts that are generated from a mining run. A typical UPA lift chart is displayed in Figure 5. The X-axis is a cumulative percentage count of the policies, sorted in order of decreasing predicted pure premium. The values therefore range from 0 to 100. The Y-axis is the cumulative percentage of actual premiums collected from, or actual claims paid to the policyholders in the order defined by the X-axis. The Y-axis therefore also ranges from 0 to 100. The chart displays three plots. The first plot is that of a hypothetical situation, in which a uniform premium is collected for each policy. This essentially represents the scenario in which an insurance firm has no insight about its policies and spreads its risk uniformly across the entire pool. The second plot displays firm's current actual premium pricing. This plot shows the actual cumulative premiums collected for the policies when sorted in descending order by predicted pure premium. The third plot displays the scenario proposed by the UPA in which the UPA-recommended pricing is plotted (which is the actual cumulative claim amounts for the policies when sorted in descending order by predicted pure premium).

In our experience, the relationships among the curves shown in Figure 5 are commonly encountered in practice. Actuaries have identified many distinct risk groups and their characteristics have been incorporated into the premiums charged. However, as the lift chart illustrates, the UPA solution has a strong likelihood of discovering previously unknown risk groups and is therefore able to suggest more competitive prices in many situations.

The lift charts provide a quick visual indication of whether a detailed analysis of mining results will uncover any nuggets. If an actual mining run results in a lift chart very similar to the one illustrated in Figure 5, then the business analyst has a basis for continuing further investigations of the rules. If the lift chart indicates very little or no difference between actual pricing and the UPA-proposed pricing, then further investigation would likely have little business value.

To uncover nuggets, the analyst needs to first understand the statistics for the entire book of business. The UPA application can present these *universal* statistics to the user:

```
for "Accs This Qtr Ult $ BI+PD"
claim rate 0.00600882
mean severity     4676.55
std dev severity  9165.3
```

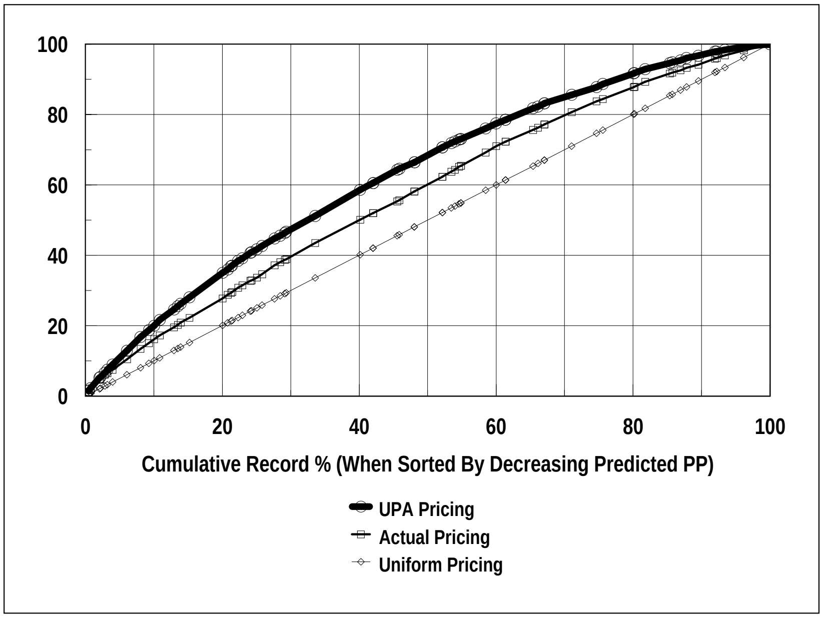

Figure 5: UPA Lift Chart

```
pure premium      28.1006
loss ratio        0.315589
3958 training claims out of 662656 training points
```

This particular book has 662,656 records, of which 3,958 records actually had claims. The claim rate for this database is 0.006, with a mean severity of $4676 and a standard deviation of $9165. The average pure premium is $28, and the loss ratio is 0.31. Using these overall statistics as a basis, an analyst needs to look for rules that predict pure premiums and/or loss ratios that differ significantly from the overall average, and still cover a sizable number of policies to be both actuarially credible and interesting from a marketing standpoint. This latter threshold will vary from business to business, and the end-user may use discretionary judgment in identifying nuggets utilizing the above criteria. For example, the rule illustrated previously has a predicted pure premium of $63, a loss ratio of 0.69, and matches 8.0% of the policies. It therefore represents a potential nugget using these criteria. The final check to confirm the quality of a nugget is to examine how these candidate rules hold up when applied to unseen data; i.e., when they are evaluated for their predictive accuracy.

The above methodology was employed in a joint development project with a major insurance company in North America. Automobile insurance data for 16 quarters from the books of business in a single state were extracted and transformed into a data mart. The mart represented about 2 million policies and was approximately 30 GB in size. The data consisted of three major books of business: preferred, high risk and standard. There were over 250 explanatory data fields, comprising demographic, agency, vehicle, and policyholder information. In addition, there were several different types of coverages to be modeled, including bodily injury (BI), property damage (PD), comprehensive (Comp), and collision (Coll).

The books of business, the different variable groups, and the different coverages could be combined in many different ways for the purpose of mining. After consulting with the firm's actuaries and marketing analysts, mining runs were conducted for 18 unique combinations of books of business, explanatory variables, and coverages. Each run generated about 40 rules. From this collection of rules, 43 nuggets were identified using the methodology described in this paper. Six of these nuggets were selected by the insurer for a detailed benefits assessment study using the insurer's internal methodology for evaluating proposed changes to their pricing and/or underwriting practices. The benefits assessment study indicated that implementing just these 6 nuggets in a single state could potentially realize a net profit gain of several million dollars. The benefits that could be realized by scaling up the business implementation of all 43 nuggets across multiple states are clearly appealing.

One of the six nuggets has already been widely publicized in the media. While it is well-known among insurers that drivers of high-performance sports cars are more likely to have accidents than are other motorists, the UPA discovered that if the sports car was not the only vehicle in the household, then the accident rate is not much greater than that of a regular car. In one estimate [11], "just letting Corvettes and Porsches into [the insurer's] 'preferred premium' plan could bring in an additional $4.5 million in premium revenue over the next two years without a significant rise in claims." Another publicly disclosed nugget relates to experienced drivers, who tend to have relatively low claim frequencies. However, the UPA turned up a particular segment of experienced drivers who are unusually accident prone.

9 Evaluation

In addition to data mining runs that were performed for the purpose of uncovering nuggets, runs were also performed to assess the UPA's ability to identify distinct risk groups as a function of the amount of training data provided, as well as to assess the predictive accuracy of the risk models produced by the UPA versus those obtained using other data mining technologies.

Figure 6 shows an example of the relationship among lift curves that was observed as the amount of training data was varied from 43 thousand records to 1.38 million records. As this figure illustrates, increasing the amount of training data increases the accuracy of the resulting model, as indicated by the increase in lift. Accurate risk models are thus obtained only from large training sets.

On the surface, these results seem to contradict the results obtained by Oates and Jensen [12] for classification tree algorithms. Their experiments demonstrate that the error rates of decision tree classifiers tend to rapidly reach a plateau as the number of training records increases. In fact, the plateau is often reached with only a few thousand training records. Once reached, further increases in the number of training records has little effect on the accuracy of the resulting classification tree.

A similar plateau almost certainly exists in the case of insurance risk modeling because, ultimately, there is always a limit to the degree of accuracy one can attain in any prediction problem. However, the start of the plateau clearly exists beyond the 1.38 million record mark, instead of the several thousand record mark observed by Jensen and Oats. The reason for this difference has to do with nature of the prediction problem. Decision tree classifiers make yes/no type predictions and model accuracy is assessed on the basis of whether those predictions are right or wrong. Risk models, on the other hand, make predictions about the values of continuous parameters (i.e., frequency, severity, and pure premium). Model accuracy is assessed not on whether the predictions are right or wrong, but on how well those predictions reflect reality. Such assessments are analogous to drawing distinctions between shades of gray, instead of the black and white distinctions made by classifiers. Moreover, insurance data is inherently noisy so that large amounts of data are needed to obtain accurate parameter estimates. Consequently, the accuracy plateau for risk models will be reached only for very large training sets.

The size of the training sets needed to obtain accurate risk models placed severe constraints on the experiments we were able to perform to compare the UPA to other data mining technologies. Except for SPRINT [4], all of the other tree-based modeling programs available to us (i.e., CART [2] and C4.5 [3]) could not handle the data volumes involved (1.38 million records constitutes roughly one gigabyte of data).

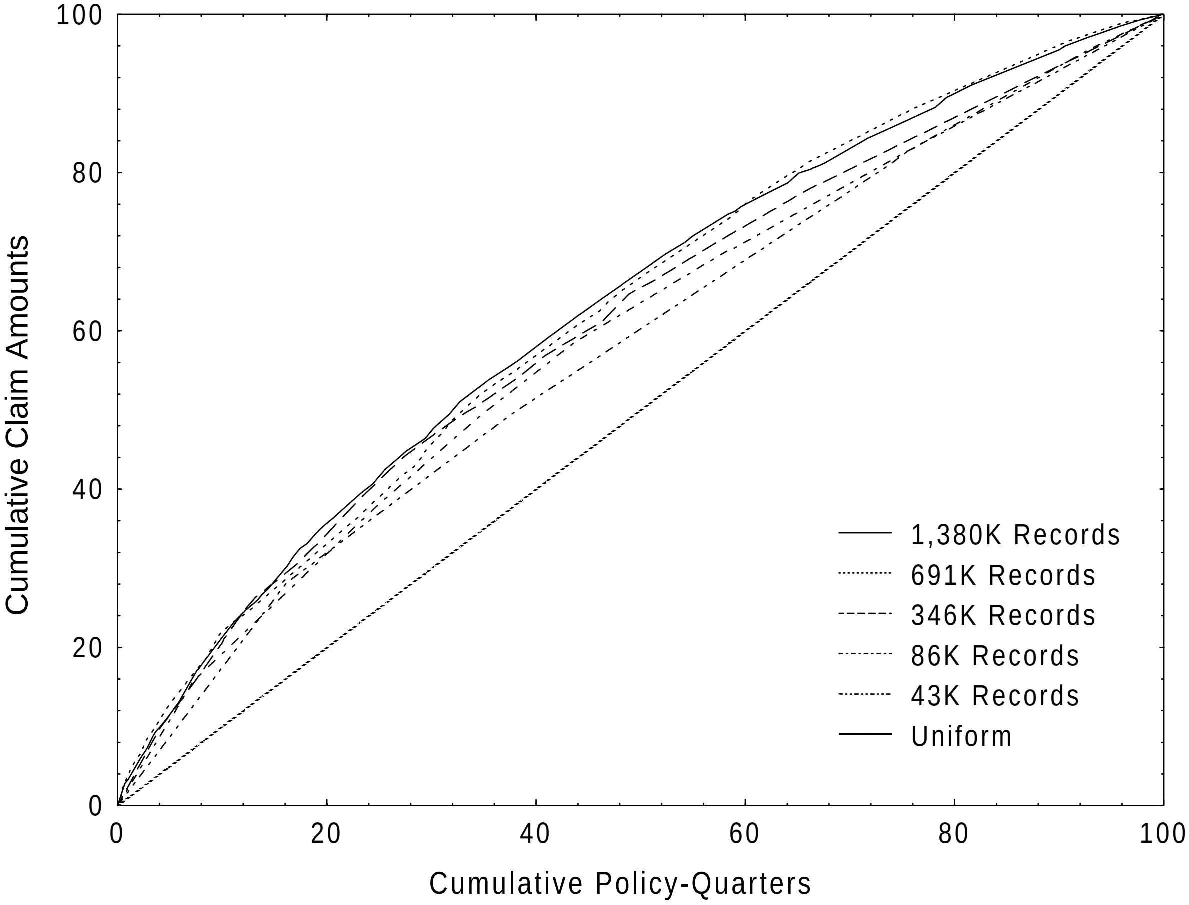

Figure 6: More is Better

Hence, comparisons were only made with SPRINT.

SPRINT permits data to be modeled in several different ways. To investigate the impact of different modeling methodologies, three different SPRINT runs were performed for every UPA run:

1. SPRINT was run in classification mode on training sets having a 1:1 ratio of claim records to non-claim records with the claim/non-claim indicator field used as the dependent variable. SPRINT thereby constructed trees that predict claim frequency.

2. SPRINT was run in regression mode on training sets containing only claim records with the claim amount used as the dependent variable. SPRINT thereby constructed trees that predict claim severity.

3. SPRINT was run in regression mode on training sets having a 1:1 ratio of claim records to non-claim records with the claim amount used as the dependent variable. SPRINT thereby constructed trees that predict pure premium.

All of the trees produced by SPRINT were converted to ProbE-style rules so that they could be loaded into the UPA for calibration and evaluation purposes. The latter step was necessary because SPRINT is simply not equipped to perform the calculations defined by Equations 3-8 for estimating insurance risk parameters. In particular, earned exposure and claim status are completely ignored by SPRINT. Actuaries

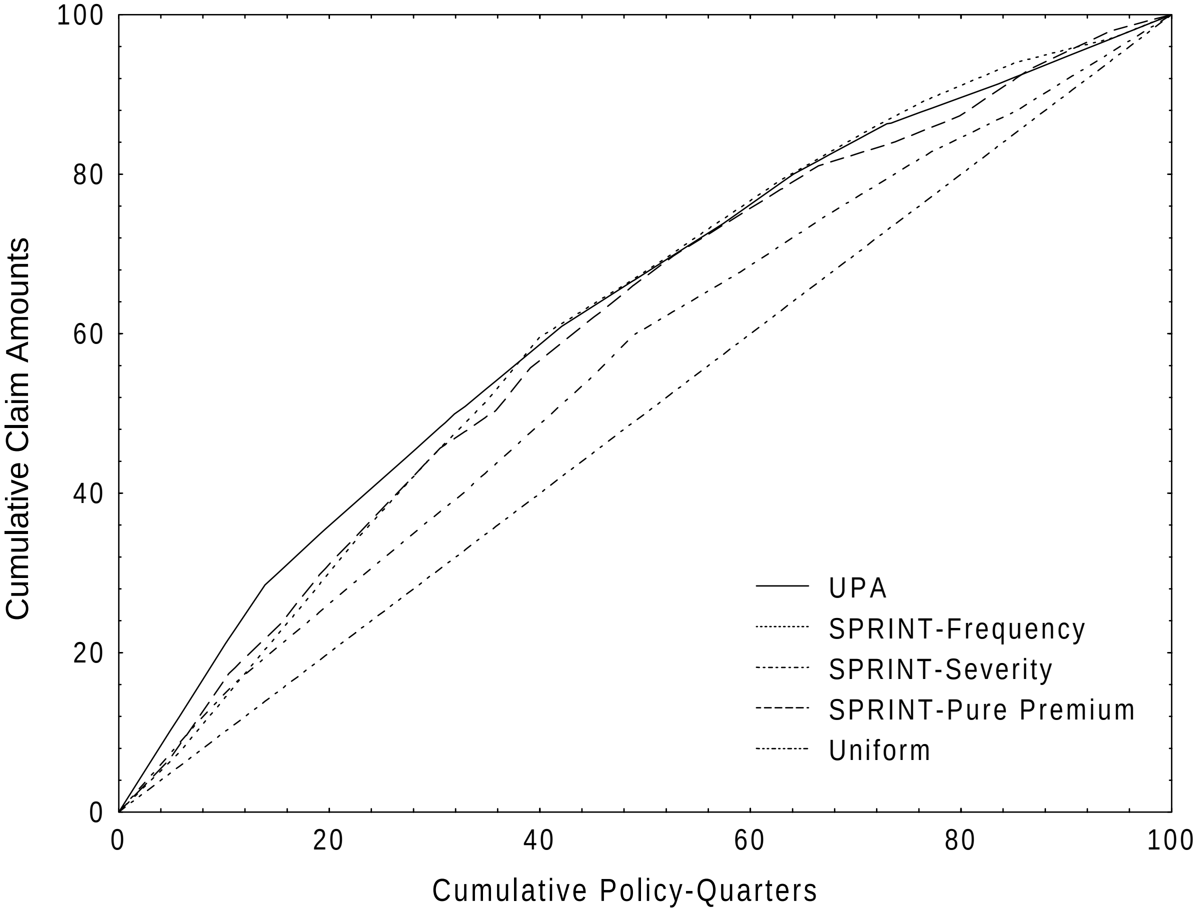

Figure 7: Bad Drivers

using SPRINT would likewise have to convert SPRINT trees into other computationally convenient forms in order to perform these calculations.

In order to obtain actuarially credible risk groups, additional pruning had to be performed on the trees produced by SPRINT, above and beyond the pruning that SPRINT itself performs, before the trees could be converted to ProbE-style rules. SPRINT does not provide a direct means of ensuring actuarial credibility. The level of credibility of a risk group depends both on the number of claim records that enter into the risk parameter estimates and on the mean and standard deviation of the claim amounts. The only comparable constraint that SPRINT is able to impose on the tree-building process is to set a threshold on the minimum number of records per leaf node. Average settings for this threshold were calculated for each data set based on the risk models produced by the UPA. However, simply imposing these thresholds on SPRINT was not sufficient to achieve actuarial credible results. SPRINT actually applies such thresholds before splitting occurs to decide whether a node should be split—it does not apply the thresholds to the nodes that would result from a split. Consequently, the number of data records that are covered by a leaf node can be arbitrarily small; in particular, the number can be much less than the threshold. Any leaf node that did not meet the threshold requirement was therefore pruned from its tree either by replacing the parent of the leaf with the sibling of the leaf, or by converting the parent into a leaf node in the case in which the sibling also did not meet the threshold requirement. Actuaries using SPRINT would likewise have to employ the same

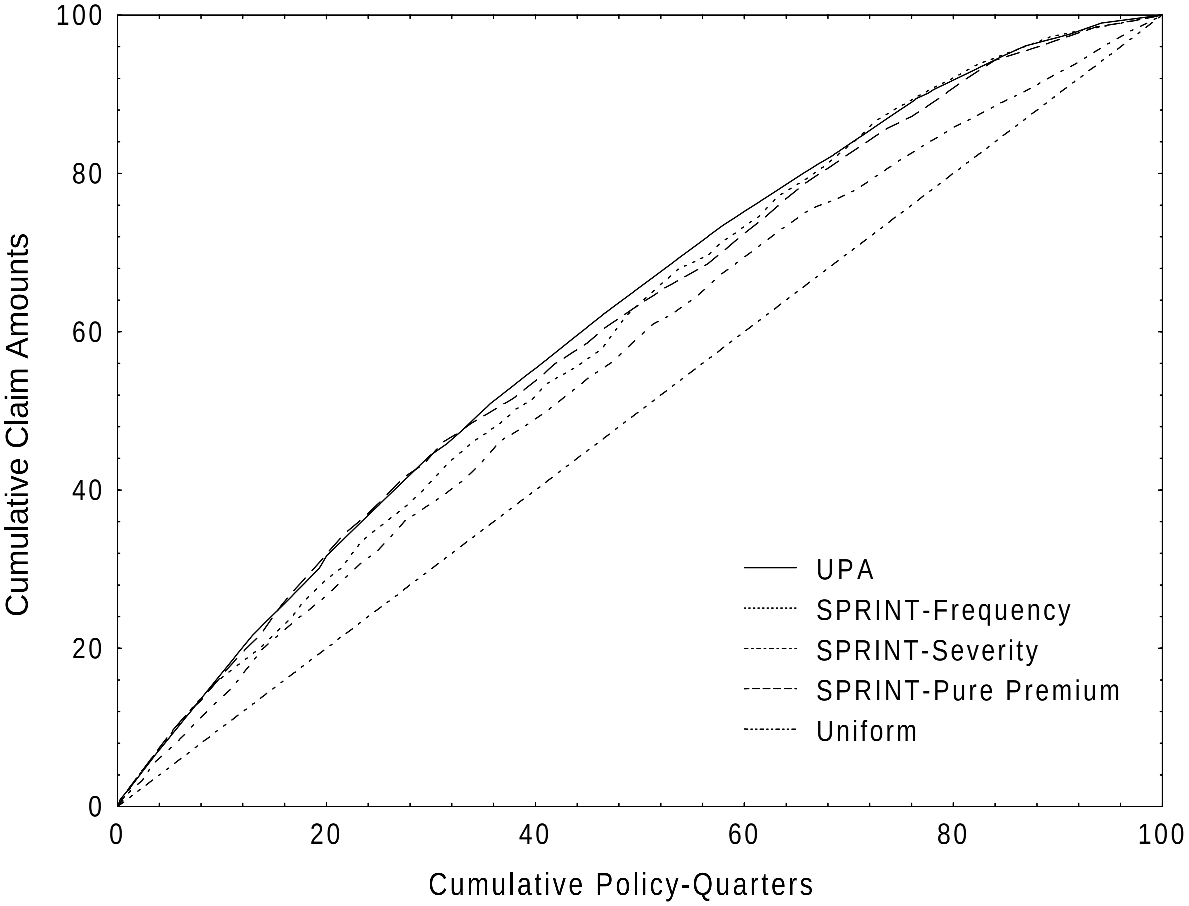

Figure 8: Good Drivers

or a similar form of pruning in order to obtain actuarially credible risk groups.

Figures 7 and 8 provide examples of the typical lift curves that were obtained in the comparison runs that were performed. As can be seen by comparing the SPRINT-Frequency to the SPRINT-Severity curves, frequency modeling is generally more predictive than severity modeling when it comes to assessing the true levels of risk posed by policyholders. Pure premium modeling can be superior to both frequency modeling and severity modeling, but not always. This effect is evident by comparing the above curves to the SPRINT-Pure-Premium curves. The lift curves obtained from the UPA, on the other hand, consistently lie in the upper ranges among all lift curves, as can be seen by comparing the UPA curves to the SPRINT curves. These general trends are likewise observed in all other comparison runs that were performed.

Although rigorous quantitative evaluations still need to be performed, qualitative assessments of the lift curves that were obtained confirm, or at least are consistent with, our expectation at the outset that the domain-specific estimation procedures and splitting criteria incorporated into ProbE would consistently produce highly predictive models for this application. There seems to be a general perception in the machine learning community that the choice of splitting criteria is a relatively unimportant difference among tree-based learning algorithms. This perception, however, runs contrary to results obtained in robust estimation [5]. When dealing with highly skewed data, standard estimators (based, for example, on assumptions of normality) can be unreliable. Robust estimators that are tolerant of skew often yield better predictions. Insurance claims data, as previously discussed, are highly skewed. Some methods of robust estimation involve deleting extreme values (i.e., outliers). Such methods are not appropriate from an actuarial standpoint because extremely high (and extremely low) claims do occur and the regularity with which they occur must be modeled in order to avoid financial ruin. Other methods of robust estimation are based on the use of probability distributions that better reflect the observed skew of the data, as well as the thickness of the tails of the observed distributions. This approach is the one preferred by actuaries, who routinely make use of a wide range of distributional models in their analyses [9]. The same approach likewise guided the development of ProbE. Because ProbE was developed from the point of view of robust estimation, our a priori expectation was that ProbE would be highly robust with respect to the risk models it produces. The lift curves presented above are consistent with this expectation, and we anticipate that extensive quantitative evaluations will further confirm our expectation.

In conclusion, we demonstrate that extra leverage can be obtained in data mining by 1) employing suitable statistical models that accurately reflect the underlying statistical properties of the data, and 2) incorporating relevant domain specific constraints, e.g., actuarial credibility for insurance risk discovery in the UPA solution. The ProbE data mining framework has enabled this approach, and will continue to serve as a robust kernel for domains where extracting maximal predictive accuracy in the mining process is at a premium.

References

[1] G.V. Kass. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29(2):119–127.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth, Monterrey, CA., 1984.

[3] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[4] J. Shafer, R. Agrawal, and M. Mehta. SPRINT: A Scalable Parallel Classifier for Data Mining. In *Proceedings of the 22nd International Conference on Very Large Databases*, 1996.

[5] R.R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, 1997.

[6] J. Hosking, E. Pednault, and M. Sudan. A Statistical Perspective on Data Mining. *Future Generation Computer Systems*, November 1997.

[7] E. Pednault. Statistical Learning Theory. *MIT Encyclopedia of the Cognitive Sciences*, 1998.

[8] C. Apte, E. Grossman, E. Pednault, B. Rosen, F. Tipu, and B. White. Insurance Risk Modeling Using Data Mining Technology. In *Proceedings of PADD99: The Practical Application of Knowledge Discovery and Data Mining*, pages 39–47, 1998. *IBM Research Division technical report RC-21314*.

[9] S.A. Klugman, H.H. Panjer, and G.E. Wilmot. *Loss Models: From Data to Decisions*. John Wiley & Sons, 1998.

[10] V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[11] L. Bransten. Looking for Patterns. *The Wall Street Journal*, page R16 and R20, June 21 1999.

[12] T. Oates and D. Jensen. Large Datasets Lead to Overly Complex Models: an Explanation and a Solution. In *Proceedings of The Fourth International Conference on Knowledge Discovery and Data Mining*, pages 294–298, 1998.