RC 21989 (98803) 13 March 2001 **Computer Science**

IBM Research Report

Probabilistic Estimation in Data Mining

Edwin P. D. Pednault, Chidanand Apte
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - T. J. Watson - Tokyo - Zurich

Chapter 1

PROBABILISTIC ESTIMATION IN DATA MINING

Edwin P. D. Pednault and Chidanand Apte

Abstract The goal of scientific inquiry is to uncover the principles that govern the world around us, and ultimately to express those principles in a mathematical form that reflects the empirical characteristics of observed data. In this regard, we have been exploring ways of modifying machine learning techniques so that the resulting predictive models likewise reflect the empirical characteristics of observed data. Following the principles of robust estimation, our methodology involves first examining the data to identify an appropriate family of statistical distributions for modeling the data, and then incorporating the corresponding maximum-likelihood estimation procedures into a decision tree algorithm. We have applied this methodology to insurance risk modeling and have obtained tree-based models superior to those obtained using conventional classification and regression tree algorithms.

Keywords: Predictive modeling, insurance risk modeling, robust estimation, decision trees, classification and regression trees

1. Introduction

The goal of scientific inquiry is to uncover the principles that govern the world around us, and ultimately to express those principles in mathematical form. The discovery process involves hypothesizing, testing, and refining potential scientific principles using both theoretical and experimental methods. Through this process, principles are eventually discovered and formulated mathematically in such a way as to reflect the empirical characteristics of observed data.

We have been exploring ways of modifying machine learning techniques so that the resulting predictive models likewise reflect the empirical characteristics of observed data. In particular, we are interested in

constructing models that not only predict the most probable outcome, but also predict the probability distribution over possible outcomes.

Our work is highly influenced by results obtained in robust estimation [R.R97]. Conventional modeling techniques are often based on certain fixed statistical assumptions; for example, on assumptions of normality in the case of least-squares regression. When the observed data violates the assumptions that underlie the modeling techniques being employed, the resulting models can be unreliable. To obtain reliable models, the statistical assumptions of the modeling techniques must be consistent with the statistical properties of the data being modeled.

Figure 1.1a illustrates some of the problems that can arise when the statistical assumptions of a modeling technique do not reflect the empirical characteristics of the data. In this example, least-squares regression is used to model a linear relationship in which the distribution of the residuals is heavy-tailed—that is, most of the data points are concentrated near a straight line that relates X to Y; however, there is also a substantial wide-spread background scatter of data points that lie quite far from the line. When least-squares regression is used to estimate the coefficients of the linear relationship, the points in the background scatter have an undue influence on the estimates of the regression coefficients. Consequently, the coefficient estimates are highly inaccurate, and so too is the resulting model.

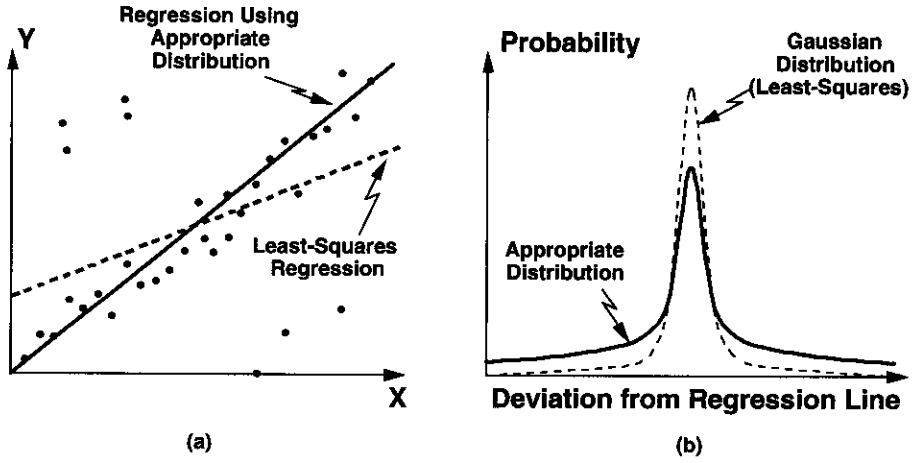

Figure 1.1. A basic principle of robust estimation is that the accuracy of regression models can be maximized by modeling the residuals using mathematical distributions that accurately reflect the observed distributions of the data.

This inaccuracy is due largely to the fact that the extent of the background scatter violates the statistical assumptions that underlie least-

squares regression. Least-squares regression implicitly assumes that the residuals are normally distributed, whereas the actual distribution of the residuals in this example is highly non-normal (i.e., non-Gaussian). Least-squares methods estimate regression coefficients by adjusting the coefficients so as to maximize the probability density of the residuals assuming that the residuals obey a normal distribution. According to a normal distribution, the probability density rapidly falls to zero for residuals that deviate more than two standard deviations from the regression line (see Figure 1.1b). If the actual probability density of the residuals does not fall as rapidly to zero as does a normal distribution, then the contributions of distant data points to the overall likelihood of the data will begin to dominate the contributions of closer data points. As the contributions of extreme data points begin to dominate, the regression line will adjust so as to reduce the distance from the regression line to the extreme points. The adjustment occurs in order to maximize the overall likelihood of the data, assuming a Gaussian noise model. The resulting effect can be observed in Figure 1.1a.

Some methods of robust estimation involve deleting extreme values by treating them as "outliers." However, such methods are analogous to assuming that extreme values represent erroneous data. Deletion of these values is then justified as a data cleansing step to eliminate erroneous values from consideration.

Other methods of robust estimation are based on the use of probability distributions that better reflect the observed distribution of the data. This approach assumes that extreme values are legitimate data values—that is, extreme values are actually observed in practice and are not in any way erroneous. Consequently, the statistical assumptions of the modeling technique must be adjusted so as to take into account (i.e., model) the extreme values that might be present.

This latter approach is the one we have adopted in our work. We presume that the data have already been cleansed of erroneous values, and that any remaining extreme values are legitimate data values actually observed in practice.

When applied to linear regression, this latter approach to robust estimation requires that a suitable probability distribution be selected that reasonably approximates the empirical distribution observed in the residuals. In the case of the example shown in Figure 1.1a, this means selecting a heavy-tailed distribution such as the one shown in Figure 1.1b. When the latter distribution is used to obtain maximum-likelihood estimates of the regression coefficients, the resulting estimates produce a much more reasonable regression line as shown in Figure 1.1a. The resulting estimation procedure is not a least-squares procedure because it

does not assume a normal distribution; however, it is still a maximum-likelihood procedure relative to a different distributional assumption.

In applying robust estimation in the context of machine learning, we have been examining extensions to decision-tree methods that allow arbitrary distributional assumptions to be incorporated into the models used in the leaves of trees. Specifically, in place of the usual classification model or Gaussian (i.e., squared-error) model, we have instead been investigating the use of more sophisticated statistical models.

Our first effort in extending decision tree methods has focused on modeling insurance risks for personal-lines (i.e., non-commercial) property and casualty insurance. The modeling technique we developed uses joint Poisson/lognormal statistical models in the leaves of decision trees to simultaneously model the frequency with which insurance claims are filed and the amounts of those claims. The choice of a joint Poisson/lognormal statistical model was made after an examination of the data to be modeled, following the principles of robust estimation. Joint Poisson/lognormal statistical models have other uses besides insurance risk modeling, and can be generally used to model random processes consisting of events that occur over time and that have positive quantities associated with those events.

The methodology we developed also demonstrates how other types of statistical models can be incorporated into the leaves of trees. Our methodology involves first examining the data to identify an appropriate family of distributions to employ, and then incorporating the corresponding maximum-likelihood estimation procedures into a decision tree algorithm. Maximum-likelihood principles are used not only to estimate the parameters of the statistical models that appear in the leaves of the trees, but also to select splitting factors during tree-building, and to prune overfitted trees after tree-building is complete.

2. Insurance Risk Modeling

The property and casualty (P&C) insurance business deals with the insuring of tangible assets, such as 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.

Insurance risk modeling can be approached as a machine learning problem; however, one is then faced with several challenges. One challenge is that specialized, domain-specific equations must be used to estimate claim frequency and claim severity. Another challenge is that some types of claims can take several years to settle, most notably bodily injury claims. Consequently, the claim amount can be a missing value. A suitable learning algorithm would therefore have to compensate for such "partially labeled" training examples. A third challenge is that insurance actuaries demand tight confidence bounds on the risk parameters that are obtained; that is, the risk groups identified by the learning algorithm must be *actuarially credible.*

The above challenges imply that standard machine learning algorithms are ill-suited for this application. If one were forced to use standard algorithms such as CHAID [Kas80, BdS80], CART [BFOS84], C4.5 [Qui93], or SPRINT [SAM96], 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 different groups of policyholders, each with their own unique filing rates, rather than attempt to discover groups that are "classified" as either always filing claims or never filing claims.

Frequency prediction is also not the same as predicting class probabilities. 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.

Severity prediction, on the other hand, is 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, as previously discussed.

The above challenges have motivated our own research [HPS97, Ped99] and have lead to the development of the IBM ProbE™ (*Prob*abilistic *E*stimation) 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.

3. Statistical Modeling of 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, [KPW98]). 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 often dictated 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 insurable risk (see Figure 1.2). Policyholders can also file multiple claims for the same type of loss over the life of a policy.

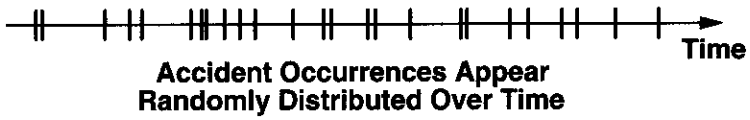

Figure 1.2. A defining characteristic of Poisson random processes is that events occurrences are distributed randomly over time.

Claim filings such as these can be modeled as a Poisson random process [KPW98], which is the appropriate mathematical model for events that are randomly distributed over time with the ability for events to reoccur (i.e., renew). The Poisson model is likewise used in the current version of ProbE.

It is important to note that Poisson processes are not appropriate for modeling catastrophic claims arising from widespread disasters—such as hurricanes, earthquakes, floods, and so forth—because such catastrophes do not produce uniform distributions of loss events over time. Instead, natural disasters give rise to clusters of claims being filed over short periods of time, where the number of claims in each cluster depends both on the number of policyholders in the region affected by the disaster and on the severity of the disaster. An appropriate statistical model should likewise take into account the geographic distribution of policyholders relative to the damage caused by disasters. Such a model would be substantially more sophisticated than the simple Poisson model currently used in ProbE, and it would have to be incorporated into ProbE in order to adequately model the true levels of risk posed by widespread catastrophes.

The Poisson model is also not appropriate for modeling other forms of insurance, such as life insurance in which at most one death benefit is ever claimed per policy. Life insurance is best modeled as a survival process, not a renewal process. Additional probabilistic models would have to be incorporated into ProbE to model such forms of insurance. Such extensions, however, should be relatively easy to accomplish compared to the extensions that would be needed to model catastrophic losses.

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. In principle, aggregate loss distributions can be derived mathematically from the distributions of the individual losses that make up the sum. However, only in a few special cases can closed-form solutions be obtained for these mathematical equations. In most cases, approximations must be employed. Fortunately, actuaries typically consider large groups of claims when analyzing aggregate loss. The central limit theorem can therefore be invoked and aggregate losses can be reasonably approximated by normal (i.e., Gaussian) distributions.

The distributions incorporated into the current version of ProbE were selected based on an examination of historical automobile claims data. Claim amounts were found to have highly skewed distributions in this data. 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 logarithmically transformed, the skewness virtually disappeared and the resulting distributions were found to be highly Gaussian in shape. These properties are the defining characteristics of lognormal distributions, an example of which is illustrated in Figure 1.3. Other statistical distributions employed by actuaries (see, for example, [KPW98]) can also be incorporated into ProbE and used in place of the lognormal distribution; however, the lognormal distribution was found to provide the best overall fit to the observed distribution of claim amounts in the available data (the degree of fit was assessed using quantile-quantile (Q-Q) plots, a standard visual method used in applied statistics for assessing the degree of fit between an assumed theoretical distribution and the empirical distribution of a set of data points). The lognormal distribution was therefore selected as the first distribution to be incorporated into ProbE for risk modeling purposes.

Unfortunately, there are no closed-form solutions for the aggregate loss distribution given that individual losses follow a lognormal distribution. In particular, a sum of lognormal random variables is not itself lognormal. An approximation must therefore be made. In ProbE, the central limit theorem is invoked and the normal distribution is used to model aggregate losses. It should be noted, however, that aggregate loss distributions are not used by ProbE during data mining, but only for post-mining analysis when estimating the aggregate parameters of each risk group. The above approximation therefore has no effect on the risk groups that are identified by ProbE.

Because different distributions are used to model individual versus aggregate losses, different statistical procedures are employed for estimating the parameters of these distributions. For the lognormal distribu-

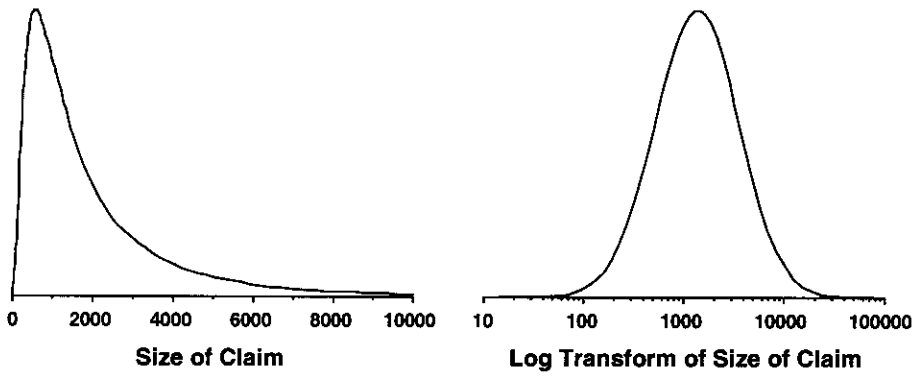

Figure 1.3. A defining characteristic of lognormal random variables is that their logarithms are normally distributed.

tions used to model individual losses, the relevant statistical parameters are the means and standard deviations of the natural logarithms of the 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.

4. 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; that is, 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 [BFOS84, Kas80, BdS80, SAM96, Qui93]. 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 pro-

duced. 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 and other classification and regression tree algorithms is that splitting factors are selected based on application-specific statistical models that appear in the leaves of the trees. In the case of insurance risk modeling, a joint Poisson/lognormal model is used to enable the simultaneous modeling of frequency and severity and, hence, pure premium. The joint Poisson/lognormal 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.

5. The Joint Poisson/Lognormal 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/lognormal 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/lognormal 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 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 1.4 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.

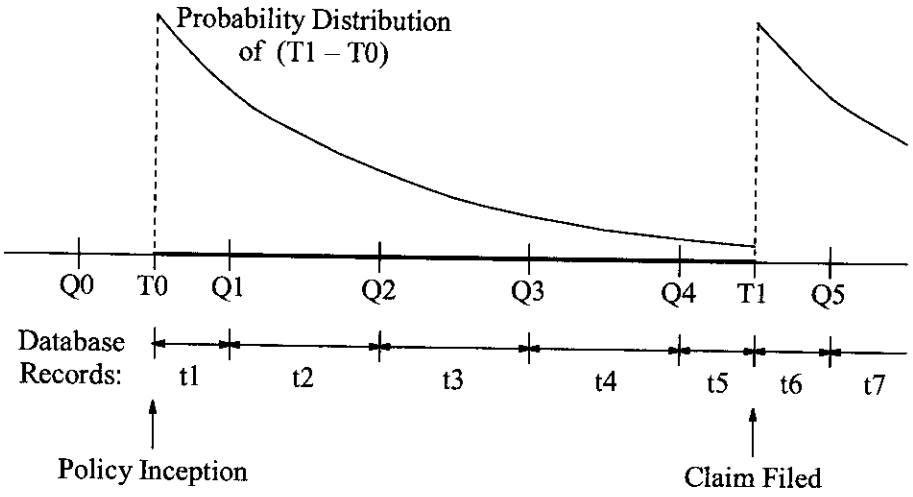

Figure 1.4. Policy data must be divided into time intervals that are short enough to support the assumption that risk characteristics remain constant within each interval.

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 the same exponential distribution as the time between claim events. For the example shown in Figure 1.4, the probability density function for the time between the policy inception and the first claim being filed is given by

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

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

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

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:

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

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:

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

The estimates of the mean log severity $\mu_{\log}$ and the variance of the log severity $\sigma^2_{\log}$ are likewise the ones typically used for lognormal 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 lognormal. 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; that is, 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 lognormal portion of the model contributes nothing to the scores of non-claim records, and an amount $\log(\sigma_{\log}) + (\log(s_i) - \mu_{\log})^2/(2\sigma_{\log}^2)$ 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 lognormal 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 lognormal 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:

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

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 mentioned 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. A modified version of CHAID's bottom-up merging technique [Kas80, BdS80] is used to construct splits. CHAID performs an iterative grouping of categories to form larger groups, and eventually a split. The algorithm is tailored to categorical factors, but numerical factors can be handled by discretizing numerical values into equiprobable categories.

We have modified CHAID's merging algorithm by using actuarial credibility constraints (discussed below) to restrict the potential merges that can be performed. At each iteration of the merging process, if one of the currently existing groups of categories does not satisfy the desired actuarial credibility constraint, then one of these nonconforming groups must be selected for merging. Subject to this constraint, pairs of groups are merged so as to minimize the sum of Equation 10 over the resulting groups of categories at each iteration.

In addition, our procedure continues merging until a binary split is constructed, whereas CHAID employs a stopping criterion that attempts to identify multiway splits during the merging process. If one of the resulting groups of categories in the binary split does not satisfy the desired actuarial credibility constraint, then the corresponding risk factor is not allowed to be used for splitting. The risk factor that is selected for dividing a larger risk group into two smaller risk groups is the one that minimizes the sum of Equation 10 for the resulting binary split, provided this split is allowed.

From the point of view of machine learning, the important thing to note about Equation 10 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 modeling 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/lognormal model presented above illustrates the general methodology one can employ to customize the splitting criteria of tree-based modeling algorithms to take into account data characteristics that are peculiar to specific applications.

6. Actuarial Credibility

ProbE's top-down modeling procedure is constrained to produce risk groups that are actuarially credible. In actuarial science, credibility [KPW98] 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 resulting 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 given by

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

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

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

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 [Kas80, BdS80], CART [BFOS84], C4.5 [Qui93], or SPRINT [SAM96].

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.

7. Predictive Accuracy

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 insurance risk modeling, 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. The use of actuarial credibility constraints helps to reduce the effects of overfitting by limiting the number of splitting factors that can be introduced, but credibility constraints in and of themselves do not eliminate the problem.

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, [Vap98]) 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. Trees are constructed using actual training data until no further splitting is allowed (e.g., because of the actuarial credibility constraints). The problem then is to identify the subtree that minimizes the estimated true score of the resulting model. The latter is obtained by evaluating Equation 10 on the validation data at each leaf of the subtree and then summing the results. The subtree that minimizes this unbiased estimate of the true score is selected as the most accurate risk model given the available data. This subtree can be identified by modifying Quinlan's reduced-error pruning method [Qui87] so that it minimizes the sum of Equation 10 over the leaf nodes, instead of minimizing the sum of the number of misclassifications at each leaf.

8. Evaluation

From the material presented in previous sections, it should be evident that using joint Poisson/lognormal models in the leaves of trees substantially increases the complexity of the calculations that are performed during tree construction. One might legitimately wonder, therefore, whether any benefits are derived from this additional level of complexity.

To answer this question, the risk models obtained using ProbE were compared to models obtained using SPRINT [SAM96], an off-the-shelf classification and regression tree program. Comparisons with other programs were not made because all of the other tree-based modeling programs available to us (i.e., CART [BFOS84] and C4.5 [Qui93]) could not handle the data volumes involved (the total available data comprised 32 gigabytes).

Risk models were compared using a variation on the idea of lift curves. In this case, the X-axis is the cumulative percentage count of policy records 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 claims paid to policyholders in the order defined by the X-axis. The Y-axis therefore also ranges from 0 to 100.

Figure 1.5 shows the lift curves that were obtained using ProbE when the amount of training data was varied from 43 thousand records (36 megabytes) to 1.38 million records (1.1 gigabytes). The curve labeled "uniform" corresponds to the hypothetical situation in which all records

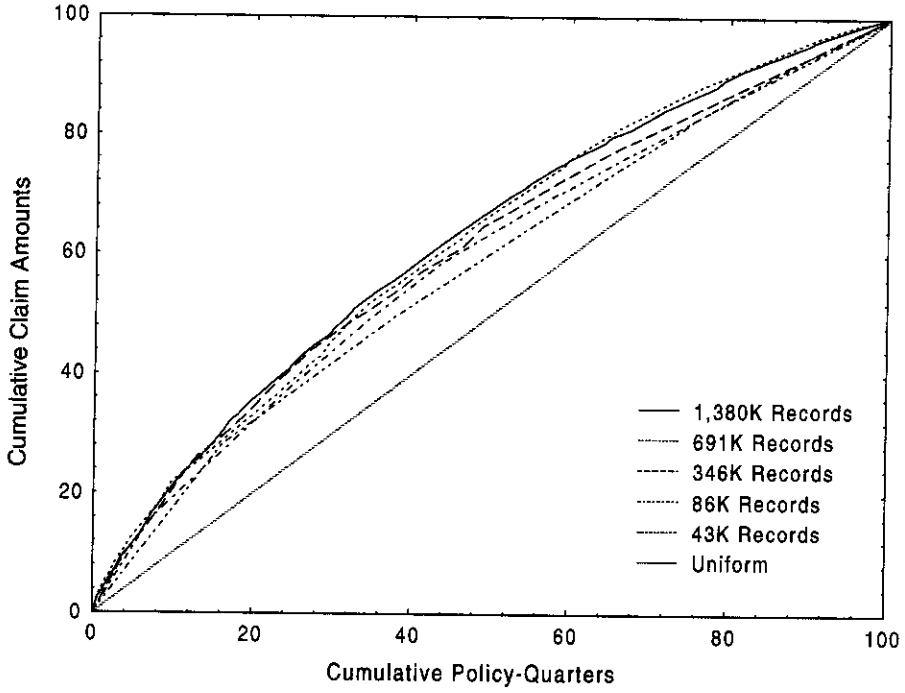

Figure 1.5. As these lift curves demonstrate, increasing the amount of training data increases the accuracy of the resulting risk model, as indicated by the increase in lift.

are assigned a uniform pure premium. 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.

Figure 1.6 provides an example of the typical lift curves that were obtained in our comparisons between ProbE and SPRINT. To investigate the impact of different modeling methodologies that could be employed when using off-the-shelf technology for insurance risk modeling, three different SPRINT runs were performed for every ProbE 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 target variable. SPRINT thereby constructed trees to predict claim frequency.

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

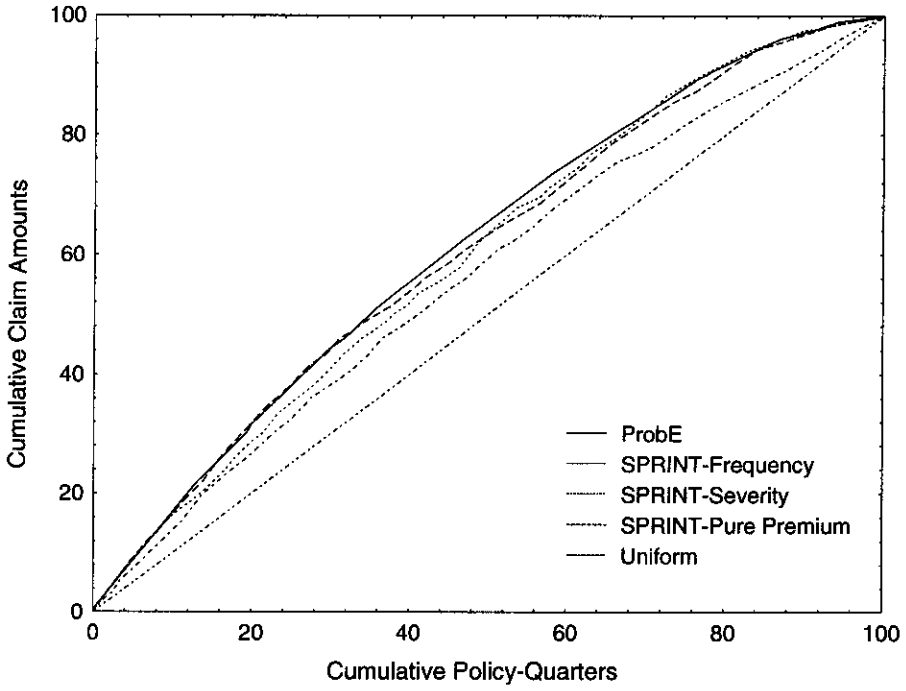

Figure 1.6. In all the experiments performed, the risk models produced by ProbE consistently outperformed those produced by SPRINT.

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 target variable. SPRINT thereby constructed trees to predict pure premium.

All of the trees produced by SPRINT were converted to ProbE-style trees so that they could be loaded into ProbE for recalibration (i.e., recalculation of the risk parameters at the leaves without modifying the branches of the trees). The latter step was necessary because SPRINT is simply not equipped to perform the calculations necessary for estimating insurance risk parameters.

As can be seen by comparing the SPRINT-Frequency curve of Figure 1.6 to the SPRINT-Severity curve, 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 using ProbE,

on the other hand, consistently lie in the upper ranges among all lift curves, as can be seen by comparing the ProbE curve to the SPRINT curves. These general trends were likewise observed in all other comparison runs that were performed.

A more complete presentation of the evaluation results can be found in [AGP+99]. 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.

9. Conclusions

There is a perception among some members of 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 [R.R97]. When dealing with highly-skewed or heavy-tailed data, standard statistical estimators (based, for example, on assumptions of normality) can be unreliable. Robust estimators that are tolerant of skew and/or heavy tails 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 [KPW98].

This latter approach to robust estimation 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.

Our emphasis on the principles of robust estimation is likewise consistent with the goals of scientific inquiry: which is to discover the princi-

ples that govern natural phenomena, and to express those principles in mathematical form, wherein the mathematical expression of the principles reflect the empirical characteristics of observed data.

References

[AGP+99] C. Apte, E. Grossman, E. P. D. Pednault, B. K. Rosen, F. A. Tipu, and B. White. Probabilistic estimation-based data mining for discovering insurance risks. *IEEE Intelligent Systems*, 14(6):49–58, 1999.

[BdS80] D. Biggs, B. deVille, and E. Suen. A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, 18(1):49–62, 1980.

[BFOS84] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth, Monterrey, CA., 1984.

[HPS97] J. Hosking, E. Pednault, and M. Sudan. A statistical perspective on data mining. *Future Generation Computer Systems*, 13(2-3):117–134, November 1997.

[Kas80] G.V. Kass. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2):119–127, 1980.

[KPW98] S.A. Klugman, H.H. Panjer, and G.E. Wilmot. *Loss Models: From Data to Decisions.* John Wiley & Sons, 1998.

[Ped99] E. Pednault. Statistical learning theory. In R. A. Wilson and F. C. Keil, editors, *MIT Encyclopedia of the Cognitive Sciences*, pages 798–801. MIT Press, 1999.

[Qui87] J. Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27:221–234, 1987.

[Qui93] J.R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1993.

[R.R97] R.R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing.* Academic Press, 1997.

[SAM96] 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*, pages 544–555. Morgan Kaufmann, 1996.

[Vap98] V.N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, 1998.