RC 21314 (94531) 31MAR98 Computer Science/Mathematics

Research Report

Insurance Risk Modeling Using Data Mining Technology

C. Apte, E. Grossman, E. Pednault, B. Rosen, F. Tipu, and B. White
IBM T. J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598

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

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 class library 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. The ProbE library itself is scalable, extensible, and embeddable. Our approach to modeling insurance risks and the implementation of that approach have been validated in an actual engagement with a P&C insurance firm. The benefit assessment of the results suggest that this methodology provides significant value to the P&C insurance risk management process.

1 Introduction

The business of insuring tangible assets, also known as P&C (Property and Casualty) insurance, deals with the insuring of things like 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. A good overview of this business appears in [6, 7].

Actuaries develop 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). Premiums are determined for each policy in a risk group based on the risk characteristics of the group as well as on the cost structure of the P&C company, its marketing strategy, competitive factors, etc.

A basic tenet in the industry is that no rating system can be perfect and competition therefore compels P&C companies to continually refine both the delineations they make among risk groups and the premiums they charge. The analytical methods employed by actuaries are based as much on statistical analysis as they are on experience, expert knowledge, and human insight. Thus, it is widely recognized that any risk model one develops is likely to overestimate the true levels of risk of some groups of policies and underestimate the risks of others. Overcharging low-risk policyholders may induce them to leave and seek lower rates from competitors, thereby reducing revenue and market share. Undercharging high-risk policyholders may attract similar high-risk customers away from competitors, thereby driving costs up and lowering profits. When one insurer refines the risk groups it has identified and adjusts its prices accordingly, market pressures eventually provoke other insurers to follow suit. To remain competitive, insurers must charge policyholders according to ever-improving assessments of their true levels of risk.

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 opportunity exists to use data mining techniques to discover previously unrecognized risk groups and thereby assist actuaries in developing more competitive rating systems that better reflect the true risks of the policies that are underwritten. The challenge for data mining is that individual policyholders file for claims very infrequently and, when they do, the individual claim amounts vary over several orders of magnitude. In addition, some of the most important data fields often have large proportions of missing values. A further challenge is that actuaries demand statistical rigor and tight confidence bounds on the risk parameters that are obtained—that is, the risk groups must be *actuarially credible*. This combination of rare events, wide variation in claim amounts, large proportions of missing values, and demand for statistical rigor is problematic for many data mining algorithms.

These challenges have motivated our own research [3, 4] and have lead to the development of the IBM ProbE$^{\mathbf{TM}}$ (Probabilistic Estimation) predictive modeling class library. This C++ library 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 IBM UPA$^{\mathbf{TM}}$ (Underwriting Profitability Analysis) application 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 recently concluded pilot engagement 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 Modeling Insurance Risk

Risk groups and their associated risk characteristics can be expressed in the form of actuarial rules such as *male drivers under age 25 who drive sports cars have a claim frequency of 25% and an average claim amount of $3200*. To be able to discover such rules from historical claims and policy data, it is intuitively natural to view the data mining task as one of predictive modeling based on rule induction. Insurance companies collect several hundred data fields for each policy they underwrite. There may be several million policies in a geographic region. Clerical verification and entry of this information into a database is common practice. Given the high dimensionality of the data and a regulatory requirement that risk groups be mutually exclusive, a decision-tree [1, 2, 5] approach is a pragmatic and practical method for rule induction.

The key variables that one must try to predict are claim frequency and claim severity, and thereby pure premium. Claim frequency is the average rate at which individual policyholders from a risk group file for claims and is expressed as the number of claims filed per policy per unit time (i.e., quarterly, annually, etc.). Sometimes the rate is 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 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 and is simply the average dollar amount per claim. Pure premium is the product of frequency and severity.

Raw policy claims data contains fields from which frequency and severity can be estimated. The fields for estimating frequency typically specify the total number of claims filed under a given policy during a specified time period. The field for estimating severity typically specifies the total dollar amount of the corresponding claims for that policy during that time period.

If one were restricted to using standard decision-tree algorithms, 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 not exactly straightforward 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 a frequency modeling problem, which needs to discover recognizable groups of policies, each with its own unique filing rates, rather than attempt to discover groups that are classified as either *always* filing for claims or *never* filing for claims.

From the point of view of standard decision-tree algorithms, severity prediction appears to be very much a regression problem, given that the fields corresponding to this variable are continuous values across a wide range. However, the distribution characteristics of claim amounts are quite different from the traditional Gaussian (i.e., least-squares optimality) assumption that most regression modeling systems make.

A further complication from the point of view of standard decision-tree algorithms is that frequency and severity must be modeled *simultaneously* because the risk parameter that ultimately determines pricing is the pure premium estimate. Developing risk groups based on frequency or severity alone and then estimating pure premium post hoc would yield suboptimal models because the risk groups would not be optimized for predicting pure premium. Developing separate rule-based models for frequency and severity and then combining them would also be unsatisfactory because the estimates for frequency and severity would be based on different subpopulations. From actuarial and regulatory standpoints, the credibility of the resulting pure premium estimates would therefore be questionable. Frequency and severity must be modeled simultaneously in order to construct risk groups that accurately predict pure premium.

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.

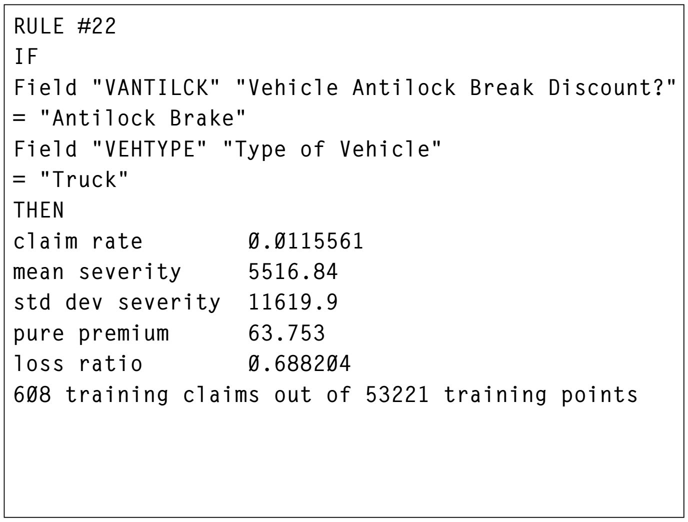
ProbE replaces manual exploration of potential risk factors with automated search. Risk groups are identified in a top-down fashion by a method similar to those employed in classification and regression tree algorithms [2, 5]. Starting with an overall population of policyholders, ProbE recursively divides the policyholders into risk groups by identifying a sequence of factors that produce the greatest increase in homogeneity within the sub-groups 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 splitting factor is used to enable the simultaneous modeling of frequency and severity and, hence, pure premium. By explicitly taking these factors into account, ProbE is able to overcome the major barriers that render standard data mining algorithms inappropriate for this application.

3 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 class library is used for actual execution of mining tasks. The client-server implementation is multi-threaded 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



```
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
```

Figure 1: Example of a UPA generated rule

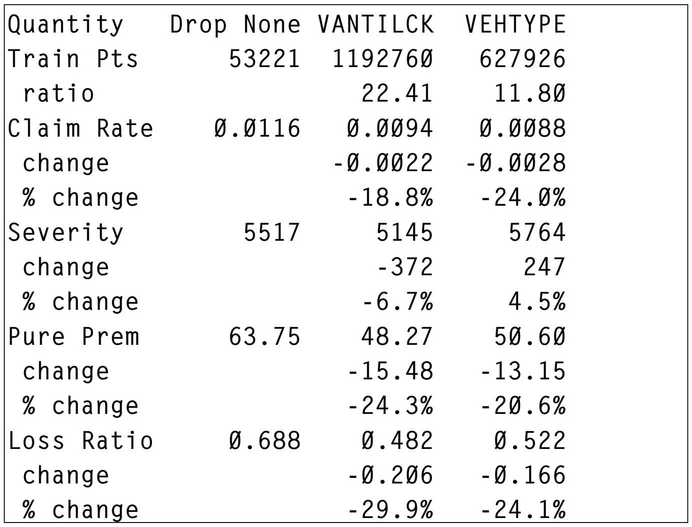
information is captured in a set of derived fields. The application is geared to predict pure premium, which is the product of claim frequency and claim severity. Though not explicitly present in the raw data, it is readily computed once mean frequency and mean severity have been estimated.

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.

Mining runs produce risk models that are represented as collections of rules. A typical rule is illustrated in Figure 1. 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),



```
Quantity   Drop None VANTILCK  VEHTYPE
Train Pts      53221  1192760   627926
 ratio                  22.41    11.80
Claim Rate    0.0116   0.0094   0.0088
 change               -0.0022  -0.0028
 % change              -18.8%   -24.0%
Severity        5517     5145     5764
 change                  -372      247
 % change               -6.7%     4.5%
Pure Prem      63.75    48.27    50.60
 change                -15.48   -13.15
 % change              -24.3%   -20.6%
Loss Ratio     0.688    0.482    0.522
 change                -0.206   -0.166
 % change              -29.9%   -24.1%
```

Figure 2: Example of sensitivity analysis of a rule

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. For example, the rule in Figure 1 matches 53,221 examples, out of which 608 had incurred claims.

Also reported for each rule is a sensitivity analysis table as illustrated in Figure 2. The sensitivity analysis tables show how the segment statistics of each rule change by dropping one clause from the "if" part of the rule. For example, the rule in Figure 1 has two clauses, VANTILCK = "Antilock Brake" and VEHTYPE ="Truck". The sensitivity analysis table in Figure 2 illustrates the effect on the predictions if either of these clauses were to be dropped individually. The table is an extremely useful analytical tool that an end-user can employ during rule editing to generalize a rule by dropping only those clauses that cause maximal increases in matching examples with minimal changes in risk characteristics.

4 Uncovering nuggets in the rules

The first step in transforming the mining results into business value begins with the lift charts that are generated from a mining run. A typical UPA lift chart is displayed in Figure 3. 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 when an insurance firm has

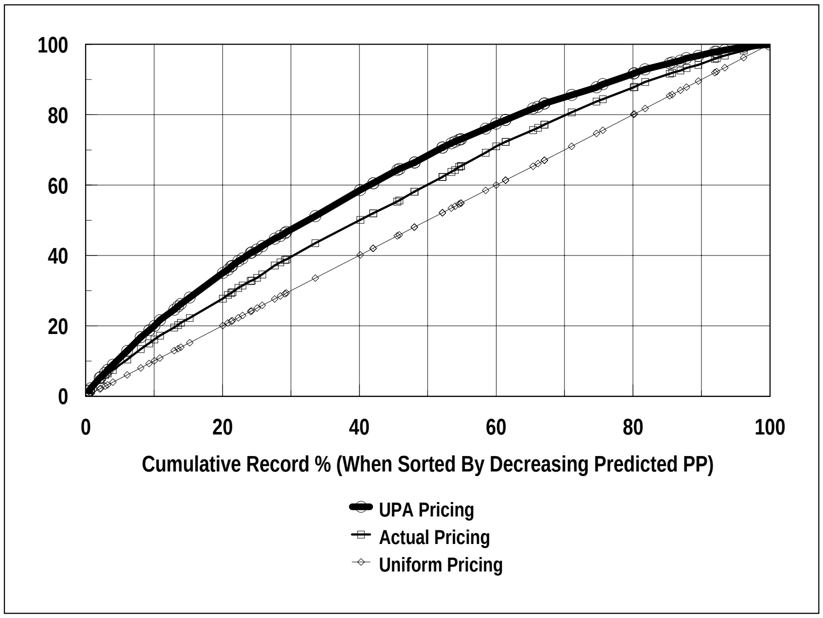


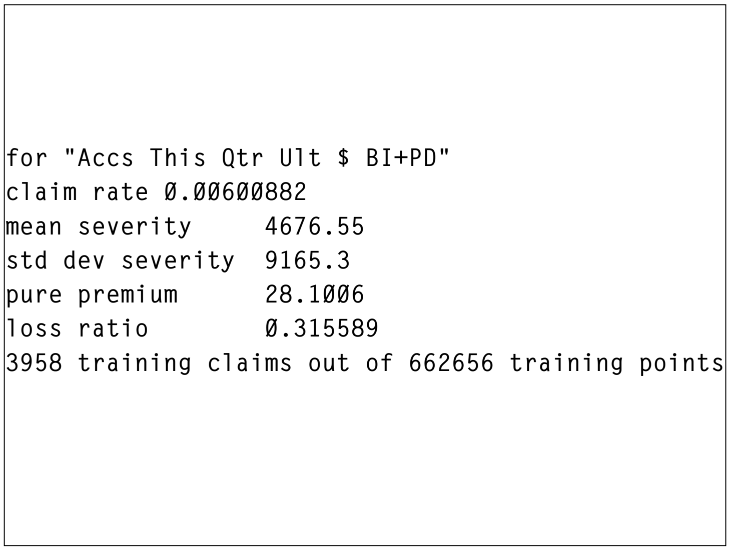
Figure 3: UPA Lift Chart

no insight about its policies and spreads its risk uniformly across the entire pool. The second plot displays the actual scenario in which the firm's current actual premium pricing is plotted (which is the actual cumulative premiums collected for the policies in descending order sorted 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 in descending order sorted by predicted pure premium).

Figure 3 illustrates the best-case scenario from a data-mining perspective in which the insurer's actual pricing reflects differences between risk groups that are not captured by uniform pricing, and the UPA-proposed pricing incorporates further distinctions among risks not currently reflected in the insurer's pricing. These relationships are also common 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, with its insurance-tuned data mining engine, has a strong likelihood of discovering previously unknown risk groups and is therefore able to suggest more competitive prices in many situations.

If an actual mining run results in a lift chart very similar to the one illustrated in Figure 3, 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 the results can be abandoned right away.

To uncover nuggets, the analyst needs to first understand the statistics for the entire book



```
for "Accs This Qtr Ult $ BI+PD"
claim rate 0.00600882
mean severity     4676.55
std dev severity  9165.3
pure premium      28.1006
loss ratio        0.315589
3958 training claims out of 662656 training points
```

Figure 4: Background Statistics for a Data Mining Run

of business. The UPA application can present these *background* statistics to the user, as shown in Figure 4. 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 determining the nuggets utilizing the above criteria. For example, the rule illustrated in Figure 1 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.

5 Conclusion

The UPA solution was recently developed as part of a joint project with a major insurance company in North America. Automobile insurance data for 16 quarters from the books of business in a single state was extracted and transformed into a data mart. The mart represented about 2 million policies and was approximately 30 GB in size.

There were many ways to mine this data. 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. A benefits assessment study indicated that implementing just 6 of these 43 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.

UPA was recently announced as an IBM product offering, as a component of the IBM Decision Edge for Insurance business intelligence solution suite.

References

[1] C. Apte and S. Weiss. Data Mining with Decision Trees and Decision Rules. *Future Generation Computer Systems*, November 1997.

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

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

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

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

[6] B.L. Webb, C.M. Harrison, and J.J. Markham. *Insurance Operations - Volume I*. American Institute for Chartered Property Casualty Underwriters, 1992.

[7] B.L. Webb, C.M. Harrison, and J.J. Markham. *Insurance Operations - Volume II*. American Institute for Chartered Property Casualty Underwriters, 1992.