3.2.3 Data Mining Techniques

5m

Several data mining techniques can be used to refine classification analysis. Some commonly used ones are described below. Generally, the data mining techniques mentioned below can enhance a ratemaking exercise by:

- narrowing down a large set of potential explanatory variables to a manageable list for use within a GLM;
- offering insights on how to categorize discrete variables;
- simplifying multi-level discrete variables (e.g., consolidating 100 levels, some with few or no claims, into 20 homogeneous levels);
- identifying potential interaction variables within GLMs by uncovering interdependent patterns between variables.

Factor Analysis

Factor analysis is a method used to reduce the number of parameter estimates in a classification analysis. The most commonly used factor analysis is ***principal components analysis***. In the context of ratemaking, this method can be applied to consolidate a long list of highly correlated variables into a single score variable, which represents linear combinations of the original variables. An example of these score variables is the vehicle symbol used in the previous subsection, which is a linear combination of correlated variables such as vehicle weight, number of cylinders, horsepower, etc.

Cluster Analysis

Cluster analysis is an exploratory technique that groups similar risks into larger homogeneous categories or "clusters." It aims to minimize the differences within each category while maximizing differences between categories. Its primary application is in geographical rating, where risks are grouped into clusters based on their location, e.g., zip codes.

CART

CART, which stands for Classification and Regression Trees and is more commonly known as decision trees, is used to create tree-building algorithms that construct a set of if-then logical conditions that help enhance classification.

Analysis of CART can be used to:

- Identify the key initial variables
- Categorize each variable
- Detect interactions between variables

MARS

The Multivariate Adaptive Regression Spline (MARS) algorithm functions as a multiple piecewise linear regression, with each breakpoint indicating a region for a particular linear regression equation. It is primarily used to determine breakpoints for transforming continuous variables into categorical ones. It can also help detect interactions between variables. MARS can even be incorporated into GLM.

Neural Networks

Neural networks are sophisticated modeling techniques in which training algorithms automatically learn the structure of a given set of data. The outcomes of a neural network can be integrated into a GLM, or vice versa. Neural networks can help identify interactions that are missing in a GLM.

3.2 Summary

10m

Early insurance classification relied on basic univariate methods, later advancing to standardized univariate approaches like minimum bias procedures. With improved computing and data capabilities, insurers adopted multivariate methods, propelling industry progress.

Multivariate methods, grounded in statistical theory, account for diverse business mixes and randomness in insurance data, offering insights through diagnostics and supporting rating algorithms with interaction effects. Generalized linear models (GLMs), now standard in developed markets, are favored for their transparency and usability in ratemaking.

Ratemaking actuaries must understand the foundations of GLMs, which build on linear models, and visualizing GLM results aids practical application. Alongside GLMs, tools like CART, factor analysis, and neural networks enhance analysis, while increased access to external data further enriches classification ratemaking.

Minimum Bias Procedures

Minimum bias procedures are iteratively standardized univariate approaches that account for an uneven mix of business. The process of a minimum bias procedure that uses a multiplicative rating structure with balance principle is summarized below.

1. Equate the exposure-weighted loss costs to the indicated loss costs for each level of each rating variable.
2. Calculate the seed relativities for all but the first variable.
3. Using the seed relativities, solve for the relativities of the first variable.
4. Discard the seed relativities for the second variable and use the result from Step 3 to calculate the relativities for the second variable.
5. Repeat the process until convergence is achieved.
6. Re-base the relativities.

Multivariate Classification

Factors leading to the use of multivariate statistical techniques in classification ratemaking:

1. Enhancement in computing power

2. Improved granularity and accessibility of data
3. Competitive pressure

Benefits of multivariate methods:

1. They automatically adjust for exposure correlations between rating variables.
2. They attempt to capture only systematic effects (signals) and ignore unsystematic effects (noise).
3. They generate model diagnostics.
4. They allow consideration of response correlation.

GENERALIZED LINEAR MODELS

Generalized linear models (GLMs) remove the assumptions of normality and constant variance on the error term and allow a link function to define the relationship between the expected response variable and the linear combination of the predictor variables. A modeling dataset, a link function, and a distribution of the underlying random process are needed to solve a GLM.

GLM analysis is typically performed on loss cost data instead of loss ratios due to the following reasons:

1. Modeling loss ratios requires on-level premiums at the granular level, which poses practical challenges.
2. Experienced actuaries have an a priori expectation of frequency and severity patterns.
3. Loss ratio models become obsolete when rates and rating structures are changed.
4. There is no standard distribution for modeling loss ratios.

GLMs achieve all the benefits of the multivariate methods and have the following additional benefits:

1. Transparent
2. Model output includes parameter estimates for every level of each explanatory variable within the model, along with a range of statistical diagnostics
3. Iterations can be tracked
4. Model output is a series of multipliers

GLM diagnostics include:

1. Standard errors: Measure the variability associated with a parameter

2. Measures of deviance: Measure the extent to which the fitted values deviate from the actual observations
3. Practical diagnostic: Run the GLM for individual years across a multi-year dataset to assess the consistency of results over time

Practical considerations when using a GLM include:

- Ensuring data adequacy to avoid GIGO
- Identify when unusual outcomes dictate further exploratory analysis
- Evaluating model results from both statistical and business perspectives
- Developing appropriate methods to communicate model results based on the company's ratemaking objectives

The following external information can be used to supplement an insurer's existing data for a GLM analysis.

- Geo-demographics
- Weather
- Property characteristics
- Information about insured individuals or businesses

Data Mining Techniques

Data mining techniques that are commonly used to refine classification analysis include:

1. Factor analysis
2. Cluster analysis
3. Classification and Regression Trees (CART)
4. Multivariate Adaptive Regression Spline (MARS)
5. Neural network