Predictive Models

A Practical Guide for Practitioners and Regulators

Don Closter ACAS, MAAA, ASA

Caryn Carmean ACAS, MAAA

Motivation. Provide guidance with respect to the creation, testing, documentation, and evaluation of predictive models, in particular Generalized Linear Models.

Approach. Compact summary of data organization & preparation, variable usage & selection, model evaluation, and algorithm building.

Keywords. generalized linear modeling; rating algorithms

1. INTRODUCTION

It has become standard practice for P&C pricing actuaries to use predictive models, specifically GLMs, in setting rates. Models are developed and presented to various customers, clients, and regulators for review. Although there are many books and articles covering model development and theory in detail, our goal is to provide a relatively short paper highlighting the most important topics to understand, and key questions to ask when evaluating a model and its output.

1.1 Objective

This white paper provides both practitioners and model reviewers (i.e. product managers and regulators) with recommendations for analysis and review of predictive models (in particular GLMs), including guidance with respect to the creation, testing, documentation, and evaluation of models. This information is meant to assist both practitioners and reviewers in their efforts to be comfortable with final model results. If more detail is desired, a short list of references has been included in Section 6.

It is assumed that the reader has a mathematics or statistical background and is familiar with actuarial ratemaking.

1.2 Outline

The remainder of the paper proceeds as follows:

Section 2 provides the background on some of the advantages of GLM over earlier methods. It also covers the input required to build a model such as data, **variables**, and **relativities** along with some of the adjustments needed to get this data into a format that can be properly modeled.

Section 3 covers model building and testing.

Section 4 covers algorithm building and final documentation.

Section 5 contains the summarized conclusion.

2. BACKGROUND

In the latter part of the 20th century, insurance companies began moving away from univariate analyses of rating factors to set rates and instead adopted multivariate predictive modeling, in particular GLM approaches. At about the same time computing power and data availability dramatically expanded. As a result, many companies increased the complexity of their rating algorithms – both in structure and number of variables. The use of GLMs served to enhance the predictive value associated with risk classification, improving the best estimate of future loss.

Some of the advantages of GLMs include:

- GLMs automatically adjust for correlations by evaluating rating variables simultaneously, thus eliminating the risk of "double counting" the influence of these correlated rating variables.
- The GLM approach allows for the use of explicit interaction tables that can appropriately account for situations where the signal from one variable is based on the level of another variable. For example, the risk differences between male and female drivers can vary based on the age of the driver, and an interaction table can address this.
- GLM models tend to be transparent. It is easy to see what variables are included and how they combine together in the regression analysis.

GLMs do have disadvantages as well, some of which include:

- Intuitive interpretations can be difficult because the models are more complex, and the results include correlations among all the variables in the model.
- GLMs will generate a relativity and standard error for every variable level in the model, regardless of the underlying volume. All relativities are assumed to be fully credible.
- GLMs assume the random portion of the risks are uncorrelated.

2.1 Definitions

Variable – any characteristic, or quantity related to an individual risk that can be measured or counted.

Variable levels – unique values a **variable** may take. For a driver age variable, driver age levels could be individual ages or a collection of age groups.

Relativities – a set of numbers associated with a **variable** comparing each **variable level** to a corresponding base **variable level.**

Regression – the analysis or measure of the association between one **variable** (the dependent variable) and one or more other **variables** (the independent **variables**).

Univariate analysis / model – analysis or model built by evaluating a single **variable** at a time.

Multivariate analysis / model – analysis or model built by evaluating two or more **variables** simultaneously.

Correlation – a number representing the degree of relationship between two **variables**.

GLM – an extension of traditional regression models that allows the mean to depend on the explanatory **variables** through a link function, and the response **variable** to be any member of a set of distributions called the exponential family (e.g., Normal, Poisson, Binomial, etc.).

Interaction – when the value of one **variable** depends on the level of another **variable**. An example would be where accident frequency differs based on the interaction of driver age and gender, or where home loss severity differs based on the interaction between construction type and protection class.

Rating variable – a **variable** included in the rating algorithm.

Rating factors – **variable levels** and **relativities** associated with each **rating variable.**

Rating algorithm – a formula used to generate rates by combining various **rating variables** and their associated **rating factors.**

2.2 Data

The starting point for good analysis is always quality data with an appropriate level of granularity. The data should be organized, verified, and documented.

Organized means the individual records associated with a risk should contain placeholders to record each of the predictor **variables** and the target **variable**. For **variables** whose levels have been grouped, a modeler should be able to show that the **variable levels** have been grouped appropriately.

Verified means variable totals and distributions have been reviewed to insure they are consistent with other collaborating data. For example, total dollars of loss should equal an external report; or there should be an explanation for the differences; the distribution of policies by age should be consistent with the company's book of business; etc.

Documented means the source of the data and any data adjustments needed to put the data in the proper format for analysis have been identified. For example, if no claim occurred on a policy, blank or NA values may need to be converted to zero in order to be analyzed properly.

Data Quality is demonstrated by investigating each **variable** to make sure the data is reasonable. For example, **variables** should have a reasonable number of missing values, **variables** and **variable levels** should be well defined, the distribution of business across the **variable levels** should be relatively stable over time, the data should be appropriate for its intended use, etc.

Granularity is the level of detail available in the data. From a different perspective, granularity is the level at which the data is grouped or summarized. In general, data should be collected at the most detailed exposure level available such as, policy, vehicle, coverage, peril, and item level. For example, population can be counted at a world, country, state, county, zip, block group, etc. An appropriate balance needs to be selected between cost, computational complexity, and the amount of detail required for effective analysis.

2.3 Data Preparation

Data Grouping – A good portion of data grouping can be done in the preparation stage. Groupings are often made for the following reasons:

- To address low credibility issues, or data fill issues (NAs, missing values, or blanks). Variable levels with insufficient data should be grouped with other similar levels if available, grouped with the base level if the volume is small enough to not distort the base value, or excluded from the analysis. Alternatively, a Missing Level could be included in the modeling dataset and analyzed separately for informational purposes.
- To reduce the number of **variable levels** by combining homogeneous levels. For example, grouping drivers age 40-45, or grouping zip codes with similar weather patterns. In the data preparation stage a modeler should be careful not to reduce levels too much, as they may inadvertently hurt the predictive power of the model. Variables can always be grouped during modeling if it is warranted.

Data Transformations are uniform conversions of data to a more convenient form. A very common method is the log transformation of loss amounts. Loss amounts are generally highly skewed with a large number of small claims and a smaller number of large claims. Modeling the log of these claim amounts allows the model to handle them more efficiently. Another example would be normalizing transformations such as dividing each dataset by its mean. This results in datasets each with a mean of 1.0 that can effectively be compared to each other.

2.4 Variable Use

Modeled Rating Variables are **variables** whose **relativities** are generated by the model and included in the final **rating algorithm**.

Offset variables are **variables** whose **relativities** are included in the model and the final **rating algorithm**, but they are generated from other studies outside the **multivariate analysis** and fixed (not allowed to change) in the model when it is run. Examples of offset **variables** include limit and deductible **relativities** that are more appropriately derived via loss elimination analysis. The resulting **relativities** are then included in the **multivariate model** as offsets.

Manufactured variables are **variables** created outside the **multivariate analysis**, often as a score. This score variable is then included in the model so that it can generate score **relativities** to be used in the final **rating algorithm**. Manufactured variables can pick up complex **interactions** from combinations of weak **variables** that would not be detectable if placed in a **GLM** as individual **variables**. Insurance (credit) Score is a good example of a **manufactured variable**.

Control variables are **variables** whose **relativities** are not used in the final **rating algorithm**, but are included when building the model. They are included in the model so that other correlated

variables do not pick up their signal. For example, state and year are frequently included in models as control **variables** so that the different experiences and distributions between states and across time do not influence the **rating factors** used in the final **rating algorithm**.

Excluded variables are **variables** (known or unknown) that are not included in the model. These can be truly unknown variables, variables that were eliminated because of lack of significant signal or instability over time, variables not included in an effort to simplify the final model, uncollected variables, or variables prohibited by statute. The **variables** included in the final model will capture any signal from **excluded variables** (whether known or unknown) to the extent they are correlated with them.

Variable Use	Variable Included in Model	Relativity Generated inside or outside Model	Relativity Included in Rating Algorithm
Modeled Rating Variable	Yes	Inside	Yes
Control Variable	Yes	Inside	No
Offset Variable	Yes	Outside	Yes
Manufactured Variable	Yes	Inside	Yes
Excluded Variable	No	N/A	No

2.5 Splitting the data into Training and Testing Datasets

Prior to modeling, the available data should be split randomly into 2 groups. The first is the training dataset that is used in creating the model and typically contains 70% to 80% of the available data. The remaining data makes up the testing (holdout) dataset that is set aside to demonstrate that the model works on a previously unseen set of data. If sufficient data is available, a validation dataset can also be created. The validation dataset can be used during the model building process to obtain additional information on potential model refinements.

3. MODEL BUILDING

The first step in building a model is to determine the type of analysis that will be done. Based on the type of analysis and characteristics of the variables involved, a statistical distribution can be selected. Typical examples of GLM insurance models include modeling frequency using a Poisson or Negative Binomial distribution, modeling severity using a Gamma or Inverse Gaussian distribution, and modeling pure premium with a Tweedie distribution. In general, for pricing algorithms, GLMs use a log-link function with each of these distributions. The indicated model relativities, then, will consist of relativities that are multiplicative.

3.1 Variable Identification & Selection

Variable selection involves choosing those **variables** (from all those available) that will be most predictive of future experience. Various statistical tests such as chi square, evaluating stability over time, and reviewing variables on a univariate basis, can help in paring down the number of **variables** needed in a model and the final **rating algorithm**. In addition, correlations between variables should be reviewed. Although GLMs automatically adjust for correlated variables, a highly correlated pair of variables can make a model unstable unless one of these variables is removed from the model. Regardless of the **variable**, those used in the final **rating algorithm** should be limited to the best predictors of the target **variable.**

3.2 Model Validation & Integrity

Building a model is an iterative process. A model should be able to differentiate between true signal in the data and random noise. Model output should be reviewed, adjustments made as needed, and re-run. Examples of adjustments include additional variable level grouping, curve fitting, adding interactions, or simplifying the model.

3.3 Final Model Testing

The purpose of model testing is to demonstrate that the model [built on a training dataset] can effectively predict results on a random, or a previously unseen, set of data [the test (holdout) dataset]. By applying the model to the test dataset and comparing model results to actual observations, we can get an idea of the predictive quality of the model. Tools that can help provide that information include predicted vs. observed graphs generated from the test dataset, lift curves, Gini curves, etc.

4. ALGORITHM BUILDING

Once a model is finalized, the resulting indicated **relativities** can be used to build a **rating algorithm**. Theoretically, the indicated model results could be implemented directly, but often other considerations influence how results should be adjusted and used in the real world. When building the final **rating algorithm**, consideration should be given to the following adjustments:

Competitive adjustments – made to insure selected factors are reasonably aligned with the competition to both minimize anti-selection, and reflect the fact that a competitor may be able to generate more credible indications from its larger volume of data.

Business adjustments – designed to meet various corporate goals. For example, forcing a **variable** to have the same discount across all coverages, or grouping similar **rating factors** together to simplify the rating structure.

Disruption adjustments – used to reduce large price swings and ensure stability in the book of business. The impact can be measured by generating a histogram comparing the change from currently charged rates to indicated rates for current customers. Capping or tempering the rate relativities are common approaches used to control disruption.

4.1 Documentation

Documentation is the key to communication, and the responsibility of the modeler. With thorough documentation, the modeler can respond efficiently and effectively to inquiries about the model. In addition, if the models need to be run again perhaps by a different person, there should be enough detail available to re-create the model.

Model documentation should include information on the software used for analysis, data sources (internal & external), data volume, and time period, as well as any material assumptions and adjustments made to the data. Variables used in the model should be defined along with any statistical tests used to verify their significance. The approach for generating the training dataset and corresponding testing dataset(s) should be well defined and documented.

A good summary overview of the model results is a set of predicted vs. observed graphs (calculated on the test dataset) along with the final model parameters and relativities. This would provide a clear picture of how all the modeling pieces come together.

5. CONCLUSIONS

Effective model evaluation can be challenging because of the large volume of data involved, as well as the need to understand the details of how the model was built. The goal of this paper is to provide a framework and structure for evaluating these models so a reviewer can feel comfortable with the key components and results of the model.

Appendix A – Two Key Questions to Ask when Evaluating a Model

A) Is the model predictive?
   a. What data was used to train the model?
   b. What data was used to test the model?
   c. How do predicted values based on test data compare to observed data? This is most easily seen via graphs:
      i. A plot of predicted values (sorted lowest to highest) relative to actual values should reasonably follow a diagonal line.
      ii. Comparing a graph of new model predicted results to actual, vs. current model predicted results to actual, should demonstrate an improved fit using the new model.
      iii. Other helpful graphs include lift curves and Gini curves.
      iv. Because of the size of the dataset, or the presence of large outliers, it may be necessary to display the data grouped into equal quantiles and/or display the results on a log scale.

B) What adjustments were made to the model when building the final rating algorithm?
   a. This would involve comparing the current algorithm to both the indicated algorithm based only on the output of the model, as well as the proposed algorithm.
   b. A description of the proposed algorithm should include:
      i. The variables and levels included in the model, their relativities, and their associated standard errors.
      ii. Adjustments for:
         1. Competitive considerations – such as tempering relativities that appear too different from a competitor to minimize anti-selection, and reflect larger volumes of data a competitor may have available to produce more credible rating factors.
         2. Business objectives – such as forcing a particular discount to be the same across all coverages to simplify rating.
         3. Disruption mitigation - Tempering relativity changes from current to limit the rate change impact seen by the customer.

Appendix B – Additional In-Depth Model Evaluation Items

A more detailed review could include questions about the data, adjustments, and judgment that went into building the model:

1) Data
    a. Sources – both internal & external.
    b. Data definitions.
    c. Volume.
    d. Data characteristics – such as range of values, identifying outliers, volume of missing values.
    e. Time period.

2) Data Preparation
    a. Adjustments – such as capping, loss development, curve fits, NA treatment.
    b. Corrections.
    c. Transformations – such as using the log of claim amounts.
    d. Groupings – such as driver age grouping, or home value grouping.

3) Model Building
    a. Model type:
        i. Frequency, Severity, Pure Premium, Retention, etc.
        ii. Statistical distribution used i.e. Poisson, Gamma, Tweedie, etc.
        iii. Link function used.
        iv. Describe any curve fitting or special grouping used during modeling.
    b. Variable Identification & Selection
        i. Identify variables available for inclusion in the model.
        ii. For all available variables:
            1. Define variable selection criteria – examples include chi square, AIC, BIC, or other goodness-of-fit measures.
            2. Identify those variables selected for the model.
        iii. Identify any interactions between variables to be included in the model.

6. REFERENCES

[1] Mark Goldburd, Anand Khare, and Dan Tevet, "Generalized Linear Models for Insurance Rating", 2016, CAS Monograph Series Number 5
[2] Geoff Werner, and Serhat Guven, "GLM Basic Modeling: Avoiding Common Pitfalls", Casualty Actuarial Society Forum, Winter 2007
[3] ACTUARIAL STANDARD OF PRACTICE NO. 23, "Data Quality", December 2016
[4] ACTUARIAL STANDARD OF PRACTICE NO. 38, "Using Models Outside the Actuary's Area of Expertise (Property and Casualty)", May 2011
[5] ACTUARIAL STANDARD OF PRACTICE NO. 41, "Actuarial Communications", December 2010

Biographies of the Authors

Don Closter is the VP of actuarial research at Horace Mann Insurance company in Springfield, Illinois. He is responsible for Pricing research on a countrywide level. Don has degrees in Mathematics and Psychology from Kent State University in Ohio. He is an Associate of the CAS and SOA, and a Member of the American Academy of Actuaries. He is also a member of the CAS Ratemaking Committee and chair of the CAS Research Grants Task Force.

Caryn Carmean is an AVP of product management at Horace Mann Insurance Company in Springfield, Illinois. In addition to years of pricing, reserving, and product management experience in personal lines insurance, she is a former pricing regulator for the Illinois Department of Insurance. Caryn has a degree in Mathematics from University of Illinois at Springfield. She is also a member of the CAS Ratemaking Committee and a member of the American Academy of Actuaries' Casualty Committee.