And The Winner Is…? How to Pick a Better Model – Part 1

2017 CAS RPM Seminar

ISO
Verisk Insurance Solutions

SERVE | ADD VALUE | INNOVATE

Antitrust Notice

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.
- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.
- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

SERVE | ADD VALUE | INNOVATE

Motivation

- Models that appear to be strong may have weaknesses
  - Fit may not be good enough
  - Model may be overfit
  - Wrong distribution may have been chosen
  - Results may not be stable across data subsets or over time
  - Results may be highly influenced by several records
  - Model may underperform the status quo

SERVE | ADD VALUE | INNOVATE

Understanding & Validating a Model

- Model Lift
  - How well does the model differentiate between best and worst risks?
  - Does the model help prevent adverse selection?
  - Does the model improve the rating plan?
- Goodness of Fit
  - What kind of model statistics are available, and how do you interpret them?
  - What kind of residual plots should you consider, and how do you interpret them?
  - What are some considerations regarding actual versus predicted plots?
- Internal Stability
  - How well does the model perform on other data?
  - How will the model perform over time?
  - How reliable are the model's parameter estimates?

SERVE | ADD VALUE | INNOVATE

Model Lift

- Ability to differentiate between low and high cost policyholders
  - Sometimes called the "economic value" of the model
- Some tools for illustrating model lift
  - Simple quantile plots
  - Double quantile charts
  - Loss ratio charts

SERVE | ADD VALUE | INNOVATE

Model Lift – Simple Quantile Plots

- Creating a quantile plot
  - Use holdout sample.
  - Sort data based on predicted value (frequency, severity, loss cost).
  - Subdivide sorted data into quantiles (quartiles, quintiles, deciles) with equal weight (exposure, claim count).
  - Calculate average actual value and predicted value for each quantile and index to overall average.
- Checking a quantile plot
  - Is there a close match between actual and predicted values?
  - Are values increasing monotonically or with few reversals?
  - How well does the model distinguish between low cost and high cost policyholders?

SERVE | ADD VALUE | INNOVATE

Quantile Chart Examples Using Simulated Data

```
#
# Simulate data
#
# use a seed value to make results reproducible
set.seed(2017)
# simulate data and store in data frame
d <- data.frame(
     policy_number = c(1:240000, 1:240000),
     year = c(rep(2015, 240000), rep(2016,240000)),
     group = rep(c(rep('G1', 80000), rep('G2', 80000), rep('G3', 80000)), 2),
     age = rep(c(rep(1, 20000), rep(2, 20000), rep(3, 20000), rep(4, 20000)), 6),
     exposure = rep(1, 480000),
     claim_count = c(rpois(480000,
                         c(rep(0.0100, 20000), rep(0.0200, 20000), rep(0.0400, 20000), rep(0.0800, 20000),
                           rep(0.0300, 20000), rep(0.0525, 20000), rep(0.0919, 20000), rep(0.1608, 20000),
                           rep(0.0600, 20000), rep(0.0900, 20000), rep(0.1350, 20000), rep(0.2025, 20000)))),
     stringsAsFactors = TRUE
)
```


SERVE | ADD VALUE | INNOVATE

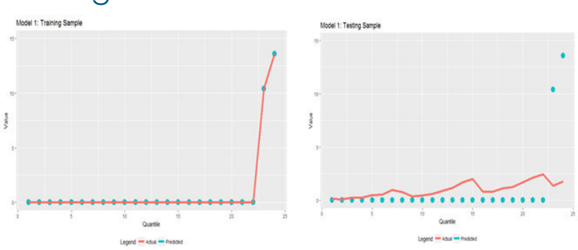
An Example of Overfitting

- Model 1

```
#
# Model 1: predicted claim count equals claim count in training sample
#
Training$m1_claim_count <- Training$claim_count
Testing$m1_claim_count <- Training$claim_count
```


SERVE | ADD VALUE | INNOVATE

Quantile Chart Examples Using Simulated Data



Model 1: For each policy, predicted claim count equals claim count in training data.

SERVE | ADD VALUE | INNOVATE

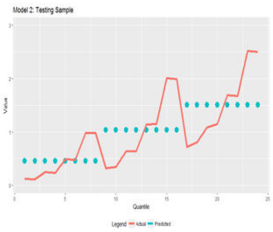
Alternative Models

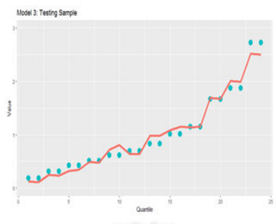
- Models 2 and 3

```
#
# Model 2: group only
#
m2 <- glm(claim_frequency ~ group,
                 family = poisson(link = log),
                 data = Training, weights = exposure)
summary(m2)
Testing$m2_frequency <- round(predict(m2, newdata = Testing, type = 'response'), 6)
Testing$m2_claim_count <- Testing$exposure * Testing$m2_frequency
#
# Model 3: group and age
#
m3 <- glm(claim_frequency ~ group + age,
        family = poisson(link = log),
        data = Training, weights = exposure)
summary(m3)
Testing$m3_frequency <- round(predict(m3, newdata = Testing, type = 'response'), 6)
Testing$m3_claim_count <- Testing$exposure * Testing$m3_frequency
```


SERVE | ADD VALUE | INNOVATE

Quantile Chart Examples Using Simulated Data





Model 2: Group
Model 3: Group and Age

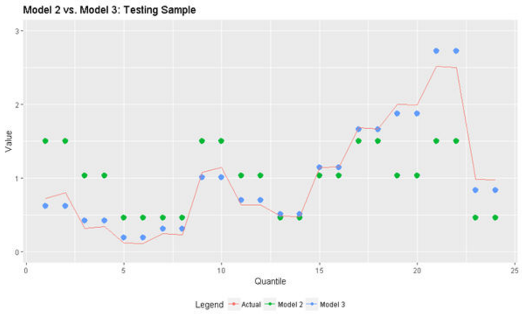
SERVE | ADD VALUE | INNOVATE

Model Lift – Double Quantile Charts

- Creating a Double Quantile chart
  - Sort data by ratio of model prediction to current premium.
  - Subdivide the sorted data into quantiles with equal exposure.
  - For each quantile calculate average actual loss cost (frequency or severity), average model predicted value, and the average value underlying the current manual premium.
  - Index the quantile averages to the overall averages.

SERVE | ADD VALUE | INNOVATE

Double Quantile Chart Example Using Simulated Data



13

SERVE | ADD VALUE | INNOVATE

Alternative Models

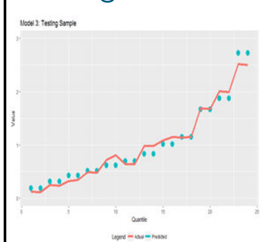
- Models 3 and 4

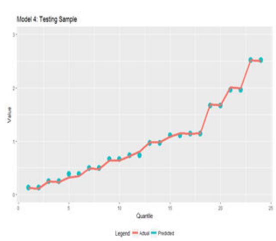
```r
#
# Model 3: group and age
#
m3 <- glm(claim_frequency ~ group + age,
          family = poisson(link = log),
          data = Training, weights = exposure)
summary(m3)
Testing$m3_frequency <- round(predict(m3, newdata = Testing, type = 'response'), 6)
Testing$m3_claim_count <- Testing$exposure * Testing$m3_frequency
#
# Model 4: group, age and interaction
#
m4 <- glm(claim_frequency ~ group + age + group*age,
          family = poisson(link = log),
          data = Training, weights = exposure)
summary(m4)
Testing$m4_frequency <- round(predict(m4, newdata = Testing, type = 'response'), 6)
Testing$m4_claim_count <- Testing$exposure * Testing$m4_frequency
```


14

SERVE | ADD VALUE | INNOVATE

Quantile Chart Examples Using Simulated Data



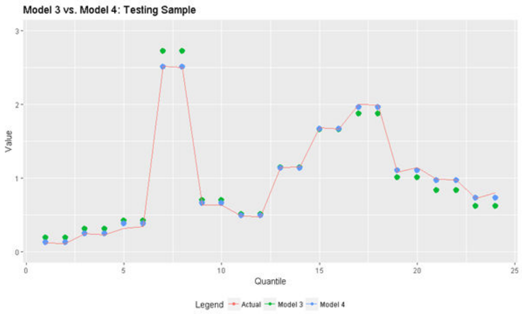


Model 3: Group and Age
Model 4: Group, Age, and interaction

15

SERVE | ADD VALUE | INNOVATE

Double Quantile Chart Example Using Simulated Data

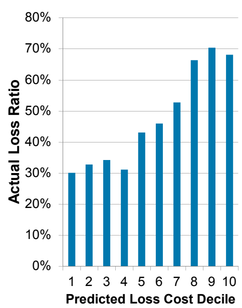


16

SERVE | ADD VALUE | INNOVATE

Model Lift – Loss Ratio Charts

- Quantile charts and Double Quantile charts may be unfamiliar to some stakeholders
- Loss ratios are widely used and understood in the industry
- Ranking by predicted loss cost
  - Rank data into quantiles by predicted model loss cost
  - Calculate loss ratio for each quantile



SERVE | ADD VALUE | INNOVATE

Model Lift – Summary

- Simple Quantile plots
  - Illustrate how well the model helps prevent adverse selection
- Double Quantile charts
  - Compare competing models
  - Compare new model against current rating plan
- Loss ratio charts
  - Puts lift in context most people in insurance industry can understand
  - Can be distorted by redundancy or inadequacy of current rating plan

SERVE | ADD VALUE | INNOVATE

References

- De Jong, P. and Heller, G. Z., *Generalized Linear Models for Insurance Data*, Cambridge University Press, 2008
- Dickey, D. A., "Finding the Gold in Your Data: An Overview of Data Mining", SAS Global Forum 2013
- Frees, E.W., Derrig, R. A., and Meyers, G., *Predictive Modeling Applications in Actuarial Science*, Cambridge University Press, 2014
- Jaffery, T. and Liu, S. X., "Measuring Campaign Performance by Using Cumulative Gain and Lift Chart", SAS Global Forum, 2009
- May, E., *Handbook of Credit Scoring*, Global Professional Publishing, 2001
- Parr Rud, O., *Data Mining Cookbook*, John Wiley & Sons, 2001

Hernan L. Medina, CPCU
Senior Principal Data Scientist
ISO Solutions
545 Washington Boulevard
Jersey City, NJ 07310-1686

Hernan.Medina@verisk.com
http://www.verisk.com/iso

This material was used exclusively as an exhibit to an oral presentation. It may not be, nor should it be relied upon as reflecting, a complete record of the discussion.

ISO
Verisk Insurance Solutions