And The Winner Is...? How to Pick a Better Model

Model Lift – 2016 CAS RPM Seminar

Verisk Analytics | ISO

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.

Motivation

- Models that appear to be strong may have weaknesses
  - Fit may not be good
  - 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

Motivation

- Models that appear to be strong may have weaknesses
  - Fit may not be good
  - 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

Some Models Used by Actuaries

- Linear regression
- Exponential regression
- Logistic regression
- Minimum bias procedures
- Generalized linear models
- Classification and regression trees
- Clustering procedures

Understanding & Validating a Model

- Model Lift
  - How well does the model differentiate between best and worst risks?
  - Does the model help prevent adverse selection?
  - Is the model better than the current 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?

Model Lift

- Ability to differentiate between low and high cost policyholders
  - Sometimes called the “economic value” of the model

- Some tools for measuring and illustrating model lift
  - Simple Quantile plots
  - Double lift charts
  - Gini index
  - Loss ratio charts

Model Lift – Simple Quantile Plots

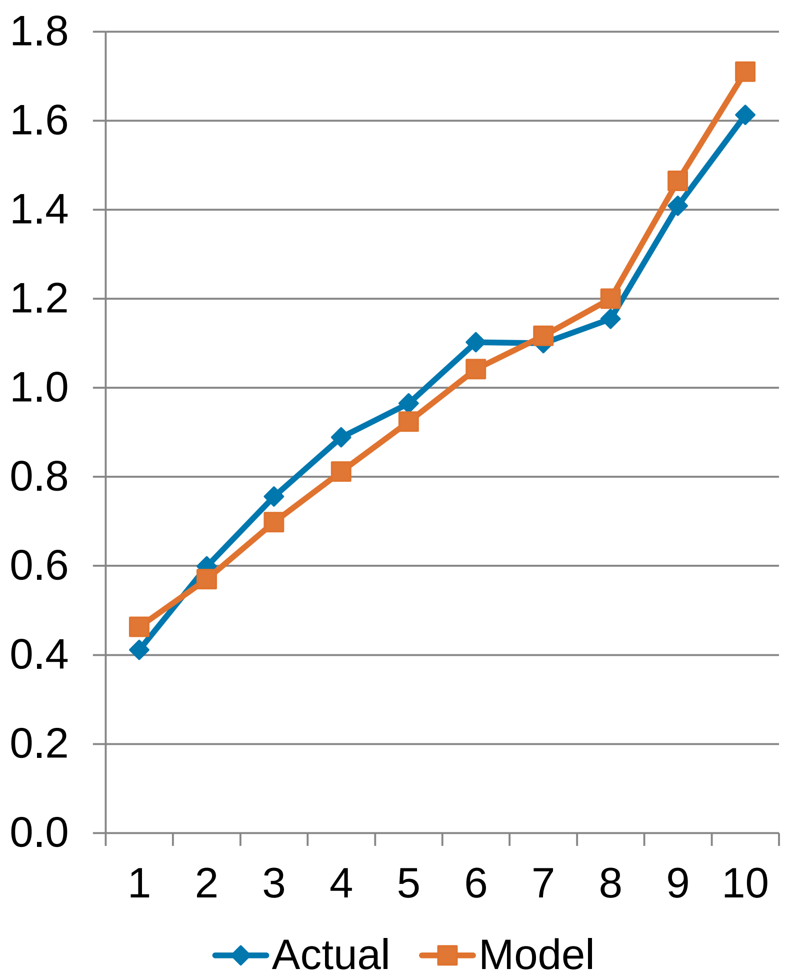
- Creating a quantile plot
  - Use holdout sample.
  - Sort data based on predicted value (frequency, severity, loss cost).
  - Subdivide sorted data into quantiles (quartiles, quintiles, deciles) with equal weight (exposure, claim count).
  - Calculate average actual value and predicted value for each quantile and index to overall average.

Decile	Actual Pure Premium	Model Pure Premium	Actual Index	Model Index
1	41.10	47.23	0.411	0.463
2	59.90	58.14	0.599	0.570
3	75.60	71.20	0.756	0.698
4	88.90	82.82	0.889	0.812
5	96.50	94.25	0.965	0.924
6	110.20	106.28	1.102	1.042
7	110.00	113.93	1.100	1.117
8	115.50	122.40	1.155	1.200
9	140.90	149.43	1.409	1.465
10	161.30	174.42	1.613	1.710
All	100.00	102.00		

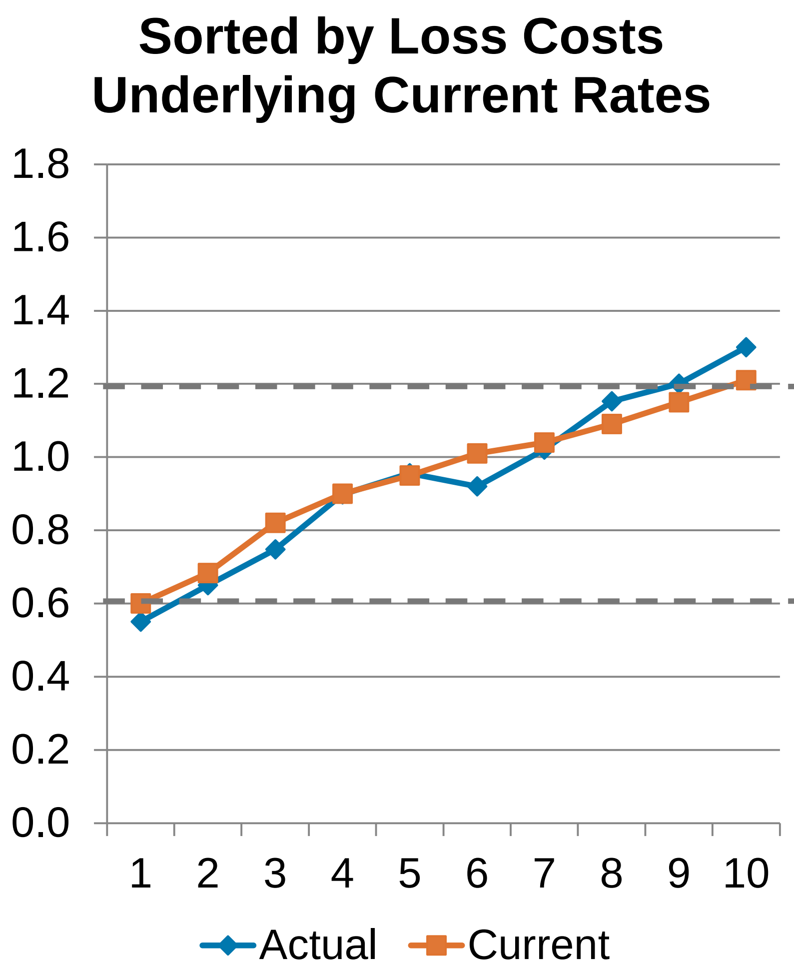
Hypothetical

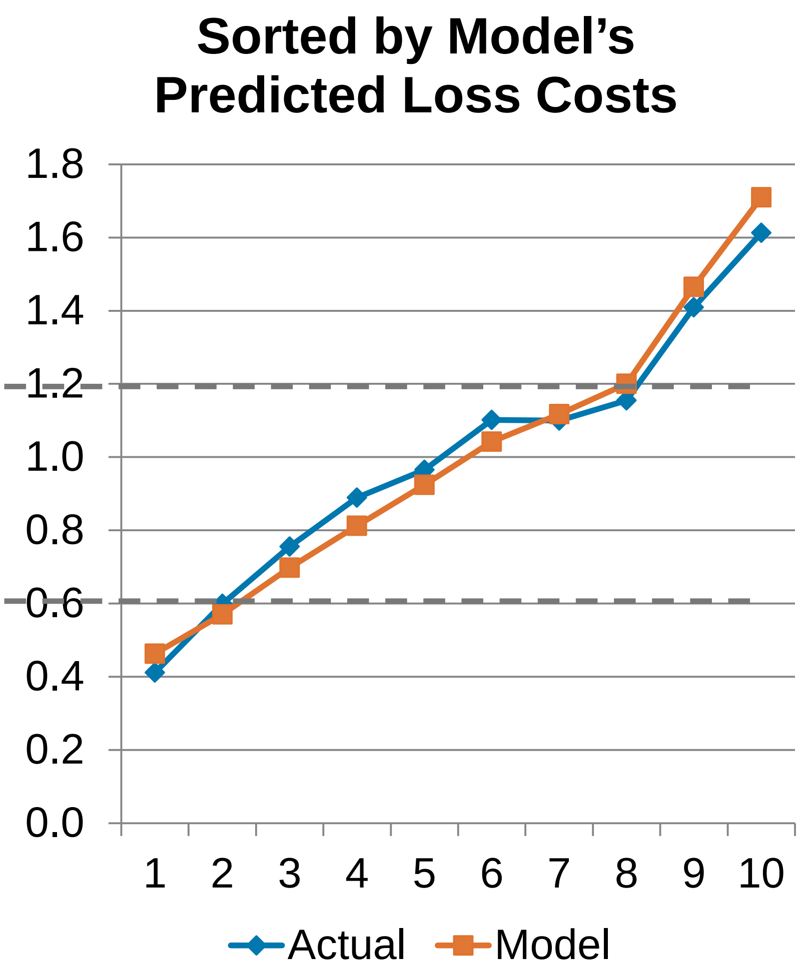
Model Lift – Simple Quantile Plots

- 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 costs policyholders?



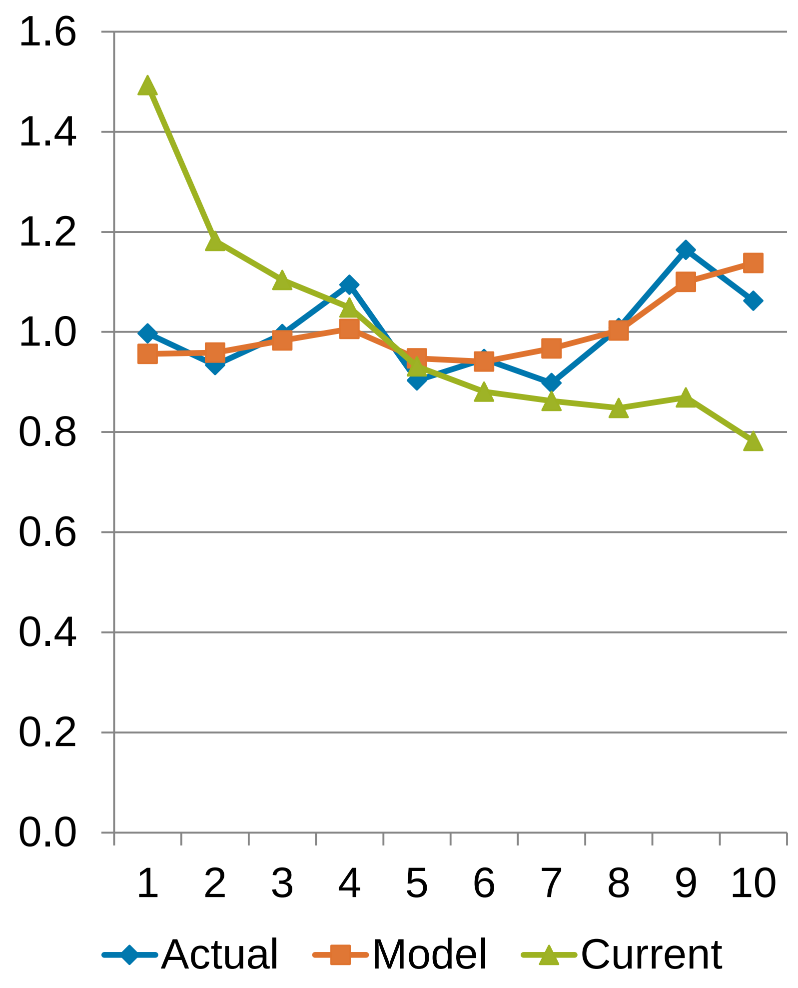
Model Lift – Simple Quantile Plots





Model Lift – Double Lift Charts

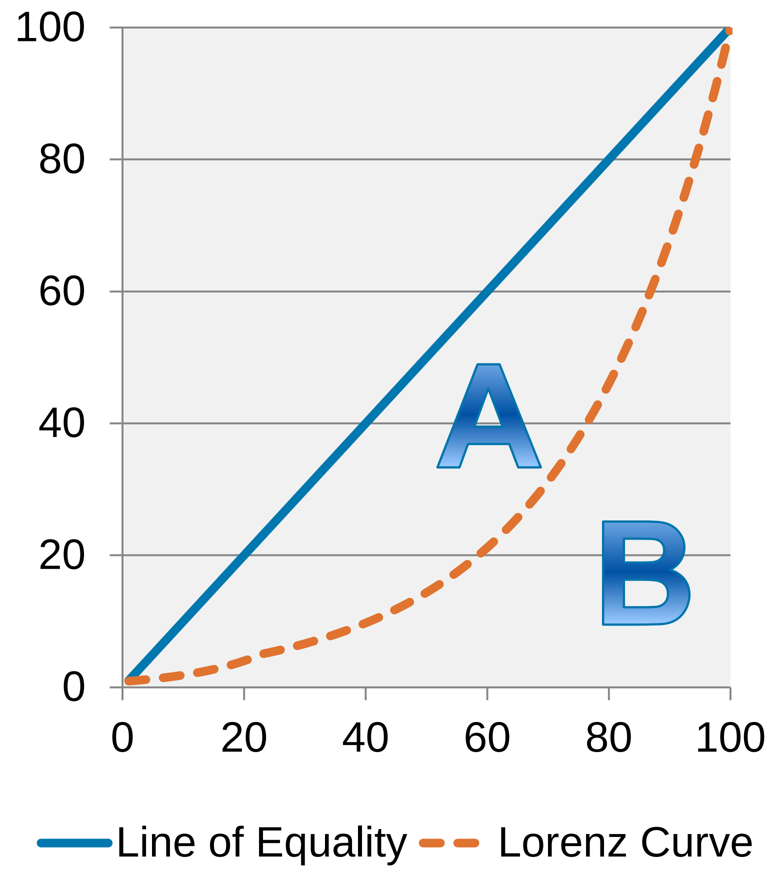
- Creating a double lift chart
  - Sort data by ratio of model prediction to current premium.
  - Subdivide sorted data into quantiles with equal exposure.
  - For each quantile calculate average actual loss cost, average model predicted loss cost and the average loss cost underlying the current manual premium .
  - Index the quantile averages to the overall averages.



Economics – The Gini Index

- Gini coefficient or Gini ratio
  - Named after Corrado Gini

- Measure of income inequality
  - Horizontal axis = percentage of country's population
  - Vertical axis = percentage of country's income
  - A = Area between line of equality and Lorenz Curve
  - B = Area beneath Lorenz Curve
  - Gini index = A / (A + B)



Model Lift – Simple Gini Index

Binary Response

- SAS Proc Logistic
- t = total pairs with different responses
- n_c = concordant pairs
- n_d = discordant pairs
- $t - n_c - n_d$ = tied pairs
- Sommer's D

  = Gini's coefficient

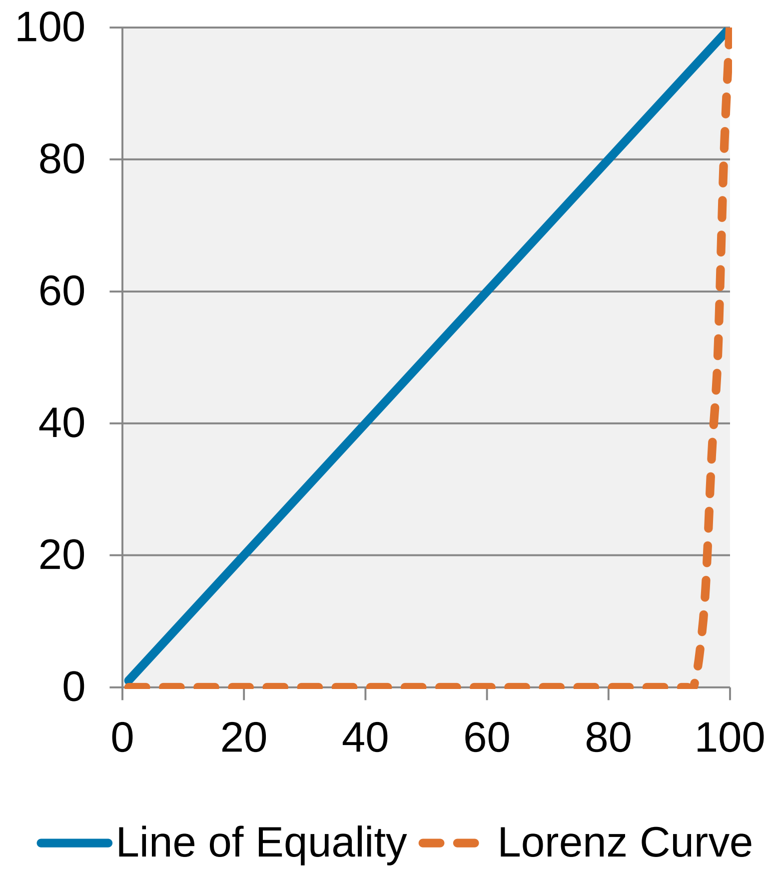
  $= (n_c - n_d) / t$

Lorenz Curve for Loss Cost

- Sort holdout data by predicted loss cost
- Calculate cumulative percentages for actual incurred loss and exposure
- Plot
  - Cumulative exposure in horizontal axis
  - Cumulative loss in vertical axis

Model Lift – Simple Gini Index

- Adapting to car insurance
  - Assume claim frequency = 5%
- "The perfect model"
  - Prediction = actual loss, which is $0 for 95% of vehicles insured.
  - Sort holdout data set by model prediction.
  - Horizontal axis = percentage of total earned car years.
  - Vertical axis = Percentage of total incurred loss.
  - Gini Index = A / (A+ B) is very high.



Model Lift – Simple Gini Index

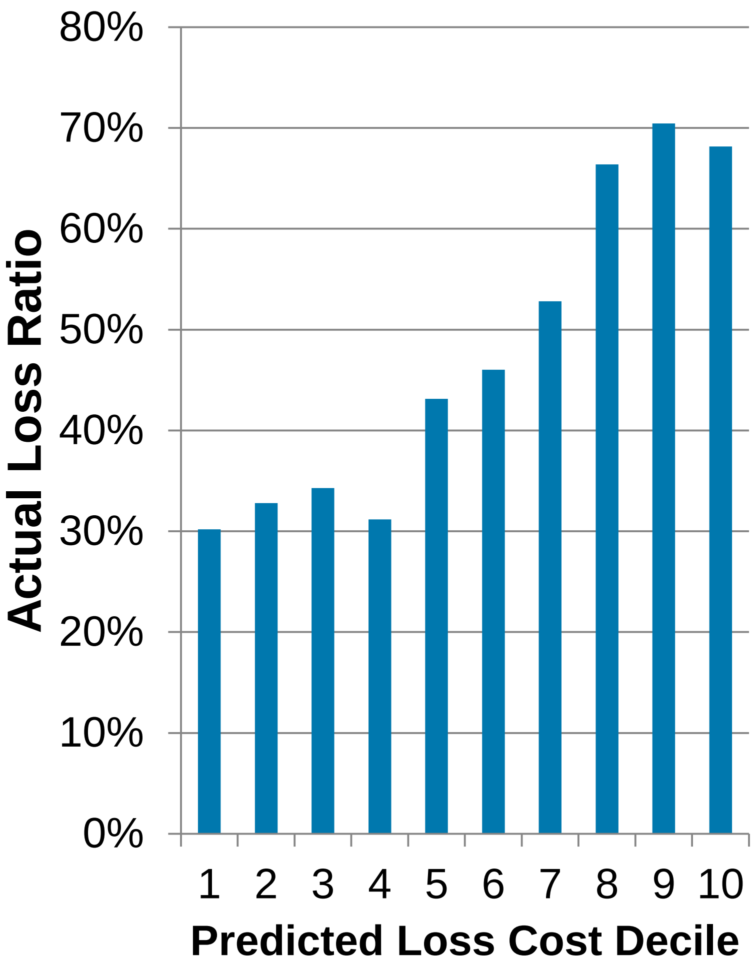
- Exercise:
  - Model X prediction = expected loss cost
  - Model Y prediction = 0.5 (expected loss cost)
  - Model Z prediction = 2.0 (expected loss cost)
  - Which model has the highest Gini index?

- Model A has a Gini index of 15.9 and B has a Gini index of 15.4
  - Is that difference significant, or is it just a quirk of the holdout data?

SERVE | ADD VALUE | INNOVATE

Model Lift – Loss Ratio Charts

- Lift charts and Gini index
  - May be unfamiliar to some stakeholders
- Loss ratios
  - 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



Model Lift – Summary

- Simple Quantile plots
  - Illustrate how well the model helps prevent adverse selection
- Double lift charts
  - Compare competing models
  - Compare new model against current rating plan
- Simple Gini Index
  - Summarizes model lift into one number
- 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

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
- Gini, C. "On the Measure of Concentration with Special Reference to Income and Wealth", Cowles Commission Research Conference on Economics and Statistics, Colorado College Publication, General Series No. 208, 73–79, 1936.
- 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
Director, Analytical Data Management
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.