The Revolution and Evolution of Predictive Modeling

Session C-24

A presentation for the 2012 CAS Spring Meeting
By Claudine Modlin, FCAS, MAAA

May 2012

TOWERS WATSON

What did pricing/underwriting look like before 2000?

- Throes of credit controversy
- Legacy systems and little attention to enterprise data initiatives
- Mainframe computers
- Consistent rating plans…which we all understood
- Pricing departments staffed with actuaries
- “High touch” underwriting in small commercial
- Machine learning (to many of us) meant learning how to use new software
- Price optimization relied on collective judgment, not mathematical algorithms

towerswatson.com

1

What changed?

- Computing power
- Better access to better data
- Regulation in support of multivariate analysis (in response to credit controversy)
- Statistical rigor
- Product management culture
- Competition

Agenda

- The predictive modeling revolution/evolution in insurance pricing (presented by Claudine)
  - Estimating claims costs
  - Understanding policyholder demand
  - The road to price integration
- Venn diagram of data science
- The revolution spreads (presented by Steve)
  - Operational efficiency
  - Underwriting
  - Marketing
  - Claims
  - Agency

Predictive modeling in insurance pricing

- Late 1990s — A major revolution in pricing analysis

Statistical framework

- Ongoing — An evolution of pricing refinement

Existing data and existing model structure → **New predictors and modified model form** → **Model refinement** → **Integration of cost and demand**

Math and statistics knowledge

Parametric Modeling	Machine Learning Tools
• Objective: build a predictive model • User makes assumptions (e.g., distribution, model structure) and specifies preliminary list of explanatory variables • User guides statistical method in order to effectively describe a particular response (e.g., claim frequency) • Result is an algorithm, a set of parameters, and diagnostics • Examples: minimum bias methods, linear regression, GLM	• Objective: learn new things (which may help in building a model) • Find patterns (often complex) in an unknown underlying distribution • Tool may be supervised, unsupervised, or blend of the two • Result might be a new variable, a tree, a grouping, a score, etc. • Examples: principal components analysis, decision trees, clustering, artificial neural networks

GLMs: the global industry standard for pricing

- Benefits of generalized linear models
  - Multivariate method accounts for exposure correlations between variables
  - Allows modeler to capture signal and remove noise within statistical framework but also infuse business knowledge
  - Provides useful diagnostics
  - Incorporates interactions
  - Transparent, easy to explain

Selecting response variable for insurance pricing models

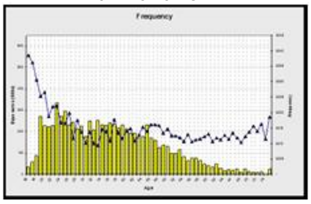
Early Thinking:	Current Practice:

• Fit models to loss ratio
• Define claim types according to how currently priced
• Use existing rating variables (or a subset)
	• Fit models to frequency and severity, or directly to loss costs
	• Define claim types, balancing homogeneity and credibility
	  – Dispersion modeling can address some degree of heterogeneity
	  – Alternative model structures can be used for low volume claim types: Frequency (3rd party claim) x prob (BI claim / 3rd party claim) x Severity (BI claim)
	• Explore various sources of data

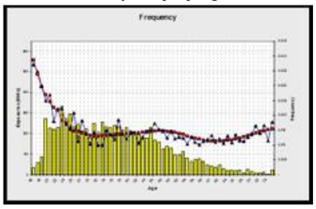
Pure premium vs. loss ratio

- When viewing frequency and severity data separately, easy to discern patterns from the noise; more difficult with loss ratio

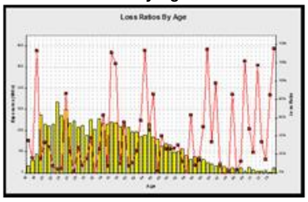
Raw Frequency by Age of Driver



Smoothed Frequency by Age of Driver



Loss Ratio by Age of Driver

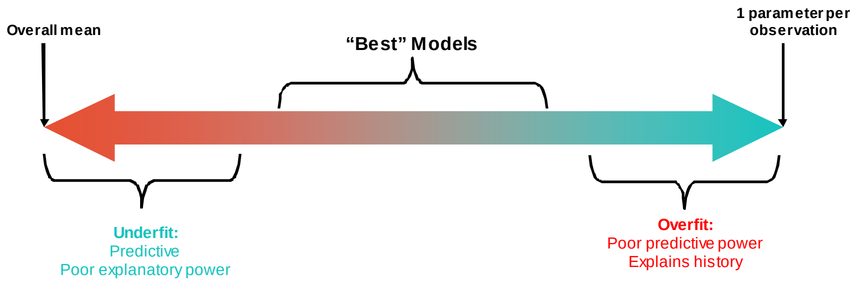


towerswatson.com

8

Factor selection

- Produce a sensible model that explains recent historical experience and is likely to be predictive of future experience

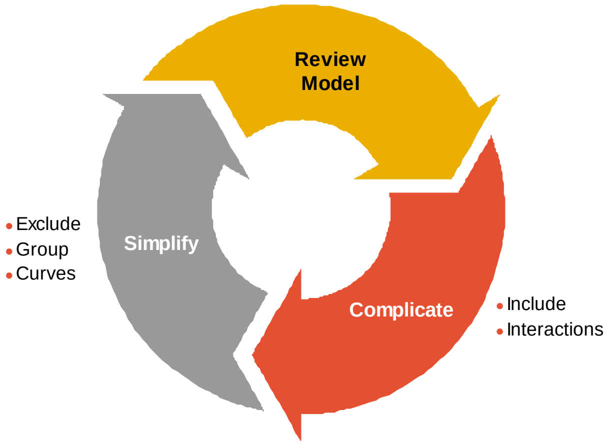


towerswatson.com

9

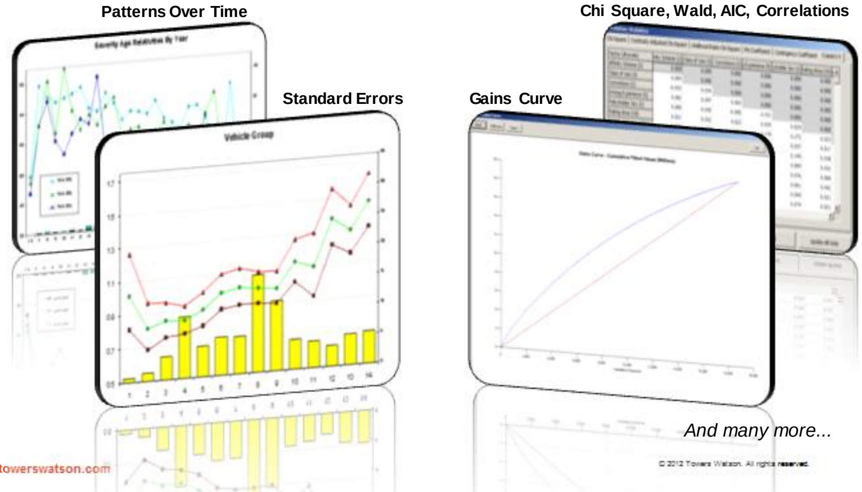
Factor selection

- Factor selection is an iterative process — involving simplification as well as complication of the model form



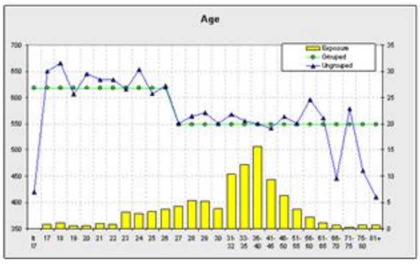
Factor selection

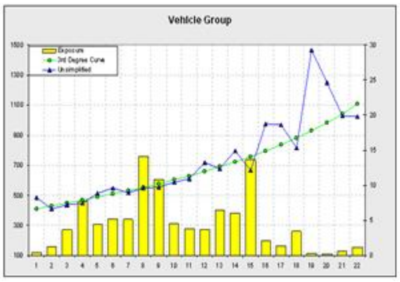
- Iterate models using statistical diagnostics, practical tests, and business knowledge to avoid overfitting



Factor selection

- Seek parsimony
  - Group similar discrete levels to reduce volatility
  - Fit curves to continuous variables with natural order





A plethora of predictors

- Early GLMs analyzed traditional rating variables (or subset)
- Current practice is to survey all sources of predictive data — including external sources

Policy (e.g., minimum age of driver on policy)	**Coverage** (e.g., limits, deductibles)	**Risk** (e.g., age of home, type of car, industry class)
Relationship with insurance company (e.g., tenure, distribution channel, affinity)	**Insured** (e.g., age)	**Financial attributes** (e.g., insurance credit score)
Payment and billing information (late pays, payment frequency)	**Prior claims experience**	**Geography/Environment** (including geo-demographics)
Other lines of business and related claims experience		

- Consider the explosion of data with usage-based insurance!

Coping with large numbers of related factors

- Can be hard to interpret output from a GLM that includes a very large number of related characteristics
- Best to prune the list using variable reduction techniques
  - One-way analysis and business judgment
  - Test "families" of predictors one at a time to find most predictive members
  - Limited forward regression
  - Principal components analysis
  - Factor analysis
  - Classification and Regression Trees (CART)
  - Random forests

Principal Components Analysis

- Example: External geodemographic data

GEO_ UNIT	GEO_POP _DENSITY	GEO_MED_ AGE	GEO_ UNEMP
A	100,000	34	5%
B	50,000	55	6%
…	…	…	…

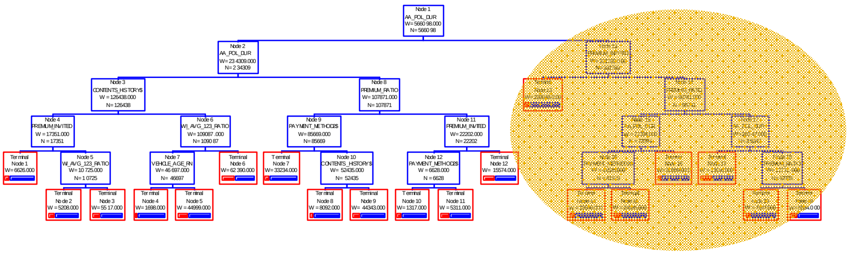
X

GEO	PC1	PC2	PC3
GEO_POP_DENSITY	1.50	0.75	0.68
GEO_MED_AGE	0.40	1.20	0.34
GEO_UNEMP	2.00	3.00	1.50

GEO_ UNIT	PC1 SCORE	PC2 SCORE	PC3 SCORE
A	150,013	75,040	68,011
B	75,022	37,566	34,018
…	…	…	…

Classification and Regression Trees (CART)

- Example: indicating localization strategies

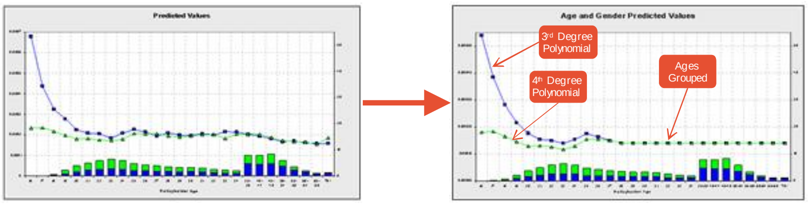


- Significantly different branch structure suggests data split and model localization

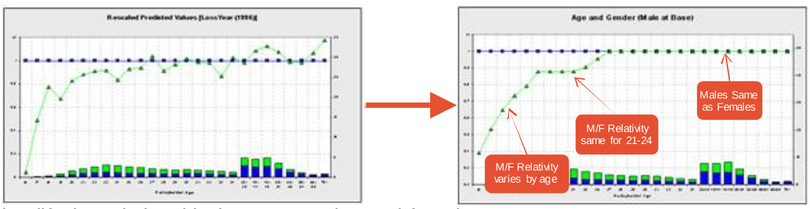
towerswatson.com

Interactions: Detection and simplification

- Complex relationships can be simplified using curves, groups, etc.



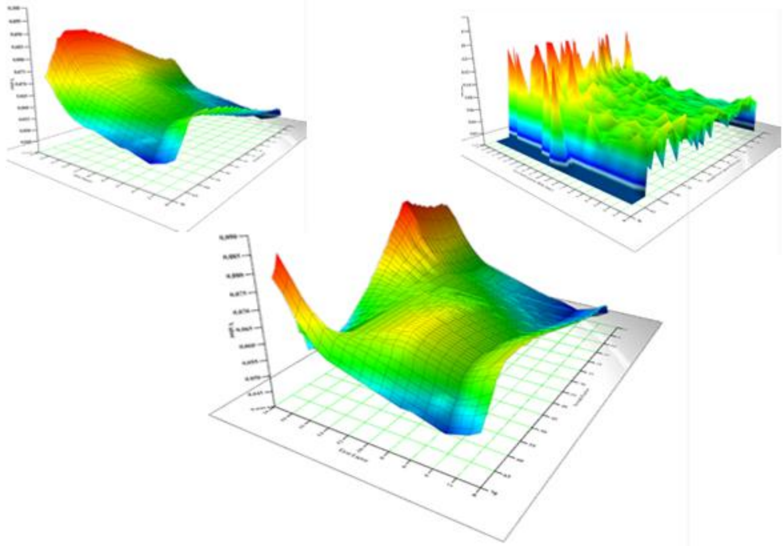
Simplify the age curve (i.e., the male age curve since male is base level)



Simplify the relationship between males and females

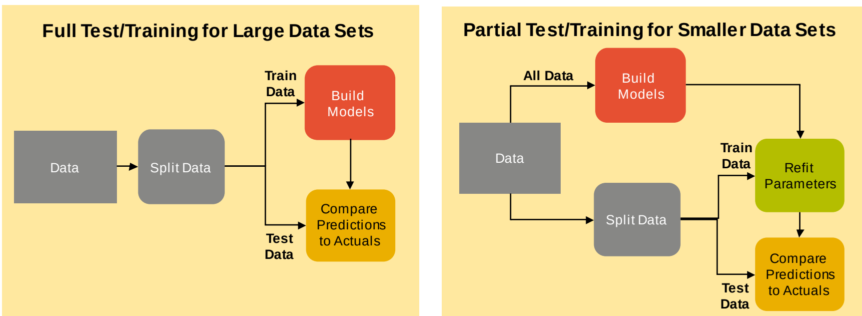
towerswatson.com

Interactions: Detection and simplification



Model validation

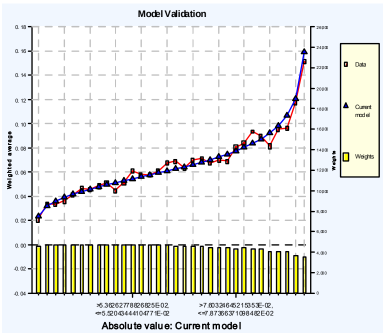
- Holdout samples are effective at validating model
  - Determine estimates based on part of data set
  - Use estimates to predict other part of data set

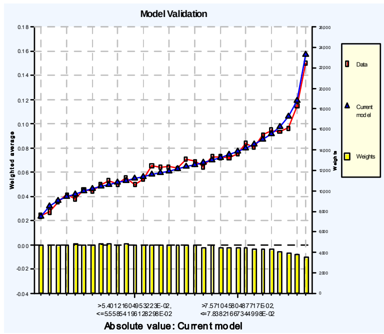


- Predicted values should be close to actual values for populated cells

Model validation

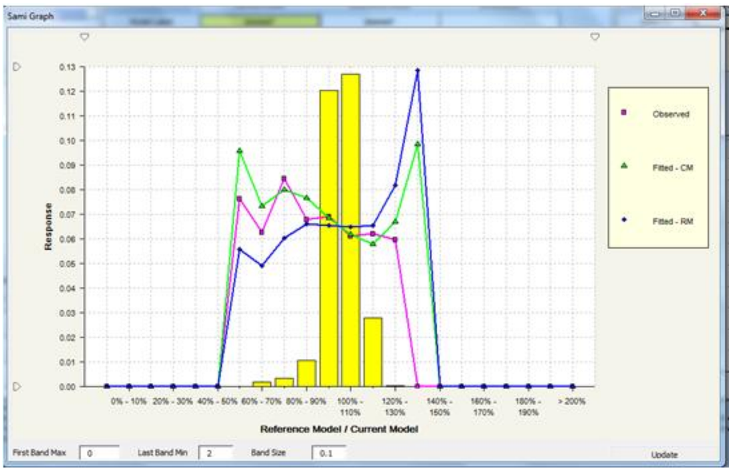
- Useful to track how well model fits to hold-out observations





- But difficult to assess performance between models

Model validation



- Comparing model performance plotted against the ratio of the two models is a more telling and less biased comparison

Noise reduction

- “Case Deleted Deviance” by Tony Lovick & Peter Lee (can be found at www.actuaries.org.uk)
- Implicitly dampens parameters in consideration of variability of parameters
- Factor selection no longer limited to in/out but rather in/out/dampened

REDEFINING THE DEVIANCE OBJECTIVE
FOR GENERALISED LINEAR MODELS

BY A.C. LOVICK AND P.K.W. LEE

[Presented to the Institute and Faculty of Actuaries: London: 28 March 2011; Norwich: 6 June 2011]

ABSTRACT

This paper defines the ‘Case Deleted’ Deviance - a new objective function for evaluating Generalised Linear Models, and applies this to a number of practical examples in the pricing of general insurance. The paper details practical approximations to enable the efficient calculation of the objective, and derives modifications to the standard Generalised Linear Modelling algorithm to allow the derivation of scaled parameters from this measure to reduce potential over fitting to historical data. These scaled parameters improve the predictiveness of the model when applied to previously unseen data points, the most likely being related to future business written. The potential for over fitting has increased due to number of factors now used, particularly in pricing personal lines business and the advent of price comparison sites which has increased the penalties of mis-estimation. New material in this paper has been included in a UK patent application No. 1020091.3.

KEYWORDS

Generalised Linear Modelling; General Insurance Pricing; Parameter Uncertainty; Case Deletion; Deviance; Non-Linear Modelling; Demand Modelling; Price Comparison Site Pricing; Winner’s Curse.

CONTACT ADDRESSES

Tony Lovick, Towers Watson, Saddlers Court, 64 - 74 East Street, Epsom, Surrey, KT17 1HB.
Tel: +44 (0)1372 751060; E-Mail: Tony.Lovick@TowersWatson.com

Peter Lee, Towers Watson, Saddlers Court, 64 - 74 East Street, Epsom, Surrey, KT17 1HB.
Tel: +44 (0)1372 751060; E-Mail: Peter.Lee@TowersWatson.com

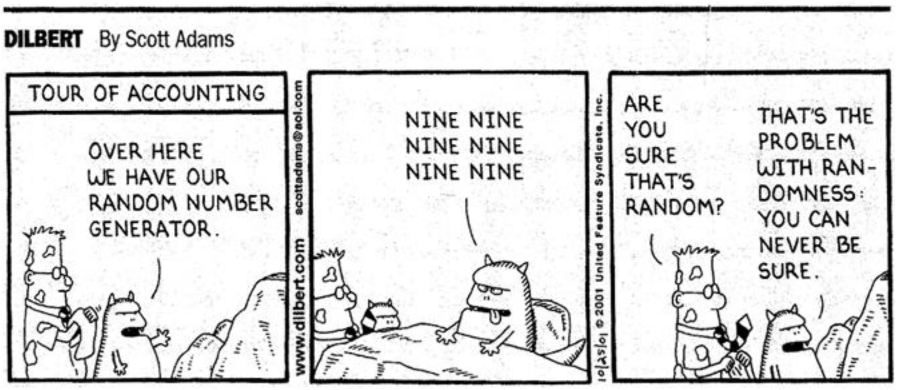
towerswatson.com 22

Mining GLM residuals

- There may be unexplained predictive power in GLM residuals
- Supervised machine learning tools can mine residuals from a GLM and develop algorithms that group risks with similar residuals
- Results can form basis of a single correction factor to the GLM
- Potential disadvantages of this approach:
  - Hard to distinguish signal from noise in the residual when no basis for evaluating residual
  - Prone to overfitting
  - Difficult to understand and explain effect on model, which can lead to implementation issues

towerswatson.com 23

Mining GLM residuals



towerswatson.com

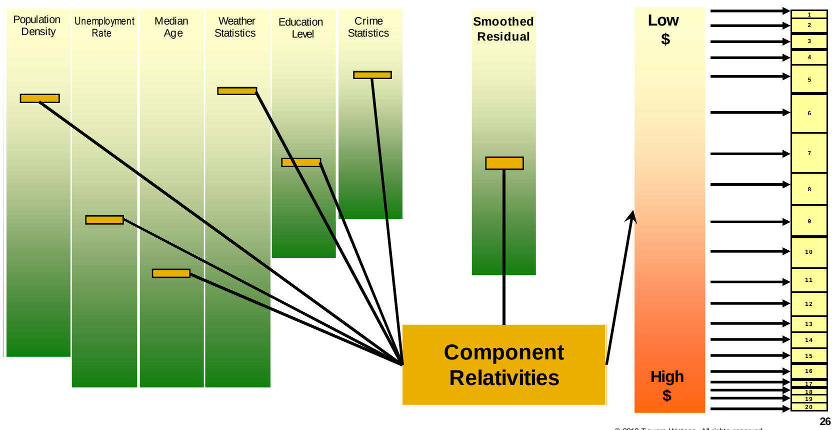
Mining GLM residuals

- Current thinking is to identify additional signal in residuals that can be attributed to a particular high-dimension factor — for example,
  - Geography (zip code)
  - Vehicle (VIN)
  - Worker compensation SIC code
  - Any factor requiring a large number of small units as building blocks — and many building blocks have little or no claims experience
- A Bayesian-based data mining method that utilizes the signal in the residuals to “correct” the GLM results for that high-dimension factor is easier to control and understand

towerswatson.com

Geographic spatial analysis

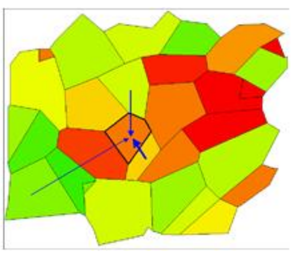
- Spatial Correction
  - Residual signal used to adjust score from the multivariate model



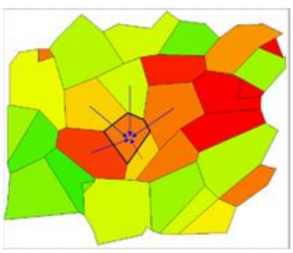
Geographic spatial analysis

- Spatial Smoothing Methods: Uses knowledge of surrounding areas to enhance estimates of the underlying risk in each area based on "Principle of Locality"

Distance-based Methods

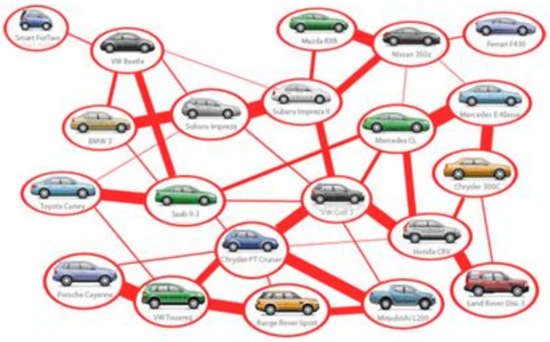


Adjacency-based Methods



Vehicle grouping analysis

- Neighboring vehicles: Instead of using latitude/longitude to build adjacency relationships, use vehicle dimensions



- Once neighbors are determined, similar techniques used for geographic analysis can be applied

towerswatson.com

28

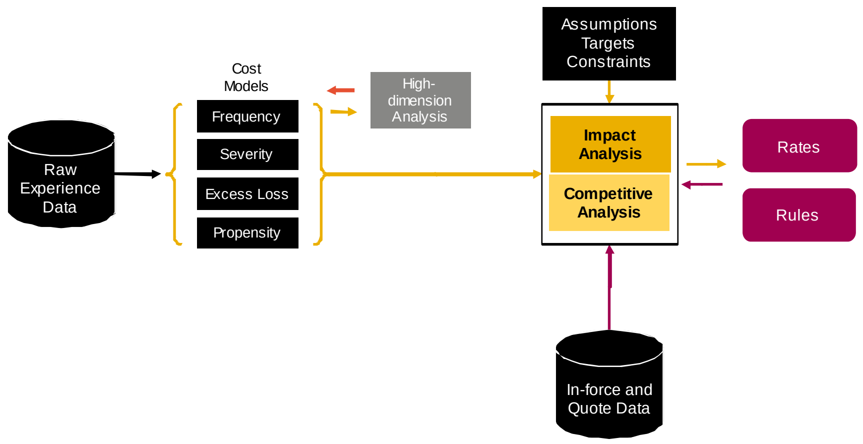
Agenda

- The predictive modeling revolution/evolution in insurance pricing
  - Cost estimation
  - Understanding policyholder demand
  - The road to price integration
- Venn Diagram of Data Science
- The revolution spreads
  - Operational efficiency
  - Underwriting
  - Marketing
  - Claims
  - Agency

towerswatson.com

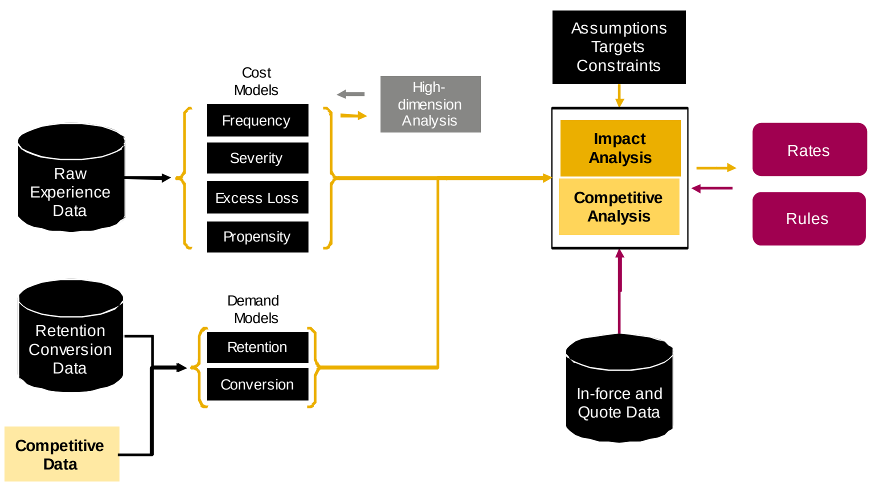
29

Traditional pricing



towerswatson.com

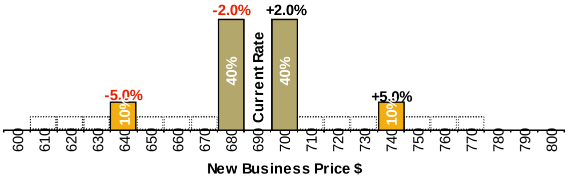
Integrated pricing



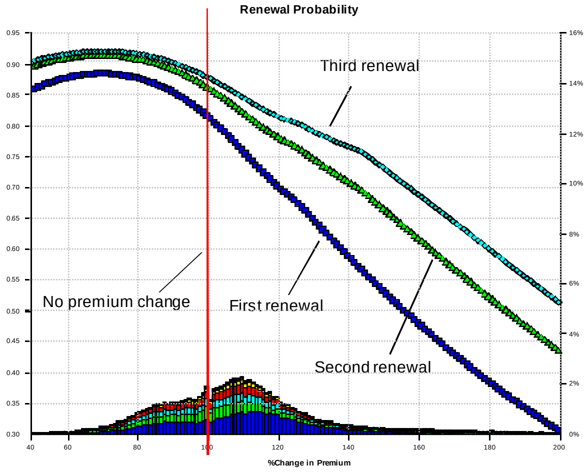
towerswatson.com

Policyholder demand models

- Fit demand models separately for new business conversion and renewal
- Demand model is a logistic regression GLM (i.e., Y-variate of GLM is "did they buy, yes/no")
- Models should include
  - Price-related variables (e.g., quoted premium, price change at renewal, competitive measures)
  - Non-price variables (e.g., policy tenure, age of insured, payment method)
- Best to have robust spread of de-correlated price changes

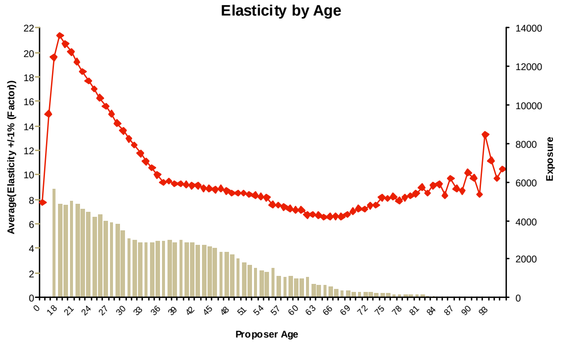


Policyholder demand models



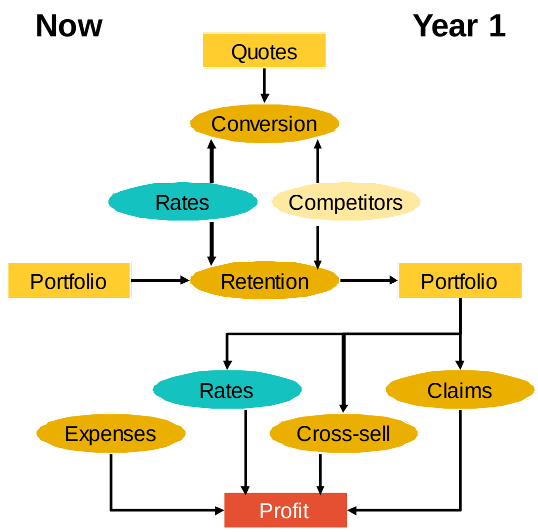
Understanding elasticity

- Focus on price-related explanatory variables different
- Can re-express as elasticity by wobbling price explanatory variables after fitting model

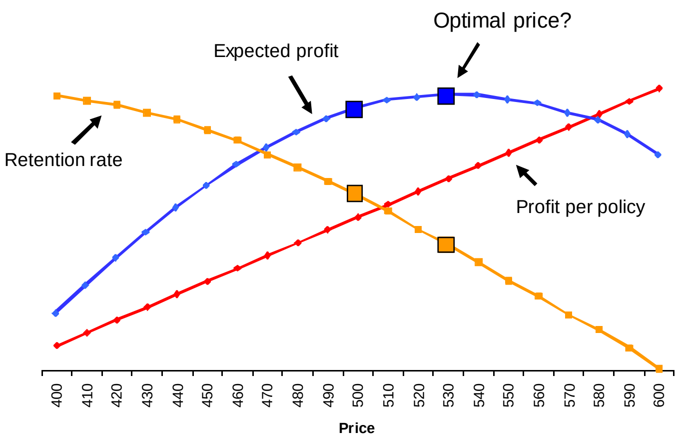


- GLMs can produce negative elasticity; requires complex interaction strategies

Integrating cost and demand to project volume and profit

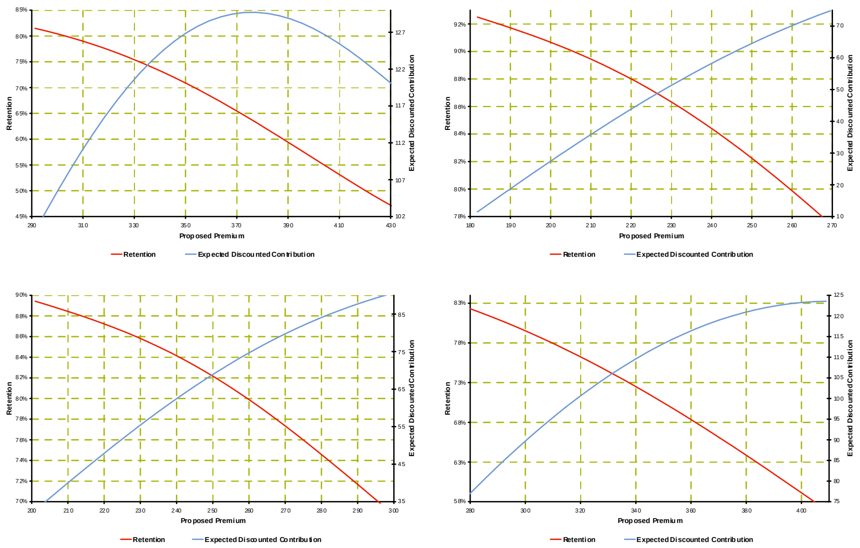


Results for one policy



Premium for policy in question

Results for multiple policies…creating a search space



Constraints provide rules to limit the search space

- The solution reflects a wide array of constraints

- Universal constraints
  - Regulatory
  - Legal
  - Corporate

- Local constraints for different optimization runs
  - Maximum rate changes
  - Competitive positions
  - Profit ranges

Strategy 1 | Strategy 2 | Strategy 3 | Strategy 4

…

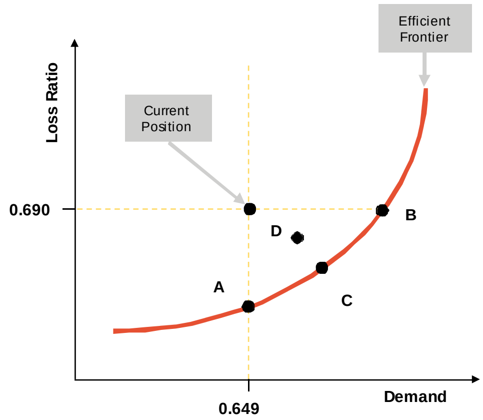
Balancing profit and volume

- Can optimize
  - Profit for a particular volume, or
  - Volume for a particular profit

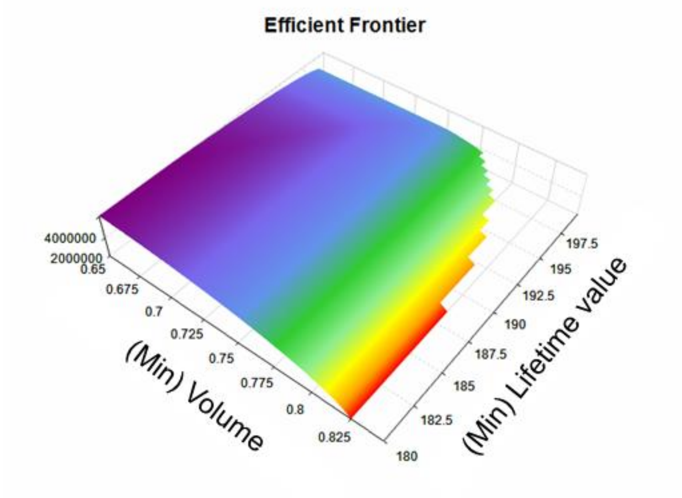
  over a defined time horizon
- Try different options to understand different balances available
- Generates efficient frontier, which aids understanding of target selection

Optimization targets

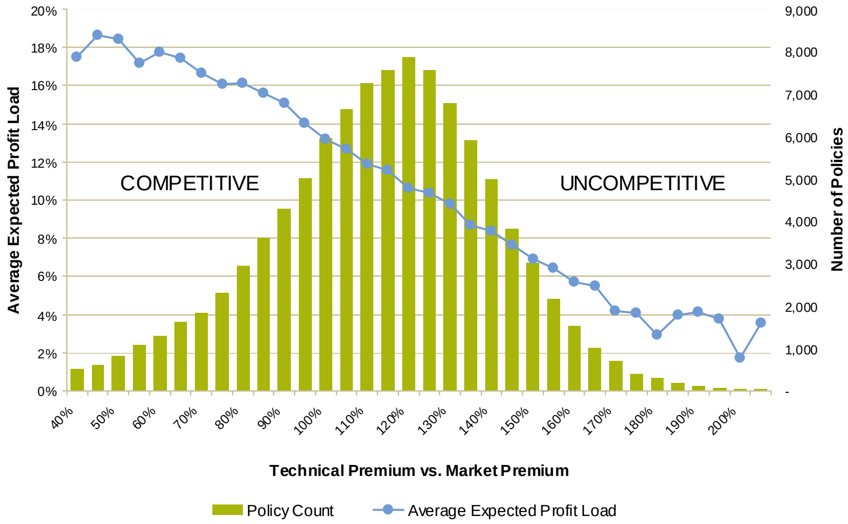
- Efficient frontier
  - Maximize profits (A)
  - Maximize volume (B)
  - Increase profits and volume (C)
  - Softer targets (e.g., business mix) (D)



Multidimensional optimization

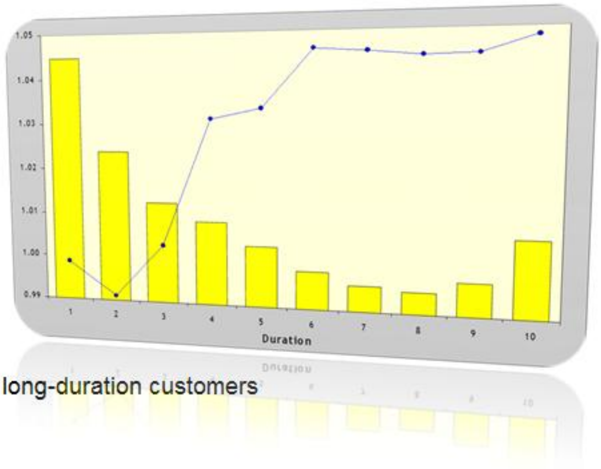


Example diagnostics



Optimization diagnostics

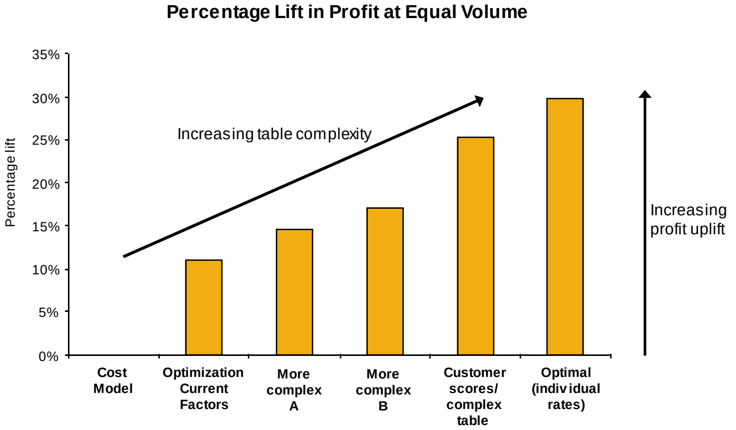
- Duration analysis



- Not overexploiting long-duration customers

Implementation

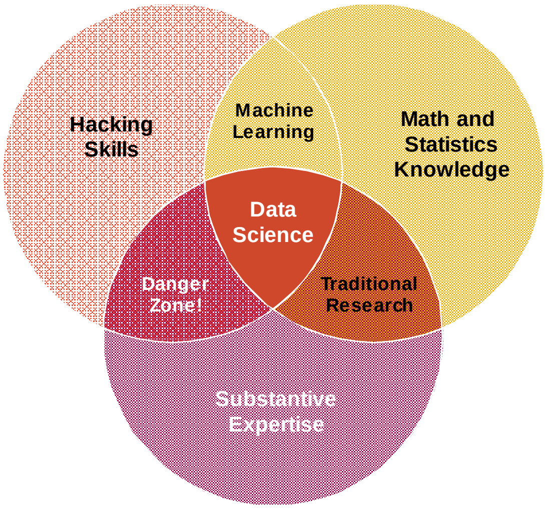
Profit Uplift Comparison



As our pricing toolkit evolves…

- GLMs continue to be an accurate and useful tool in pricing
- Make sure your GLMs follow best practices and refinements
- Investigate other analytical methods in order to understand data better and to improve accuracy of models
- Consider practical implications (usefulness) of new pricing tools
  - Easy to understand and communicate
  - Available in a timely manner
  - Capable of implementation
- Strive to understand the policyholder's reaction to price through demand modeling
- Consider price optimization as a scientific approach to select deviations from cost-based indications that achieve volume/profit targets within specified constraints

Data Science Venn Diagram



The Data Science Venn Diagram by Drew Conway in Zero Intelligence Agents blog, September 30th, 2010.

Contact details

TOWERS WATSON

Claudine Modlin, FCAS, MAAA

901 North Glebe Road
Arlington, VA 22203
312.261.9631
claudine.modlin@towerswatson.com