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.**

EagleEye Analytics

Enhancing Generalized Linear Models using Rule Induction

CAS In Focus Seminar 4 October 2011 **Christopher Cooksey, FCAS, MAAA**

EagleEye Analytics

Agenda…

1. Signal Beyond GLMs – Theory
2. Machine Learning & Rule Induction
3. Signal Beyond GLMs – Case Study
4. Possible Changes to the GLM Development Process
5. Model Development Case Study
6. Other Variations
7. Summary

1.

Signal Beyond GLMs - Theory

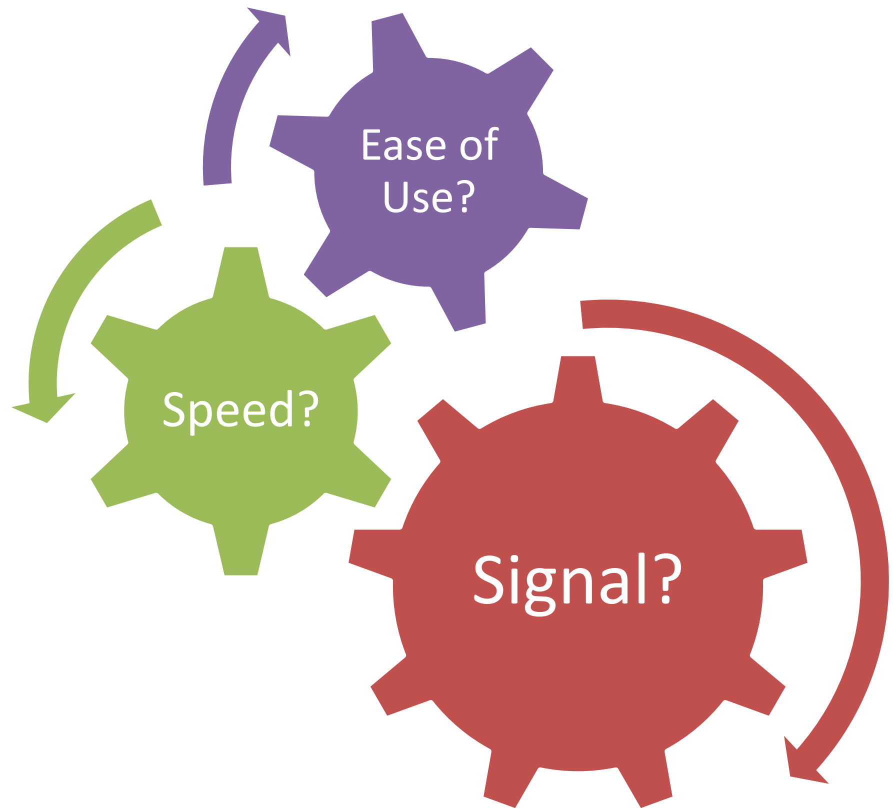
Enhancing GLMs – Core Issue

What do we mean by "enhancing" GLMs?

Thanks to constraints on time and energy, these three enhancements are related.

Enhancing the ease of use or the speed of the process leaves more time to search for additional signal.

Regardless of improvements, there is always a practical limit to time & energy.



Enhancing GLMs – Core Issue

The problem is that the GLM framework is fundamentally limited by its linear structure and the lack of an algorithmic approach to finding significant higher order interactions.

The implicit claim is that relevant higher order interactions do exist in insurance data; that insurance signal does consist of both linear and non-linear parts.

> By “signal” I mean that portion of variation in the response that can be related to a predictor and which will persist (reasonably well) over time.
>
> By “noise” I mean that portion of variation in the response that is random and will manifest itself differently from one dataset to another.

Enhancing GLMs – Core Issue

If higher order interactive effects exist in insurance data, then…

- …a naturally non-linear machine learning approach…
- …which algorithmically explores the solution space…

…would be more efficient in capturing that portion of the signal.

Rule Induction, a type of Machine Learning which includes trees, fits both of these descriptions.

We applied a Rule Induction approach to GLM residuals to see if they are indeed non-random – to see if we can create stable models.

2.

Machine Learning & Rule Induction

What is Machine Learning?

"Machine Learning is a broad field concerned with the study of computer algorithms that automatically improve with experience."

Machine Learning, Tom M. Mitchell, McGraw Hill, 1997

"With algorithmic methods, there is no statistical model in the usual sense; no effort made to represent how the data were generated. And no apologies are offered for the absence of a model. There is a practical data analysis problem to solve that is attacked directly…"

"An Introduction to Ensemble Methods for Data Analysis", Richard A. Berk, UCLA, 2004

What is Rule Induction?

Just what it sounds like – an attempt to induce general rules from a specific set of observations.

The procedure we used partitions the whole universe of data into "segments" which are described by combinations of significant attributes, a.k.a. compound variables.

- Risks in each segment are homogeneous with respect to the model response, in this case loss ratio.
- Risks in different segments show a significant difference in expected value for the response.

What is Rule Induction?

In contrast to GLMs, Rule Induction…

- …is non-parametric in nature; it makes no assumption about the underlying error distribution.
- …is algorithmic in that the computer does the "heavy lifting" of identifying significant combinations of fields.
- …uses a mild set of assumptions call "Probably Approximately Correct". The only requirement is that future unseen data have reasonably similar distributions to the training data.
- …does not provide p-values for testing individual fields.

3.

Signal Beyond GLMs – Case Study

Personal Auto Portfolio

Specifics of the original GLM:

- Australian insurer of moderate size
- 2 years of data
- Comprehensive motor vehicle coverage
- An independent actuarial firm developed the GLM.
- The GLM was designed without having to consider filing constraints.
- The GLM was built on total data.

Personal Auto Portfolio

Specifics of the Rule Induction analysis:

- Data was split into a training and validation dataset – one year each.
- Analysis was conducted on the GLM residuals.
- Only the variables used in the original GLM were considered – no predictors were added to the data.
- Output segments were required to have at least 3000 claims.

Signal Beyond GLMs – Case Study

Personal Auto Portfolio

Model built on training data:

Segment	Exposure	GLM Premium	Incurred Loss	Claim Count	Loss Ratio
1	40,088	9,677,889	7,223,230	5,730	75%
8	26,642	8,770,620	7,454,508	4,717	85%
3	35,946	8,036,238	7,298,945	5,178	91%
4	20,954	6,699,637	6,353,455	3,664	95%
6	26,212	6,754,957	6,534,512	4,127	97%
10	29,558	7,868,872	8,109,686	5,018	103%
9	20,049	5,636,667	5,935,182	3,576	105%
2	33,043	10,830,010	11,614,780	6,287	107%
7	23,203	8,181,896	10,125,938	4,356	124%
5	30,163	7,419,663	9,590,068	5,081	129%

Signal Beyond GLMs – Case Study

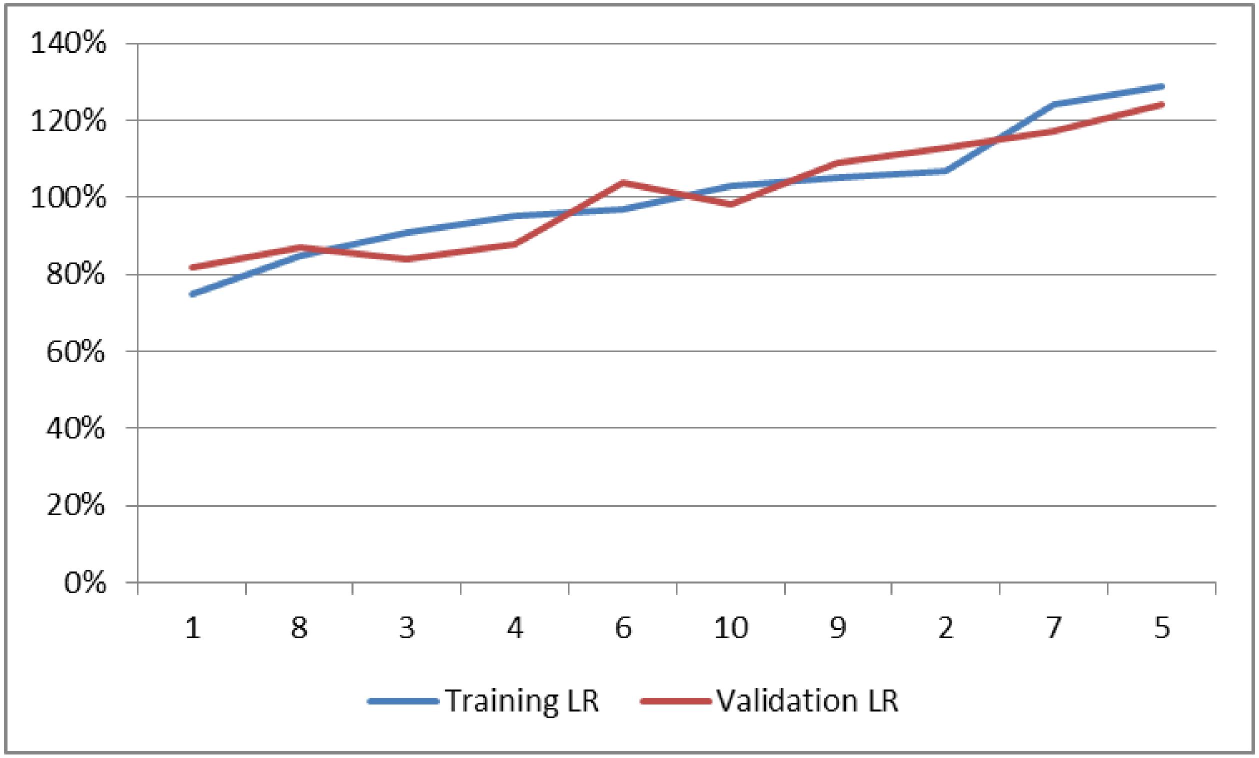
Personal Auto Portfolio

Model applied to validation data:

Segment	Exposure	GLM Premium	Incurred Loss	Claim Count	Loss Ratio
1	39,262	9,511,229	7,767,501	5,913	82%
8	20,083	6,415,686	5,565,564	3,784	87%
3	35,105	7,505,323	6,283,145	5,073	84%
4	15,379	4,749,230	4,195,864	2,822	88%
6	29,387	6,935,811	7,187,731	4,688	104%
10	33,141	8,311,156	8,171,977	5,761	98%
9	20,488	5,266,095	5,748,663	3,720	109%
2	34,729	10,911,435	12,336,791	6,768	113%
7	24,679	8,140,954	9,532,883	4,641	117%
5	25,717	5,925,355	7,358,740	4,570	124%

Personal Auto Portfolio

92.6% correlation of loss ratios between training and validation data.



Signal Beyond GLMs – Case Study

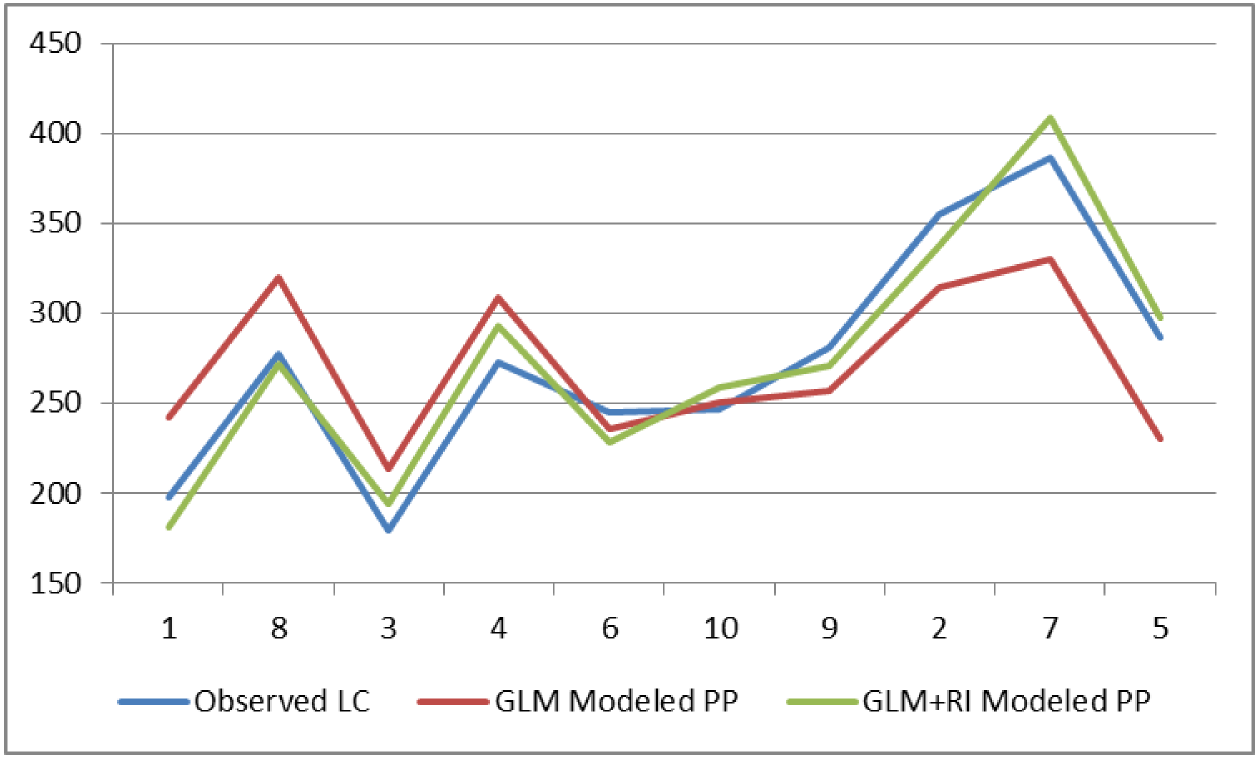
Personal Auto Portfolio

We also compared the observed loss cost to both modeled pure premiums – on validation data.

Segment	Observed Loss Cost	GLM Modeled PP	GLM+RI Modeled PP	% Diff – GLM to Observed	% Diff – GLM+RI to Observed	% Improvement
1	198	242	181	-18.3%	9.4%	8.9%
8	277	319	272	-13.3%	2.1%	11.2%
3	179	214	194	-16.3%	-7.8%	8.5%
4	273	309	293	-11.7%	-6.8%	4.8%
6	245	236	228	3.6%	7.1%	-3.5%
10	247	251	258	-1.7%	-4.6%	-2.9%
9	281	257	271	9.2%	3.7%	5.5%
2	355	314	337	13.1%	5.4%	7.6%
7	386	330	408	17.1%	-5.4%	11.7%
5	286	230	298	24.2%	-3.9%	20.3%

Personal Auto Portfolio

We also compared the observed loss cost to both modeled pure premiums – on validation data.



4.

Possible Changes to the GLM Development Process

Possible Changes to the GLM Development Process

First way to "enhance" GLMs – simply add Rule Induction

Rule Induction can enhance the signal of the combined model. In this case study, there were no changes to the GLM development process.

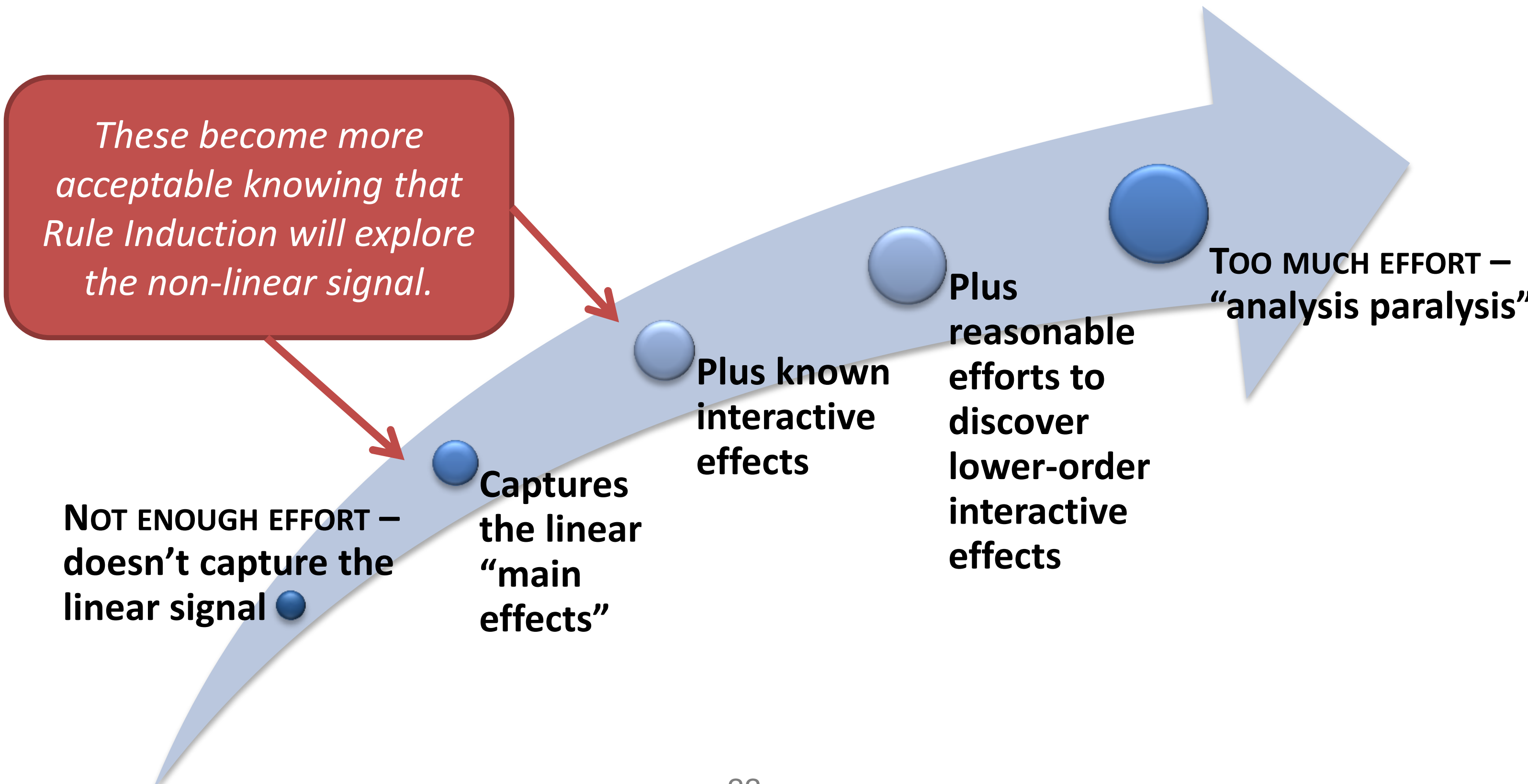
This approach leaves you doing everything you did before, plus development of the Rule Induction model.

Open question: Does going into the modeling process knowing you have both GLM and Rule Induction change how you build the total model?

Possible Changes to the GLM Development Process

Second way to "enhance" GLMs – rebalance the workload

The first place to look is in how much effort is put into building the initial GLM.



Third way to "enhance" GLMs – variable identification

Rule Induction can be useful to reduce the number of potential predictors. There are a couple of methods…

- Use Rule Induction on frequency and severity, and note which fields are used first to split the data.

- Use one of several methods to "shake the tree" to create multiple output models. *[For example, randomly incorporate something other than the optimal splits in the data.]* Over the course of many iterations, note which fields are used across many models regardless of the random perturbations.

Fourth way to “enhance” GLMs – use hold-out data

Non-parametric methods, because they do not have p-values and significance testing, rely on hold-out data for model selection.

The accuracy of significance testing depends on the extent to which sample means tend toward a normal distribution. For insurance data, with its inherent volatility, this convergence is slow.

Using hold-out data as a part of model selection provides a test for over-fitting which does not rely on distributional assumptions.

Fourth way to "enhance" GLMs – use hold-out data

This approach would look something like this:

- Use forward regression techniques to build an array of GLMs to consider. *[Our method used the training deviance to find the next "best" predictor. This is only one approach.]*
- Select the best model based on multiple metrics – validation data deviance improvement; AIC/BIC on training data; etc.
- As the model form solidifies, one can confirm the validity of predictors through normal statistical and consistency tests.

Can still develop model from there (known interactions, etc.). Look for good statistics on the training data as well as improvement in the validation model metrics.

Fourth way to "enhance" GLMs – use hold-out data

Advantages of this approach:

- Provides a test of model performance that is independent of any error distribution assumption.
- Gets to the final model form faster – only those predictors which are part of the best model get the full array of tests.
- Evaluates the model as a whole, not just individual pieces.

Disadvantages of this approach:

- "Contaminates" the validation data. Using the hold-out data this extensively makes it unfit for a final test. Model will be biased to perform well on this dataset.
- Ideally would be used in conjunction with a 3rd hold-out dataset – training/test/validation.

5.

Model Development Case Study

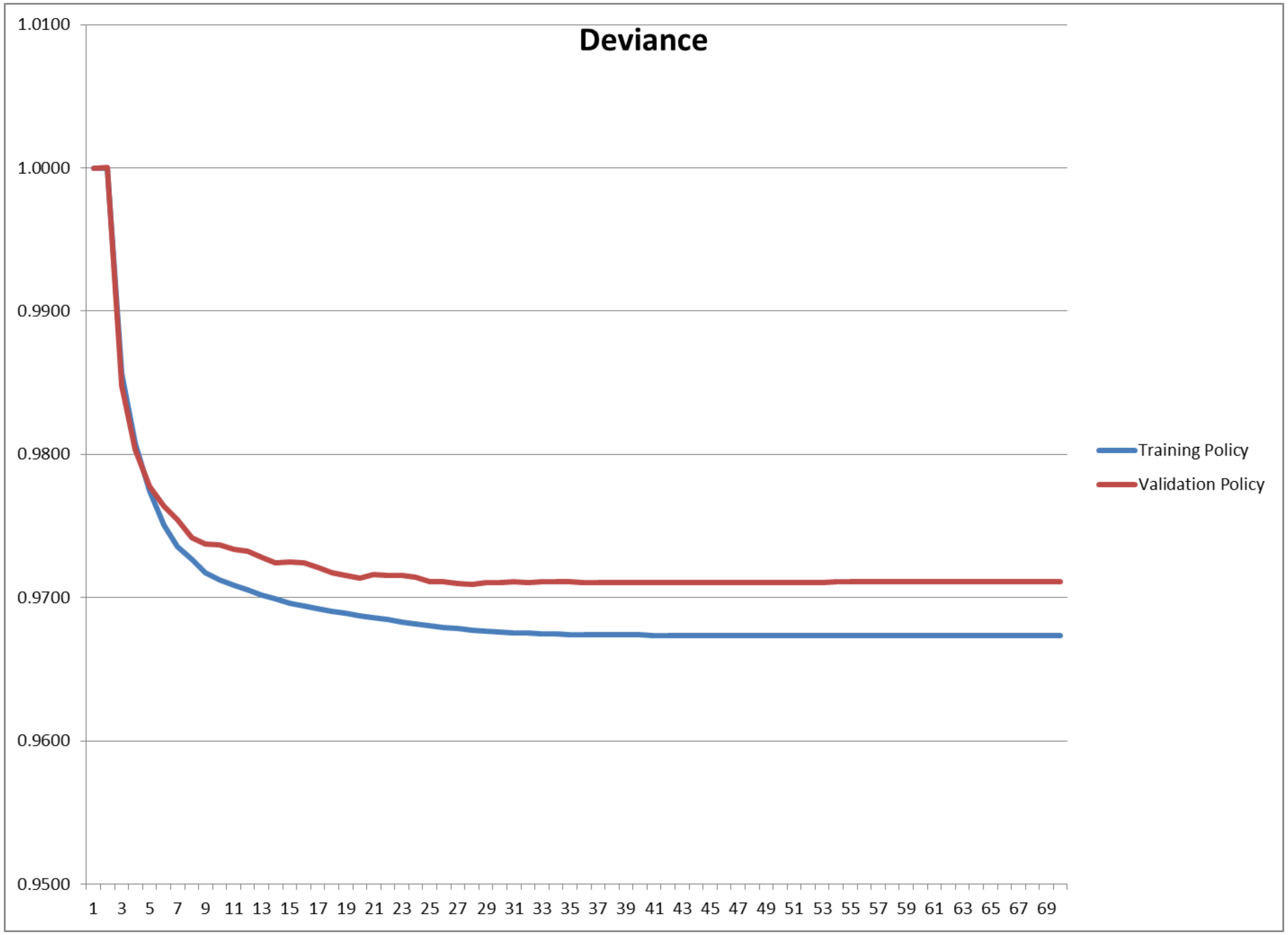
Homeowners

Specifications:

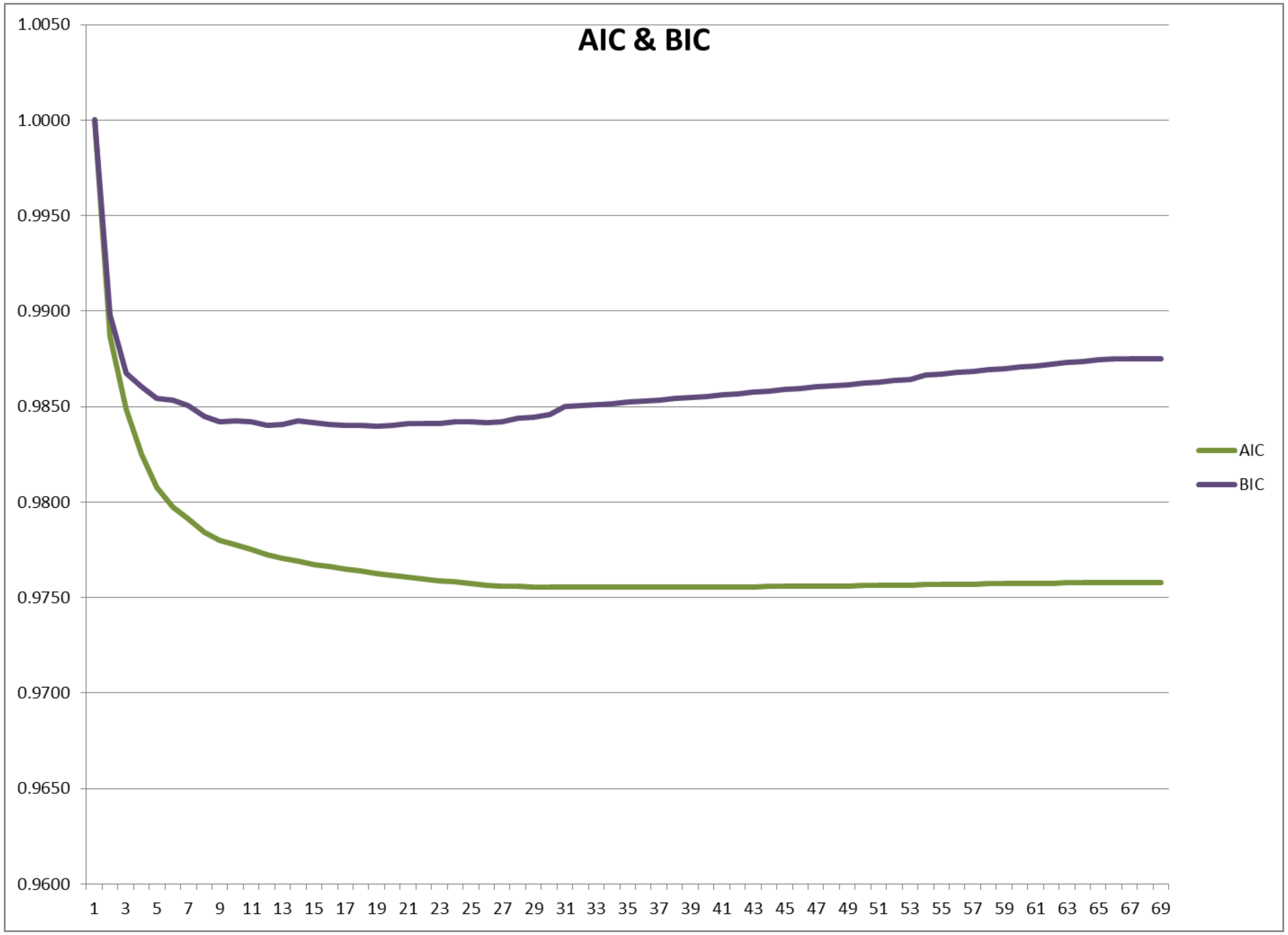
- Moderately small homeowners book – multi-state
- 6.25 years of data
- "Other perils" only – Wind/Hail & CATs removed
- Modeled frequency and severity separately
  - Frequency used Poisson error distribution.
  - Severity used gamma error distribution.
- Log link function

After verifying initial model assumptions, and after exploring capping levels (none was used), we ran a forward regression routine to explore main-effect models. This routine selected predictors based on the improvement in training deviance.

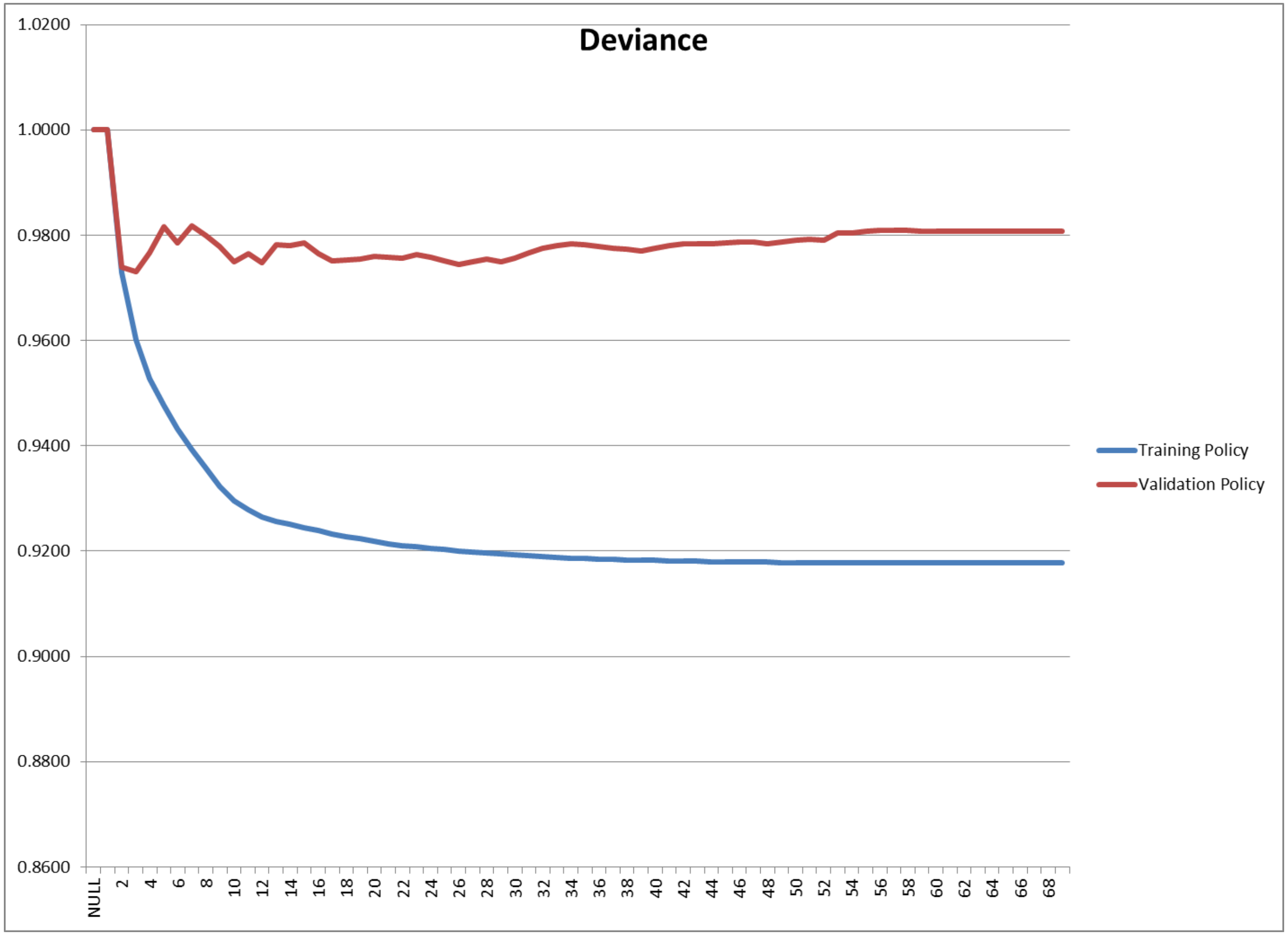
Frequency – Forward Regression – Deviance Improvement



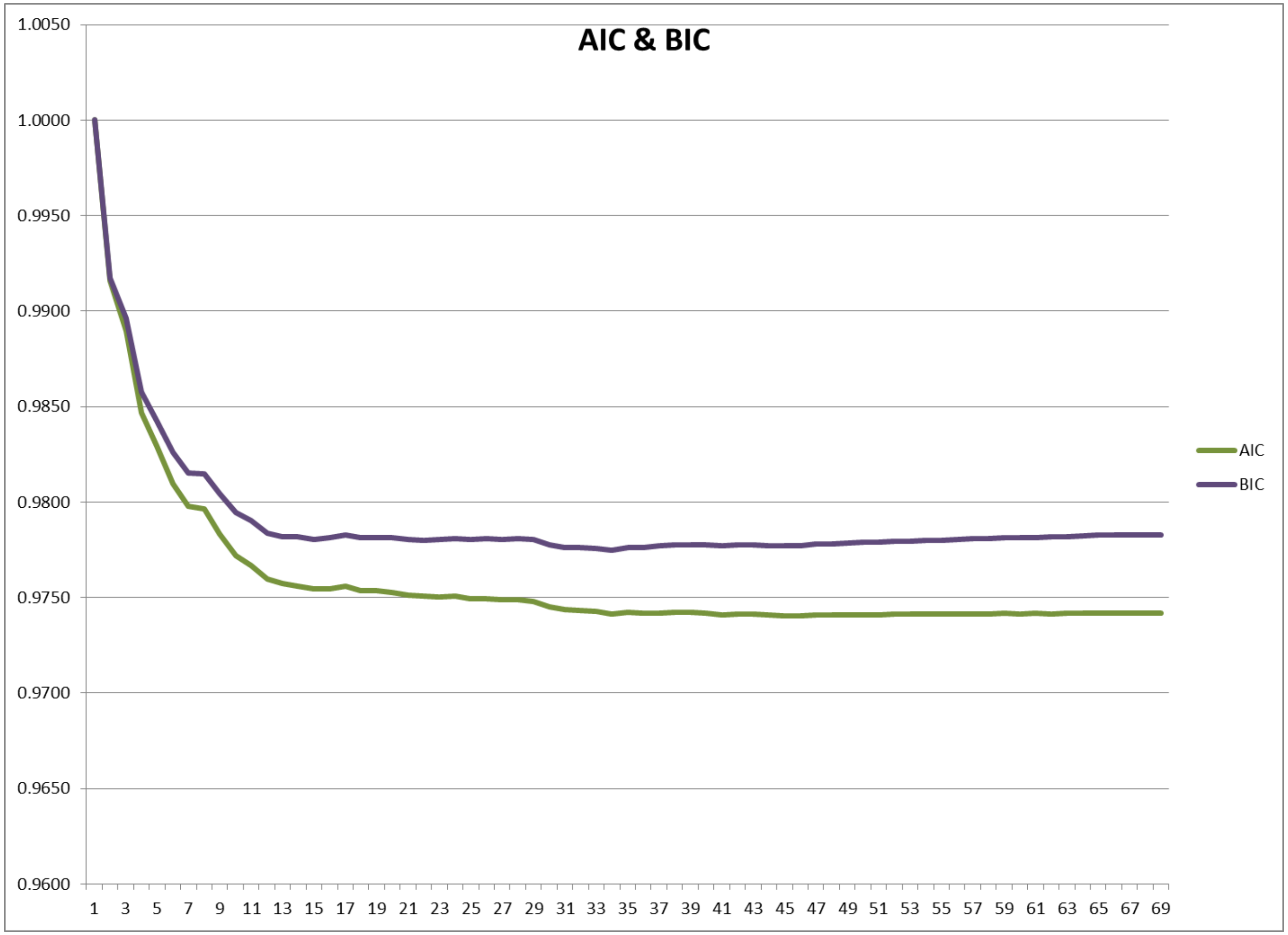
Frequency – Forward Regression – AIC & BIC Improvement



Severity – Forward Regression – Deviance Improvement



Severity – Forward Regression – AIC & BIC Improvement



Further Model Refinement

This (or any other) version of forward regression just gives a starting point. Other model refinement included…

- Evaluation of included fields:
  - Statistical significance
  - By-year consistency
  - Business sense and utility
- Evaluation of excluded fields when they were of particular interest to the business or for regulatory reasons – example: credit no-hit
- Creation of predictor concatenations which better reflect the business reality – example: mortgage & paid-in-full
- Evaluation of known or suspected interactions
- Grouping & simplification of fields

Analysis of residuals – Frequency

Once we had final initial models, we used Rule Induction to analyze the frequency model residuals.

Our methodology controls the granularity of the model by specifying the minimum number of claims required for a segment to be identified.

We examined models with a minimum of 1000 claims up through a minimum of 5000 claims.

We also looked at allowing any of the original 68 potential predictors versus limiting to only those fields used in the frequency GLM.

Analysis of residuals – Frequency

The table to the right shows results when all available fields were used. In general, the lift was superior.

Min Claims	# Segments	Obj. Range	Correlation
1000	14	0.0164	94.7%
2000	8	0.0077	90.7%
3000	5	0.0052	84.2%
4000	4	0.0052	86.3%
5000	3	0.0040	94.2%

Min Claims	# Segments	Obj. Range	Correlation
1000	14	0.0137	94.9%
2000	8	0.0080	95.2%
3000	4	0.0053	98.2%
4000	4	0.0045	96.4%
5000	3	0.0035	99.6%

This table shows the results when only GLM fields were used. In this case the match between training and validation is better and more consistent.

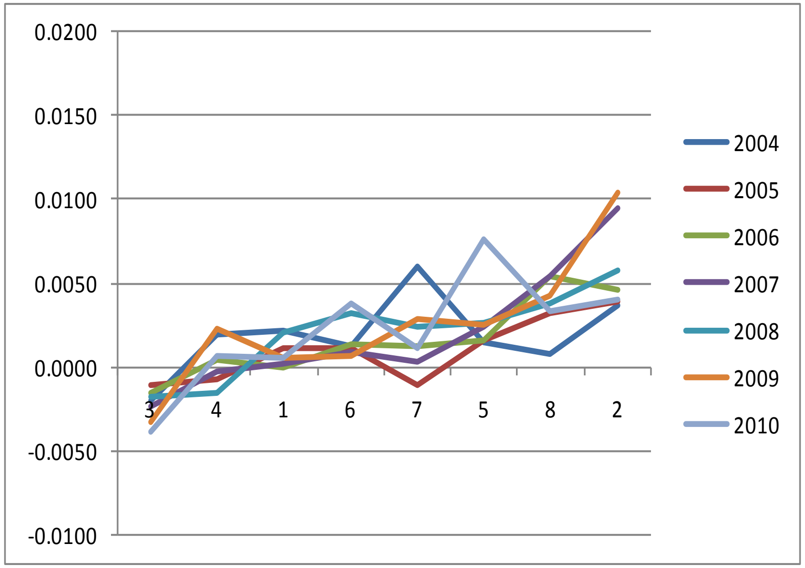
Analysis of residuals – Frequency

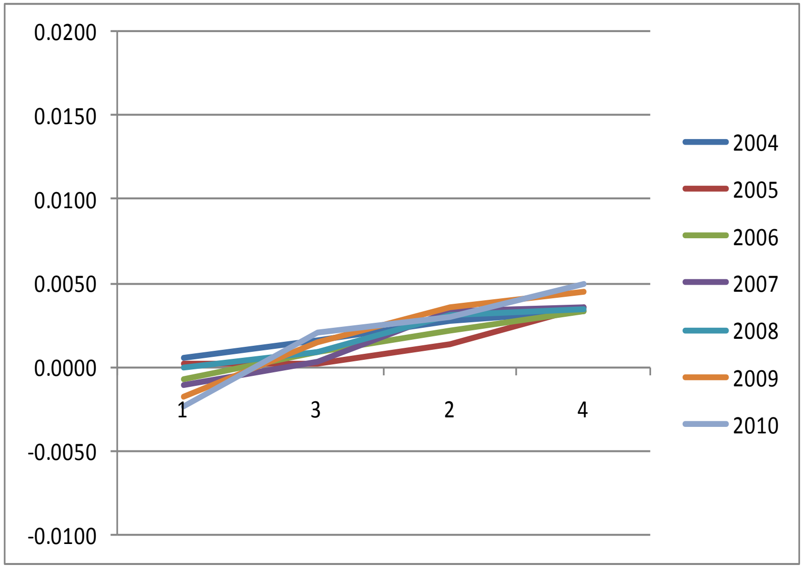
Choices, choices, choices…

- What level of model complexity fits your appetite?
  - Is there value in simplicity over lift?
  - Will the interactions be acceptable to agents, regulators, and upper management?
- Do you limit yourself to only those fields in the GLM, or expand the model with other fields?
- After exploring what other fields may have been used in the more expansive models, does anything lead you back to refine your underlying GLM?
- Or, do you choose a model and move forward?
  - Does the model cause reversals which need to be smoothed?
  - Does the model unwind anything the GLM does?

Analysis of residuals – Frequency

What is the consistency of the model over time?





The lift of the second model is half the first. The training/validation correlation is not much higher (96.4% versus 95.2%).

Analysis of residuals – Frequency

Another choice: If you choose a Rule Induction model to use along with the GLM, do you simply add it, or do you rerun your GLM with the model as a predictor?

Rerunning the GLM has one very significant advantage. All the statistics that we, and regulators, are used to seeing will now be generated for the levels of the Rule Induction model.

We put the model with 8 segments into the frequency GLM as a new predictor.

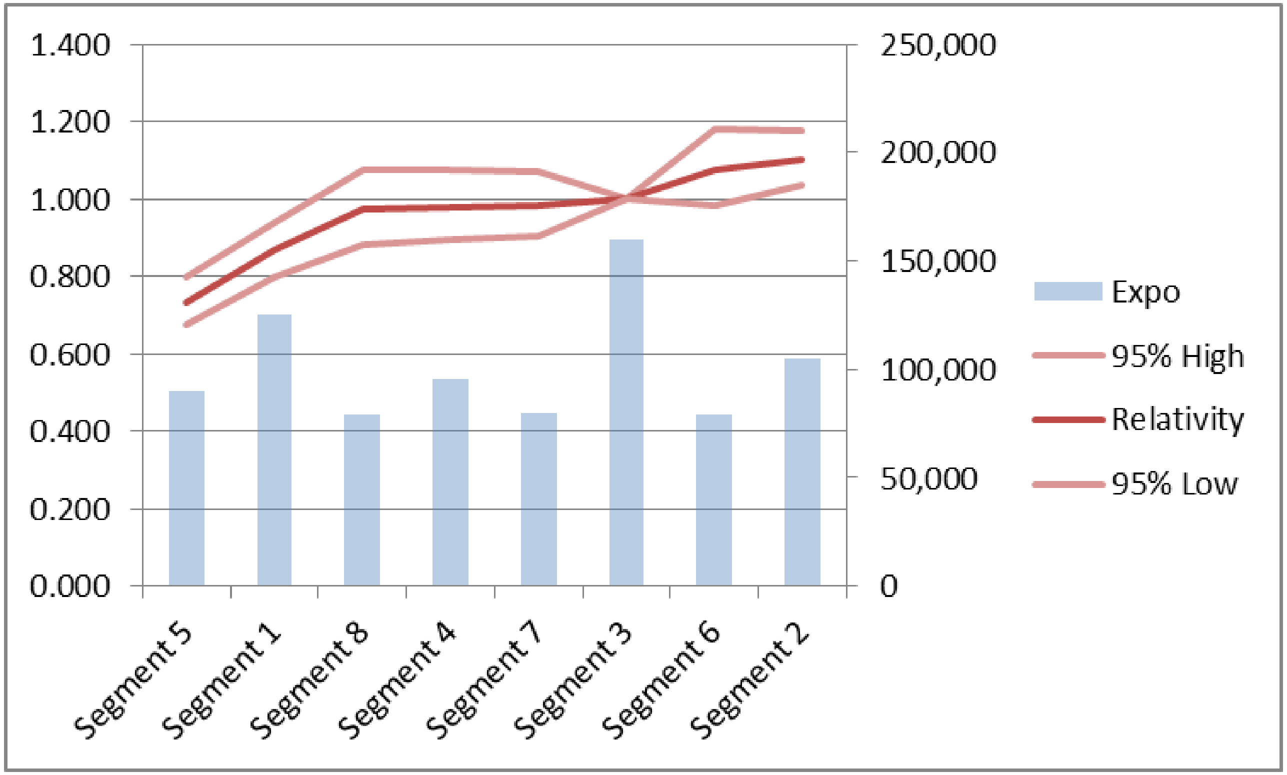
Analysis of residuals – Frequency

Rule Induction model with minimum of 2000 claims per segment.

Segment	Beta	Std Err	z	P > |z|	95% Low	Relativity	95% High	Exposures
5	-0.3092	0.0438	-7.054	0.00%	0.674	0.734	0.800	89,627
1	-0.1432	0.0414	-3.461	0.05%	0.799	0.867	0.940	125,459
8	-0.0283	0.0505	-0.561	57.49%	0.880	0.972	1.073	78,804
4	-0.0198	0.0471	-0.421	67.41%	0.894	0.980	1.075	95,863
7	-0.0156	0.0433	-0.360	71.86%	0.904	0.985	1.072	79,434
3	0	NA	NA	NA	1.000	1.000	1.000	160,190
6	0.074	0.0471	1.573	11.58%	0.982	1.077	1.181	78,865
2	0.0983	0.0322	3.055	0.22%	1.036	1.103	1.175	105,168

Analysis of residuals – Frequency

Rule Induction model with minimum of 2000 claims per segment.



6. *Other Variations*

Variations on a Theme

In the model development case study, the result of the Rule Induction analysis of residuals was put directly into the GLM as a new predictor.

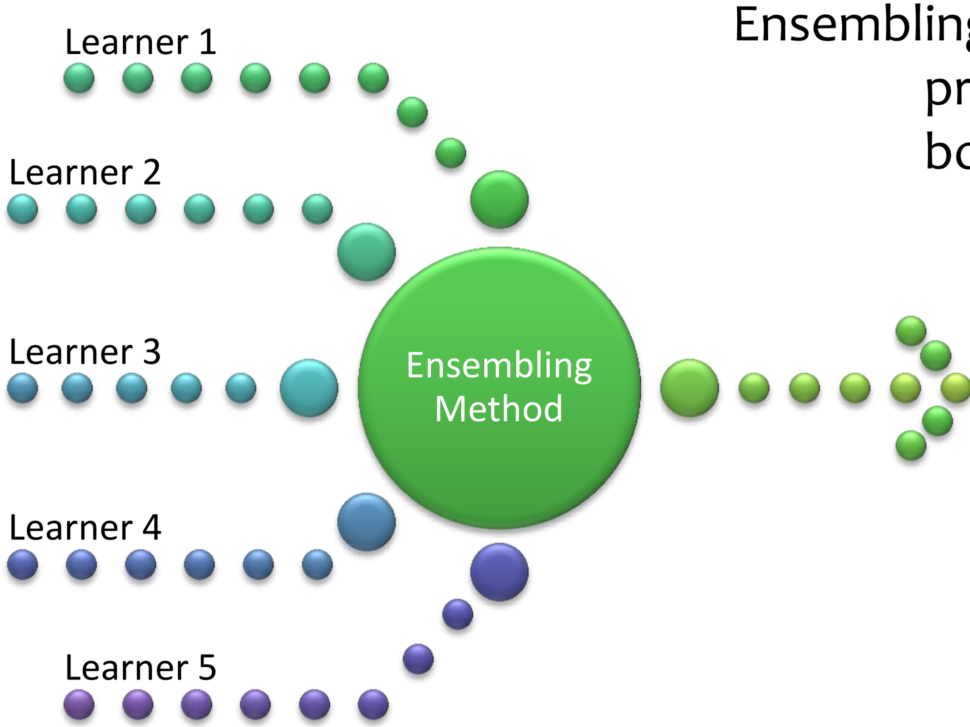
Another version is to take the Rule Induction model as information about relevant variable interactions.

- Global interactions between these same fields can be put into the GLM and the results evaluated.
- An analysis of the residuals can be repeated on the new model, and the new information utilized in the same manner.
- This process can be iterated until no stable model can be found in the GLM residuals.

Scoring based on Rule Induction

Rule Induction can be the base learner in an ensembling approach.

Each base learner provides an estimate of risk. These separate models are combined into a single model. Ensembling methods are shown to provide superior models in both stability and lift over the base learners.



There are various techniques to build models on different versions of the data.

Other Variations

Scoring based on Rule Induction

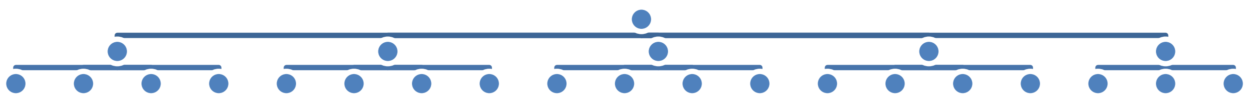
Boosting and bagging are a couple of ways to take a single leaner and single set of data, and still produce multiple estimates to be ensembled.

Boosting – use the base learning model iteratively, but change the weights such that future iterations focus more heavily on examples that are misclassified.

Bagging *(bootstrapping aggregation)* – use the base learning model on different sets of data generated by randomly sampling (with replacement) from the original data.

Scoring based on Rule Induction

Through ensembling, simplicity is traded in for better lift and stability. Even just layering a 4 segment model on top of a 5 segment model produces 20 unique segments.



With added complexity, simply trade in segments for a 3-digit score, and bands the scores appropriately.

The pros and cons of this approach vary by situation, but this option can always be explored.

Do it all at once – Fusion Algorithms

The process described here – build an initial GLM, analyze the residuals, and incorporate the results into a final GLM – is a sequential process. However…

- …any series of choices sets up a dependency of later choices on earlier ones, and…
- …optimal results are not guaranteed by a series of optimal choices.

Fusion algorithms are an alternative which combine two types of algorithms (in this case a linear model and rules) into one model, solving for each piece simultaneously.

"Predictive Learning via Rule Ensembles", Jerome Friedman and Bogdan Popescu (Stanford University)

7. *Summary*

Summary

- **Higher-order interactive effects exist within insurance data.**
- **GLMs are effective models for capturing linear signal and lower-order interactive effects.**
- **Rule Induction is effective at quickly finding compound variables which capture the high-order interactive effects within insurance data.**
- **Hold-out data and a model validation approach can also be used to specify the predictors used in a GLM.**
- **Using GLM and Rule Induction in a complementary manner can change both the process of model building and the resulting models as well.**

<u>Questions?</u>

<u>Contact Info</u>

Christopher Cooksey, FCAS, MAAA
EagleEye Analytics
ccooksey@eeanalytics.com
www.eeanalytics.com