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

Getting More Out of Your Existing Data

CAS In Focus Seminar 3 October 2011 **Christopher Cooksey, FCAS, MAAA**

EagleEye Analytics

Agenda…

1. Setting up the issue
2. Loss Ratio versus Pure Premium
3. Machine Learning and Rule Induction
4. Model Validation
5. Case Studies
   - Private Passenger Auto
   - Homeowners
   - Commercial Auto
6. Other Issues
7. Summary

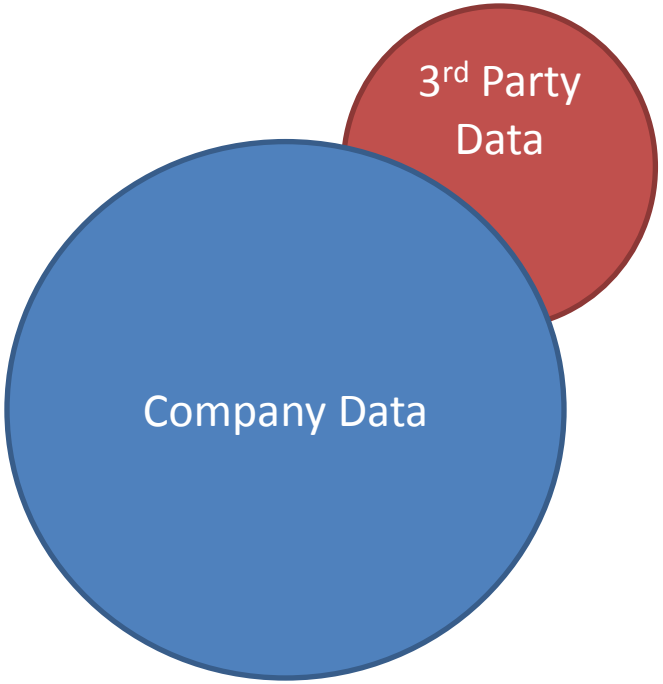
1. *Setting up the issue*

Setting up the issue

Many companies are looking at 3rd party data to enhance their modeling efforts.

Additional predictors attached to company data certainly have the potential to increase the predictive power of company models.

- Credit with auto & home
- MVR data
- Census data
- NOAA weather info
- Commercial data aggregators



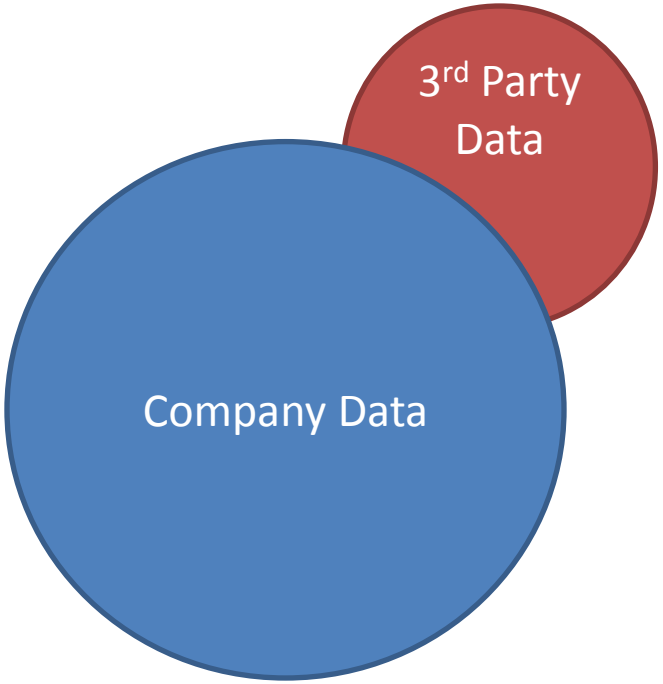
Setting up the issue

There is also a cost to getting, analyzing and using 3rd party data.

Predictive utility of potential data needs to be verified.

- Purchased cost of getting bulk data for analysis
- Development cost of determining how it should be used in conjunction with current rating

There may also be on-going costs, including purchasing data at point-of-sale.



Setting up the issue

Another alternative is to spend those resources getting more signal out of existing data.

The signal in existing company data goes deeper than most companies have mined.

- Many companies can improve their class plans simply by beginning to use analytics.
- Companies who have modeled the signal with GLMs can explore the higher order non-linear signal.
- Companies can also explore the signal at different levels of the data (for example, policy level).

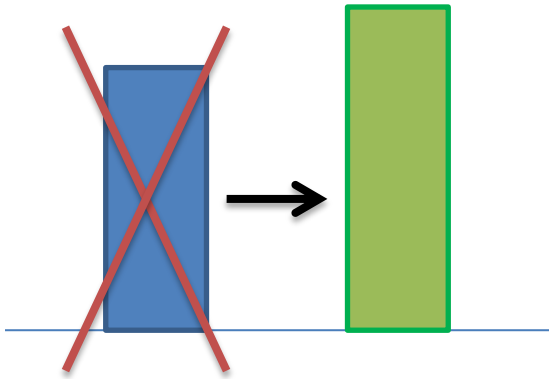
2. *Loss Ratio versus Pure Premium*

Loss Ratio versus Pure Premium

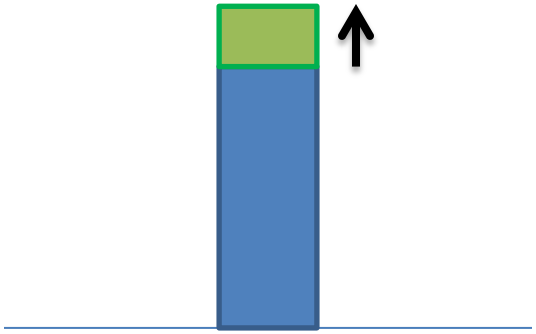
Statewide indication – should it be expressed as…

…a new rate? (pure premium approach)

…a change to existing rates? (loss ratio approach)



Pure Premium



Loss Ratio

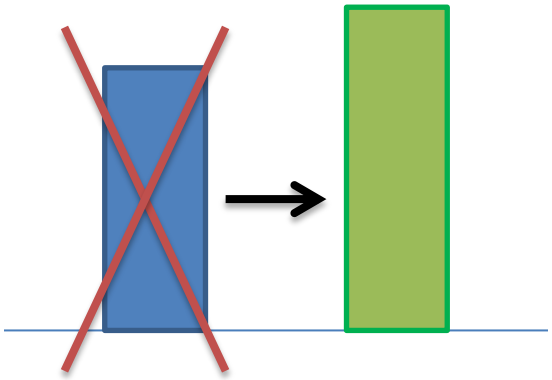
GLM modeling is (usually) a pure premium approach

There is no reference to existing rating or existing premium.

Modeling is done at the frequency/severity or loss cost level.

Advantages of a pure premium analytical approach include…

- An understanding of from-the-ground-up relationships
- An understanding of frequency and severity effects



Pure Premium

Disadvantages of the same include…

- Significant analytical effort
- Significant implementation issues

Loss Ratio versus Pure Premium

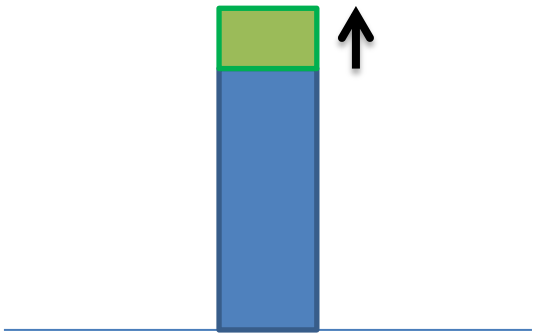
Loss ratio modeling is an under-explored approach

Results are relative to existing rating plan

Modeling is done using loss ratios – residual modeling

Advantages of a loss ratio analytical approach include…

- Easier implementation (modify what you have)
- An understanding of profitable (and unprofitable) customers



Loss Ratio

Disadvantages of the same include…

- Significant data prep issues – you must have rerated premiums!

How to model loss ratios?

GLM is not an effective approach for modeling loss ratios

- A priori information is helpful when creating class plans using GLMs, but there is no a priori info on mispriced segments – if we already knew, we'd change it!
- Most class plans capture primarily the linear signal and lower-order interactive effects, so using a linear modeling approach will continue to miss higher-order interactive signal.

Rule Induction, a type of Machine Learning which includes trees, is an effective approach because it…

- …algorithmically explores the solution space.
- …naturally finds non-linear, interactive effects.

3. *Machine Learning and 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.

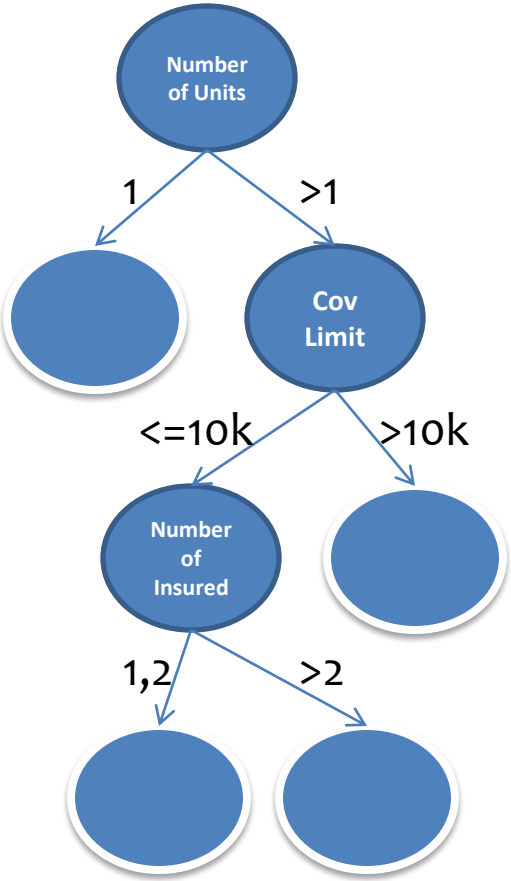
- Risks in each segment are homogeneous with respect to the model response, in this case loss ratio.
- Risks in different segments show a significant difference in expected value for the response.

What is Rule Induction?

Branches of the tree are segments of the book; each segment with a common definition for all business with in that branch.

Utilized two versions…

- Segmentation – a greedy approach which makes optimal selections at each split
- Multiple Splits – a non-greedy approach which explores a variety of non-optimal splits in the data



4. *Model Validation*

Why validate models?

With Machine Learning, the computer does the "heavy lifting" of model development.

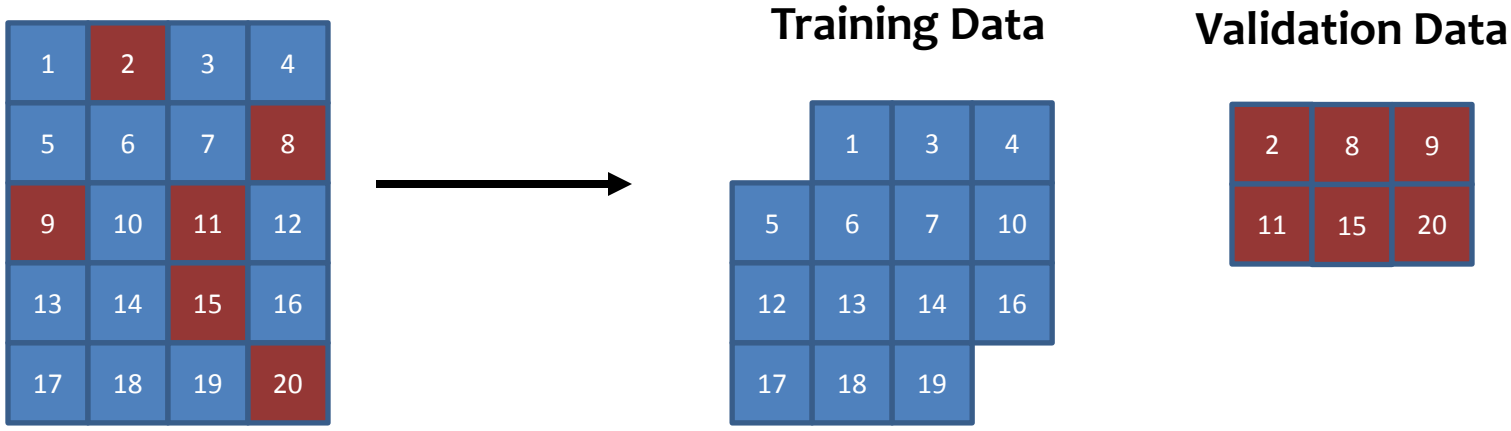
This obviates the need for significance testing as a means of model development – which is good because we have no error distribution!

However models are built, there is a need to evaluate their generalization power by validating them against unseen data.

Hold-out datasets

Used two methods –

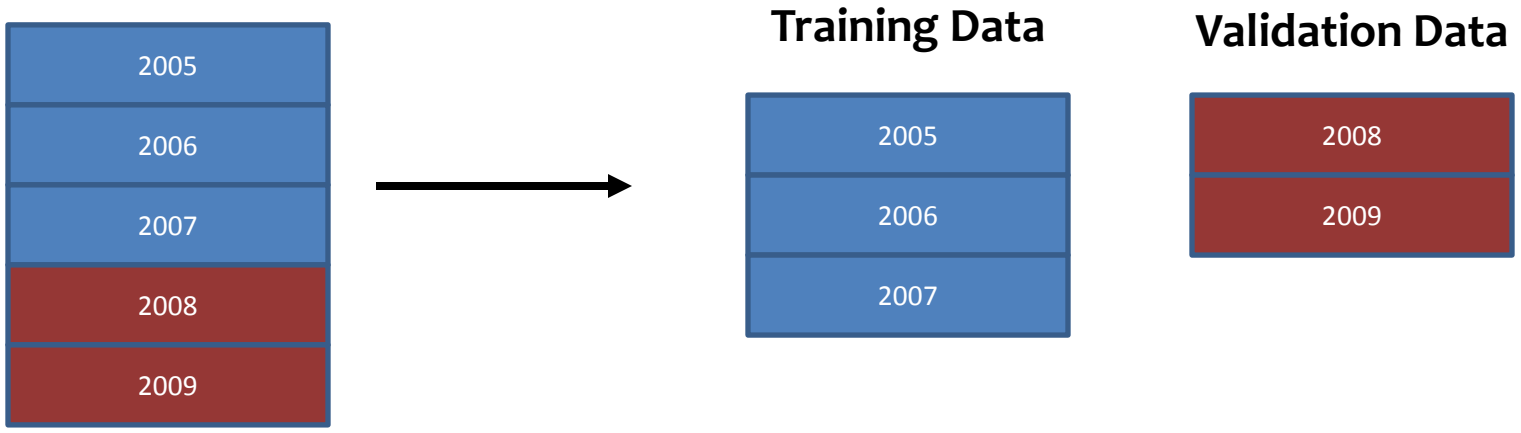
- Out of sample: randomly trained on 70% of data; validated against remaining 30% of data.



Hold-out datasets

Used two methods –

- Out of sample: randomly trained on 70% of data; validated against remaining 30% of data.
- Out of time: trained against older years of data; validated against newest years of data.



Hold-out datasets

Models were built using training data. Once built, models were applied to validation data.

Model performance on this unseen data was used to select the most appropriate model form.

- **Lift** – ratio of the worst loss ratio to the best loss ratio
- **Correlation** – weighted Pearson correlation between training data and validation data loss ratios
- **Deviance improvement** – reduction in deviance on validation data when model is applied
- **Performance by year** – consistency of model loss ratios when data is split by year

5. *Case Studies*

Private Passenger Auto

Small, US regional auto insurer – 5 years of data
End goal was to take pricing actions
Current rating not based on a GLM analysis
Out of sample validation – 70% training, 30% validation
Separate analyses by coverage – BI, PD, MP, COMP, COLL

Coverage	Earned Exposures	Claim Count	Loss Ratio*
BI	2,018,527	6,617	47.6%
PD	2,017,525	26,594	54.4%
MP	1,149,735	3,875	52.2%
COMP	1,167,903	28,069	54.3%
COLL	1,163,388	24,683	60.1%

**Loss Ratio was calculated using rerated premium*

Private Passenger Auto – Bodily Injury

First issue was to identify potential predictors:

- 32 fields on the file
- 9 fields identified as inappropriate
  - Agent number: highly dimensional & unrelated to loss
  - Some fields didn't discriminate data: 98% was 'N'
  - Other fields exhibited data integrity issues: 20% of policies have 6 drivers?!?
- Remaining 23 fields were considered potential predictors

Ordinal fields were bucketed based on the univariate signal in loss ratio.

The same approach was used for each coverage.

Private Passenger Auto – Bodily Injury

Second issue was to identify the best lower limit on segment size:

- If segments are too small, the model will be too granular and will not generalize well to unseen data.
- If segments are too large, the model will be too simple and will miss signal in the training data.
- Correct (or rather, useful) segment sizes depend not only on total volume, but also the internal volatility of the data and the amount of signal to be measured.

We ran Segmentation (greedy) models at four thresholds for the minimum number of claims and examined the results.

Private Passenger Auto – Bodily Injury

In this example we see what we would expect – the lower the number of claims, the more segments, the higher the lift, but the lower the consistency with validation data.

Note: overall lift and correlation are aggregate measures of model performance. They do not tell the whole story, but are good for deciding on a useful balance between fit and generalization.

Claims Threshold	# of Segments	Lift	Correlation
500	8	2.664	89.1%
750	5	2.159	94.8%
1000	4	2.111	99.5%
1250	3	1.746	99.8%

Private Passenger Auto – Bodily Injury

Once a useful minimum segment size was identified, we used the Multiple Split (non-greedy) approach to look at a larger array of possible models. In this case we looked at 40 models.

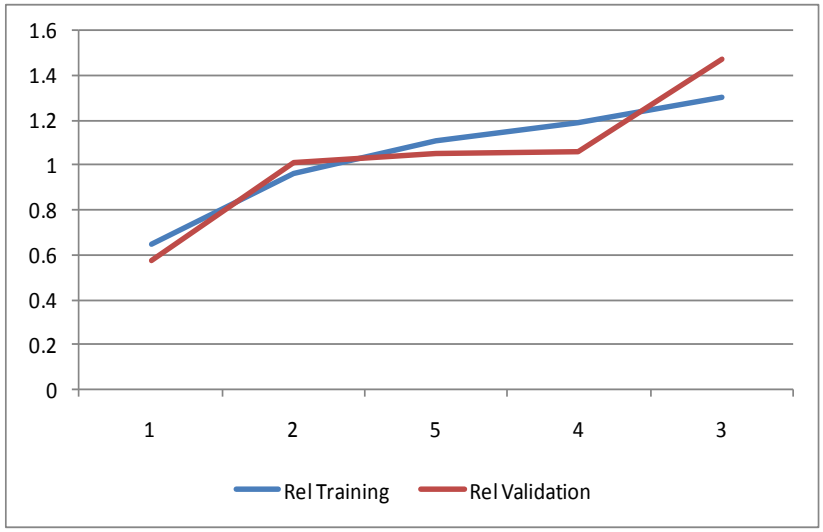
Claims Threshold	# of Segments	Lift	Correlation
750	5	2.31	98.3%
750	4	2.24	97.3%
750	5	2.15	99.0%
750	4	2.03	93.3%
750	5	2.00	99.3%
750	5	1.99	98.0%
750	5	1.98	98.0%
…	…	…	…

Private Passenger Auto – Bodily Injury

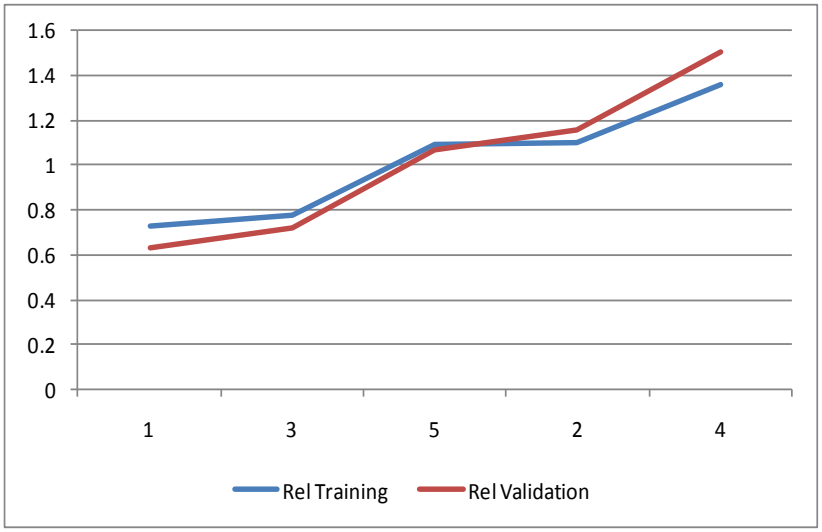
The final model was chosen by considering more than just lift and correlation.

Visual representations of correlation

Correlation: 94.8%



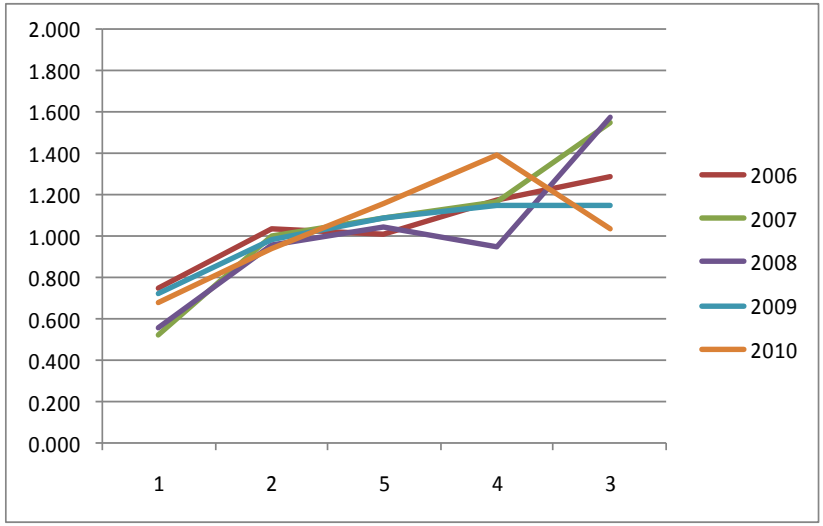
Correlation: 99.3%

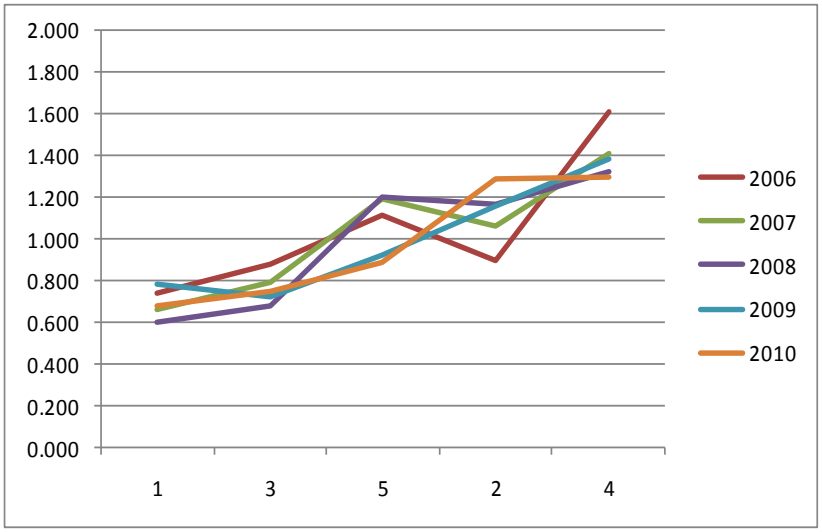


Private Passenger Auto – Bodily Injury

The final model was chosen by considering more than just lift and correlation.

Visual representations of consistency by year – All Data





Private Passenger Auto – Bodily Injury

The final model was chosen by considering more than just lift and correlation.

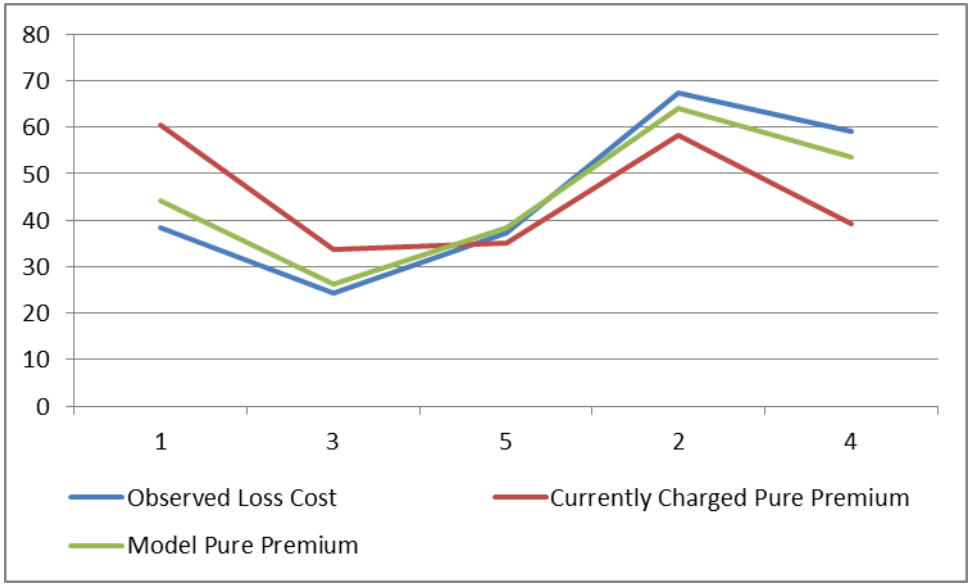
Percent improvement of the fit by segment – Validation Data

Segment	Observed Loss Cost	Currently Charged Pure Premium	Modeled Pure Premium	% Diff – Current to Observed	% Diff – Modeled to Observed	% Improvement
1	38.3	60.3	44.1	-57.7%	-15.4%	42.3%
3	24.4	33.8	26.2	-38.8%	-7.4%	31.4%
5	37.4	35.0	38.3	6.7%	-2.5%	3.9%
2	67.4	58.2	64.0	13.6%	5.0%	8.5%
4	59.1	39.2	53.4	33.7%	9.7%	24.0%

Private Passenger Auto – Bodily Injury

The final model was chosen by considering more than just lift and correlation.

Visual improvement of the fit by segment – Validation Data



Private Passenger Auto – Bodily Injury

The final model was chosen by considering more than just lift and correlation.

Total segment-level deviance improvement – Validation Data

Statistic	Average Deviance - Current to Observed	Average Deviance - Modeled to Observed	% Improvement
Simple Deviance	11.1	3.0	72.5%
Sum of Squares Deviance	173.7	13.3	92.4%
Chi-square Deviance	3.8	0.3	92.7%

Private Passenger Auto – Bodily Injury

The final model was chosen by considering more than just lift and correlation.

Actual model definitions

Model 1 used **Model Year, Multi-policy, and BI Limit**
Model 2 used **Model Year, Driver Age, and Multi-Policy**

Though the statistics for each model were similar, BI Limit can be manipulated by agents and insureds. Model 1 was removed from consideration.

Some models are more implementable than others!

Private Passenger Auto – Bodily Injury

The final chosen model:

Lift: 2.00 *Correlation: 99.3%*

Loss Ratio:	33.4%	36.1%	51.7%	53.1%	66.9%
Exposures:	269,074	552,999	535,194	304,605	356,656
Variables	**Segment 1**	**Segment 3**	**Segment 5**	**Segment 2**	**Segment 4**
Driver Age	0-Adult	Adult+	Young Adult+	0-Young Adult	Young Adult+
Model Year	Old	Old	Not Old	Not Old	Not Old
Multi-policy			Y		N

Note: results are specific to the given underlying class plan and should not be generalized to other companies. The model has also been modified for display.

Private Passenger Auto

Across coverages, significant lift was found along with notable generalization power.

Coverage	Earned Exposures	Lift	Correlation
BI	2,018,527	2.00	99.3%
PD	2,017,525	2.08	99.0%
MP	1,149,735	2.83	99.7%
COMP	1,167,903	1.36	97.1%
COLL	1,163,388	1.65	97.0%

Predictive fields included marital status, credit, age, model year, at-fault claims, and rating tier.

Homeowners

Regional US homeowners insurer – 5 years of data
End goal was to understand issues – considered pricing and underwriting actions
Three data sets – 2006-2009 training/validation (random split); 2010 for testing
Analysis for "Other Perils" only – Wind/Hail & CATs removed
Minimum claims per segment set at 1000

Peril	Earned Exposures	Claim Count	Loss Ratio*
Other Perils	171,917	15,080	63.8%

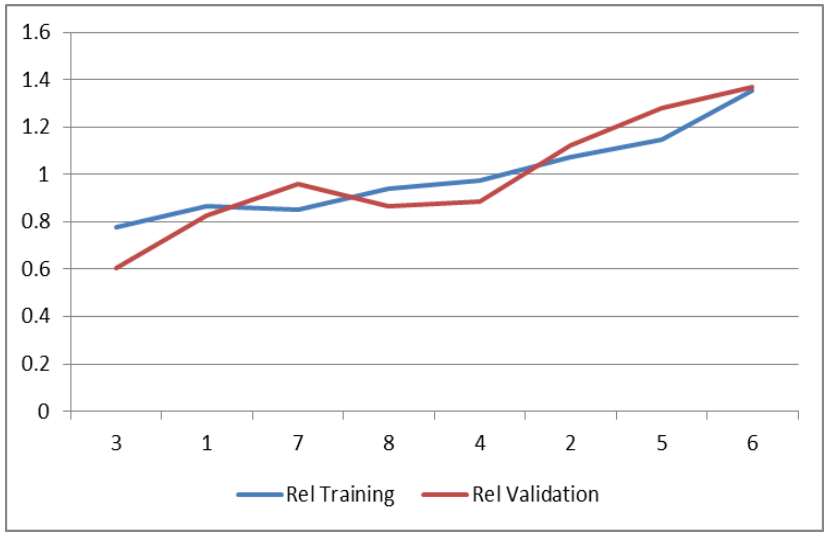
**Loss Ratio was calculated using rerated premium*

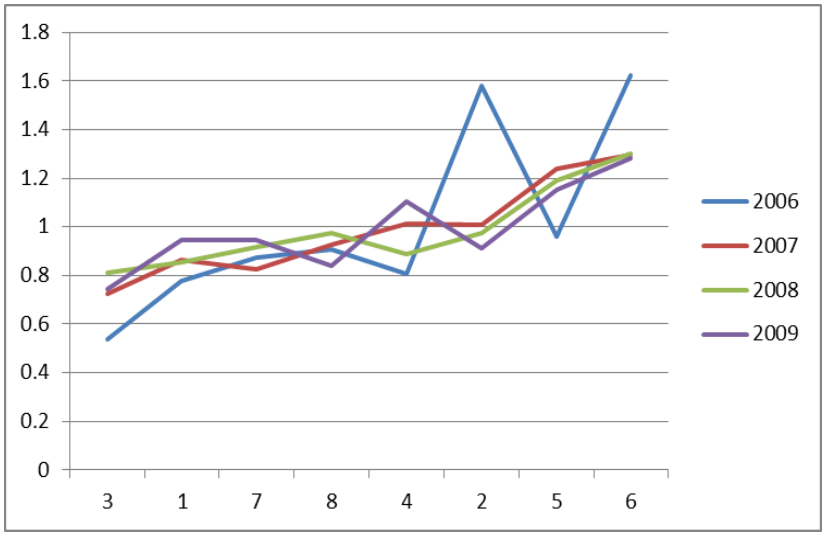
Homeowners

The final model was chosen by considering more than just lift and correlation. In this case, the lift was 1.87.

Visual representations of correlation and consistency

Correlation: 92.6%





Homeowners

The final model was chosen by considering more than just lift and correlation.

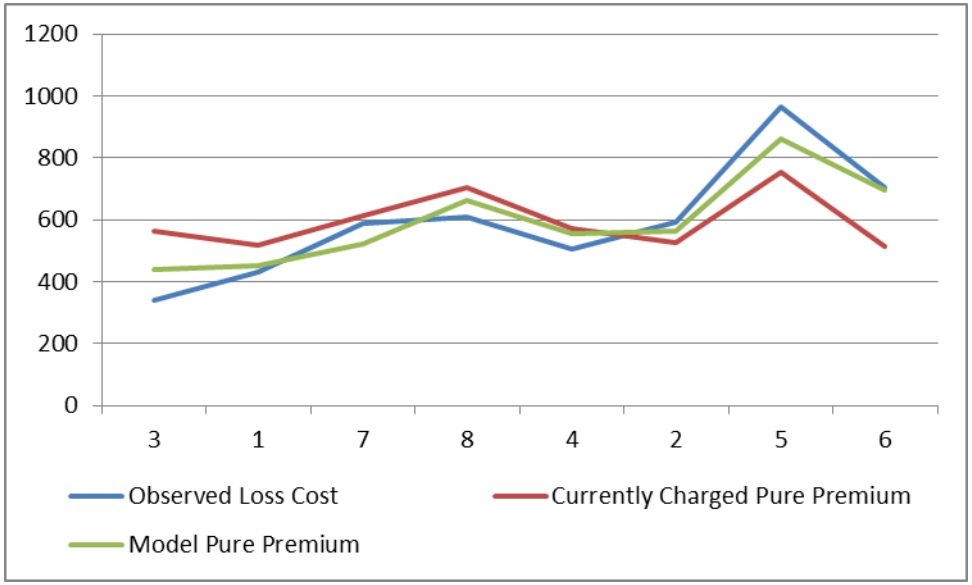
Percent improvement of the fit by segment – Validation Data

Segment	Observed Loss Cost	Currently Charged Pure Premium	Modeled Pure Premium	% Diff – Current to Observed	% Diff – Modeled to Observed	% Improvement
3	341.1	564.2	438.2	-65.4%	-28.5%	37.0%
1	429.4	518.8	449.7	-20.8%	-4.7%	16.1%
7	588.8	613.9	523.3	-4.3%	11.1%	-6.9%
8	608.9	704.9	661.4	-15.8%	-8.6%	7.1%
4	504.8	570.9	556.4	-13.1%	-10.2%	2.9%
2	591.0	526.2	565.2	11.0%	4.4%	6.6%
5	965.4	754.6	862.9	21.8%	10.6%	11.2%
6	705.7	515.4	696.8	27.0%	1.3%	25.7%

Homeowners

The final model was chosen by considering more than just lift and correlation.

Visual improvement of the fit by segment – Validation Data



Homeowners

The final model was chosen by considering more than just lift and correlation.

Total segment-level deviance improvement – Validation Data

Statistic	Average Deviance - Current to Observed	Average Deviance - Modeled to Observed	% Improvement
Simple Deviance	119.8	52.7	56.0%
Sum of Squares Deviance	19,347	3,819	80.3%
Chi-square Deviance	32.2	6.4	80.0%

Homeowners

The final chosen model:

Lift: 1.87 *Correlation: 92.6%*

Loss Ratio:	46.3%	54.6%	56.4%	58.5%	60.6%	69.6%	75.7%	86.7%
Exposures:	18,363	23,696	23,415	19,460	24,131	17,827	22,706	22,317
Variables	**Seg 3**	**Seg 1**	**Seg 7**	**Seg 8**	**Seg 4**	**Seg 2**	**Seg 5**	**Seg 6**
Policy Tenure	Long time	Not Long	Not New	Not New	Not New	New	New	New
Mortgage	No Mort	No Mort	Mortgage	Mortgage	Mortgage	Mortgage	Mortgage	Mortgage
Coverage A						Cheap	Not Cheap	Cheap
Age Roof						Not Old		Old
Endorsement1			Y	Y	N			
Prot Class			1-3	4+				

Note: results are specific to the given underlying class plan and should not be generalized to other companies. The model has also been modified for display.

Commercial Auto

Regional US commercial auto insurer – 6 years of data
End goal was to evaluate book profitability
Out of sample validation – 70% training, 30% validation
Analysis done at the policy level
Minimum claims per segment set at 750

Level	Earned Exposures	Claim Count	Loss Ratio*
Policy	271,239	7,339	43.2%

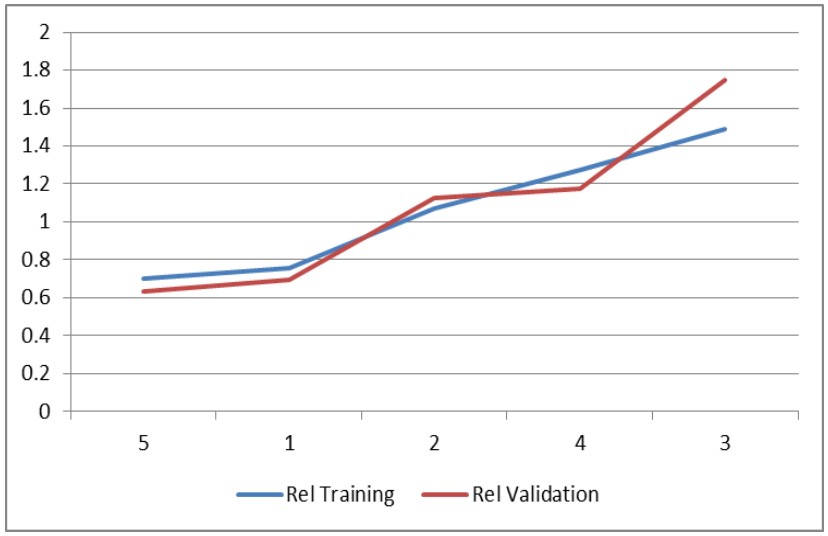
**Loss Ratio was calculated using rerated premium*

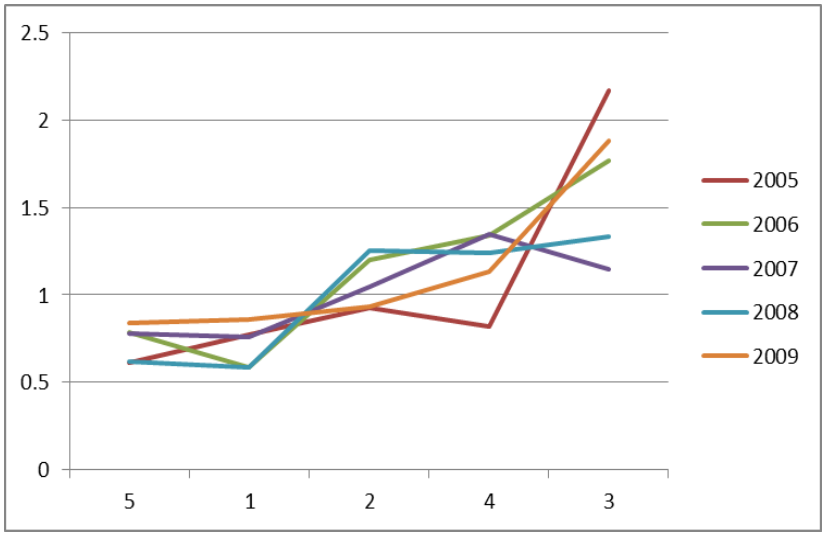
Commercial Auto

A similar collection of criteria was used to select a model. In this case, lift was 2.31.

Visual representations of correlation and consistency

Correlation: 97.7%





Commercial Auto

A similar collection of criteria was used to select a model.

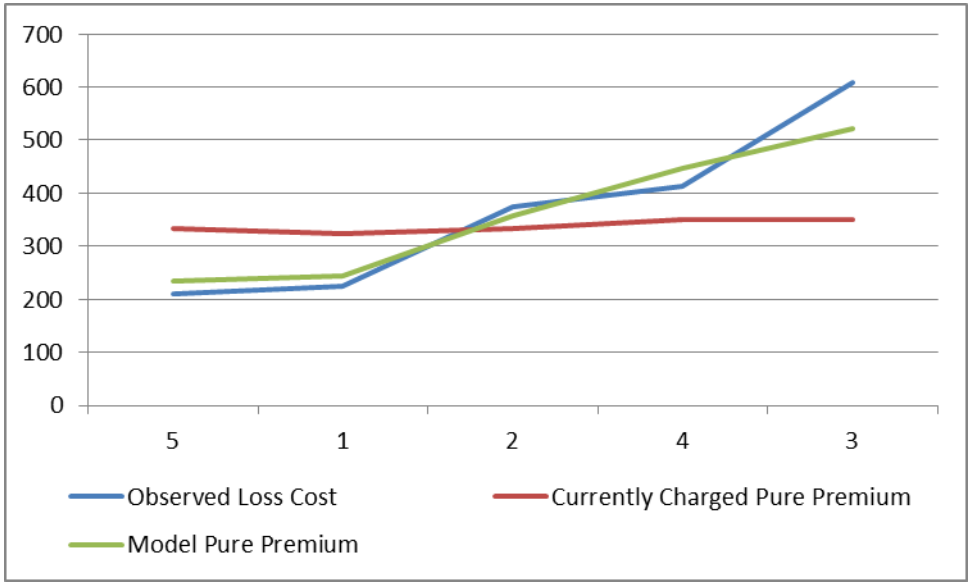
Percent improvement of the fit by segment – Validation Data

Segment	Observed Loss Cost	Currently Charged Pure Premium	Modeled Pure Premium	% Diff – Current to Observed	% Diff – Modeled to Observed	% Improvement
5	211.3	334.4	234.2	-58.3%	-10.8%	47.4%
1	224.0	323.3	244.0	-44.3%	-8.9%	35.4%
2	375.3	333.7	358.2	11.1%	4.5%	6.5%
4	412.1	350.9	446.8	14.9%	-8.4%	6.4%
3	609.5	349.1	521.0	42.7%	14.5%	28.2%

Commercial Auto

A similar collection of criteria was used to select a model.

Visual improvement of the fit by segment – Validation Data



Commercial Auto

A similar collection of criteria was used to select a model.

Total segment-level deviance improvement – Validation Data

Statistic	Average Deviance - Current to Observed	Average Deviance - Modeled to Observed	% Improvement
Simple Deviance	101.7	30.0	70.5%
Sum of Squares Deviance	14,606	1,386	90.5%
Chi-square Deviance	42.9	3.3	92.4%

Commercial Auto

The final (?) chosen model:

Lift: 2.31 *Correlation: 97.7%*

Loss Ratio:	29.3%	31.7%	46.9%	53.6%	67.7%
Exposures:	52,881	71,079	76,179	38,736	32,364
Variables	**Segment 5**	**Segment 1**	**Segment 2**	**Segment 4**	**Segment 3**
Pay Plan Change	Y or New Bus	N			Y or New Bus
% Drv with Viols	Low	Low	Moderate	High	Low
Ave Yrs Driving	Experienced				Not Experienced

Note: results are specific to the given underlying class plan and should not be generalized to other companies. The model has also been modified for display.

6. *Other Issues*

Limitations and some cautions

The three case studies showed how the amount of signal varies both due to the volatility of the underlying data and also due to the sophistication of the underlying class plan.

What are the limits of this approach? Can the underlying class plan be too good or too bad to find signal? Can the data be too volatile?

With respect to volatility… of course! Some datasets are too small with respect to their inherent volatility to be modeled.

Where that line is depends on the data.

Limitations and some cautions

It is also possible that the existing class plan captures the non-linear, interactive portion of the signal. This is not common in the industry today.

Finally, existing class plans can be so poor as to pose additional difficulties.

- When the underlying class plan does not capture the linear signal, Rule Induction tends to focus on that to the exclusion of other compound effects.
- In these cases, the best results are found by first modeling the linear signal through traditional methods, and then modeling the new residuals with Rule Induction.

7. *Summary*

Summary

- Insurer class plans, in general, do not capture the higher-order non-linear portion of the signal.

- Rule Induction is an effective technique for exploring the residuals (a.k.a. loss ratios) of existing class plans.

- Loss Ratio modeling allows insurers to identify customers their existing rating plan writes profitably/unprofitably. This can either…
  - … minimize implementation issues by finding adjustments to their current rating plan.
  - … allow for underwriting or other actions besides rating.

- These same techniques can be used in combination with 3rd party data to find even more signal in insurer data.

Questions?

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