Variable Selection

A Comparison of Various Techniques

A Presentation to RPM
March 11, 2015

Ben Williams
Greg Hansen
Ari Baraban

WESTFIELD INSURANCE
Sharing Knowledge. Building Trust.®

TOWERS WATSON

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.

This presentation summarizes the work detailed in the paper, "*A Practical Approach to Variable Selection – a Comparison of Various Techniques*", by Benjamin Williams, Greg Hansen, Aryeh Barbaran, and Alessandro Santoni.

The paper will be appearing in an upcoming edition of the CAS E-Forum (http://www.casact.org/pubs/forum/)

Agenda

- Variable Selection
  - The problem in general
  - How we framed it
- Methods
  - Lists of methods used
  - Overview of selected methods
- Approach
  - Data used
  - Process
  - Ranking criteria
- Conclusions
  - Results
  - Interpretation and next steps

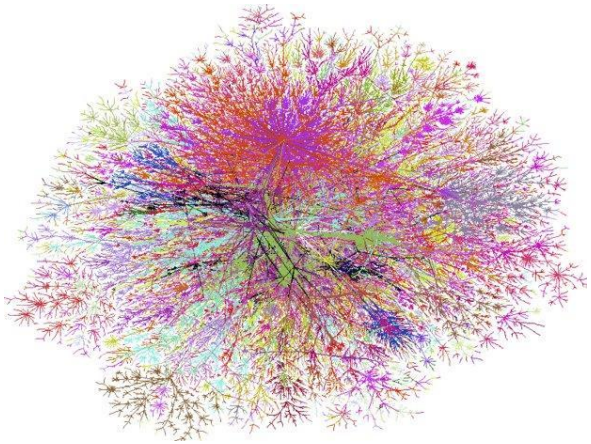
Variable Selection

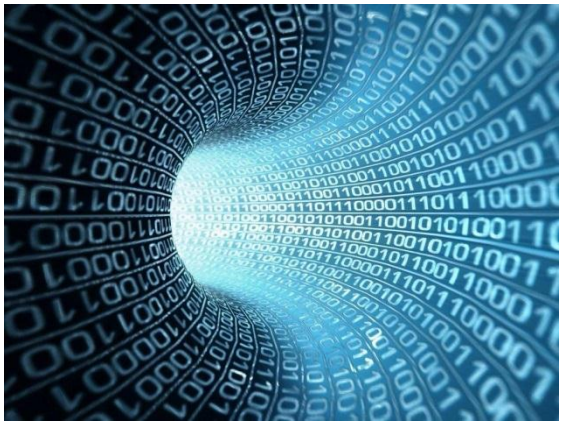
WESTFIELD INSURANCE
Sharing Knowledge. Building Trust.®

TOWERS WATSON

The Problem

- Between the data stored by companies and that available from external providers, modelers now have access to hundreds or even thousands of variables
- So many variables are available that it is often impractical to consider all of them in a formal predictive modeling context.
- This situation will only be accentuated in the future, as the number of candidate variables continues to grow (Big Data!)
- Recognizing which variables to consider in predictive modeling becomes an important problem for which automated approaches are required.





The Problem

- Note that we are not talking about Variable Reduction, which we consider to be the creation of “Super Variables”, functions of combinations of the original variables
- A classical example of Variable Reduction is Principal Component Analysis

The Problem

- Stated formally, the question we investigated was:

  How to select, from a (potentially very) long list of variables, a shortlist which will be useful for current predictive modeling techniques

- In particular, we limited our investigation to ordinal, pre-banded, anonymized geo-demographical variables
- Context was homeowners fire and water, frequency and severity
- We used a list of variable selection techniques to pick shortlists from the list of variables available, and compared the results, in terms of
  - Quality of the resulting shortlist in terms of usefulness for predictive modeling
  - Ease of use of the technique

  More details on both of these later…

Methods

WESTFIELD INSURANCE
Sharing Knowledge. Building Trust.®

TOWERS WATSON

Data Summary

Model Variable	Description
Y	Response Variable (frequency or severity for specific peril)
$X_1 – X_{15}$	"Core" rating attributes (Amount of Insurance, Deductible, Fire Protection, etc.)
$X_{16} – X_{377}$	362 Geo-demographic, ordinal, pre-banded variables (zip or census block level)

- 1.9 million observations
  - 5 years of Homeowners experience
    - Water & Fire Perils
    - Frequency & Severity
- 2/3 of observations used to train models, 1/3 held out for testing

Variable Selection Methods

- AIC Improvement Rank (AICRank)
- Stepwise GLM based on AIC Improvement With Correlated Variables Removed (GLMCorr)
- Stepwise Least Squares Regression with Correlated Variables Removed (LSRCorr)
- Elastic Net (ENet)
- Variable Clustering (Varclus)
- Classification and Regression Trees (CART)
- Random Selection

Base vs. Residual

- "Core Model" fit using the 15 "core rating attributes."
- When searching for additional candidate variables, should we control for the original 15?
- If yes, how?
  - If modeling in the same environment, the clear choice is to add the new candidate variables to the Core Model
  - If modeling in a different environment, options include:
    - More pure: Introduce the prediction from the initial model as an offset (presuming a model structure and software solution that supports offsets)
    - Less pure, but comparably effective: Divide the original response variable by the Core Model prediction to create a new response variable
- We tested several of the methods on both a Base and Residual basis to see if controlling for the Core Model made a material difference.

Method: AIC Improvement Rank (AICRank)

- Similar to the first pass of a stepwise process
- Create a distinct model for each of the candidate variables
  - In our experiment, the 1st, 2nd, and 3rd degree polynomials were concurrently introduced for each variable
- Rank order the models / variables based on the best (lowest) AIC statistic
- Strengths:
  - Relatively simple to program within the existing GLM framework
  - Runs reasonably efficiently with the right hardware and software
- Challenges:
  - Assumes all of the candidate variables are independent. If that's false, then the shortlist is likely to produce redundant candidates.

AICRank Setup

Residual

Step	Predictors
1	$X_1, X_2 \ldots X_{14}, X_{15}, X_{16}, X_{16}^2, X_{16}^3$
2	$X_1, X_2 \ldots X_{14}, X_{15}, X_{17}, X_{17}^2, X_{17}^3$
3	$X_1, X_2 \ldots X_{14}, X_{15}, X_{18}, X_{18}^2, X_{18}^3$
4	$X_1, X_2 \ldots X_{14}, X_{15}, X_{19}, X_{19}^2, X_{19}^3$
…	…
362	$X_1, X_2 \ldots X_{14}, X_{15}, X_{377}, X_{377}^2, X_{377}^3$

Base

Step	Predictors
1	$X_{16}, X_{16}^2, X_{16}^3$
2	$X_{17}, X_{17}^2, X_{17}^3$
3	$X_{18}, X_{18}^2, X_{18}^3$
4	$X_{19}, X_{19}^2, X_{19}^3$
…	…
362	$X_{377}, X_{377}^2, X_{377}^3$

Recall that $X_1 – X_{15}$ are the “core” rating variables. If modeling on a residual basis, then they should be included in every step.

AICRank – Correlation Matrix (1st 20 variables)

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19
v1																			
v2	-0.40																		
v3	0.89	-0.38																	
v4	0.83	-0.49	0.79																
v5	0.87	-0.36	0.84	0.72															
v6	-0.36	0.99	-0.35	-0.44	-0.33														
v7	-0.01	-0.10	0.01	0.09	0.01	-0.09													
v8	-0.32	0.54	-0.27	-0.16	-0.28	0.55	-0.16												
v9	-0.36	0.89	-0.34	-0.40	-0.32	0.91	-0.13	0.65											
v10	-0.09	0.08	-0.04	0.12	-0.07	0.11	-0.12	0.72	0.20										
v11	-0.25	0.69	-0.24	-0.29	-0.26	0.70	-0.10	0.46	0.72	0.13									
v12	-0.42	0.66	-0.37	-0.30	-0.38	0.68	-0.02	0.68	0.75	0.31	0.53								
v13	-0.37	0.78	-0.35	-0.38	-0.35	0.78	-0.20	0.53	0.78	0.12	0.66	0.67							
v14	0.41	-0.78	0.39	0.51	0.36	-0.77	0.34	-0.63	-0.79	-0.18	-0.58	-0.63	-0.75						
v15	0.53	-0.59	0.63	0.76	0.54	-0.56	0.18	-0.20	-0.53	0.11	-0.39	-0.36	-0.50	0.58					
v16	0.33	-0.68	0.32	0.42	0.29	-0.67	0.24	-0.51	-0.67	-0.12	-0.51	-0.53	-0.69	0.78	0.54				
v17	-0.34	0.79	-0.31	-0.33	-0.31	0.82	-0.08	0.68	0.90	0.24	0.65	0.90	0.75	-0.73	-0.46	-0.62			
v18	-0.31	0.86	-0.30	-0.33	-0.28	0.89	-0.09	0.65	0.96	0.23	0.71	0.74	0.75	-0.74	-0.48	-0.63	0.90		
v19	-0.53	0.59	-0.53	-0.75	-0.47	0.55	-0.23	0.26	0.53	-0.04	0.40	0.35	0.52	-0.62	-0.75	-0.52	0.45	0.47	
v20	0.48	-0.64	0.41	0.27	0.43	-0.66	0.04	-0.80	-0.76	-0.47	-0.56	-0.77	-0.65	0.63	0.31	0.51	-0.79	-0.77	-0.32

Stepwise GLM based on AIC Improvement With Correlated Variables Removed (GLMCorr)

- Extends AICRank to a full forward stepwise process
- After each step, remove highly correlated candidates from all subsequent steps
  - Reduces processing time
  - Reduces likelihood of unstable, offsetting variables (a common affliction of high-dimensional stepwise processes)
  - User must select an "acceptable" correlation threshold
  - Used 0.35 for this exercise
- Strengths:
  - Addresses shortcomings of AICRank
  - Stays within GLM framework
- Issues:
  - Impractically slow for large datasets / variable lists (solve thousands of individual models for this example – more than 24 hours in SAS)

GLMCorr Setup (shown on a residual basis)

Step	Predictors
1.1	$X_1, X_2 \ldots X_{14}, X_{15}, X_{16}, X_{16}^2, X_{16}^3$
1.2	$X_1, X_2 \ldots X_{14}, X_{15}, X_{17}, X_{17}^2, X_{17}^3$
1.3	$X_1, X_2 \ldots X_{14}, X_{15}, X_{18}, X_{18}^2, X_{18}^3$
1.4	$X_1, X_2 \ldots X_{14}, X_{15}, X_{19}, X_{19}^2, X_{19}^3$
…	…
1.362	$X_1, X_2 \ldots X_{14}, X_{15}, X_{377}, X_{377}^2, X_{377}^3$

Step 1.82 (Variable X_{97}) produces the lowest AIC.

Retain X_{97}, eliminate all correlated variables, and proceed to Step 2

Step	Predictors
2.1	$X_1, X_2 \ldots X_{14}, X_{15}, X_{97}, X_{97}^2, X_{97}^3, X_{16}, X_{16}^2, X_{16}^3$
2.2	$X_1, X_2 \ldots X_{14}, X_{15}, X_{97}, X_{97}^2, X_{97}^3, X_{17}, X_{17}^2, X_{17}^3$
2.3	$X_1, X_2 \ldots X_{14}, X_{15}, X_{97}, X_{97}^2, X_{97}^3, X_{19}, X_{19}^2, X_{19}^3$
2.4	$X_1, X_2 \ldots X_{14}, X_{15}, X_{97}, X_{97}^2, X_{97}^3, X_{20}, X_{20}^2, X_{20}^3$
…	…
2.350	$X_1, X_2 \ldots X_{14}, X_{15}, X_{97}, X_{97}^2, X_{97}^3, X_{376}, X_{376}^2, X_{376}^3$

Step 2.23 (Variable X_{40}) produces the lowest AIC.

Retain X_{97}(Step 1) and X_{40}(Step 2), eliminate all correlated variables, and proceed to Step 3

Note: In this (fabricated) example, 11 variables were eliminated for Step 2 due to high correlation with X_{97}. X_{18} and X_{377} were among the eliminated variables

Stepwise Least Squares Regression with Correlated Variables Removed (LSRCorr)

- Nearly identical concept to GLMCorr
- Substitutes simple regression (Ordinary Least Squares) for GLM model structure
- Strengths:
  - Replicates most of the strengths of GLMCorr
  - Much faster (well under an hour to produce 50 candidate variables)
  - Shortlist is very similar to GLMCorr's
- Issues:
  - Inconsistent modeling assumptions (identity link function and normal error structure not typically associated with ratemaking)
    - Given the purpose, this may not be an issue
  - May need to leave existing modeling environment

Is it necessary to control for correlation when using stepwise methods?

- Are we redundantly addressing the problems with AICRank by introducing both a stepwise process and a correlation censor?
- It depends
  - The shortlist created without the correlation adjustment might perform as well or better
  - BUT: stepwise processes can still produce highly correlated variables, often with competing / unintuitive signs
    - Can produce unstable models (volatile parameter estimates)
    - Can produce extreme (and likely inaccurate) predictions at the edges

Many Statisticians Hate Stepwise Methods

"A very popular technique for many years, but if it had just been proposed as a statistical method, it would likely be rejected because it violates every principle of statistical estimation and hypothesis testing"

– Frank Harrell

"Treat all claims based on stepwise algorithms as if they were made by Saddam Hussein on a bad day with a headache having a friendly chat with George Bush"

– Steve Blinkhorn

"I don't know what knowledge we would lose if all papers using stepwise regression were to vanish from journals at the same time as programs providing their use were to become terminally virus-laden"

– Ira Bernstein

Why Statisticians Hate Stepwise

- Regression coefficients are biased too high
  - A predictor is more likely to be included if its coefficient is overestimated
  - A predictor is less likely to be included if its coefficient is underestimated
- Too many "false positives" for unimportant predictors (particularly when number of candidate predictors is high)
- Unstable -- Small changes in the data can produce significant changes in predictions
  - Particularly when there are redundant predictors
- Incorrect distributional assumptions (The F and χ^2 test statistics do not have the claimed distribution.)

paraphrased from Regression Modeling Strategies, Harrell (2001)

Elastic Net (ENet)

- A form of penalized regression (extension of GLM)
  - Parameter estimates are "penalized" as they get larger or as additional predictors are added to the model.
  - Addresses some of the bias issues inherent in stepwise processes
- Variables enter the model one at a time. This order indicates the relative importance of variables in explaining the signal which was used to create our shortlist of candidate variables.
- Strengths:
  - Handles redundant / highly correlated variables
  - Handles high-dimensional modeling datasets
  - Theoretically more stable and computationally efficient
- Challenges:
  - Most robust implementations are in R
  - R has memory-based limitations
  - Learning curve if you are unfamiliar with the tool or with method

Variable Clustering (Varclus)

- Used VARCLUS procedure (Base SAS) – similar procedures exist in other packages
- Closely related to Principal Component Analysis
- Finds clusters of variables that are closely correlated with each other and as uncorrelated as possible with variables in other clusters
- For our shortlist, we selected the variable from each cluster with the lowest 1-R^2 ratio (that is, the variable most representative of its cluster)
- Strengths
  - Very fast – typically runs in seconds, even with large datasets
  - Especially useful in large datasets with redundant variables
- Issues
  - Not influenced by the response variable
  - May still produce highly correlated variables for the shortlist

Correlation Matrix: Visualizing VarClus Clusters

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19	v20
v1																				
v2	0.95		Cluster 1																	
v3	0.96	0.90																		
v4	-0.90	-0.95	-0.86																	
v5	-0.89	-0.88	-0.85	0.85																
v6	-0.30	-0.27	-0.21	0.25	0.31															
v7	-0.32	-0.29	-0.22	0.27	0.33	0.99		Cluster 2												
v8	-0.32	-0.29	-0.22	0.27	0.31	0.98	0.97													
v9	-0.21	-0.16	-0.14	0.15	0.19	0.94	0.94	0.92												
v10	0.20	0.17	0.10	-0.16	-0.20	-0.77	-0.77	-0.76	-0.69											
v11	-0.32	-0.37	-0.24	0.37	0.26	0.32	0.33	0.35	0.28	-0.29										
v12	-0.34	-0.38	-0.25	0.39	0.28	0.33	0.35	0.35	0.31	-0.28	0.91		Cluster 3							
v13	-0.29	-0.34	-0.20	0.34	0.24	0.32	0.34	0.34	0.31	-0.26	0.88	0.91								
v14	-0.33	-0.35	-0.26	0.36	0.30	0.35	0.37	0.36	0.32	-0.27	0.82	0.78	0.77							
v15	-0.20	-0.22	-0.15	0.23	0.16	0.24	0.25	0.26	0.22	-0.18	0.46	0.45	0.44	0.41						
v16	0.26	0.36	0.17	-0.38	-0.34	-0.04	-0.07	-0.03	0.01	0.03	-0.28	-0.30	-0.27	-0.21	-0.13					
v17	-0.19	-0.25	-0.11	0.27	0.28	0.14	0.17	0.14	0.09	-0.10	0.18	0.23	0.21	0.23	0.07	-0.49		Cluster 4		
v18	-0.22	-0.29	-0.14	0.32	0.30	0.17	0.20	0.18	0.09	-0.12	0.23	0.25	0.23	0.25	0.12	-0.49	0.96			
v19	0.48	0.52	0.42	-0.50	-0.55	-0.01	-0.05	0.00	0.01	0.01	-0.25	-0.28	-0.25	-0.23	-0.09	0.72	-0.39	-0.37		
v20	0.38	0.45	0.30	-0.44	-0.45	0.09	0.04	0.09	0.11	-0.05	-0.27	-0.32	-0.29	-0.21	-0.07	0.69	-0.40	-0.35	0.74	

Classification and Regression Trees (CART)

- Standard implementation of CART by Salford Systems
- Binary splitting algorithm – not related to GLM
- Produces a "variable importance" ranking, which we used to generate our shortlists
- Strengths
  - Non-parametric / not constrained by GLM (or any) model structure
  - Easy and fast to set up and run
- Issues
  - Potentially requires separate software purchase and expertise
    - In addition to Salford Systems, there are implementations of CART in Matlab and R (and maybe others)
  - Prediction space is not continuous, so results can be unintuitive / difficult to interpret

Random Selection

Discussion

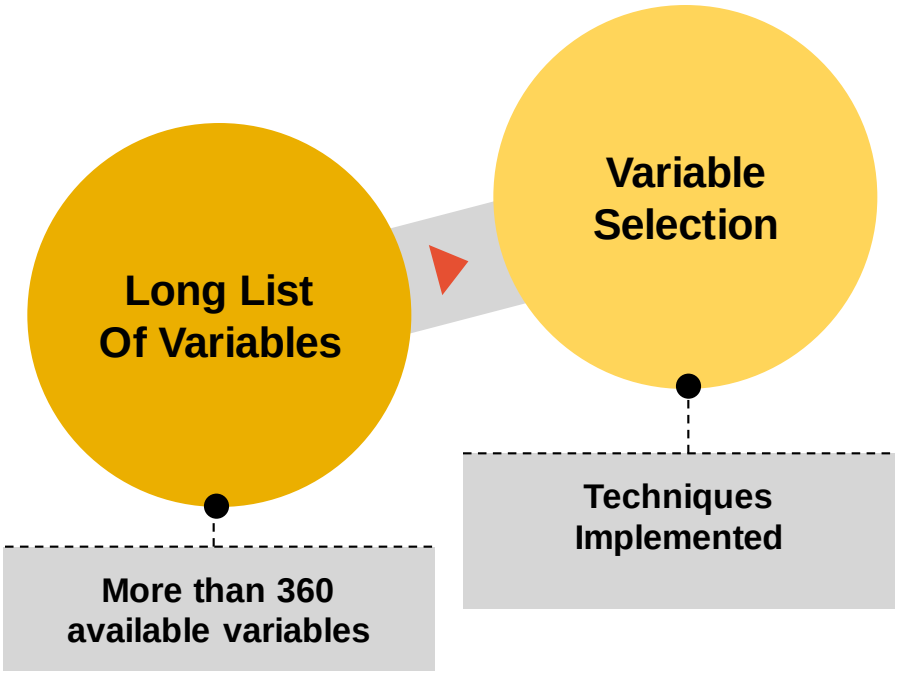
- Goal is to produce candidates for further analysis
  - Through that lens, purity of modeling assumptions may be less important
- Techniques are not mutually exclusive
  - Can (and probably should) be used in conjunction
- Not meant to be an exhaustive list of methods to produce shortlist
  - Our goals are to start the conversation and provide immediate, practical options

Approach

WESTFIELD INSURANCE
Sharing Knowledge. Building Trust.®

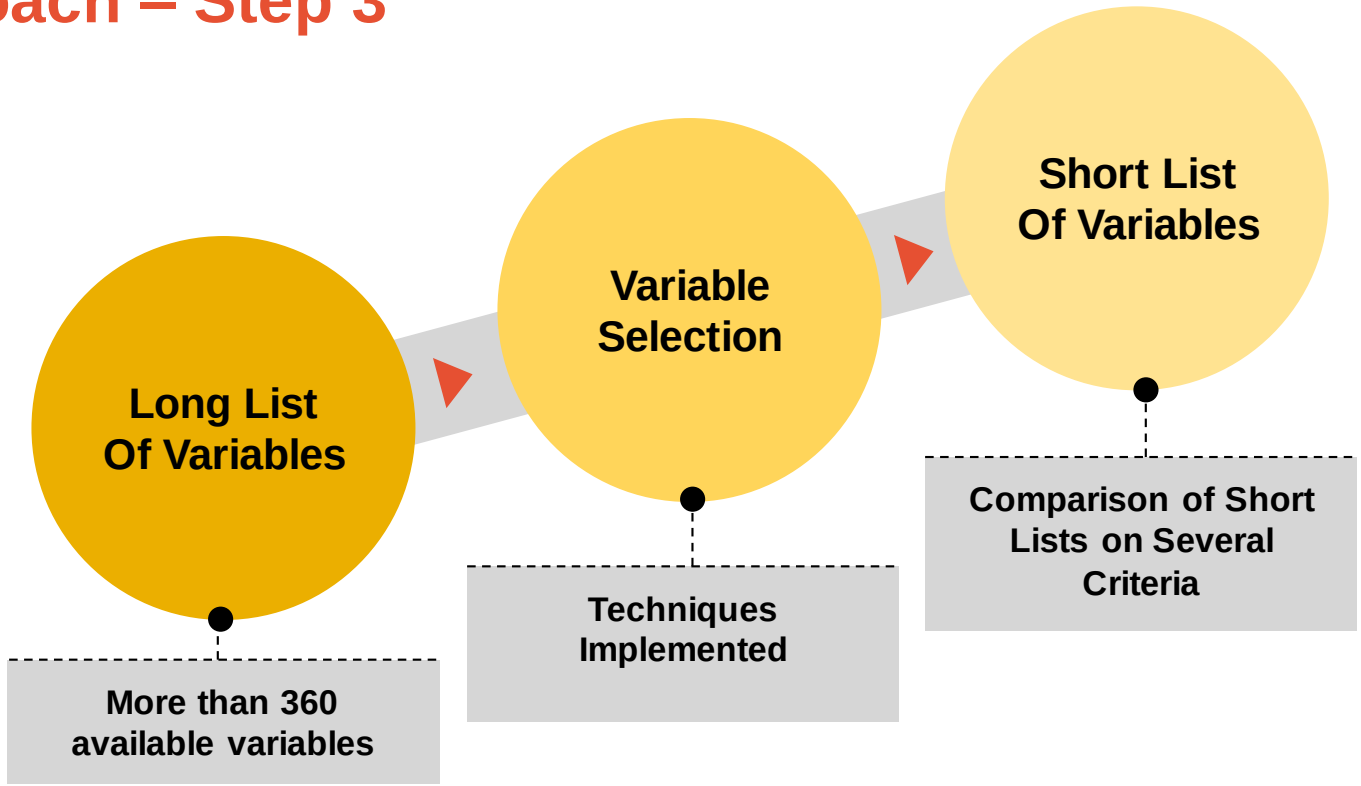
TOWERS WATSON

Approach – Step 1 & 2



- The techniques were implemented on the full data set in different software environments
- This process generated a shortened list of variables
  - Narrow down “the best” 50 variables
    - 50 was chosen to represent a manageable number of variables
- Short list of variables for consideration in a GLM model
- This short list will represent the technique throughout the process

Approach – Step 3



- Using each short list, fit a model with an automated stepwise process
- Frequency and severity models fit separately

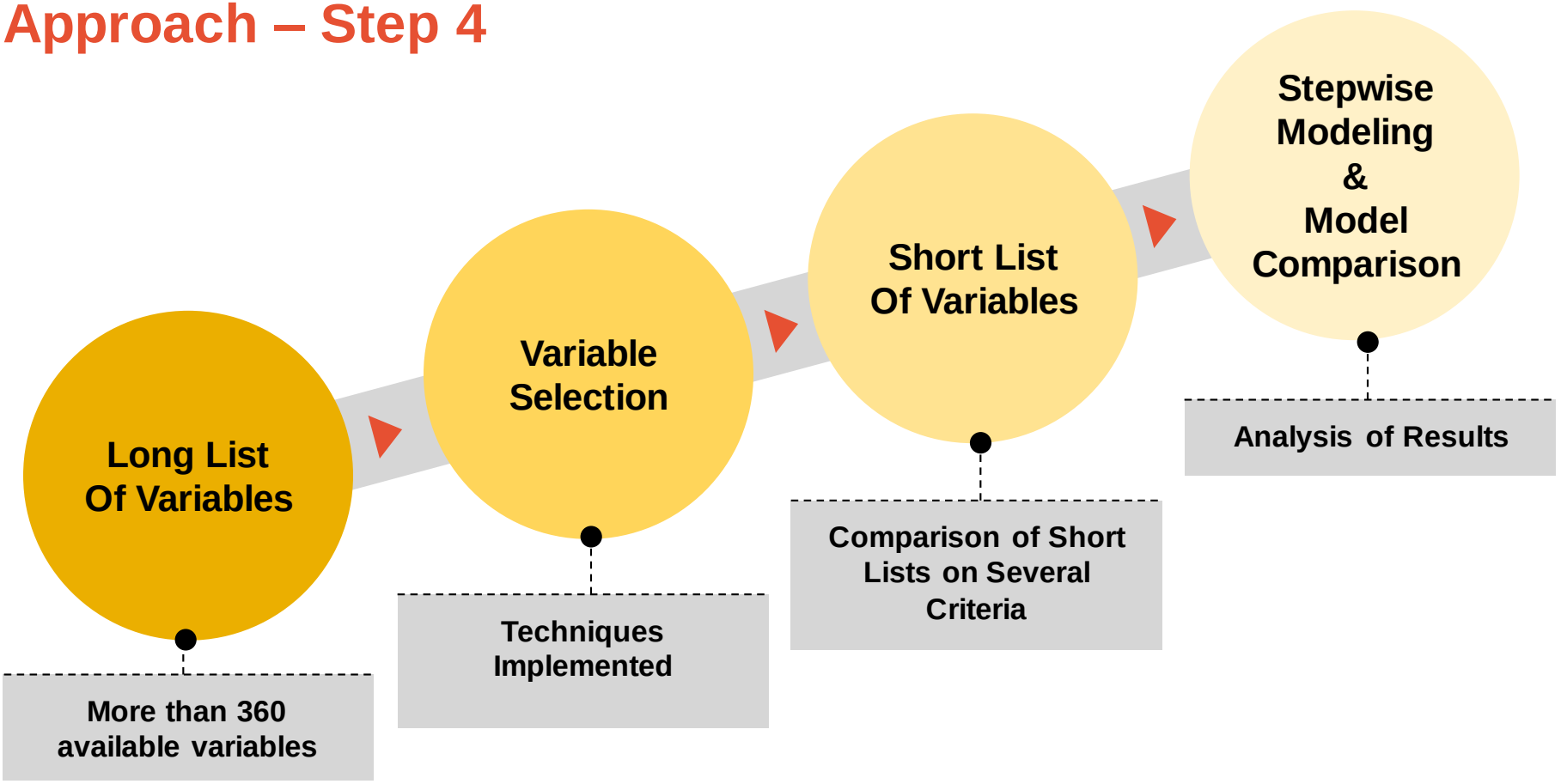
Modeling the Shortlists

- Starting with a base model from the policy-related variables
- Using the shortlist to supplement this base model
- All variables in the shortlist were considered
- Custom stepwise regression performed in Emblem
  1. Variables were “checked into” the model one at a time (as linear effects)
  2. The change in AIC was recorded
  3. Variables were then “checked out”
  4. The variable with the greatest decrease in AIC is added to the model
- Process is repeated with remaining variables on the shortlist until there are no further decreases in AIC (or the decrease is less than .05%)

Modeling the Shortlists

- Automated process was designed to eliminate potential bias from a human modeler
- Under normal circumstances, we would not recommend modeling in a "blind" manner
- Issues associated with using a stepwise process
- Some choices we made could have affected outcome
  - Using AIC as selection criteria
  - Only considering first order polynomial effects
  - Setting length of 50 was an arbitrary choice
    - Shorter lists were explored in some cases

Approach – Step 4



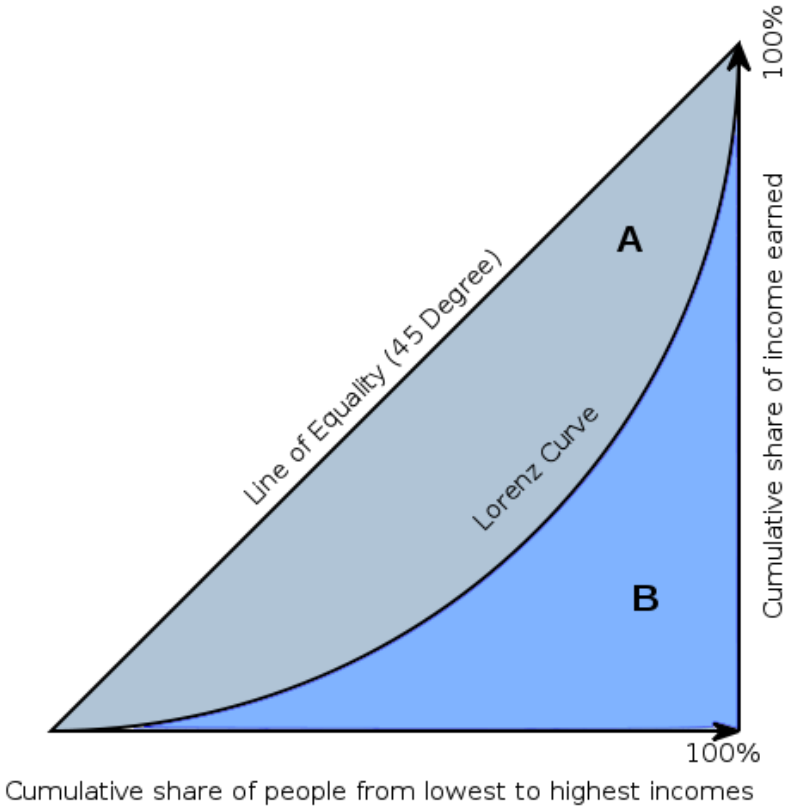
- At this point there is a frequency and severity model associated with each variable selection technique
- These models can be compared against one another

Ranking of Models

- Several Criteria Used
  - Gini Coefficients
  - Double Lift Charts
  - Ranking of Deviance
  - Modeling Considerations

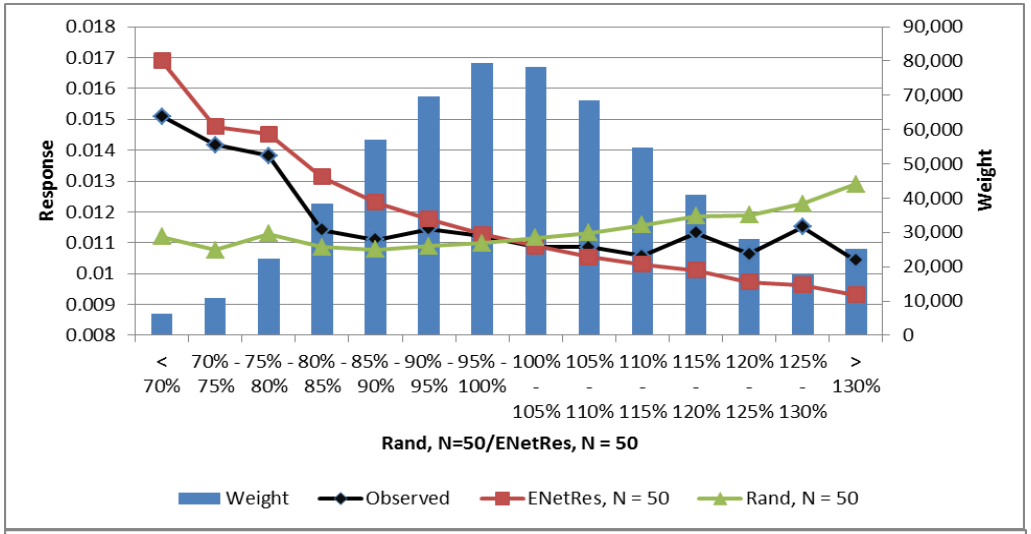
Ranking of Models – Gini Coefficient

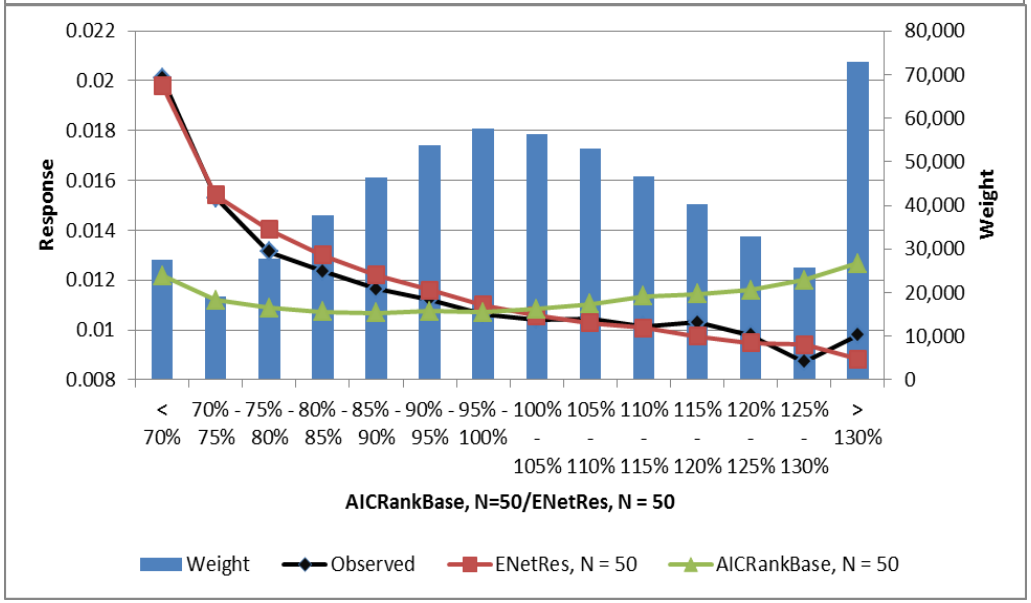
- Gini coefficient is based on Lorenz Curve
  - Quantification of Lorenz Curve
  - Equal to [area of section A]*2
- Prefer steeper curve
- Select model with larger value
- Issues
  - Distribution of values is unknown
  - Hard to tell if differences are significant
  - Range of values is dependent on the dataset
  - Only assesses model's lift



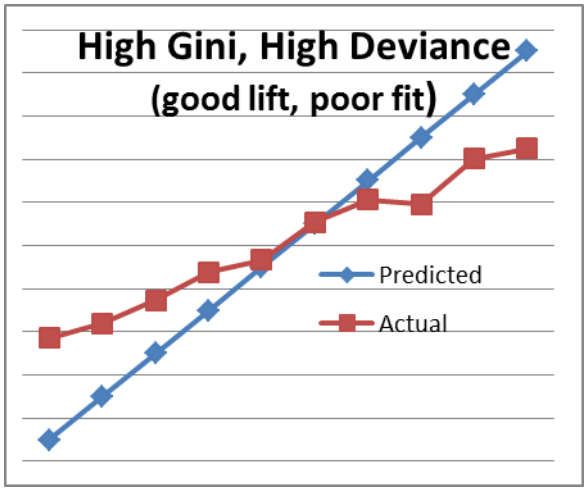
Ranking of Models – Double Lift Chart

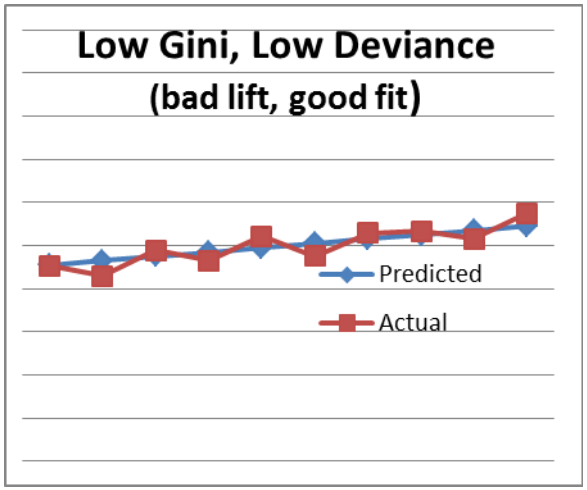
- Calculates ratio of fitted values from 2 models
- These ratios are divided into bands (displayed on x-axis)
- Histogram shows distribution across bands
- Within each band:
  - Red line shows model 1's average prediction
  - Green line shows model 2's average prediction
  - Dark line shows observed response
- Prefer model that is closer to observed line
- Qualitative approach
- Administered subjectively
- Only compares 2 models to one another

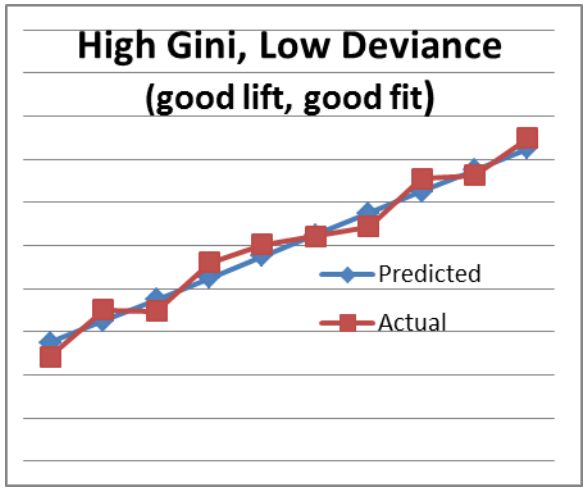




Lift vs. Goodness of Fit

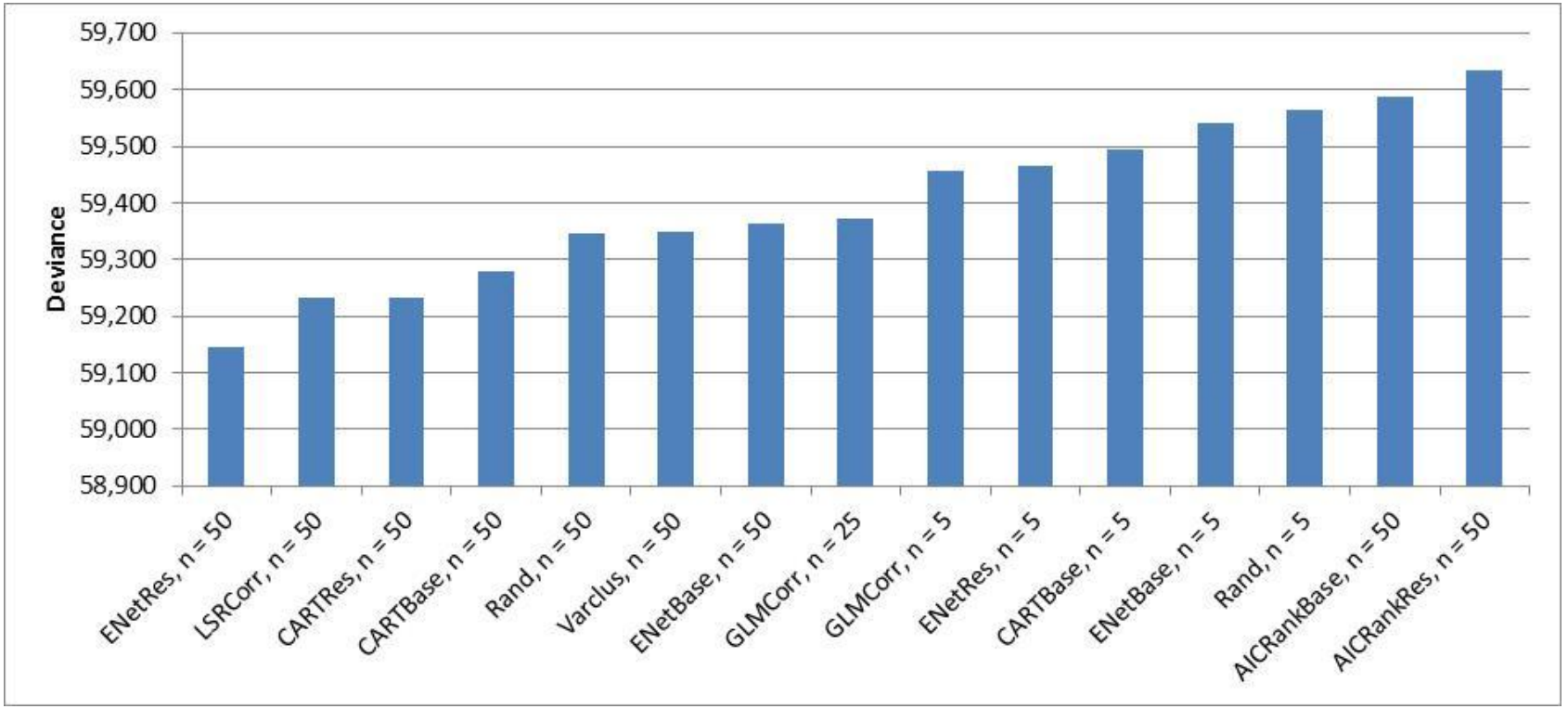






- Previous methods only consider the lift of a model, but lift is not the only relevant measure
- First and third graphs have the same lift (Gini), but much different goodness-of-fit
- Deviance, a measure of goodness-of-fit, was selected as another statistic on which to rank the methods

Ranking of Models – Ranking of Deviance



- Calculate the deviance of each model on testing dataset
- Simply rank the models in order of deviance
- Select model with lowest deviance

Results

WESTFIELD INSURANCE
Sharing Knowledge. Building Trust.®

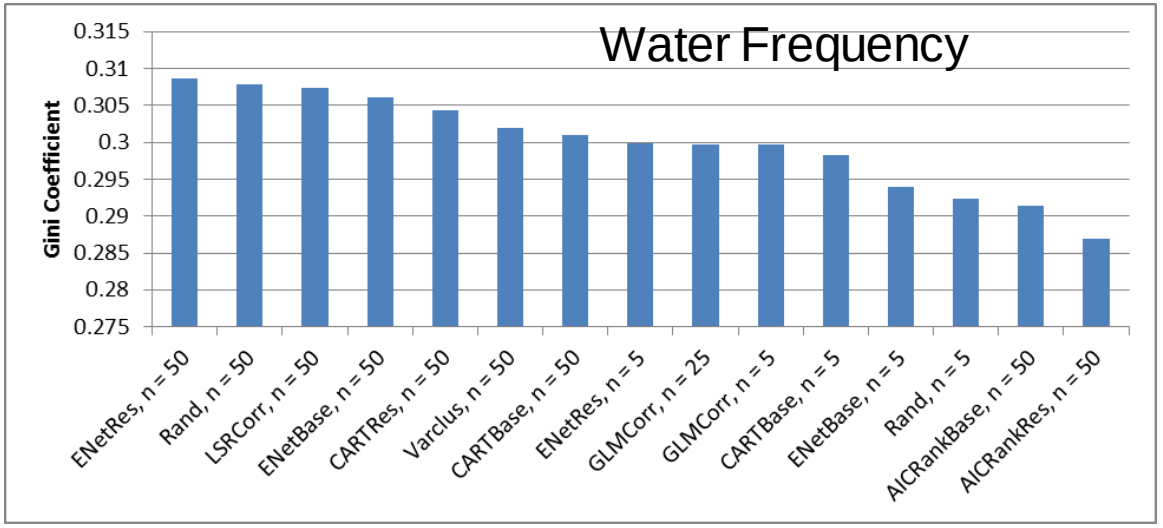
TOWERS WATSON

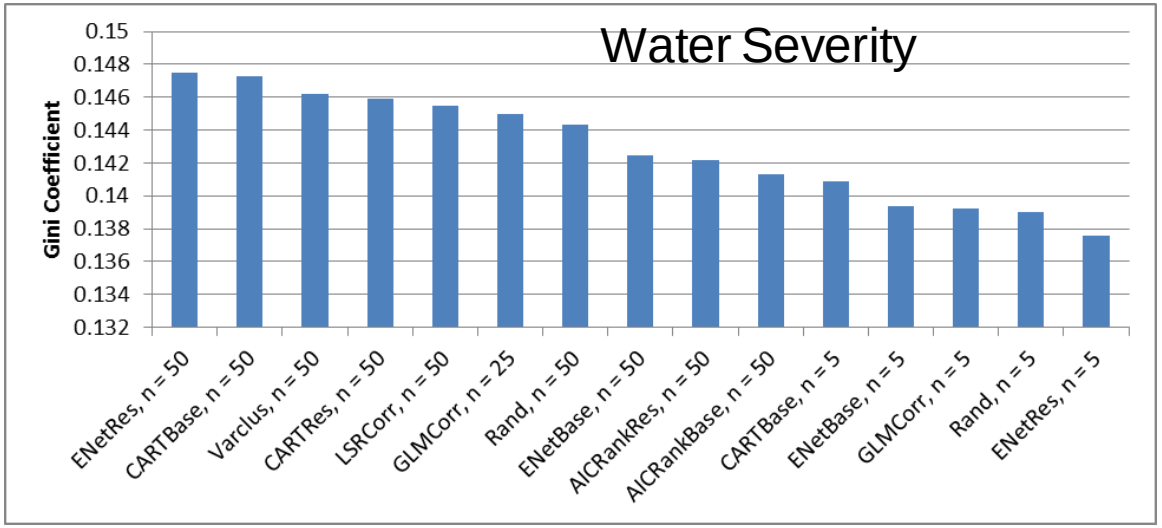
Results of Analysis – Gini Coefficient

		Water		Fire	
Method	N=	Frequency	Severity	Frequency	Severity
AICRankBase	50	0.2914	0.1413	0.2764	0.1018
AICRankRes	50	0.2869	0.1422	0.2766	0.1273
CARTBase	5	0.2982	0.1409		
CARTBase	50	0.3010	0.1473		
CARTRes	50	0.3043	0.1459		
ENetBase	5	0.2939	0.1394		
ENetBase	50	0.3060	0.1425	0.2806	0.1257
ENetRes	5	0.2999	0.1376		
ENetRes	50	0.3086	0.1475	0.2820	0.1168
GLMCorr	25	0.2997	0.1450		
GLMCorr	5	0.2997	0.1392		
Rand	5	0.2924	0.1390		
Rand	50	0.3079	0.1443	0.2711	0.1585
VarClus	50	0.3020	0.1462	0.2692	0.1521
LSRCorr	50	0.3073	0.1455		
LSRCorr	48				0.1166
LSRCorr	45			0.2781	

Results of Analysis – Gini Coefficient

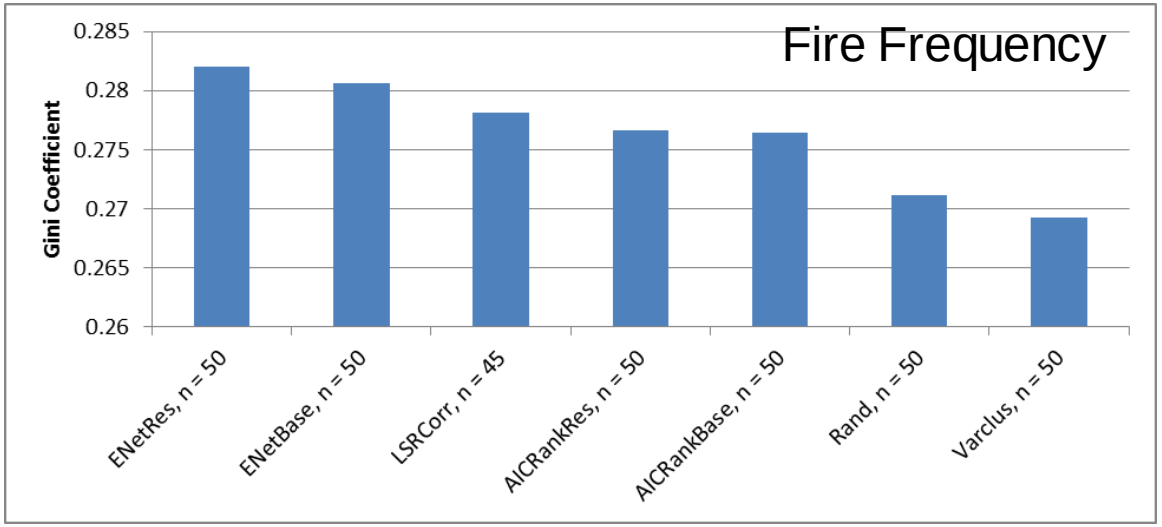
- Elastic Net is top performing for Frequency (top chart) and Severity (bottom)
- Generally, longer shortlists perform better

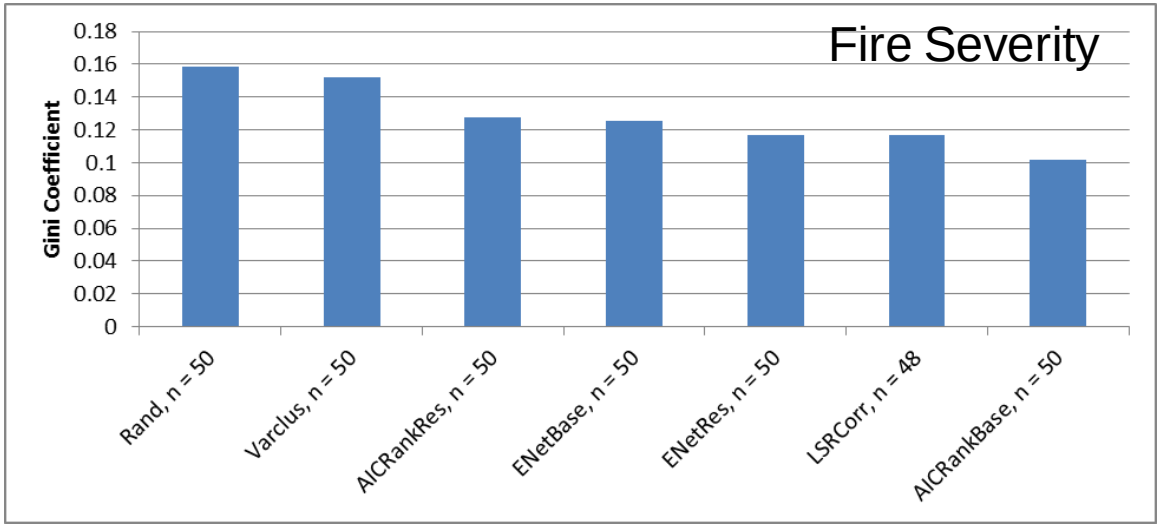




Results of Analysis – Gini Coefficient

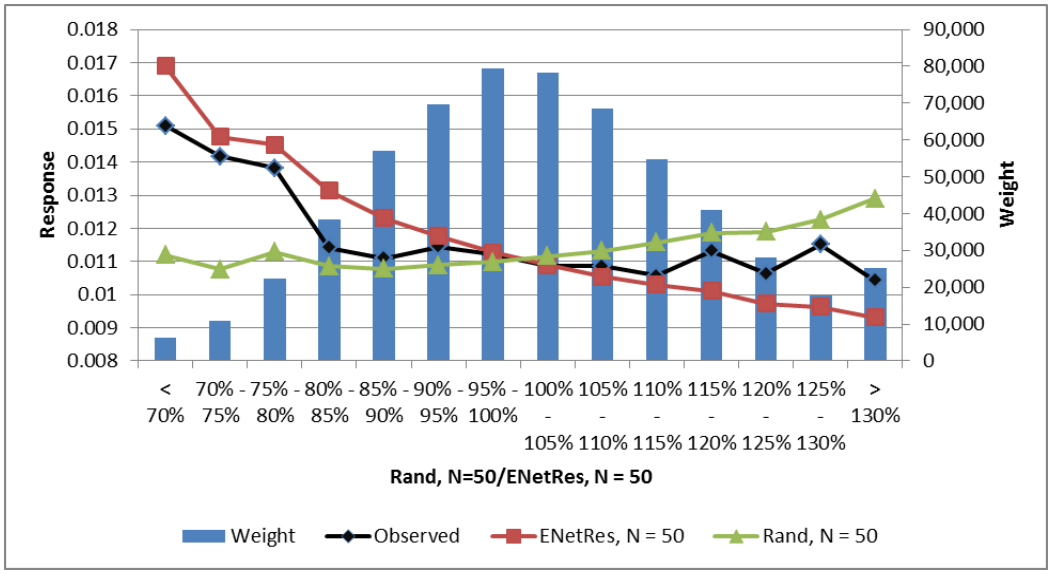
- Elastic Net is top performing for frequency (top chart)
- Random is top performing for severity (bottom chart)
- Fewer claims than water peril
- Severity has less data than frequency

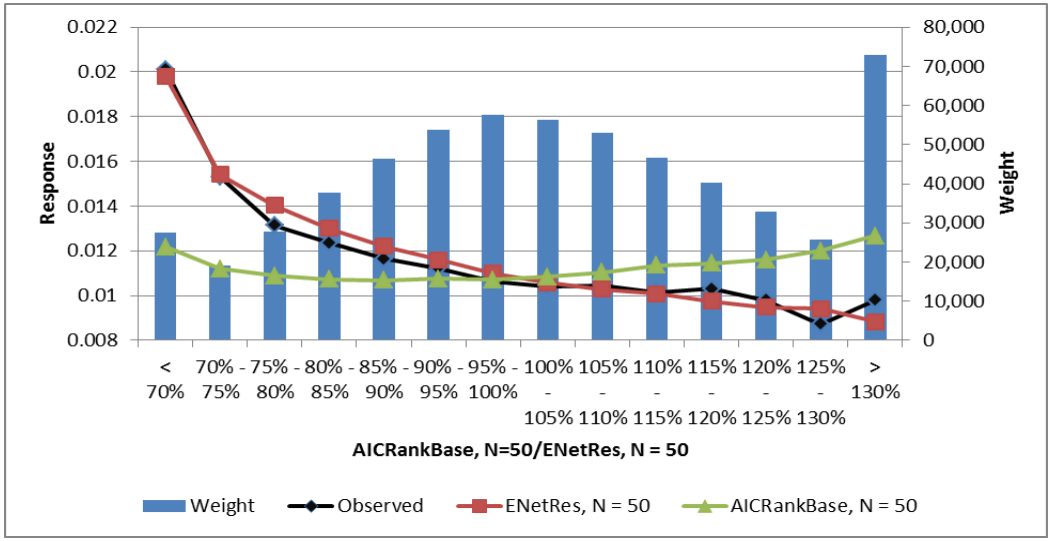




Results of Analysis – Double Lift Charts

- Best Gini coefficients from water frequency compared
- Shows that these models perform similarly
  - Random (green) performs well in less populated bandings 80%-90% & 115%-120%
  - ENet (red) performs well in highly populated bands, but the difference is closer
- When comparing best & worst models, the result is clearer
- Seems to confirm Gini rankings



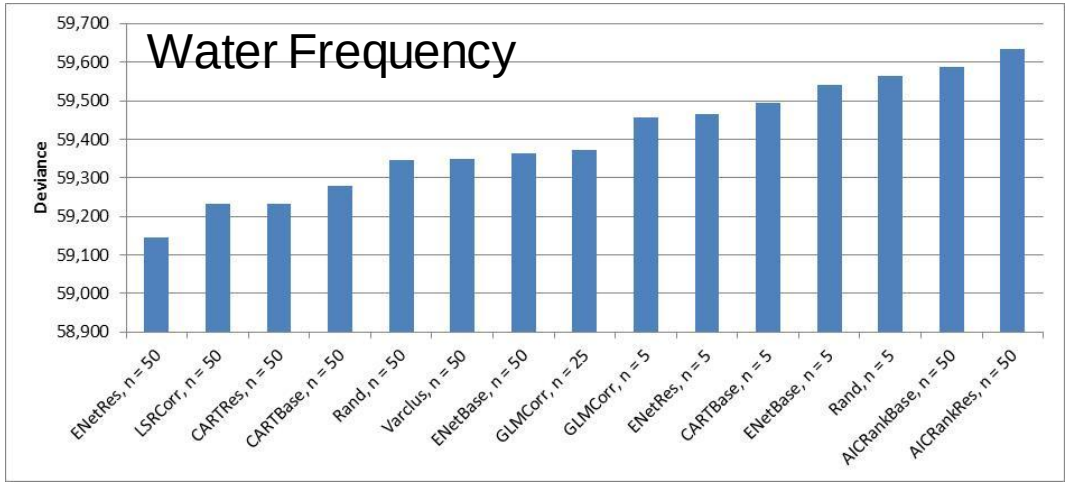


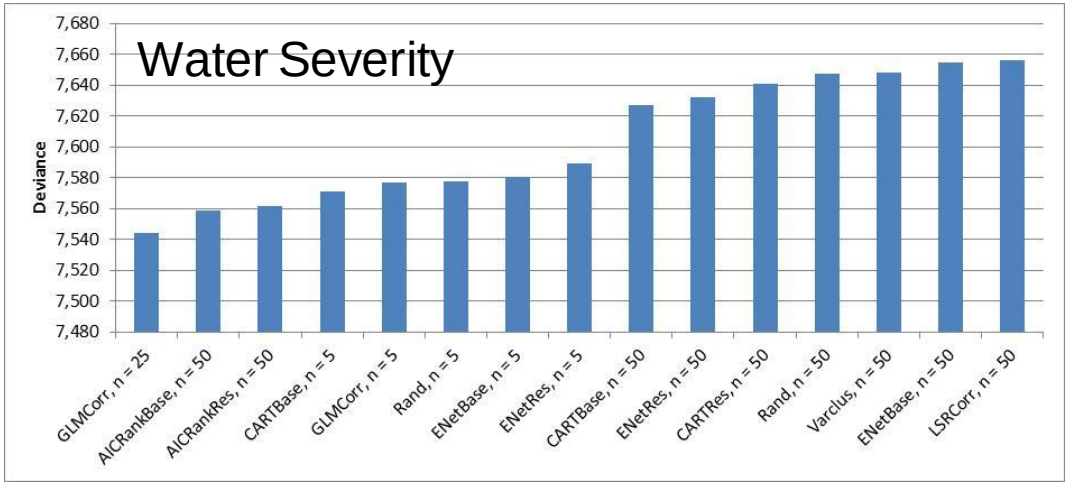
Results of Analysis – Model Deviance

		Water		Fire	
Method	N=	Frequency	Severity	Frequency	Severity
AICRankBase	50	59,587	7,559	14,251	3,738
AICRankRes	50	59,634	7,562	14,251	3,712
CARTBase	5	59,495	7,571		
CARTBase	50	59,279	7,627		
CARTRes	50	59,234	7,641		
ENetBase	5	59,541	7,581		
ENetBase	50	59,362	7,655	14,237	3,715
ENetRes	5	59,466	7,589		
ENetRes	50	59,145	7,632	14,230	3,749
GLMCorr	25	59,374	7,544		
GLMCorr	5	59,457	7,577		
Rand	5	59,564	7,577		
Rand	50	59,345	7,648	14,259	3,735
VarClus	50	59,350	7,648	14,260	3,733
LSRCorr	50	59,233	7,656		
LSRCorr	48				3,703
LSRCorr	45			14,238	

Results of Analysis – Model Deviance

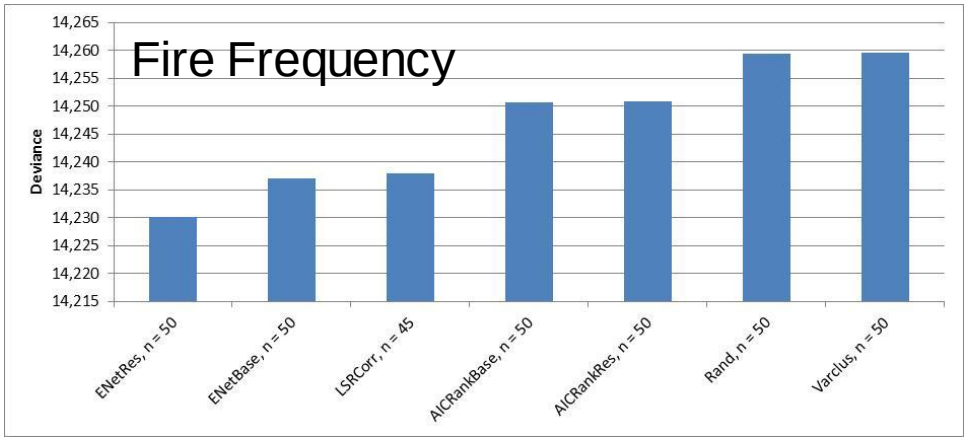
- Elastic Net is top performing for Frequency
- GLMCorr is top performing for severity

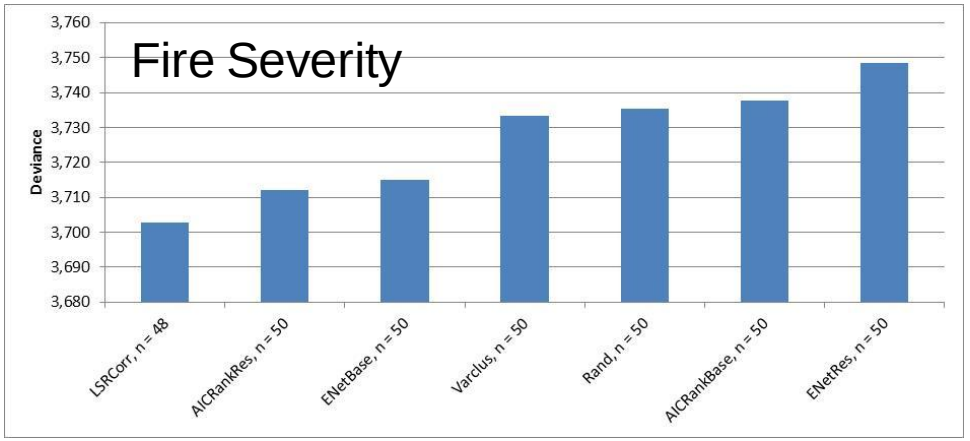




Results of Analysis – Model Deviance

- Elastic Net is top performing for Frequency
- LSRCorr is top performing for severity





Ranking of Models

- We also took into consideration
  - Software requirements
  - Ease of implementation
  - Processing Speed
  - Specialist knowledge required

Method	Software Used	Complexity to Set-Up	Processing Speed
AICRank	Emblem	Easy	Average
CART	CART	Easy	Fast
ENetBase	R	Average	Fast
GLMCorr	SAS	Easy	Slow
LSRCorr	SAS	Easy	Fast
Rand	-	Trivial	None required
VarClus	SAS	Easy	Fast

Conclusions and Next Steps

WESTFIELD INSURANCE
Sharing Knowledge. Building Trust.®

TOWERS WATSON

Conclusions

- If we had to pick a winner, we would go with Elastic Net on Residuals
- Other strong performers were
  - Elastic Net on Response
  - Stepwise GLM based on AIC Improvement With Correlated Variables Removed (GLMCorr)
  - Stepwise Least Squares Regression with Correlated Variables Removed (LSRCorr)
- We were surprised by the performance of the Random
- Suggestion that residual methods outperform base methods

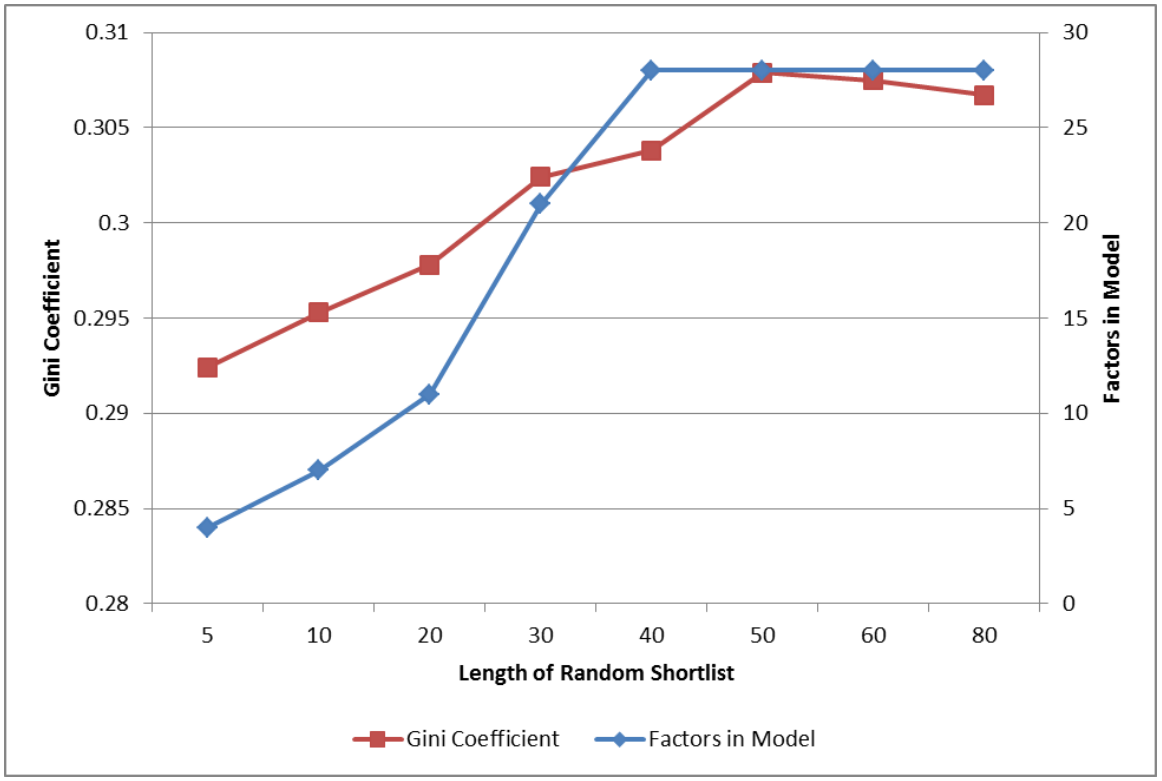
Next Steps: Broaden results

- Our conclusions were cautious because we only analyzed
  - Two perils (fire and water)
  - One line of business (personal lines- homeowners)
  - One type of variable (ordinal geo-dems)

  For other perils/lines/types of variables, the results could be very different
- A natural next step is to try the same approach on other perils/lines/types of variables

Next Steps: Variable Length Shortlists

- We analyzed nested random shortlists of various lengths
- The number of variables retained by the automatic modeling technique leveled off, as did the lift of the model



Next Steps: Variable Length Shortlists

The “leveling out” of the variable length random shortlist suggests to us two possible approaches to investigate

- Start with a random shortlist, and keep adding variables to it until the “leveling out” occurs, i.e. until no more variables “stick” to the model
- Investigate other techniques using variable length shortlists. The technique that levels out fastest can be considered the best

Next Steps: Data Structure

- Usefulness of random shortlists depends on the underlying data, in particular the lift provided by each variables, and the correlations between all variables
- In an extreme case in which all variables are perfectly correlated, a random list of any length will work as well as any other technique
- The opposite extreme is a dataset with hundreds of uncorrelated variables, only one of which provides any lift at all. Most of the variable selection techniques so far investigates should successfully find the "needle in the haystack", whereas a random shortlist would only find it by chance
- Investigating the relationship between predictiveness of variables, correlations, and the usefulness of random shortlists is a worthwhile line of future research.

Next Steps: Compound Techniques

Investigate processes such as:

- Fit a base model, using traditional techniques on a subset of variables believed to be relevant
- Employ ENetRes to create a shortlist.
- Incorporate the shortlist into the model, exploring traditional techniques such as splines, interactions, and spatial smoothing
- Employ LSRCorr, residual to the model developed in step 3, to seek out any additional variables that may have been missed in Step 2.

Appendix

WESTFIELD INSURANCE
Sharing Knowledge. Building Trust.®

TOWERS WATSON

References

1. Maitra, Saikat, and Jun Yan, “Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression”, Casualty Actuarial Society, 2008 Discussion Paper Program, http://www.casact.org/pubs/dpp/dpp08/08dpp76.pdf
2. Sanche, Robert and Kevin Lonergan, .”Variable Reduction for Predictive Modeling with Clustering” FCAS, in Casualty Actuarial Society Forum, Winter 2006 http://casualtyactuarialsociety.net/pubs/forum/06wforum/06w93.pdf
3. Kolyshkina, Dr. Inna, Sylvia Wong, and Steven Lim, “Enhancing Generalised Linear Models with Data Mining”, in Casualty Actuarial Society Discussion Paper Program, Casualty Actuarial Society - Arlington, Virginia, 2004, http://www.casact.org/pubs/dpp/dpp04/04dpp279.pdf
4. Guo, Lijia, Ph.D., “Applying Data Mining Techniques in Property/Casualty Insurance”, Casualty Actuarial Society Forum Casualty Actuarial Society, Arlington, Virginia, Winter, 2003 http://www.casact.org/pubs/forum/03wforum/03wf001.pdf
5. Nelson, Brian D., “Variable Reduction for Modeling using PROC VARCLUS, in Statistics, Data Analysis, and Data Mining”, http://www2.sas.com/proceedings/sugi26/p261-26.pdf
6. Anderson, Duncan, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, Neeza Thandi “A Practitioner’s Guide to Generalized Linear Models, Third Edition, February 2007, http://www.casact.org/pubs/dpp/dpp04/04dpp1.pdf
7. Frees, Edward W., Glenn Meyers, A. David Cummings, “Predictive Modeling of Multi-Peril Homeowners Insurance”, Volume 6, Issue 1, Casualty Actuarial Society
8. Tevet, David, “Exploring Model Lift: Is Your Model Worth Implementing?”, The CAS Actuarial Review, Volume 40, Number 2, May 2013 http://www.casact.org/newsletter/index.cfm?fa=viewart&id=6540
9. Hindawi, Mohamad PhD, FCAS “Variable Selection Using Elastic Net A Gentle Introduction to Penalized Regression” https://cas.confex.com/cas/rpms12/webprogram/Presentation/Session4724/Variable_Selection_Using_Elastic_Net.pdf