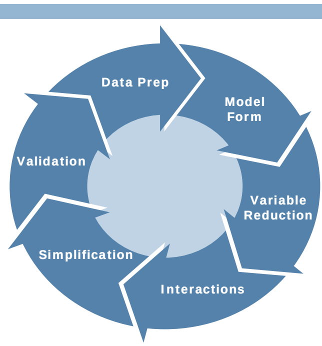
1

C-14 FINDING THE RIGHT SYNERGY FROM GLMS AND MACHINE LEARNING

CAS Annual Meeting
November 7-10

GLM Process

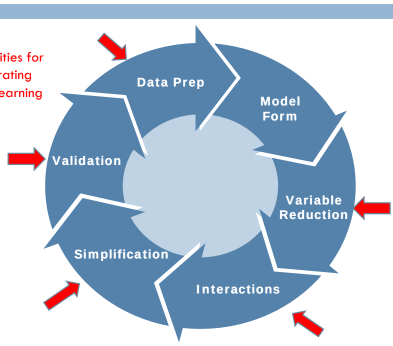
2



GLM Process

3

Opportunities for incorporating machine learning



GLM best practices
Gathering data

4

- Know the data
  - Analyze one-way cuts (exposures, frequencies, severities, loss ratios, etc.)
  - Monitor changing distributions over time
  - Identify outliers and determine whether capping or removing extreme values might be appropriate
- Clean data
  - Validate between sources
  - Consider impact of nulls
  - Account for changing level definitions (i.e. territory boundary redefinition)

Data Prep

Incorporating machine learning
Mining the data

5

Incorporate data mining to supplement data knowledge and identify adjustments.

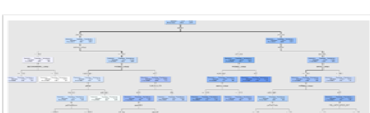
- Often requires little up-front data prep
- Provides valuable insight into your data

Data Prep

Incorporating machine learning
Using decision trees

6

Advantages	Disadvantages
• Easy to understand • Relatively quick to run • Makes no prior assumptions about the data • Able to process both numeric and categorical data	• Tree structure is unstable (not in terms of prediction) • Can be complex • Prediction is not smooth/continuous



Data Prep

Incorporating machine learning

Identifying important variables

7

Variable	Importance
Age	100
Limits	98
Prior Accidents	93
Tier	72
Vehicle Symbol	26
Prior Convictions	22
Mileage	21
Territory	19
Model Year	18
Gender	16
…	

- Data mining output usually includes variable importance, useful for:
  - Communication
  - Determining where to start GLM
  - Gaining insight into new variables

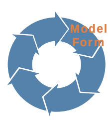
GLM best practices

Selecting the model form

8

- Goal of GLM is to identify the signal and remove the noise.

$$y = \overbrace{h(\text{Linear Combination of Variables})}^{\text{Signal}} + \overbrace{\text{Error}}^{\text{Noise}}$$



GLM best practices

Selecting the model form

9

- Error term:
  - Reflects the noise of the process.
  - Typically within the exponential family.

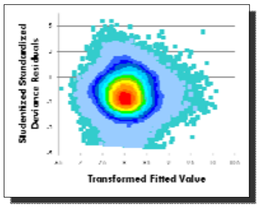
Distribution	Common Uses
Poisson	Frequency
Gamma	Severity (left skewed)
Inverse Gaussian	Severity (right skewed)
Tweedie	Pure Premium
Binomial	Response

- The link function determines how variables relate to one another.

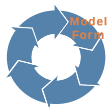
Link Function	Common Uses
Log	Multiplicative algorithm
Identity	Additive algorithm
Logit	Retention/Close Rate Studies

GLM best practices
Validating the error term

10

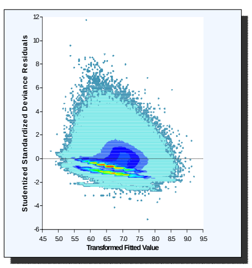


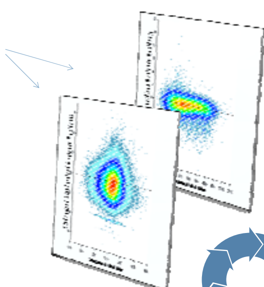
- Residuals should:
  - Be symmetric around 0
  - Have constant variance across fitted values
  - Be pre-grouped for frequency ("crunched")



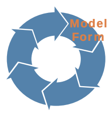
GLM best practices
Example: Modeling heterogeneous risks

11





- Bi-modal distribution can be corrected by modeling perils separately



GLM best practices
Building the GLM

12

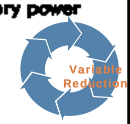
A balanced model is both refined and robust.

Under-fit

Over-fit

Most robust model:
- **Mean only**
- **No explanatory power**
- **No noise**

Most explanatory model:
- **Restating history**
- **Full explanatory power**
- **Full of noise**



GLM best practices

Example: Deciding to include/exclude variables

13

Variable	Chi-squared %	Deviance/AIC/BIC	Business judgment	Confidence intervals	Standard errors	Time consistency	…	Decision
Model Year								

Statistic	Impact
Chi-squared %	0.0%
Deviance	-995
AIC	-1026
BIC	-1019

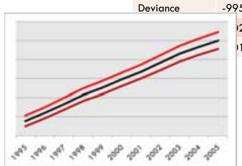
GLM best practices

Example: Deciding to include/exclude variables

14

Variable	Chi-squared %	Deviance/AIC/BIC	Business judgment	Confidence intervals	Standard errors	Time consistency	…	Decision
Model Year								

Statistic	Impact
Chi-squared %	0.0%
Deviance	-995
	26
	19



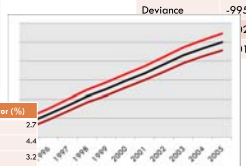
GLM best practices

Example: Deciding to include/exclude variables

15

Variable	Chi-squared %	Deviance/AIC/BIC	Business judgment	Confidence intervals	Standard errors	Time consistency	…	Decision
Model Year								

Statistic	Impact
Chi-squared %	0.0%
Deviance	-995
	26
	19

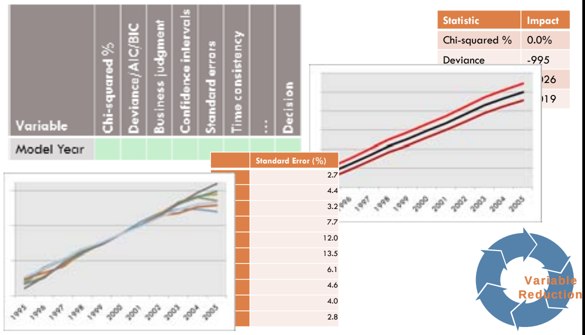


	Standard Error (%)
1995	2.7
1996	4.4
1997	3.2
1998	7.7
1999	12.0
2001	13.5
2002	6.1
2003	4.6
2004	4.0
2005	2.8

GLM best practices

Example: Deciding to include/exclude variables

16



GLM best practices

Example: Inclusion/exclusion decisions

17

Variable	Chi-squared %	Deviance/AIC/BIC	Business judgment	Confidence intervals	Standard errors	Time consistency	…	Decision
Model Year								
Vehicle Size								
Vehicle Class								
Horsepower								
…								

Incorporating machine learning

Bulk variable reduction

18

Machine learning can be particularly helpful in reducing a long list of potential variables.

- Examples:
  - Principal Components Analysis (PCA)
  - k-Means clustering
  - Decision trees
  - Pairwise correlations
  - Forward Stepwise Regression
  - …

Incorporating Machine Learning

Example: Bulk reduction of new variables

19

Initial Variables	• Hundreds of new variables
Need	• Bulk reduction of variables that may or may not be predictive
Response	• Frequency • Severity
Primary Method	• **Pairwise correlations** to identify variables that pass some minimum R^2 threshold
Secondary Method	• **Forward Stepwise Regression**
Output	• <100 variables selected for further study in GLM

Incorporating Machine Learning

Example: Reducing highly correlated variables

20

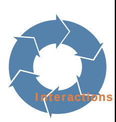
Initial Variables	• Thousands of geo-demographic variables
Need	• Bulk reduction of highly correlated variables
Primary Method	• **Principal Components Analysis** to explain the majority of the variation with a few variables
Secondary Method	• **k-Means Clustering** to pick the 'best' representative variable of the components
Output	• Few hundred unique variables that capture the majority of the variation

GLM best practices

Identifying interactions

21

- Develop a list of potential interactions:
  - Brainstorm with business partners
  - Use filed rating manuals to investigate what the competition is doing
  - Study most predictive variables, especially with a wide range of predicted values
- Guess and check!

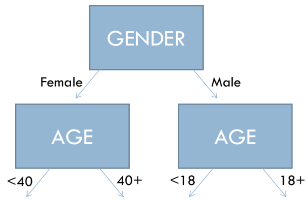


Incorporating machine learning

Example: Identifying interactions with trees

22

Decision trees can be used to identify potential interactions.



Interactions

GLM best practices

Simplifying variables and interactions

23

- A good GLM describes the signal with as few parameters as possible.
- Reduce the number of parameters by fitting curves to continuous variables and logically grouping categorical variables.
- Diagnostics help validate simplification decisions.

Simplification

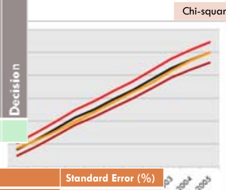
GLM best practices

Example: Simplifying variables

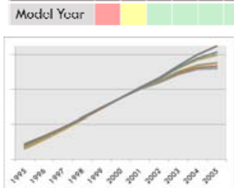
24

Variable	Chi-squared %	Deviance/AIC/BIC	Business judgment	Confidence intervals	Standard errors	Time consistency	…	Decision
Model Year								

Statistic	Impact
Chi-squared %	0.0%
ance	-132
AIC	-109
BIC	+18



	Standard Error (%)
1st Degree	1.7
2nd Degree	10.2



Simplification

Incorporating machine learning
Variable Simplification

25

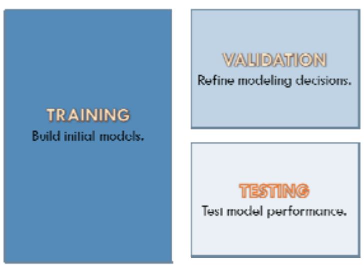
Machine learning can help determine how best to simplify the GLM.

- Examples:
  - Identify potential binning of categorical variables
  - Test whether groupings or curves are more appropriate for continuous variables
  - How to handle nulls

GLM best practices
Validating with hold-out samples

26

- Split data for modeling and validation.



GLM best practices
Starting with a solid model

27

- Employ GLM best practice techniques.
- Be prepared to make multiple iterations through your model.

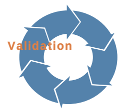
GLM best practices

Validating modeling decisions

28

VALIDATION
Refine modeling decisions.

- Fit the **model structure** developed to a new set of data.
- Compare predicted values between validation and training.
- Refine and validate model decisions.



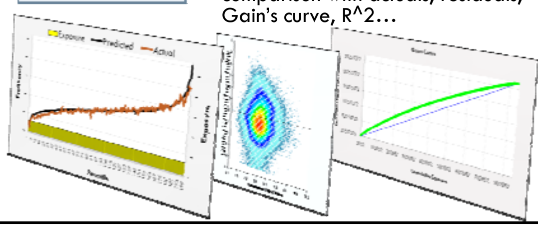
GLM best practices

Validating model performance

29

TESTING
Test model performance.

- Test **model prediction** on a new set of data.
- Consider an out-of-time hold-out.
- Validate the model through a comparison with actuals, residuals, Gain's curve, R^2…



Incorporating machine learning

Validating against neural net

30

- Compare predictions from a well-built GLM to a neural network.

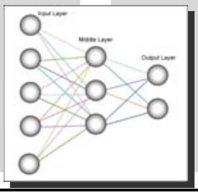
- The neural net should offer some additional lift at the cost of transparency.
- If lift is significant, consider further complication of GLMs.

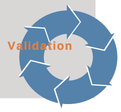
Incorporating machine learning
Using neural networks

31

Advantages
- Very flexible
- Will pick up important interactions even if not specified
- Can model complex non-linear relations
- Should always give equal or better model fit than a GLM

Disadvantages
- Can easily be over-fit
- Black box, difficult to decipher useful info.





Incorporating machine learning
Example: Validating GLM vs. neural net

- Used neural net to validate performance of a less complex GLM.
- The comparison resulted in the following conclusions:
  - The less complicated model held up well.
  - The slight improvement in predictive power achieved using a neural net was not enough to justify further complicating the model.

Validation Statistics	Additional lift using neural net
Log Likelihood	2.7%
Classification rate	1.3%

Summary

33