TRAVELERS

Model Validation Techniques

Kevin Mahoney, FCAS
kmahoney@ travelers.com

CAS RPM Seminar

March 17, 2010

Uses of Statistical Models in P/C Insurance

Examples of Applications

- Determine expected loss cost for an account (by line-of-business, peril, etc.)
- Determine likelihood to defect for an account
- Determine effectiveness of advertising
- Identify "ripe" targets for cross-sell attempts
- Triage for further treatment (risk engineering, inspection, etc.)
- Identify claims that may be fraud
- Identify claims that need experienced adjusters

Examples of Techniques

- Generalized Linear Models
- Generalized Additive Models
- Cox Regression
- Decision Trees
- Ensemble Methods
- Text Mining
- Spatial Models
- Mixed Models
- Neural Networks

TRAVELERS

What Models Need to Be Validated?

- **All** models need to be validated
- However, unlike many other statistical diagnostics,
  - **THE SAME CONCEPTS APPLY REGARDLESS OF THE TYPE OF MODEL**
- You can apply the concepts you learn today to any of the above, plus just about any other type of predictive model you may encounter

TRAVELERS

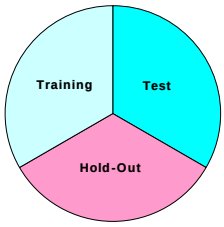
The Many Meanings of Model Validation

- **Primary Meaning / Use—Quantifying Model Performance**
  - **How well can we expect this model to perform in the future**
  - **The only objective test is unseen data**
- Secondary Meanings/Uses—Using Similar Procedures for Other Goals
  - Looking at out-of-sample data during the modeling process to determine:
    - the "right" choice of predictor variables [feature selection], and/or
    - the "right" type of model, and/or
    - The "right" value of a tuning parameter

TRAVELERS

Important Caution

- The same out-of-sample data cannot serve both purposes above
  - If used in feature selection, then it influenced the model
  - So need additional out-of-sample data to quantify performance



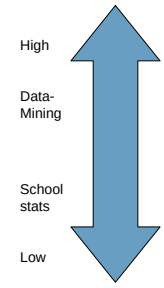
TRAVELERS

Important Caution

- Data terminology
  - Data used in building the model (in-sample) are "training" data
  - Out-of-sample data used in guiding the modeling process are "test" data
  - Out-of-sample data against which the predictive power of the ultimately chosen model is tested are "validation" data or "holdout" data. It is sometimes important for this data to be out-of-time as well.
    - E.g., if you are modeling severities of homeowners losses, you don't want claims from the same storms in the training/test data and the validation data

TRAVELERS

Danger Scale



High

Data-Mining

School stats

Low

Type of Search

Lookup Table

Neural Network

Adaptive methods: MARS, GAMS, splines

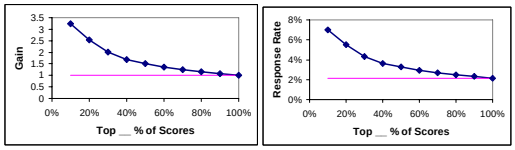
Automated feature searches (stepwise)

Search "by hand"

No search

TRAVELERS

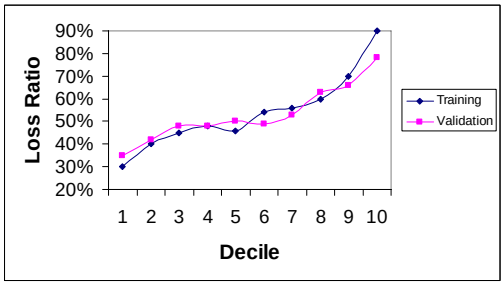
Gains Charts for Direct Marketing Model



- Classic chart to diagnose a response model
- Use the model to "score" the validation data
  - I.e., assign the modeled probability of a response
- SORT by the modeled probability from most likely to least likely
- Plot the actual response rate in the validation among the highest scoring x%
- Typically do NOT even plot the training data…goal is purely to show objective performance

TRAVELERS

Decile Charts



TRAVELERS

Decile Charts

- Again, you sort by predicted value
- You show actual value
- The validation line is key
  - This shows the actual predictive power of the model
- The discrepancy between the validation and actual lines is useful
  - In modeling (using test rather than final validation data), to diagnose overfit
  - In implementation: If implementing as a rating algorithm, discrepancies between the training line and validation line suggest "shrinking" extreme estimates
- Nothing "magical" about deciles: Use quintiles, vingtiles, whatever your data will support

TRAVELERS

The Many Meanings of Model Validation

- Primary Meaning—Quantifying Model Performance
  - How well can we expect this model to perform in the future
  - The only objective test is unseen data
- **Secondary Meanings—Using Similar Procedures for Other Goals**
  - **Looking at out-of-sample data during the modeling process to determine:**
    - **the "right" choice of predictor variables [feature selection], and/or**
    - **the "right" type of model, and/or**
    - **The "right" value of a tuning parameter**

TRAVELERS

Choice of Predictors

- Do NOT use validation data for this
  - Just training and test
- Divide dataset into training and test data
- Check that predictors still show up as significant if you model the test data
- Or divide the training data into many pieces
  - Say 5 pieces
  - Model each 1/5 (or each 4/5)
  - Only include predictors that were significant in at least 2 (or 3, or 4) of these 5 models

TRAVELERS

The Right-sized Model

- Why not use only seen data but penalize the goodness-of-fit measure for the number of parameters and/or degrees of freedom?
  - The "information criteria", AIC, BIC, etc., do this
  - Limitations:
    - The number of parameters may be the wrong basis for the penalty
      - E.g., if using shrinkage techniques, like ridge regression, or credibility, or hierarchical or mixed models, the effective dfs may be much smaller than the number of parameters
      - Even if you have a good way to compute the effective degrees of freedom, that doesn't penalize for the size of the search…

        If you have 20 features, the "best" 8 feature model implies a search of 125,970 models; "an" 8 feature model implies a search of *1* model.

TRAVELERS

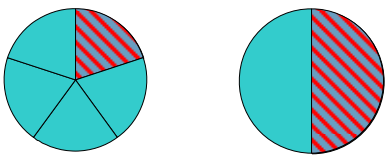
The Right-sized Model

- Why not use significance tests to decide whether to include a variable:
  - Yes, but. . .
  - Raw significance tests also do nothing to adjust for the size of the search
  - Tests that are directionally correct may not function correctly in absolute terms when modeling assumptions are violated (i.e., always)
    - For example, widths of confidence intervals are very sensitive to the scale parameter in most GLMs
    - But the scale parameter has to be estimated from the data and may not itself be very certain

TRAVELERS

Cross-Validation

- Divide the data into N pieces
  - N=5 or 10 typical; N=2 convenient if hurried



TRAVELERS

Cross-Validation

- Run the model on each 4/5 or each 9/10
  - This results in N models, each on a high percentage of the data
  - Each datapoint has been left out in building exactly 1 model
  - Compare each actual observed value to the value predicted by the model that didn't see it
    - Use this to compute goodness-of-fit (squared error, misclassification rate, etc.)
    - Use this to compare models of varying complexity
      - Fewer or more predictors
      - Different values of a tuning parameter (e.g., K in a Bayesian credibility setup)

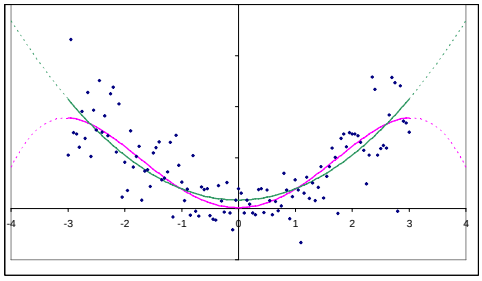
TRAVELERS

Cross-Validation

- In the data mining and machine learning community, often used to do the objective validation of model power
  - Only works because the model-building process is entirely automated
    - Each 9/10 model and the model on the entire dataset are built without knowledge of the other 10 models
      - Not just the fitting of parameters is independent
      - So is the choice of variables, indeed the entire process
      - If the process was open to alternate feature-selection methods (e.g., CART or MARS) before looking at the first 9/10, technically even that decision must be remade 10x
    - This is not possible when human beings are part of the modeling process

TRAVELERS

$A+Bx^2$ or $A+Bx^2+Cx^4$?



TRAVELERS

Parameter Estimates

	Estimate	Std Dev	p-value
$A+Bx^2$			
A	0.68	0.28	0.015
B	0.88	0.07	2.26E-37
$A+Bx^2+Cx^4$			
A	0.09	0.34	0.803
B	1.53	0.23	5.08E-11
C	-0.083	0.029	0.003

So you need an x^4 term, right?

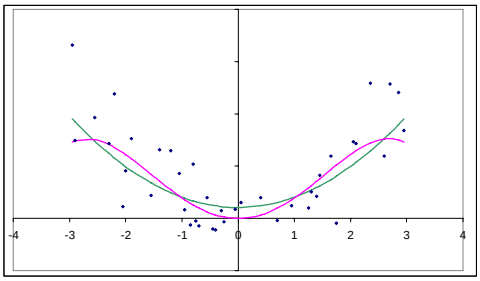
TRAVELERS

Hypothesis Test is OK?

- Or do you need an x^4 term?
- Do the errors look identically distributed?
  - Or are the data heteroskedastic?
- So assumptions are violated that may severely impact the hypothesis tests
- Let's look at another comparison of the two models:

TRAVELERS

1/3 of the data vs models fit on other 2/3



This is done 3 times, once for each 1/3

TRAVELERS

Cross-Validation to the Rescue

Sum of Squared Errors

	$A+Bx^2$	$A+Bx^2+Cx^4$
Full Model	505.6	471.3
3-Fold Cross-Validation	564.2	565.6

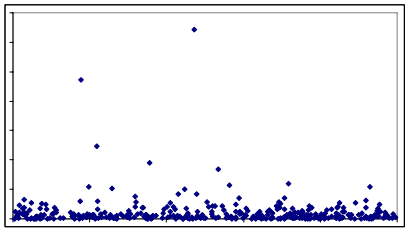
Mean Squared Error

	$A+Bx^2$	$A+Bx^2+Cx^4$
Full Model	4.18	3.90
3-Fold Cross-Validation	4.66	4.67

- Note the optimism of the full model (in-sample) errors
- Note that the 4th power term is completely unnecessary
- Data generated from a quadratic (with added heteroskedastic errors)

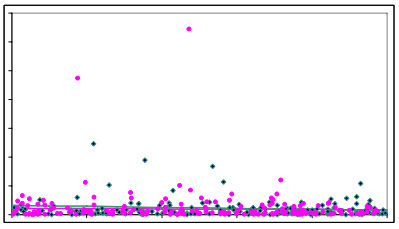
TRAVELERS

Simplified Version of Property-Casualty Insurance Data



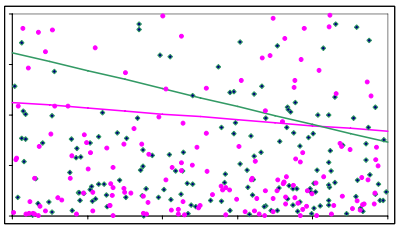
TRAVELERS

Split into two and added lines



TRAVELERS

Zooming on the lines



TRAVELERS

Lesson of the above

- Insurance data often has a few observations that are outliers
  - But we can't throw them out because they are the observations that matter most
- Therefore:
  - Choose your model carefully (linear regression without transforming y should be an obviously bad idea with the above)
  - Remember that you don't have as much information as the size of your dataset might indicate
  - Remember that you can overcome optimism in classical confidence intervals using cross-validation

TRAVELERS

Testing on Seen vs Unseen Data

In-sample tests	Out-of-sample tests
•Must adjust for "degrees of freedom"	•No need to adjust for degrees of freedom
•Many tests oriented toward inferential power	•Tests typically oriented toward predictive power
•Tests sensitive to fussy statistical assumptions	•Tests purely empirical; only simple assumptions involved
•May need deep statistical knowledge to interpret	•Have commonsense interpretations
•Difficult to present results to management	•With modest effort, usually presentable
•May require adjustments if observations are correlated	•In some cases, may need to be an out-of-time as well as out-of-sample test

TRAVELERS

Model Validation Today

- Model validation is a serious topic
- Regulators require some financial institutions to have a separate department that validates, for example, consumer creditworthiness models
- Should there be an actuarial standard of practice addressing validation of statistical models
  - Topics such a standard might address
    - When is out-of-time validation rather than just out-of-sample validation critical?
    - What steps should be taken to ensure knowledge of the validation data has not crept into the model-building process?
      - For instance, split off the validation data before or after EDA?
      - Splitting it too early makes balancing to control-totals difficult

TRAVELERS

Q&A

TRAVELERS

TRAVELERS