

Data Mining

Bob Stine
Department of Statistics

www-stat.wharton.upenn.edu/~bob

Overview

- Applications
 - Marketing: Direct mail advertising (Zahavi example)
 - Biomedical: finding predictive risk factors
 - Financial: predicting returns and bankruptcy
- Role of management
 - Setting goals
 - Coordinating players
- Critical stages of modeling process
 - Picking the model <-- My research interest
 - Validation

2

Predicting Health Risk

- Who is at risk for a disease?
 - Costs
 - False positive: treat a healthy person
 - False negative: miss a person with the disease
 - Example: detect osteoporosis without need for x-ray
- What sort of predictors, at what cost?
 - Very expensive: Laboratory measurements, "genetic"
 - Expensive: Doctor reported clinical observations
 - Cheap: Self-reported behavior
- Missing data
 - Always present
 - Are records with missing data like those that are not missing?

3

Predicting Stock Market Returns

- Predicting returns on the S&P 500 index
 - Extrapolate recent history
 - Exogenous factors
- What would distinguish a good model?
 - Highly statistically significant predictors
 - Reproduces pattern in observed history
 - Extrapolate better than guessing, hunches
- Validation
 - Test of the model yields sobering insight

4

Predicting the Market

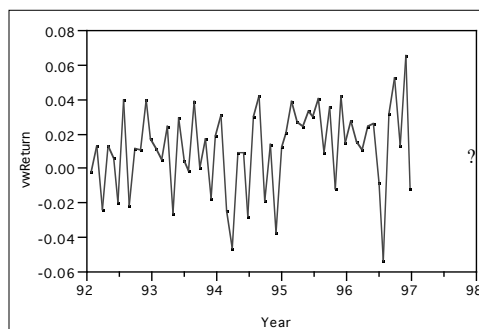
Wharton
Department of Statistics

- Build a regression model
 - Response is return on the value-weighted S&P
 - Use standard forward/backward stepwise
 - Battery of 12 predictors
- Train the model during 1992-1996
 - Model captures most of variation in 5 years of returns
 - Retain only the most significant features (Bonferroni)
- Predict what happens in 1997
- Another version in Foster, Stine & Waterman

5

Historical patterns?

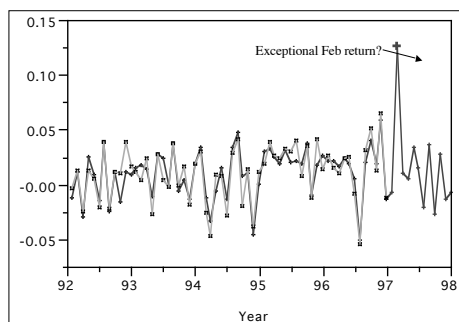
Wharton
Department of Statistics



6

Fitted model predicts...

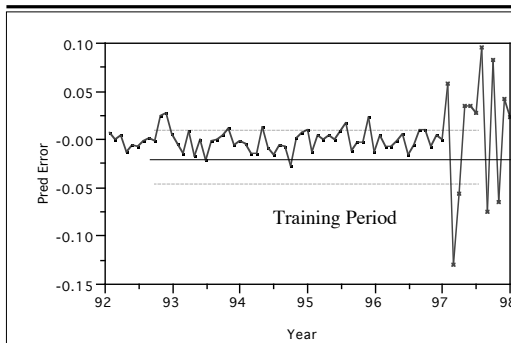
Wharton
Department of Statistics



7

What happened?

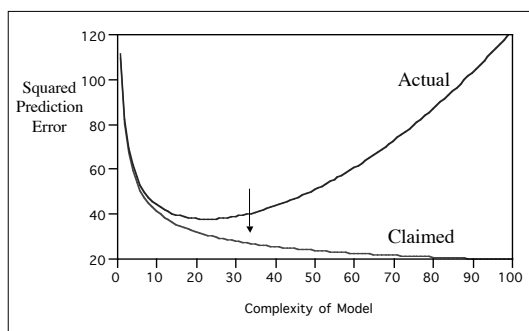
Wharton
Department of Statistics



8

Claimed versus Actual Error

Wharton
Department of Statistics



9

Over-confidence?

Wharton
Department of Statistics

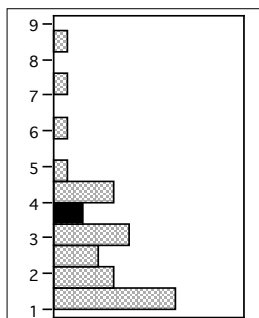
- Over-fitting
 - DM model fits the training data too well – better than it can predict when extrapolated to future.
 - Greedy model-fitting procedure
 - “Optimization capitalizes on chance”
- Some intuition for the phenomenon
 - Coincidences
 - Cancer clusters, the “birthday problem”
 - Illustration with an auction
 - What is the value of the coins in this jar?

10

Auctions and Over-fitting

Wharton
Department of Statistics

- Auction jar of coins to a class of students
- Histogram shows the bids of 30 students
- Some were suspicious, but a few were not!
- Actual value is \$3.85
- Known as “*Winner’s Curse*”
- Similar to over-fitting: best model like high bidder



11

Roles of Management

Wharton
Department of Statistics

Management determines whether a project succeeds...

- Whose data is it?
 - Ownership and shared obligations/rewards
- Irrational expectations
 - Budgeting credit: “How could you miss?”
- Moving targets
 - Energy policy: “You’ve got the old model.”
- Lack of honest verification
 - Stock example... Given time, can always find a good fit.
 - Rx marketing: “They did well on this question.”

12

What are the costs?

Wharton
Department of Statistics

- Symmetry of mistakes?
 - Is over-predicting as costly as under-predicting?
 - Managing inventories and sales
 - Visible costs versus hidden costs
- Does a false positive = a false negative?
 - Classification
 - Credit modeling, flagging “risky” customers
 - Differential costs for different types of errors
 - False positive: call a good customer “bad”
 - False negative: fail to identify a “bad”

13

Back to a real application...

Wharton
Department of Statistics

How can we avoid some of these problems?

I'll focus on

- * statistical modeling aspects (my research interest), and also
- * reinforce the business environment.

14

Predicting Bankruptcy

Wharton
Department of Statistics

- “Needle in a haystack”
 - 3,000,000 months of credit-card activity
 - 2244 bankruptcies
 - Best customers resemble worst customers
- What factors anticipate bankruptcy?
 - Spending patterns? Payment history?
 - Demographics? Missing data?
 - Combinations of factors?
 - Cash Advance + Las Vegas = Problem
- We consider more than 100,000 predictors!

15

Stages in Modeling

Wharton
Department of Statistics

- *Having framed the problem, gotten relevant data...*
- *Build the model*
Identify patterns that predict future observations.
- *Evaluate the model*
When can you tell if its going to succeed...
 - During the model construction phase
 - Only incorporate meaningful features
 - After the model is built
 - Validate by predicting new observations

16

Building a Predictive Model

Wharton
Department of Statistics

So many choices...

- *Structure*: What type of model?
 - Neural net (projection pursuit)
 - CART, classification tree
 - Additive model or regression spline (MARS)
- *Identification*: Which features to use?
 - Time lags, “natural” transformations
 - Combinations of other features
- *Search*: How does one find these features?
 - Brute force has become cheap.

17

My Choices

Wharton
Department of Statistics

- Simple structure
 - Linear regression with nonlinear via interactions
 - All 2-way and many 3-way, 4-way interactions
- Rigorous identification
 - Conservative standard error
 - Comparison of conservative t-ratio to adaptive threshold
- Greedy search
 - Forward stepwise regression
 - Coming: Dynamically changing list of features
 - Good choice affects where you search next.

18

Bankruptcy Model: Construction

Wharton
Department of Statistics

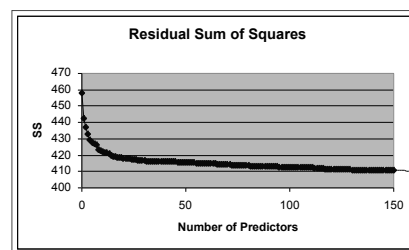
- Context
 - Identify current customers who might declare bankruptcy
- Split data to allow validation, comparison
 - Training data
 - 600,000 months with 450 bankruptcies
 - Validation data
 - 2,400,000 months with 1786 bankruptcies
- Selection via *adaptive thresholding*
 - Analogy: Compare sequence of t-stats to $\sqrt{2 \log p/q}$
 - Dynamic expansion of feature space

19

Bankruptcy Model: Fitting

Wharton
Department of Statistics

- Where should the fitting process be stopped?

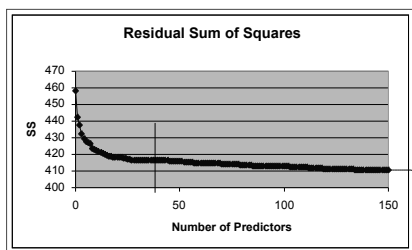


20

Bankruptcy Model: Fitting

Wharton
Department of Statistics

- Our adaptive selection procedure stops at a model with 39 predictors.

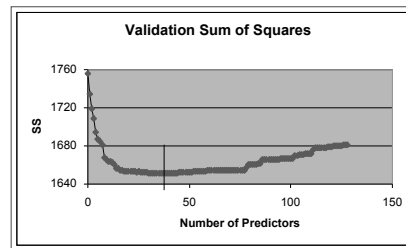


21

Bankruptcy Model: Validation

Wharton
Department of Statistics

- The validation indicates that the fit gets better while the model expands. Avoids over-fitting.



22

Lift Chart

Wharton
Department of Statistics

- Measures how well model classifies sought-for group

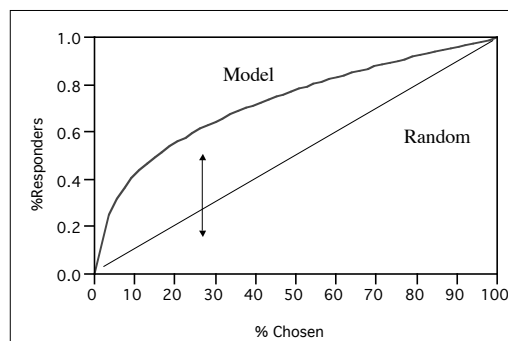
$$Lift = \frac{\% \text{ bankrupt in DM selection}}{\% \text{ bankrupt in all data}}$$

- Depends on rule used to label customers
 - Very high probability of bankruptcy
Lots of lift, but few bankrupt customers are found.
 - Lower rule
Lift drops, but finds more bankrupt customers.
- Tie to the economics of the problem
 - Slope gives you the trade-off point

23

Example: Lift Chart

Wharton
Department of Statistics

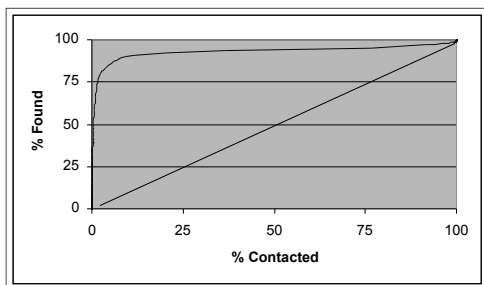


24

Bankruptcy Model: Lift

Wharton
Department of Statistics

- Much better than diagonal!

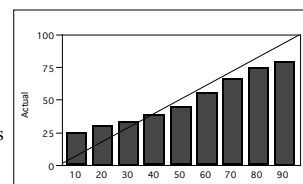


25

Calibration

Wharton
Department of Statistics

- Classifier assigns $\text{Prob}(\text{"BR"})$ rating to a customer.
- Weather forecast
- Among those classified as 2/10 chance of "BR", how many are BR?
- Closer to diagonal is better.

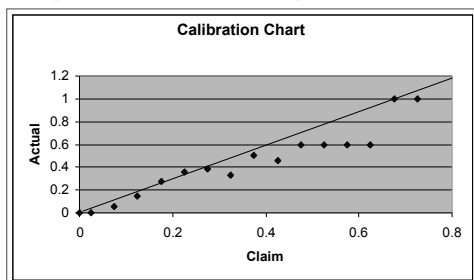


26

Bankruptcy Model: Calibration

Wharton
Department of Statistics

- Over-predicts risk near claimed probability 0.3.



27

Modeling Bankruptcy

Wharton
Department of Statistics

- Automatic, adaptive selection
 - Finds patterns that predict new observations
 - Predictive, but not easy to explain
- Dynamic feature set
 - Current research
 - Information theory allows changing search space
 - Finds more structure than direct search could find
- Validation
 - Remains essential only for judging fit, reserve more for modeling
 - Comparison to rival technology (we compared to C4.5)

28

Wrap-Up Data Mining

Wharton
Department of Statistics

- ♦ Data, data, data
 - Often most time consuming steps
 - Cleaning and merging data
 - Without relevant, timely data, no chance for success.
- ♦ Clear objective
 - Identified in advance
 - Checked along the way, with “honest” methods
- ♦ Rewards
 - Who benefits from success?
 - Who suffers if it fails?

29