Data Mining

A regression modeler's view
on where it is and where it's likely to go.

Bob Stine
Department of Statistics
The Wharton School of the Univ of Pennsylvania
March 30, 2006

Acknowledgments

- Colleagues
  - Dean Foster in Statistics
  - Lyle Ungar in Computer Science
- Support from Wharton Financial Institutions Center
- Cooperation of Federal Reserve Bank of Philadelphia

Overview

- Some examples of data mining
  - More detail on some than others
- Methods used in data mining
  - Lots of choices!
- Challenges faced in data mining
  - Common to all methods, old and new
- Directions

Examples

- Finance
  Can I predict the stock market?
  Which loans are most likely to default?
- Management
  Which applicants to hire and train?
- Health
  Who is at greater risk of a disease?
- Images
  Is there a face in this image?

Lots of Data

- Once upon a time...
  - A large data set had 50 to 100 rows and perhaps 5 to 10 columns.
  - A big multiple regression had 4 or 5 predictors
- That's changed...
  - Modern data sets are immense, with thousands to millions of rows and hundreds to thousands of columns.
  - The models have grown as well

Lots of Data

- Credit
  - Millions of credit card users
  - History, economics, transactions
- Hiring
  - Several thousand past employees
  - Numerous application characteristics
- Health
  - Thousands of patient records at one hospital
  - Genetic markers, physician reports, tests
- Images
  - Millions of images from video surveillance
  - All those pixel patterns

Similar Goals

- Numerous, repeated decisions with asymmetric costs attached to mistakes.
- Hiring
  - Firm trains 250 new employees monthly
  - Which are the best candidates (need to rate them, then pick the best)
  - Miss a good candidate: Lose sales for the firm (≈ $100,000/month)
  - Train a poor candidate: Wasted the seat and the $10,000 training fee

Similar Goals

- Numerous, repeated decisions with asymmetric costs attached to mistakes.
- Credit
  - Manage thousands of accounts in each line
  - Which accounts are going bad?
  - Miss a bad account: Defaults typically on the order of $10,000 to $30,000
  - Annoy a good customer: Might lose that customer and the 18% interest you're earning.

Similar Use of Models

- Predictive models
  - Better predictions mean a competitive advantage
  - Classification
  - Prediction
- But you sacrifice interpretation...
  - Realize that the model is not causal.
  - Collinearity among features makes interpretation of the model a risky venture.
- Lure of finding cause and effect

Similar Problems, Too

- Rare events
  Relatively few "valuable" decisions in the mix, buried among the more common cases.
- Numerous explanatory features
  Often have more ways to explain the event than cases to check them (ie, more columns that rows in data)
- Plus familiar complications
  Missing data, dependence, measurement error, changing definitions, outliers...

Wide Data Sets

Application	Rows	Columns
Credit	3,000,000	350
Faces	10,000	1,400
Genetics	1,000	10,000
CiteSeer	500	∞

Choices in Modeling

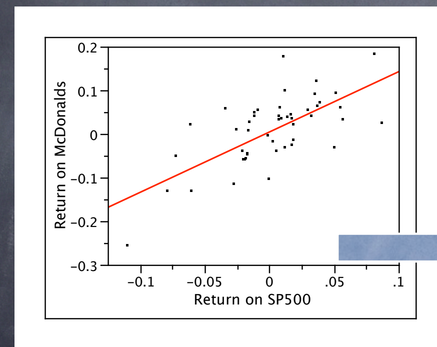
- Structure of the model
  - Regression $Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots$
  - Projection pursuit $Y = c_0 + c_1 D(X_1, X_2, \ldots) + \ldots$
  - Trees $Y = \text{if}(X_1 < a)$ then ...
- Scope of the search
  - Raw features, observed measurements
  - Combinations of features, interactions
  - Transformation of features
- Selection
  - Which features to use?

Hands-on Example

- Small model for pricing stocks suggests most of the key issues
- Context
  - Theory in Finance known as the Capital Asset Pricing Model says that only one predictor explains returns on a stock...
  - namely returns on the whole market.
- Day traders **know** this is wrong!
  - Devise "technical trading rules" based on turning points, patterns in recent history

Wharton
Statistics Department

CAPM Relationship



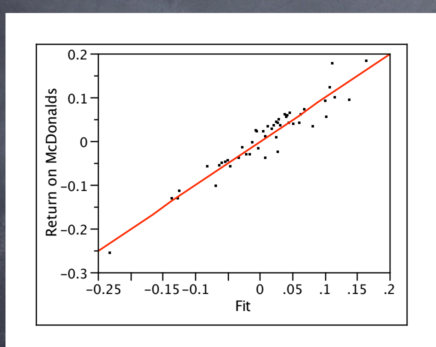
Fit = 0.006 + 1.4 S&P

- Returns on McDonalds vs Returns on S&P 500
- 48 months,2002-2005
- Slope is called "beta" of the stock
- R^2 = 46.5%
- t-stat for slope is 6.3

We can do better than that!

Wharton
Statistics Department

A Better Model



Fit = 0.017 + 0.7 S&P +...

- Add 16 features that implement variety of technical trading rules.
- Doubled R^2 to 91%
- Overall F = 17.8
- "Beta" about half prior size
- t-statistic for slope still impressively large (t = 4.9)
- Seven other predictors have p-values less than 0.0001.

Wharton
Statistics Department

Other Features

Term	Est	|t|	p
SP500	0.7	4.9	0
X_{22}	0.2	3.7	.0009
X_{34}	0.4	5.8	0
X_{36}	0.3	5.0	0
X_{37}	-.4	7.8	0
X_{39}	0.3	6.3	0
X_{44}	0.3	4.2	.0003
X_{46}	-.4	6.5	0

- Seven additional predictors add significant variation to the model
- Many have larger t-statistics than the SP500 index
- Model looks great from variety of perspectives.
- Statistician says "great model"

What are these other predictors?

Wharton
Statistics Department

Better Mousetrap?

- Added predictors are random noise!
- So why do they look so good?
  - Selection bias
  - Pick variables to add from suite of 50 columns of random noise.
    - Forward stepwise regression
    - Greedy search adds most significant next predictor to the current model
      "Optimization capitalizes on chance"
- Result
  Biased estimate of noise variance inflates t-stat and produces "cascade" of features

Consequences

- Expanding the model
  - Claims better structure, higher accuracy
  - Replaces $\beta > 1$ to $\beta < 1$.
- But in reality the expanded model is junk...
  - Adding random predictors ruins predictions
  - Conveys wrong impression of the role of the market on the returns of this stock
- Stepwise regression... Evil?

Feature Selection

- Don't blame stepwise for these problems
- Failure: uncontrolled modeling process
  - The final model looks great on paper, if you don't know how the predictors were chosen.
  - Cannot wait "until the end" and use classical methods to evaluate a model
- Flaws in this example happen elsewhere
  - Automatic methods expand the scope of the search for structure to wider spaces

Easy to Fix

- Once you recognize the problem, it is relatively easy to control the modeling
  - Must keep random features out of model
- Cross-validation
  - Use a "hold-back" or "test" sample to evaluate the model.
  - Painful to give up data when you don't have many cases ($n = 48$ here, or in genetics)
- Bonferroni methods
  - Use **all** data to fit and evaluate model

Second Example

- Classification problem
  Identify onset of personal bankruptcy
- Illustrate
  - Scope of data and size of models
  - Control greedy modeling process without using cross validation
  - Save validation data to show that "it works" rather than to pick the model itself
- Make a claim about regression

Building a Predictive Model

- Claim
  Regression is competitive with other types of predictive models
- Keys
  - Expand the scope of features
    - Interactions: subsets, nonlinearity
    - Missing data treated as interaction
  - Cautious control of selection of features
    - Avoid bias in noise variance
    - Don't trust CLT to produce accurate p-value

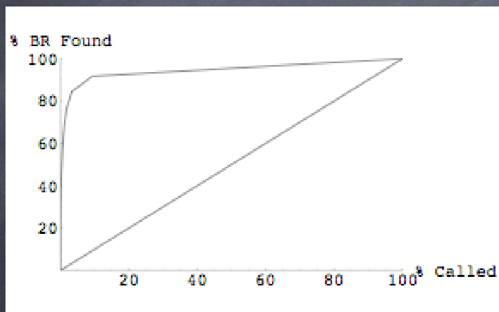
Goals for Model

- Goal
  Reduce loss from bankrupt accounts without irritating profitable customers
- Ideal customer
  Borrow lots of money, pay back slowly
- Business strategy: triage
  Contact customers who are "at risk" and keep them paying

Data

- Rows
  - 3,000,000 months of activity
  - 2200 bankruptcies
- Columns
  - 350 basic features
    - Credit application
    - Location demographics
    - Past use of credit
  - Missing data indicators
  - Add **all** interactions... 66,430 more predictors

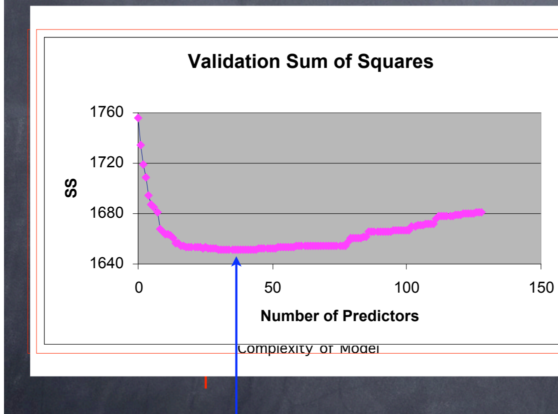
Results



- Use cross-validation to evaluate the model
- Fit on 600,000, and then classify the other 2.4 million
- Lift chart displays ordering of cases compared to random selection
- If call 1,000, find 400 bankrupt cases.
- Triage becomes economically viable

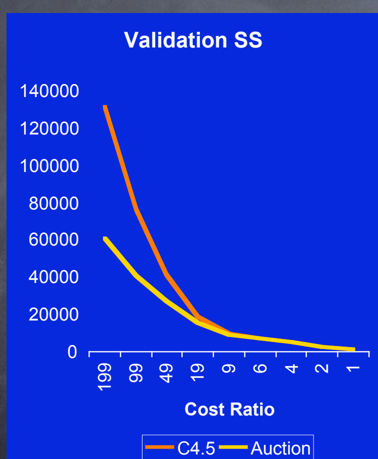
Every added variable improved the results!

Controlling Selection



- Where to stop the addition of variables?
- Over-fitting occurs when the model begins to add random features that are predictive in-sample
- Our method stopped after adding 39 predictors
- Avoids over-fitting: Error increases if the model is expanded further.

Comparison to Tree



- Always good to have a benchmark
- C4.5 is a commercial classifier that builds trees
- Cost ratio is the ratio of the cost of missing a bankrupt customer to the cost of annoying a good customer.
- Regardless of the ratio of costs, regression achieves lower costs

How does it work?

- Basically stepwise regression
  - Caution: Dont' try this with standard SAS/R
- Three ingredients
  1. Rearrange order of computing
  2. Hard thresholding rule
     - Compare p-value to ≈ 1/67000
     - AIC would let in about 16% of all features!
  3. Cautious standard error
     - Use residuals from fit without predictor
     - Allow for Poisson-like variation (Bennett)even though n is large (recall spare nature of data)

Conclude from Example

- Regression is competitive with other methodologies for data mining... if you adapt it to the context
  - Ability to study residuals and other diagnostics facilitated improvements
- Details
  - Other adjustments include calibration
  - Foster and Stine, 2004, JASA
  - Portions of data are available from Dean's web page

Challenges

Lots of room for improvement!

Challenges

- "That's the way we used to work"
  - Population drift, moving target
  - Model in business changes the population
    - Credit: effective screening removes features
    - Hiring: model changed data collection
  - Cross-validation is optimistic!
    - In CV, you truly predict new observations from the same population
- How to fix this one?
  - Can you detect this problem?

Challenges

- "Simple models are better"
  - Often find that complex models offer little that not found with simpler model (Hand, 2006, forthcoming Stat Science)
  - Not our experience: Linear models do not find predictive structure in BR application, fare poorly compared to trees
- Still suggests room to improve...
  - Yuk: All but one predictor is an interaction
  - A different type of search finds linear terms

Challenges

- "You missed some things"
  - Knowledgeable modelers with years of experience can suggest features that improve the model
  - Simple feature space omits special features that use domain-specific transformations
- Can do better...
  - Alternative methods allow additional expert input and do find richer structure

Challenges

- There's a lot more data!
  - Transaction information in the credit model
    - We only used total spending and payments, not the nature of what was being bought
  - Semi-supervised modeling
    - Billions of "unmarked" cases: images, text
    - Too expensive to mark them all
- Room to improve...
  - How to use the vast number of unmarked cases to improve the modeling of those that have been classified or scored?

Overcoming Challenges

- Still building regression models
- Problems
  - Population drift
  - ✓ Better mix of simple features
  - ✓ Incorporate expert guidance
  - ✓ Explore richer spaces of predictors
  - ✓ Run faster
- Come back tomorrow!