Feature Selection in Models For Data Mining

Robert Stine

Statistics Department

The Wharton School, Univ of Pennsylvania

January, 2005

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

Questions Asked by Data Miners

- Anticipate bankruptcy
  - Which borrowers are most likely to default?
- Adverse effects
  - Are patients at risk of adverse side effects from medication?
- Facial recognition
  - How can we train computers to find faces in images?
- Other domains…
  - Employee evaluation: Who should we hire?
  - Genomics: Which genes indicate risk of a disease?
  - Document classification: Which papers resemble this one?

Common Answer is Prediction

- Regardless of the context
  - Anticipating default on loan
  - Identifying presence of unexpected side effect
  - Deciding if there's a face in an image
- Want the model with the best predictions
  - Best prediction = smallest costs
- Desire for accuracy motivates numerous methods
  - Equations: regression, logistic regression
  - Combined equations: graphical models, neural networks
  - Trees
  - Clustering, nearest neighbor

Similar Issues to Overcome

Wharton
Statistics Department

- Rare events
  - Few cases frequently dominate costs
  - Lots of images, but few faces most of the time
  - Numerous credit cards, few that will default
- Wide data sets: more features than cases
  - Cheaper to get measurements than cases
  - Categorical data, networks, missing data…
- Synergies add further possibilities
  - Long lists of database features, none predictive
  - Combinations are predictive, but so many.

Data’s getting obese!

Wharton
Statistics Department

Application	*Number of Cases*	*Number of Raw Features*
Bankruptcy	3,000,000	350
Faces	10,000	1,400
Genetics	1,000	10,000
CiteSeer	500	∞

Key Challenge for Modeling

Which features belong in the model?

- Regardless of the modeling technology, how do you decide which features to add to the model.
- Add the right features, and you get better predictions.
- Add the wrong features, and you think you've done well but only fooled yourself.

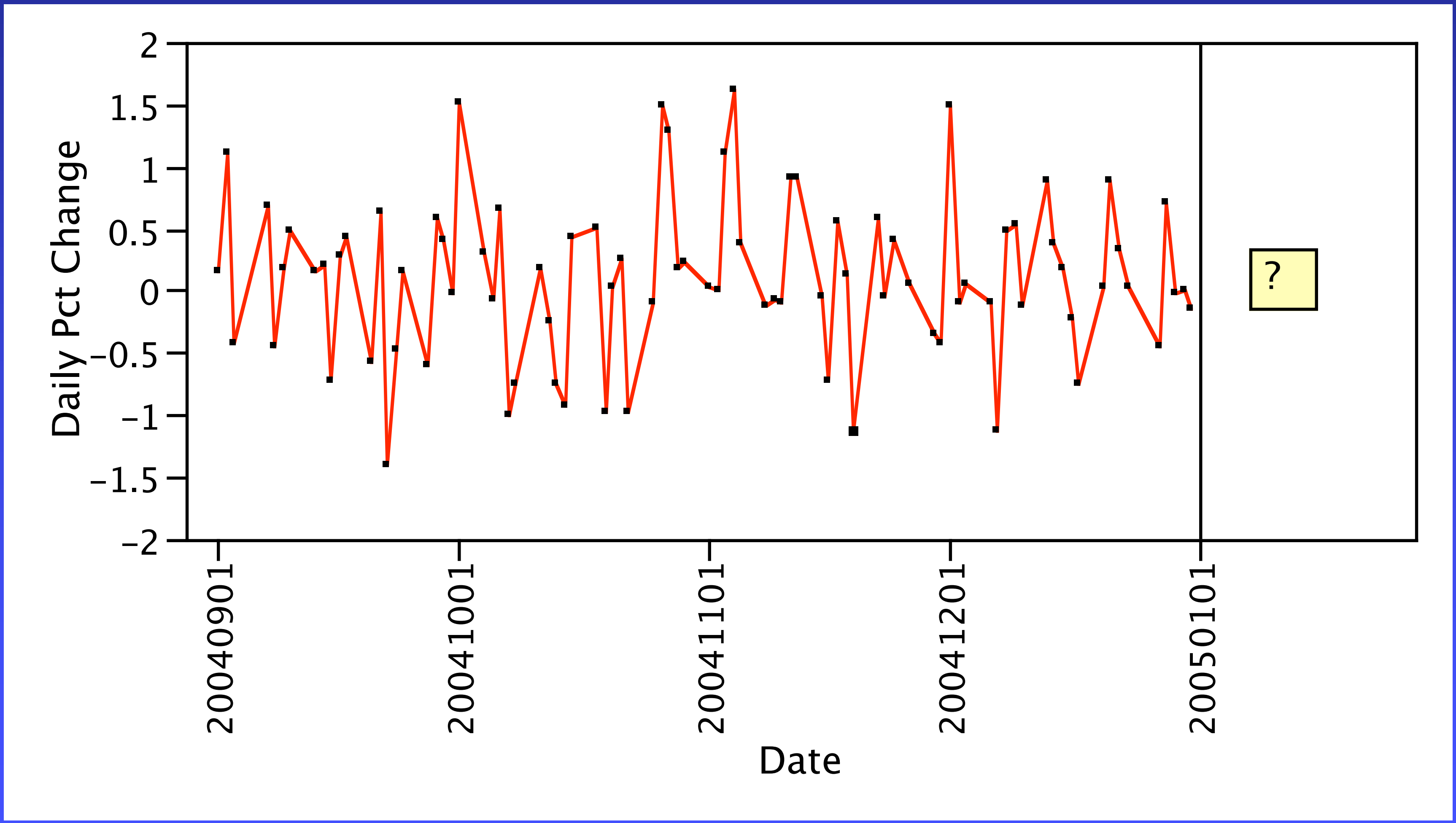
Example

Wharton
Statistics Department

- Predict the direction of the stock market
  - Use data from 2004 to predict market returns in 2005.
- Data
  - Daily returns (percentage changes) on the S&P 500 index during the last 3 months of 2004.
- Predictors
  - 12 technical trading rules
  - These are known for January 2005 ahead of time and so can be used to predict future returns.
- Next slides show plots, then the model…

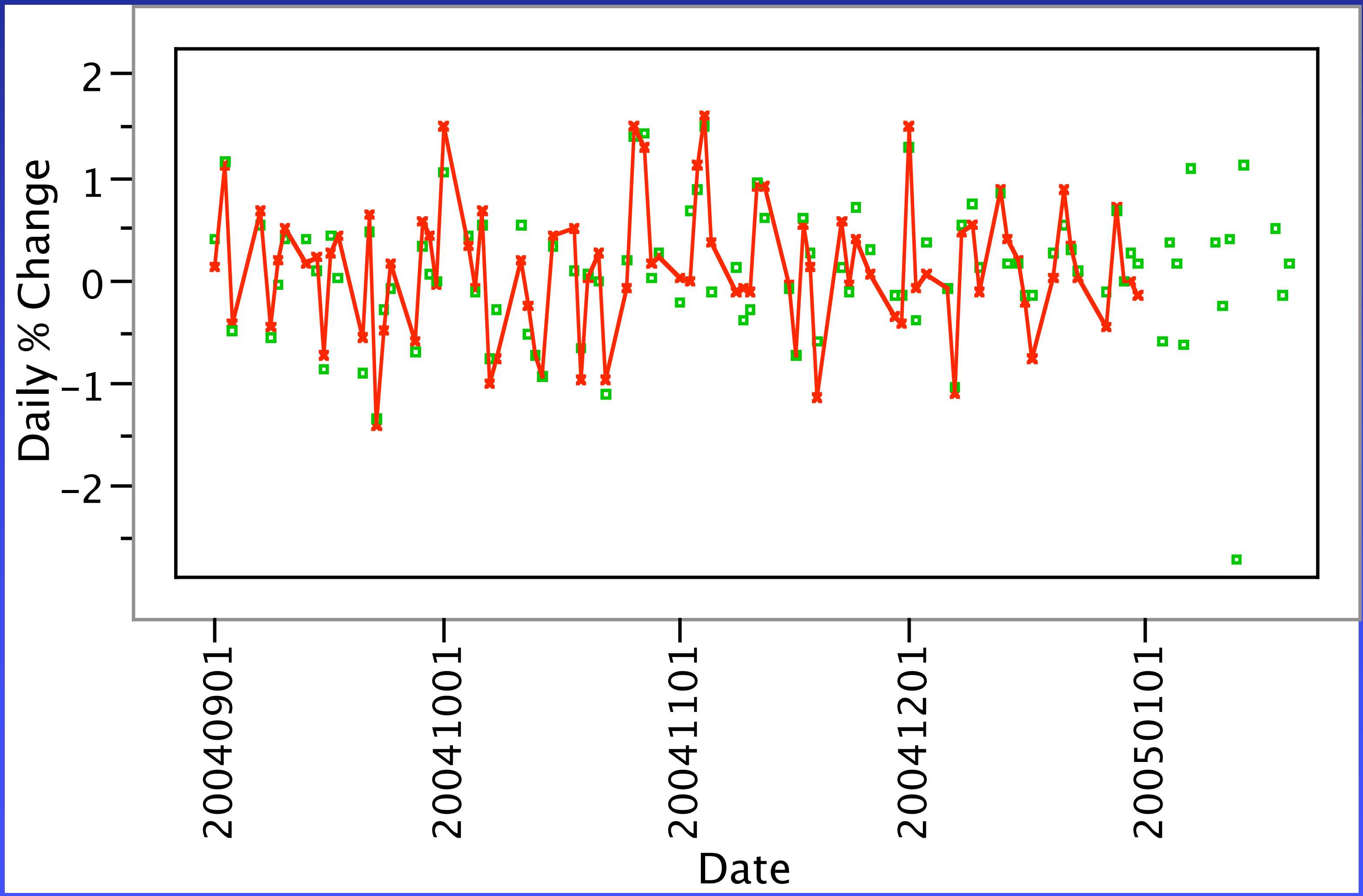
Last 3 Months of 2004

Wharton
Statistics Department



Predictions from a Model

Wharton
Statistics Department



Model Summary

Wharton
Statistics Department

- Data
  - $n = 85$ trading days in October through December, 2004
  - Search selects 28 predictors constructed from 12 trading rules.
- Statistical attributes
  - $R^2 = 84.8\%$ variation explained (adjusted $R^2 = 77.2\%$)
  - Overall F-ratio = 11.2 ($p < 0.001$)
- Individual coefficients
  - Almost all have p-value < 0.0001
- Model passes the usual statistical diagnostic tests with flying colors, even passing Bonferroni rules.

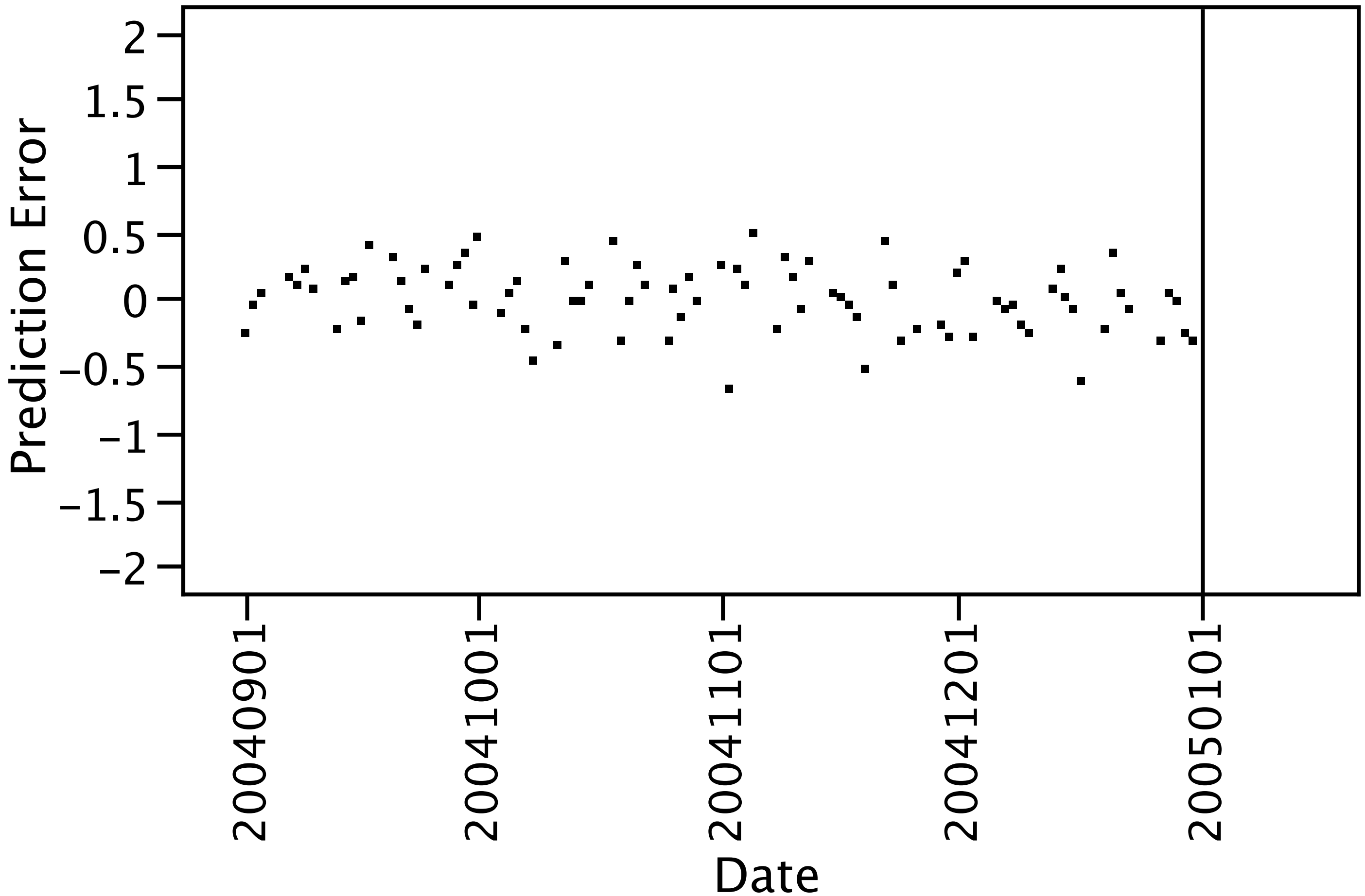
Parameter Estimates Look Great

Wharton
Statistics Department

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.323	0.078	4.14	0.0001
X4	0.172	0.040	4.34	<.0001
(X1)*(X1)	-0.202	0.039	-5.16	<.0001
(X4)*(X4)	0.126	0.036	3.52	0.0009
(X1)*(X5)	0.256	0.048	5.34	<.0001
(X2)*(X6)	0.289	0.044	6.59	<.0001
(X4)*(X6)	-0.222	0.050	-4.43	<.0001
(X4)*(X7)	-0.213	0.047	-4.54	<.0001
(X6)*(X8)	-0.243	0.048	-5.02	<.0001
(X5)*(X9)	-0.192	0.044	-4.35	<.0001
(X7)*(X9)	0.249	0.046	5.37	<.0001

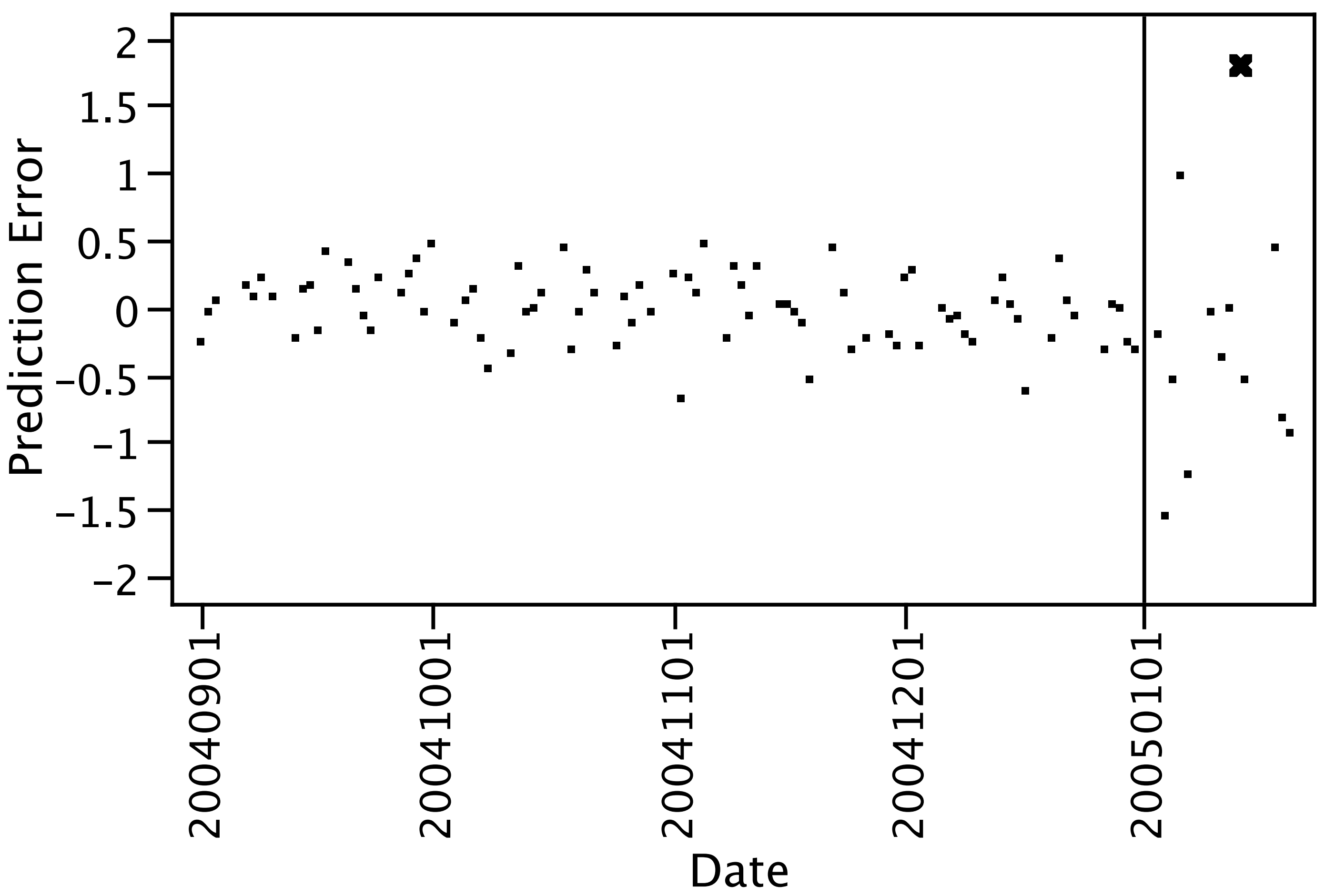
Prediction Errors, In Sample

Wharton
Statistics Department



Out-of-Sample Errors Larger

Wharton
Statistics Department



So, how’d you lose the house?

Wharton
Statistics Department

- How can a model look so good (summary statistics, in-sample fit), but predict the future so poorly?
- *Overfitting*

“Optimization capitalizes on chance.” (Tukey)

  - Overfitting describes a model that captures random patterns in the observed data as if these patterns can be extrapolated to other data.
- All those significant coefficients… these cannot be random, not with these statistics! Can they?

What are those predictors?

- Random noise
  - Filled columns with a normal random number generator.
- Model built predictors from 12 columns of random noise, plus
  - Squares of the columns
  - Cross-products of the columns
- Total of 12 + 12 + 66 = 90 predictors considered
  - Random patterns in these predictors match patterns in the S&P so well that it fools the standard diagnostics.
  - More predictors to consider than observations

Moral of the Story

Wharton
Statistics Department

- Shouldn't leverage the house,

  but if you do,

- Only trust a model if you understand the *process* used to choose the form of the model.
  - Automated modeling procedures have to be carefully monitored, or the results are likely to be spurious.

- In this example, it's easy to avoid the problem.
  - Cross-validation is not so appealing.
  - Bonferroni can control the process.
  - Ensure that the model never adds noise.

Stepwise Regression

- Where'd that model come from?
  - Ran stepwise regression in "forward selection mode" to select predictors from the list of 90 features.
  - "Promiscuous" threshold for adding variables kept the default p-to-enter = 0.25 criterion.
  - Ran backward elimination to clean up the model so the final structure looks impressive.
- Process generates a biased estimate of noise variation and a cascade of noisy predictors in the model.
- Better way to run software avoids the problem
  - Set the p-to-enter to 0.05/90 at the start.
  - Nothing added to the model, the right choice.

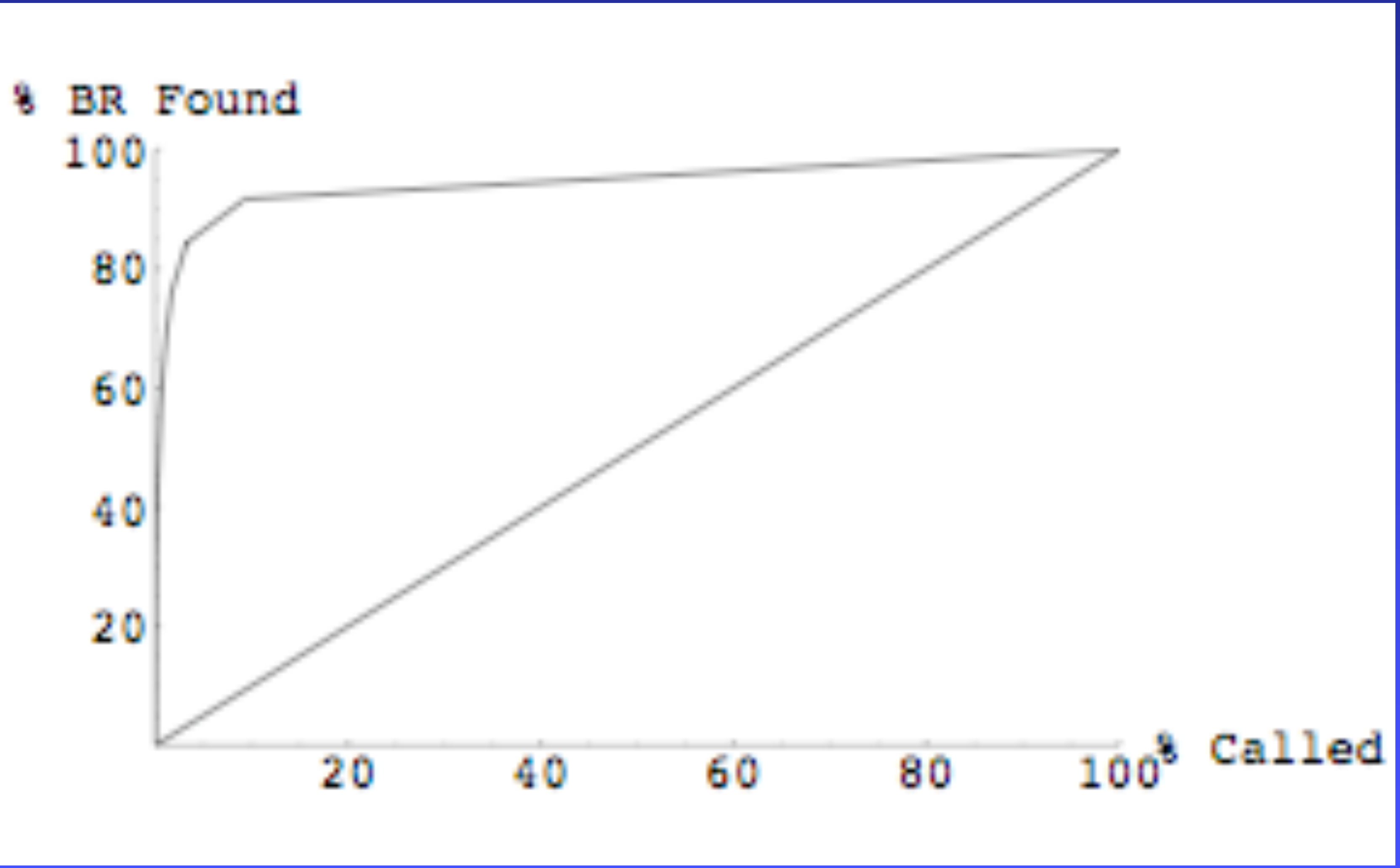
Don't blame stepwise regression!

Wharton
Statistics Department

- Predicting personal bankruptcy
  - Lots of good customers that you don't want to harass.
  - Few who won't pay you back that you'd like to find.
- Regression model predicts incidence of bankruptcy with lower costs than modern classification tree.
- Test results
  - Five-fold cross validation, with 600,000 cases in each fold.
  - Regression generates better decisions than C4.5, with or without boosting.
  - Huge lift (next slide)
- To be successful, regression needs a little help.

Impressive lift results

Wharton
Statistics Department



Helping Regression

Wharton
Statistics Department

- Lessons from regression applicable to any model and fitting process
- Expand the scope of features to find structure
  - Don't pretend the right features are the ones in the database.
  - Recognize there's not a true model.
  - Consider the possibility of higher-order interactions, subsets, and nonlinearity.
- Evaluate features to avoid overfitting
  - Estimate *standard errors* using the fit computed *before* adding a predictor rather than after.
  - Construct p-values to allow for rare, high leverage points.

Expanding the Scope

- Began bankruptcy modeling with 350 predictors
  - These include categorical factors, such as region.
  - Missing data indicators
- Add *all* possible interactions
- Use forward stepwise to search the collection of

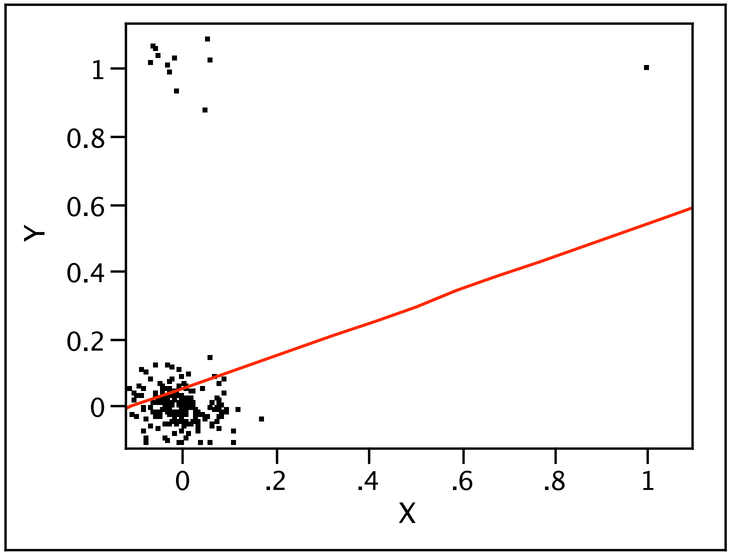
	350 base predictors
+	350 squares of predictors
+	350*349/2 = 66,430 interactions
=	67,610 features

Evaluating the Features

- Selection from a large collection of features requires a different method for deciding what it means to be "statistically significant"
  - Proliferation of features overwhelms standard method.
  - Large n $\neq$ normal sampling distribution (no CLT)
- Approaches
  - Cross-validation: Save some data to test the model to help you decide if you've really done better.
  - Thresholding: Use an in-sample test to avoid the sacrifice of data and the time to compute.

Example of the Problem

Wharton
Statistics Department



- P(Y=1) = 0.001, ind of X
- p-value ought to be?
- Usual summary
  - n = 10000, t = 14
  - p-value < 0.000001
- Interactions can concentrate leverage in rare combination
- Need a different sampling model, or a better p-value.
- Bennett's inequality does well (Foster & Stine, 2004)

Regression Can Succeed, but

- Fits as well as modern classifier, but…
- “Rigid and clumsy” search of interactions
  - Begins with the list of *all* features to consider.
  - If X1 and X2 are in model, why not try their interaction? No!
- Slow
  - “Breadth-first” search for next predictor
- Omits substantive features, domain knowledge
  - If you were to talk to an expert, they could offer ideas.
    - Genomics, credit modeling, database structure
  - Can you use this knowledge to find better models?

Each domain has many experts

Wharton
Statistics Department

Who offers the best advice?

Experts ⇔ Auction ⇔ Model

Wharton
Statistics Department

Any Predictive Model

Feature Auction

Domain Expert

Domain Expert

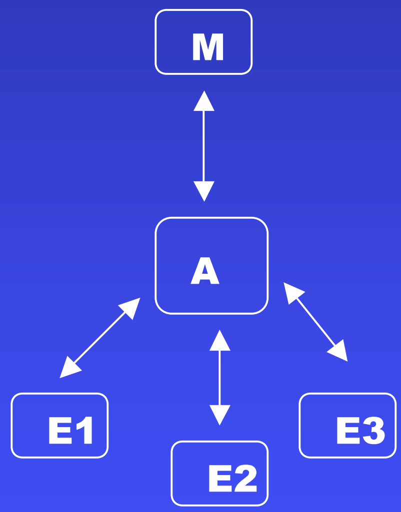
Domain Expert

Domain Expert

Different Modeling Process

Wharton
Statistics Department

- *Experts* recommend features based on context.
- *Auction* takes feature with highest bid.
- *Model* tests this feature.
  - Bid determines p-value threshold
  - Accepts significant predictors, rejects others
- *Auction* passes results back to experts.
  - Winning bids earn wealth for expert.
  - Losing bids reduce wealth.
- *Information* flows both ways.



Experts recommend features

Wharton
Statistics Department

- *Substantive* experts *order* features
  - Offer a sorted list of features to consider, or
  - Propose a strategy to generate "next" predictors
- *Automatic* experts
  - Interactions piggy-back on success of others
    - Allows search to consider high-order interactions
  - Principal components
  - Feature bundles that combine several variables to include as one
    - Allows search to include parameter shrinkage
  - Nearest neighbor predictors
    - Singular value decompositions

Auction is sequential

Wharton
Statistics Department

- Each expert offers a predictor to the auction given the history and state of the model.
  - Each expert has wealth as allowed Type 1 error rate.
  - Experts bid on predictors.
  - Each bid is a p-to-enter threshold.
- Auction takes the predictor with the highest total bid.
  - It collects the bids on this feature from the experts.
- Auction passes the chosen predictor to model.
  - Model assigns p-value to feature.
  - If p-value < bid, add the feature and "pay" bidders.
- Continue

Theory: Sequential Selection

- Sequential tests:
  Evaluate *next* feature rather than *best of all* features.
  - Essential when the choice of the next feature depends on what has worked so far, as in CiteSeer application.
- Fast, even when experts are dumb.
- SDR: the sequential discovery rate
  - Resembles an alpha-spending rule as used in clinical trials
  - Works like FDR, but allows an infinite sequence of tests.
- More theory…
  - Ordering captures prior information on size of effects

Sequential vs. Batch Selection

Sequential

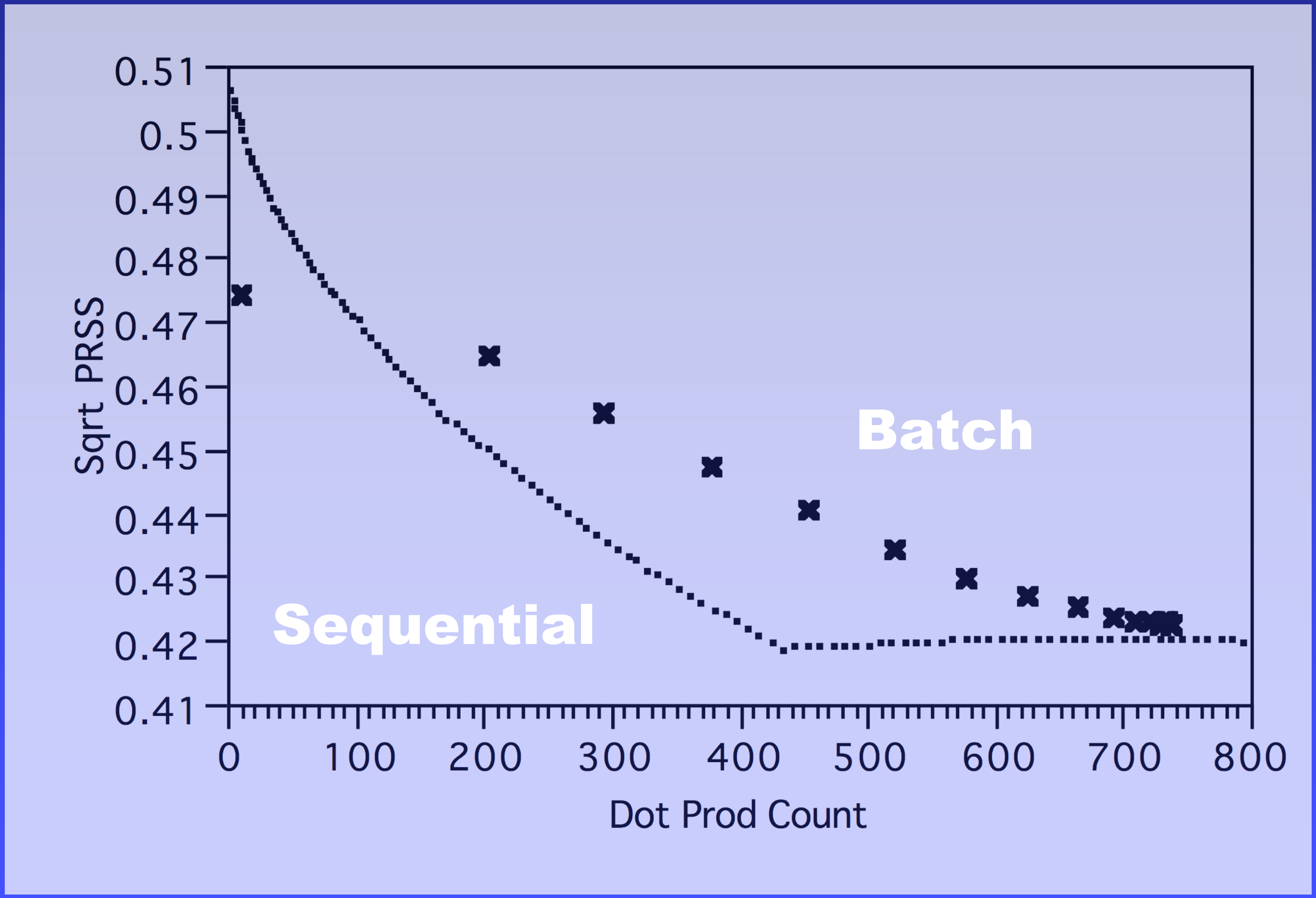
- Search features in order identified by domain expert
- Allows an infinite stream of features.
- Adapts search to successful domains.
- Reduces calculations to a sequence of simple fits.

Batch

- Search "all possible" features to find the best one.
- Needs all possible features before starts.
- Constrains search to those available at start.
- Requires onerous array manipulations.

Sequential works…

Wharton
Statistics Department



Theory: Bidding Strategy

Wharton
Statistics Department

- Auction prevents “strategic betting”
  - Experts offer honest estimate of value of the predictor.
- Multiple bidders represent each expert
  - Geometric bidder: Spend λ% of current wealth on next bid.
  - Use mixture of bidders with varying λ.
- Auction adaptively discovers smart experts
  - Auction rewards the bidder/expert with the right rate
  - Wipes out the others.
- Universal bidding strategies (universal Bayes prior)

Calibration and Models

Wharton
Statistics Department

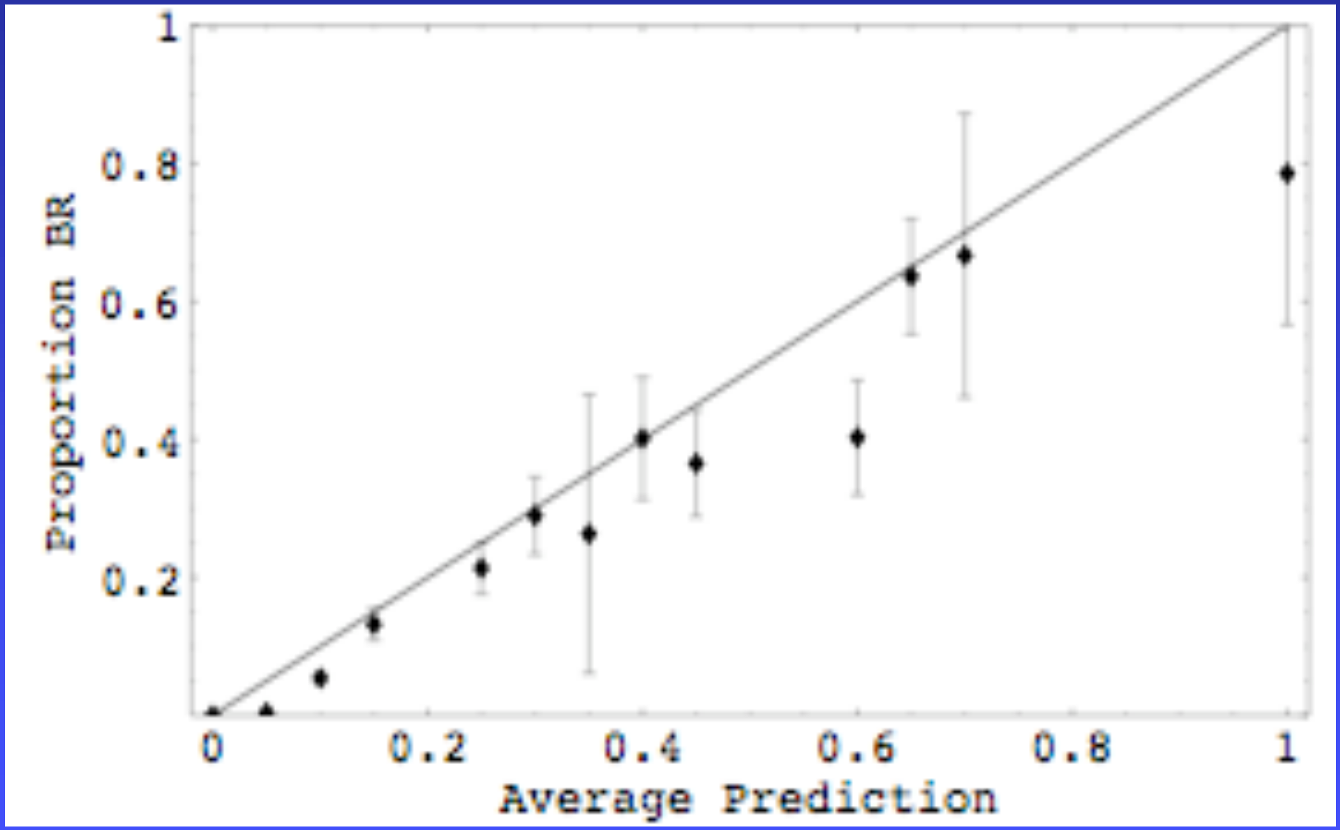
- Calibration
  - First-order calibration
  - Predictor is “right on average”
  - Examples
    - Doctors?
    - Weather predictions?

$$E\left(Y \middle| \hat{Y}\right) = \hat{Y}$$

- Automatic
  - Improve predictor with no knowledge by calibrating.
  - Simple scatterplot smoothing.
  - Incorporate as part of the modeling process.

Calibration plot

Wharton
Statistics Department



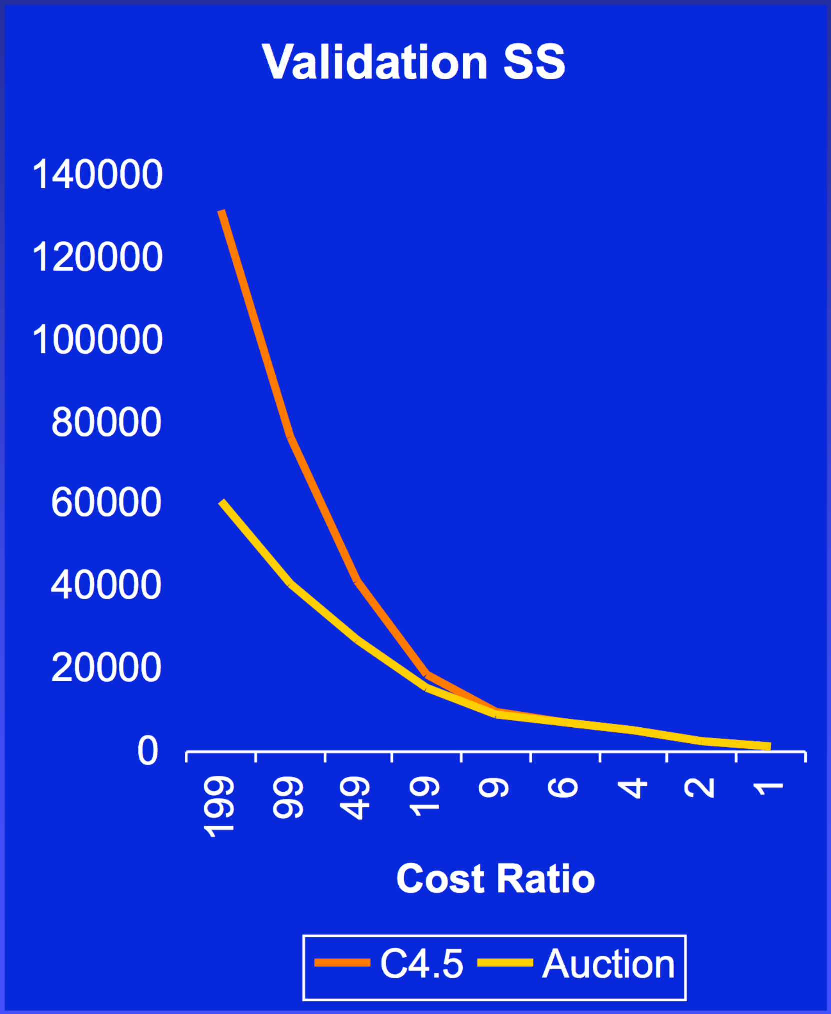
Auction: Some Results

Wharton
Statistics Department

- Rare events data
- Five-fold "reversed" cross-validation
  - 100,000 cases per fold
  - Fit on one fold, predict other 4 folds
- Methods
  - C 4.5 with boosting
  - Auction with calibrated logistic regression and multiple experts using SDR to spend alpha rate.
- Goal: Minimize costs of classification errors in the validation data.

Cross-validation Comparison

Wharton
Statistics Department



- At higher cost ratios, auction produces much lower costs.
- If the two errors have equal cost, either method does well.
- For each fold, use one logistic regression for all cost ratios.
- C4.5 uses a new tree for each fold and for each cost ratio within a fold.

Comments on Computing

- Prior code
  - Monolithic C program
- Auction
  - Written in C++, using objects and standard libraries
  - Modular design
    - Templates (e.g., can swap in different type of model)
    - Runs as a unix command-line task
    - Separate commands for data processing, modeling, and validation
    - Adopt C4.5 data file format

Closing Comments

Wharton
Statistics Department

- Key problem of data mining
  Find the right features without over-fitting
- Can learn from study of what it takes to adapt familiar methods like regression to data mining
  - Thresholding allows you to avoid extra cross-validation.
  - p-values are powerful way to communicate effect size.
- Auction modeling offers a framework that
  - Exploits domain knowledge if it exists
  - Combines various automated methods of feature creation
  - Runs quickly with any type of underlying model
- More information…www-stat.wharton.upenn.edu/~stine