Auctioning Experts in Credit Modeling

Robert Stine
Statistics Department
The Wharton School, Univ of Pennsylvania
May, 2004

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

- Anticipate default
  - Who are most likely to default in the near future?

- Detect fraudulent applications
  - Which loan applications are made up?

- Segment corporate bond market
  - Which companies are most risky?

- Other domains…
  - Employee evaluation: Who should we hire?
  - Disease prognosis: Who are most at risk?
  - Document classification: Can you find one like this?
Similarities

Different contexts, but common characteristics…

- Rare events
  - Few cases dominate costs.
  - Millions of accounts, thousands of defaults.

- Synergies
  - Linear models find little. Interactions work.
  - Many combinations seem plausible.

- Wide data: more features than cases
  - Interactions, transformations, categories, missing data…
  - Too many to find the best at each stage.
Common Objective

- Regardless of the context
  - Credit default
  - Detecting fraudulent loan applications
  - Segmenting corporate bond market

- Pragmatic goal remains prediction.

- Best model generates highest revenue
  - Asymmetry of costs, presence of rare events

- Many schemes for building a predictive model
  - Algorithms, features, criteria…
Which model to use?

Every domain has experts…

But which offer good advice?
Automated Methods

- Expense of custom modeling hard to justify

- Automate process
  - Higher productivity
  - “Objective”
  - “Rigorous”
  - Convenient

- But what about expert know-how?
  - Is the loss of their insight worthwhile?
Comparison

Substantive

- Pick model “by hand”
  - Advantages
    - Leverage domain knowledge
    - Can “interpret” for regulator
  - Disadvantages
    - Did we miss something?
    - Time consuming to
      - Construct
      - Maintain

Automatic

- Computer search
  - Advantages
    - Scans entire data warehouse
    - Hands-off, fast
      - Construction
      - Maintenance
  - Disadvantages
    - Lost domain expertise
    - Hard to explain or interpret
Best of Both Approaches

**Substantive**
Pick model “by hand”

- Advantages
  - Leverage domain knowledge
  - Can “explain” to regulator

- Disadvantages
  - Did we miss something?
  - Time consuming to
    - Construct
    - Maintain

**Automatic**
Computer search

- Advantages
  - Scans entire data warehouse
  - Hands-off
    - Construction
    - Maintenance

- Disadvantages
  - Lost domain expertise
  - Hard to explain or interpret
Best of Both
Auction = Experts + Model
**Awktion Modeling**

- *Experts* recommend features.
  - Bid reflects strength of “conviction” (Bayes prior)

- *Auction* identifies feature with highest bid.

- *Statistical model* tests 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

- Experts recommend predictive features

- *Substantive* experts order features
  - Domain knowledge of specific area
  - Prior models in similar problems

- *Automatic* experts
  - Interactions based on other experts
  - Transformations
    - Segments, nearest-neighbor, principal components
    - Nonlinearity
  - Feedback adjustments for calibration
Underlying Theory

- Streaming feature selection
  - Sequential, not all at once
    - “Depth-first” rather than “breadth-first”
  - Overcomes width constraints
  - Ordering captures prior information

- Universal bidding strategies

- Multiple testing without overfitting
  - False discovery rate (FDR) for infinite sequence of tests.

- Calibration
  - Ensures predictions track reality.
  - Adaptive link function
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...
Example

- Predicting default
  - Logistic regression model
  - 15,000 cases, 67,000 possible features (most interactions).

- Standard model finds linear predictor
  - Higher risk with lower line allowance.
  - Statistically significant
Example: Nonlinear pattern

- **Auction model**
  - Experts recommendations based on state of model.
  - Look for combinations of extant predictors.

- **Discovers nonlinear effect**
  - Nonlinear effect for size of credit line
  - Statistically significant “bump” in risk
Example: Synergies

- Feedback expert
  - Builds interactions among predictors in current model.
  - Limited search does not obscure simple predictors.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Found in Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral score</td>
<td>Marginally linear</td>
</tr>
<tr>
<td>Missing data</td>
<td>Behavior score affects these differently</td>
</tr>
<tr>
<td>Non-linear</td>
<td>Larger for high scores</td>
</tr>
<tr>
<td>Synergies</td>
<td>Changes with payment</td>
</tr>
</tbody>
</table>
Summary

- Auction modeling combines
  - Domain knowledge
  - Automatic search procedures
- Offers
  - Fast, guided search over complex domains
  - Strategies for constructing features in parallel.
  - Flexible statistical models
- More information…

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