Enhancing Compliance with Predictive Analytics—

FTA 2007

Revenue Estimation and Research Conference

Reid Linn

Tennessee Department of Revenue

reid.linn@state.tn.us

Sifting through a Gold Mine of Tax Data

“Discovering the patterns, trends, and anomalies in massive data is one of the grand challenges of the information age.”

[Kantardzic, M. and Zurada, J., *Next Generation of Data Mining Applications*, John Wiley & Sons, New York, 2005]

Outline

- Introduction
  - What, Why, How
  - Caveats—Currently in Development Stage…stay tuned for results
  - …and, Where Does this Fit into a Research Division?

- Predictive Modeling and Data Mining
  - Definitions, Examples and Methods

- Logistics & (preliminary) Lessons Learned
  - Going from Modeling to Selection
  - Bridging the Gap to Other Research Responsibilities

- Discussion, Questions, Feedback

Enhancing Audit Selection—
Strategies and Goals

- Tax administration exists to ensure compliance with the tax laws
- Audit selection is one of the major enforcement mechanisms for—
  - Sustaining and enforcing compliance
  - Maintaining the stream of revenue
- Project Goal: to help achieve maximum taxpayer compliance

- *What is the best combination of strategies for achieving this goal?*
- *How do we efficiently use limited resources in this effort?*

- The Department already profiles potential audit candidates
  - Utilizing data in our transactional revenue integrated tax system
  - Rule-based selection, Leads, Relationships (e.g., subsidiary audits)
- Predictive modeling provides an additional advanced method
  - ...in a systematic & automated way, and on a very big scale

Predictive Platform—
Leveraging Existing Technologies & Acquiring Additional Tools

- Existing Platform
  - Long-time SAS shop (from mainframe to mouse)
  - Client-server setup
    - SAS BASE, STAT, ETS and others for complete data manipulation and statistical analysis
  - Raw transactional data extracted from mainframe (data warehouse in future) and is stored, manipulated & processed on a single dedicated Windows server (primarily using SAS/CONNECT)
  - ESRI ArcGIS ArcView—sophisticated geocoding & mapping (client side)

- Additional Technologies
  - Additional CPU (2-CPUs total), RAID technology with multiple controllers, and a lot more disk storage for new data & new SAS temp files
  - SAS Enterprise Miner—full, heavyweight data mining solution but integrated with existing modules
  - SAS Text Miner—additional module to incorporate case notes containing significant & otherwise uncoded predictor information
  - SAS DataFlux—data quality and cleansing tool for record linkage, address verification, geocoding

Predictive Platform—
New Tools: What Fits versus Other Options

- Toolbox Enhancement
  - We chose to ramp up our capabilities with additional SAS modules/products
  - Obvious benefits of seamless and straightforward integration, smaller learning curve, easier pursuit of advanced techniques (getting into the guts of the software) that scale to large sets of data
    - Perl regular expressions (PRX), Hash Data Step Component Objects, Pipes
    - Scalable Performance Data Engine (SPDE)
    - Macros, Macros & more Macros please (and of course, PROC SQL)
  - Predictive analytics doesn't require SAS EM
    - PROC LOGISTIC can do the trick to get us started
    - ...BUT, on a limited and much smaller scale without other major EM benefits

- Alternative Solutions
  - Other data mining packages available, from enterprise level to standalone and even freeware
  - Business intelligence tools, but data mining as described in BI tools is different (OLAP, queries, cubes, etc.)
  - Contractors offering solutions from full compliance software deployments to single audit selection efforts

Predictive Modeling—
Explanation and Examples

- What is data mining?
  - Many definitions that aren't all synonymous
  - Advanced methods for analyzing large data sets in search of consistent patterns not readily apparent
- Predictive models are the most frequently used form of data mining
  - Allows organizations to predict the outcome of a given process
  - Produces a numeric score indicating the most likely outcome
  - In audit selection the score ranks the expected outcome of an audit
- Countless examples exist in industry
  - Used commonly in finance and insurance industries for assessing credit risk and identifying fraudulent activities
- Over 200 data mining initiatives ongoing in the federal government
  - IRS has at least 9 planned and 2 operational efforts (based on 2005 review)
  - Including, its Electronic Fraud Detection System (EFDS)

Data Mining Overview—
Not a black box...

- Integrated process with several different phases
  - Data identification and preparation
  - Exploration
  - Model building and validation
  - Deployment
- Preparing the data and combining different sources into a “high quality” data set is critical
- The software technology will allow advanced exploration and visualization to identify patterns and trends on a very large scale
- Different models provide diverse answers
  - Choose the set of models producing the most accurate and stable results
  - Ongoing process with constant fine-tuning and adjustments
- It is important to assess the predictive model’s success and how it improves audit selection

Data Preparation—
Obtaining, Cleansing, Standardizing, Gleaning

- Easy to be fooled into thinking it’s simple & won’t take long
  - Triple or quadruple the allotted time (even if a data warehouse is available)
  - Translating intended structure into efficient computer code is daunting
- Choosing, obtaining and processing multiple data sources
  - DOR tax data, other state agencies, private vendors, census
- Pre-modeling Goal: to create a single, whole picture of the taxpayer with as much information as possible
  - Then, to aggregate and condense into the best possible structure
- Requires record linkage techniques for the best possible match
  - Address standardization, name cleansing, phonetic matching
  - Geocoding for mapping and adding demographic information
- Errors, missing data, outliers, undercoverage, high dimensionality
  - Complexities can be overcome
  - SAS EM is customized to correct for many of the problems

Data Preparation—
Aggregating into a Single Useful Picture

- Selecting the best structure depends on analytic strategy
  - Most predictive modeling condenses multiple units into one observation
  - Two necessary tasks: choosing meaningful aggregates and then developing (coding) appropriate transformations
- Entity versus Account level data
  - Bank analogy: multiple household members with multiple accounts/products
  - Entities are wholly audited, but individual accounts submit tax returns
  - Can hierarchical modeling work?
- Deciphering existing audit results
  - Target window for analysis
  - Significant assessments, tax types, reasons, adjustments, settlements
- Dual analytic sets: Target data versus Scoring data
  - With audit selection, a target set is small (rare observations)
  - Massive scoring set requires special coding for scaling to millions of records

Predictive Modeling—
Wait, more preparation?

- Oversampling (or, stratified sampling on the target)
  - Non-compliance is “rare” relative to millions of returns
    - Create a biased model to over-represent the target class, then adjust prediction results using prior probabilities
- Categorical inputs—Collapse by Recoding, or even better by Clustering
- Missing values—more common as data & dimensionality increase
  - Disregard, Impute, or Create new “unknown” levels
    - Imputation techniques: simple (e.g., medians) vs. complex (e.g., tree-based)
- Variable redundancy and irrelevancy—reducing dimensionality
  - Principal Components or Variable Clustering
    - Clustering groups most correlated variables and selects representative
  - Variable screening—eliminate irrelevant variables, attentive to interactions
- Transformations—nonlinearities, skewed distributions, outliers
- Almost there—Split data into Training and Validation sets
  - Treat validation as unknown data but stratify for equal proportions

Predictive Modeling—
Fitting Different Models

- Prediction—SAS EM provides multiple statistical methods
  - For audit selection, we'll focus on Logistic Regression, Decision Trees, and Neural Networks
  - Test and possibly incorporate various combinations, or Ensembles
    - Boosting—model averaging with error correction weights (future capability?)
  - Two-Stage Models—model both outcome probability & expected profit
    - Identifying candidates with highest propensity for non-compliance is valuable
    - But, also would like to maximize productivity (and ROI) by pursuing those with greatest assessment potential
- Model Comparison—Let the best fitted model(s) win
  - Connect all models to a comparison node for sophisticated analysis and visual representations, and select one having the most predictive power
    - ROC & Lift charts provide comparisons of each model's classification accuracy
    - Summary statistics provide rankings, misclassification, profit/loss, error statistics
- Score!
  - Attach best model to a score node for internal or external scoring

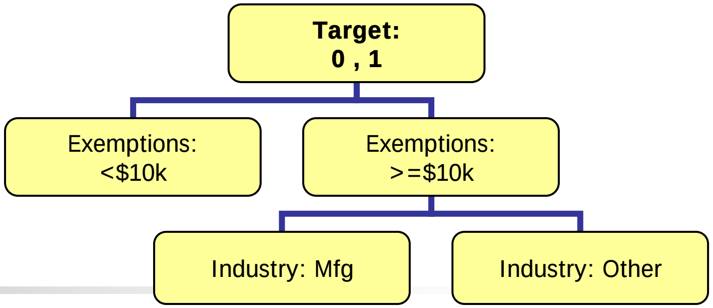
Predictive Modeling—
Logistic Regression

- A “binary” shift from regular revenue estimation of linear dependent variables
  - Modeling probability of non-compliance and ranking highest

$$\boxed{\log(\frac{\hat{p}}{1-\hat{p}}) = \hat{w}_0 + \hat{w}_1 \cdot x_1 + \hat{w}_2 \cdot x_2}$$

- Estimated probabilities obtained by taking the log of the odds to restrict the outcome between 0 and 1
- Odds Ratios obtained by exponentiating the coefficients
  - Measures odds of being non-compliant relative to predictors
  - e.g., a $1,000 change in exemptions increases odds of non-compliance by 5%
- Assessing predictive ability (no R-squareds)—
  - Want to maximize the “Percent Concordant”, or the estimate of correctly predicted probabilities (also, referred to as a C-stat)
- Subset selection—Stepwise versus Best-Subsets (or All-Subsets)
- Assessing classifiers on the validation data with multiple measures
  - ROC curves, profit, sensitivity, mean squared error (MSE)

Predictive Modeling— Decision Trees



- Multiple Variable Analysis Technique that Creates a Set of Decision Rules
  - Starts at the top, or Root node, and autonomously or interactively creates Branches that split, or segment, input variables
  - The Splits and corresponding decision rules are based on traditional statistical tests of significance and other machine learning algorithms (CHAID, CART, etc.)
- Decision Tree Split Search
  - Algorithm segments each input variable according to a chi-square test to obtain the greatest independence between the two branches
  - After the best split on each input is determined, the input split having the highest logworth score (or lowest adjusted p-value) is chosen as the next partition
- Process continues until a Maximal Tree is developed
  - This tree predicts the training data well, but must be adjusted, or Pruned, to generalize well on the independent validation data
  - Different assessments available: decisions (accuracy), rankings (correct ordering), estimates (low average squared error)
- Multiple Benefits: ease of development and interpretation
  - Powerful assessment of most important variables
  - Selects inputs, accommodates nonlinearities, detects interactions, handles missing

Predictive Modeling—
Other Model Types

- Neural Networks—a real black box
  - Powerful and flexible nonlinear models used for supervised prediction
  - Is like a regression model on a set of derived inputs, called Hidden units, that are regressions on various linear combinations of the original inputs
  - Not necessarily better—tendency to overfit, difficult to understand, and extreme lack of interpretability, but…
    - May provide good predictions by itself or as part of an Ensemble
    - Decision trees or surrogate models on results can demystify complexities
- Ensembles—combination of different model techniques
  - Combines predictions from multiple models (multiple samples or methods)
- Two-Stage Models
  - Estimate noncompliance propensity & assessment amount
  - Prediction from the binary outcome model becomes an input for the expected profit (assessment) model
  - Can be constructed separately or SAS EM contains a two-stage node

Predictive Modeling—
Model Assessment and Scoring

- After tuning each individual model with the best fit statistics, it's time to choose the model with best predictive performance
- SAS EM major benefit—provides extensive summary statistics and associated graphics for easy comparison of all models
  - ROC charts & Lift charts can show best models for separating outcomes
  - Default best model is one with the smallest validation misclassification rate
  - But, can choose from a multitude of other measures
- Profit (assessment) optimization is an alternative selection method
  - Useful if decent assessment/profit information available
  - Best if we can model the expected value of the profit random variable for each outcome
- Scoring can occur inside or outside of SAS EM on data structured similar to training & validation (SAS, C, Java)
  - Better to score outside of SAS EM because of magnitude of tax returns
  - Score code modules run all prediction code and transformations

Data Mining—
Future Techniques and Research Possibilities

- Text Mining and Cluster Analysis
  - SAS Text Miner runs as a node within SAS EM and allows for analysis and pattern searching of unstructured text
  - Case notes are a source of important predictor information describing assessment reasons and audit outcomes (the real scoop)
  - Must incorporate other pattern discovery techniques because obviously no notes for non-audited returns
  - Cluster analysis—unsupervised classification technique that groups data based on similarities in input variables
  - Attempts to find segments with similar attributes that can be applied to new data for classifying potential new targets
- Other Possibilities:
  - Overall Secondary benefits—great, extensive set of taxpayer data
  - Time series—hope to discover trends/patterns in corporate taxes
  - Other predictive or pattern discovery type questions—sure to be many

Audit Selection Process—
Logistics & Lessons Learned

- Domain versus analytical and data expertise
  - Not feigning audit expertise, but trying to assist in enhancement of audit selection (& hopefully incur consequential benefits for other research tasks)
  - Collaboration with the audit division is an integral and continuous part
  - Crucial to have the domain experts in audit identify relevant processes, business questions and data necessary to start modeling
- Predictive model results and audit selection scores will be provided to audit division management and departmental executives for ultimate decisions, implementation and deployment
- Realism about effort and time involved
  - Difficult to envision the programming difficulty of seemingly simple concepts related to the targeted business problem and its requisite data
- Learning curve—Steep, but not setting out to reinvent the wheel
  - Software training, other state efforts, existing credit scoring models
  - Not an automatic process, but instead a formidable discipline

Discussion, Questions & Feedback—