ACTS Phase 2 Model

Prepared by Fair Isaac Corporation.

San Diego, CA

6-11-2003

Bruce Harris

Dr. Frank Elliott

Presented at the 2003 IRS Research Conference, June 2003

Introduction

The objective of this project was to provide a ranking of taxpayers by their probability of utilizing Abusive Corporate Tax Shelters (ACTS). We have utilized IRS taxpayer data and financial data for companies with assets exceeding $250 million dollars to achieve these objectives.

The model developed during this project utilizes contemporary data mining technology to process the data, select variables for the model and to measure the accuracy and generality of the result.

The result obtained in this phase provides a very strong workload selection model with expected shelter detection rates in the CIC as high as 55%.

Previous Project

This project repeats an effort that concluded in January, 2002. The previous project had the same goal, a ranking model of the abusive tax shelters in the LMSB taxpayer population. The data for the present project is richer in several ways. The match rate between SOI data and financial data has doubled, increasing the effect and value of the financial data. The tags, indicating the presence or absence of shelter activity have been refined by the IRS, combining four surveys and a new source of shelter tags, disclosures. The improved quality of the data along with more accurate and complete tagging enables a more accurate and useful ranking model.

Technology Descriptions

This section describes the assortment of data mining and modeling technology and technique use for this project.

Neural Networks

Neural networks are used by the ranking model to form a ranking of taxpayers. Neural networks are used as a form of regression, used to develop a probability of a taxpayer utilizing abusive tax shelters in a specific year. The neural network used for the ranking model has 29 input variables, 2 hidden neurons and an output neuron. Figure 1 illustrates the network developed for the ranking model.

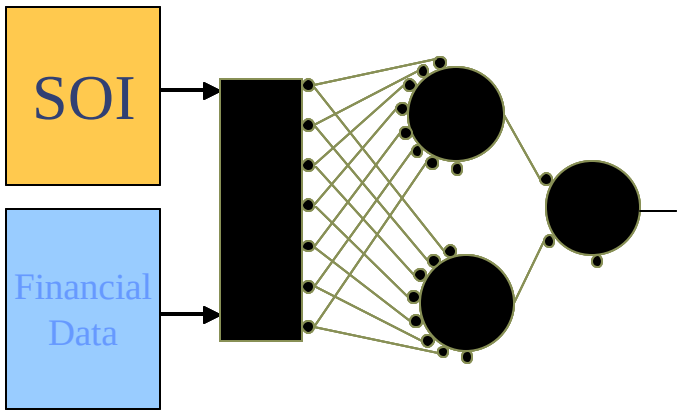

Figure 1 - Ranking Neural Network

Variable Selection Techniques

Variables in the neural network ranking model were selected with a technique that combines sensitivity analysis and principal component analysis. This variable selection technique is described in detail in the report for the previous project.

Cross Validation

The ranking model was developed with a relatively small data set. 5 fold model development and testing were used to assure an accurate and general result. The details are reported in the ranking model development section.

Data Development

Data Sources

SOI Data

SOI data were provided relatively complete coverage for tax years 1990 through 1999 and sparse coverage for 1981 through 1989. The data was provided for activity codes 219, 221,223, and 225. As shelter data was available for 223 and 225, the modeling effort was restricted to 223 and 225. The SOI data set provided 206 fields.

BRTF Data

BRTF data were provided relatively complete coverage for tax years 1994 through 2000 and sparse coverage for 1990 through 1993 and 2001. The data was provided for activity codes 219, 221,223, and 225. Many of the fields used in the models were not available in the BRTF so it was not used during development.

Financial Data

Financial data was provided by the IRS for public companies covering the date range of 1991 through 2001. 51% of the CIC taxpayer years matched to financial data. This match rate is up

from 30% in the earlier project. This match rate meets our expectations of the number of public companies in the CIC. The financial data set provided 172 fields.

Tags

The objective of the ranking model is to find tax filings affected by abusive corporate tax shelters. The construction of this model required examples of returns affected by ACTS ("bads") as well as returns unaffected by ACTS ("goods"). The determination of whether or not a return is affected by ACTS and, if so, the value of the adjustment to taxes due to ACTS are called "tags."

There are two sources of good tags, both surveys of IRS personnel. A 1999 survey (the "Buffalo Survey") conducted by the IRS Buffalo office for gathering examples of returns unaffected by ACTS, and a 2002 survey conducted by LMSB for finding booth goods and bads. In the latter survey, goods are simply EIN-years are documented as having zero adjustment to taxes due to ACTS.

There are also two different sources of bad tags. The first is a combined registry of adjustments due to ACTS compiled in 2002 by LMSB. This registry includes the adjustment information from the 2002 LMSB ACTS survey. The second is a log of disclosures of tax returns affected by ACTS; these disclosures were made to the IRS by corporate taxpayers under an ACTS amnesty for disclosure program conducted by the IRS in 2002. While the LMSB registry has well-documented adjustment amounts for the affects of ACTS on tax returns, the disclosure log frequently includes only the information that a particular tax return was affected by ACTS with no attempt to estimate the dollar amount of the adjustment. Hence, while both sources of bads are useful for training the ranking model, only bads from the LMSB ACTS registry are useful for the training of an estimate model.

The process of creating a unified collection of tags for model training and evaluation proceeds in three stages. First all the goods documented in the Buffalo survey and the 2002 LMSB ACTS survey are merged into the tag set with the tag "G" for "good." Second all tax filings affected by ACTS disclosures are merged into the tag set with the tag "D" for "disclosure" with the new tag over-writing the previous tag when necessary. Third all tax returns with non-zero adjustments from the LMSB registry are added with the tag "S" for "Shelter" with the new tag overwriting the old tag as necessary.

Variable Generation

Overview

Several techniques were used to develop candidate modeling variables from the IRS data sources. From the original data fields in the SOI dataset of 206 fields we generated about 1200 variables. These candidate variables were analyzed and processed in the variable selection steps of the model.

Profile variables

Some of the variables in the model were derived using a time-dependent weighted-average technique called "profiling." Profiling reduces the effect of short-time variations in the value of a variable. A profiled variable contains information about the trend over time of the original variable. Profiling does not give all values in the history of an input variable equal weight; rather, profiling weighs current data more heavily than past data. The decay time-constant a of the profile determines the relative dependence of the profiled variable on past

data versus current data. For instance, for a two-year profile variable ($a = 1/2$), the weight given to two-year-old data is only 37% ($1-1/e$) of that given to current data. Similarly, for a five-year profile variable ($a = 1/5$), the weight given to five-year-old data is only 37% ($1-1/e$) of that given to current data

Shelter Risk Rule development

We now describe the process for developing risk rules to aid in identifying the likely presence of particular shelter types. The purpose for this effort is twofold: first, we expect that informative rules will boost model performance, and second, the rules provide explanations to auditors as to why a return was ranked as high-risk, and can thus assist in their investigations.

For the tax years under study, we chose to develop rules for the most commonly occurring listed transactions. This choice was only mildly restrictive, as the commonly occurring shelters comprised nearly 80% of the total. The overwhelming majority of listings occur less frequently, and while we were able to identify potential rules for some of these, the data did not permit us to conclude with certainty that those rules were of use.

A number of techniques were used to develop risk rules. In all cases, we first attempted a straightforward identification of the fields commonly used to claim deductions for each shelter. Secondly, we attempted to define economic ratios which differentiate shelter users from non-users. In some cases we attempted to define time-dependent variables meant to identify sudden changes in certain fields. In other cases, we used principal components analysis to properly weight multiple fields. The data fields we use are the aggregate amount reported—we did not have detailed schedules for each field. Hence these variables do contain noise (not surprisingly, given the desire to mask such activity).

Throughout this process we examined IRS listed transactions and worked with the OTSA technical advisers to get a deeper understanding of shelter activity. Expert input was critical because most shelters are disguised on tax returns in widely varying ways. For example, deductions claimed using the LILO shelter seldom appears in the same fields on form 1120. Moreover, some shelters occur more frequently within certain industries (as with LILOs), or are associated with certain preparers, or may have other identifying characteristics. Finally, some listed transactions have variations with very different 'fingerprints', and for these it is important to focus on the right common features. As a final test for each rule, we held a conference call with the technical advisers to review the work. The advisers often suggested additional (unexpected) fields to test, which we evaluated during the call.

We evaluated rules against a set of known goods, bads, and disclosures from the merged shelter database. In general, the number of known bads varied from ten to a hundred, while the known goods amounted to a few thousand. Rules which clustered the highest percentage of bads in the top 5 percent were kept. Finally, the model itself selected which rules we developed to incorporate. Of seven major shelter types for which we developed rules, the model accepted three.

The three rules accepted were designed to identify specific shelter types. These three comprised 176 of the 530 shelters; however, the variables used in these three rules overlap with many commonly occurring shelters. The reason is that there are a limited number of fields commonly used to claim large deductions, which are used for different shelter types. Hence these rules will tend to identify more that just the three shelter types for which they were developed.

Data Integration

Tax and Financial Data Merge

The SOI data set is indexed by EIN-year and the financial data set is indexed by EIN-period, where the period consists of both year and month. When there was financial data for a particular EIN-year, the IRS provided the matching financial period in the SOI data set. After computing the necessary derived variables in the SOI data set and the financial data set, the data sets are merged using EIN-period and records with duplicate keys are eliminated.

Ranking Model

Target Description

Because shelters based on corporate owned life insurance (COLI) are now criminalized, the need for the ranking model to find tax returns affected by such shelters is moot. Since COLIs are so numerous in historic data that they would dominate the data set and distort the model, we exclude them. We do, however, include disclosures as a primary source of non-COLI tax shelters.

Data Corpus

The model is developed from the 1200 generated variables from the 206 SOI fields and 172 financial data fields. The development data set contains 8289 taxpayer years. Within this data set 786 observations are tagged as abusive tax shelters. Of these, 633 were from disclosures, and 153 were from the shelter surveys. All of the non-shelter observations were from surveys.

The number of fields in the data set was reduced to 200 fields by removing variable with extremely low correlation to the target.

Five fold segmentation was used to manage the training and validation of the model. A system was coded that generated hundreds of pseudo random splits. Each of these 5 way splits was evaluated and ranked to find a split that had very similar ratios of goods to bads. This process is used to assure generalization during model training.

Model Description

The Database Mining Marksman hardware and software was used to calculated many hundreds of neural network models, exploring the neural network complexity, and set of model variables that provide the best accuracy respecting the performance on a 20% test set, used to test for generality.

Industry risk table

A concerted effort was made to factor industry specific risk into ranking model. The IRS supplied NAICS code for all taxpayers. We obtained the legitimate hierarchy of NAICS codes from their website, and built a hierarchical risk table. We computed the odds ratio at each branch and leaf of the NAICS tree. We created a flexible algorithm that weighted the leaf ratios and the higher populated levels above it in the hierarchy and set about to optimize the weighing between local and global influence. Extensive experimentation revealed that we could not find a setting that provided generality over the model data set. This over fitting was detected in two ways. First, we found that that the industry risk factor dominated the model during neural network training, but the resulting models had very low Kolmogorov-Smirnov (KS) statistics, in the range of 8 to 10%. However, when we dropped the risk factor from the model the KS jumped to the mid 30% level, on preliminary models. That indicated that the risk factor was a system that memorized the risk of shelter abuse in non-general ways. After observing this effect, we examined the risk factors alone, and found that we could not form a global-local weighting scheme from 75% of the taxpayer-years that would be confirmed in the remaining 25% of the data. Apparently, there are not enough taxpayer-years in each industry to use industry as a predictive factor.

Variable Selection

We reduced the number of candidate variables from the 1200 generated in data development to the 200 variables with the strongest correlation to abusive tax shelters. Those 200 variables were processed with the Database Mining Marksman hardware and software by building a set of eight neural network models with between one and eight hidden neurons. Each model was built on 80% of the data, and tested on 20%. After a set of models was trained and evaluated, a blending of Principal Component Analysis and Sensitivity Analysis was used to eliminate several redundant variables or variables that contributed very little to the model. Those variables, usually three to five per session were eliminated from the next found of eight models. That process was repeated until there were no variables remaining. After this large set of about 400 models, the model that produced the best results, moderated by the significance of the number of weights in each model was identified. That model involved 29 variables, with 2 hidden neurons.

5 Fold Training Procedure

After selecting the 29 variables and a two hidden neuron neural network architecture, a five fold training and analysis was performed to determine a training regimen that assured generality. Five samples of the training data were developed by generating hundreds of potential splits of the data. Each split with five segments was analyzed and ranked for the consistency between each segment with respect to the percentage of shelters and non-shelters in each segment. The actual split selected for the five fold analysis had nearly identical numbers of shelters and non-shelters in each of the five segments.

The five splits provide for five separate training and analysis sets. A 29 variable, two hidden neuron was trained for each set. Each training episode was carefully observed to determine the number of presentations of taxpayer data that provided optimum results. The number of training iterations was consistent between the five training sets. After determining the optimum training regimen, all of the data was utilized in a final training. That data was then analyzed for performance. The table below shows that the performance of the final model is consistent with the performance of the five training sets.

Model Performance

	Training Segments	Test Segments	KS
Set 1	ABCD	E	0.24
Set 2	BCDE	A	0.31
Set 3	CDEA	B	0.38
Set 4	DEAB	C	0.38
Set 5	EABC	D	0.38
Final	ABCDE	ABCDE	0.39

Figure 2 - Five Fold Analysis

The following diagram shows the detailed KS analysis for the final model:

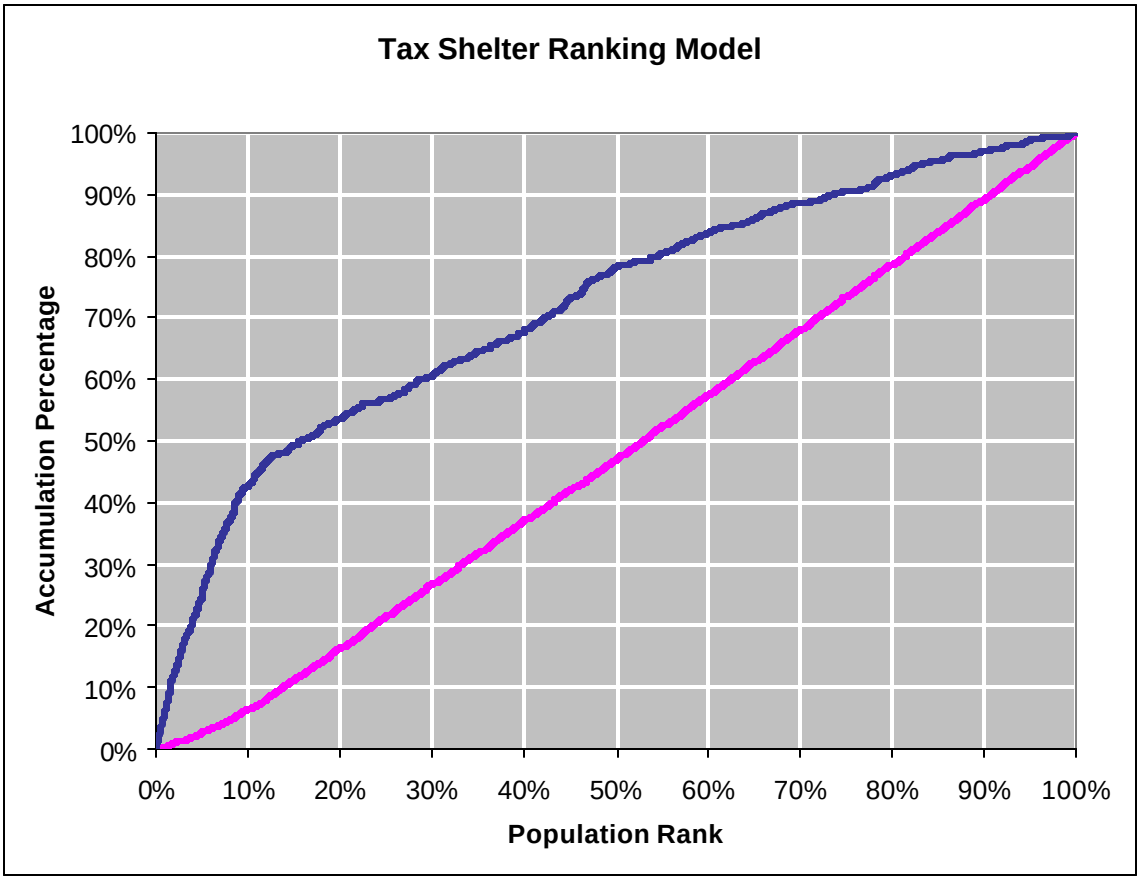

Figure 3 - Final Ranking Model KS Graph

Segment Performance Analysis

The performance of the Ranking Model was studied on specific grouping of asset code and CIC/IC status. The performance in the 225 asset group is attractive for both IC and CIC. However, the asset class 223 does not perform well. The lower performance in the class is

attributed to the small number of 223 cases in the data base. Only 23 shelters and 18 non shelters were available in our data base for asset class 223.

The following graph illustrates the strength of the model, expressed as an odds ratio, compared to random selection at any percentile of the Ranking model:

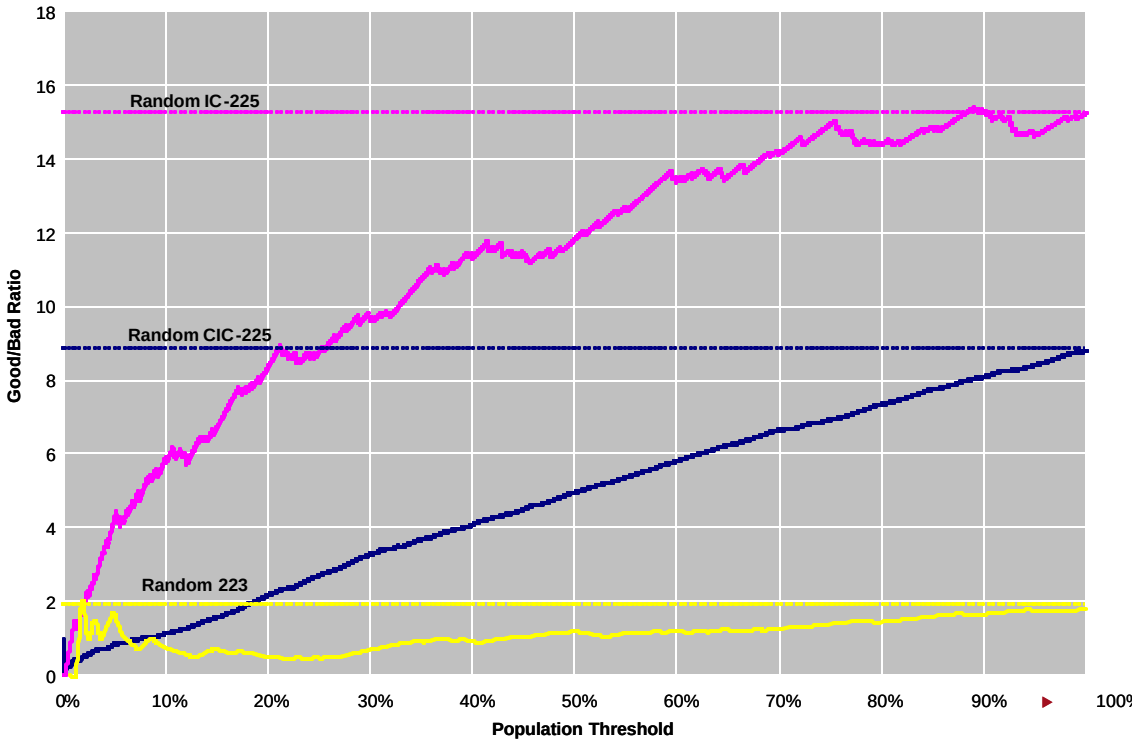

Figure 4 - Odds Ratios for Ranking Model

The vertical distance between the random lines for each segment and the plotted curve for the model shows the lift available from the model for each studied segment.

Conclusions and Recommendations

The Ranking Model developed in this project is very strong in the upper quintile and decile and can be the basis of a productive workload selection process. It shows the strength of combining multiple years of taxpayer data and financial data.

Adding the disclosures to the model, in addition to the detected shelters has provided a dramatic improvement in performance compared the previous project, even thought the project eliminates COLIs.

We recommend that the Ranking Model be applied to the 2000 and 2001 data as soon as it is available.