

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

The Honors Program
Senior Capstone Project
Amber Torrey
Faculty Advisor: Alan Olinsky
April 2008

Table of Contents

Abstract	1
Introduction	2
Background	2
Prior IRS Methods	3
Current IRS Methods	5
2002 National Research Program.....	5
2007 National Research Program.....	6
Unreported Income	7
Effectiveness	7
History of Unreported Income	8
Current Audit Rates.....	9
Taxpayer Compliance and the Tax Gap.....	9
Tax Areas Most Likely to be Included in the DIF	11
Discriminant Analysis Methodology	12
Assumptions and Limitations.....	12
Previous Independent Study on DIF by Amir Aczel	13
Data Mining as an Alternative	14
Decision Trees.....	15
Titanic Example Using a Decision Tree	16
Regression.....	17
Neural Networks	18
Comparing the Methods.....	18
Credit Data Using a Decision Tree, Regression and a Neural Network	18
Credit Data Using Discriminant Analysis.....	19
Discriminant Analysis versus Decision Tree	19
Comparing the Results from all Four Methods.....	20
Appendices	22
Appendix A – Chart of what Items make up the Tax Gap.....	23
Appendix B - Credit Data Decision Tree.....	24
Appendix C – Diagram of Data Mining.....	25
Appendix D - Credit data discriminant analysis output	26
References	31

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

ABSTRACT

This paper discusses past and current methods the IRS uses to determine which individual income tax returns to audit. The IRS currently uses the discriminant function to give all individual tax returns two scores; one based on whether it should be audited or not and one based on if the return is likely to have unreported income. The discriminant function is determined by the IRS's National Research Program, which takes a sample of returns and ensures their accuracy. Previously, the function was determined by the IRS's Taxpayer Compliance Measurement Program. However, this was too burdensome and time consuming for taxpayers. The data mining techniques of decision trees, regression, and neural networks were researched to determine if the IRS should change its method. Unfortunately IRS tax data were not obtainable due to their confidentiality; therefore credit data from a German bank was used to compare discriminant analysis results to the three new methods. All of the methods were run to predict creditworthiness and were compared based on misclassification rates. The neural network had the best classification rate closely followed by regression, the decision tree, and then discriminant analysis. Since this comparison is not based on IRS tax data, no conclusion can be made whether the IRS should change its method or not, but because all methods had very close classification rates, it would be worthwhile for the IRS to look into them.

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

INTRODUCTION

The IRS is responsible for ensuring that all individuals and businesses accurately report their income and deductions and pay their tax liabilities. In fiscal year 2006, a total of 228,145,029 income tax returns were filed, of which 133,917,068 were individual returns (SOI Tax Stats, 2007). In order to maximize its efficiency, the IRS needs some type of method to help ensure all tax returns are accurately stated. The IRS currently uses a statistical method called the discriminant function to decide if a tax return is accurate and if it should be audited. It is possible for other statistical methods to work just as well if not better than the discriminant function.

Background

“The IRS’s mission is to provide America’s taxpayers top quality service by helping them understand and meet their tax responsibilities and by applying the tax law with integrity and fairness to all.” They have five guiding principles which include to, “understand and solve problems from taxpayers’ point of view; enable IRS managers to be accountable to taxpayers; use balanced measures of performance to measure taxpayer satisfaction business results, and our employees’ satisfaction; foster open, honest communications; and insist on total integrity.” (The IRS Mission)

There are three main reasons why the IRS measures taxpayers’ compliance. The first reason is to determine the IRS’ progress toward meeting its mission and goals. Measuring taxpayers’ compliance helps “decision-makers to size potential compliance problems and to make strategic resource decisions accordingly,” (Brown & Mazur, 2002, p.2). Secondly, the IRS can figure out what taxpayers are having problems with and can redesign forms, provide education about tax laws, or propose law changes. Lastly, compliance measures can be used to help determine which tax returns should be selected for an audit or other follow-up. Knowing where problems exist make it much easier to disburse resources. (Brown & Mazur, 2002)

Compliance with the tax system can be measured based on payments, filing, and reporting. Payment compliance is the proportion of tax liability that is both reported and remitted in a timely manner. This measure of compliance was at about 98 percent in the years prior to

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

2002. Filing compliance is the proportion of timely filed returns to the number of required returns. The IRS uses the Consumer Population Survey to estimate the number of individuals that are required to file a return and compares that number to the actual number of individuals who filed a return. In the years prior to 2002 about 91 percent of individuals required to file a return actually did so in a timely manner. (Brown & Mazur, 2002) In 2003 it was estimated that more than 11 million people are either filing their returns late or not at all, (Brown & Mazur, 2003). Reporting compliance is the proportion of correctly reported tax liability filed in a timely manner. (Brown & Mazur, 2002) This type of compliance will continue to be discussed further in this paper.

PRIOR IRS METHODS

In 1962 the IRS used a computer for the first time ever to select tax returns to audit. No specific methods were used with the computer at this point and it did prove to work somewhat better than manually selecting returns to be audited. It is assumed the IRS used the computer to randomly select returns for an audit. Unfortunately, it was not completely efficient in recognizing all returns with a high audit potential. (Hunter & Nelson, 1996) The IRS was still in need of an efficient method for selecting tax returns to audit.

In order to help itself develop efficient methods for selecting tax returns to audit, the IRS instituted its Taxpayer Compliance Measurement Program (TCMP) in 1964. It began by examining the 1963 tax year's returns with detailed field audits. The IRS randomly selected individual tax returns, statistically representing the taxpaying public, and went through each of the returns, line by line to ensure accuracy. The IRS would do this procedure about every three years. The TCMP also helped the IRS to estimate overall compliance. (Brown & Mazur, 2002)

One such method that was developed as a way to select returns to audit was through the use of the discriminant function (DIF), which was first used in 1969. The DIF is a computerized, statistical method of selecting individual and corporate returns to audit. They are selected by specific tax criteria that are weighted for probability of evasion or error. (Hunter & Nelson,

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

1996) The tax criteria used was based on the TCMP data. If the TCMP data showed there was one tax item most individuals made an error with, then that tax item would appear in the DIF. The incoming individual tax returns were and still are sent to the IRS National Computer Center where they are analyzed by the discriminant function and given a DIF score. The higher the score is, the higher the probability of misstatement; therefore the higher the chance of being selected for an audit. (Daily, 1999)

At first, only some of income classes were reviewed by the DIF. More income classes were added in 1970 and additional refinements were made to the DIF system in the years to follow. The IRS was also able to improve the audit process with the use of computerized matching of third party documents and computer scanning for mathematical accuracy during the 1980s. (Hunter & Nelson, 1996)

The development and use of the TCMP and DIF enabled the IRS to efficiently use its resources and focus on returns with a high chance of leading to more taxable income. This is measurable in two ways, the first being the percent change of audits resulting in “no tax change.” (Hunter & Nelson, 1996) Before the development of the DIF the outcome of about half of all audits was “no tax change” which was a huge waste of both the IRS’ and taxpayers’ resources. After the development, only about one fifth of all audits resulted in “no tax change.” (Brown & Mazur, 2002) The second measurable aspect is the additional taxes and penalties assessed. Before the development of the TCMP the average additional taxes and penalties assessed was 700 dollars per audit. Right after the development, the average went up to a little more than 1000 dollars per audit. (Hunter & Nelson, 1996)

Observations from the use of the TCMP led to many policy changes. One example is that it was found that many taxpayers were misreporting their number of dependents; therefore taxpayers were then required to report a taxpayer identification number for each dependent claimed. The year after this requirement was put into effect, about 5 million dependents disappeared. (Brown & Mazur, 2002)

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

The last tax year of this procedure was 1988. The IRS attempted multiple times to restart the program but never succeeded due to the objections of the Treasury, the White House, and the Congress. They were all concerned about the burden on taxpayers of having to confirm each line of their tax return to the IRS, the amount of IRS resources devoted to the TCMP, not knowing how much it cost versus the benefit in additional tax dollars it brought in, “and a general negative feeling towards the IRS on the part of Congress and a substantial portion of the American public,” (Brown & Mazur, 2002).

CURRENT IRS METHODS

2002 National Research Program

The National Research Program (NRP) began in the fall of 2002 to take the place of the TCMP. Since the last TCMP audits looked at the 1988 tax year’s returns, the economy and tax laws had changed enough in those thirteen years for the IRS to need new, updated tax information. TCMP audits were often criticized due to the amount of time and scrutiny they required since the IRS literally went line by line through the tax return. The NRP was developed to be a less intrusive means of getting tax compliance information from individuals. Without the updated data, the IRS would not be able to efficiently audit nor decrease the amount of audits that result in “no tax change,” (IRS Sets New Audit Priorities, 2002). Since the last set of TCMP audits, the number of “no tax change” audits had dramatically risen. (Brown & Mazur, 2002) New data would also allow the IRS to update its estimate of the tax gap and to “provide information about the number of US taxpayers who voluntarily file their returns in a timely manner,” (Brown & Mazur, 2003).

The NRP audited fewer than 50,000 individual returns out of the 132 million filed in 2001. About 8,000 of the returns audited were done so by verifying the information the IRS already had. There was no communication with the taxpayer or any third parties. (IRS Sets New Audit Priorities, 2002) About 9,000 of the audits were done through the mail with limited questions, (Pilla, 2002). For roughly 30,000 of the audits, the IRS collected information from agency records and only focused on specific parts of the return. Finally, about 2,000 of the audits were done by examining each line of the return. Fortunately, the taxpayers did not

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

have to provide confirmation of each line as was done with the TCMP audits. (IRS Sets New Audit Priorities, 2002)

The NRP tries to acquire high quality data while minimizing the burden on taxpayers. The IRS is trying to use its own data as much as possible, called case-building, in order to limit contact with the taxpayers. With the use of case-building, the IRS should be able to determine if a taxpayer is complying without having to contact the individual. Case-building involves using “income tax returns from the current and prior years, information documents (such as W2s or 1099s), currency and banking transaction reports, the dependent database, and returns from related entities such as closely-held corporations or partnerships.” This should allow the IRS to focus on non-third party reported items since it can confirm third party reported items with the actual third party. (Brown & Mazur, 2002)

2007 National Research Program

The IRS planned to start a second round of audits under their national research program (NRP) in October 2007 for the 2006 tax year. The new study will examine 13,000 individual tax returns selected over time so that returns filed on extension may also be included in the study. This “will be the first of an ongoing series of annual individual studies using an innovative multi-year rolling methodology.” (IRS to Launch New National Research Program Study in October, 2007) Similar sample sizes will be used in the years to come and the results will be combined over rolling three year periods. The current sample includes individuals from all income levels and those involved with farm and sole proprietor business activities. Most of the individuals involved in the study will only have to confirm a few specific lines of their return, versus every line that was required in the TCMP, while some will not be contacted at all as long as the IRS has third party data to confirm their return. (IRS Updates National Research Program for Individuals) The IRS probably chose to combine the results over three year periods because it only has three years to audit a return; therefore after the three years is up, the data is no longer relevant.

This will be a much more efficient way of conducting research since it will be done on a continuous basis versus every couple of years. It will enable the IRS to maintain up to date

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

estimates of the tax gap and of taxpayer compliance. Tax laws, the economy and patterns of noncompliance are continuously changing; therefore it is vital for the IRS to maintain current data in order to efficiently target individuals for audits and “to improve the detection of underreported income and overstated deductions and credits. The data also enables the IRS to audit fewer taxpayers with accurate tax returns, which lessens the burden on compliant taxpayers,” (IRS Updates National Research Program for Individuals, 2007).

The IRS is also in the final stages of an S-corporation compliance research project. Since profits and losses flow down to the stockholders and are reported on their individual returns, this project will also help to improve the individual income tax gap. (IRS Updates National Research Program for Individuals, 2007)

The IRS still applies the NRP data to its discriminant function. It then puts new returns coming in through the model and the top ten percent of returns with the highest DIF scores are initially selected by a computer for an audit. IRS examiners then look at the returns and choose about ten percent to conduct an audit on, this equates to only about one percent of all tax returns filed actually being audited. (Daily, 1999) It is unsure what techniques the examiners use to decide which returns to actually audit.

Unreported Income

Unreported income represents the largest component of the tax gap and as of the fall of 2002, the IRS now has a direct technique to select the returns with the highest probability of having unreported income. The IRS now uses the Unreported Income Discriminant Index Formula (UI DIF) to give each individual tax return a score based on its probability of having unreported income. Prior to the UI DIF, the IRS only had indirect examination methods to identify unreported income. The UI DIF score will be given along with the regular DIF score. (IRS Sets New Audit Priorities)

Effectiveness

The IRS’s methods have proved to be very effective. Additional revenue received due to enforcement activities in the fiscal year 2007 amounted to \$59.2 billion. This amount is up

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

from fiscal year 2006's of \$48.7 billion and 2002's of \$34.1 billion. (Fiscal Year 2007 Enforcement and Services Results, 2008)

HISTORY OF UNREPORTED INCOME

The Office of Examination Planning and Research originally came up with the UI-DIF using the 1974 tax year's TCMP data. It was later updated with 1985's TCMP data and further revised with the TCMP data from 1988. (Cyr, Eckhardt, Sandoval, & Halldorson, 2002) It is believed this was just research for potential use in the future. The UI-DIF was first used as a means to select returns to audit in 2002.

Before the IRS could use UI DIF scores to select returns to examine, it had to prove that the scores were actual indicators of unreported income. The IRS Restructuring and Reform Act of 1998 states "the Secretary shall not use financial status or economic reality examination techniques to determine the existence of unreported income of any taxpayer unless the Secretary has a reasonable indication that there is a likelihood of such unreported income," (Office of Research: Research, Analysis, & Statistics, 2002). The IRS believed that high UI scores may be reasonable enough indication of the likelihood of unreported income for it to legally look into the tax returns, but first it had to be proven. The IRS proved this by testing the UI scores of 400 returns from the 1988 TCMP dataset. It did not want to wait for new data because that would have taken a couple of years to complete and evaluate. Fifty returns were classified within each of the eight activity codes, such as non-business, non-farm business, or farm business, which were all used in formulating the UI-DIF. Half of the 50 to be classified were previously given the top two percent of UI scores and the other half had been given the bottom 50 percent UI scores. The classifiers had no knowledge of the UI scores and were given the task of answering the question: "Should the return be examined for unreported income?," 'Yes' or 'No.' Two conditions were necessary to certify the UI scores, first a strong relationship between yes and high UI scores, and second, a strong relationship between no and low UI scores. Results showed that 188 of the 200 returns with high UI scores received a 'yes' to the question asked and 160 of the 200 returns with low UI scores received a 'no' to the question asked. The relationships were very high; therefore the UI-DIF

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

for each of the eight activity codes was certified and high UI scores could be used by the IRS to select returns to audit. (Office of Research: Research, Analysis, & Statistics, 2002)

A second question that arose while the testing was being conducted was if the returns with the high UI DIF scores also had high DIF scores. This was considered since both scores were developed using the discriminant function, only the criterion for each score is different. Of the 200 returns tested that had the highest two percent of UI DIF scores, only 39 had extremely high DIF scores. On the opposite end, of the 200 returns with the lowest 50 percent of UI DIF scores, seven had extremely high DIF scores. It was found that less than ten percent of returns have high scores in both the DIF and UI DIF. Therefore the UI DIF selects different tax returns to audit and should be used in addition to the DIF. (Cyr, Eckhardt, Sandoval, & Halldorson, 2002)

CURRENT AUDIT RATES

Individual tax audits reached a ten year high in 2007 as the IRS was feeling additional pressure from Congress to reduce the \$290 billion tax gap, (Herman, 2008). In 2007, 9.25 percent of taxpayers with incomes over \$1,000,000 were audited for the 2006 tax year, or roughly one out of eleven, up from 6.3 percent the previous year, (Don't Mess with Taxes, 2008). That came out to be 31,382 returns, up from 17,015 (Herman, 2008). Additionally, 2.87 percent of taxpayers, or 113,105, with incomes between \$200,000 and \$1,000,000 were audited up from 2.57 percent, or 87,558, in 2006. (Don't Mess With Taxes, 2008) Still, for most taxpayers, the chance of getting audited remains at about 1%, (Herman, 2008). At the bottom of the income brackets, only 0.93 percent of taxpayers making less than \$100,000 were audited, up from 0.89 percent, (Don't Mess With Taxes, 2008).

TAXPAYER COMPLIANCE AND THE TAX GAP

The tax gap is the difference between the amount of taxes that should be collected and the actual amount collected. Most Americans do pay their taxes on time, however there is still \$312 to \$353 billion that goes uncollected annually. The overall tax gap for 2001 was \$345 billion, of which \$55 billion was collected through late payments and IRS enforcement

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

activities, netting out to a \$290 billion tax gap. (IRS Updates Tax Gap Estimates) This estimate is up from the 1998 tax gap estimate of \$280 billion (Brown & Mazur, 2002). Taxpayer compliance is one of the reasons why the NRP began, for the IRS to get a better estimate. The NRP showed that more than 80 percent of the tax gap comes from underreporting taxes. Nonfiling and underpaying taxes make up the other 20 percent of the tax gap. (IRS Updates Tax Gap Estimates)

Compliance is the highest when there is a third-party reporting, such as an employer reporting their employees' income (IRS Updates Tax Gap Estimates). Based on the Individual Income Tax Underreporting Gap Estimates, Tax Year 2001 chart in Appendix A, wages, salaries, and tips are only misreported by one percent and both interest and dividend income are misreported by four percent. All three of these categories are reported to the government by third parties, making compliance much higher. Items that are not reported by third parties are much more likely to be misstated, such as form 4797 income, other income, farm income, rents and royalties, and nonfarm proprietor income. "IRS research indicates much of the tax-noncompliance is committed by self-employed workers, such as consultants and small-business owners, whose taxes aren't withheld from their pay and whose income isn't reported separately to the government," (Herman, 2008).

Linda Stiff, the IRS's acting commissioner said in an interview, that this year's audits will continue to focus on high income individuals and in addition the IRS will be strengthening its focus on abusive tax shelters. Tax shelters are created to avoid or reduce taxes. The IRS has also increased audits of individuals in partnerships and S corporations. (Herman, 2008) Both of these types of business entities involve passing the profits or losses to either the partners or the shareholders to be reported on the individual's return.

Foreign athletes and entertainers are also being targeted by the IRS. Officials say they have discovered significant noncompliance by such individuals and have about 60 open cases dealing with them. The cases deal with both prize money won by the athletes competing in US events and product-endorsement money. Since many of these athletes and entertainers

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

perform in many different countries each year, this is a very complicated issue. (Herman, 2008)

The easiest way to attract the IRS' attention is to claim there is no law that you have to pay federal income taxes or file a return. The IRS refers to these claims as "frivolous" arguments, (Herman, 2008). There are many other frivolous claims such as misinterpreting the US Constitution's 9th Amendment with regards to objecting to military spending and claims that taxes are only owed by people with a fiduciary relationship to the United States or the IRS, (IRS Names Four New Frivolous Claims to Avoid). Courts always reject these cases and the maximum penalty is \$5,000. The US Tax Court judges have inflicted penalties of up to \$25,000 on some individuals who persisted. (Herman, 2008)

TAX AREAS MOST LIKELY TO BE INCLUDED IN THE DIF

Douglas Gross, a CPA, believes that historically the IRS looks closely at home office deductions and that form 2106 issues and legal expenses are the most common 2% deductions claimed that it will challenge, (personal communication, January 23, 2008). The earned income tax credit is also a big audit trigger because 36.5 percent of the total number of returns audited in 2007 from the 2006 tax year were selected because they claimed the credit (Federal Taxes Weekly Alert, 2008).

Brynes (2002) believes that if a taxpayer reports higher deductions, that are unusually higher than their averages from the past years, the taxpayer is more likely to be selected for an audit. Another way to trigger an audit is if a taxpayer's reported income does not match the attached W-2, 1099 etc. The IRS also receives copies of these forms and will match the taxpayer's attached forms and reported income to the copies the IRS already has. Schedule Cs are a big audit trigger. Since there is no third party reporting, it is easier to misreport business income and expenses. The IRS will be looking for receipts. Brynes agrees with Gross that taking the home office deduction is an audit trigger because many taxpayers have wrongly taken it in the past. Lastly, if a taxpayer is supposed to be receiving a bigger tax refund than they have in the past, the IRS will check to make sure the taxpayer actually deserves this larger tax refund.

DISCRIMINANT ANALYSIS METHODOLOGY

Discriminant analysis was developed in the 1930s in order to separate data into two groups. From there it was further developed to separate data into multiple groups and to describe differences between the groups after a multivariate analysis of variance. (Mertler & Vannatta, 2005) It is “used to analyze relationships between a non-metric dependent variable and metric or dichotomous independent variables,” (Schwab, 2006). A metric variable is one that is represented by a number. A dichotomous variable is one that has two possibilities. A dependent variable is the one that is being predicted while the independent variables are the predictors.

If the discriminant analysis is being used to predict group membership versus to describe differences between the groups, then “the goal is to determine dimensions that serve as the basis for reliably-and accurately-classifying subjects into groups,” (Mertler & Vannatta, 2005). Once the model is formed, new data is applied and given a score to predict which group the case belongs to. The score is calculated by multiplying assigned weights by the independent variables. The weights are determined by how well each of the variables is at predicting the dependent variable. The scores are standardized, meaning if they are greater than the set boundary, they are members of one group and if they are less than the set boundary, they are members of the other group. (Schwab, 2006)

Assumptions and Limitations

There are four assumptions of discriminant analysis. The first is that observations must be randomly sampled and independent of one another. They also must be from a normal distribution. Third, “the population covariance matrices for the predictor variables in each group must be equal (the assumption of homoscedasticity),” (Mertler & Vannatta, 2005, p.287). Lastly, there must be linear relationships among all pairs of predictors within each group.

It is unsure whether or not the IRS meets all four assumptions. The IRS conducts a random sample of audits on individual returns and they are independent of one another, which satisfies the first assumption. However, it is impossible to prove that the IRS meets the other

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

three assumptions without having any of its data to run and test. It is very likely that the IRS does meet the assumptions based on the amount of data it has; datasets tend to normalize the larger they become.

A limitation of discriminant analysis is that it can be sensitive to sample size. If the sample size is not large relative to the number of variables, “both the standardized coefficients and the correlations are very unstable,” (Stevens, 1992, p.277). The ratio of total sample size to the number of variables should be at least twenty to one or else one should use caution in interpreting the results. For example, if there are three variables used, then there should be at least 60 subjects in order for the discriminant analysis model to be accurate. There were thirteen thousand 2006 individual tax returns audited under the NRP for the IRS to use in their most recent discriminant analysis (IRS to Launch New National Research Program Study in October, 2007). It is not possible to know how many variables the IRS uses in their analysis; however it is safe to assume that they satisfy the sample size requirement. The IRS’s model certainly exceeds the twenty to one ratio and therefore its discriminant analysis model most likely satisfies the assumptions.

PREVIOUS INDEPENDENT STUDY ON DIF BY AMIR ACZEL

A current professor at Boston University, Amir Aczel, conducted a study to estimate the IRS’s DIF in the mid-1990s by using logistic regression and regression trees. Aczel examined 1,289 returns and developed a DIF that will show if the return has no, some, or a high risk of being audited by the IRS. Unfortunately, due to confidentiality and the changes in the economy and tax laws in the past twelve years, Aczel failed to share his data from the returns he examined. However he did publish his results. Aczel found when Schedule A’s itemized deductions are less than 35 percent of adjusted gross income, the taxpayer has virtually no audit risk. When the deductions are between 35 and 44 percent, there is some risk and when they are greater than 44 percent, there is a very high risk of an audit. He also found that when filing a Schedule C, deductions of up to 52 percent of revenues will result in virtually no audit risk, however deductions of more than 63 percent of revenues result in a very high audit risk. Aczel also found what numbers on a Schedule F, for farm income, may result in an audit and

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

a list of other items such as missing schedules or lines left blank that may also result in an audit. (Johnston, 1996)

Aczel argues the discriminant analysis is an out-of-date statistical method that lets the guilty cheaters and tax evaders go free because the DIF does not catch non-filers or taxpayers who should have very low deductions, but cheat and still come in unnoticed. He believes the IRS should use a more modern approach of selecting individual returns to audit. Wayne Thomas, the IRS national compliance research director, responded by saying “‘DIF just works well for us.’ He added that the IRS had repeatedly retained outside consultants to test newer statistical techniques, but had never found one as reliable as its method of discriminant function,” (Johnston, 1996).

DATA MINING AS AN ALTERNATIVE

Data mining, first introduced in the 1990s, is the process of extracting information from large sets of data. It finds patterns and relationships among the data using data analysis tools and techniques to build models. There are two main types of models in data mining. The first is predictive modeling, which creates a model based on data that has a known result and applies it to new data to predict the result. The second type of model is a descriptive model, which describes patterns in data. (Data Mining) Both of these types of models can apply to what the IRS is trying to do. It could use predictive modeling with its NRP audits, which the outcome of whether or not the audit resulted in more tax revenue is known, and then use the model to predict the outcome of all the other tax returns coming in. The IRS could also use descriptive modeling to find the patterns of what different taxpayers report. It already does this to an extent when it finds the averages of what each different income level reports in each specific category of income and deductions and credits. The extent of this paper will focus on the IRS using predictive modeling.

There are two main reasons why data mining would be used, each applies to the IRS. The first is when there is too much data and too little information. (Data Mining) The IRS certainly has too much data to work with, considering each line of a tax return. It may not

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

have too little information since the IRS has been researching compliance since the 1960s, but since the IRS is continuously updating its model, it could always use more information in order to optimize efficiency. The second reason is that there is a need to extract information from and interpret the data, (Data Mining). The IRS certainly has the need to extract useful information from all the tax returns filed each year and to interpret the data so that it can efficiently apply it to the model in order to select the proper returns to audit.

Predictive modeling requires a target variable, what the modeling is trying to predict, and at least one input variable. There are few assumptions to worry about, unlike with discriminant analysis. There are three main types of predictive modeling in data mining that are regularly used. They are decision trees, logistic regression, and neural networks. They will each be discussed further in this paper.

Decision Trees

Decision trees use values of the input variables, arranged hierarchically in an upside-down tree like structure, to predict the target variable. The rules, which transforms measurements into predictions, are represented by nodes and are all connected by lines to keep themselves in order. The root node is the first rule at the top of the tree. All of the rules to follow are called interior nodes. Leaf nodes occur when a node has only one connection. (Georges, 2007)

Predicting the outcome of a new case is very easy, just look at the values of the input variables and follow them through the rules of the tree. This will lead to a single leaf in the tree which tells the prediction for the target variable. While setting up the model, there is the option of stopping it after a specific number of depths. For example, if the model naturally comes up with a depth of six, it can be set to stop at the fourth. This could be done if the model comes up with a better prediction of the target variable at the fourth node versus the sixth. In this case the model is already good at predicting so there is no need to continue through more nodes.

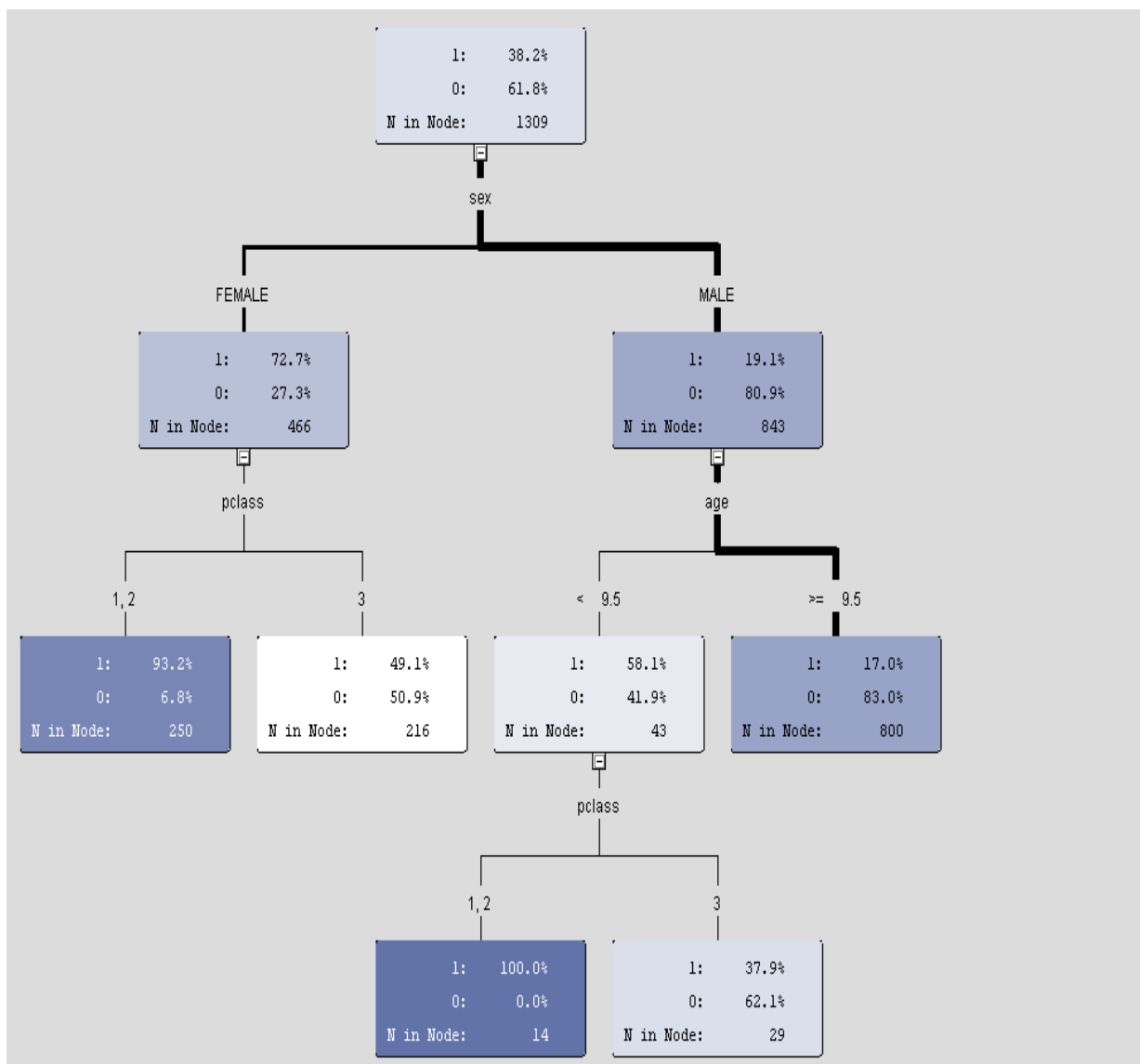
The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

Titanic Example Using a Decision Tree

An extremely easy to understand example of a decision tree is to use the Titanic dataset from Encyclopedia Titanica, which has the target variable of whether or not an individual survived. Neither the data nor decision tree are applicable to predict anything because it is very unlikely that another catastrophe simulating the Titanic would occur, however the data can still be used as an example of how a decision tree works.

In the decision tree below, the very top box shows the actual percentages for each outcome of the target variable where a '1' means the individual survived and a '0' means they did not.



The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

The root node is gender, meaning this is the most important variable for predicting whether or not an individual survived. The boxes underneath both of the decisions from the root node represent the prediction of the target variable. A female had a 72.7 percent survival rate, while males only had a 19.1 percent survival rate. Breaking it down further by continuing to the next node under female, first and second class females had a 93.2 percent survival rate while third class females only had a 49.1 percent survival rate. Continuing with the males, age was the next best predictor of survival; males younger than 9.5 had a 58.1 percent survival rate while males 9.5 and older only had a 17 percent rate of survival. The last node under males younger than 9.5 show that first and second class had a 100 percent survival rate and third class had a 37.9 percent rate of survival. If Titanic was not an event of the past and required future predictions of survival to be made, predictions could easily be made by following this decision tree through and looking at the different rates of survival based on specific characteristics.

Regression

There are both linear and logistic regressions. Regression creates a mathematical formula used to score new data in order to predict the data's outcome. The formula is created by using values of the input variables to fit the target variable. A linear regression is created by taking linear combinations of the input variables to predict the target variable. This type of model usually works best when the target outcome is on an interval measurement scale versus being a binary outcome. (Georges, 2007) Logistic regression on the other hand, can be used when the target variable is binary, since it is used to predict group membership. In some cases it is used instead of discriminant analysis; therefore it could be a potential substitute for the IRS. Logistic regression creates a formula based on values of the input variables to come up with a probability of the outcome being one target variable over the other. For example, if the IRS used logistic regression, the model would predict the probability of whether a tax return will fall into the needs to be audited category or the accept as is category. "Mathematically speaking, logistic regression is based on probabilities, odds, and the logarithm of the odds," (Mertler & Vannatta, 2005). The IRS may not like to use logistic regression because it is very difficult to interpret the equation; it takes a lot of experience, therefore the IRS sticks to discriminant analysis, which is easier to interpret.

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

Neural Networks

A neural network is a data mining technique “modeled after the processes of learning in the cognitive system and the neurological functions of the brain and (is) capable of predicting new observations from other observations after executing a process of so-called learning from existing data,” (StatSoft, 2008). The network represents a pattern in the data.

Unlike with other methods though, the relationships it finds between the variables cannot be stated in a simple equation like with regression. Statisticians refer to neural networks as representing a “black box” approach because no one really knows how the model or relationships within are formed. (StatSoft, 2008) Neural networks are great predictors of almost all target variables based on the inputs. This concerns some statisticians as overfitting can occur when the model has been trained too much to fit the dataset given, which might take away from predicting the outcomes of new datasets. Fortunately there is an optimization algorithm called stopped training to reduce this risk. (Georges, 2007) The IRS probably does not use this type of model because it would not know which characteristics the model is basing its decision on whether or not to audit the tax return.

COMPARING THE METHODS

Credit Data Using a Decision Tree, Regression and a Neural Network

Since tax data from the IRS is confidential and cannot be used, credit data from a German bank provided by Georges (2007) will be used to compare the four statistical methods. Each of the methods was run to predict credit worthiness. The variables are difficult to interpret, however they are not important because misclassification rates will be used to compare each of the methods. A statistical model comes up with a misclassification rate by only using a portion of the data to form the actual model and then by testing the model with the unused data. The misclassification rate is the percent of time the model wrongly predicted the outcome.

SAS’s Enterprise Miner, a data mining package, was used to run a decision tree, regression, and neural network with the credit data. The decision tree can be found in Appendix B. The

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

misclassification rate is 15.17 percent. The misclassification rate using data mining regression is 14.3 percent. The misclassification rate using a neural network is 13.9 percent.

Using the same data mining package, all three models and a new set of data could be connected to a model comparison. A diagram of this can be found in Appendix C. This would compare the three models and use the best of the three to score the new data. The data mining program chose the neural network as the best fit to the data because it has the lowest misclassification rate.

Credit Data Using Discriminant Analysis

The same German bank credit data (Georges, 2007) were run using SPSS, a statistical program, to form a discriminant analysis. Assumptions and limitations of this data will not be discussed since this is not data the IRS would actually use and is only being used as an example of the method. The output from running this model can be found in Appendix D. Using stepwise discriminant analysis, classification results show that 83.7 percent of the cases were correctly classified; therefore 16.3 percent were misclassified, which can be compared to the three data mining results. Using a neural network still has the best misclassification rate of 13.9 percent.

There are 13 variables that make it into the model, they are percent satisfactory to total trade lines, number trade lines 75 percent utilized, time since first trade line, number trade lines 30 or 60 days 24 months, number bad debt plus public derogatories, number trade lines currently satisfactory, number finance inquires 24 months, number trade lines opened three months, total open trade lines, percent trade line balance to high credit, percent trade lines open 24 months, number public derogatories, and time since last trade line.

Discriminant Analysis versus Decision Tree

Since the neural network and regression use multiple combinations of the input variables to predict the target variable, it is simpler to just compare the variables used in the discriminant analysis and decision tree. As stated before, the interpretation of the variables used in the models are not important, however there are some differences concerning the variables used

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

in each of the models. Only two of the variables found in the decision tree are not in this discriminant analysis, which are number trade lines 60 days or worse and number trade lines 50 percent utilized. The variable number trade lines 60 days or worse is the root node in the decision tree, meaning it is most important in predicting group membership and it is not a part of the discriminant analysis. This is not a negative aspect, just merely an observation. The rest of the variables found in the decision tree are also in the discriminant analysis.

Based on the discriminant analysis the variable percent trade line balance to high credit is lower when the case belongs to group zero and higher when belonging to group one. The decision tree agrees with this because when this variable is less than 1.01575 the case has a 93.1 percent chance of belonging to group zero. When this variable is greater than or equal to 1.01575, it has a 66.7 percent chance of belonging to group one. Both models are in agreement this way with all of their overlapping variables except for part of two of them. The discriminant analysis believes the variable percent satisfactory to total trade lines is greater when it belongs to group zero and is lower when belonging to group one. The decision tree agrees that when it is greater it has a higher probability of belonging to group zero, however, when it is lower it still has a higher chance of belonging to group zero versus group one. The other discrepancy is with the variable percent trade line balance to high credit after the node percent satisfactory to total trade lines. Again, the discriminant analysis believes when the variable is higher, it belongs to group one. This decision in the tree believes when the variable is higher, it has a greater chance of belonging to group zero. These are the type of discrepancies that could occur with the IRS's data. Using a decision tree could mean that different tax item variables are more important in predicting whether or not the return should be audited than the current variables used in the discriminant analysis. Also, the way the decision tree is set up, could result in different values of the variables predicting group membership versus the values that currently predict group membership in the discriminant analysis.

Comparing the Results from all Four Methods

Comparing misclassification rates, the neural network had the lowest at 13.9 percent, followed by regression with 14.3 percent, the decision tree with 15.17 percent and then

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

discriminant function with 16.3 percent. Nothing can be concluded pertaining to whether the IRS should continue to use the discriminant function or look into using a data mining method since IRS data was not available. These results do show that all four methods result in very close misclassification rates; therefore the IRS should look into the data mining methods to determine if one would be a better fit than its current discriminant function.

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

APPENDICES

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

Appendix A – Chart of what Items make up the Tax Gap

Individual Income Tax Underreporting Gap Estimates, Tax Year 2001

<i>Type of Income or Offset</i>	<i>Tax Gap (\$B)</i>	<i>NMP [†]</i>
Total Underreporting Gap	197	18%
Underreported Income	166	11%
Non-Business Income	56	4%
Wages, salaries, tips	10	1%
Interest income	2	4%
Dividend income	1	4%
State income tax refunds	1	12%
Alimony income	*	7%
Pensions & annuities	4	4%
Unemployment Compensation	*	11%
Social Security benefits	1	6%
Capital gains	11	12%
Form 4797 income	3	64%
Other income	23	64%
Business Income	109	43%
Nonfarm proprietor income	68	57%
Farm income	6	72%
Rents & royalties	13	51%
Partnership, S-Corp, Estate & Trust, etc.	22	18%
Overreported Offsets to Income	15	4%
Adjustments	-3	-21%
SE Tax deduction	-4	-51%
All other adjustments	1	6%
Deductions	14	5%
Exemptions	4	5%
Credits	17	26%
Net Math Errors (non-EITC)	*	

[†] NMP = Net Misreporting Percentage

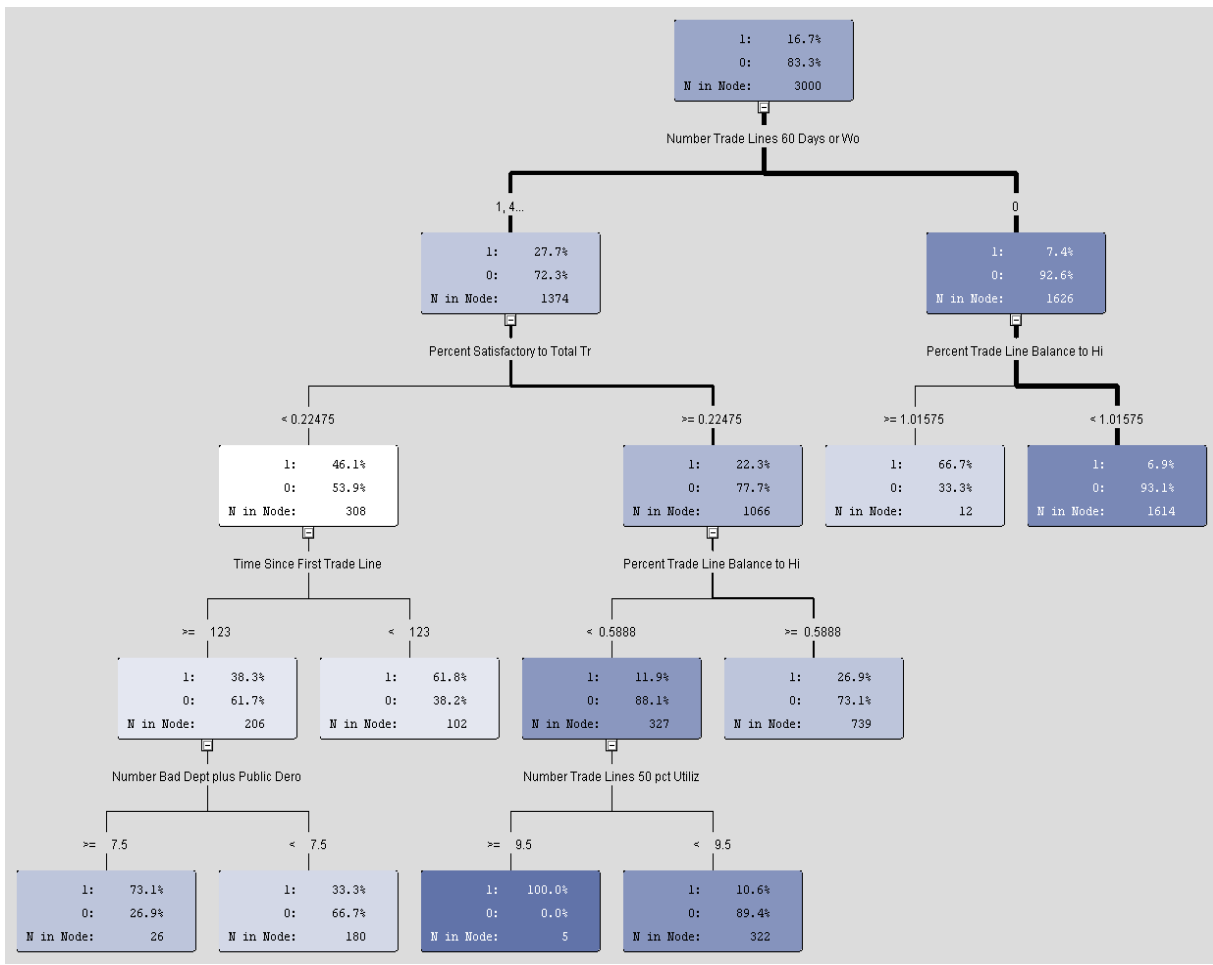
* Less than \$0.5 billion.

Source: IRS Updates Tax Gap Estimates

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

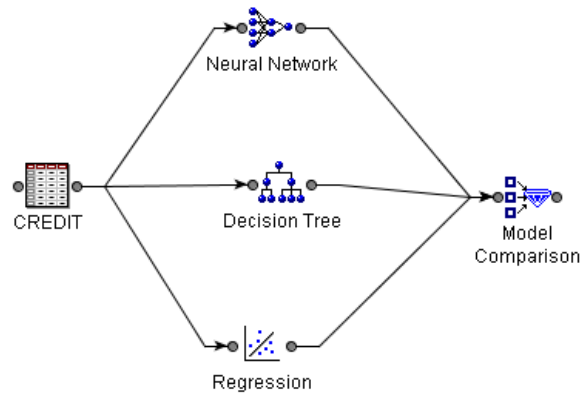
Appendix B - Credit Data Decision Tree



The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

Appendix C – Diagram of Data Mining



The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

Appendix D - Credit data discriminant analysis output

Classification Results(b,c)

			Predicted Group Membership		Total
			0	1	0
Original	Count	0	2331	80	2411
		1	378	108	486
	%	0	96.7	3.3	100.0
		1	77.8	22.2	100.0
Cross-validated(a)	Count	0	2324	87	2411
		1	384	102	486
	%	0	96.4	3.6	100.0
		1	79.0	21.0	100.0

a Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b 84.2% of original grouped cases correctly classified.

c 83.7% of cross-validated grouped cases correctly classified.

Variables Entered/Removed(a,b,c,d)

Step	Entered	Min. D Squared					
	Statistic	Exact F				Statistic	Between Groups
	Statistic	df2	Sig.	Statistic	df1	df2	Sig.
1	Percent Satisfactor y to Total Trade Lines	.475	0 and 1	185.074	1	2726.000	7.89E-041
2	Number Trade Lines 75 pct Utilized	.672	0 and 1	130.851	2	2725.000	5.47E-055
3	Time Since First Trade Line	.794	0 and 1	103.091	3	2724.000	3.27E-063
4	Number Trade Lines 30 or 60 Days 24 Months	.936	0 and 1	91.080	4	2723.000	9.12E-073

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

5	Number Bad Dept plus Public Derogatories	1.038	0 and 1	80.750	5	2722.000	3.15E-079
6	Number Trade Lines Currently Satisfactory	1.114	0 and 1	72.233	6	2721.000	8.31E-084
7	Number Finance Inquires 24 Months	1.165	0 and 1	64.721	7	2720.000	1.63E-086
8	Number Trade Lines Opened 3 Months	1.197	0 and 1	58.162	8	2719.000	7.19E-088
9	Total Open Trade Lines	1.229	0 and 1	53.077	9	2718.000	2.89E-089
10	Percent Trade Line Balance to High Credit	1.274	0 and 1	49.476	10	2717.000	1.64E-091
11	Percent Trade Lines Open 24 Months	1.302	0 and 1	45.946	11	2716.000	1.32E-092
12	Number Public Derogatories	1.328	0 and 1	42.957	12	2715.000	1.33E-093
13	Time Since Last Trade Line	1.341	0 and 1	40.013	13	2714.000	1.20E-093

At each step, the variable that maximizes the Mahalanobis distance between the two closest groups is entered.

a Maximum number of steps is 56.

b Maximum significance of F to enter is .05.

c Minimum significance of F to remove is .10.

d F level, tolerance, or VIN insufficient for further computation.

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

Group Statistics

TARGET		Mean	Std. Deviation	Valid N (listwise)	
		Unweighted	Weighted	Unweighted	Weighted
0	Number Public Derogatories	1.3199	2.55377	2257	2257.000
	Number Collections	.7630	2.00798	2257	2257.000
	Bankruptcy Indicator	.1520	.35907	2257	2257.000
	Number Inquiries 6 Months	3.1639	3.40807	2257	2257.000
	Time Since Last Inquiry	3.1830	4.68812	2257	2257.000
	Number Finance Inquires 24 Months	3.5560	4.32937	2257	2257.000
	Time Since First Trade Line	173.5534	91.43009	2257	2257.000
	Time Since Last Trade Line	10.3930	13.75914	2257	2257.000
	Number Trade Lines Opened 3 Months	.3035	.61205	2257	2257.000
	Number Trade Lines Opened 12 Months	1.9318	1.93872	2257	2257.000
	Number Trade Lines Opened 24 Months	4.1276	3.41339	2257	2257.000
	Total Open Trade Lines	8.2782	5.43193	2257	2257.000
	Total Balance All Trade Lines	20911.8028	20105.75993	2257	2257.000
	Total High Credit All Trade Lines	32871.8516	29828.03716	2257	2257.000
	Number Trade Lines Currently Satisfactory	14.3332	8.92585	2257	2257.000
	Number Trade Lines Currently 60 Days or Worse	1.3181	2.52642	2257	2257.000
	Number Trade Lines Bad Debt 24 Months	.4710	1.18021	2257	2257.000
	Number Trade Lines 75 pct Utilized	3.0656	2.58339	2257	2257.000
	Number Trade Lines 50 pct Utilized	4.0470	3.08055	2257	2257.000
	Percent Trade Line Balance to High Credit	.6454	.24439	2257	2257.000
	Percent Satisfactory to Total Trade Lines	.5440	.21869	2257	2257.000
	Number Trade Lines 30 or 60 Days 24 Months	.6176	1.07127	2257	2257.000
	Number Trade Lines 90+ 24 Months	.6735	1.42675	2257	2257.000
	Number Trade Lines 60 Days or Worse Ever	2.2822	3.15914	2257	2257.000

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

1	Percent Trade Lines Open	.4995	.19794	2257	2257.000
	Number Bad Dept plus Public Derogatories	1.1856	2.26827	2257	2257.000
	Number Trade Lines 60 Days or Worse 24 Months	.8826	1.61825	2257	2257.000
	Percent Trade Lines Open 24 Months	.5649	.44010	2257	2257.000
	Number Public Derogatories	1.8875	2.89131	471	471.000
	Number Collections	1.2527	2.36145	471	471.000
	Bankruptcy Indicator	.1741	.37960	471	471.000
	Number Inquiries 6 Months	4.1040	3.79639	471	471.000
	Time Since Last Inquiry	2.7771	4.46299	471	471.000
	Number Finance Inquires 24 Months	5.0722	5.39494	471	471.000
	Time Since First Trade Line	156.4331	91.44083	471	471.000
	Time Since Last Trade Line	11.8089	12.74503	471	471.000
	Number Trade Lines Opened 3 Months	.2357	.52357	471	471.000
	Number Trade Lines Opened 12 Months	1.8535	2.06124	471	471.000
	Number Trade Lines Opened 24 Months	3.9894	3.52588	471	471.000
	Total Open Trade Lines	7.5860	5.30080	471	471.000
	Total Balance All Trade Lines	20933.5966	18749.40246	471	471.000
	Total High Credit All Trade Lines	28579.5966	24976.49296	471	471.000
	Number Trade Lines Currently Satisfactory	12.2611	8.82632	471	471.000
	Number Trade Lines Currently 60 Days or Worse	2.7622	3.91597	471	471.000
	Number Trade Lines Bad Debt 24 Months	1.0828	1.81470	471	471.000
	Number Trade Lines 75 pct Utilized	3.7028	2.81424	471	471.000
	Number Trade Lines 50 pct Utilized	4.5945	3.38506	471	471.000
	Percent Trade Line Balance to High Credit	.7564	.27129	471	471.000
	Percent Satisfactory to Total Trade Lines	.3908	.23873	471	471.000
	Number Trade Lines 30 or 60 Days 24 Months	1.3163	1.31429	471	471.000

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

Number Trade Lines 90+ 24 Months	1.5924	2.23213	471	471.000
Number Trade Lines 60 Days or Worse Ever	4.0892	4.39542	471	471.000
Percent Trade Lines Open	.4677	.19633	471	471.000
Number Bad Dept plus Public Derogatories	2.5159	3.04213	471	471.000
Number Trade Lines 60 Days or Worse 24 Months	2.0934	2.38251	471	471.000
Percent Trade Lines Open 24 Months	.6171	.57398	471	471.000

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

REFERENCES

- Brown, R.E. & Mazur, M.J. (2003, June). *IRS's Comprehensive Approach to Compliance Measurement*. Retrieved February 20, 2008, from <http://www.irs.gov/pub/irs-soi/mazur.pdf>
- Brown, R.E. & Mazur, M.J. (2002, June 12). *The National Research Program: IRS's New Approach to Measuring Taxpayer Compliance*. Retrieved February 1, 2008, from http://www.iadb.org/INT/Trade/1_english/4_SpecialInfo/Conference/2002/d_Jun1102-Tax-Compliance/Mazur.pdf
- Brynes, T. (2002). 10 Ways to Trigger an IRS Attention. *CBS MarketWatch*. Retrieved Feb 17, 2008 from <http://www.marketwatch.com/News/Story/10-red-flags-you-dont/story.aspx?guid={B52A776D-A5CC-403F-A2B8-C6C843772E2D}>
- Cyr, D., Eckhardt, T., Sandoval, L.A., & Halldorson, M. (2002, June). *Predictors of Unreported Income: Test of Unreported Income (UI) DIF Scores*. Retrieved February 13, 2008, from <http://www.irs.gov/pub/irs-soi/puidif2.pdf>
- Daily, F.W. (Oct 1999). *Surviving an IRS Tax Audit*. California: Nolo.
- Data Mining. (n.d.). Retrieved February 12, 2008, from <http://www.unc.edu/~xluan/258/datamining.html>
- Don't Mess With Taxes. (2008, January 18). *Uh Oh Audits are Up!* Retrieved February 5, 2008, from http://dontmesswithtaxes.typepad.com/dont_mess_with_taxes/irs/index.html
- Encyclopedia Titanica (n.d.) *Titanic Passengers and Crew Listings Dataset*. Retrieved March 18, 2008, from <http://www.encyclopedia-titanica.org/titanic-passengers-and-crew/>
- Federal Taxes Weekly Alert (2008, March 20). *What are your chances for being audited? IRS's 2007 data book provides some clues*. 54(12).
- Georges, J. (2007). *Applied Analytics Using SAS Enterprise Miner 5*. USA: SAS Institute Inc.
- Herman, T. (2008, January 30). High Earners Face Surge In Tax Audits. *The Wall Street Journal*, p. D1, D2.
- Hunter W.J. & Nelson M.A. (1996, March). An IRS Production Function. *National Tax Journal*, 49(1), 105-115.
- IRS. (2008). *Fiscal Year 2007 Enforcement and Service Results*. Retrieved March 30, 2008, from <http://www.irs.gov/newsroom/article/0,,id=177701,00.html>

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

- IRS. (2008, January 14). *IRS Names Four New Frivolous Claims to Avoid*. Retrieved March 2, 2008, from <http://www.irs.gov/newsroom/article/0,,id=177519,00.html>
- IRS. (2002, September). *IRS Sets New Audit Priorities*. Retrieved March 3, 2008, from <http://www.irs.gov/newsroom/article/0,,id=105695,00.html>
- IRS. (2007, September 12). *IRS to Launch New National Research Program Study in October*. Retrieved February 4, 2008, from <http://www.irs.gov/businesses/small/article/0,,id=173874,00.html>
- IRS. (2007, June). *IRS Updates National Research Program for Individuals*. Retrieved February 16, 2008, from <http://www.irs.gov/newsroom/article/0,,id=171023,00.html>
- IRS. (2006, February 14). *IRS Updates Tax Gap Estimates*. Retrieved February 12, 2008, from <http://www.irs.gov/newsroom/article/0,,id=154496,00.html>
- IRS. (2007). *SOI Tax Stats*. Retrieved March 12, 2008, from <http://www.irs.gov/taxstats/indtaxstats/article/0,,id=103106,00.html>
- IRS. (n.d.) *The IRS Mission*. Retrieved March 2, 2008, from <http://www.irs.gov/newsroom/article/0,,id=98186,00.html>
- IRS. (2005, March). *Understanding the Tax Gap*. Retrieved November 11, 2008, from <http://www.irs.gov/newsroom/article/0,,id=137246,00.html>
- Johnston, D.C. (1996, February 25). Your Taxes; Some New Tricks To Help Filers Avoid An Old Audit Trap. *New York Times*. Retrieved February 12, 2008, from <http://query.nytimes.com/gst/fullpage.html?res=9500E1D71039F936A15751C0A960958260&sec=&spon=&pagewanted=1>
- Mertler C.A. & Vannatta R.A. (2005). *Advanced and Multivariate Statistical Methods* (3rd ed.). Glendale, California: Pyrczak Publishing.
- Office of Research: Research, Analysis, & Statistics. (2002, May 1). *Testing the UI-DIF Formulas*. Retrieved February 16, 2008, from <http://www.irs.gov/pub/irs-soi/uidiffor.pdf>
- Pilla, D.J. (2002). The IRS's intensified audit attack. *USA Today*, 131(2686), 30-31.
- Schwab, A.J. (2006). *Discriminant Analysis-Basic Relationships*. University of Texas at Austin. Retrieved March 23, 2008, from http://www.utexas.edu/courses/schwab/sw388r7_spring_2006/SolvingProblems/DiscriminantAnalysis_BasicRelationships_spring2006.ppt

The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits

Senior Capstone Project for Amber Torrey

StatSoft. (2008). *Data Mining Techniques*. Retrieved February 16, 2008, from <http://www.statsoft.com/textbook/stathome.html?stdatmin.html&1>

Stevens J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ:Lawrence Erlbaum Associates.