Developing Error Prone Profiles using Administrative Data with a Control Group

Pedro J. Saavedra and Hoke J. Wilson

Pedro J. Saavedra and Hoke J. Wilson – ORC Macro
Pedro J. Saavedra, ORC Macro, 11785 Beltsville Dr., Calverton, MD 20705 Pedro.J.Saavedra@orcmacro.com

Introduction

The practice of error-prone profiling to select applicants for various entitlement programs has made use of a variety of prediction techniques, from multiple linear regression to logistic regression to regression tre3s. In most programs, data exists where a random sample of applicants or participants have been audited, and a resulting outcome can yield a disposition indicating whether the original data provided by the program participant was in error. In the Student Financial Aid system, one only has access to administrative records indicating if a person was selected to be audited or not, if their data changed and if he was paid. But students change their data spontaneously or drop out for reasons unrelated to the error or the audit. This paper examines various methodologies that have been explored to derive profiles from administrative data under these constraints.

For more than two decades the U.S. Department of Education (ED) – first under Health, Education and Welfare's Bureau of Student Financial Assistance, and now through the Office of Federal Student Aid (FSA) - has strived to protect the integrity of the Pell Grant program through the use of error prone profiling (EPP). EPP involves the use of statistical or econometric techniques to identify error and reached its highest point of usage in the 1980s. With the advent of more powerful computer database hardware and software, the use of EPP has declined somewhat in favor of data mining and database cross-validation. However, EPP is still effectively employed by the Internal Revenue Service, the US Department of Agriculture, and other State and Federal Agencies to protect the integrity of their programs.

EPP Methods in Other Programs

The most widely used methodologies for Error Prone Profiling are various forms of regression. The Illinois Department of Public Aid's Bureau of Fraud Research (BFR), for example, employs logistic regression in an EPP model of error and fraud in the Medicaid program. For the last ten years, the Social Security Administration's Office of Disability and Income Security Programs (ODISP) has also used logistic regression for the analysis of any dichotomous event (e.g., error or no error, disability benefit allowance or denial, continuance or cessation of disability, etc.) [Personal correspondence with Susan David, Director, Disability Program Information and Studies, ODISP, SSA, September 24, 2003].

The BFR model incorporates approximately 100 variables to predict the probability of Medicaid applications being in error. In particular, the model is used to identify benefits applications with a low probability of error. These applications are then removed from the caseloads of auditors (Illinois Department of Public Aid, 1999).

Another example of the use of non-linear regression in EPP is the USDA Food and Nutrition Service's Violation Prone Profiling (VPP) system. The VPP is designed to detect retailer trafficking in the Food Stamp program and employs a close cousin of the logistic regression methodology – probit analysis. Developed in 1994, it uses administrative data on retailers and Census-based demographics of the geographic locale within which retailers do business to generate a probability that represents an estimation of the retailers' propensity to traffic. While still functional, the VPP has fallen from favor with FNS Compliance Branch investigators because they do not believe that its predictions point to retailers who they would otherwise not consider appropriate targets for investigation (Mantovani and Wilson, 2002).

Decision trees are methodologies used to examine complex relationships between a set of predictors and an outcome measure through the classification of like observations. A Decision Tree methodology (AID – Sonquist et. Al., 1974) was used by FSA to probabilistically identify errors in student aid applications in the early 1980s. The State of West Virginia and the Social Security Administration were also early users of Decision Trees (Applied Management Sciences, 1980). More recently, the Department of Housing and Urban Development (HUD) used a decision tree to determine how various tenant characteristics were related to the magnitude of rent error (ORC Macro International, 2001). At about the same time as the HUD study, the US Department of Agriculture's Food and Nutrition Service (FNS) used a decision tree to validate its estimate of retailer trafficking in the Food Stamp program, as well as to evaluate the ability of its Anti-Fraud Locator using EBT Retailer Transactions (ALERT) system to detect retailer trafficking (ORC Macro International, 2002).

Discriminant Analysis, also known as Discriminant Function Analysis has also been a widely used methodology in error prone profiling. The states of South Carolina and New Mexico, and the District of Columbia have used it to identify errors in applications for unemployment insurance, food stamps, and TANF/AFDC benefits[1]. However, the Internal Revenue Service (IRS) is probably the most long-standing practitioner. The IRS recently completed an evaluation of a new discriminant analysis application to unreported income and found the results of the model to be statistically, significantly correlated with outcomes generated by a panel of experienced IRS auditors. After further refinement, the evaluation recommends the incorporation of the scores generated by the model into the criteria used to select tax returns for audit (Cyr, *et al*, 2002).

Artificial intelligence (AI) algorithms include Fuzzy Logic, Evolutionary Computation, Expert Systems, and Computational Finance. However, the most broadly used AI, especially in the domain of error prone profiling, is Neural Networks. In the public sector, the most well-known application of a neural network is that employed by the Texas Health and Human Services Commission (HHSC) as part of its Medicaid Fraud and Abuse Detection System (MFADS). On line since December of 1997, the Texas HHSC identified 7,262 cases of Medicaid fraud and estimates that it has recovered almost $7.5 million in fiscal year 2000 through fiscal year 2002 (Texas Health and Human Services Commission, 2003). At the federal level, during the development stages of its Violation Prone Profiling system, the USDA FNS tested a neural network as a candidate methodology. While the neural network was not without its merits, FNS rejected it in favor of the probit regression methodology described earlier (Mantovani and Wilson, 2002). In the private sector, Capital One Bank uses a neural network procedure that it believes reduces credit card fraud by 96 percent (McCue, 2001).

Most of these programs conduct random audits and obtain information of whether a particular form was in error or not. Then the task becomes one of identifying the characteristics of the forms most likely to be in error. There are, of course, a number of difficulties in obtaining such a prediction. While the forms many have a finite number of variables, it is in the combination of these variables that one is likely to find good predictors. This is why techniques that emphasize the identification of interaction effects are likely to be more effective. A second, less serious, problem is that often the variable in question is far from normal, there being a larger number of cases with no error, and a smaller number with varying degrees of error, often capped by a maximum award or entitlement in the case of social programs. However, in the end, there usually exists a well-defined variable to be predicted, whether it be a dichotomous error/no-error variable, error in a particular direction or the absolute value of the discrepancy between what should have been reported or calculated and what actually was.

The Student Aid Problem

The situation for Student Financial Aid is entirely different from that of the above mentioned programs. Applications in the Student Financial Aid program are processed through the Title IV Central Processing

[1] Personal correspondences with Linda Martin, Director of South Carolina's Department of Social Services, Division of Planning and Research, September 9, 2003; Marise McFadden, Deputy Director for the New Mexico Department of Human Services Income Support Division; and Alvaro Galvin, Chief of Corrective Action and Analysis for the District of Columbia's Department of Human Services.

System. An application, known as a Free Application for Federal Student Aid (FAFSA) may be filled out electronically or on paper. A whole system of edits is in place, and students are at times requested to correct their application and submit the correction. The initial application and its subsequent corrections are referred to as transactions. If correction is mandatory, the transaction is rejected, and the student must resubmit. If a transaction is not rejected, an Expected Family Contribution (EFC) is calculated. Students are paid based on their EFC, their cost of education and their enrollment status (full or part time). Students also make spontaneous corrections.

A percentage of students (at present, about 30%) are selected for verification, and schools are instructed to demand a tax form or some other documentation of the information students provided on their FAFSA . If a student goes to the school and subsequently discovers in consultation with the institution's Financial Aid Office that he is not eligible for a Pell Grant, or if he reaches the conclusion on his own, the student may simply not submit a correction or pursue his request for aid. A separate system processes payments to the students, and indicates on which transaction a student was paid.

There are a number of difficulties that affect the effort to develop Error Prone Profiles in the context of the Pell Grant program:

1) There is no study that has verified a random sample of students and has the results of the verification as is the case for most programs that use EPPs.

2) Many students apply for aid and choose not to enroll, and this information is not available to the CPS system.

3) Awards depend on the Expected Family Contribution, a formula derived from income, family size, and a number of other variables, but awards also depend on cost of education and enrollment status. Cost of Education and enrollment status information are only available for students who actually receive a Pell Grant.

4) When a student is subjected to verification, and he or she becomes ineligible for an award, a correction may or may not make its way to the Central Processing System. If a student is notified that he or she must present evidence of the income reported and the student decides not to go through with the process, because of knowledge that the reported income is in error, that change will not make it to the CPS and the student will be indistinguishable from one who chose not to enroll for reasons unrelated to verification.

5) Many students make corrections spontaneously or in response to edits that suggest that their data may be in error. Students and parents who estimate their taxes are asked to correct their application if they discover that their estimate was wrong. Thus, corrections by a student selected for verification cannot be attributed solely to the verification process.

6) The development of profiles relies on data from two years before the academic year in which the EPP model is to be implemented, and the timing is such that some students will receive payment after the analysis is conducted (and hence will not appear in the payment system).

7) Schools are not required to verify more than 30% of their students. While the payment system records whether a student was verified, a student who was selected for verification and is not paid is likely to have been verified, but it is possible that they did not undergo verification.

8) Some schools receive permission to use their own verification criteria, and for students who do not receive an award, the exact school they attend has to be inferred from their first choice, so the individual practices of the school they attend or even what school was attended, can only be inferred.

9) The award has both an upper bound and a lower bound. Students receiving the maximum award can make any number of errors without affecting the award. A linear equation with continuous variables as predictors is unlikely to be effective unless it uses a Tobit model or some other procedure that handles censored data.

10) The most effective criteria are obtained from the interaction of variables or from cutoff points. That makes the number of possible criteria extreme large.

11) The EPP must be programmable as part of a system that processes several million applicants every year. And there must be a way to monitor what percent of the students are being selected and to modify the parameter of the model (e.g. by changing the cutoff points of an equation) in order to bring the proportion to 30%.

12) The application form, the formulas that define the EFC, the maximum award and the tax laws (Taxes paid is one highly error-prone variable) change often, and these changes can undermine the validity of a model in a given year.

The task confronting developers of EPPs for the Pell Grant and other SFA program is how to use the CPS and payment databases, which are used primarily to process applications and payments of grants, in order to build Error Prone Profiles for financial aid applicants. One reason for the development of profiles is that Department of Education policy requires that 30% of students be verified and EPPs are used to maximize the identification of error within that 30%.

The Research Design

In the early 1980s, the EPP for the SFA programs were derived by randomly selecting a sample of applicants, requiring that the school verify them and observing who made corrections. It became obvious that one drawback of this approach was that it could not differentiate between the applicant who had nothing to correct (and got paid), the one who was verified and did not re-enter the system (having become ineligible as a result of the verification), and the one who did not re-enter the system for reasons unrelated to verification. As a result, two innovations were implemented in the middle 1980s:

1) The CPS database was merged with the payment system. This merger had to coincide with the timing for the analysis – late May or June. The data used was from the academic cycle, then in its second year (i.e. in 2004 the cycle 2003-2004 would have been used).

2) In addition to a randomly selected group whom schools were instructed to verify, a second group was rendered immune from verification. That is, even if any member of this "immune group" met the criteria set by the existing Error Prone Profiles, it would not be selected for verification.

3) For each student, two values were calculated. First, the award the student would have received had he attended full-time based on the first transaction within which an EFC could be properly determined was calculated (referred to as the award at selection). Second, the award he would have received (not the one actually received) had he attended full-time to at cost, and had the EFC of the transaction on which he was paid. This second number was referred to as the award at payment and was set to zero for students who did not get paid.

The dependent variable in the error-prone profiles has varied, sometimes using misallocated dollars (absolute value of award at payment minus award at selection), sometimes using over-awards only (maximum of award at payment minus award at selection and zero), and sometime using any error (defined as a change in award, EFC, income, dependency status or family size between the selection and payment transactions, or a zero award at payment).

The use of the immune group was made necessary by the nature of the system and by an empirical examination of the data. It was noticed that individuals who make corrections or fail to get paid when selected, would have made unprompted corrections to their applications, but to a lesser degree, when not selected. Given the inability to distinguish between changes related to verification and changes that would have been made had the applicant not been selected for verification, it was decided that we had to identify characteristics of students with a high score on the error measure if verified, but lacking such a score when not verified.

The problem is conceptually similar to one of identifying persons who would benefit most from a vaccine. Some people will contract the illness whether or not they are vaccinated. Others will not contract the disease whether or not they are vaccinated. Those who will benefit most from the vaccine will be those who will contract the illness if and only if they are not vaccinated. Now suppose one had a very large sample and randomly gave the vaccine to half and a placebo to the other half. We would expect the predicted incidence of the illness to be lower in the treatment group, but how do we identify the characteristics of people who would benefit *most* from the vaccine? An answer to the question calls for some sort of data mining with a control group.

Approaches Attempted Over the Years

A source of current success for the FSA EPP model may reside in the dynamic approach taken to its estimation. The first statistically based attempt to create an EPP model that could probabilistically identify aid applicant error was conducted over twenty years ago and resulted in two models that made use of variants of a methodology known as Automatic Interaction Detection (AID). Since that time, a number of different methodologies have been used in the modeling process, and the variable used to indicate the presence of error (dependent variable), as well as the set of variables used to predict the dependent variable (the independent variables) have changed over the last two decades.

In 1986 a long-term verification plan was submitted to ED, randomly selecting 2.5% of all applicants for verification, and rendering immune from verification another 2.5%. Applicants in the immune group are randomly selected *not* to be verified, again without regard to their FAFSA responses.

The plan also made provision for the verification for cause of 30% of applicants, establishing the need for Error Prone Profiles to select 27.5% of students in addition to the 2.5% in the randomly selected group. The plan also provided for the merger of the application and payment files, and the conduct in the spring of the second year of a cycle of an analysis deriving Error Prone Profiles. In some years, when the application form has been relatively stable, files representing multiple years have been combined in the analysis. The EPPs derived in the spring are examined in the fall using data from the following cycle to make certain that the proportions meeting each criterion have not been altered considerably as a result of system or application changes.

There has been one constant throughout the nearly two decades of this approach to the development of SFA Error Prone Profiles. The method always relied on obtaining a pool of possible criteria, and then using a method to combine the criteria into a model.

The nature of the criteria is has changed over time. The first approach classified each student into one or more criteria, and then a prioritizing algorithm led to his classification into the highest priority criterion he met. Only students meeting high priority criteria were selected for verification. The second approach classified each student into mutually exclusive and collectively exhaustive criteria. Again, a subset of these was selected for verification. The third approach (based on a regression with dummy variables) added or subtracted points for each criterion a person met. Which system was used not only affected the model, but it also affected the Management Information Systems (MIS) reports associated with the verification process.

The early years of the use of the immune group relied upon a system for prioritizing criteria. The dependent variable was, as mentioned before, the change in scheduled award from the selection to the payment award, with the change equal to the award at selection if the student was not found in the payment file. Diverse methods were used to define criteria, from regression trees to multiple linear regression to simply reason what combinations might be indicative of error. A pool of criteria was compiled and for each criterion an index was obtained, equal to the mean change for the randomly selected applicants meeting the criterion less the mean change for the immune applicants meeting the criterion. A minimum number of applicants had to meet the criterion (hence eliminating the use of criteria which, while appropriate, were met by very few students). The criterion with the highest index was given Priority 1. Then the students who met that criterion were removed from the sample, and the same procedure was repeated.

One drawback of the above method was that combinations of criteria not previously considered could not emerge. As a result of a review of the methodology it was concluded that what was needed was a regression tree approach that could take into account the immune group. An examination of commercially available procedures such as C&RT (Breiman et. Al., 1984) did not yield an obvious method, so an algorithm was programmed. At the same time the decision was made to move away from trying to predict the magnitude of the change (or the difference in change between the random and the control group) to trying to predict the fact that there was a change. This means that every applicant in the random or the immune group could be classified as a "changer" or a "non-changer". In some years, the concept of a changer was expanded from merely award change to a change in award, EFC, income and/or household size.

The algorithm was based on phi, the correlation coefficient for two dichotomous variables. Possible partitions were defined just as criteria were defined before. The difference is that this system combined criteria. For each partition, the phi correlation between being selected for verification and being a changer was calculated for each side of the partition. For each partition with a certain minimum of applicants in the sample for both parts, the difference between the two phi coefficients was calculated. Then the partitioning with the largest absolute value of the difference was used to split the sample in two. The same procedure was applied to each side of the partition separately.

This procedure created a regression tree, but there was a need to validate the model. As a result, it was decided to apply the procedure to a training sample and subsequently to validate in a test sample. This almost always led to some pruning of the tree created by the algorithm. The analyses were conducted separately for dependent and independent students.

Around this time, there emerged some concern that the results were often driven by students who did not get paid. They constituted an important component of the population, since one of the purposes of the EPP was to prevent ineligible students from being paid. At the same time, their effect is confounded with the fact that many do not get paid for reasons entirely unrelated to verification, and thus their effect on the analysis is most subject to sampling variance. To minimize the effect, fifty percent of the students who did not get paid were randomly eliminated from each group.

The phi-based regression tree was far more effective than random selection, but less effective than what one might wish. It was estimated that if one verified 30% of applicants using the model, one would have verified 50% of students who would not have changed had they not been verified. In addition, the models were not very stable, and often substantial pruning was necessary.

The next change in the EPP modeling procedure took place when a match of FAFSA data and IRS tax returns was mandated, as a way of using IRS data in the EPPs was desired. Unfortunately, since the IRS has strict rules about the release of data, a system that required extensive manual intervention, as did the phi-based regression tree, was impractical. It was also necessary that the results of any IRS analysis be integrated with analysis from payment data. This was for two reasons. First, the IRS analysis could not include persons who did not file a tax return. Second, while a solid IRS match could identify profiles of misreporters, there is no guarantee that a Financial Aid Administrator requesting proof could identify these misreporters. On the other hand, identification of applicants who are misreporting and passing verification could at least lead to the re-examination of the verification procedures.

The Current EPP

The current EPP model was developed through the creation of six separate linear equations, three for dependent students and three for independents. In all six instances the dependent variable is the amount of Pell Grant overaward, or a transformation of the Pell Grant overaward. Negative differences, or "underawards", are presently set to zero.

Two hundred independent, predictor variables are used in the development of all six equations. As already stated, a principal difference between the six equations is that one set of three equations is calculated for

dependent students and another set of three equations for students with an independent status. The predictors for dependent students are similar, but not identical to those used for the independent students. Within each dependency status (dependent or independent), the predictors are identical for all three equations. For each set of equations based upon dependency status, the first equation is the result of a match with the IRS. The adjusted gross income (AGI) and tax liability information from the IRS files replace the corresponding figures reported on the FAFSA. Earned Income Credit replaces Worksheet A Income on the FAFSA if the former is larger. These figures are used to create a new EFC and scheduled award. The overaward is calculated using these latter figures.

A second method has been changed several times. We will describe the approach that was used in the analysis of the 2002-2003 and of the 2003-2004 applicant data

In 2003, using the 2002-2003 applicant database, the change in award was given a negative value for members of the immune group. More precisely, for each dependency status, the second equation used payment data for the randomly selected and immune groups. The mean award change (for both groups) was first subtracted from the applicant's absolute overaward, and the difference multiplied by negative 1 for the immune group. Then the regression equation was run.

In 2004, using the 2003-2004 database, only the random group was run, and applicants who had only one transaction and were not paid were removed from the analysis. This was based on a claim by some schools that they verified every student they were supposed to verify, and that they reported every change. The analysis made the assumption that applicants who never corrected and were never paid (if eligible for an award) also never enrolled.

The third equation for each dependency status uses the same data, but to each predictor adds the interaction of the predictor and a dummy variable indicating membership in the randomly selected group. The predictor and its interaction term are brought into the equation at the same time. Only the interaction terms are used in the final equation.

The three equations are then normed so that all three of them yielded the same mean prediction and have the same standard deviation. The normed coefficients are averaged (counting a coefficient as zero if a predictor was not in an equation). This yields the coefficients used in the final models. These coefficients were then applied to the records in the validation sets and used to generate scores that, ostensibly, are positively correlated with errors.

The random and immune groups are then sorted by their predicted scores and, for each percentile, the difference in award changes if one verified all students up to that one are calculated.

This general approach has one advantage and one main drawback. The advantage is that it yields an easily understandable point system. For each criterion one meets, so many points are added, and a student who exceeds a certain cutoff is selected for verification.

But it also has one large disadvantage. One has to start with the pool of criteria. Historical useful criteria, as well as new ones, based on combinations one sees as likely to be error-prone, are incorporated. Cutoff points are arbitrarily designated and new combinations do not emerge. The need to build a more systematic pool of criteria became apparent.

This led back to an old problem. How to conduct systematic data mining with a control group? The idea of using C&RT to obtain a regression tree seemed clear-cut, but the desirability of building in the changes that a student would make if not selected made a C&RT analysis with award change as the dependent variable seemed to make it unfeasible.

The C&RT analysis

Using the most current data available, the model is re-estimated every year, potentially resulting in the selection of different predictor (independent) variables combined in a manner that varies from one year to the

next. C&RT was suggested in the early 1990s by a team from the National Academy of Sciences, including Dan Carr and Ron Fecso. What was not obvious was how to include the information from the control group into the C&RT analysis.

It was important in 2005 that the regression system not be changed, since the processing system and the MIS system associated with Student Financial Aid application processing was in place and would have to be changed radically. So the decision was made to preserve the existing methodology, but to generate the criterion pool using regression tree methods, preserving only the historical criteria that have repeatedly appeared in the equations.

At first the plan was to use the immune group only at the time the regression equations were developed. C&RT would be implemented on the randomly selected group and each split and node would become a potential criterion. But then the suggestion arose that deriving regression equations using the immune groups and subtracting the change predicted had one not been verified (i.e. using the equation derived from the immune group) would yield a regression tree which, to the extent that the equation explained the changes in the immune group, would indeed remove some of the effect that was unrelated to verification.

Two predictions of the absolute value of the change in Pell awards were estimated; one each for dependent and independent applicants. Due to the clustering of the dependent variables at or near zero, and $4,050 (the maximum award), Tobit regressions were first attempted. Separate Tobit regressions were conducted with censoring from the left ($0), the right ($4,000), and from both sides. The resulting regressions, when applied to the full distribution of the applicants, explained less variation than linear regression, and their error distributions were not markedly improved. For these reasons, we defaulted to linear regressions to provide estimates of self-corrected errors.

For independents (about 58,000 students), preliminary C&RTs were performed on the immune group. Dichotomous variables were then coded that represented node membership, as well values of predictor variables that optimally determined splits in the population. By means of forward, stepwise regression, estimates of self-corrected errors were constructed. Continuous variables derived directly from the applicant FAFSA were added to the models as appropriate. The adjusted R-square was 0.063. As can be seen, the equations did not explain that much of the variance of change.

At that point, the immune group was set aside and the C&RT procedure was implemented on the verified group using the difference between actual change and predicted change, setting negative differences to zero.

A similar analysis was prepared for the dependents, but it will not be presented here. Currently a large percentage of dependent students are selected, so the task of find good predictors is less urgent. Nevertheless one result of the dependent analysis was the identification of one group with very large errors, identified by particularly high taxes (combined with being eligible for a Pell Grant).

Exhibits 1, and 2 present the results of the C&RT analysis, without the criteria. Exhibit 3 presents sample criteria on which nodes were split (for several alternate analyses, not just for the tree presented). Continuous variables defined splits at cutoff points.

The Next Steps

The use of nodes and splits as dichotomous variables in the regression process is being implemented as this is written. It is but the latest effort to improve the EPP. The C&RT model will not be used directly in the verification process, but will merely serve to generate criteria for the pool. Then the same regressions as were conducted last year will be implemented, except that in the second regression only the randomly selected for verification group will be used, and the equation derived from the immune group will be extracted first. Unfortunately, the true test for a model is not only stability within samples, but its ability to withstand the test of time. The new approach will be implemented for the 2006-2007 academic year, and cannot be tested in production until the spring of 2007.

Summary

The Federal Student Aid Error Prone Profiles have been developed for over twenty years in spite of many limitations. The FSA EPP models use administrative data, attempting to infer misreporting by comparing change patterns between students who have been selected for verification with those of students who have not. Central to this approach is the random selection of 2.5% of students for verification and of 2.5% of students for immunity for verification. In addition, data from the disbursement system has to be merged with the application system files, and the analysis takes place using data from two cycles previous to the cycle in which the model is to be implemented.

The modeling process consists of two parts. The first is the derivation of a pool of possible predictors, and the second is the selection of predictors and their integration into an equation. The problem is compounded because the model must not merely identify characteristics of applicants that change their award, but rather it must identify the characteristics of applicants who would change their award only if verified.

In the last two decades the model has gone through three major revisions: prioritizing of criteria, a regression tree based on differences in the phi correlation, and stepwise regressions with dichotomous predictors. The use of C&RT as a means of identifying predictors and the derivation of an equation to predict change in the immune group are currently among the alterations being implemented in the model targeting the 2006-2007 cycle.

Exhibit 1: Table of Nodes for Decision Tree for Independent Pell Applicants Randomly Selected to be Verified					
Node	**Parent Node**	**Mean**	**Std. Dev.**	**n**	**% of Total**
0	----------	671.67	1193.64	29217	100
1	0	823.38	1227.16	11137	38.12
2	0	578.21	1162.74	18080	61.88
3	1	908.38	1317.78	8748	29.94
4	1	512.15	733.80	2389	8.18
5	2	690.97	1252.29	7400	25.33
6 (terminal)	2	500.09	1089.65	10680	36.55
7	3	987.35	1310.04	5961	20.4
8 (terminal)	3	739.47	1318.64	2787	9.54
9 (terminal)	4	609.32	764.47	1298	4.44
10 (terminal)	4	396.53	677.98	1091	3.73
11	5	751.12	1336.98	6108	20.91
12	5	406.59	658.89	1292	4.42
13	7	1089.61	1301.62	2767	9.47
14	7	898.75	1131.05	3194	10.93
15	11	722.15	1327.66	5684	19.45
16 (terminal)	11	1139.46	1401.01	424	1.45
17 (terminal)	12	622.45	802.22	316	1.08
18 (terminal)	12	336.7	589.01	976	3.34
19	13	1149.13	1302.47	2124	7.27
20 (terminal)	13	893.01	1280.29	643	2.2
21 (terminal)	14	868.39	1316.18	2755	9.43
22 (terminal)	14	1089.26	1263.27	439	1.5
23	15	844.44	1381.49	1773	6.07
24 (terminal)	15	666.71	1298.92	3911	13.39
25	19	1082.51	1308.45	1490	5.1
26 (terminal)	19	1305.69	1275.69	634	2.17
27	23	988.44	1460.2	1109	3.8
28 (terminal)	23	603.94	1202.08	664	2.27
29 (terminal)	25	1149.56	1302.11	1152	3.94
30 (terminal)	25	854	1306.09	338	1.16
31 (terminal)	27	1185.61	1149.76	423	1.45
32 (terminal)	27	866.86	1422.8	686	2.35

Exhibit 2: Tree Diagram for the Analysis in Exhibit 1

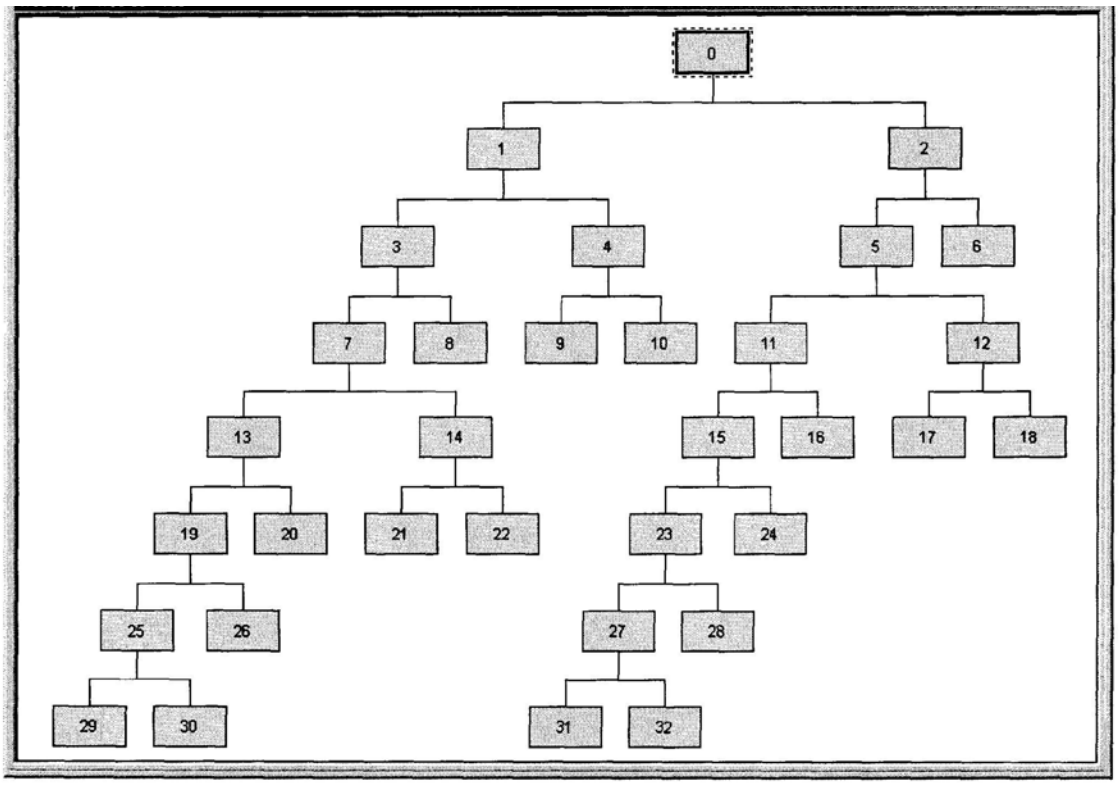

Exhibit 3: Partitions Associated with the non-terminal Nodes

- Did the student apply the previous year?
- Expected Family Contribution
- Year in School
- Control of first choice institution (public, private or proprietary)
- Worksheet A (nontaxable income)
- Reported taxes paid
- Estimated taxes paid
- Cash available
- Prior rejected transaction
- Date application received
- Adjusted Gross Income
- Difference between reported and estimated taxes
- Total family income
- Tax filing status (has filed, will file or will not file)
- Tax form used
- Eligibility for simpler tax form
- Income difference from previous year
- Earned Income
- Difference in income from previous year

References

Applied Management Sciences. *Quality Control Analysis of Selected Aspects of Programs Administered by the Bureau of Student Financial Assistance: Task 1 and Quality Control Sample, Error-Prone Modeling Analysis Plan* (Bureau of Student Financial Assistance, Office of Education, Department of Health, Education and Welfare: March 31, 1980).

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. **Classification and Regression Trees** (Belmont, Ca.: Wadsworth, 1984).

Cyr, Dennis, Thomas Eckhardt, Lou Ann Sandoval, and Marvin Halldorson. *Predictors of Unreported Income: Test of Unreported Income (UI) DIF Scores*. (Internal Revenue Service: Prepared for the Internal Revenue Service Research Conference, June 11-12, 2002).

Illinois Department of Public Aid. Long Term Care Asset Discovery Initiative (Office of Inspector General, September, 1999, http://www.state.il.us/agency/oig/docs/ltcadire.pdf)

Mantovani, R., and Wilson, H.. Measuring the Extent of Food Stamp Trafficking Using EBT Data (QRC Division of ORC Macro, prepared for the USDA Economic Research Service, January 7, 2002).

McCue, Andy. Neural Nets Join the Fight Against Fraud. (Computing, http://www.vnunet.com/News/1123414, June 25, 2001).

ORC Macro, *Quality Control for Rental Assistance Subsidies Determinations* (US Department of Housing and Urban Development, Office of Policy Development and Research, June 20, 2001).

ORC Macro, *The Extent of Trafficking in the Food Stamp Program after Welfare Reform: Analysis of Trafficking Outcomes using ALERT Scans of EBT Data* (USDA Food and Nutrition Service, April 30, 2002).

Sonquist, J. A., E. L. Baker and J. N. Morgan. **Searching for Structure**, revised ed. (Ann Arbor: Institute for Social Research, University of Michigan, 1974).

Texas Health and Human Services Commission. Texas Health Care Claims Study, March 2003. http://www.window.state.tx.us/specialrpt/hcc2003/section1/2activities.html