# Identifying Characteristics of High School Dropouts: Data Mining With A Decision Tree Model

**William R. Veitch, Ph.D.**
**Colorado Springs (CO) School District 11**

**Presented at the Annual Meeting of the American Educational Research Association**

**San Diego, CA**
**April, 2004**

# Identifying Characteristics of High School Dropouts: Data Mining With A Decision Tree Model

**William R. Veitch**
**Colorado Springs District 11**

## Introduction

The notion that all students should finish high school has grown throughout the last century and continues to be an important goal for all educational levels in this new century. Non-completion has been related to all sorts of social, financial, and psychological issues (see Kaplan et. al. 1994, for example). Many studies have attempted to put together a process that will identify students at risk of dropping out by using various research methodologies. The purpose of this study is to investigate correlates of high school dropping out through the use of data mining of existing data sources with decision trees.

## Dropout Research

As stated above, various methods have been used to predict high school non-completion. Hess and Copeland (2001) measured the use of various coping strategies by students to build a prediction model with discriminant analysis in what is, in essence, a personnel classification process. They found that the use/non-use of certain coping strategies (centering around social activities and seeking of professional support) significantly predicted high school dropout status. Similarly, another discriminant analysis by Streeter and Franklin (1991) found that students from low socio-economic status (SES) backgrounds were more likely to drop out than students from higher SES backgrounds.

On the whole, minority students tend to drop out of high school at higher rates than non-minorities (NCES, 1999). Considerable research on dropout behaviors and ethnicity are in the current literature. Pursley and Lan (2003) offer an excellent, recent review on the ethnicity topic, as well as references to dropout research undertaken from many different perspectives including academic achievement, motivation in school work, participation in activities, educational aspiration, perceptions of school, relations with peers, and self esteem.

Other recent research into the correlates of early high school dropping out include Wayman (2001), a highly technical study that looked at many measures on students through the use of logistic regression as well as multiple imputation[1] to account for the missing data. It found a predictor set of student achievement (as the strongest predictor), SES, and age. A structural equation modeling study by Battin-Pearson. et. al. (2000)

[1] In multiple imputation, missing values for any variable are predicted using existing values from other variables. The predicted values, called "imputes", are substituted for the missing values, resulting in a full data set called an "imputed data set".

found also that poor academic achievement was the best predictor of dropping out of school before completing 10th grade.

## Decision Trees[2]

Data mining is used extensively in the business field, especially in the area of marketing, where, for example, Internet companies analyze hits on their web sites. One data mining methodology involves decision trees (see Cabena et. al., 1998; and Groth, 1998), which are used to determine characteristics of best customers (perhaps, in education, students most in need of help in a particular academic area), or to determine which of a group of predictor test variables are most highly related to a target test variable (see Bowman, 2002) in order to offer a course of action concerning instructional emphasis. Extant educational applications are difficult to find and reported most often in post-secondary settings (see Luan and Willett, 2000). It is difficult to find very many K-12 organizations that have tried to make the conceptual jump from what are most often thought of as sophisticated business processes to an attempt at a data mining application in public education.

Decision tree methods are designed to sift through a set of predictor variables and successively split a data set into subgroups in order to improve the prediction (classification) of a target (dependent) variable. As such, these methods are valuable to data miners faced with constructing predictive models when there may be a large number of predictor variables and not much theory or previous work to guide them.

Traditional statistical prediction methods (for example, regression and discriminant analysis) involve fitting a model to data, evaluating fit and estimating parameters that are later used in a prediction equation. Decision tree models take a different approach. They successively partition a data set based on the relationships between predictor variables and a target (outcome) variable. When successful, the resulting tree indicates which predictor variables are most strongly related to the target variable. It also describes subgroups that have concentrations of cases with desired characteristics (for example: those students most in need of instructional assistance in math).

The general decision tree approach is to find the best single predictor of the dependent (target) variable at the root of the tree. Finding this predictor usually involves recoding or grouping together several of the original values of the predictor to create at least two nodes. Each node then defines a new branch of the tree that is being created. Within each branch, the process repeats itself. The algorithm looks for the best predictor among the remaining set of variables. Again, it will create at least two nodes with that best predictor. When no predictor can be found that improves the accuracy of prediction, the tree can be "grown" no further.

Decision tree models can offer some advantages over traditional statistical models. First, they are designed to be able to handle a large number of predictor variables, in some cases far more than the corresponding parametric statistical model would permit. Secondly, many tree-based models are entirely non-parametric and can

[2] SPSS training manual "Data Mining: Modeling"

capture relationships that standard linear parametric models do not easily handle, if at all (nonlinear relationships, complex interactions).

## CHAID Analysis

Chi-squared[3] Automatic Interaction Detection (CHAID) is a heuristic[4] tree-based statistical method that examines the relations between many categorical, ordinal or continuous predictor variables (which are grouped into ordered categories: either by the program or the user) and a categorical outcome (target) measure. The computer routine used in this study, Answer Tree (SPSS, 2001), provides a summary diagram (tree) depicting the predictor categories that make the greatest difference in the desired outcome, a summary table reporting which nodes have the greatest concentration of the trait of interest (gains analysis) and a table of misclassification information (risk analysis).[5]

## Analysis

Such analyses as described in the research section above, at once sophisticated and elegant, can have a major drawback: they are difficult to explain to non-statisticians. Individuals who posses little, or no, statistical training find regression and its other parametric cousins daunting methodologies to interpret. Since CHAID analyses involve nothing more complicated than frequency counts and percentages plus a more easily explained statistic (Pearson chi-squared - $\chi^2$ - procedure), explanations to methodological neophytes tend to be more manageable.

CHAID performs pair-wise comparisons in order to find the most effective predictor variable(s) - most highly related to - the criterion variable. In the case of many predictor variables, having this function performed automatically by computer software is essential when dealing with large data sets. In addition, since the CHAID procedure involves multiple chi-squared tests of independence, Answer Tree uses the Bonferonni[6] adjustment as its default for hypothesis testing.

In an effort to put as little stress on building and district resources as possible, this study utilized only extant data sources. No additional assessment (achievement or psychological) nor demographic data were gathered. All variables used in the study were extracted directly from district electronic databases. High school students recorded as "dropped" (with no transfer record) over the course of the 2001-2002 academic year were matched with a random sample of non-dropouts. Although some dropout research (e.g. Barrington and Hendricks, 1989) finds little in the way of gender-related dropping out behaviors, student gender was included in this study so as to further test that finding.

Appendix B lists the variables used in the analysis. The variables "Student ID Number" through "Ethnic Group" are self-explanatory. "Socio-economic Status" is

---

[3] See Appendix A for a brief overview of this statistical procedure

[4] *Able to change*: used here to describe a computer program that can modify itself in response to the user

[5] http://www.kdnuggets.com/software/classification-tree-rules.html is one site that lists various tree-based software packages.

[6] To maintain $\alpha_e$ at nominal $\alpha$, Bonferonni adjusts $\alpha$ for each comparison by the total number of comparisons. In this manner, $\alpha_e$ becomes independent of the number of comparisons.

measured by school free and reduced lunch eligibility. The "Discipline Infraction" entries are frequency of occurrence with level one being the least serious and level four the most serious.

"Advanced Classes" through "Unexcused Absences" are frequency counts with absences measured in class periods missed. Grade point average is measured on a standard four point scale. Colorado Student Assessment Program (CSAP) results are measured on a four point scale: (1) unsatisfactory, (2) partially proficient, (3) proficient, and (4) advanced.

As a result of a preliminary run using all variables as predictors of dropping out, a number of variables were recoded so as to reduce the number of categories in hopes of making the resulting trees more interpretable. "Age", Ethnic Group", "GPA", as well as number of Math, Science, and Advanced classes were recoded. In addition, since the "unsatisfactory" and "advanced" categories occur relatively infrequently, the CSAP results in reading, writing and math were recoded into dichotomies of (1) unsatisfactory/ partially proficient and (2) proficient/advanced. As it turned out, however, the best predictors were the variables in their original non-recoded states.

**Findings**

Figure 1, found below, is the result of a tree-growing effort that included all of the original variables as potential predictors. Appendix C is a larger, more readable version of this table. The results were cross-validated[7] by the software with 25 random samples. The misclassification matrix for this tree model is presented in Table 1 below. The critical cell (a dropout misclassified as a non-dropout) is held to a minimum (only 65

Table 1. Misclassification Matrix for 2201-2002 Dropout - Full Tree

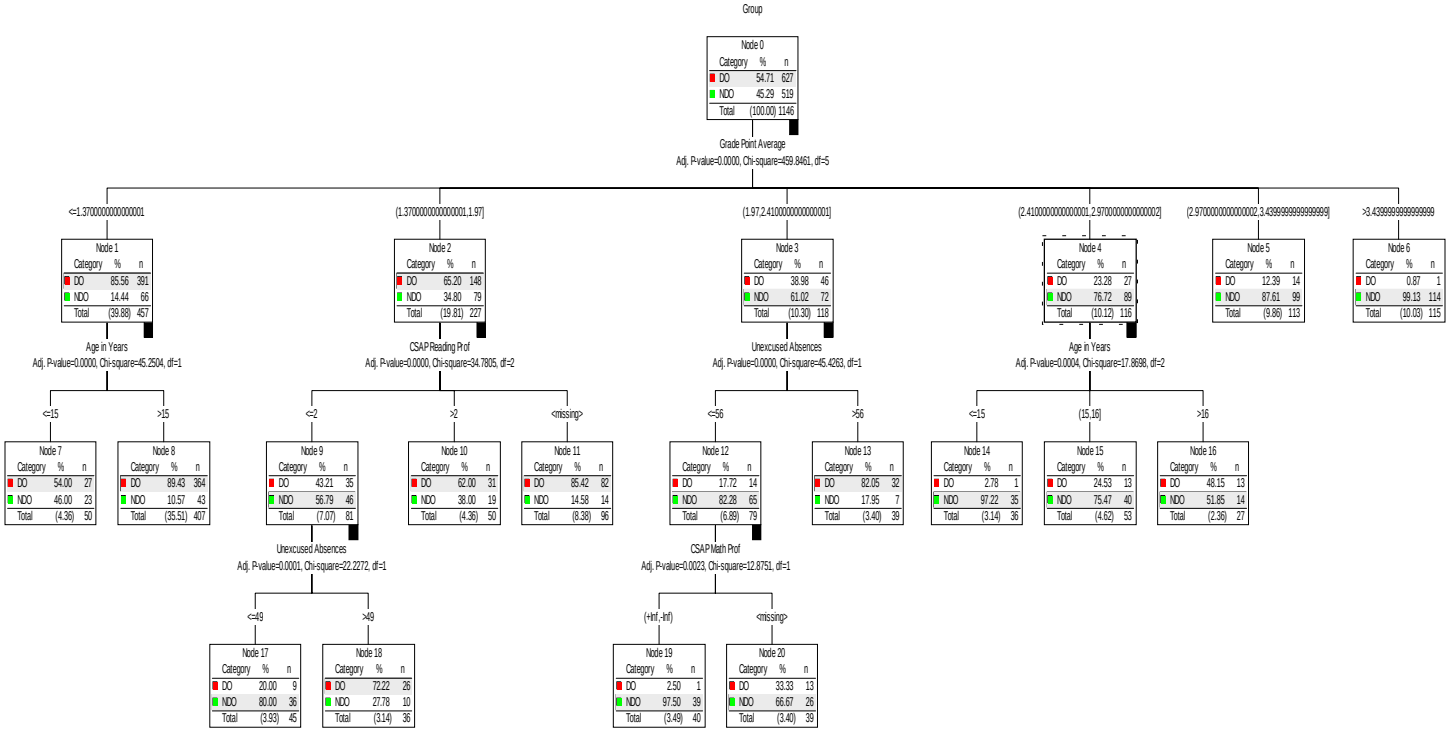
| | | Actual Category | | |
|---|---|---|---|---|
| | | Drop | No Drop | Total |
| Predicted Category | Drop | 562 | 116 | 678 |
| | No Drop | 65 | 403 | 468 |
| | Total | 627 | 519 | 1146 |
| | | Risk Estimate 0.158 | Cross-Validation 0.172 | |

students) in this model. The somewhat down-side of the model is that the risk estimate (likelihood of all types of misclassification) is nearly 16%. This number is reinforced by

[7] Cross-validation involves splitting the sample into a number (specified by the user) of smaller sub-samples (called folds) and generating a tree (excluding the current sub-sample) against which that hold-out sub-sample is applied. The cross-validated "risk estimate" is the average misfit across all the sub-samples generated.

the cross-validation risk estimate, from the 25 iterations of sub-sampling, at just over 17%. On the other hand, one's prediction should be correct over 80% of the time.

Figure 1. 2001-2002 Dropout Data – Full Tree




An inspection of Figure 1 indicates, as in many studies, that the variable most related to dropping out behaviors is academic performance. The first variable (that with the largest chi-squared statistic) is grade point average. Nodes 1 through 6 show a steady decrease in dropout percentage the higher the grade point average grouping (85.56% down to 0.87%). Node 1 splits according to age: older students are more likely to dropout than younger (see Nodes 7 and 8). Even though the node has split, it isn't telling us much we didn't know already; since almost everyone with a GPA below 1.37 drops out. This branch could probably be pruned.

Node 2 (GPA between 1.37 and 1.97) splits according to CSAP Reading performance. In essence, what this branch says is that students are more likely to stay in school, if they have a test score – regardless of how good or bad it is (see Node 11). The split of Node 3 (GPA between 1.97 and 2.41) involves unexcused absences. Those with more than 56 hours of unexcused absences are much more likely to drop out (see Node 13). In a similar manner, Node 4 (GPA between 2.41 and 2.97) splits according to age. Older students (age greater than 16) are more likely to drop out (see Node 16). Nodes 5 and 6 do not split. High achieving students tend to drop out in small numbers.

The tree contains a third level with just two branches. Low achieving students who score unsatisfactory or partially proficient on the CSAP and have high numbers of unexcused absences (see Node 18) even more likely to drop out than those students with a better attendance record. Lastly, Node 12 splits along CSAP Math performance. As

with the split of Node 2, students with test scores (regardless of how well they did) are more likely to stay in school.

## Conclusions

As stated in the opening paragraph, the purpose of this study was two-fold: first to investigate the existence of variables related to dropping out behavior; and second, to introduce the data mining of existing sources with decision trees. The tree presented in this paper does exhibit a certain ability to predict which students may drop out of school. Knowing which ones does no good without the capability to easily deploy the model.

Answer Tree can produce the programming language to identify members of any specific node within a tree. For instance, to see the students in Node 16, which was discussed above, the software generates the following SPSS syntax:

```
* Node 16.
SELECT IF ((GPA GT 2.41  AND  GPA LE 2.97)  AND  (AGE GT 16)).
EXECUTE.
```

The same node may be accessed through SQL with the following syntax:

```
/* Node 16*/
SELECT * FROM <TABLE>
WHERE (NOT(GPA IS NULL)  AND  (GPA > 2.41  AND  GPA <= 2.97))  AND  (NOT(AGE IS NULL)
AND  (AGE > 16));
```

Finally, Answer Tree will produce a set of logical statements describing the node that might be used in a written report:

```
/* Node 16*/
IF (GPA NOT MISSING   AND  (GPA > 2.41  AND  GPA <= 2.97))  AND  (AGE NOT MISSING   AND
(AGE > 16))
THEN
        Node = 16
        Prediction = 2
        Probability = 0.518519
```

With a little explanation and some guided practice, school level personnel could begin to use the results of this study to take a closer look at students who may be at risk of dropping out. It probably isn't necessary to publish the tree itself, rather only the results. Applying the generated SPSS syntax from a tree a node of particular interest to a data set of current students can quickly and simply produce an alphabetized list of at-risk students, by building, along with their accompanying demographics.

## REFERENCES CITED


Barrington, B., & Hendricks, B. (1989). Differentiating characteristics of high school graduates, dropouts, and non-graduates. *Journal of Educational Research, 82*, 309-319

Battin-Pearson, S., Newcomb, M., Abbott, R., Hill, K., Catalano, R., & Hawkins, J. (2000). Predictors of early high school dropout: a test of five theories. *Journal of Educational Psychology, 92*(3), 568-582.

Bowden, C. (2002). *Report on LEAP test strategies for the 2000-2001 school year for all principals*. Lafourche Parish, LA: Lafourche Parish Schools.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: From concept to implementation*. Upper Saddle River, NJ: Prentice-Hall PTR.

Groth, R. (1998). *Data mining: A hands-on approach for business professionals*. Upper Saddle River, NJ: Prentice-Hall PTR.

Hess, R., & Copeland, E. (2001). Students' stress, coping strategies, and school completion: a longitudinal perspective. *School Psychology Quarterly, 16*(1), 389-405.

Kaplan, D., Damphouse, K., & Kaplan, H. (1994). Mental health implications of not graduating from high school. *Journal of Experimental Education, 62*(2), 105-123.

Luan, J., & Willett, T. (2000). *Data mining and management: A system analysis for establishing a tiered knowledge management model* (ERIC 450 818). Aptos, CA: Cabrillo College.

National Center for Educational Statistics. (1999). *The condition of education* (NCES Report 1999-022). Washington, DC: NCES.

Pursley, M., & Lan, W. (2003). *Changes in personal characteristics of Mexican-American high school graduates and dropouts during the transition from junior high to high school.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

SPSS. (2001). *Answer Tree 3.1*. Chicago: SPSS, Inc.

Streeter, C., & Franklin, C. (1991). Psychological and family differences between middle class and low income dropouts. *The High School Journal, 74*, 211-219.

Wayman, J. (2001). Factors influencing GED and diploma attainment of high school dropouts. *Educational Policy and Analysis Archives, 9*(4), 18 electronic pages at http://epaa.asu.edu/epaa/v9n4/.

# APPENDIX A

## Chi-squared ($\chi^2$) Tests of Independence (Optional Reading)

Chi-squared tests of independence are extremely useful non-parametric statistical procedures for determining whether two nominal/ordinal (or categorized continuous) measures are related. If, for instance, one of the variables is group membership and the other a criterion of some sort, the test may be used to determine whether two or more populations are distributed in the same fashion with respect to the criterion.

Such data are organized into a bivariate frequency table (also called a pivot table), and the statistical test is made to determine whether the row variable is independent of classification on the column variable. As an example, suppose the row variable were used to classify subjects with respect to political affiliation and the column variable with respect to gender. The chi-squared test of independence would be used to determine if there is a relationship between these two nominal variables. We seek to answer the question "With respect to political affiliation, do males and females represent two different populations or a single population?" That is to say, is party affiliation independent of, or related to gender?

# APPENDIX B

## Research Variables Used in the Analyses

Group (Dropout/Not Dropout)

Age in Years

Gender

Ethnic Group

Socio-economic Status

Level 1 Discipline Infractions (number)

Level 2 Discipline Infractions (number)

Level 3 Discipline Infractions (number)

Level 4 Discipline Infractions (number)

Advanced Classes Taken (number)

Math Classes Taken (number)

Science Classes Taken (number)

Excused Absences (hours)

Unexcused Absences (hours)

Grade Point Average

CSAP Reading Proficiency Level (4 levels)

CSAP Writing Proficiency Level (4 levels)

CSAP Math Proficiency Level (4 levels)

# APPENDIX C

## Table 1. 2001-2002 Dropout Data—Full Tree