Learning Patterns of University Student Retention

Ashutosh Nandeshwar[a,*], Tim Menzies[b], Adam Nelson[c]

[a] *Kent State University, 126 Lowry Hall, Kent, OH 44242 Phone: 330-672-82222*
[b] *West Virginia University, 841a Engineering Sciences Building, Morgantown, WV 26505*
[c] *West Virginia University, Engineering Sciences Building, Morgantown, WV 26505*

Abstract

Learning predictors for student retention is very difficult. After reviewing the literature, it is evident that there is considerable room for improvement in the current state of the art. As shown in this paper, improvements are possible if we (a) explore a wide range of learning methods; (b) take care when selecting attributes; (c) assess the efficacy of the learned theory not just by its *median* performance, but also by the *variance* in that performance; (d) study the *delta* of student factors *between* those who stay and those who are retained. Using these techniques, for the goal of predicting if students will remain for the first three years of an undergraduate degree, the following factors were found to be informative: family background and family's social-economic status, high school GPA and test scores.

Keywords: data mining, student retention, predictive modeling, financial aid

1. Introduction

This article uses data mining to find patterns of student retention at American Universities. Such an analysis is urgently required. In our work, we have seen a disconnect between accepted best practices and the data available to support those practices:

- Based on our discussion with the university administrators, we assert that there is much *informal* agreement on the factors that influence retention (for the most part, the financial status of the student is considered to be the most important factor after the student's high-school GPA).
- However, as shown below, when we look recent experiments with student records, we find little clear support for that informal belief. In fact, we know of many universities that try to improve retention with a wide range of programs such as:
  - attracting students with high performance indicators (such as their test scores)

*Corresponding author
Email addresses: anandesh@kent.edu (Ashutosh Nandeshwar), tim@menzies.us (Tim Menzies), rabituckman@gmail.com (Adam Nelson)
URL: www.nandeshwar.info (Ashutosh Nandeshwar), www.menzies.us (Tim Menzies)

– developing student success programs (such as first-year experience)

– or encouraging tenured-faculty to teach undergraduates

Given the large levels of public support allocated to universities, it is important that we check the validity of these *informal* intuitions as well as the utility of the various retention programs such as those conducted at Kent State.

This article applies data mining methods to the problem of studying student retention. Our general conclusions will be:

- It is possible to find patterns of student retention, using data mining;
- Previous data mining studies on these student records can be greatly improved using discretization, attribute selection and cross-validation over various algorithms.

More specifically, we will show that data mining can uncover a rich level of detail about particular universities. For example, while mining data from Kent State, we found:

- A small and specific population of students at high risk of dropping out of university.
- That the above programs (using tenured-faculty for lecturing and focusing on student performance data) is far less important than the financial status of a student.

Hence, we would recommend:

- Focusing more resources on the high-risk group of students, in order to improve their chances of completing a university degree.
- Discontinuing the retention programs that primarily focus on student performance indicators or that advocate using tenured-faculty for lecturing.

While these conclusions are specific to Kent State, the method for finding them is quite general and could be applied to other universities in order to find their most specific and most important student retention patterns. We welcome contacts from other researchers who wish to repeat our analysis on their local data.

2. Literature Review

It is no news that higher education institutions are facing the problem of student retention, which affects graduation rates as well. Colleges with higher freshmen retention rate tend to have higher graduation rates within four years. The average national retention rate is close to 55% and in some colleges fewer than 20% of incoming student cohort graduate (Druzdzel & Glymour, 1994), and approximately 50% of students entering in an engineering program leave before graduation (Scalise et al., 2000). Tinto (1982) reported national dropout rates and BA degree completions rates for the past 100 years to be constant at 45 and 52 percent respectively, except for the World War II period (see Figure 1 for the completion rates from 1880 to 1980). Tillman & Burns (2000) at Valdosta State University (VSU) projected lost revenues per 10 students, who do not persist their first semester, to be $326,811. Although gap between private institutions and public institutions in terms of first-year students returning to second year is closing, the retention rates have been constant for both types of institutions (ACT, 2007, see

Figure 2). National Center for Public Policy and Higher Education (NCPPHE) reported the U.S. average retention rate for the year 2002 to be 73.6% (NCPPHE, 2007). This problem is not only limited to the U.S. institutions, but also for the institutions in other countries such as U.K and Belgium. The U.K. national average freshmen retention for the year 1996 was 75% (Lau, 2003), and Vandamme (2007) found that 60% of the first generation first-year students in Belgium fail or dropout.

[Figure 1 about here.]

[Figure 2 about here.]

Various researchers have studied this problem extensively, using theoretical models (Tinto, 1975, 1988; Spady, 1970, 1971; Bean, 1980), traditional models (Terenzini & Pascarella, 1980; Pascarella & Terenzini, 1979, 1980), and data mining techniques (Druzdzel & Glymour, 1994; Sanjeev & Zytkow, 1995; Massa & Puliafito, 1999; Stewart & Levin, 2001; Veitch, 2004; Barker et al., 2004; Salazar et al., 2004; Superby et al., 2006; Sujitparapitaya, 2006; Herzog, 2006; Atwell et al., 2006; Yu et al., 2007; DeLong et al., 2007). As shown below, we can improve those prior results by augmenting standard data mining with *discretization*, *attribute selection*, and *cross-validation* over various algorithms.

As documented in Adam & Gaither (2005), the literature on retention in higher education is extensive, and although various researchers have tested theoretical models and noted attributes critical to student retention, these theories need to be tested from time-to-time. New generation of data miners make the testing easier, and possibly can find new theories or reject old theories using state-of-the art learning algorithms. In this section, we focus on the literature relating to data mining and the student retention problem. The lesson of this section is that learning patterns of student retention is very difficult and, despite decades of effort, there is much room for improvement in the current state of the art.

2.1. Data Mining for Student Retention

Druzdzel & Glymour (1994) were among the first researchers to apply knowledge discovery algorithm to study the student retention problem. The authors applied TETRAD II, a casual discovery program developed at Carnegie Mellon University, to the U.S. news college ranking data to find the factors that influenced student retention, and they found that the main factor of retention was the average test score. Using linear regression, the authors found that test scores alone explained 50.5% of the variance in freshmen retention rate. In addition, they concluded that other factors such as student-faculty ratio, faculty salary, and university's educational expense per student were not casually (directly) related to student retention; and suggested that to increase student retention universities should increase the student selectivity.

Sanjeev & Zytkow (1995) used 49er, a pattern discovery process developed by Żytkow & Zembowicz (1993), to find patterns in the form of regularities from student databases related to retention and graduation. The authors found that academic performance in high school was the best predictor of persistence and better performance in college, and that the high school GPA was a better predictor than the ACT composite score. In addition, they found that no amount of financial aid influenced students to enroll for more terms.

Massa & Puliafito (1999) applied Markov chains modelling technique to create predictive models for the student dropout problem. By tracking the students for 15 years, the authors created state variables for the number of exams appeared, average marks obtained, and the continuation decision. Using data mining, Stewart & Levin (2001) studied the effects of student characteristics to persistence and success in an academic program at a community college. They found that the student's GPA, cumulative hours attempted, and cumulative hours completed were the significant predictors of persistence, and that young males were a high risk group.

Veitch (2004) used decision trees (CHAID) to study the high school dropouts. Using 25-fold cross-validation, the overall misclassification rates of drop-outs who were classified as non-dropouts were 15.79% and 10.36%. In this study, GPA was the most significant predictor of persistence. Salazar et al. (2004) used clustering algorithms and C4.5 to study graduate student retention at Industrial University of Santander, Colombia. The authors found that the high marks in the national pre-university test predicted a good academic performance, and that the younger students had higher probabilities of a good academic performance.

Barker et al. (2004) used neural networks and Support Vector Machines (SVM) to study graduation rates; the first-year advising center (University College at University of Oklahoma) collected data via a survey given to all incoming freshman. It is worthwhile to note that Barker et al. (2004) excluded all the missing data from the study, which constituted for approximately 31% of the total data. Overall misclassification rate was approximately 33% for various dataset combinations. The authors used principal component analysis to reduce the number of variables from 56 to 14, however, reported that the results using the reduced datasets were "much worse" than the complete datasets.

Superby et al. (2006) applied discriminant analysis, neural networks, random forests, and decisions trees to survey data at the University of Belgium to classify new students in low-risk, medium-risk, and high-risk categories. The authors found that the scholastic history and socio-family background were the most significant predictors of risk. The overall classification rates for decision trees, random forests, neural networks, and linear discriminant analysis were 40.63%, 51.78%, 51.88%, and 57.35% respectively.

Using the National Student Clearinghouse (NSC) data, Sujitparapitaya (2006) differentiated between stopout, retained, and transfer students. The overall classification rates for the validation sets using logistic regression, neural networks, C5.0 were 80.7%, 84.4%, and 82.1% respectively. Herzog (2006) used American College Test's (ACT) student profile section data, NSC data, and the institutional student information system data for comparing the results from the decision trees, the neural networks and logistic regression to predict retention and degree-completion time. The author substituted mean average ACT scores for missing scores. Decision trees created using C5.0 performed the best with 85% correct classification rate for freshmen retention, 83% correct classification rate for degree completion time (three years or less), 93% correct classification rate for degree completion time (six years or more) for the validation datasets.

Atwell et al. (2006) used University of Central Florida's student demographic and survey data to study the retention problem with the help of data mining. In this study, university retained approximately 82% of the freshmen from the study, and it used 285 variables to create data mining models. The authors used nearest neighbor algorithm to impute more than 60% observations with missing values. Using decision trees with the entropy split criterion, the authors obtained precision of 88% for the not-retained

outcome using the test data, and the actual retention rate for this test data set was 82.61%.

Yu et al. (2007) studied the data from Arizona State University using decision trees, and included variables, such as demographic, pre-college academic performance indicators, current curriculum, and academic achievement. Some of the important predictor variables were accumulated earned hours, in-state residence, and on campus living.

To study the retention problem using data mining for the admissions data, DeLong et al. (2007) applied various attribute evaluation methods, such as Chi-square gain, gain ratio, and information gain, to rank the attributes. In addition, the authors tested various classifiers, such as naïve Bayes, AdaBoost M1, BayesNet, decision trees, and rules, and noted that AdaBoost M1 with Decision Stump classifier performed the best in terms of precision and recall, hence, used this classifier for further experimentation. The authors balanced the class variable (retained and not retained) and obtained over 60% classification rates for both retained and not retained outcome. The authors concluded that the number of programs that the student applied to that specific institution and the student's order of program admit preference were the most significant predictors of retention.

Pittman (2008) compared various data mining techniques (artificial neural networks, logistic regression, Bayesian Classifiers, and decision trees) applied to the student retention problem, and also used attribute evaluators to generate rankings of important attributes. The author concluded that logistic regression performed the best in terms of ROC-curve area.

[Table 1 about here.]

2.2. Assessing the State of the Art

Table 1 lists techniques used in the studied literature, where the cohort sizes were available, along with the reported performance measures. Clearly, there is much room for improvement in the current state of the art:

- It is a recommended data mining practice to divide the data into a train and test set, learn on the train set, then assess the learned theory on the test set (Witten & Frank, 2005). Otherwise, if a theory is tested via the data used to build that theory, this test can over-estimate theory performance. For example, the Glynn et al. result of Table 1 seems impressive (a 83% accuracy on a data set with 49.08% a retention rate); however, that result should be repeated using some *hold-out* test set.

- All the regression studies from 1971 to 1999 report R^2 values under 0.6. This R^2 value is a measure of how well future outcomes are likely to be predicted by the model. The maximum value of R^2 is one and R^2 values under 0.6 indicate very weak predictive abilities.

- The accuracy reports are very close to the *ZeroR* theoretically lower-bound on performance. ZeroR is a baseline classifier that simply returns the majority class. For example, Herzog studied a data set with a 83.5% retention rate (see Table 1). ZeroR, applied to this data set, would be correct in 83.5% of cases. Therefore, the 85.4% accuracy of Herzog's data miners is very close to the ZeroR lower-bound;

i.e. the sophisticated analysis of that paper could be very nearly replicated using the dumbest of learners (ZeroR).

The last three results of Table 1 do not report their accuracies. However, these can be calculated in the following way. Let A, B, C, D be the true negatives, false negatives, false positives, and true positives respectively of a predictor that some student will attend some year of university. From Zhang & Zhang (2007) and Menzies et al. (2007), we say that (A,B,C,D) can be used in the following performance measures:

$$\begin{aligned} pd = recall &= D(B+D) && (1)\\ pf = false\ alarm &= C/(A+C) && (2)\\ prec = precision &= D/(C+D) && (3)\\ acc = accuracy &= (A+D)/(A+B+C+D) && (4)\\ neg/pos &= (A+C)/(B+D) && (5) \end{aligned}$$

Note that all these performance measures assess subtly different aspects of the performance of data miner:

- "Recall" measures how much of the target was found.
- The "false alarm" rate measures what fraction of non-targets triggered the learned theory.
- "Precision" comments on how many targets are found in the data selected by the theory.
- "Accuracy" comments on how many of the targets and non-targets were accurately labeled by the learned theory.

In an ideal result, we can obtain high recall, low false alarms, high precision, and high accuracies. However, as discussed by Zhang & Zhang (2007) and Menzies et al. (2007), these values are inter-related. Hence, the ideal result is not possible. These inter-relationships are shown below:

$$\left(prec = \frac{D}{D+C} = \frac{1}{1+\frac{C}{D}} = \frac{1}{1+\frac{neg}{pos}\cdot\frac{pf}{recall}} \right) \Rightarrow \left(pf = \frac{pos}{neg}\cdot\frac{(1-prec)}{prec}\cdot recall \right) \quad (6)$$

If a publication misses a particular performance measure, it is possible to use these equations to infer the missing value. For example:

$$\begin{aligned} D &= recall * pos && (7)\\ C &= pf * neg && (8)\\ A &= C * 1/(pf-1) && (9)\\ acc &= (A+D)/(neg+pos) && (10) \end{aligned}$$

Using these equations, we can comment that the last three results of Table 1 can be significantly improved:

- In Atwell et al. (2006), the the precision varied from 73% to 88%. Using our equations, we can estimate false alarm values (pf) ranging from 2% to 8% (assuming recall values of 65% to 90%). In our experience, it is very rare to achieve such very low false alarm rates, especially from noisy data relating to student retention. Hence, the Atwell et al., results are somewhat surprising.

- In DeLong et al. (2007), the precision varied from 57% to 60%. From our equations, we can estimate their false alarm rates in the range of 49% to 63% (assuming recall values of 65% to 90%). Such high false alarm rates are deprecated.
- In Pittman (2008), the reported precision varied from 44% to 63%. Our equations comment that these values are numerically unobtainable. For $0.78 \leq acc \leq 0.81$, $neg = 17139$ and $pos = 21136 - neg$, the equations only solve for $prec \leq 50$. That is, half the precision values reported by Pittman need to be reviewed.

In summary, learning predictors for student retention is very difficult. Despite decades of work, there is considerable room for improvement in the methods used to find patterns in student retention. As we show below, such improvements are possible if we augment standard data miners with some extra pre-processors.

3. Data

Data used in this study were from a mid-size public university, and were extracted from the student information system on official census dates. These data consisted all first-year freshmen's demographic, academic, and financial aid information (more than 100 attributes), as of the census reporting dates (after two weeks of semester starting date). As the higher education administrators may design effective policies when the students begin their studies, it is important to note that our emphasis was on detecting patterns based only on the first-term data, and that too only beginning of the term data. We created three dependent variables: RET1, if the student returned after one year; RET2, if the student returned after two years; and RET3, if the student returned after three years. The overall distribution of these dependent variables is given in Table 2. For the studied time period, the overall first-year retention rate was 71.3%, the second-year persistence rate was 60.4%, and the third-year persistence rate was 54.8%.

[Table 2 about here.]

In the Integrated Postsecondary Education Data System (IPEDS), for the U.S. only, degree-granting, Doctoral degree offering, 4-year and above institutes (excluding University of Phoenix-Online Campus), and cohort size greater than 3,000, we found that the full-time freshmen retention rate had a range from 59% to 96%, and the cohort size had a range from 3,117 to 8,025. In this list of institutions, Kent state university ranked 38 in the full-time retention percentage and 26 in the cohort size (Department of Education, 2010). Thus, Kent state data are representative of other similar size universities, and the data mining approach could be generalized to other universities.

3.1. Attribute Groups

The data mining methods discussed below used *attribute selection* to prune measurements that are poor predictors for the target class. Therefore, our data miners can be used to assess various hypotheses relating to student retention:

- If an hypothesize claims that attributes X, Y, Z are important...
- ... and if our learners prune those attributes ...

- ... then that is evidence against that hypothesis.

Accordingly, before applying our data miners, we take care to divide our attributes into the active hypothesis that they supprot:

H1: The *financial aid* hypothesis. Sanjeev & Zytkow (1995) found that no amount of financial aid influenced students to enroll for more terms; whereas Herzog (2005) found that upper-income students had reduced dropout odds compared to those from middle and lower incomes. According to John (2000), "the research literature remains ambiguous" regarding the influence financial aid on recruitment and retention.

H2: The *academic performance* hypothesis. Although there is no doubt that high school GPA and high school preparedness has a significant impact on persistence, researchers have often questioned the effects of standardized college entrance examinations (ACT/SAT). Waugh et al. (1994) found that SAT and ACT scores had no relationship with retention, whereas Murtaugh et al. (1999) found that SAT scores had some predictive value, although inferior compared to high school GPA. DesJardins et al. (2002) noted that high GPA lowered the risk of dropout, but the effect diminished over time, and that the financial aid was an insignificant factor for increasing graduation, however, it indeed reduced the student stopout. In their comprehensive literature review, Lotkowski et al. (2004) found that high school GPA had the strongest relationship with college retention in the academic factors, but ACT assessment scores had a moderate impact.

H3: The *faculty tenure and experience* hypothesis. Ehrenberg & Zhang (2005) found that for every 10 percentage point increase in the percentage of part-time faculty and not on tenure-track full-time faculty, there was a 3-5 percentage point reduction in the institution's graduation rate. Jacoby (2006) found similar results at community colleges that increase in the ratio of part-time faculty had a negative impact on the graduation rates.

In the sequel, we will return to these hypotheses to comment on which were most useful for predicting student retention. Table 7 lists attributes that we grouped together under each hypothesis.

[Table 3 about here.]

4. Building the Experiment

In Section 2.2, we assert that a good data miner should do better than the simplistic ZeroR learner. Table 2 tell us that that lower bound is:

- For first year retention: 71.3%.
- For second year retention: 60.4%.
- For third year retention: 54.8%.

As discussed below, we will be able to do much better than some, but not all, of these targets. This was achieved by

- Removing spurious attributes using *feature subset selection*;
- Exploring *a large range of classifiers*;
- Assessing the learned theories by their *variance*, as well as their *median* performance.
- Assessing the learned theories by their *variance*, as well as their *median* performance.
- Study the *delta* of student factors *between* those who stay and those who are retained.

4.1. Feature Subset Selection

Table 7 shows a sample of the 103 attributes used in this study. Our pre-experimental suspicion was that some of the attributes were "noisy"; i.e. contain signals not related to the target of prediction retention. Therefore, before we learn a theory, we first explored *attribute selection*.

Note that the number of attributes to select is crucial in the analysis of the data, because it allows us to comment on the hypotheses shown in the last section. If removal of attributes from an hypothesis does not change the performance of the prediction, then that hypothesis is spurious.

In this experiment, we ranked the 103 attributes from most informative to least informative. We then built theories using the top $n \in \{5, 10, .., 100, 103\}$ ranked attributes. Attributes were then discarded if adding them in did not improve the performance of our retention predictors.

The attributes were ranked using one of four methods: CFS, Information Gain, chi-squared, and One-R. *Correlation-based feature selection* constructs a matrix of feature to feature, and feature-to-class correlations (Hall, 2000). CFS uses a best first search by expanding the best subsets until no improvement is made, in which case the search falls to the unexpanded subset having the next best evaluation until a subset expansion limit is met.

Information Gain uses an information theory concept called *entropy*. Entropy measures the amount of uncertainty, or randomness, that is associated with a random variable. Thus, high entropy can be seen as a lack of purity in the data. Information gain, as described in Mitchell (1997) is an expected reduction of the entropy measure that occurs when splitting examples in the data using a particular attribute. Therefore an attribute that has a high purity (high information gain) is better at describing the data than the one that has a low purity. The resulting attributes are then ranked by sorted their information gain scores in a descending order.

The *chi-squared* statistic is used in statistical tests to determine how distributions of variables are different from one another (Moore & Notz, 2006). Note that these variables must be categorical in nature. Thus, the chi-squared statistic can evaluate an attribute's worth by calculating the value of this statistic with respect to a class. Attributes can then be ranked based on this statistic.

The *One-R* classifier, described below, can be used to deliver top-ranking attributes. One-R constructs and scores rules using one attribute. Feature selectors using One-R sort the attributes based on these scores.

4.2. Classifiers

In data mining, classifiers are used to learn connections between independent features and the dependent feature (called the *class*). Once these patterns are learned, we can predict outcomes in new data by reflecting on data that has already been examined.

This study tried six different classifiers: One-R, C4.5, ADTrees, Naive Bayes, Bayes networks, and radial bias networks. These are some of the well-known and standard classifiers in the machine learning field, except for ADTrees. *One-R*, described in Holte (1993), builds rules from the data by iteratively examining each value of an attribute and counting the frequency of each class for that attribute-value pair. An attribute-value is then assigned as the most frequently occurring class. Error rates of each of the rules are then calculated, and the best rules are ranked based on the lowest error rates.

A *radial basis function network* (RBFN) is an artificial neural network (ANN) that utilizes a radial basis function as an activation function (Bors, 2001). An ANN's activation function is used in order to offer non-linearity to the network. This is important for multi-layer networks containing many hidden layers, because their advantages lie in their ability to learn on non-linearly separable examples.

C4.5 (Quinlan, 1993) is an extension to the ID3 (Quinlan, 1986) algorithm. A decision tree (shown in Figure 3) is constructed by first determining the best attribute to make as the root node of the tree (Mitchell, 1997). ID3 decides this root attribute by using one that best classifies training examples based upon the attribute's information gain (described above) (Quinlan, 1986). Then, for each value of the attribute representing any node in the tree, the algorithm recursively builds child nodes based on how well another attribute from the data describes that specific branch of its parent node. The learning stops when the tree perfectly classifies all training examples, or when all attributes used. C4.5 extends ID3 by making several improvements, such as the ability to operate on both continuous as well as discrete attributes, training data that contains missing values for a given attribute(s), and employ pruning techniques on the resulting tree.

[Figure 3 about here.]

ADTrees are decision trees that contain both decision nodes, as well as prediction nodes (Freund & Mason, 1999). Decision nodes specify a condition, while prediction nodes contain only a number. Thus, as an example in the data follows paths in the ADTree, it only traverses branches whose decision nodes are true. The example is then classified by summing all prediction nodes that are encountered in this traversal. ADTrees, however, differ from binary classification trees, such as C4.5, where those trees only traverses a single path down the tree.

[Figure 4 about here.]

A *naive Bayes* classifier uses Bayes' theorem to classify training data. Bayes' theorem, as shown in Equation 11, determines the probability P of an event H occurring given an amount of evidence E. This classifier assumes feature independence; the algorithm examines features independently to contribute to probabilities, as opposed to the assumption that features depend on other features. Surprisingly, even though feature independence is an integral part of the classifier, it often outperforms many other learners (Rish; Domingos & Pazzani, 1997).

$$Pr(H|E) = \frac{Pr(E|H) * Pr(H)}{Pr(E)} \tag{11}$$

Bayesian networks, illustrated in Figure 4, are graphical models that use a directed acyclic graph (DAG) to represent probabilistic relationships between variables. As stated in Heckerman (1996), Bayesian networks have four important elements to offer:

1. Incomplete data sets can be handled well by Bayesian networks. Because the networks encode a correlation between input variables, if an input is not observed, in will not necessarily produce inaccurate predictions, as would other methods.
2. Causal relationships can be learned about via Bayesian networks. For instance, we can find whether a certain action taken would produce a specific result and to what degree.
3. Bayesian networks promote the amalgamation of data and domain knowledge by allowing for a straightforward encoding of causal prior knowledge, as well as the ability to encode causal relationship strength.
4. Bayesian networks avoid over fitting of data, as "smoothing" can be used in a way such that all data that is available can be used for training.

4.3. Cross-Validation

The value of different attributes can be assessed using equations one to four. If we use multiple *hold out* test sets, we can also discover the variance in these performance figures. In this experiment, we performed a 5×5 cross-validation i.e. we partitioned the data five times into a testing set consisting of $\frac{1}{5}$-th of the data and a training set of $\frac{4}{5}$-ths of the data. After the five rounds, we recorded the median values of recall and false alarm rates.

4.4. Contrast Set Learning

After determining the subset of the attributes that best predict for student retention, we conducted a *contrast set study*. Contrast set learners like TAR3 (Menzies & Hu, 2007) seek attribute ranges that are most *different* in various outcomes. One way to read these contrast sets are as *treatments* that promise if action X was applied to a domain, then this would favor outcome X over outcome Y. In our case, we used TAR3 in two ways:

- Firstly, we will use TAR3 to find which treatments most select for retention;
- Secondly, we will run TAR3 in the opposite direction to find the treatments that most select for students leaving university.

. In the first case, TAR3 is being used to find actions that most encourage retention. In the second case, TAR3 is being used to find the worst possible actions that most increase the probability of a student leaving.

5. Analysis of Experimental Results

5.1. Evaluation Metrics

The evaluation metrics used in this experiment are standard data mining performance measures of a method. They are:

- Probability of detection (PD);
- Probability of false alarm (PF);
- And variance PD and PF seen over the our cross-validation study.

Variance in these values provides insight into how much reliability a classifier supports on the data. For example, if a method's PD values ranges from very low to very high, we can conclude that the particular method is inconsistent in its probabilities of detection. For our studies, we rejected anything with a variance greater than $\pm 25\%$.

The above statistics were collected over 1500 experiments, which were repeated 20 times (to check for conclusion stability). In all, we conducted

$$5 * 5 * 4 * 6 * 3 * 20 = 36,000$$

experiments; i.e. 5×5 cross-validation using four feature subset selectors and 6 different learners, for the 3 data sets of Section 3 (recall from §3 that those three data sets contained data about first, second, and third year retention). This was repeated 20 times using the top $n \in 5, 10, 15, .., 100, 103$ attributes as found by the feature selector.

5.2. First Results

After rejecting all results with (1) a PD lower than the ZeroR limit; (2) a PD variance greater than $\pm 25\%$; and (3) a PF higher than 25%, we found that we had no predictors for Year1 or Year2 retention. This is the first major finding for this research: *it is very difficult to predict lower year retention.* Note that this result is consistent with prior results discussed above in our literature review.

For the rest of this study, we will focus only on third year retention. The case for focusing on third year retention is quite clear:

- If the goal is to provide a complete university education for a student, then predicting survival till second year is less interesting than lasting till third year.
- Third year retention implies second and first year retention.

5.3. Ranking with the Mann-Whitney Test

After pruning results with low PD, high PF, or high PD variance, we ranked the remaining results via a Mann-Whitney test (95% confidence). We determined the ranks by counting how many times a combination won compared to another combinations. The method that won the most number of times was then given the highest rank. The table in Figure 4 shows the top ten ranking combinations based on a PD performance measure. Note: we gave identical ranks to those treatments whose win value was equal in magnitude.

[Table 4 about here.]

Since similar results were achieved using 30 or 50 attributes, we applied Occam's Razor and focused on the 30 attributes found to be best for oneR/bnet. For these 30 attributes, we studied all their ranges. Figure 4 shows the ranges which, in isolation, select for retention at a probability greater than the ZeroR limit for (for third year, that ZeroR limit is 55%). In terms of assessing different hypothesis, the third column of Figure 4 is most informative:

- The ranges shown at the top of the table are most predictive for third year retention. Note the dominance of "Financial Aid" attributes from Figure 5.
- Attributes related to student "Performance" are rarer.
- None of the attribute ranges include the "Faculty Type and Experience" attributes of Figure 5.

From this analysis, we made two tentative conclusions:

- Using experienced faculty-level instructors is *not* predictive for third year retention.
- Issues relating to financial aid dominate over student performance.

5.4. Ranking with Contrast Set Learning

The counter-case to this conclusion might be that Figure 4 only discusses the effect of attribute ranges in *isolation*. It is possible that combination of factors might lead to different conclusions. The TAR3 treatment learner was used to test this possibility. We let TAR3 build rules of up to size 10 (i.e. ten combinations of attribute ranges) from the 30 attributes selected by the best learning combination of Figure 4. It turned out that this max size of 10 ranges was much larger than necessary: TAR3 never found combinations larger than three ranges.

[Table 5 about here.]

6. Results

Figure 5 lists the rankings of all attribute ranges which, in isolation, predict for third year retention at a probability higher than the ZeroR limit (55%), and are supported by good number of records. The top six attributes affecting third-year retention were from the financial aid hypothesis: student's wages, parent's adjusted gross income, student's adjusted gross income, mother's income, father's income, and high school percentile. Of those students who reported their wages, students who made between 7,850 and 9,958 had a 79% retention. Similar rules were found for parent's income and adjusted gross income. It means that the students with stronger financial support usually stay in college than the students with weaker financial support.

After these top six attributes, high school percentile of 81 or greater was an important attribute with 69% of students returning after three years. Some other "performance" attributes were ACT scores and ranks. This supports the argument that scores do have some predictability of student retention.

TAR3 results, given in Figure 5, produced simple theories (treatments) that combined ranges of various attributes that maximized the student retention. For example, the student retention was very high for students with the AGI in the range from $7,000 to $724,724 and father's wages were in the range from $56,289 to $999,999. One more interesting theory that predicted high retention was where father's education level was 3 (college) and student's rank amongst the freshmen cohort was between 66.3 and 98.4.

Treatments that predicted student drop-out were based on the total number of classes student was enrolled, English 10000, an introductory college writing and supplemental instruction class, and on-campus living. Students who took less than five class, enrolled in the English 10000 class, and did not live on-campus were at high risk of drop-out. Chart on the bottom of Figure 5 shows the retention percentage of each treatment. For example, students enrolled in English 10000 had a 40% retention in their third year.

Key findings were:

- Student's and parent's income capacity and levels affected student retention. Third-year retention was higher for the students with high income than the students with low income. According to treatment 1, approximately 82% of students who had at least $7,000 AGI and their fathers' income was at least $56,289 returned after three years. Similarly, according to treatment 5, approximately 79% of students who made at least $5,383 and their parents' AGI was at least $87,744 returned after three years.

- Students with better high school performance amongst their peers had higher chances of retention. According to treatment 2, approximately 81% of students who had at least $7,000 AGI and had high school percentile of 72 and better returned after three years. Approximately 79% students who had at least 3.34 HS GPA and whose parents had an AGI of at least $84,744 stayed after three years, given in treatment 4.

- ACT scores, rank of these scores amongst peers, and COMPASS scores affected student retention. Students with higher scores and rank had higher chances of retention. According to treatment 3, approximately 80% of students who had at least $7,000 AGI and had ACT math score of 21 or better returned after three years. Similarly, 77% of students who had at least 23 in ACT composite (or SAT equivalent) and had an income of at least $5,383 and less than $561,500 returned after three years, given in treatment 6.

- Parent's education level had a positive effect on student retention. Students whose parents did not attend college had a lower retention compared to students whose parents did attend college. As given in treatments 7 and 10, a student was highly likely (77%) to return after three years: (7) if the mother of that student attended college, the student had a ACT composite score of 22 or better, the parents' AGI was at least $84,744; (10) if the father of that student attended college and the student's percentile rank amongst other freshmen in the cohort was at least 66.3.

- Enrolling in fewer classes (less than five), enrolling in English 10000 (an introductory college writing class), and living off-campus had a negative effect on student retention, as given in treatments 11, 12, and 13. It is important to note that enrolling in that English course itself is not a predictor of non-retention, but the

sample of the students that attended this class were at high-risk of dropping out. Given funding for further investigation, we would focus more data collection on this high-risk group.

[Figure 5 about here.]

6.1. Strategic Actions

This study provides insights in student retention domain using beginning of term data. These insights can be used to design effective policies and strategic actions, such as:

- Most of the attributes were related to socio-economic levels and capacities of students and their parents; however, this cannot be controlled while admitting students, but better support programs and calculated financial-aid packaging for students with lower economic capacities can be created.
- First-year students should be encouraged to live on-campus by providing some incentives, as on-campus students have higher chances of retention.
- Special guidance and supplemental instruction in writing and reading should be provided to first-generation students. Parents of first-year generation students have considerably low-incomes than the parents of non-first-generation students, and according to the results of this study, income of parents is a critical factor in student retention even if the students had similar academic performance.
- Students are placed in the supplemental instruction classes, such as English 10000, based on their COMPASS and ACT scores. As these students' scores indicated lack of academic preparedness in some areas, academic advisers correctly place students in such classes; however, if the students fail or perform poorly in such classes, it leaves a lasting impression and sets the students to for future drop-out, even after three years. Therefore, it is paramount that advisers not only place students in supplemental instruction classes, but also ensure the success of students in these classes and improve the skills that students lack. Out of all classes considered in this study, English seemed to have the greatest impact. Intuitive as it may be, to succeed in college, students need good writing and reading skills.

7. Conclusion

Although our techniques could not predict first or second year retention with significantly higher accuracies than the baseline, these techniques obtained probability of detection approximately 15% higher for the class value of Y and 20% higher for the class value of N than the baseline percentages for third-year retention, based on the first-year beginning of the term data. In the studied literature, we have not found any studies with such a significant improvement over the baseline for the third-year retention. In addition, if policies are designed to improve third-year retention rate (using this predictive model), not only will they improve first and second year retention rates, but also the six-year graduation rates.

For the studied institution, family background and family's social-economic status are critical for student's third-year persistence. Using feature subset selection methods, we found that the attributes from the "financial aid" hypothesis were selected the most as predictors of retention, and although the attributes from the "performance" hypothesis were selected, their predictability, in isolation, was lesser than the attributes from the "financial aid" hypothesis. None of the attributes from the "faculty tenure and experience" were selected by the feature subset selectors.

These results could very well be true only for the studied institution; however, if the approach detailed in this study is followed, other institutions can find top performing classifier and important attributes. We recommend: (a) data discretization; (b) feature subset selection with cross-validation and evaluation the performance over various learners; (c) treatment learners, such as TAR3 to find succinct strategic actions in complex data. We welcome the opportunity to study data from other institutions and willing to share the experiment platform used in this study.

References

ACT (2007). ACT National Collegiate Retention and Persistence to Degree Rates. `http://www.act.org/research/policymakers/reports/retain.html`.

Adam, A. J., & Gaither, G. H. (2005). Retention in higher education: A selective resource guide. *New Directions for Institutional Research*, *2005*, 107–122.

Atwell, R. H., Ding, W., Ehasz, M., Johnson, S., & Wang, M. (2006). Using data mining techniques to predict student development and retention. In *Proceedings of the National Symposium on Student Retention*.

Barker, K., Trafalis, T., & Rhoads, T. R. (2004). Learning from student data. *Systems and Information Engineering Design Symposium*, (pp. 79–86).

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, *12*, 155–187.

Bors, A. (2001). Introduction of the radial basis function (rbf) networks. In *Online Symposium for Electronics Engineers* (pp. 1–7). volume 1.

Bresciani, M. J., & Carson, L. (2002). A study of undergraduate persistence by unmet need and percentage of gift aid. *NASPA Journal*, *40*, 104–123.

DeLong, C., Radcliffe, P. M., & Gorny, L. S. (2007). Recruiting for retention: Using data mining and machine learning to leverage the admissions process for improved freshman retention. In *Proceedings of the National Symposium on Student Retention*.

Department of Education (2010). Integrated Postsecondary Education Data System (IPEDS). `http://nces.ed.gov/ipeds/datacenter/`.

DesJardins, S. L., Ahlburg, D. A., & McCall, B. P. (2002). A temporal investigation of factors related to timely degree completion. *The Journal of Higher Education*, *73*, 555–581.

Dey, E. L., & Astin, A. W. (1993). Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Research in Higher Education*, *34*, 569–581.

Domingos, P., & Pazzani, M. J. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, *29*, 103–130.

Druzdzel, M. J., & Glymour, C. (1994). Application of the TETRAD II program to the study of student retention in u.s. colleges. In *Working notes of the AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94)* (pp. 419–430). Seattle, WA.

Ehrenberg, R., & Zhang, L. (2005). Do Tenured and Tenure-Track Faculty Matter? *Journal of Human Resources*, *40*, 647.

Freund, Y., & Mason, L. (1999). The alternating decision tree learning algorithm. In *In Machine Learning: Proceedings of the Sixteenth International Conference* (pp. 124–133). Morgan Kaufmann.

Glynn, J., Sauer, P., & Miller, T. (2003). Signaling student retention with prematriculation data. *NASPA Journal*, *41*, 41–67.

Hall, M. (2000). Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000): June 29-July 2, 2000, Stanford University* (p. 359). Morgan Kaufmann.

Heckerman, D. (1996). *A Tutorial on Learning With Bayesian Networks*. Technical Report Learning in Graphical Models.

Herzog, S. (2005). Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. *Research in Higher Education*, *46*, 883–928.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis--vis regression. *New Directions for Institutional Research*, *131*.

Holte, R. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*, 63.

Jacoby, D. (2006). Effects of part-time faculty employment on community college graduation rates. *Journal of Higher Education*, *77*, 1081–1103.

John, E. P. (2000). The impact of student aid on recruitment and retention: What the research indicates. *New Directions for Student Services*, (pp. 61–76).

Lau, L. K. (2003). Institutional factors affecting student retention. *Education*, *124*, 126–137.

Lotkowski, V., Robbins, S., & Noeth, R. (2004). The role of academic and non-academic factors in improving college retention. *ACT Office of Policy Research*, .

Massa, S., & Puliafito, P. (1999). An application of data mining to the problem of the university students' dropout using markov chains. In *Principles of Data Mining and Knowledge Discovery. Third European Conference, PKDD'99* (pp. 51–60). Prague, Czech Republic.

Menzies, T., Dekhtyar, A., Distefano, J., & Greenwald, J. (2007). Problems with precision. *IEEE Transactions on Software Engineering*, . `http://menzies.us/pdf/07precision.pdf`.

Menzies, T., & Hu, Y. (2007). Just enough learning (of association rules): The TAR2 treatment learner. In *Artificial Intelligence Review*. Available from `http://menzies.us/pdf/07tar2.pdf`.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Moore, D., & Notz, W. (2006). *Statistics: concepts and controversies*. WH Freeman & Co.

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education*, *40*, 355–371.

NCPPHE (2007). Retention rates - first-time college freshmen returning their second year (ACT).

Pascarella, E. T., & Terenzini, P. T. (1979). Interaction effects in spady and tinto's conceptual models of college attrition. *Sociology of Education*, *52*, 197–210.

Pascarella, E. T., & Terenzini, P. T. (1980). Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *The Journal of Higher Education*, *51*, 60–75.

Pittman, K. (2008). *Comparison of data mining techniques used to predict student retention*. Ph.D. thesis Nova Southeastern University.

Quinlan, J. R. (1986). *Induction of decision trees*. (1st ed.).

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. (1st ed.). Morgan Kaufmann.

Rish, I. (). An empirical study of the naive bayes classifier. In *IJCAI-01 workshop on "Empirical Methods in AI"*. `http://www.intellektik.informatik.tu-darmstadt.de/ tom/IJCAI01/Rish.pdf`.

Salazar, A., Gosalbez, J., Bosch, I., Miralles, R., & Vergara, L. (2004). A case study of knowledge discovery on academic achievement, student desertion and student retention. *Information Technology: Research and Education, 2004. ITRE 2004. 2nd International Conference on*, (pp. 150–154).

Sanjeev, A., & Zytkow, J. (1995). Discovering enrolment knowledge in university databases. In *First International Conference on Knowledge Discovery and Data Mining* (pp. 246–51). Montreal, Que., Canada.

Scalise, A., Besterfield-Sacre, M., Shuman, L., & Wolfe, H. (2000). First term probation: models for identifying high risk students. In *30th Annual Frontiers in Education Conference* (pp. F1F/11–16 vol.1). Kansas City, MO, USA: Stripes Publishing.

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, *1*, 64–85.

Spady, W. G. (1971). Dropouts from higher education: Toward an empirical model. *Interchange*, *2*, 38–62.

Stage, F. (1989). Motivation, Academic and Social Integration, and the Early Dropout. *American Educational Research Journal*, *26*, 385–402.

Stewart, D. L., & Levin, B. H. (2001). A model to marry recruitment and retention: A case study of prototype development in the new administration of justice program at blue ridge community college.

Sujitparapitaya, S. (2006). Considering student mobility in retention outcomes. *New Directions for*

Institutional Research, *2006*.

Superby, J. F., Vandamme, J. P., & Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. In *8th International Conference on Intelligent Tutoring Systems (ITS 2006)* (pp. 37–44). Jhongli, Taiwan.

Terenzini, P. T., & Pascarella, E. T. (1980). Toward the validation of tinto's model of college student attrition: A review of recent studies. *Research in Higher Education*, *12*, 271–282.

Tillman, C., & Burns, P. (2000). Presentation on First Year Experience. `http://www.valdosta.edu/ cgtillma/powerpoint.ppt`.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*, 89–125.

Tinto, V. (1982). Limits of Theory and Practice in Student Attrition. *The Journal of Higher Education*, *53*, 687–700.

Tinto, V. (1988). Stages of student departure: Reflections on the longitudinal character of student leaving. *Journal of Higher Education*, *59*, 438–455.

Vandamme, J. (2007). Predicting Academic Performance by Data Mining Methods. *Education Economics*, *15*, 405–419.

Veitch, W. R. (2004). Identifying characteristics of high school dropouts: Data mining with a decision tree model.

Waugh, G., Micceri, T., & Takalkar, P. (1994). Using ethnicity, SAT/ACT scores, and high school GPA to predict retention and graduation rates.

Witten, I., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. (2nd ed.). San Francisco: Morgan Kaufmann Publishers.

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., Lo, W., & Kaprolet, C. (2007). A data-mining approach to differentiate predictors of retention between online and traditional students.

Zhang, H., & Zhang, X. (2007). Comments on 'data mining static code attributes to learn defect predictors'. *IEEE Transactions on Software Engineering*, .

Żytkow, J., & Zembowicz, R. (1993). Database exploration in search of regularities. *Journal of Intelligent Information Systems*, *2*, 39–81.

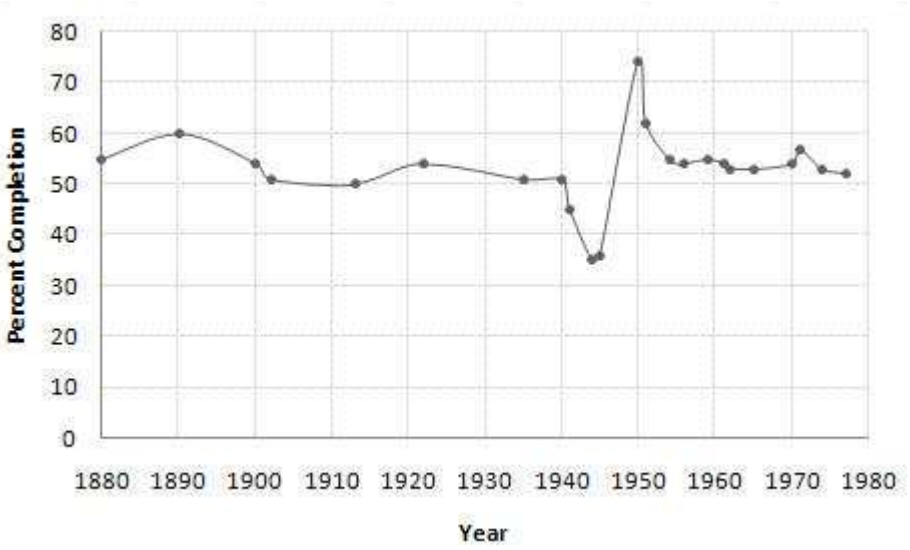
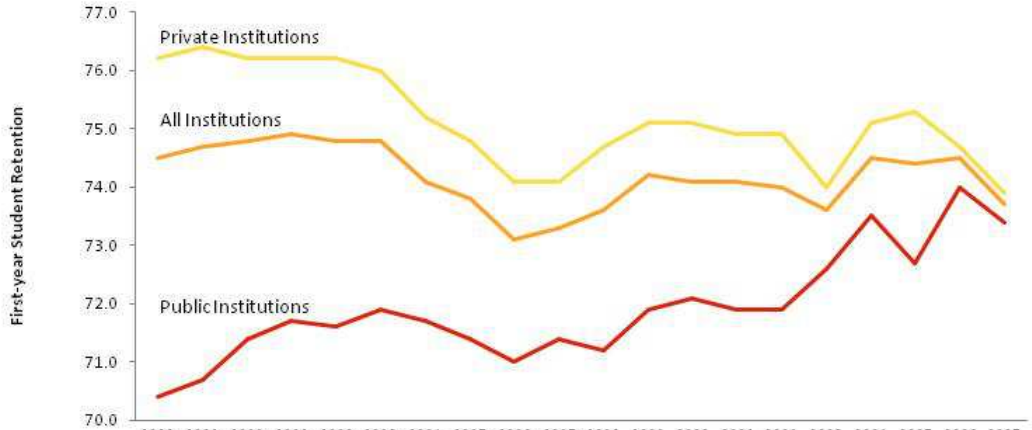


Figure 1: BA Degree Completion Rates for the period 1880 to 1980, where Percent Completion is the Number of BAs Divided by the Number of First-time Degree Enrollment Four Years Earlier (Tinto, 1982)



	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Private Institutions	76.2	76.4	76.2	76.2	76.2	76.0	75.2	74.8	74.1	74.1	74.7	75.1	75.1	74.9	74.9	74.0	75.1	75.3	74.7	73.9
All Institutions	74.5	74.7	74.8	74.9	74.8	74.8	74.1	73.8	73.1	73.3	73.6	74.2	74.1	74.1	74.0	73.6	74.5	74.4	74.5	73.7
Public Institutions	70.4	70.7	71.4	71.7	71.6	71.9	71.7	71.4	71.0	71.4	71.2	71.9	72.1	71.9	71.9	72.6	73.5	72.7	74.0	73.4

Figure 2: Percentage of First-Year Students at Four-Year Colleges Who Return for Second Year (ACT, 2007)

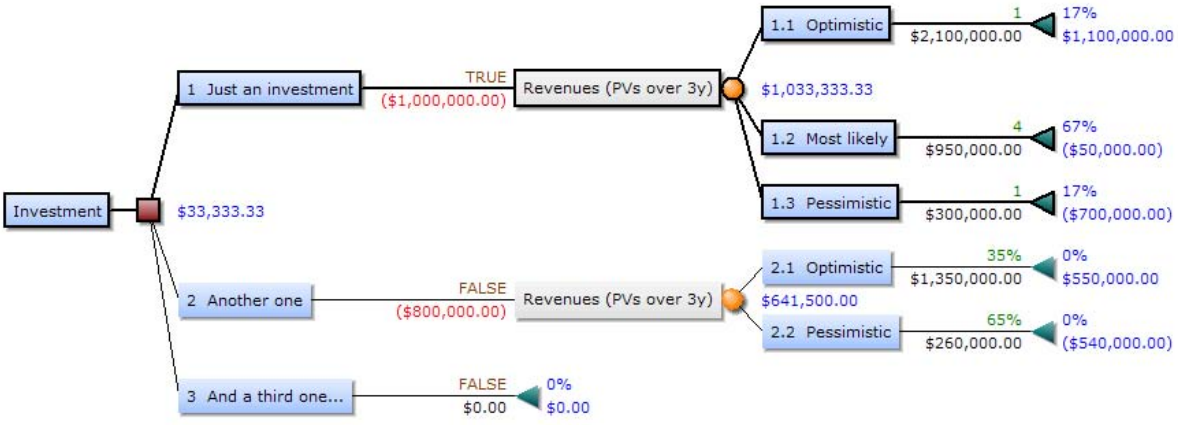


Figure 3: A decision tree consists of a root node and descending children nodes who denote decisions to make in the tree's structure. This tree, for example, was constructed in an attempt to optimize investment portfolios by minimizing budgets and maximizing pay-offs. The top-most branch represents the best selection in this example.

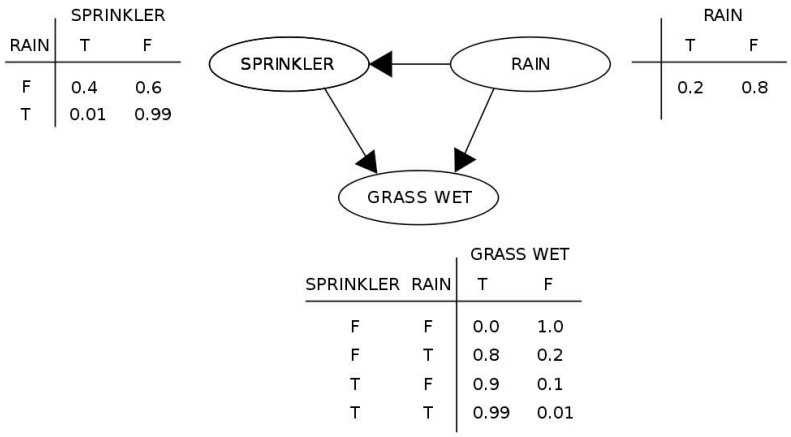


Figure 4: In this simple Bayesian network, the variable *Sprinkler* is dependent upon whether or not it's raining; the sprinkler is generally not turned on when it's raining. However, either event is able to cause the grass to become wet - if it's raining, or if the sprinkler is caused to turn on. Thus, Bayesian networks excel at investigating information relating to relationships between variables.

#	Treatment
1	7000 $\leq$ FinAidSTUDENT_AG $<$ 724,724 and 56,289 $\leq$ FinAidFATHER_WAG $<$ 999,999
2	7,000 $\leq$ FinAidSTUDENT_AG $<$ 724,724 and HS_PERCENT $\geq$ 72
3	7,000 $\leq$ FinAidSTUDENT_AG $<$ 724,724 and 21 $\leq$ ACT1_MATH $<$ 36
4	84,744 $\leq$ FinAidPARENT_AGI $<$ 999,999 and HS_GPA $\geq$ 3.34
5	84,744 $\leq$ FinAidPARENT_AGI $<$ 999,999 and 5383 $\leq$ FinAidSTUDENT_WA $<$ 561,500
6	23 $\leq$ MaxACT $<$ 35 and 5383 $\leq$ FinAidSTUDENT_WA $<$ 561,500
7	22 $\leq$ ACT1_COMP $<$ 35 and 84,744 $\leq$ FinAidPARENT_AGI $<$ 999,999 and FinAidMOTHER_ED=3
8	5383 $\leq$ FinAidSTUDENT_WA $<$ 561,500 and 21 $\leq$ ACT1_MATH $<$ 36
9	HS_GPA $\geq$ 3.34 and 32,570 $\leq$ FinAidMOTHER_WAG $<$ 533,395
10	FinAidFATHER_ED=3 and 66.3 $\leq$ PercentileRankHSGPA $<$ 98.4
11	1 $\leq$ TotalClass$\leq$ 5
12	ENG10=Y
13	LIVE.ON.CAMP=N

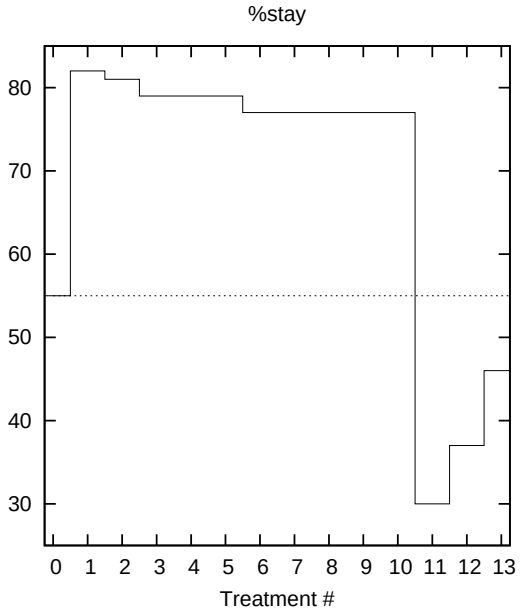


Figure 5: Treatments 1 to 10 are the top ten treatments found by this analysis that increases the third year retention rates. Treatments 11,12,13 are the worst three treatments found by this analysis that *most decrease* the third year retention rates. The effects of each treatment, is shown on the bottom plot.

Author (Year)	**Notes**	**Cohort Size**	**Retained (#)**	**Retained (%)**	**Measure of Accuracy**	**Coeffes Used?**	**Techniques Used**
Spady (1971)		683	615	90.04%	R^2 of .3132 for men and .3879 for women	Yes	Multiple regression
Bean (1980)		906	769	84.88%	R^2 of .22 for women and 0.09 for men	Yes	Multiple regression
Terenzini (1980)	study 1	379	60	15.80%	R^2 of .246	Yes	discriminate analyses
	study 3	518	428	82.63%	R^2 of .256	Yes	Multiple regression
	study 5	763	673	88.20%	R^2 of .309	Yes	discriminate analyses
	study 6	763	673	88.20%	R^2 of .476 for men and .553 for women	Yes	discriminate analyses
Stage (1989)		323	294	91.00%		Yes	Logistic regression
Dey & Astin (1993)		947	152	16.00%	Multiple R 0.354, 0.351, and 0.323	Yes	logit, probit, and regression
Murtaugh et al. (1999)		8667	5200	60%	estimated ret prob 59.3%	Yes	Survival Analysis/ Hazard regression
Bresciani & Carson (2002)		3535	3121	88.30%	R^2 of 0.022	Yes	Logistic regression
Glynn et al. (2003)	any dropout; not only first-year; accuracies based on the training data	3244	1592	49.08%	overall accuracy of 83%	Yes	Logistic regression
Herzog (2005)		5261	4014	76.30%	77.4% accuracy	Yes	Logistic regression
		4298	3314	77.10%		Yes	
		4671	4040	83.50%	85.4% accuracy	Yes	
Sujitparapitaya (2006)		2,444	1943	79.50%	81.6% accuracy on training; 80.7% on validation	Yes	Logistic regression
		2,445	1994	79.50%	83.9% accuracy on training; 82.1% on validation		Neural Network
		2,445	1994	79.50%	85.5% on training; 84.4% on validation		C4.5
Herzog (2006)		8,018	6037	75.29%	accuracy close to 75%		Neural Networks; CHAID, C4.5, CR&T; Logistic regression
Atwell et al. (2006)	training	3,829	3149	82.24%	precision for drop-outs 91, 84, 84, 78		decision trees (entropy, chi-sq, gini) and
	test	5,990	4,881	81.49%	precision for drop-outs 88, 82,82,73		logistic regression
DeLong et al. (2007)				50%	precision varied from 57% to 60%		AdaBoost M1 with Decision Stumps
Pittman (2008)		21136	17139	81.10%	overall accuracy of 78-81%; not-retained precision from 44-63%		Logistic regression, neural network, bayes, J48

Table 1: Techniques and Accuracies Reported in Literature

	RET1		**RET2**		**RET3**	
	Count	Percentage	Count	Percentage	Count	Percentage
retained=**Y**	24,039	71.3%	18,055	60.4%	14,362	54.8%
retained=**N**	9,673	28.7%	11,857	39.6%	11,854	45.2%
Total	33,712	100%	29,912	100%	26,216	100%

Table 2: Distribution of Dependent Variables

Category	Financial Aid	**Category**	Performance Indicators	**Category**	Faculty Type & Experience
Attribute	**Description**	**Attribute**	**Description**	**Attribute**	**Description**
FinAidAwardType_G	Financial aid amount of grants	ACT_COMP	ACT comprehensive score (old)	FacExpLT1Cnt	Count of courses taught by faculty [CCTF] w/ less than 1 yr experience
FinAidAwardType_J	Financial aid amount of jobs	ACT_ENGL	ACT english score (old)	FacExpLT1Ratio	Ratio of courses taught by faculty [RCTF] w/ less than 1 yr experience to the total courses
FinAidAwardType_L	Financial aid amount of loans	ACT_MATH	ACT math score (old)	FacExp1to5Cnt	CCTF w/ experience between 1 & 5
FinAidAwardType_S	Financial aid amount of scholarship	ACT1_COMP	ACT comprehensive score (new)	FacExp1to5Ratio	RCTF w/ experience between 1 & 5 to the total courses
FinAidAwardType_W	Financial aid amount of waiver	ACT1_ENGL	ACT english score (new)	FacExp6to10Cnt	CCTF w/ experiencAe between 6 & 10
FinAidDEPENDENCY	Dependency status	ACT1_MATH	ACT math score (new)	FacExp6to10Ratio	RCTF w/ experience between 6 & 10 to the total courses
FinAidFATHER_ED	Father's education level	ACTEQUIV	ACT equivalent of the sat score	FacExp11to15Cnt	CCTF w/ experience between 11 & 15
FinAidFATHER_WAG	Father's income	MaxACT	Max of ACT score and ACT equivalent	FacExp11to15Ratio	RCTF w/ experience between 11 & 15 to the total courses
FinAidMOTHER_ED	Mother's education level	COMP_READ	Compass read score	FacExp16to20Cnt	CCTF w/ experience between 16 & 20
FinAidMOTHER_WAG	Mother's income	COMP_WRITE	Compass write score	FacExp16to20Ratio	RCTF w/ experience between 16 & 20 to the total courses
FinAidOfferedInd	Financial aid offered indicator	SAT_TOT	SAT total score	FacExp21to25Cnt	CCTF w/ experience between 21 & 25
FinAidPARENT_AGI	Parent's adjusted gross income	SAT_VERB	SAT verbal score	FacExp21to25Ratio	RCTF w/ experience between 21 & 25 to the total courses
FinAidPARENT_HOU	Parents's household size	HS_CODE	High school code	FacExp25to30Cnt	CCTF w/ experience between 25 & 30
FinAidPARENT_MAR	Parents's martial status	HS_GPA	High school gpa	FacExp25to30Ratio	RCTF w/ experience between 25 & 30 to the total courses
FinAidPARENT_TAX	Parents's tax form type	HS_PERCENT	High school percentile	FacExpGT31Cnt	CCTF w/ experience greater than 31 years
FinAidSPOUSE_WAG	Spouse's wages	HS_RANK	High school rank	FacExpGT31Ratio	RCTF w/ experience greater than 31 years to the total courses
FinAidSTUDENT_AG	Student's adjusted gross income	HS_SIZE	High school class size	NoTenureFacCnt	Count of courses taught by no rank faculty
FinAidSTUDENT_HO	Student's household size	RankHSGPA	Percentile of hs gpa among all freshmen	NoTenureFacRatio	Ratio of courses taught by no rank faculty to the total courses
FinAidSTUDENT_MA	Student's martial status	RankMaxACT	Percentile of max ACT among all freshmen	NTTFacCnt	Count of courses taught by ntt faculty
FinAidSTUDENT_TA	Student's tax form type	ANTH18	Enrolled in anthropology course	NTTFacRatio	Ratio of courses taught by ntt faculty to the total courses
FinAidSTUDENT_WA	Student's wage	BSCI10	Enrolled in biological science course	TTFacCnt	Count of courses taught by tenured/tenure-track faculty
FirstGenInd	First generation indicator	CHEM10	Enrolled in chemistry course	TTFacRatio	Ratio of courses taught by tenured/tenure-track faculty to the total courses
TotalFinAidOffered	Total financial aid offered	ENG10	Enrolled in English course		
		ENG11	Enrolled in English course		
		GEOL11	Enrolled in geology course		
		LEST16	Enrolled in leisure studies course		
		MATH10	Enrolled in math 100 level course		
		MATH11	Enrolled in math 110 level course		
		MATH12	Enrolled in math 120 level course		
		MATH14	Enrolled in math 14 level course		
		PHY11	Enrolled in physics 11 level course		
		PEP15	Enrolled in physical ed 15 level course		

Table 3: List of Attributes by Stated Hypotheses

Rank	Number of Attributes	FSS	Classifier
61	30	oneR	bnet
61	50	cfs	adtree
57	50	oneR	adtree
56	30	oneR	adtree
55	30	cfs	adtree
52	50	oneR	bnet
51	30	infogain	adtree
51	30	cfs	bnet
48	50	infogain	adtree

Table 4: The top ten ranking treatments for third year retention. Ranks represent how many times a particular treatment wins over all other treatments in the experiment.

#	$P(Ret3\|X)$ (percent)	Support (#students)	Hypothesis	X (Feature = Range) Feature	Range
1	79	1,752	Financial Aid	Student's Wage	7850 to 9958
2	73	3,751	Financial Aid	Parent's Adjusted Gross Income	96636 to inf
3	71	4,152	Financial Aid	Student's Adjusted Gross Income	4830 to 7916
4	71	3,148	Financial Aid	Mother's Income	42957 to inf
5	70	5,873	Financial Aid	Father's Income	52366 to inf
6	69	5,622	Financial Aid	Student's Wage	4093 to 7851
7	69	5,838	Performance	High School Percentile	81 to inf
8	68	2,523	Financial Aid	Student's Dependency Status	I
9	68	7,502	Financial Aid	Father's Education Level	3
10	67	6,045	Financial Aid	Parent's Adjusted Gross Income	58551 to 96636
11	66	12,215	Financial Aid	Student's Tax Form	2
12	66	7,710	Financial Aid	Mother's Education Level	3
13	66	2,057	Financial Aid	Student's Wage	1.5 to 1000
14	66	10,370	Financial Aid	First Generation Student	N
15	65	7,082	Performance	ACT Math Score (new)	23 to inf
16	65	2,780	Financial Aid	Student's Adjusted Gross Income	3336 to 4830
17	65	4,676	Performance	ACT English Score (new)	25 to inf
18	65	5,669	Performance	ACT Comprehensive Score (new)	24 to inf
19	65	13,101	Financial Aid	Parent's Tax Form	1
20	65	11,328	Financial Aid	Parent's Marital Status	M
21	64	6,658	Performance	Percentile Of Max ACT Among Freshmen	71 to inf
22	64	6,952	Performance	Max Of ACT Score And ACT Equivalent	24 to inf
23	64	2,697	Financial Aid	Student's Tax Form	1
24	63	17,254	Financial Aid	Student's Marital Status	U
25	63	13,063	Financial Aid	Mother's Wages	-inf to 42957
26	63	3,126	Financial Aid	Parent's Tax Form	2
27	63	1,022	Financial Aid	Student's Adjusted Gross Income	16714 to inf
28	62	15,154	Financial Aid	Dependency	D
29	62	9,459	Financial Aid	Father's Income	-inf to 52366
30	61	8,792	Financial Aid	Mother's Education Level	2
31	61	8,461	Financial Aid	Father's Education Level	2
32	61	4,176	Financial Aid	Student's Wage	1904 to 4093
33	60	14,523		Total Enrolled Hours	15 to 19
34	60	2,540	Financial Aid	Student's Adjusted Gross Income	1895 to 3336
35	59	3,780	Performance	Compass Writing Score	-inf to 10
36	59	8,271	Performance	ACT English Score (new)	20 to 25
37	59	7,311		FirstGenInd	Y
38	59	6,980	Performance	HS_PERCENT	61 to 81
39	59	15,021	Performance	Total Number of Enrolled Classes	6 to inf
40	59	5,598	Financial Aid	Parent's Adjusted Gross Income	1838 to 58551
41	58	3,127	Financial Aid	Parent's Marital Status	S
42	58	8,667	Performance	ACT Composite	20 to 24
43	58	4,767	Performance	ACT Math Score (new)	20 to 23
44	58	13,887	Performance	Compass Writing Score	74 to inf
45	58	5,769	Performance	High School GPA	3.02 to 3.4
46	58	10,281	Performance	RankMaxACT	31 to 71
47	58	10,044	Performance	MaxACT	20 to 24
48	57	20,087		On-Campus Indicator	Y
49	57	11,36	Financial Aid	Father's Education Level	4
50	56	24,826		Age of Student at Matriculation	-inf to 19.5
51	56	24,407	Performance	Enrolled in English Courses	N
52	56	3,660	Performance	Percentile Of Hs Gpa Among Freshmen	46 to 60

Table 5: Ranking all attribute ranges which, in isolation, predict for third year retention at a probability higher than the ZeroR limit (55%). From the above, the strongest predictor for third year retention is a student's wage (at 79%). On the other hand, the bottom line of this table says that the percentile of a student amongst their Freshmen cohort is little better than ZeroR (at 56%).