Data Mining and Student Persistence

Ashutosh Nandeshwar
Kent State University

Tim Menzies
West Virginia University

Adam Nelson
West Virginia University

January 17, 2010

Abstract

Higher education institutions are plagued with the problem of student persistence. Close to 45% of first-year freshmen do not return their second year (Druzdzel & Glymour 1994). Numerous researchers have studied this problem using traditional and non-traditional predictive modeling techniques. In this paper, we present data mining experiments on the first-year freshmen data from a mid-size public institution. Results of these experiments have performed better than any reported model in the studied literature.

KEYWORDS: data mining, student persistence, predictive modeling

1 Introduction

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 with the exception of 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 a long period 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.

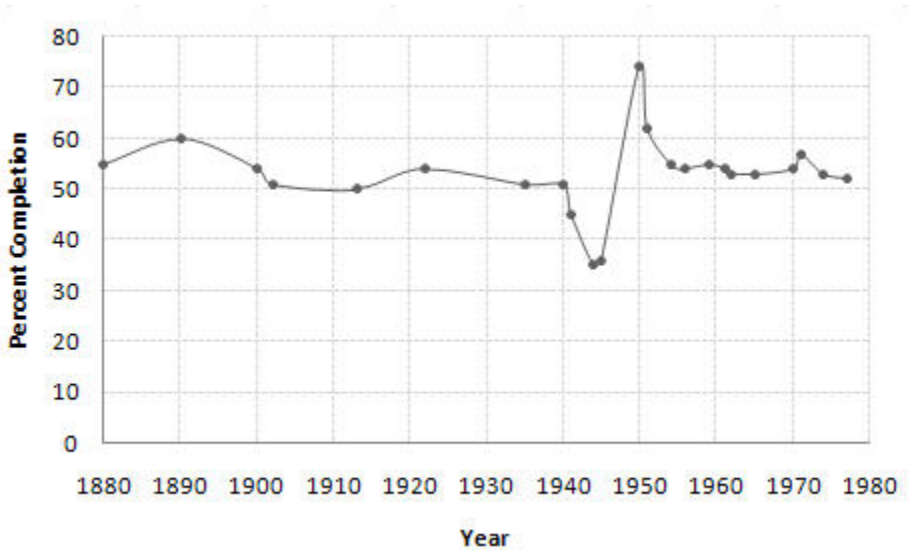
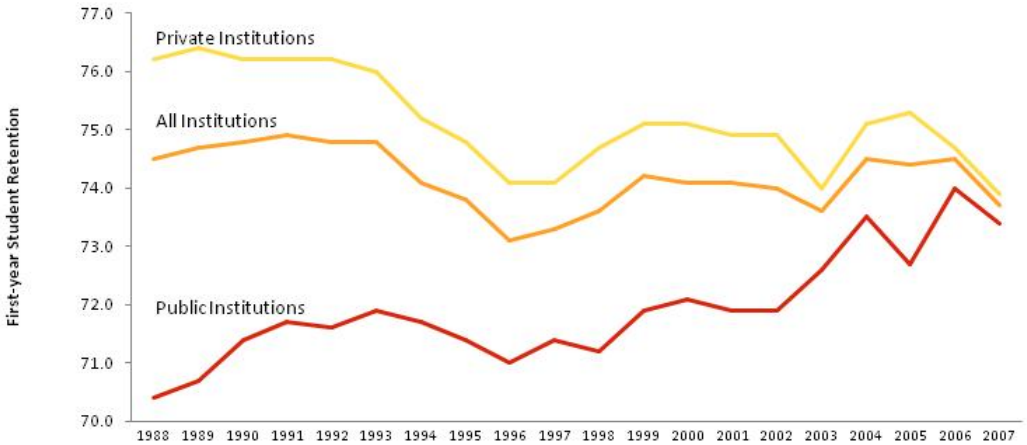


Figure 1: BA Degree Completion Rates for the period 1880 to 1980, where Percent Completion is the Number of BAs Divided by the Number of First-time Degree Enrollment Four Years Earlier (Tinto 1982)

Various researchers have studied this problem extensively, using theoretical models (Tinto 1975, 1988;



	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)

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). In the studied literature, however, we noted that there was a need of a thorough data mining experiment, which covered three main areas: discretization, attribute selection, and cross-validation over various algorithms. This paper presents a brief review of literature, research experiment, results, and discussion on the results.

2 Literature Review

Literature on retention in higher education is extensive. Rather than rehashing that information we suggest readers, who are unfamiliar with this topic and its research, to go through the excellent resource guide on retention in higher education (Adam & Gaither 2005). We, however, present literature review on the application of data mining to the student retention problem.

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 rate was 15.79%, and 10.36% of students, who did drop out were classified as non-dropouts. 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 lists techniques used in the studied literature, where the cohort sizes were available, along with the reported accuracies or measures of accuracy. Apart from results obtained by Glynn et al. (2003), all other studies performed worse or marginally better than the baseline retention percentage, and Glynn et al. (2003) reported results based only on the training data and not on the test data. Regardless of poor fits, many authors used regression coefficients to indicate the relationship of an attribute to the retention outcome.

2.1 Measures of Performance

In predictive modelling and machine learning, some of the common measures of performance are: probability of detection (PD), also called as recall, probability of false alarm (PF), precision, and accuracy. These measures are defined in Equations (1) to (4) respectively (Zhang & Zhang 2007; Menzies et al. 2007). According to Menzies et al. (2007), high recall and precision values can only be achieved if the probability of false alarm is very low, because of the relationship of PF to PD and precision, as defined in Equation (7).

$$pd = recall = \frac{D}{B+D} \tag{1}$$

$$pf = \frac{C}{A+C} \tag{2}$$

$$prec = precision = \frac{D}{D+C} \tag{3}$$

$$acc = accuracy = \frac{A+D}{A+B+C+D} \tag{4}$$

$$\frac{neg}{pos} = \frac{A+C}{B+D} \tag{5}$$

Where A, B, C, D are the true negatives, false negatives, false positives, and true positives respectively. Using the Equation(5), a relationship of pf and PD and precision can be found, as shown in Equation (6),

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		91.00%		Yes	Logistic regression
Dey & Astin (1993)		947		16.00%	Multiple R 0.354, 0.351, and 0.323	Yes	logit, probit, and regression
Waugh et al. (1994)						Yes	
Murtaugh et al. (1999)		8667		60%	estimated ret prob 59.3%	Yes	Survival Analysis/ Hazard regression
Bresciani & Carson (2002)		3535		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		76.30%	77.4% accuracy	Yes	Logistic regression
		4298		77.10%		Yes	
		4671		83.50%	85.4% accuracy	Yes	
Sujitparapitaya (2006)		2,444		79.50%	81.6% accuracy on training; 80.7% on validation	Yes	Logistic regression
		2,445		79.50%	83.9% accuracy on training; 82.1% on validation		Neural Network
		2,445		79.50%	85.5% on training; 84.4% on validation		C4.5
Herzog (2006)		8,018		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

which after rearranging becomes Equation (7).

$$prec = \frac{D}{D+C} = \frac{1}{1+\frac{C}{D}} = \frac{1}{1+\frac{neg}{pos}\cdot\frac{pf}{recall}} \tag{6}$$

$$pf = \frac{pos}{neg}\cdot\frac{(1-prec)}{prec}\cdot recall \tag{7}$$

Using these relationships, we estimated PD and PF values for given accuracy, precision, and pos/neg values. For Atwell et al. (2006), where the precision varied from 73% to 88%, we estimated PF values ranging from 2% to 8%, when we used the values of PD from 65% to 90%. As it is rare to achieve very low PF values, the estimated PF values were alarming. For DeLong et al. (2007), where precision varied from 57% to 60%, we estimated the PF values in the range of 49% to 63% using the PD values of 65% to 90%. Similarly, for Pittman (2008), where the precision varied from 44% to 63%, we estimated the PF values from 1% to 24%. These estimated PF values using state-of-the-art techniques for studying retention problem warranted a thorough study.

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 71.31%, the second-year persistence rate was 60.36%, and the third-year persistence rate was 54.78%.

	RET1		**RET2**		**RET3**	
	Count	Percentage	Count	Percentage	Count	Percentage
Y	24,039	71.31%	18,055	60.36%	14,362	54.78%
N	9,673	28.69%	11,857	39.64%	11,854	45.22%
Total	33,712	100.00%	29,912	100.00%	26,216	100.00%

Table 2: Distribution of Dependent Variables

4 Building the Experiment

To construct the experiment, we first determined certain aspects to be pertinent in the final selection of top, actionable attributes in the data. The following section provides brief explanations of each method used.

4.1 Number of Attributes

An attribute in the data could be something such as GPA, or ZIPCODE. The number of attributes to select is crucial in data analysis, because it allows us to conclude how many of the selected attributes we should concentrate on. For example, suppose a data set consists of 1,000 attributes, but the results from experimentation find that only 15 of the 1000 are actually important. We can then pay subsequent attention on what actions to take based on the 15 important attributes, as opposed to the rest of the 985 attributes.

In this experiment, we chose n as the number of attributes selected in increments of 5. Thus, with a maximum of 103 attributes in each data set used in the experiment, our feature subset selectors (described below) chose 20 different intervals of n.

4.2 Classifiers

In data mining, researchers employ machine learning techniques to learn patterns in the data. Using these learned patterns, we can attempt to predict the outcomes. We can also determine how well a classifier predicts for the data. This is done by learning on a certain portion of the data, and reflecting on how well the predictions are made on the unseen data. By examining overall performance, we can make a statement about how much better one classifier predicts on a specific data set than another.

- Naive Bayes - A naive Bayes classifier is a simple and fast probabilistic classifier that uses Bayes' theorem to classify training data. Bayes' theorem, as shown in Equation 8, determines the probability P of an event H occurring given an amount of evidence E. The classifier also 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 2001).

$$Pr(H|E) = \frac{Pr(E|H) * Pr(H)}{Pr(E)} \tag{8}$$

- C4.5 - C4.5 is a decision tree classifier (Quinlan 1993), and is an extension to the ID3 algorithm (Quinlan 1986). A decision tree (shown in Figure 3) is constructed by first determining the best attribute as the root node of the tree (Mitchell 1997a). ID3 decides the root attribute that best

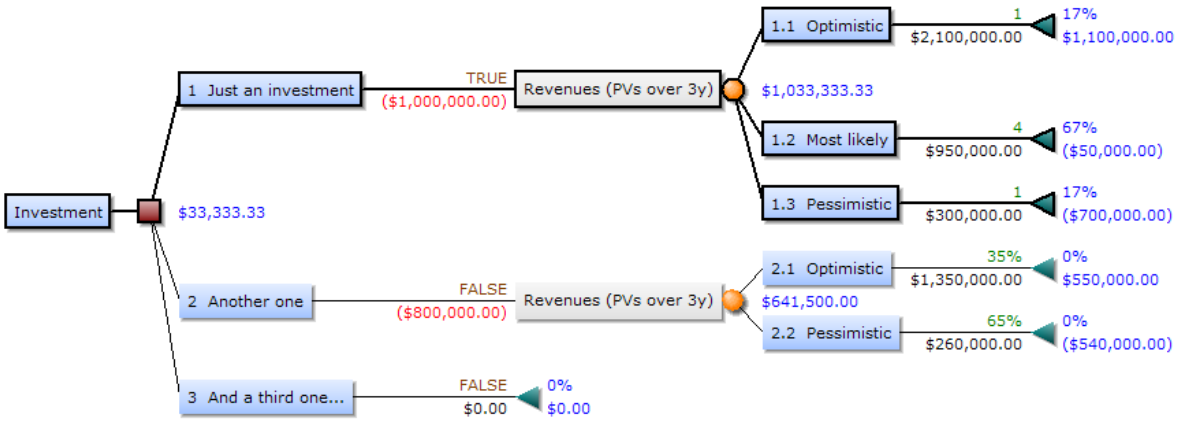


Figure 3: A decision tree consists of a root node and descending children nodes who denote decisions to make in the tree's structure. This tree, for example, was constructed in an attempt to optimize investment portfolios by minimizing budgets and maximizing payoffs. The top-most branch represents the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

classifies training examples using information gain of the attributes(described below). 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 stopping criteria are either when the tree perfectly classifies all training examples, or until no attribute remains unused. C4.5 extends ID3 by making several improvements, such as operating on both continuous as well as discrete attributes, handling training data that contains missing values for a given attribute(s), and employing pruning techniques on the resulting tree.

- One-R - 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 the most frequently occurring class. Error rates of each of the rules can then be calculated, and the best rules can be ranked based on the lowest error rates.

- Zero-R - Often used to evaluate the success of other classification algorithms, Zero-R is an extremely simple algorithm that returns the majority class from the training data.

- Alternating Decision Trees - 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, in that in those trees an example only traverses a single path down the tree.

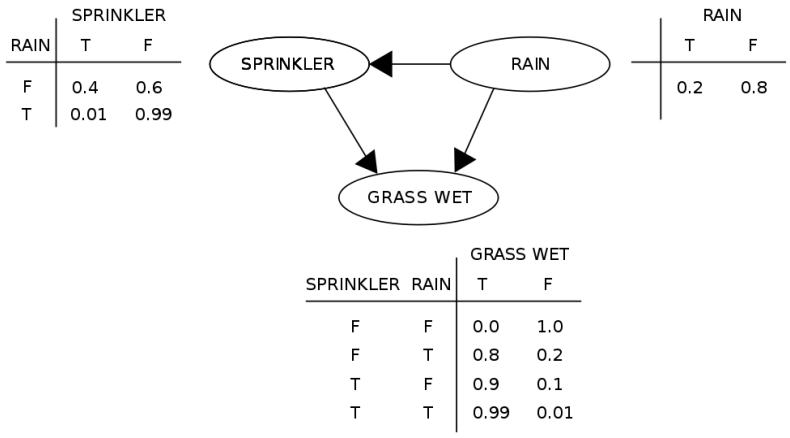


Figure 4: In this simple bayesian network, the variable *Sprinkler* is dependent upon whether or not its raining; the sprinkler is generally not turned on when it's raining. However, either event is able to cause the grass to become wet - if it's raining, or if the sprinkler is caused to turn on. Thus,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

- Bayesian Network - 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, if an analyst wished to know if a certain action taken would produce a specific result, and also 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.

- Radial Basis Function Network - A radial basis function network (RBFN) is a type of an artificial neural network (ANN) (Bors 2001), and they utilize a radial basis function as an activation function. An ANN's activation function is used to introduce 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.

4.3 Feature Subset Selectors

Feature Subset Selection (FSS) methods provide ways to determine how important the attributes (or features) are in the data set, and how we can keep the best scoring ones, and throw out the rest. However, we must experiment with varying FSS procedures, because each method can return strikingly different results. Thus, just by experimenting with attributes selected from a handful of FSS, we are not left with a sense of how well attributes were selected from a data set compared to other feature selection tools.

A brief overview of the FSS methods used in this study were as follows:

- CFS - Correlation-Based Feature Selection begins by constructing a matrix of feature to feature and feature-to-class correlations (Hall 2000). It then performs 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 - Information Gain uses entropy, a concept from information theory. 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 (1997b), 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 one with a low purity. The resulting attributes are then ranked by their information gain scores in a descending order.

- Chi-squared - Attributes can also be ranked using the chi-squared statistic. 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.

- One-R - One-R (as described above), can also be used to deliver top-ranking attributes. Since each rule contains one attribute and a corresponding value, we can then evaluate attributes by sorting them based on the error rate of the rule associated with that attribute-value pair. Using this method, we can determine top ranking attributes whose rules have the lowest error rates.

4.4 Cross-Validation

In the process of experimentation, it is crucial to determine a method's performance. Using performance criteria, further analysis can be conducted on experimental results to aid in the search for an optimal solution.

Cross-validation provides the ability to discover how well a classifier performs on any given data set or a treatment of that data set. This is conducted by randomly partitioning the data into two subsets: the training set and the testing set. Specifically for this experiment, the data prior to partitioning has been reduced given n attributes selected using an FSS method.

In the learning phase, the classifier uses only the training subset. The testing set is then used to determine how well the concepts learned from the training phase can be applied to unseen data. However, to reduce variability, the partitioning of the data and reclassification of resulting subsets is generally conducted multiple times. In this experiment, for example, we perfomed 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 examined the median values of the validation results, and assigned to a particular combination of the above facets.

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 denotes the probability that the classifier will predict correctly for a given class, given both its correct and incorrect predictions. PF, on the other hand, is the probability that the classifier will predict incorrectly for a given class, also given its correct and incorrect predictions. We evaluated all combinations of FSS method and classifiers that maximized PD values and minimized PF values.

We also used variance in the experiment based on PD and PF values independently as an extra means of determining performance. 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. Therefore, the selected methods should have a very small variance in both PD and PF values.

5.2 Visualizing the Results

Figures 5, 6, and 7 show the PD and PF median results for first, second and third year retention against the variance of these values. Each point represents a specific combination of the number of attributes selected, the feature subset selector used to select them, and the classifier used to train on the resulting data. For example, one point on a graph could be seen as 50/Information Gain/Naive Bayes, where 50 denotes the

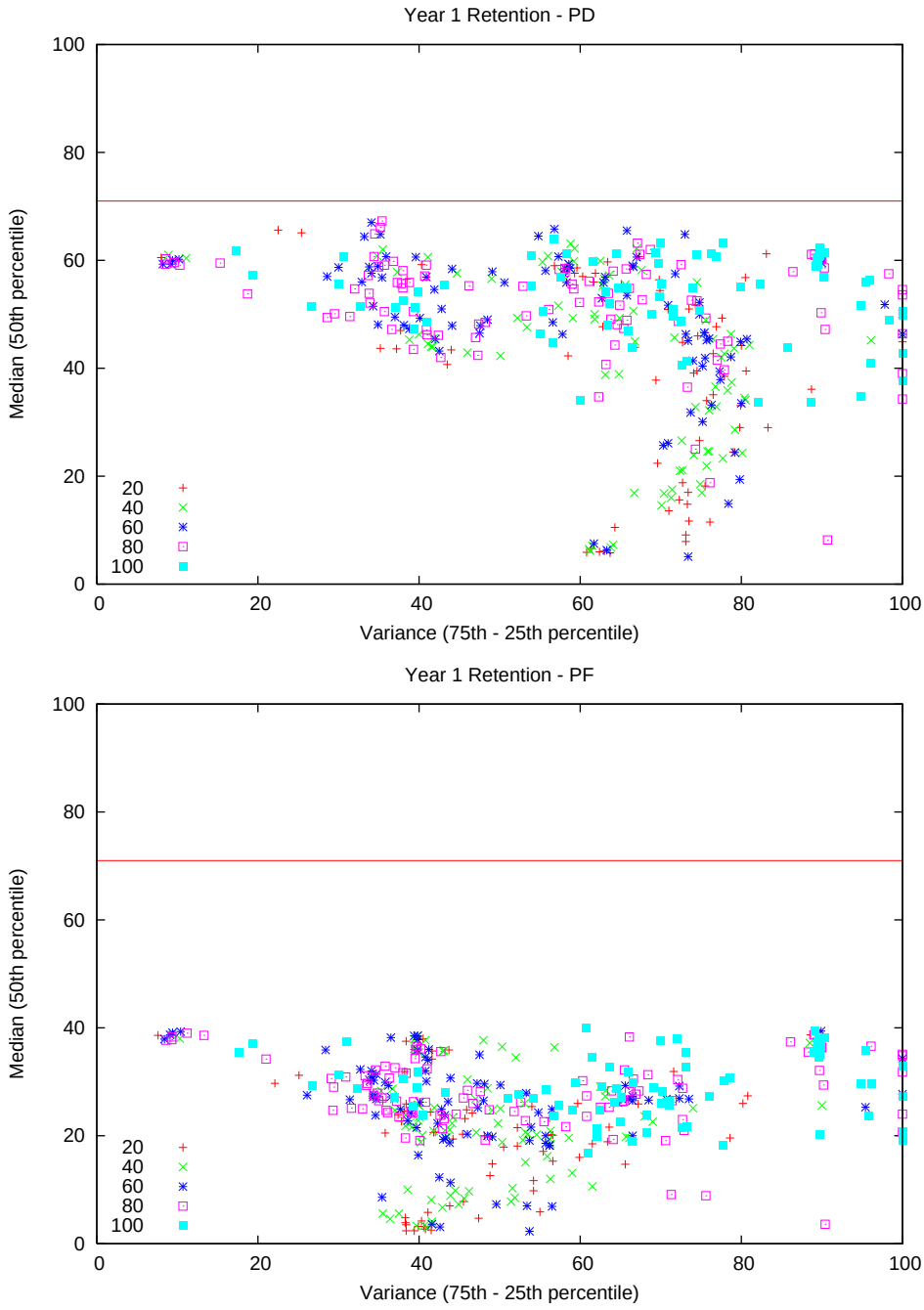


Figure 5: Probability of Detection (PD) and Probability of False Alarm (PF) with variances for first year retention.

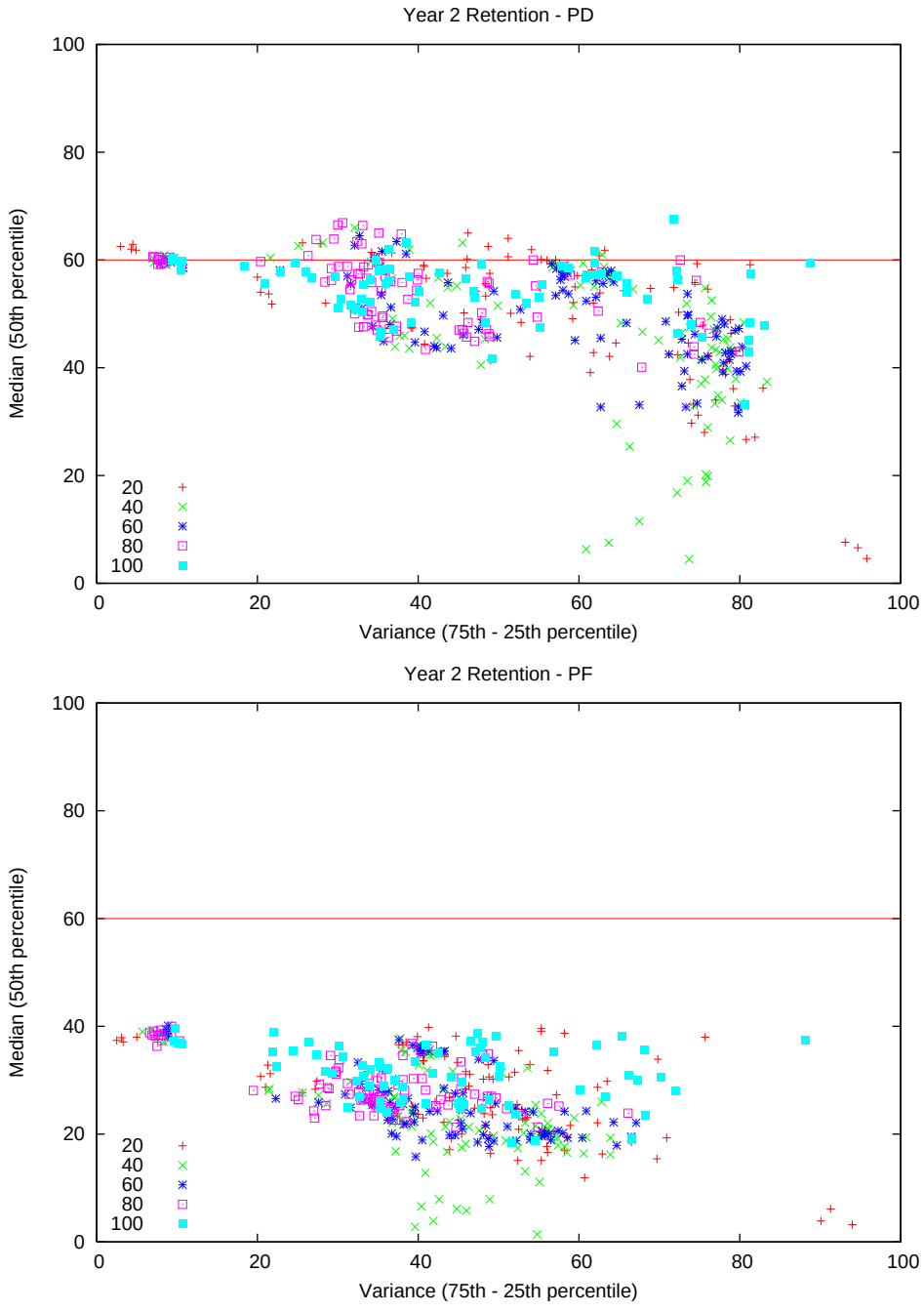


Figure 6: Probability of Detection (PD) and Probability of False Alarm (PF) with variances for second year retention.

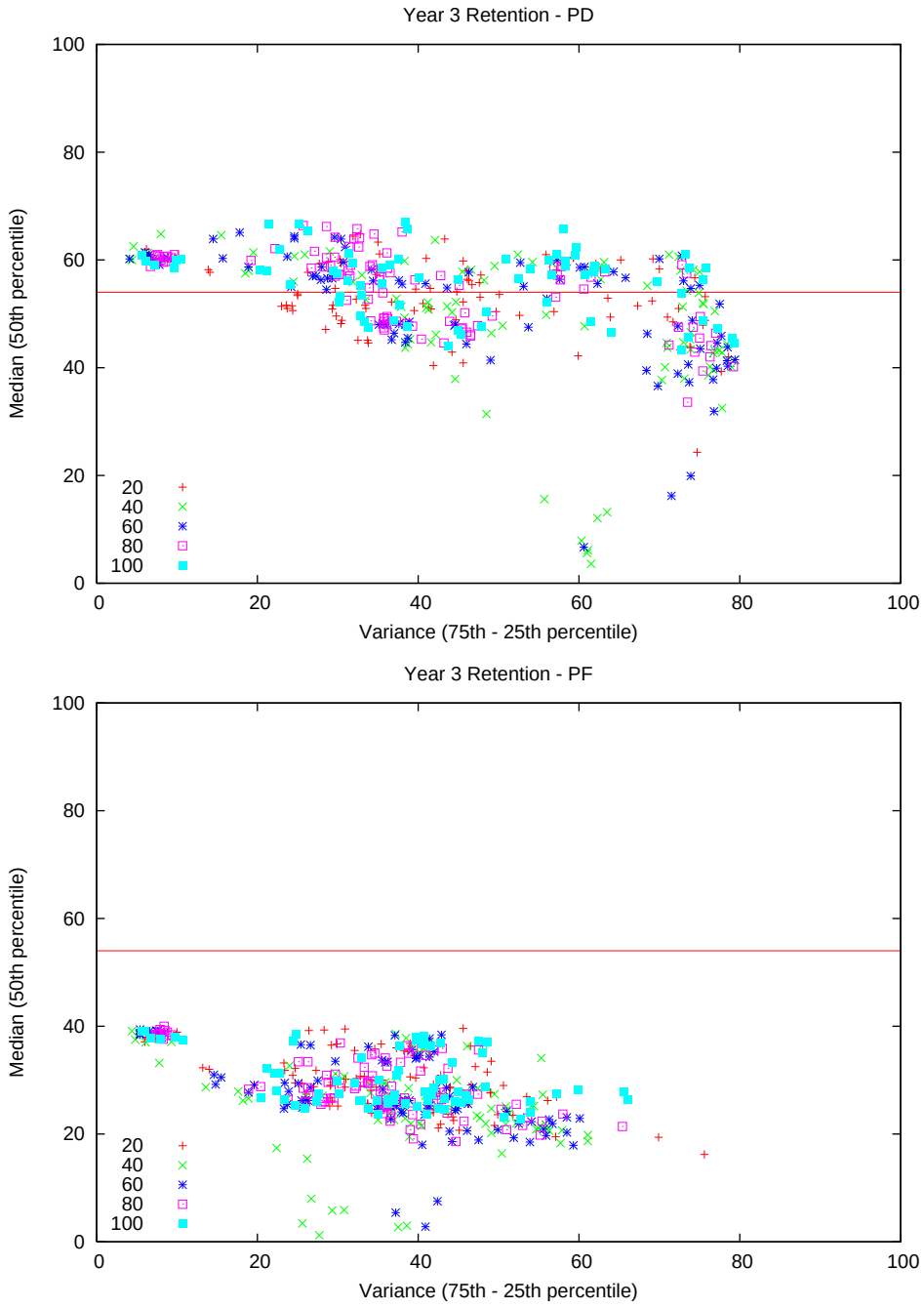


Figure 7: Probability of Detection (PD) and Probability of False Alarm (PF) with variances for third year retention.

number of attributes used. The color of each point shows the number of attributes used for that particular combination representing that point.

The horizontal line segmenting the PD graphs is a baseline reference to the existing retention rates in the data. Thus, to predict for retention in a given year, it is desirable to yield results higher than the baseline. As can be seen in the figures, the median probability of detection of retention values for the first year do not meet the baseline, and therefore we can assume using our methods, we cannot accurately predict first-year retention. Prediction of second year retention had better results than first year retention, but these results did not improve the baseline significantly. For example, most of the points lie at or below the baseline. For this reason, we did not consider second-year retention in further analysis. However, third year PD values successfully exceeded the baseline, and required more thorough examination.

5.3 Narrowing the Search

Using the visualizations described above, we narrowed our space of possible combinations to examine for third year retention. The graphs for PD and PF medians show that the range of number of attributes that maximizes PD and minimizes PF values while maintaining minimal variance is approximately 20 to 60. We performed further analysis on the reduced datasets with 20 to 60 attributes.

5.3.1 Ranking with the Mann-Whitney Test

At the moment of pruning the results based on attribute ranges, we are left with many combinations to be analyzed. In order to rank each combination, we performed a statistical Mann-Whitney test at 95% confidence in order to rank a treatment. 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 8 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.

5.4 Selected FSS and Classifier

Figure 8 shows the top-most ranking combination of FSS and classifier is obtained by either using 30 or 50 attributes. Since, the two numbers of attributes (along with their own FSS and classifier) resulted in the same Mann-Whitney rank, we concluded that the results obtained using One-R/Bayes Netork and CFS/ADTree are not statistically different. As we selected top 30 attributes critical to third-year persistence, we concentrated on approximately 1/3 of the original data. Table 3 lists the performance measures obtained for RET3 using OneR and Bayes Net.

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

Figure 8: 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.

Class Value	**Baseline**	**Probability of Detection (PD)**	**Probability of False Alarm (PF)**	**Precision**	**Accuracy**
Y	54.78%	70.0%	34.9%	70.8%	67.8%
N	45.22%	65.1%	30.0%	64.2%	67.8%

Table 3: Performance Measures obtained for RET3 using OneR as the FSS and Bayes Net as the classifier.

6 Results

Using data mining techniques, we were unable to significantly improve the classification rates for first-year and second-year retention prediction over the baseline, but we achieved approximately 20% higher probability of detection for third-year retention over the baseline. As we can predict third-year retention probability with high accuracy, based only on the first-year, beginning of term data, this result is significant in student persistence research.

Attribute	**Description**	**Value**	**Instances**	$P(RET3) = Y$
				0% 20% 40% 60% 80% 100%
FinAidSTUDENT_TA	Student's Tax Form Type	4	35	
		3	24	
		2	12,215	
		1	2,697	
FinAidMOTHER_ED	Mother's Education Level	3	7,710	
		4	814	
		2	8,792	
		1	289	
FinAidSTUDENT_MA	Student's Marital Status	M	386	
		U	17,254	
		S	24	
FinAidFATHER_ED	Father's Education Level	3	7,502	
		2	8,461	
		4	1,136	
		1	436	
FinAidDEPENDENCY	Student's Dependency Status	I	2,523	
		D	15,154	
FirstGenInd	First Generation Student	N	10,370	
		Y	7,311	
FinAidPARENT_TAX	Parent's Tax Form Type	4	24	
		1	13,101	
		2	3,126	
				0% 20% 40% 60% 80% 100%

Table 4 continued on next page

Attribute	Description	Value	Instances	$P(RET3) = Y$
				0% 20% 40% 60% 80% 100%
		3	16	
FinAidSTUDENT_AG	Student's Adjusted Gross Income	4829.5-7915.5	4,152	
		3335.5-4829.5	2,780	
		16713.5-inf	1,022	
		1894.5-3335.5	2,540	
		-inf-1894.5	2,106	
FinAidSTUDENT_WA	Student's Wage	7850.5-9958	1,752	
		4092.5-7850.5	5,622	
		1.5-999.5	2,057	
		1903.5-4092.5	4,176	
		-inf-1.5	1,721	
HS_GPA	High School GPA	3.015-3.345	5,769	
		2.905-3.015	1,990	
		2.645-2.905	4,541	
		2.035-2.645	4,758	
		-inf-2.035	545	
FinAidPARENT_MAR	Parent's Marital Status		464	
		W	394	
		M	11,328	
		S	3,127	
		U	637	
PercentileRankHSGPA	Percentile Of Hs Gpa Among Freshmen Cohort	45.75-59.65	3,660	
		33.7-45.75	3,165	
		15.35-33.7	4,803	
		2.35-15.35	3,479	
		-inf-2.35	637	
FinAidPARENT_AGI	Parent's Adjusted Gross Income	96636-inf	3,751	
		58550.5-96636	6,045	
				0% 20% 40% 60% 80% 100%

Table 4 continued on next page

Attribute	Description	Value	Instances	$P(RET3) = Y$
				0% 20% 40% 60% 80% 100%
		18376.5-58550	5,598	
		-inf-18376.5	1,167	
FinAidFATHER_WAG	Father's Income	52366-inf	5,873	
		-inf-52366	9,459	
FinAidMOTHER_WAG	Mother's Income	42957-inf	3,148	
		-inf-42957	13,063	
HS_PERCENT	High School Percentile	80.5-inf	5,838	
		60.5-80.5	6,980	
		43.5-60.5	5,624	
		-inf-43.5	7,774	
MaxACT	Max Of ACT Score And ACT Equivalent	23.5-inf	6,952	
		19.5-23.5	10,044	
		15.5-19.5	7,001	
		-inf-15.5	2,219	
PercentileRankMaxACT	Percentile Of Max ACT Among Freshmen Cohort	71.35-inf	6,658	
		30.55-71.35	10,281	
		8.35-30.55	6,763	
		-inf-8.35	2,514	
CUR_ERLHRS	Total Enrolled Hours	14.5-18.5	14,523	
		13.5-14.5	6,964	
		10.5-13.5	4,016	
		-inf-10.5	532	
		18.5-inf	181	
ACT1_COMP	ACT Comprehensive Score (new)	23.5-inf	5,669	
		19.5-23.5	8,667	
		17.5-19.5	4,043	
		-inf-17.5	7,837	
ACT1_MATH	ACT Math Score (new)	22.5-inf	7,082	
				0% 20% 40% 60% 80% 100%

Table 4 continued on next page

Attribute	Description	Value	Instances	$P(RET3) = Y$
				0% 20% 40% 60% 80% 100%
		19.5-22.5	4,767	
		16.5-19.5	6,611	
		-inf-16.5	7,756	
ACT1_ENGL	ACT English Score (new)	24.5-inf	4,676	
		19.5-24.5	8,271	
		16.5-19.5	4,877	
		-inf-16.5	8,392	
AGE	Age of Student at Matriculation	-inf-19.5	24,826	
		19.5-inf	1,390	
ENG10	Enrolled in English Courses	N	24,407	
		Y	1,809	
LIVEONCAMP	On-Campus Indicator	Y	20,087	
		N	6,129	
ADMIT_MAJ	Admit Major	ADV	38	
		AERN	433	
		AEDG	208	
COMP_WRITE	Compass Writing Score	-inf-9.5	3,780	
		73.5-inf	13,887	
		49.5-73.5	5,299	
		9.5-49.5	3,250	
TotalClasses	Total Number of Enrolled Classes	5.5-inf	15,021	
		4.5-5.5	10,237	
		-inf-4.5	958	

Table 4: Top 30 attributes with values. Only five attribute values with at least 10 records are shown.

After selecting the best combination of FSS (oneR) and classifier (Bayes Network) based on Mann-Whitney test rankings, we found that attributes given in Table 4 are critical to third-year persistence. Out of these 30 attributes, top ten attributes described student's family background and family's economic condition, and the most selected attribute was the student's tax form type, which came from the FAFSA

submission and had these values:

1. IRS 1040
2. IRS 1040A, 1040EZ
3. A foreign tax return
4. A tax return with Puerto Rico, another U.S. territory or a Freely Associated State

A person is eligible to file 1040A or 1040EZ if he or she makes less than $100,000, does not itemizes deductions, does not claim dependents, etc. As shown in Figure 9, there is a positive correlation between tax form type 2 and third-year retention for lower high school GPA ranges with the exception of the range: 2.645 to 2.905. Third-year retention percentages are significantly higher for the students who (or their parents) have filed a foreign tax return (type 3) or a U.S. territory tax return (type 4) than those who have filed U.S tax return (type 1 or 2).

Second attribute in the list was the parent's household size, which had a positive correlation with third-year retention percentage as shown in Figure 10 along with the distribution of the parent's household size. The sample size was low for student's with large number of people in the household, therefore, retention percentages in such cases is meaningless.

As previous research has concluded that parent's education level plays an important role in student's dropout decision (Spady 1970; Tinto 1975; Bean 1979), Figure 11 shows that chances of student's persistence are higher if the parent's education level is higher. If the parents did attend college and beyond, father's education level has greater impact than mother's education level on student's persistence.

As shown in the Table 4, student's marital status does play a role in persistence, especially if the student is separated (denoted by S in the table). Out of 24 students, who indicated in FAFSA as separated, only four students persisted till the third year. Students income (FinAidSTUDENT_WA) also affect their persistence; students with wages in the range of $7850.5-$9958 had the highest percentages of return (close to 80%).

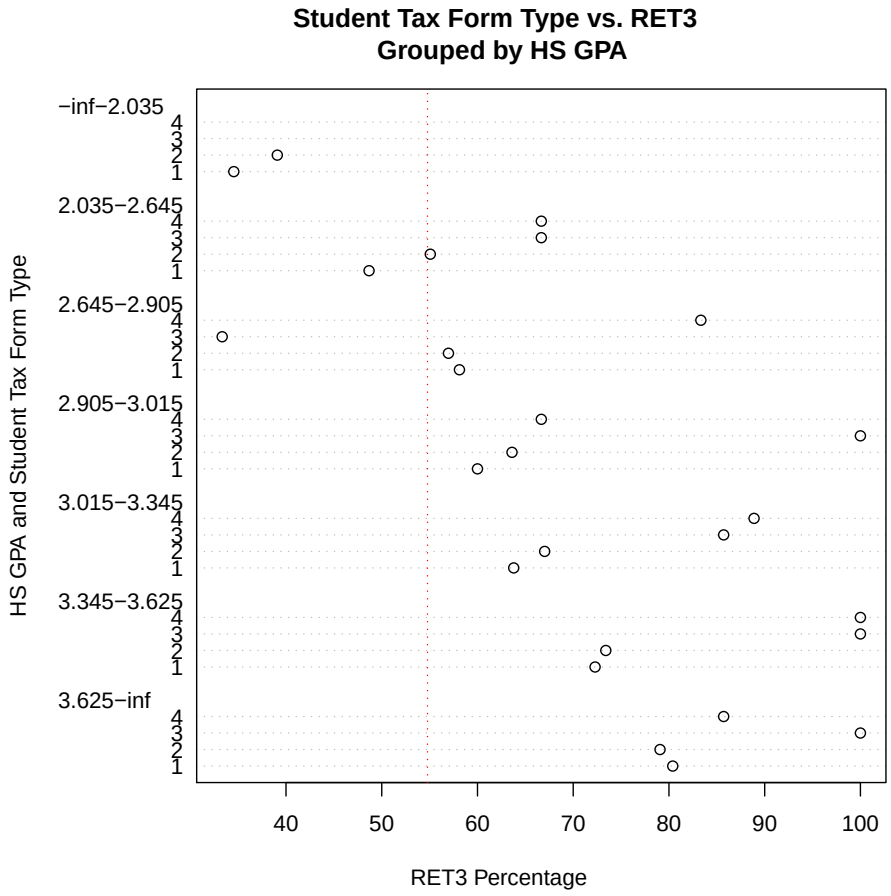


Figure 9: Student Tax Form Type vs. RET3 Percentage, Grouped by High School GPA. Red Dashed Line Represents the Baseline RET3 Percentage

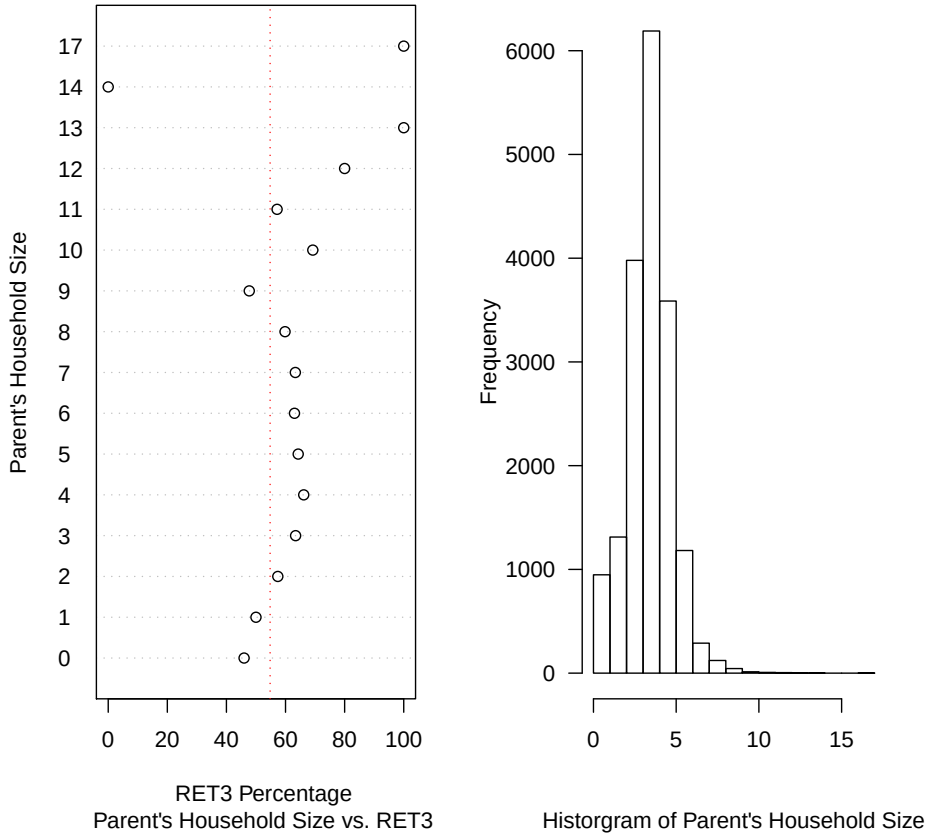


Figure 10: Parent's household size vs. RET3 percentage (left), and distribution of parent's household size (right). Red dashed line represents the baseline RET3 percentage

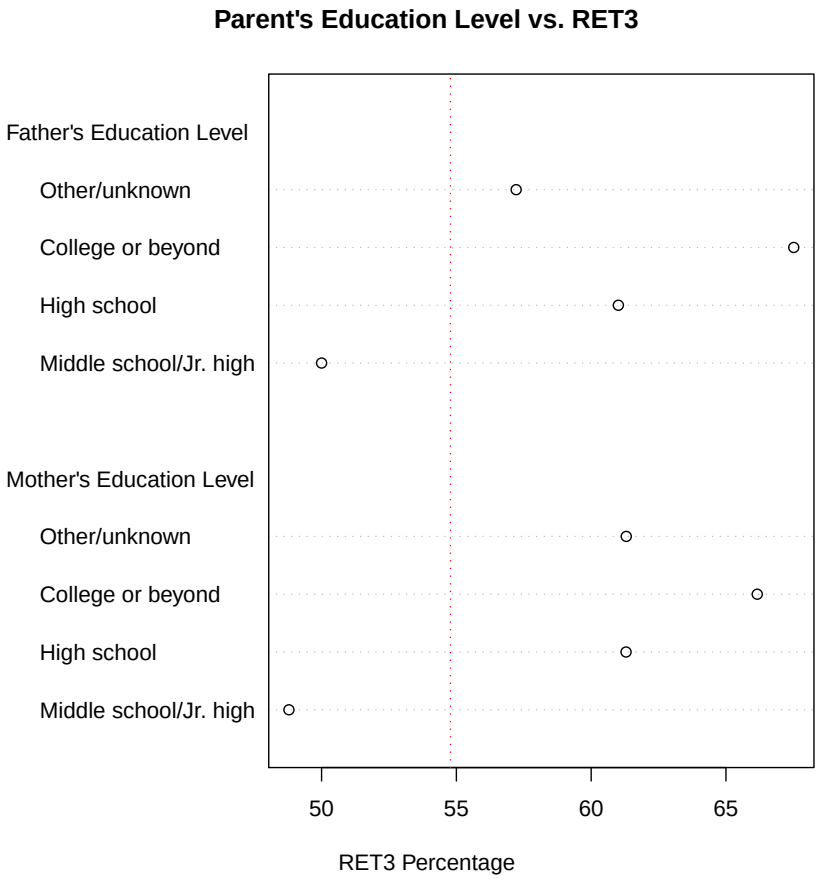


Figure 11: Parent's education level vs. RET3 percentage. Red dashed line represents the baseline RET3 percentage

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.

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. For the studied institution, family background and family's social-economic status are critical for student's third-year persistence.

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*(125), 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. (1979). Path Analysis: The Development of a Suitable Methodology for the Study of Student Attrition. Paper presented at the Annual Meeting of the American Educational Research Association,San Francisco, California.

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, *12*(2), 155–187.

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

Bresciani, M. J., & Carson, L. (2002). A study of undergraduate persistence by unmet need and percentage of gift aid. *NASPA Journal*, *40*(1), 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*.

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*(5), 569–581.

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.

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*(1), 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. Tech. rep., 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*(8), 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*(2006).

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

Lau, L. K. (2003). Institutional factors affecting student retention. *Education*, *124*(1), 126–137.

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: A Response to. *IEEE Transactions On Software Engineering*, (pp. 637–640).

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

Mitchell, T. M. (1997b). *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*(3), 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*(4), 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*(1), 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. *Machine Learning*, (pp. 81–106).

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

Rish, I. (2001). An empirical study of the naive bayes classifier. In *IJCAI-01 workshop on "Empirical Methods in AI"*.
URL 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*(1), 64–85.

Spady, W. G. (1971). Dropouts from higher education: Toward an empirical model. *Interchange*, *2*(3), 38–62.

Stage, F. (1989). Motivation, Academic and Social Integration, and the Early Dropout. *American Educational Research Journal*, *26*(3), 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*(131).

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*(3), 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*(1), 89–125.

Tinto, V. (1982). Limits of Theory and Practice in Student Attrition. *The Journal of Higher Education*, *53*(6), 687–700.

Tinto, V. (1988). Stages of student departure: Reflections on the longitudinal character of student leaving. *Journal of Higher Education*, *59*(4), 438–455.

Vandamme, J. (2007). Predicting Academic Performance by Data Mining Methods. *Education Economics*, *15*(4), 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.

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*, *33*(9), 635.

Żytkow, J., & Zembowicz, R. (1993). Database exploration in search of regularities. *Journal of Intelligent Information Systems*, *2*(1), 39–81.