International Journal of Engineering & Technology, 7 (2.28) (2018) 337-343

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Identifying factors for student retention of higher ed institutions using decision tree

Jason Gierman [1] *, Oliver Strong [1], Gongzhu Hu [1]

[1] Department of Computer Science, Central Michigan University, Mount Pleasant, MI 48859, USA
**Corresponding author E-mail: hu1g@cmich.edu*

Abstract

Student retention is an issue of high priority for many colleges and universities. Keeping students in school is the very basic condition for them to achieve their goals for going to colleges in the first place. A lot of research and practices have been done across institutions to improve student retention rates, but colleges and universities are still trying to figure out what are the factors that are most important to student retention. In this paper, we present our experiments of building predictive models, particularly decision tree models, to fit in the overall prediction of full time student retention. The data set of 1,965 cases from 1987 to 2000 obtained from the Delta Cost Project Database of the American Institutes for Research has 541 variables. We used variable selection measures like R-Squared to reduce to 45 variables and build decision tree models to fit the training data. Eight variables were identified to be most influential to the retention rates. Our experiments show that the decision trees with moderate depth are suitable for creating retention model.

Keywords: *Student Retention; Predictive Models; Variable Selection; Regression; Decision Tree.*

1. Introduction

To many high school students and their parents the only true way to prosperity is through a higher education. Higher education prepares the young generation for the most rewarding careers that the high technology world has to offer. Yet, with all this promise of intellec-tual and financial recompense, many students are not completing the degree that they seek. This leaves the students and potentially their parents, commonly the financial backer, to suffer both the expense and the consequence of failing to prepare for the competitive career market.

This is the reason that many colleges and the higher education system as a whole are collecting data and taking measures in order to guarantee that students do not just attend, but finish with their degrees in a timely manner.

It is not a stretch of imagination to assume that with soaring cost, relatively declining sources of financial assistance that it is even more imperative that students leave the university with the degree that they spent so much time and effort to seek. However, six year graduation rates (150% normal time for graduation) are hovering just under 60% in USA even colleges and universities made a great efforts to retain students. The rates of students graduating in 4, 5, and 6 years in USA is shown in Fig. 1(a) according to the National Center for Education Statistics data [11], and the freshman retention rates of 4-year public institutions from 2010 to 2015 are shown in Fig. 1(b), Selective universities, those that accept 25% or less, retain 95.9% of their freshmen in the 2014-2015 academic year. Compared to a more reflective measure of success, universities with open admissions that retain 62.3% freshmen and graduation rate of 59% [11].

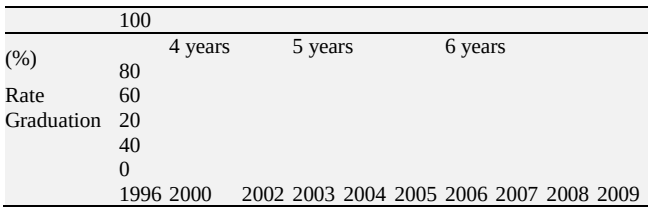
There are many factors affecting the retention and graduation rates. It is clear from the data and also naturally expected that the competi-tiveness of admission is a decisive factor. Students admitted to more selective universities tend to be better prepared, more motivated, have clearer goals for their life, and hence are more likely to stay in school to finish their college education. The question is, though, what are the factors that most schools should pay attention to to retain their already-admitted students regardless of how tough or how easy the students get admitted.

Research on the topic of student retention has been conducted mostly in the fields of education and social study. Researchers in the field of computer technology have also studied this topic as an application of computational algorithms. In this paper, we apply data mining technology, particularly decision trees, to identify the factors that have the high influence on the student retention rate.

2. Related work

We give a brief review of studies on student retention in two fields:

Education and social study, and computer technology.



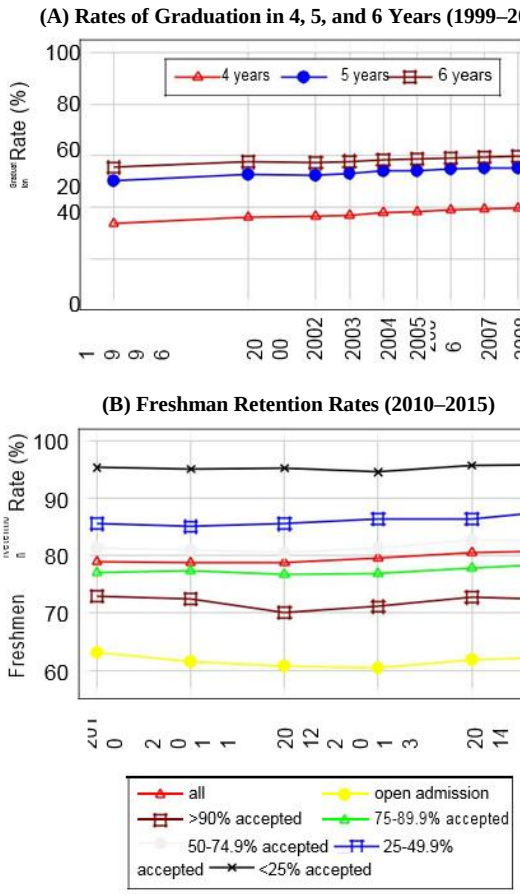


Fig. 1: Graduation and Retention Rates of 4-Year Institutions in USA.

2.1. Education and social study

Research on student retention and graduation rate is mostly in the fields of education and social study. It focuses on two main areas that have some inherent overlap. The first area is that of minority and underrepresented groups. This research looks to bridge the gap between retention and more importantly the graduation of minority groups and other selective groups. Secondly, and more importantly to this work, research focuses on the economic situation in student retention and graduation. This area is quickly gaining momentum as the price of higher education is making it equally inaccessible for the underprivileged and the middle class [5].

Retention rates in their simplest form are how many students return from the previous year. Starting in the late 1980s governmental decision makers have become interested in the count and reasons that students are not finishing what they started. Some of these reasons include misplaced emphasis in athletics, others cite affirmative action bringing under performing students who are doomed to fail and yet others cite the sheer cost stopping even the most well meaning student from completing what they started [4], [1].

Olbrecht et al. made an attempt at determining what keeps students in school. Using data from a selective liberal arts university in New Jersey the researchers compare school financial contribution, student contribution and unmet need for five cohorts in an attempt to determine the pertinent factors. Somewhat surprisingly to the authors they determined that while the ability of a student's family to contribute their first year is important in retention the unmet need that keeps the student's financial well being part of the equation has a positive correlation to retention [12]. The authors conclude by making the blanket assessment that not only shaping the student body is possible with financial contributions, the university to some degree has the ability to chose who succeeds at it.

Many studies do not even focus or apply emphasis to the fact that monetary adjustment may, whether through higher tuition or through gift aid, influence how many students are retained. Instead these researchers chose to focus more on plans and policies such as fresh-men orientation programs, past academic performance measures and enrollment policies [3], [9], [14]. From these studies researchers have made some consistent findings:

- High school GPA has a higher correlation with retention than ACT/SAT score [3, 14].
- Proper orientation programs have a positive impact on student retention [6].
- Proper pre-university preparation is key to student success and retention [13].

2.2. Computer technology

Researchers in computer technology field have also studied the prob-lem of student performance and retention by applying data mining methods to collected academic data.

Laur´ıa et al. built several predictive models (logistic regression, support vector machine, decision trees) for predictions of student performance as at risk or not. They applied and compared these models on a data set collected from student records system of their college and the open source Sakai CMS (Course Management Sys-tem) [8].

A similar study was reported in [15] that also built models for predicting at-risk students using data mining methods (SVM, decision tree). However, the data set was collected as clicks in the Virtual Learning Environment (VLE) and Tutor Marked Assessment (TMA) scores. Their study showed that VLE clicks is a better predictor than TMA scores, but combine the two yielded the best prediction results measured by precision, recall, and f-measure.

Nandeshwar et al. conducted a study on finding predictors on student retention [10] using a data set from a mid-size public university that had more than 100 attributes. Six classifiers (One-R, C4.5, ADTrees, Naive Bayes, Bayes networks, and radial bias networks) were applied to the data set, and the results indicated that the most significant factors for student retention are family background and family's social-economic status, high school GPA and test scores.

Yu et al. [16] applied three data mining techniques (classification trees, multivariate adaptive regression splines (MARS), and neural networks) to explore the issue of student retention. The data set contains 6690 sophomore students enrolled at their university. They found that the transferred hours, residency, and ethnicity as crucial factors to retention.

3. Methodology

Our analysis relied on data set from the Delta Cost project database by the American Institutes for Research [2]. The Delta Cost project is a research project that curates and analyzes a longitudinal data set with data starting in 1987 and extending through 2012. The data consists of 974 variables. Data included in the set contains measures of student and institution related variables, such as gift aid contribution, student return rate and graduation rates.

3.1. Data preparation

In order to best focus on the pertinent variables when asking the question of what the impact of financial backing to student retention and graduation rates are, the variables that were not of interest were removed from the data set. 451 variables remained from the original 974 variables once the data preparation completed. Due to the number of years and sheer amount of data still required to perform the analysis the study focuses on a single year. The year chosen is 2012, the most recent year available in the data set.

For the focused exploration of variables we chose to work exclusively with four of the five hundred and forty one variables as input variables and a single target. The variables were chosen based on the fact that they could be tied directly to funding source and should, though not modeled, show an even distribution across the student body. The variables chosen and their respective de-

scriptions are included in Table 1. It should be noted that other variables may have represented some aspect of funding dichotomy between the student and external benefactors, but they did not have the same internal consistency that the chosen variables had in regards to direct and uniformness.

Table 1: Variables Used in First Model of Retention Rates

Variable	Description
Net Student Tuition	Net tuition revenue coming directly from students.
Federal Grant Percent	Percentage of full-time, first-time degree/certificate-seeking undergraduate students who received federal grants (grants/educational assistance funds).
State Grant Percent	Percentage of full-time, first-time degree/certificate-seeking undergraduate students who received state/local grants (grants/scholarships/waivers).
Institutional Grant Percent	Percentage of full-time, first-time degree/certificate-seeking undergraduate students who received institutional grants (scholarships/fellowships).
Loan Percent	Percentage of full-time, first-time degree/certificate-seeking undergraduate students who received student loans.
Full Time Retention Rate	The percent of the previous year's fall first-time full-time cohort (minus exclusions) that re-enrolled at the institution as either full-time or part-time the following fall.

The variables were then transformed into categorical variables by calculating the mean and standard deviation of each of the variables. For the predictors a three level model is defined as followed. High: which is more than two standard deviations over the mean, average: within two standard deviations of the mean and low: which is two or more standard deviations below the mean. The target variable was only divided once. Making drop out or transfer likely above the mean and drop out or transfer unlikely below the mean and representing this using yes and no respectively. A portion of the data is shown in Table 2.

Table 2: Sample Data

net tuition	fed grant	st grant	inst grant	load pct	drop out
low	average	low	low	average	yes
low	low	low	low	average	no
low	low	low	low	low	no
low	average	low	low	average	yes
high	low	low	low	low	no
low	average	low	low	average	yes
average	low	low	low	low	no
low	low	low	average	average	no
low	high	high	average	average	yes
low	average	low	low	average	yes
low	average	average	average	average	yes

The top node is the percent of students receiving federal grants. It branches quickly with schools in the high category having a higher likelihood of student drop out or transfer. Though it is pertinent to note that there are few schools deviating at least two standard deviations above the mean. Also from this node, low is automatically classified as low dropout/transfer likelihood. Moving up the tree loan percent breaks into two paths with average loan distribution seeing high drop out/transfer rates, with state grant percentage at the top of the tree. The classification completely disregarded the institutional grant percentage.

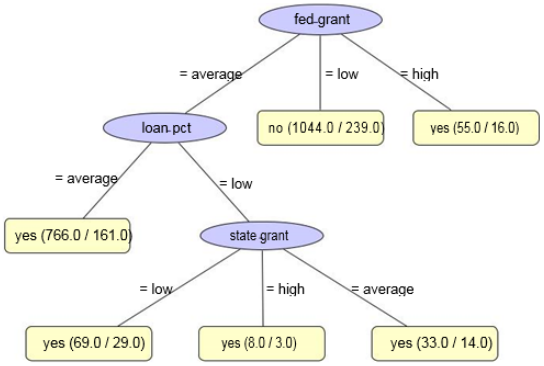


Fig. 2: Decision Tree Model.

```
Size of the tree:

Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly classified instances        1493               75.9796 %
Incorrectly classified instances       472               24.0204 %
Kappa statistics                         0.516
Mean absolute error                      0.3573
Root mean squared error                  0.4242
Relative absolute error                 71.6993 %
Root relative squared error             84.9728 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)     100      %
Total number of instances             1965

== Detailed accuracy by class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.706    0.192      0.766      0.706     0.735      0.751    yes
                 0.808    0.294      0.755      0.808     0.78       0.751    no
Weighted Avg:    0.76     0.246      0.76       0.76      0.759      0.751

=== Confusion Matrix ===

   a   b   <-- classified as
 654 272 |  a = yes
 200 839 |  b = no
```

Listing 1: Validation Result of Decision Tree.

3.2. Creating and evaluating the model

Once the data was transformed into a categorical representation Weka was used to develop a decision tree. The classifier that was chosen is Naive-Bayesian with 10 fold cross validation. When running analysis using Weka the tree shown in Fig. 2 was produced. A quick visual examination of this tree will reveal a couple interesting points. The validity of the classification model will be discussed in a later.

This model would fall under the moderate category as proposed by Landis and Koch [7]. A model of the moderate categorization would have an reliability great enough to rely on for many decision making exercises. At this level the model was able to classify 76% of the records correctly. However, the missing institutional grant percentage and the fact the heavy bias on student loan percent is a cause of concern, in that the model is very heavy at the root and places most of its decision making on the root node. This lead to the exploration of the whole data dictionary in order to determine if any of the variables were part of the larger picture.

3.3. Getting the bigger picture

As mentioned previously a model of student retention was successful derived from the variables in Table 1. However, the weight placed on the root node poses a problem of heavy model bias on the single factor of federal grant percent. It seems prudent to also see how the factors selected fit into the overall model of retention. What if the entirety of the 541 variables were taken into account when developing the model of student retention? Would the importance of financing source show through or will it not play a larger part in the overall model.

To do this the data was moved from Excel and Weka and into SAS Enterprise Miner in order to handle the processing of the much larger dataset. SAS also provided a more robust and productive manner in which to select variables, partition the data, create

the model and compare the created models. Fig. 3 enumerates the nodes utilized. The first node is data partition and was configured to split the model 60, 20, 20. With 60% of the records being used for training, 20% used for validation and 20% used for for testing.
The next step after partitioning the data was variable selection. Not all of the 541 variables in the data set were relevant to model creation. A variable selection node was employed so reduce the number of variables used for model creation. During this process, variable having over 50 percent missing values were removed. The variable selection node also removed variables with extremely low individual and incremental R^2 relationships to the target. The lower bound cutoff for individual R^2 was set at .005. This removed variables which did not meet the individual R^2 threshold when not taking into account other variable. The incremental R^2 lower bound was set at. 0005. This removed variables if they did not provide an additional R^2 above the incremental R^2 threshold when compared with the other variables. After variable selection, forty five variables were retained and considered for model creation.
The retained variables were then passed to group of six modeling nodes. The models created are as follows:

1) Linear Regression: forward selection.
2) Linear Regression: stepwise selection.
3) Decision Tree 1.
- Assessment measure: Average Squared Error (ASE)
- Splitting criterion: ProbF
- Maximum depth adjustment: 6

4) Decision Tree 2.
- Assessment measure: Average Squared Error (ASE)
- Splitting criterion: ProbF

Maximum depth adjustment: 5

5) Decision Tree 3.
- Assessment measure: Average Squared Error (ASE)
- Splitting criterion: ProbF
- Maximum depth adjustment: 4

6) Decision Tree 4.
- Assessment measure: Average Squared Error (ASE)
- Splitting criterion: ProbF
- Maximum depth adjustment: 3

Since our target was interval, the assessment measure used for all decision trees was Average Squared Error (ASE). An adjustment to the maximum depth was made to each decision tree. The maximum depth determine the maximum number of splits allowed, from root node to child node. Decrease the maximum depth reduces the com-plexity of the model. This was done to test the predictive power of simpler models as well as reduce the risk of over-fitting.

4. Results

Recall the goal the of the project was to create the simplest model of retention possible, while still retaining predictive power. During the exploration phase a model of retention was produced that was of acceptable reliability. But, what is this the best model that can be created?
The six models described were put through a model comparison node and ranked based on their Average Squared Error (ASE) when predicting the test set. The respective ASEs for the models on the training and testing set are giving in Table 3. As shown in table, the top two models have fairly similar ASE when predicting the test set. In terms of ASE the most accurate model is the Decision Tree with a maximum depth of 5, followed by the Decision Tree with a maximum depth of 6. Due to the lowest ASE when predicting the test set, as well as being a simpler model, the Decision Tree (MaxDepth5) was chosen as our final model. While the simplest of our models were not chosen we did save some efficiency over the largest model generated.

Table 3: Absolute Standard Errors, Sorted in Ascending Order

Model	Average Square Error	
	Test	Training
Decision Tree (MaxDepth5)	0.015862	0.010708
Decision Tree (MaxDepth6)	0.016094	0.010354
Decision Tree (MaxDepth4)	0.016333	0.011629
Decision Tree (MaxDepth3)	0.016762	0.012872
Regression (Forward)	0.024938	0.021004
Regression (Stepwise)	0.024938	0.021004

We compared the model-predictive retention rates and the actual (target) retention rates for the regression model and decision tree models. The result for stepwise linear regression model is shown in Fig. 4.
For the decision tree models, other than the improvement in ASE seen between the decision tree of max depth-5 and the other models, the 5-level model provided a better fit between the mean actual target of test set and predicted means, displayed in Fig. 5. The x-axis is model score, only shown in the last subfigure Fig. 5(d).
There is some divergence at around fifty percent retention, but after this the average is very similar to the actual, which is the main reason that this method had the lowest ASE. The other models were unable

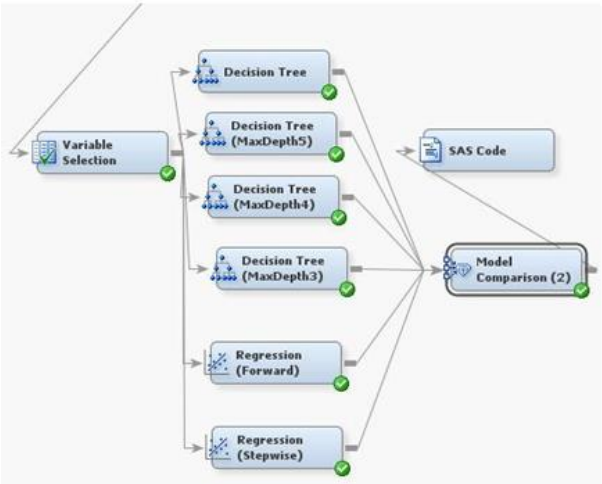


Fig. 3: SAS Nodes Used in Generating the Model to Provide the Same Level of Consistency Over the Range of Student Retention Rates.

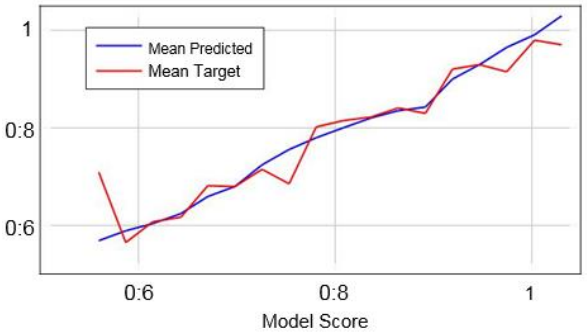


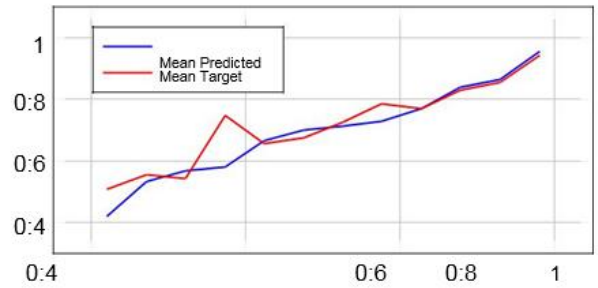
Fig. 4: Actual vs. Predicted for Stepwise Linear Regression.

After the Decision Tree (MaxDepth5) was chosen based on ASE and having the most consistent representative mean target versus pre-dicted plot, the tree was examined, as well as the variables involved.
The full tree of our final mode, is displayed in Fig. 6 and contains 5 tiers of splitting. Splits were allowed to occur on the same variable multiple times in order to retain the models ability to predict the test set.
In Fig. 6, each leaf node contains the count of cases and the average classification accuracy in the training set and testing set, given in Table 4.

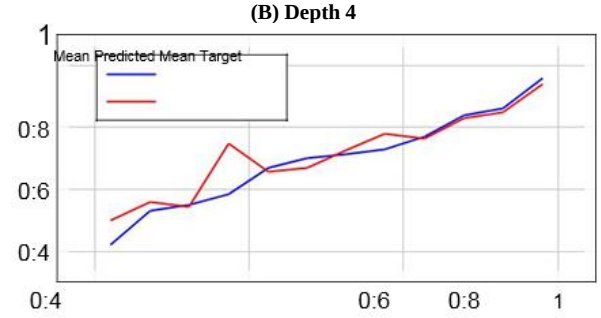
Table 4: Information in the Leaf Nodes in Fig. 6

Node ID	Count		Statistic Average	
	Training	Testing	Training	Testing
9	196	63	0.6987	0.6598
10	40	17	0.4210	0.4424
11	91	29	0.5690	0.5955
12	84	35	0.8531	0.8194
15	21	5	0.5800	0.4960
27	54	10	0.9561	0.9400
30	13	7	0.7162	0.7000
31	69	20	0.5370	0.5505
32	12	3	0.5217	0.5167
33	153	48	0.6663	0.6519
44	6	0	0.8350	—
45	31	6	0.9029	0.9050
48	63	34	0.7121	0.7288
49	169	49	0.7692	0.7698
50	26	6	0.7377	0.7133
51	124	37	0.8373	0.8200

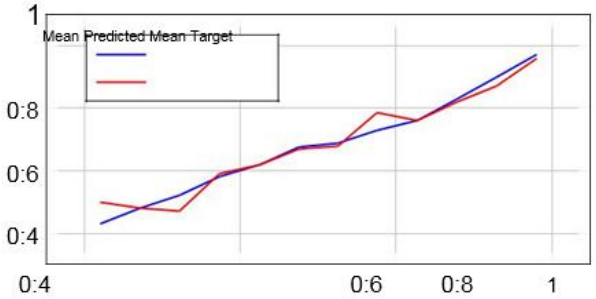
(A) Depth 3



(B) Depth 4



(C) Depth 5



(D) Depth 6

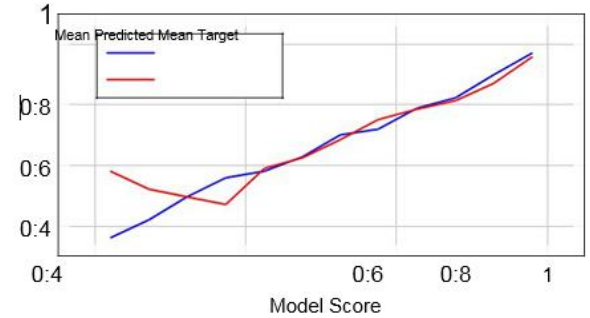


Fig. 5: Actual vs. Predicted for Decision Trees.

There were 8 important variables in the models, reduced from the 45 variables used in model creation. The most important variable in the model was Graduation Rate 150% which is the percentage of students graduating within 150% of normal time, e.g., if a student is enrolled in a 4 year program then they would have graduated within 6 years. This variable, as shown in Table. 5, splits 6 times and it is by far the most important variable in the model. However, there are many more variables involved in the model which have significant splits depending on a colleges graduation rate. The first major split in on a graduation rate of around 50%. On the left side of the decision tree, we have all 4 year colleges with a graduation rate of below 50%. The significant splits for these colleges include attributes like the number of enrolled part-time students between the ages of 35-39, total number of applicants, and total salary for full time institutional faculty.

On the right side of the decision tree, we have colleges with graduation rates over 50%. The significant splits for these colleges contain Federal Grant Percent, which was shown to be relevant in our previ-ous model with hand picked variables, as well as in previous research. It is interesting to see that only colleges with above a 50% gradu-ation rate had a significant split on Federal Grant Percent. This may suggest that Federal Grant Percent is a more important factor in retention rates for these higher graduation rate colleges. Further research could be conducted to explore this relationship between graduation rates, federal grants, and retention rates. For these higher graduation rate colleges we also see faculty salary variables, includ-ing total salary for full time instructional faculty, and average salary for full time faculty. These types of variables appear for both high and low graduation rate colleges, and may suggest the importance of faculty salary when attempting to increase retention rates. Again, more research is needed to explore the nature of these relationships.

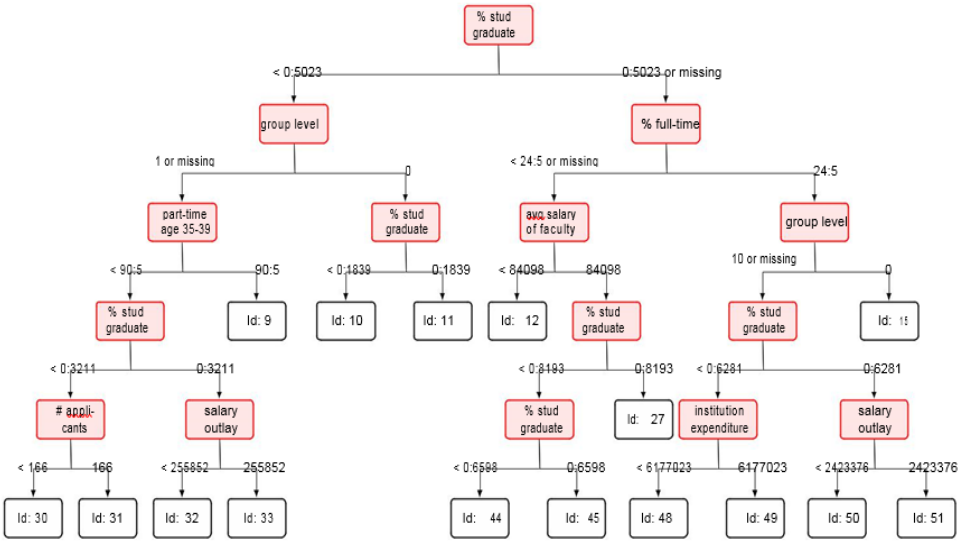


Fig. 6: Decision Tree (Depth 5).

Table 5: Variables with High Importance Measures

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
grad rate 150 p4yr	Percentage of students graduating within 150% of normal time	6	1.0000	1.0000	1.0000
G sector revised	Grouped levels for sector revised	2	0.4913	0.4450	0.9059
fed grant pct	Percentage of full-time first-time degree-seeking undergraduates	1	0.4152	0.3637	0.8759
pta1110	Part-time age 35-39 all	1	0.2316	0.0000	0.0000
salarytotal	Total salary outlays of full-time instructional faculty	2	0.2029	0.1753	0.8641
applcn	Total number of applicants	1	0.1800	0.1944	1.0797
ft faculty salary	Average salary of full-time faculty	1	0.1544	0.2371	1.5359
instsupp01 fasb	Expenditures for institutional support – current year total (FASB)	1	0.1175	0.0852	0.7249

5. Conclusion

We have learned from this exploration and prediction of the data that student retention is a multifaceted problem. We first created a model with a greater than acceptable kappa statistic that placed the federal grants at the root and immediately was able to classify a large number of records. However, this weight on a single factor was somewhat disconcerting in the lack of diversity of the selective criteria. This lead us to the investigation of the other factors. Con-ducting a complete analysis of the all the 541 factors using SAS Enterprise Miner, we were able to create a more complete decision tree model, which highlighted a more complex and interconnected list of important variables that effect retention rates.

From the discussions in the literature, particularly of researchers in the area of education and social study, the problem of student retention is a lot more complex. Although we can select specific variables for creating a retention model, this may not give a full picture of what is involved in retention rates. Collaborative work among researchers in education and social study and researchers in computer and information technology is much needed to create better models.

We used the most recent available academic year in our data set, which was 2012. Future research would benefit from increasing the number of academic years included in the analysis to explore more long term trends involving retention rates. This research also chose to only examine 4-year institutions, and further work could be done to expand the analysis to 2-year institutions as well. Exploring retention rates in colleges and universities outside USA could also highlight some cultural similarities and differences involved in retention rates.

References

[1] S. Alon and M. Tienda. Assessing the “mismatch” hypothesis: Differ-ences in college graduation rates by institutional selectivity. Sociology of education, 78(4):294–315, 2005.

[2] American Institutes for Research.Delta cost project database. http://www.deltacostproject.org/ delta-cost-project-database, 2012.

[3] P. Brotherton. It takes a campus to graduate a student: A look at seven academic retention programs and what makes them effective. Diverse Issues in Higher Education, 18(18):34, 2001.

[4] B. Cook and N. Pullaro. College graduation rates: Behind the numbers. American Council on Education, 2010.

[5] K. Crockett, M. Heffron, and M. Schneider. Targeting financial aid for improved retention outcomes. Targeting Financial Aid for Improved Retention Outcomes, 2011.

[6] E. Jamelske. Measuring the impact of a university first-year experience program on student gpa and retention. Higher Education, 57(3):373—-391, 2009.

[7] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. biometrics, pages 159–174, 1977.

[8] E. J. M. Laur´ıa, J. D. Baron, M. Devireddy, V. Sundararaju, and S. M. Jayaprakash. Mining academic data to improve college student reten-tion: An open source perspective. In Proceedings of the 2nd Interna-tional Conference on Learning Analytics and Knowledge, LAK ’12, pages 139–142. ACM, 2012.

[9] V. A. Lotkowski, S. B. Robbins, and R. J. Noeth. The Role of Academic and Non-Academic Factors in Improving College Retention. ACT Policy Report. American College Testing ACT Inc, 2004.

[10] A. Nandeshwar, T. Menzies, and A. Nelson. Learning patterns of uni-versity student retention. Expert Systems with Applications, 38:14984– 14996, 2011.

[11] National Center for Educational Statistic. Digest of education statistics. https://nces.ed.gov/programs/digest/2016menu_ tables.asp, 2016.

[12] A. M. Olbrecht, C. Romano, and J. Teigen. How money helps keep stu-dents in college: The relationship between family finances, merit-based aid, and retention in higher education. Journal of Student Financial Aid, 46(1):2, 2016.

[13] S. M. Weiss and T. L. Robinson. An investigation of factors relating to retention of student–athletes participating in ncaa division ii athletics. Interchange, 44(1):83–104, 2013.
[14] P. A. Westrick, H. Le, S. B. Robbins, J. M. Radunzel, and F. L. Schmidt. College performance and retention: A meta-analysis of the predictive validities of ACT® scores, high school grades, and SES. Educational Assessment, 20(1):23–45, 2015.
[15] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek. Improving reten-tion: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In Proceedings of the Third International Conference on Learning Analytics and Knowledge, LAK '13, pages 145–149. ACM, 2013.
[16] C. H. Yu, S. DiGangi, A. Jannasch-Pennell, and C. Kaprolet. A data mining approach for identifying predictors of student retention from sophomore to junior year. Journal of Data Science, 8:307–325, 2010.