Predictive Modeling for Student Retention at St. Cloud State University

Hasith Dissanayake[1] , David Robinson[2], and Omar Al-Azzam[1]
[1]Department of Computer Science and Information Technology, St Cloud State University, Saint Cloud, Minnesota, 56301, USA
[2]Department of Mathematics & Statistics, St Cloud State University, Saint Cloud, Minnesota, 56301, USA
{diha1201, dhrobinson, oalazzam}@stcloudstate.edu

***Abstract*— Student graduation rates have always taken prominence in academic studies since they are considered a major factor in the performance of any university. Accurate models for predicting student retention plays a major role in university strategic planning and decision making. Students' enrollment behavior and retention rates are also relevant factors in the measurement of the effectiveness of universities. This paper provides a comparison of predictive models for predicting student retention at Saint Cloud State University (SCSU). The models are trained and tested using a set of features reflecting the readiness of students for college education, their academic capacities, financial situation, and academic results during their freshman year. Principle Component Analysis (PCA) was used for feature selection. Six predictive models have been built. A comparison of the prediction results has been conducted using all features and selected features using PCA analysis.**

Keywords— Retention probabilities; predictive modeling; student retention; Principal Component Analysis; Bayesian Networks; k-Nearest Neighbor; Random Forest; Artificial Neural Networks.

I. Introduction

Student enrollment behavior remains a significant focus in institutional research. Student retention is a serious national issue, and some academic areas experience it more than others. Knight et al. (2003) argued student retention became a nationwide problem in the 1980's, when retention reached 40%. This statistic suggests every 4th student leaves their college or university before graduation, and Knight called this "leakage from the engineering pipeline" [1].

While having an understanding of the reasons behind student drop-out or transfer behavior is crucial for effective enrollment management, what is even more important is the ability to predict these types of behavior to develop preventive measures. This will allow more precise tuition revenue forecasts, enrollment rate predictions, and help develop successful intervention and recruiting programs [2]. Theoretical data indicates using algorithmic approach and data mining can provide more accurate results when predicting student retention compared to traditional statistical methods [3].

This research uses data mining and an algorithmic approach to predict retention with high accuracy, while identifying the variables that could affect student dropout. Those identified variables were used to create predictive models to predict the likeliness of students' retention in their sophomore year. Considering the nature and the diversity of data available for this study, it was assumed highly accurate predictions could be made.

Data related to the admissions process was obtained from the Admissions Office at St Cloud State University (SCSU). This data was originally collected by SCSU from 2006-2010. Data related to academic records of students was obtained from the Office of Records and Registration. This dataset was extracted from ISRS in 2013 by the Office of Strategy, Planning and Effectiveness at SCSU.

The models used in this study are k-NN (k-Nearest Neighbor), Classification Tree, Random Forest, Binomial GLM, Neural Network, and Bayesian Neural Network. During the process of training the predictive models, the data was first examined for outliers and missing values. Next, these values were replaced with appropriate values to minimize the effect on incorrect predictions. Since the data was obtained through different sources, the datasets were then aggregated to one dataset linking them through the Student ID of each dataset. Out of the combined dataset, 70 variables were identified as relevant factors to student retention. At this point it was determined grouping these variables through a Principal Component Analysis (PCA) could yield better results than using the entire set of variables to train the models. However, the models were trained both ways to test the validity of this hypothesis. Finally the models were compared using their accuracy, sensitivity, specificity, positive predictive value, and negative predictive value, to identify the best model that suited the set of variables that was available.

II. Literature Review

Student retention is a widely researched area in the higher education sector, and it spans over four decades of research. Tinto (2006) states "there has also been a concomitant increase in the number of businesses and consulting firms that have sprung up, each of which claims unique capacity to help institutions increase the retention of their students." This shows the amount of research that goes into student retention in the higher education sector is large. It has eventually become a business to consult institutions on how to retain students. The 2005 National Center for Education Statistics revealed the "national rate of student retention has shown disappointingly little change over the past decade".

According to Tinto (2006), there is still much research work to be done in this field. There is a lack of translation of the research and theory into effective practice. Tinto (2006) further states before the 1970s the reason identified for low student retention rates was the failure on the part of students and not the institution. Students who dropped out were thought to be less able, less motivated, and less willing to defer the benefits of college education. Research carried out after the 1970s, mostly by Alexander Astin (1975, 1984), Ernest Pascarella (1980), and Patrick Terenzini (1980), focused more on the environment, the institutions and the people who govern them, when determining reasons for student retention [4]. With the focus shifting to the institutions, attention to the student-faculty relationships was increased, mostly outside the classroom. Also, more focus was placed on the transition to college, by introducing extended orientation, freshman seminars, and a variety of extracurricular programs for first-year students [4]. These programs were introduced to make students feel welcome to the new culture, community or the new environment in college. Around the 1970s, institutions held the view students needed to break away from the society they were in, in order to adapt to college and thereby remain in college. But later, they discovered the gap between breaking away from society and adapting to college should be bridged through orientation programs and extracurricular programs, making them feel part of their past communities, families, churches or tribes [4].

Predictive modeling for student retention can be seen as early as 1975 with Tinto's model. Following Tinto's model, there have been many models introduced by researchers considering different factors and variables to predict student retention. Some of these models focused on identifying students with high risk of dropping out from college [5] [6]. Furthermore, Alkhasawneh & Hargraves (2014) credit Ben Gaskin (2009), with traditional methods of statistical analysis such as logistic regression being used to predict student retention. In most recent years, data mining, which is recognizing patterns in large data sets and then understanding those patterns, has been used to study student retention because of high accuracy and the robustness of missing data [5].

Tinto's model considered social and academic impacts on a student's decisions (voluntarily or involuntarily) to drop out from college. The model is based on Durkheim's (1961) theory of suicide, especially its notions of the cost-benefit analysis of individual decisions about investment in alternative educational activities, which comes from the field of economics of education. Tinto (1975) makes the connection with Durkheim's suicide theory by considering the case of dropping out as committing suicide and views college as a social system. He then relates all the reasons behind committing suicide to that of dropping out of a college when it is viewed as a social system. Furthermore, as colleges also consist of an academic system, he combines the academic factors to the model to be more effective and to shape it to be more suitable to the college structure [7]. These factors are valid even for today's college structure and should be considered as inputs in the predictive models even today.

In addition to Tinto's ideas in 1975, Astin (1993), in his I-E-O (Input-Environment-Output) model, suggests in addition to factors affecting student retention during their college life, researchers should also explore factors affecting student retention before entering college, such as race/ethnicity, gender, family background (which he calls "pre-college characteristics"), high school GPA, and student self-reported data. Astin discusses how these factors affect one's academic and social life while that person is in college and how it could affect the decision she/he makes [8]. Other than the factors Tinto focused on in his research, the factors Astin discusses in his research also could be variables that might change the outcome of a retention model and make it more accurate.

This study is an attempt to identify factors contributing to student retention, primarily taking into consideration the models of both Tinto (1975) and Astin (1993), data obtained from the SCSU Admissions Office on student admission and first year grades, and data from first year student surveys. Knowledge and research of time to degree (TTD) completion remains one of the blind spots of student retention studies. A number of models have been developed to build successful predictions: k-NN, decision trees, Bayesian networks and neural networks, among others [9].

III. Research Methodology

When conducting a study to predict student retention in colleges it is important to choose the appropriate input data and variables, as well as a modelling framework. The main variables used to predict student retention can be categorized into the following groups: financial, encouragement from friends and family, college/university integration, ethnicity [1], academic performance, social integration, institutional commitment, goal commitment, and persistence [10]. Accounting for all of the important variables is crucial for building a successful model with a strong prediction power. Herzog (2006) identified the institutional support factor as a key element in predicting student retention, while Skhakil (2002) observed differences in behavior among commuters and resident students and proposed a degree of connectedness as a most likely explanatory variable.

A. Predictive models and their accuracy

Comparison of the efficiency and accuracy of various methods to predict retention rate and TTD has been performed by multiple studies, revealing contradictory results. Luan (2002) suggests the decision tree analysis method is a better predictor of community college student transfer rate, when compared to neural network analysis. It is important to identify a specific research goal and have a good understanding of the available data type, prior to choosing a specific method [9]. Neural network models have been criticized for poor performance, however, they are widely utilized for data mining (see Table 1) and are considered to have a strong predictive power [14].

Kabakchieva (2012) has simultaneously compared three models (neural network, decision trees and k-NN) to evaluate their prediction power when studying student performance. The highest accuracy was observed when using the neural network model (73.59%), followed by the decision tree demonstrating 72.74% accuracy. The k-NN model resulted in 70.49% accuracy [12].

TABLE I. COMPARISON OF ARTIFICIAL NEURAL NETWORKS (ANN), TREES, K-NN AND BAYESIAN NETWORKS MODELS (AFTER: HASTIE, TIBSHIRANI & FRIEDMAN, 2009)

Characteristics	ANN	Trees	k-NN	BNN
Handling of mixed type data	poor	good	poor	good
Handling missing values	poor	good	good	good
Robustness of outliers	poor	good	good	good
Computational scalability	poor	good	poor	good
Handling irrelevant input data	poor	good	poor	medium
Interpretability	poor	medium	poor	medium
Predictive power	good	poor	good	good
Extracting linear combination of features	good	poor	medium	medium

Decision trees are considered one of the best "off-the shelf" methods of data analysis in terms of speed, interpretability, computational scalability, etc. [14].

B. *Data used in this research*

The data used in this research was obtained through the Admissions Office and the Office of Strategy, Planning, and Effectiveness at SCSU. This data includes the following categorical data which was used to identify the input variables for the models. The following subsections use example variables from the set of variables available, to better explain the categorization.

1) Readiness for a college education: The first category we considered was the likelihood of student retention in school relative to the applicant's level of readiness when applying for college. For example, the variable *AdmitDaysBeforeTerm* can be used as a parameter to measure this likelihood. It would entail the number of days starting from the day the student receives admission to college, to the academic year start date. It is possible to assume the earlier the student gets admission, the earlier s/he would have started the application process. And this would help us conclude that the student's decision to get a college education was well thought out.

2) Academic capacity: The variable *ACE_Student*, identifies a student who is provisionally admitted and needs more preparation for college studies. This can be used to determine the academic capacity of students. It is more likely that students who are not provisionally admitted will perform better in their college education. This would lead to academic satisfaction; a factor directly linked to student retention at any stage.

3) Financial stability: There are a few variables indicating financial stability of the student. Apart from the variables related to Federal student aid or scholarships provided by the college, there are other variables such as, *MilesToSCSU,* that is, the number of miles to SCSU from the applicant's home. If a student thinks that s/he is too far from school, in which case the student must rent an apartment closer to the school or seek university housing, this could impose additional financial stress on the student, and could eventually lead to the student dropping out from college.

4) Other variables: These are variables that do not represent any of the above mentioned categories, but are still important for the models. Some examples are as follows.

a) ClosestToSCSU: This parameter determines if the closest college to the applicant's home is SCSU or not. This could later have an effect if the student decides to transfer to a closer school. In that case, SCSU fails to retain the student since the student has another school in closer proximity than SCSU.

b) FirstGenStudent: Being a first generation college student could have a negative effect on continuing college studies. Reasons for this could be lack of guidance or motivation from family.

c) FAFSA number: This is the number applicants mark SCSU in the precedence order of the FAFSA list. If this number is low, it is likely that the student will try to transfer to a school which s/he prefers.

d) Age: The student's age could be a determining factor when deciding if s/he wants to continue studying. It is highly likely the student would have commitments and responsibilities other than her/his education, the older in age the student is. This could be another reason why adult students drop out from college.

C. *Data cleaning process*

A data cleaning process was conducted using the following steps:

- Performed data type conversion (character to numeric, character to factor)
- Aggregated data to student ID level
- Identified outliers and removed them appropriately
- Cleaned/replaced missing values
- Merged data into one dataset

1) Identifying outliers: The following variables were identified as ones with outliers and were treated accordingly.

a) ACT_Composite scores: Outliers were replaced by the mean value of the said dataset taken without the outliers.

b) MilesToSCSU: Outliers were rectified using a log function on the variable. Before the log function was calculated, all the data points were increased by one to avoid log(0) resulting in NA values after the calculation.

c) HousingAppDaysBeforeTerm: This variable had outliers both from positive and negative sides. Outliers were replaced with the maximum values of either sides.

d) AdmitDaysBeforeTerm: All the outliers were replaced with the maximum value of the dataset taken without the outliers.

e) T1_TermCumulativeLocalCreditsEarned: The outliers which were identified for this variable were replaced by the maximum value of the dataset when taken without the outliers.

f) T2_TermCumulativeLocalCreditsEarned: The outliers which were identified for this variable were replaced by the maximum value of the dataset when taken without the outliers.

2) Cleaning missing values: The variables which had missing values and the respective number of missing values are shown in the Table 2.

TABLE II. NUMBER OF MISSING VALUES FOR EACH VARIABLE IN THE DATASET

Variable	Number of missing values
ACT_Composite	5456
HS_GPA_4Scale	4179
HSPct	5653
QPP	5
EnglTotal	3229
ClosestToSCSU	464
MilesToSCSU	464
HousingAppDaysBeforeTerm	9044
FAFSADaysBeforeTerm	2690
FAFSADayOfYear	2690
FAFSA_Nbr	2690
TotalCRHR_TRSF	8312
TransferGPA	8484
TransferQP	8312
HS_MnSCURegion	502
T1_TermGPA	299
T2_TermGPA	2502
T2_TermLocalCreditsAttempted	2249
T2_TermLocalCreditsEarned	2249

a) QPP: The missing values of this variable was replaced with the mean value of the QPP dataset.

b) ClosestToSCSU: Since this is a Boolean variable all the missing values were considered as not reported and substituted with a zero (false).

c) MilesToSCSU: After considering the retained records out of these missing value records, the maximum value of the dataset was substituted for the missing values.

d) FAFSADaysBeforeTerm: The effects of the records with missing values in the *FAFSADaysBeforeTerm* variable on retention shows 69.5% of the time a student was retained when a value was present for *FAFSADaysBeforeTerm* variable. The same number was at 75.7% when the variable had a missing value.

Furthermore, by looking at the boxplot of the *FAFSADaysBeforeTerm* vs *T3_Enrolled* which is shown in Fig. 1, it can be concluded the *FAFSADaysBeforeTerm* variable has very low impact on retention.

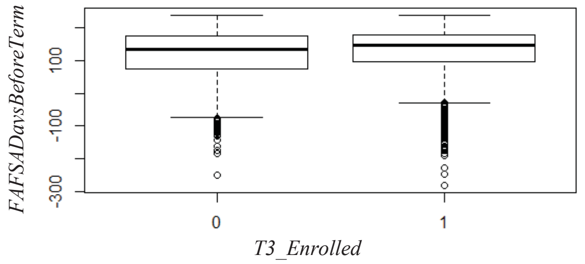

Fig. 1. Boxplot of the *FAFSADaysBeforeTerm* vs *T3_Enrolled*

Considering these facts, it can be assumed that using a fixed value for missing values, rather than a statistic calculated from the sample (like mean), can make the results more robust. This is because when a new dataset is processed, it will have the same replacement values for the missing data, and thus will make the prediction more consistent.

All the variables that had missing values related to FAFSA behaved in an identical way. Therefore, *FAFSADayOfYear* and *FAFSA_Nbr* were also treated the same way as *FAFSADaysBeforeTerm.* Missing values were replaced with a zero.

e) EnglTotal: Whenever there is no value for *High school English subject total*, -1 is assigned to that data point. Hence all the missing values were treated as instances, the student did not have an English score to report and were given a value of -1.

f) HS_GPA_4Scale, ACT_Composite, HSPct(High School Percentage): These missing values were replaced by mean value of the available dataset of the same variable.

g) Other variables: The remaining variables could be considered as instances where data were not reported. These variables were replaced with a value of zero (0).

D. Building Models

Once the dataset was cleaned, there were 68 variables excluding Student ID and T3_Enrolled variables, meaning there were 68 covariates that could be used as predictors with T3_Enrolled being the target variable. The distribution of the target variable is shown in Table 3.

TABLE III. DISTRIBUTION OF T3_ENROLLED VARIABLE

T3_Enrolled	Frequency	Rate
TRUE	10968	0.7067
FALSE	4551	0.2933

1) Preprocessing: Seventy five percent (75%) of the available data was used to train the selected model. The other 25% of the data was used as the test dataset to verify the validity of the final trained model. Choosing a model was done by comparing the number of models for their accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. In the comparison stage, the models were built and tested using the same sample dataset.

2) Identifying near-zero variance variables: In some situations, the data has predictors that only have a single unique value (i.e. a “zero-variance predictor”). For many models (excluding tree-based models), this may cause the model to crash or the fit to be unstable. Similarly, predictors might have only a handful of unique values that occur with very low frequencies. These predictors may become zero-variance predictors when the data are split into cross-validation/bootstrap sub-samples or when a few samples may have an undue influence on the model. These “near-zero-variance” predictors may need to be identified and eliminated prior to modeling[24]. After removing the near zero variance variables, there were 64 variables left as predictors.

3) Principal Component Analysis: In some cases, there is a need to use principal component analysis (PCA) to transform the data to a smaller sub–space where the new variables are uncorrelated with one another. Using a number of correlated predictors may lead to over fitted models. PCA transformation eliminates this problem as it results in uncorrelated variables which will improve the results of the predictions.

Principal Component Analysis is carried out usually to achieve the following:

- Extract the most important information from the data table.
- Compress or reduce the size of the data set by only keeping this important information.
- Simplify the description of the data set.
- Analyze the structure of the observations and the variables.

When Principal Component Analysis is carried out, a new set of variables is computed. These new variables, called Principal Components, are attained as linear combinations of the original variables[25].

After removing the near-zero variance variables from the data set, a PCA was performed to compress the data set with a threshold of 95%. After the Principal Component Analysis, the number of variables were reduced from 64 to 35.

4) Training Models: Once the preprocessing of data was done, each of the models were built to use in the comparison stage. The models which were built in this phase were KNN (K Nearest Neighbor), Classification Tree, Random Forest, Binomial GLM, Neural Network, and Bayesian Neural Network. After building each model, accuracy, sensitivity, specificity, positive prediction value, and negative prediction value was calculated using the confusion matrix (refer comparing models phase for description of these terms).

Building models were done in two different phases. First, the models were built using PCA compressed data, and secondly using the entire variable set. These two phases were done to compare and identify which model, using which set of variables, had the best statistics. It was also used to test the hypothesis from before, the Principal Component Analysis should improve the results of prediction.

IV. Experimental Results

A. Comparing Models

The results obtained during the model building phase were used in this stage to determine if the hypotheses made about using PCA will result in better predictive models. Secondly, the best model was selected by looking at the statistics of the trained models.

1) Measures used to compare models: A few measures were used to compare the models: accuracy, sensitivity, specificity, positive predictive value, and negative predictive value.

a) Confusion Matrix: Confusion matrix is used to measure the performance of a predictive model. These are often used in classification models. A confusion matrix is composed of true positives, true negatives, false positives, and false negatives, which are statistics calculated using predicted values and actual values.

TABLE IV. Example of a confusion matrix

n=1000		**Prediction**	
		FALSE	TRUE
Actual	FALSE	175	35
	TRUE	115	675

b) Accuracy: Accuracy is how probable it is on average for the prediction of the model to be correct. This is calculated as the proportion of the number of correctly classified cases to the total number of cases. Correctly classified cases are the total of true positives and true negatives. Equation (1) represents the calculation of accuracy, n being the total number of cases.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{n} \tag{1}$$

c) Sensitivity: Sensitivity is how probable it is to classify a case as true when it is actually true. This is calculated as the proportion of correctly classified true cases within actual true cases. Equation (2) represents the calculation of sensitivity.

$$Sensitivity = \frac{True\ Positives}{Actual\ Trues} \tag{2}$$

d) Specificity: Specificity is how probable it is to classify a case as false when it is actually false. This is calculated as the proportion of correctly classified false cases within actual false cases. Equation (3) represents the calculation of specificity.

$$Specificity = \frac{True\ Negatives}{Actual\ Falses} \quad (3)$$

e) Positive Predictive value: Positive predictive value is how probable it is for a case to be actually true when the model predicts it to be true. In the context of this thesis problem, the probability would be how likely it is that a student will retain, when the model identifies the student as a student who will retain. This is calculated as the proportion of correctly classified true cases within predicted true cases. Equation (4) represents the calculation of positive predictive value.

$$Pos.\ Pred.\ Value = \frac{True Positives}{Predicted\ Trues} \quad (4)$$

f) Negative Predictive Value: Negative predictive value is how probable it is for a case to be actually false when the model predicted it to be false. In the context of this thesis problem, it is how likely it would be for a student to dropout, when the model identifies the student as a student who will dropout. This is calculated as the proportion of correctly classified false cases within predicted false cases. Equation (5) represents the calculation of negative predictive value.

$$Neg.\ Pred.\ Value = \frac{True\ Negatives}{Predicted Falses} \quad (5)$$

2) Statistics of trained models: Trained models were tested using the same sample dataset. Table 5 and 6 display the statistics of the results.

TABLE V. CALCULATED STATISTICS OF PREDICTIONS OF MODELS USING PRINCIPAL COMPONENT ANALYSIS STRUCTURE

	Accuracy	Sensitivity	Specificity	Pos.pred. val	Neg.pred. val
KNN	0.8337	0.9634	0.5139	0.8301	0.8506
Classification Tree	0.8136	0.8915	0.6215	0.8531	0.6992
Random Forest	0.8587	0.9535	0.625	0.8624	0.8451
Binomial GLM	0.8307	0.9352	0.5729	0.8437	0.782
Neural Network	0.8487	0.9493	0.6007	0.8542	0.8277
BNN	0.8527	0.9577	0.59375	0.8532	0.8507

TABLE VI. CALCULATED STATISTICS OF PREDICTIONS OF MODELS USING ORIGINAL SET OF VARIABLES

	Accuracy	Sensitivity	Specificity	Pos.pred. val	Neg.pred. val
KNN	0.7405	0.9437	0.2396	0.7536	0.633
Classification Tree	0.8357	0.9563	0.5382	0.9362	0.8334
Random Forest	0.8477	0.9479	0.6007	0.8541	0.8238
Binomial GLM	0.8307	0.9324	0.5799	0.8485	0.7767
Neural Network	0.8407	0.9338	0.6111	0.8555	0.7892
BNN	0.8507	0.9507	0.6042	0.8555	0.8325

The KNN algorithm had an accuracy improvement of 0.0932 when the PCA structure was used to train the model. Particularly, the specificity of KNN when using PCA was improved substantially. Improvement of specificity was 0.2743. That is a 27% increase when PCA was used. Sensitivity was also improved by 0.0197.

The Classification Tree did not show an overall performance improvement when the PCA structure was used. However, specificity was 0.0833 higher when PCA was used whereas sensitivity was 0.0648 higher when the original set of variables was used. Overall, there was a 0.0221 or 2.21% improvement of accuracy when the original set of variables was used to train this model.

Random Forest indicated higher performance in all measures when the PCA structure was used for training the model. It had a 0.0056 higher sensitivity value and a 0.0243 improvement of the specificity value. The accuracy improvement was at 0.011 when the PCA structure was used. Even though it did not show drastic improvements when the PCA structure was used, considering all measures had higher values than using the original set of variables, it is logical to determine Random Forest had higher performance when the PCA structure was used.

Binomial GLM did not have any difference in accuracy value in either cases. Even though sensitivity was 0.0028 higher when PCA was used, specificity was reduced by 0.007. Overall, this model did not show any significant difference when the PCA structure was used over the original set of variables.

Neural Networks showed improvement in both accuracy and sensitivity with 0.012 and 0.007 respectively when PCA structure was used. Yet, specificity dropped by 0.0104. Overall, the Neural Network model performed well when the PCA structure was used.

BNN model also behaved in a similar way to Neural Network model. It had higher accuracy and sensitivity with improvements of 0.002 and 0.007 respectively, but, specificity decreased by 0.0105. The overall performance was better when the PCA structure was used to train BNN.

V. CONCLUSION

By comparing the statistics in Table 5 with the same statistics in Table 6, it can be concluded there are no clear cut results to confirm the hypothesis made about the principal component analysis always being true. Or that PCA yields better results than the results obtained by using the original set of variables. However, when the accuracy value is compared, it is evident in most cases, models built using the PCA structure yielded better results. But, in some cases, the models built using the original set of variables were better. Namely, KNN had the most significant difference using the PCA structure, while Random Forest had a slight advantage over the original variable list. Binomial GLM, Neural Network, and BNN had very little to no advantage using the PCA structure. The Classification Tree model had a considerable advantage when the original set of variables was used over the models trained using PCA structure.

Apart from the Classification Tree model, all the other cases had improvements of sensitivity when the PCA structure was used in training models. Half the models showed improvements in specificity when the PCA structure was used. Namely, KNN, Classification Tree, and Random Forest had better specificity values with PCA while Binomial GLM, Neural Networks and BNN had dropped in the same category.

Out of the six models that were used, Random Forest was the only model that showed improvement in all areas when the PCA structure was used. It was also the model with the highest accuracy value. The overall average of accuracy of all models was approximately 84%. Binomial GLM showed the lowest accuracy at 83.07%, while Random Forest showed the highest accuracy at 85.87% when the PCA structure was used.

An analysis of the results reveal Random Forest and BNN had the highest accuracy values and they were very close to each other. Both models had the highest accuracy when the PCA structure was used. However, BNN had a slightly higher positive predictive value when the model was trained using the original set of variables. It can be concluded either of these models can be used with PCA data structure to predict retention at St Cloud State University with a very high accuracy. Furthermore, it can be concluded the use of data mining and an algorithmic approach to predict retention can yield results with high accuracy.

REFERENCES

[1] Knight, D. W., Carlson, L. E., & Sullivan, J. F., Staying in Engineering: Impact of a Hands-On, Team-Based, First-Year Projects Course on Student Retention. American Society for Engineering Education Annual Conference & Exposition. Session 3553. 2003.

[2] Richardson, J. & Dantzler, J., Effect of a Freshman Engineering Program on Retention and Academic Performance. Frontiers in Education Conference Proceedings, ASEE/IEEE, pp. S2C-16-S2C-22. 2002.

[3] Education – US News & World Report. “Freshman Retention Rate”. Retrieved from http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/national-universities/freshmen-least-most-likely-return

[4] Tinto, V., “Research and practice of student retention: What next?”. Journal of College Student Retention: Research, Theory and Practice, 8(1), 1-19. 2006.

[5] Alkhasawneh, R., & Hargraves, R. H., “Developing a Hybrid Model to Predict Student First Year Retention in STEM Disciplines Using Machine Learning Techniques”. Journal of STEM Education: Innovations & Research, 15(3), 35-42. 2014.

[6] Herzog, S., Estimating Student Retention and Degree-Completion Time: Decision Trees and Neural Networks Vis-à-Vis Regression. New directions for institutional research, 131, pp. 17-33. 2006.

[7] Tinto, V., “Dropout from Higher Education: A Theoretical Synthesis of Recent Research”. Review of Educational Research, (1). 89. 1975.

[8] Astin, A. W. (1993). Assessment for excellence: the philosophy and practice of assessment and evaluation in higher education. Phoenix, AZ: Oryx Press.

[9] Bogard, M., Helbig, T., Huff, G., & James, C. A comparison of empirical models for predicting student retention. White paper. Office of Institutional Research, Western Kentucky University. 2011.

[10] Cabrera, A., Nora, A., Castaneda, N., College persistence: Structural equations modeling test of an integrated model of student retention. The Journal of Higher Education, 64(2), 123–139. 1993.

[11] Yadav, S. K., Bharadwaj, B., & Pal, S., Mining Education data to predict student's retention: a comparative study. arXiv preprint arXiv:1203.2987. 2012.

[12] Kabakchieva, D., Student performance prediction by using data mining classification algorithms. International Journal of Computer Science and Management Research, 1(4), 686-690. 2012.

[13] Cover, T. M., & Hart, P. E., Nearest neighbor pattern classification.Information Theory, IEEE Transactions on, 13(1), 21-27. 1967.

[14] Hasti, Tibshirani and Friedman. Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Springer-Verlag. 2009.

[15] Arifovic, J., & Gencay, R., Using genetic algorithms to select architecture of a feedforward artificial neural network. Physica A: Statistical mechanics and its applications, 289(3), 574-594. 2001.

[16] Alkhasawneh, R., & Hobson, R., Modeling student retention in science and engineering disciplines using neural networks. In Global Engineering Education Conference (EDUCON), 2011 IEEE (pp. 660-663). IEEE. 2011.

[17] Yang, J., Ma, J., Berryman, M., & Perez, P., A structure optimization algorithm of neural networks for large-scale data sets. In Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on (pp. 956-961). IEEE. 2014.

[18] Brown, J. L., “Developing a freshman orientation survey to improve student retention within a college”. College Student Journal, 46(4), 834-851. 2012.

[19] CollegeBoard Advocacy. “How Colleges Organize Themselves to Increase Student Persistence: Four-Year Institutions”. Retrieved from https://professionals.collegeboard.com/profdownload/college-retention.pdf, April 2009.

[20] Student Financial Aid Services Inc. “What is FAFSA?” Retrieved from http://www.fafsa.com/understanding-fafsa/what-is-fafsa

[21] Luan, J., & Serban, A., Knowledge Management: Building a Competitive Advantage in Higher Education. New Directions for Institutional Research, 113. San Francisco: JosseyBass, 2002.

[22] Skahill, M., The role of social support network in college persistence among freshmen students. Journal of College Student Retention, 4(1), 39-52, 2002.

[23] U.S. Department of Education. “Federal Pell Grant Program”. Retrieved from http://www2.ed.gov/programs/fpg/index.html

[24] Kim, J.-H.. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics and Data Analysis, 53(11), 3735–3745. doi:10.1016/j.csda.2009.04.009

[25] H. Abdi and L.J. Williams, "Principal Component Analysis", Wiley Interdisciplinary Reviews: Computational Statistics, 2010.