Finding the Causes of Student Attrition in U.S. Higher Education

Ryan Kallicharran[1] and Tereza Shterenberg[2]

Abstract— **Nationwide, one-third of enrolled students leave the educational entity they entered without a degree and the school is unable to identify those problematic student cases. It is crucial to be able to predict the future outcome of those ambiguous cases, because the failure to do so leads to immense amounts of wasted resources such as money and time on both student and school sides. This task becomes arduous due to the diverse student body. By analyzing CUNY Hunter College's students from 2005-2015, this paper offers models of student attrition prediction through the application of machine learning algorithms such as logistic regression, Sequential Minimal Optimization and Bayes Network. We aim to find consistencies and correlations among the proposed methods, and by doing so disentangle the core reasons for attrition. Application of our findings can be proposed to CUNY and other educational establishments as a means of attrition detection and possible ways of retention intervention.**

I. INTRODUCTION

Student attrition is one of the top concerns for many college establishments. To put it simply, dropouts indicate wasted resources invested in students and loss of revenue. A lot of research has been done to determine the causes of college drop-out but this remains a phenomenon. Because of the complexity of drop-out patterns, each college needs to analyze the extent of their own attrition problem. "As colleges scramble for students, then, it becomes increasingly important to characterize...the potential dropout; to determine the reasons why he or she might withdraw, and to see if procedures or programs could be established to help reduce those numbers that are going back out the open door." [4] The complexity of the issue lies with the uncertainty of the patterns within the data. For example, one would think that students who have low grades would be more likely to drop out, but it turns out that it is equally likely that students with high grades drop out.

We aim to tackle this complex problem by considering all the reasons as to why a student may drop out and choosing only the ones that can be used analytically. Ultimately, a students age, semester GPA, and admission scores made the list of factors that contribute to student attrition. In this paper:

- We applied three machine learning algorithms (i.e Bayes net, logistic regression, SMO) and briefly explained their methods of prediction.
- We explained in detail the specifics of each measures of the outputs of the algorithms.
- Described preprocessing strategies applied to the original data set.

The purpose of our study is to help colleges identify students that will potentially drop out, giving an opportunity for the colleges to intervene, thus lowering student attrition.

II. EXPERIMENT

The problem

The task was identifying a pattern within the data that suggest student attrition. Many of the research papers we reviewed restated this fact. The issue is that many students, who dropped out, had unpredictable causes. At times, colleges will ask students that were leaving to fill out a survey as to assess why they decided to leave, but numerous students do not to fill out the survey. This makes it even more difficult to identify why a student decided to leave. Here lies another problem, where students who transfer or dropped out are not categorized as such. However, we are certain of those who graduated. Let us have a look at our actual data set to illustrate this issue.

TABLE I
SAMPLE DATA

The following data is a representation of the final data pre-possessed [1].

admission_sc = Admission Score
grade_num_x = Semester GPA
semester = current semester

studentid	age	admission_sc	grade_num_x	semester	graduated
1	75.21	81.18	3.23	18	yes
2	19.07	99	3.75	5	no
3	27.59	68	2.57	3	no
4	22.28	81.18	4	9	no

The variables in Table Is can be interpreted as follows. The stidentid is not sorted in any particular fashion and the four cases displayed are picked out at random, for the purpose of illustrating the examples below. The numbers under studentid variable are used as a reference tool in order to be able to point out the specific cases and explain the underlying logic. The age variable is not a whole number, but instead has two decimal points, due to the data collectors preference to record the student's age in the most precise manner, by

*This work was not supported by any organization

[1] R. Kallicharran is a student working towards undergraduate computer science and mathematics degree, Hunter College of The City University of New York `R. Kallicharran at Ryankall.github.io`

[2] T. Shterenberg is a student working towards undergraduate computer science, mathematics and economics degree, Hunter College of The City University of New York

[1] Data pre-processing: an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: 100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. [5]

including the month of students birth into the calculation. The admission_sc variable is the score received by a student on the admission tests. However, not all students have this information recorded (i.e. transfer students), which causes missing data. Dealing with missing data when outputting predictions is a sensitive issue. A common practice used in these types of scenarios, is to use the mean imputation technique, and by doing so replace the missing data with the mean value of the given attribute. In case of the admission_sc variable, the mean value was 81.18, which is the dummy variable inputted into the slots of missing data. Some of the research we have read indicates that age reflects students' motivation to succeed in college. Students below the age of 25 are less motivated and do not know what they really want to study, which causes them to perform poorly. As you can see from Table 1 that students 2 and 4 fall into that age group, however, they perform well. Although, they did not graduate, the data describes them as being motivated. Again, the problem with non-graduating students is that we do not have the accurate data to tell us if they transferred or dropped out. Students above age of 25 and below 37 are more motivated because they are likely to have responsibilities, more focus, family dependents, and have real world experiences. Student 3 defies this notion. In addition, students over the age of 37 are expected to stay in college for a short duration and are not expected to complete a degree. Student 1 not only stayed in college for 18 semesters, but graduated. As the reader may have observed from the example, CUNY Hunter Colleges student population is very diverse and defies conventions in several ways.

Other research has shown that students who perform well on admission tests perform well academically and are likely to graduate. However, student 2 and many other students in the data set are good examples of this going against that idea.

One may argue, that the data presented in Fig. 1 are outliers, but it is not the case. Our data has shown these type of students in large quantities that completely go against all of the research findings.

III. Objective

One of the goals was to read as many research studies done on student attrition and attempt to arrive at a plausible assumption that would predict student's academic performance. This became a difficult task quickly due to the amount of factors that were involved for student attrition. We eventually chose to use the papers that spoke to us the most.

GPA :

Why They Leave: Understanding Student Attrition from Engineering Majors by D. Raj Raman & Brandi N. Geisinger. This paper mostly focuses on engineering majors and why they decided to leave the program. This paper aims to explain why students may leave the majors in detail with different categories. One category is classroom and academic climate, which include inadequate teaching and advising: lack of faculty guidance (encouragement, support, and attaching), competitive or hostile environment and inadequate teaching style. Another category is individualistic culture: lack of sense of engagement or belonging and sense of isolation. While this may be true it is difficult to quantify this category because it is more of a psychological view point of the student. Another category is self-efficacy and self-confidence this includes high school preparation, inadequate mathematics, science, physics, and chemistry preparation. Also inadequate overall high school GPA [1], inadequate high school class rank, ACT [2]/SAT[3] scores. The GPA and ACT/SAT scores were considered but this does not reflect the older students well since they are coming back to school after such a long time. Other categories include interest, goals, race and gender. Again these are not quantitative entities. Lastly, the category that stands out the most are grades and conceptual understanding. Here we only care about the grades (GPA) because there is no ambiguity, from a mathematical perspective, therefore it is easier to work with. Low course grades drive students away, thus a drop-out is more likely to occur (regardless of conceptual understanding).

Age:

Understanding Student Attrition in the Six Greater Toronto Area (GTA) Colleges by Tet S. Lopez-Rabson (Seneca College) et al. Study conducted in Canada across six colleges that seeks to better understand the factors motivating college departure and identify post-attrition pathways that college dropouts undertake. This paper has a great deal of information, but most of it cannot be used since we do not care so much about the post-attrition. However, the information about student age is very insightful. The age group below 25 has a common reason for leaving. It shows them having a high rate of attrition due to financial, academics (changes/issues) and interest in program while personal reasons are low. Personal reasons being external factors such as family related or factors unrelated to college. The ages between 26 to 36 has a high dropout rate in personal reasons while the others were low. The students older than 37 are of a smaller percentage and most of them are just going for fun rather than obtaining a degree. With the small amount of students above the age of 37, we believe that it will only help when trying to build a prediction model. However, if it affects the prediction model, it will have a very minor impact.

Admission Scores:

Student Attrition: Consequences, Contributing Factors, and Remedies by Ascend Learning, LLC. This paper mostly focuses on nursing schools and explains the interactions between program and student. Also it mentions several strategies, policies and processes that help improve student

[1] GPA: Grade Point Average

[2] ACT: The ACT test is a curriculum-based education and career planning tool for high school students that assesses the mastery of college readiness standards.

[3] SAT: a test of a student's academic skills, used for admission to US colleges.

attrition. However, the most profound aspect of the paper is about admission tests. The study showed that it can improve the student's success rate in the program, therefore lowering student attrition. Admission scores work perfectly with out data set as well. However, some students are exempt from taking the admission test because of special reasons. This results in missing data within our data set. We found a solution for this problem, which will be explained later.

Summary

Considering all these factors we came up with our hypothesis, which is: Students age, GPA, and admission test scores predict students attrition.

IV. Procedure

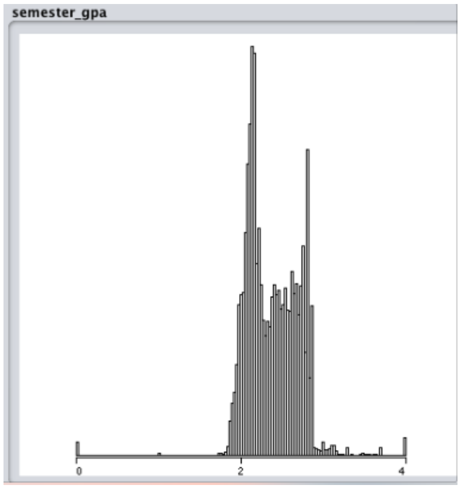
The original data set consisted of 969,104 instances and 234 attributes. To help readers understand common terms in machine learning, we provide some definitions. In the discipline of machine learning an instance is essentially an example, or a case, or a piece of record of a single object of the world, from which a model will be learned, or from which the future prediction will be extrapolated. The attributes in a data set are the feature values, which describe the given instances. The information contained in the attributes defines which type the attribute belongs to. There are several types of attributes, but we will describe only those relevant to the data being analyzed in this work.

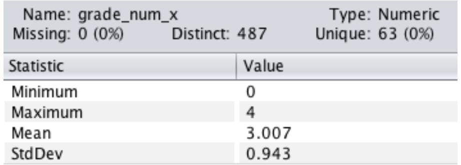
- Nominal- this type of an attribute denotes that there is no ordering between the values, such as last names and colors.
- Numeric- attribute variables are expressed as numbers instead of letters.

Given the fact, that having too many attributes puts the modeling algorithms at risk of poor predictive performance, as well as increases the likelihood of overfitting, removing some of the attributes that are not relevant to this particular work would increase the chances of having more accurate models and predictions. Therefore, the attributes that were left are as follows:

- studentid - serves as an identifier variable to each particular student.
- age - corresponds to a student's age at the time of enrollment.
- start_dt_x - chronologically monitors the time of enrollment in a given class (typically corresponds to a start of a semester).
- admission_sc - lists the admission scores of an enrolled student.
- grade_num_x - numerical grade received by a student in a given class.
- graduated_2005_2015 - indicates whether a student graduated in the time period between 2005 and 2015.
- semester - depicts which semester a given class was taken (e.g. Fall, Spring and etc).
- admissiontypedesc - identifies a type of a student, i.e. transfer within CUNY, or outside of it, or transfer internationally, or previously non-matriculated freshman.

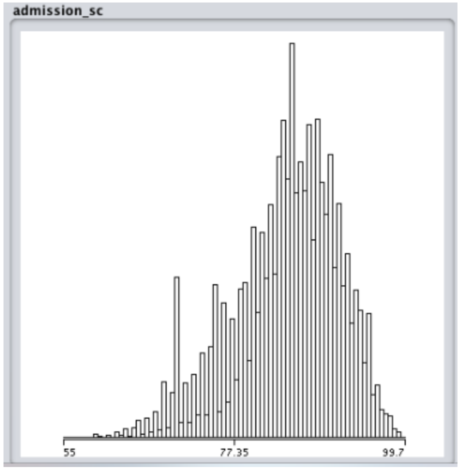
Below are the visual depictions of the data, particularly the variables of interests (e.g Semester GPA, Admission Score, and Age), along with their statistical measures.

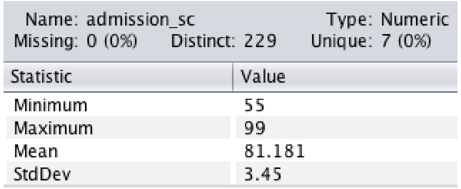




Name: grade_num_x Type: Numeric
Missing: 0 (0%) Distinct: 487 Unique: 63 (0%)

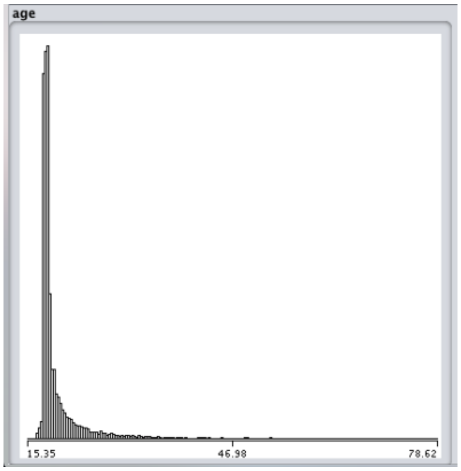
Statistic	Value
Minimum	0
Maximum	4
Mean	3.007
StdDev	0.943

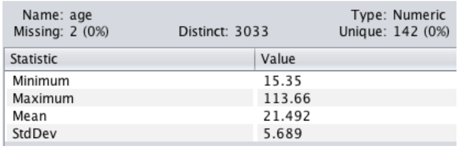




Name: admission_sc Type: Numeric
Missing: 0 (0%) Distinct: 229 Unique: 7 (0%)

Statistic	Value
Minimum	55
Maximum	99
Mean	81.181
StdDev	3.45




Name: age Type: Numeric
Missing: 2 (0%) Distinct: 3033 Unique: 142 (0%)

Statistic	Value
Minimum	15.35
Maximum	113.66
Mean	21.492
StdDev	5.689

Data Pre-processing

The main objective for the data file was to filter out all the unnecessary information. This was done more efficiently by utilizing pandas API for Python. As mentioned in the previous section, the missing data for *admission score* was a concern. After considering the variety of ways to handle the missing data, mean imputation technique was chosen and worked well. Mean imputing simplify acquired the mean of admission scores and replaced the missing values.

The Python script took approximately 1.5 hours to run, mainly because of size of the data.

Algorithms

There is a large number of algorithms in the realm of machine learning that are capable or incorporating these various data types into its modeling techniques and outputting accurate predictions, but the ones being included in this work are Logistic regression, which is a learner, that falls under a regression type of modeling category and its underlying idea is to make linear regression produce probabilities, and by doing so it builds and uses a multinomial logistic regression model with a ridge estimator, taking batches of size 100; Sequential Minimal Optimization (SMO), which is aclassification algorithm, which serves as a fast training method of support vector machines and solves the quadratic programming (QP) problem by using heuristics to partition the training problem into smaller problems, that can be solved analytically; and, BayesNet, which is a Bayesian network learning method, that builds probabilistic graphical model using various search algorithms and quality measures.

The software used to provide all of the predictions below is Weka, which is an open source tool implemented by the Machine Learning Group at the University of Waikato, which provides a collection of machine learning algorithms.

The type of analysis embedded into the modeling techniques and algorithms described above is the 10-fold stratified cross-validation, which divides the given data into two subsets, which are called training and testing sets. In short, training is the process of providing feedback to the algorithm in order to adjust the predictive power of the classifier it produces. While testing is the process of determining the realistic accuracy of the classifier, which were produced by the algorithm. During testing, the classifier is given never-before-seen instances of data to do a final confirmation that the classifier's accuracy is not drastically different from that during training. Validation is performed after each training step and it is performed in order to help determine if the classifier is being overfitted. The validation step does not provide any feedback to the algorithm in order to adjust the classifier, but it helps determine if overfitting is occurring and it signals when the training should be terminated.

V. RESULTS

Below are the results gathered from running the three algorithms described above. The running times for stratified 10-fold cross-validation on each learner are 30 seconds for Logistic regression, 64 hours for Sequential Minimal Optimization, and 10 seconds for Bayes Network. All three learners consistently yielded 75% accuracy. Every measure outputted as the algorithm's result is discussed in detail further.

Logistic Regression

Correctly classified instances 201,841 = 75.53%
Incorrectly classified instances 65,4025 = 24.47%
Kappa statistic 0.50
Mean absolute error 0.24

Fig. 1. Summary of Logistic

As Figure 2 shows, the Logistic learner has produced approximately 75% accuracy and its Kappa statistic is 50%. The accuracy of an algorithm is its degree to which the result of a measurement conforms to the correct value or a standard.

$$accuracy = \frac{\sum_{i=1...N}\left(1-(target_i - threshold(f(\vec{x}_i)))\right)^2}{N}$$

Additionally, the Kappa measure (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). The kappa statistic is used not only to evaluate a single classifier, but also to evaluate classifiers among themselves. In addition, it takes into account random chance (agreement with a random classifier), which generally means it is less misleading than simply using accuracy as a metric (an Observed Accuracy of 80% is a lot less impressive with an Expected Accuracy of 75% versus an Expected Accuracy of 50%). Computation of Observed Accuracy and Expected Accuracy is integral to comprehension of the kappa statistic, and is most easily illustrated through use of a confusion matrix. The Logistic regression produced a kappa statistic of 0.50.

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

The Mean Absolute Error measures how close the prediction was to the actual result. The Logistic learner's MAE was 0.24.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n} |e_i|$$

TABLE II

LOGISTIC REGRESSION OUTPUT

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.79	0.29	0.76	0.79	0.77	0.50	0.83	0.82	yes
0.71	0.21	0.74	0.71	0.72	0.50	0.83	0.82	no

In Table 2 are the precise measures outputted by the Logistic regression. The measure of True Positive (TP) rate measures the proportion of positives among the examples, which were classified as positives (e.g., the percentage of students, who truly graduated and were predicted to graduate by the given algorithm). True Positive rate is equivalent to Recall. The top row corresponds to the students, who were predicted to graduate and, in fact, did. There Logistic regression yielded 79% for a ”yes” graduating class, and 71% for a ”no” class. In the confusion matrix, this is the diagonal element divided by the sum over the relevant row, i.e. 113,269/(113,269+30,389)=0.79 for class ”yes” and 87,549/(87549+36,036)=0.71 for class ”no” (all values rounded up to two decimal places).

The False Positive (FP) rate is the proportion of examples which were classified as graduating, but in actuality did not graduate, In the matrix, this is 36036/(36036+87549)=0.29 for class ”yes” 30389/(113269+30389)=0.21 for class ”no” The Precision is the proportion of the examples which truly have class ”yes” among all those which were classified as class ”yes” and vice verse. In the matrix, this is the diagonal element divided by the sum over the relevant column, i.e. 113269/(113269+36036)=0.76 for class yes and 87549/(87549+30389)=0.742 for class no.

The F-Measure is simply 2*Precision*Recall/(Precision+Recall), a combined measure for Precision and Recall. The Matthews correlation coefficient (MCC, a.k.a phi coefficient) takes into account true and false positives and negatives and is considered to be a balanced measure, which can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications. It can be calculated directly from the confusion matrix using the formula provided below. In both classes the MCC in Logistic regression yielded 0.50.

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Receiver operating characteristic or ROC area quantifies the overall ability of the test to discriminate between those individuals who graduated and those who did not. Logistic regression produced 83% on both classes. The precision-recall (PRC) plot shows precision values for corresponding sensitivity (recall) values. Similar to the ROC plot, the PRC plot provides a model-wide evaluation. The logistic learner outputted 82% on the PRC measure across both classes.

TABLE III

CONFUSION MATRIX FOR LOGISTIC

	yes	no
yes	113,269	30,389
no	36,036	87,549

Sequential Minimal Optimization

Correctly classified instances 201,841 = 75.53%
Incorrectly classified instances 65,402 = 24.47%
Kappa statistic 0.50
Mean absolute error 0.24

Fig. 2. Summary SMO

TABLE IV

SEQUENTIAL MINIMAL OPTIMIZATION

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.83	0.33	0.75	0.83	0.78	0.51	0.75	0.71	yes
0.67	0.17	0.77	0.67	0.72	0.51	0.75	0.67	no

TABLE V

CONFUSION MATRIX FOR SMO

	yes	no
yes	118,958	24,700
no	40,702	82,883

A. Bayes Network

Correctly classified instances 202,509 = 75.78%
Incorrectly classified instances 64,734 = 24.22%
Kappa statistic 0.51
Mean absolute error 0.31

Fig. 3. Summary BayesNet

As the results of the stratified 10-fold cross-validation on all three algorithms show, there was a consistent *75%* accuracy on all learners. The BayesNet predictor earned the highest TP of *86%* on the ”yes” class, followed by SMO with an *83%*, and Logistic with *79%*. On the ”no” class Logistic

TABLE VI
BAYESNET

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.86	0.37	0.73	0.86	0.79	0.52	0.84	0.84	yes
0.63	0.14	0.80	0.63	0.71	0.52	0.84	0.84	no

TABLE VII
CONFUSION MATRIX FOR BAYESNET

	yes	no
yes	124,157	19,501
no	45,233	78,352

yearned *71%*, SMO gathered 67%, and BayesNet *63%*. The FP rate on ”yes” class had BayesNet on the first place with *37%*, SMO with *33%*, and Logistic with *29%*. Pertaining to ”no” class, the FP rate was the highest on Logistic, yielding *21%*, followed by SMO with *17%*, and BayesNet’s *14%*. Precision on ”yes” class was highest on Logistic outputting *76%*, SMO earned *74%*, whereas in the ”no” class BayesNet got an *80%*, followed by SMO’s *77%*, and Logistic’s *74%*. The Recall measure is equivalent to TP Rate. F-Measure on the ”yes” class was highest on BayesNet with a *79%*, then SMO with *78%*, and Logistic’s *77%*. The ”no” class’ results of the F-measure are *71%* on BayesNet, *72%* on Logistic, and *71%* on SMO. The highest MCC measure was *52%* achieved by the BayesNet, followed by SMO’s *51%*, and Logistic’s *50%*. BayesNet’s ROC area was *84%*, with an *83%* on Logistic and SMO’s *75%*. The highest PRC area was *84%* on the BayesNet, followed by *82%* on the Logistic, and *71%* on the SMO.

Related Works

- Why They Leave: Understanding Student Attrition from Engineering Majors by D. Raj Raman & Brandi N. Geisinger : *Low course grades drive students away (regardless of conceptual understanding).*
- Understanding Student Attrition in the Six Greater Toronto Area (GTA) Colleges by Tet S. Lopez-Rabson (Seneca College) et al. : *insightful information about student age and reason for drop outs*
- Student Attrition: Consequences, Contributing Factors, and Remedies by Ascend Learning, LLC : *admission score linked to successful students*
- The ”Big Picture”: Key Causes of Student Attrition & Key Components of a Comprehensive Student Retention Plan, by Joe Cuseo (Marymount College) : *root causes of attrition*
- The Institutional Costs of Student Attrition by Nate Johnson, Postsecondary Analytics LLC : *cost analysis of an education entity pertinent to Attainment status across USA*
- College Student Attrition and Retention, by Leonard Ramist, College Board Report No.81-1: *Reasons students drop out*

VI. CONCLUSION

As the reader may have observed, various machine learning algorithms used to test this work’s hypothesis have consistently shown a 75% accuracy with additional measures displaying proper results. This in turn can be interpreted as follows. It can be said with a 75% certainty that a student’s GPA, age and admissions score impact one’s potential future attrition. Therefore, having consistently high GPA throughout one’s academic career, or at least one that doesn’t decrease below a certain threshold, in addition to having a high admission score and having set goals, reduces the risk of attrition.

Future works

Upcoming extension of this work includes the following: Identifying additional attributes that could have a significant contribution to the student attrition prediction rate. Additionally, testing the current prediction models against other school data to see if the approach presented in this work can be applied universally. Finally, the application of the prediction model into other educational establishments would have its challenges but would have potential great outcomes for the school.

REFERENCES

[1] Angelino, Lorraine M., Frankie K. Williams, and Deborah Natvig. Strategies to Engage Online Students and Reduce Attrition Rates (n.d.): n. pag. Web.

[2] Ascend Learning, LLC. (2012). Student attrition: Consequences, contributing factors, and remedies.

[3] Geisinger, Brandi N., and D. Raj Raman pag. Why They Leave: Understanding Student Attrition from Engineering Majors. Iowa State University. Web.

[4] College, 1980. ED 198 851 Rounds, J.C. Attrition and Retention of Community College Students: Problems

[5] Cuseo, Joe. ”The ”BIG PICTURE”: Key Causes of Student Attrition & Key Components of a Comprehensive Student Retention Plan.” PDF. N.p., n.d. Web.

[6] Johnson, Nate. The Institutional Costs of Student Attrition. The Findings and (n.d.): n. pag. Bill & Melinda Gates Foundation. Web.

[7] Gill, Brian P., Christina C. Tuttle, and Ira Nichols Barrer. ”Does Student Attrition Explain KIPP’s Success? - Evidence on Which Students Leave.” RSS. Stanford University, Fordham, Harvard Kennedy School, 29 Sept. 2014. Web.

[8] Pyle, D., 1999. Data Preparation for Data Mining. Morgan Kaufmann Publishers, Los Altos, California.

[9] Ramist, Leonard. College Student Attrition and Retention. New York: College Entrance Examination Board, 1981. Research College Board. Web.

[10] Lopez-Rabson, T. S. and McCloy, U. (2013). Understanding Student Attrition in the Six Greater Toronto Area (GTA) Colleges. Toronto: Higher Education Quality Council of Ontario.

[11] Raisman, Neal. The Cost of College Attrition at Four Year Colleges & Universities (2013): n. pag. The Educational Policy Institute. Web.