Engagement vs Performance: Using Electronic Portfolios to Predict First Semester Engineering Student Retention*

Everaldo Aguiar[†‡]
University of Notre Dame
Notre Dame, Indiana 46556
eaguiar@nd.edu

Nitesh V. Chawla[†‡]
University of Notre Dame
Notre Dame, Indiana 46556
nchawla@nd.edu

Jay Brockman[†§]
University of Notre Dame
Notre Dame, Indiana 46556
jbb@nd.edu

G. Alex Ambrose[¶‖]
University of Notre Dame
Notre Dame, Indiana 46556
gambrose@nd.edu

Victoria Goodrich[§]
University of Notre Dame
Notre Dame, Indiana 46556
vfroude@nd.edu

ABSTRACT

As providers of higher education begin to harness the power of big data analytics, one very fitting application for these new techniques is that of predicting student attrition. The ability to pinpoint students who might soon decide to drop out[1] of a given academic program allows those in charge to not only understand the causes for this undesired outcome, but it also provides room for the development of early intervention systems. While making such inferences based on academic performance data alone is certainly possible, we claim that in many cases there is no substantial correlation between how well a student performs and his or her decision to withdraw. This is specially true when the overall set of students has a relatively similar academic performance. To address this issue, we derive measurements of engagement from students' electronic portfolios and show how these features can be effectively used to augment the quality of predictions.

Categories and Subject Descriptors

J.1 [**Administrative Data Processing**]: Education; K.3.0 [**Computer Uses in Education**]: General

General Terms

Measurement, Performance

Keywords

Electronic Portfolios, Student Retention, Early Intervention, Data Fusion, Learning Analytics, Predictive Analytics

1. INTRODUCTION

Over the course of many years, the education field has gone through several transformations. As new techniques for both teaching and assessing students emerge, universities and other post-secondary institutions are expected to quickly adapt and begin to follow the new norms. Further, as the needs of our society shift, we often see increased demands for professionals in particular disciplines. Most recently, this phenomenon can be observed with respect to the areas of Science, Technology, Engineering, and Mathematics (STEM).

While creating an environment that stimulates student enrollment in these particular fields is a challenge in itself, preserving high retention rates can be a far more complicated task. As [41] highlights, our understanding of retention has considerably changed over time, and efforts to address the issue are ubiquitous in higher education today. Yet, despite the rapid growth of this subject over the last few years, there are clear indications that the complexities involved with helping a highly diverse array of students to succeed are far from being understood.

It is estimated that nearly half of the students that drop out of their respective programs do so within their first year in college [17]. Consequently, a clear focus has been directed towards early identification and diagnose of *at-risk* students, and a variety of studies using statistical methods, data mining and machine learning techniques can be found in recent literature (e.g., [14, 49, 9, 48, 32, 27, 26, 16, 50, 4, 42]).

A downside of these proposed models is that they frequently rely strictly on academic performance, demographic and financial aid data. There is a wide recognition, however, that the reasons for student dropouts can range based on several other factors outside that scope [5, 35, 45, 20, 34, 31, 26]. Moreover, a number of dropout students do not exhibit any early signs of academic struggle as per their grades. The inverse is also true, as there are often highly engaged

[†]Department of Computer Science and Engineering
[‡]Interdisciplinary Center for Network Science & Applications
[§]College of Engineering
[¶]First Year of Studies
[‖]Notre Dame ePortfolio Engagement Program
*This material is based upon work supported by the National Science Foundation under Grant No. DUE 1161222
[1]For the remainder of the paper, the term *dropout* is loosely used to denote both the students that withdraw from the institution and the students that opt out of the College of Engineering.

LAK '14, March 24 - 28 2014, Indianapolis, IN, USA

students who despite performing below the expectations, remain enrolled. Figure 1 illustrates these two specific groups of students.

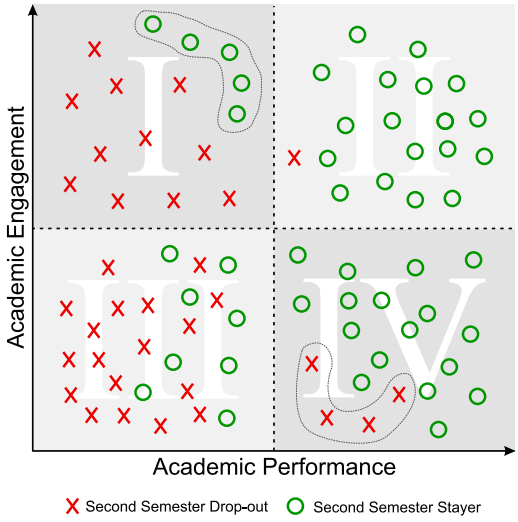

Figure 1: Low performing/highly engaged students (quadrant I) are often retained. High performing/disengaged students (quadrant IV) may drop out.

In this paper, we focus on remedying the shortcomings that arise when classification models are trained using only student academic performance and demographic data. We collected data that describe the access patterns of first-year engineering students to their personal electronic portfolios, which are dynamic web-based environments where students can list and describe their skills and achievements, and we show how these features correlate to and can help enhance the prediction accuracy of student attrition. In particular, we investigate how measurements of student engagement can be used to decrease miss-prediction rates of instances belonging to the groups highlighted in Figure 1.

The remaining portion of the paper is organized as follows. Section 2 gives an overview of the most recent related literature. Section 3 describes the context in which this study was carried out and gives insights as to our decision to utilize electronic portfolios to measure student engagement. Following, section 4 describes our dataset in detail. The methodology and experimental results are covered in sections 5 and 6 respectively, and a brief discussion of our findings concludes this paper in section 7.

2. RELATED WORK

From a sociological standpoint, student attrition has been studied in great detail. Seidman [41] and Tinto [43] provide comprehensive studies that investigate the causes and consequences of this issue. Though related, this ramification of the literature is outside the scope of this paper. Following, we provide a more elaborate description of the most recent works that utilize student data to create and evaluate prediction models for student attrition.

Early work by DeBerard et al. [14] combined academic performance, demographics and self-reported survey data of students in an attempt to forecast cumulative GPA using linear regression, and retention rates via logistic equations. The former achieved commendable results while the outcomes of the later were not statistically significant. Contemporary to that, a study by Zhang et al. [49] showed that high school GPA and math SAT scores were positively correlated to graduation rates of engineering students, while verbal SAT scores correlated negatively with odds of graduation. Similar findings are reported by Mendez et al. [32].

A key premise of our work is highlighted by Burtner [9]. After monitoring a group of incoming engineering students over a three year period, the author concludes that while a predictive model based on cognitive variables such as the students' math and science ability can perform relatively well, it would greatly benefit if non-cognitive factors developed during the freshman year were to be incorporated. Lin et al. [27] validate that idea by showing that the accuracy of their classification models improves after the inclusion of non-cognitive features extracted from a survey.

Yu et al. [48] utilize decision trees to predict student retention, and among other discoveries, the authors report that in their context, student persistence was more closely related to the students' residency status (in/out of state) and current living location (on/off campus) than it was to performance indicators. Likewise, a sensitivity analysis exercise performed on neural networks, decision trees, support vector machine and logistic regression models by Delen [16, 17] ultimately concluded that several important features utilized to predict student retention were not related to academic performance.

With the intent of developing a long-term intervention system to enhance student retention, Zhang et al. [50] tested three different classifiers and observed that the best prediction accuracy for student retention was yield by naive Bayes. Alkhasawneh [4] utilizes neural networks to predict first year retention and provides an extensive analysis of his models' performance. Finally, we highlight the recent work by Thammasiri et al. [42], in which the problem of predicting freshmen student attrition is approached from a class imbalance perspective, and the authors show how oversampling methods can enhance prediction accuracy.

3. CONTEXT

3.1 The College of Engineering

The University of Notre Dame is a medium sized, Midwestern, private institution with a traditional student composition, i.e. the vast majority of students complete their undergraduate studies in four years and are in the age range of 18 - 22. The overall student body is 53% male and 47% female, while the College of Engineering is approximately 75% male and 25% female. First-year students are admitted to the First-Year of Studies program regardless of their intended future major. Students select their major (whether engineering or something else) near the end of their first-year when they register for classes for the upcoming fall semester. Beyond admission / selection into the university as a whole, there are no admission or selection criteria for entering any of the disciplines of engineering; rather, it is based on student interest alone.

With few exceptions, first-year students that are considering an academic pathway within engineering complete a standard first-year curriculum, including the two-semester

course sequence of "Introduction to Engineering," taught within the College of Engineering. Each year the course sequence has enrollments of approximately 450 - 550 students. The course has two main objectives: 1) to expose students to the engineering profession and engineering major options, and 2) to demonstrate the processes of planning, modeling, designing, and executing specified project deliverables. The course curriculum uses a project based learning approach, with students completing a total of three group projects across the two semester sequence. Students are required to attend large lecture sections which introduce basic concepts needed to complete the projects and small group (30 - 35 students) learning centers that focus on hands on learning. For over a decade, the course sequence has included similar material and project based course assignments, including: homework, quizzes, exams, technical reports and presentations.

3.2 ePortfolios for Engagement

The ePortfolios serve as a creative space and a recording system that utilizes digital technologies to allow learners to collect artifacts and examples of what students know and can do, in multiple media formats; using hypertext to organize and link evidence to appropriate outcomes/skills, goals, or standards [6]. ePortfolios capture and document students' learning and engagement through their reflection, rationale building, and/or planning. Chen and Black [11] found ePortfolios generate shared responsibility and ownership of learning between students and instructors since they can be used inside and outside the classroom. They are also available and can be used on and off campus, in face-to-face and virtual environments, and during and after the student's time in college (as a way of practically demonstrating what ABET [1] refers to as "life-long learning" achievements). Atabi et al. [3] found the use of ePortfolios to be valuable as an advising tool, allowing students to track the progress of their learning outcomes, to provide documentary evidence, and used when they meet regularly with their academic advisors for feedback. Significantly, the use of ePortfolios generates intentional and active learners since students become self-aware and take ownership of their academic progress.

Higher education institutions such as Bowling Green State University [23], La Guardia Community College [18], University of Barcelona [29], Spelman College [37], Clemson [40], Penn State and Florida State Universities [47] have begun to implement ePortfolio initiatives to enhance engagement and measure impact through integrating life-wide academic, personal, and professional contexts. Student engagement is a construct that measures the alignment between what effective institutions purposefully do (a range of teaching practices and programmatic interventions) to induce and channel students to desired outcomes, compared with what students actually do with their time and energy towards achieving these educationally purposeful activities [24].

The ePortfolio platform of our choice is supported by Digication [2] and its Assessment Management System (AMS). The Digication paid subscription account not only offers an ePortfolio platform but also provides a powerful back-end course, program, institution, or inter-institution (AMS). Within individual, and across our partnering, institutions, the AMS tracks, compares, and generates customizable reports on student progress and performance by standards, goals, objectives, or assignments.

3.3 ePortfolio Engagement as an Analytic

For too long much of the emphasis on improving retention has focused solely on the binary metric of retention (yes/no). By focusing on student engagement rather than just predictive variables, after-the-fact outcome of retention or a subjective measurement of learning, the ePortfolio provides a window into the time, energy level, and commitment exhibited by students throughout the trajectory of a given course. The assessment focus on retention is too late to interdict and improve learning within a course, especially during the first semester of college.

An ePortfolio engagement analytic has important implications to the emerging field of learning analytics. Johnson et al. [22] define learning analytics as the interpretation of a wide range of data produced by and gathered on behalf of students in order to assess academic progress, predict future performance, and spot potential issues. The goal of learning analytics is to enable educators to understand and optimize learning via an environment tailored to each student's level of need and ability in close-to-real time. Up until now, most of the data sources have been limited to learners' tacit digital actions inside the learning management systems (i.e., discussion forum posts, downloading content to read, login rates, and duration). The ePortfolio tool and platform offers a more authentic environment that could provide a week-by-week measure to identify if and when students are losing engagement and explore why, where, and what is engaging students as well as how they spend their time and energy outside of the class. Therefore, data mining the ePortfolios could generate more effective learning analytics to improve the understanding of teaching and learning, and to tailor education to individual students more effectively.

3.4 ePortfolio Use in the First-Year Engineering Course

In the 2012 - 2013 academic year, ePortfolio assignments were integrated with the traditional course deliverables as a means to guide students' reflections on their education. A total of eleven ePortfolio updates were assigned throughout the academic year. For the course, all students were required to create an ePortfolio following an instructor designed template. The ePortfolio template included three main sections, which were each updated over the course sequence:

1. Engineering Advising – Required reflection on their engineering major choice and their progress towards engineering skill areas. Seven skills areas were defined, each relating to ABET accreditation required outcomes (a - k).

2. Project Updates – Required updates following the completion of each project. Minimally, students were asked to include a picture of their project and a reflection on skills developed through the project.

3. Engineering Exploration – Required reflections after attendance at eight engineering related events that took place outside of the course. Events included seminars, engineering student group meetings, professional development activities, etc. that were delivered by various groups within the university.

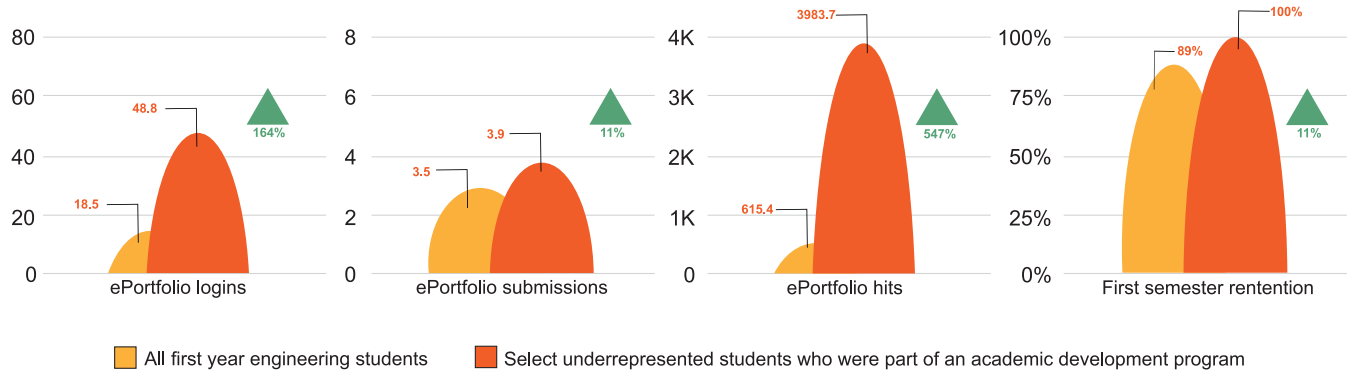

Figure 2: The effect of ePortfolio engagement on first semester retention

Although ePortfolio assignments were a required portion of the course, they were graded largely for completion. Therefore, student effort towards their ePortfolio assignments had wide variability. In addition, students were encouraged to personalize their ePortfolios to include additional pages and information not required by the course. Because students were asked to share this ePortfolio with their advisors after matriculating into engineering departments, they were encouraged to keep any additional content professional in nature.

Given this particular context of how ePortfolios are utilized, we believe that an important correlation between the students' engagement using this tool and retention levels exists and can be potentially mined for predictive analysis. While section 4 will provide more details on our datasets and describe each of the features we analyze, a preliminary case study illustrated by Figure 2 showed that a select group of students who were more exposed to electronic portfolios as part of an enhanced academic program exhibited markedly higher levels of engagement using that tool, and more importantly, was retained in its entirety.

The particular group highlighted was composed of 12 students from underrepresented groups that were selected prior to the fall semester to be part of an academic development program which, among other components, exposed them more extensively to the use of ePortfolios. In addition to maintaining an ePortfolio as part of their Introduction to Engineering class requirements, these students also enrolled a seminar course that made use of the tool. Note, however, that all data collected refers only to the their Introduction to Engineering portfolio. Even so, this group appears noticeably more interactive and engaged than the remaining portion of the students. As Figure 2 shows, they not only exhibited much higher levels of interactivity to their ePortfolios based on all three metrics, but also displayed a retention rate of 100%.

4. DATASET

This study used data collected from a single cohort of incoming freshmen students who were registered in a first semester Introduction to Engineering course. This particular group was made up of 429 students, the vast majority of which had engineering majors listed as their first year intent and remained in the program for the subsequent semester, leading to a very imbalanced dataset. While majors are not formally declared until their sophomore year, students are asked to inform their intended majors when submitting their application package and prior to their first semester on campus.

4.1 Description

A variety of features that describe each student's academic performance, engagement and demographic background were made available to this project from multiple sources. These were then matched student-wise and merged into a single dataset. After an initial analysis of the data, we decided to exclude a number of features that either (1) had no apparent correlation to the outcome variable, (2) directly implied it, or (3) provided redundant information. Further, we also removed 10 instances that had a very considerable amount of missing data. These particular instances corresponded to students that dropped out early in the semester and hence had no academic performance or engagement data available. Table 1 lists and describes each feature available in our final dataset and Table 2 groups these into their respective categories.

It is worth noting that this particular dataset has a highly imbalanced class distribution wherein only 11.5% of the instances belong to the minority class (student dropped out). As described in [42], predicting student retention becomes more challenging when the available training sets are imbalanced because standard classification algorithms usually have a bias towards the majority class. While a wide range of sampling techniques can be used to artificially balance datasets of this kind (e.g., SMOTE [10]), we reserve those optimizations as future work.

4.2 Feature selection

As a second step to preparing our dataset, we carried out a series of tests to investigate how strongly correlated to the outcome each feature was. In general, performing feature selection as a means for reducing the feature space provides some benefits when building classification models. Namely, the model becomes more generalizable and less prone to overfitting, more computationally efficient and easier to interpret.

The following feature selection methods were used: information gain (IG) [38], gain ratio (GR) [39], chi-squared (CS) and Pearson's correlation (CR). The first evaluates the worth of each attribute by measuring its information gain with respect to the class. Gain ratio works in a similar

Name	Type	Description
Adm Intent	Nominal	Intended college major as specified by the student in his/her application package
Adm Type	Nominal	Type of admission earned by student (e.g., early, regular, waiting list)
AP Credits	Numeric	Number of credits earned through AP courses taken prior to college enrollment
Dormitory	Nominal	Name of dorm where the student resides (note: all first-year students are required to live on campus)
EG 111 Grade	Nominal	Letter grade obtained in the introduction to engineering course
ePort Hits	Numeric	Hit count for the student's ePortfolio pages during the fall semester
ePort Logins	Numeric	Number of times the student logged in to his/her ePortfolio account during the fall semester
ePort Subm	Numeric	Number of assignment submitted via ePortfolio during the fall semester
Ethnicity	Nominal	The student's self-declared ethnicity
First Gen	Binary	A flag to denote first-generation college students
FY Intent	Nominal	Intended college major as specified immediately prior to the beginning of the fall semester
Gender	Binary	The student's gender
Income Group	Numeric	A numeric value ranging from 1-21, each corresponding to a different income group segment
SAT Comb	Numeric	Combined SAT scores
SAT Math	Numeric	SAT score for the math portion of the test
SAT Verbal	Numeric	SAT score for the verbal portion of the test
Sem 1 GPA	Numeric	The student's overall GPA at the end of the fall semester
Retained	Binary	A flag identifying students that dropped out immediately after the fall semester

Table 1: Dataset features described

Type	Subtype	Feature
Academic	Pre-Adm	Adm Intent
		SAT Math
		SAT Verbal
		SAT Comb
		AP Credits
	Post-Adm	FY Intent
		EG 111 Grade
		Sem 1 GPA
		Retained
Demographics	-	Adm Type
		Gender
		Ethnicity
		Income Group
		First Gen
		Dormitory
Engagement	-	ePort Logins
		ePort Subm
		ePort Hits

Table 2: Dataset features categorized

	IG	GR	CS	CR
Adm Intent	2	4	2	-
SAT Math	7	2	5	3
SAT Verbal	-	-	-	-
SAT Comb	9	10	9	-
AP Credits	8	8	8	8
FY Intent	5	7	6	9
EG 111 Grade	4	5	3	6
Sem 1 GPA	-	-	-	**1**
Adm Type	-	-	-	-
Gender	10	9	10	5
Ethnicity	-	-	-	-
Income Group	-	-	-	7
First Gen	-	-	-	-
Dormitory	3	6	4	-
ePort Logins	6	3	7	4
ePort Subm	-	-	-	10
ePort Hits	**1**	**1**	**1**	2

Table 3: Feature selection rank

manner while adopting a different metric. CS and CR compute chi-squared and Pearson's correlation statistics for each feature/class combination.

The results of our experiments are summarized in Table 3, where the top 10 features ranked by each method are listed, and the highest correlated one is highlighted for each column.

Several interesting observations can be derived from these results. First, we emphasize that all but one method reported *ePort Hits* as being the most important feature of the dataset. In other words, there appears to be a strong correlation between the number of times a certain student's electronic portfolio pages are visited and that student's decision to stay or withdraw from the College of Engineering. Note that these hits originate from both external visitors and the students themselves. While the current data does not allow us to discern the two scenarios, we suspect that the majority of the hits do in fact come from the portfolio owner. If that is indeed the case, this noticeable correlation could be explained simply by the fact that students whose portfolios exhibit larger number of hits are likely to be those who spend more time editing their pages, creating new content, and submitting assignments (as those actions directly contribute to that student's page hit count). It would then make reasonable sense that this display of engagement should, in some cases, correlate with the chances of this particular student being retained.

Further, we noticed that some of the features had no substantial individual correlation to the class values. For instance, in our particular context, ethnicity, admission type, first generation status, income, and the number of assignments a student sent through his/her ePortfolio did not appear to be closely related with that student's retention in the program. As reported by [49, 32], we also observed minor negative correlations between verbal SAT scores and engineering student retention.

So as to effectively compare the performance of classification models based on traditional academic data to that of

models based on student engagement features, we created four subsets from the original data. These are described bellow:

- **all-academic**: This subset contained all academic and demographics features listed in Table 2.
- **top-academic**: Following the feature selection process described above, this subset contains only the top three academic and demographics features. Multiple wrapper methods (i.e., which can score feature subsets rather than individual features alone) were used, and the final subset chosen contained the following: *admin intent*, *EG 111 grade*, and *sem 1 GPA*.
- **all-engagement**: Contained the three ePortfolio engagement features.
- **top-academic+engagement**: This final subset contained the optimal three-element combination of features across all initially available. These were: *EG 111 grade*, *ePort logins*, and *ePort hits*.

5. METHODOLOGY

For this study, we selected a range of classification methods that have been previously utilized in this particular domain, or that are suitable to work with imbalanced datasets. Following is a brief description of each classifier and the evaluation measurements we use to compare their performance.

5.1 Classification methods

5.1.1 Naive Bayes

Among the simplest and most primitive classification algorithms, this probabilistic method is based on the Bayes Theorem [7] and strong underlying independence assumptions. That is, each feature is assumed to contribute independently to the class outcome. In predicting student attrition, naive Bayes classifiers have been used by [36, 50, 33]. Notably, the best results reported in [50] were achieved via this method.

5.1.2 C4.5 Decision trees

Another very popular classification method, C4.5 decision trees [39] have been used to predict student retention multiple times in recent literature (e.g., [46, 33, 25, 28]). This method works by building a tree structure where split operations are performed on each node based on information gain values for each feature of the dataset and the respective class. At each level, the attribute with highest information gain is chosen as the basis for the split criterion.

5.1.3 Logistic regression

Logistic regression is often used as a classification method wherein a sigmoid function is estimated based on the training data, and used to partition the input space into two class specific regions. Given this division, new instances can be easily verified to belong to one of the two classes. This approach has been used to predict student retention in [30, 19, 25, 27, 49, 21, 44], and it often achieved highly accurate results.

5.1.4 Hellinger distance decision trees

When applying learning algorithms to imbalanced datasets, one often needs to supplement the process with some form of data sampling technique. Hellinger distance decision trees [12] were proposed as a simpler alternative to that. This method uses Hellinger distance as the splitting criterion for the tree, which has several advantages over traditional metrics such as gain ratio in the context of imbalanced data. To the best of our knowledge, this method has not yet been used to predict student retention, but given that our dataset is highly imbalanced, we chose to investigate its performance.

5.1.5 Random forests

Random forests [8] combine multiple tree predictors in an ensemble. New instances being classified are pushed down the trees, and each tree reports a classification. The "forest" then decides which label to assign to this new instance based on the aggregate number of votes given by the set of trees. Recent work by Mendez et al.[32] used this method to predict science and engineering student persistence.

5.2 Evaluation measures

In order to compare the results obtained by each of the classifiers as well as the four different data "subsets", we utilize a variety of measures. A very popular standard used to evaluate classifiers is the predictive accuracy. Note, however, that utilizing this metric to evaluate classification that is based on imbalanced datasets can be extremely misleading.

To illustrate this issue, suppose that upon being given our dataset, an imaginary classifier predicts that all students will be retained in the engineering program following their first semester enrolled on campus. This will result in a remarkable 88.5% accuracy (recall that only 11.5 % of the students in this dataset dropped out). It is obvious, however, that such a classifier should not be awarded any merit since it fails to identify all students that should have been labeled as being *at risk*.

Instead, it is more appropriate to analyze the prediction accuracy for each individual class, or to use ROC curves to summarize the classifier performance. These and other measures can be calculated using confusion matrices (see Figure 3).

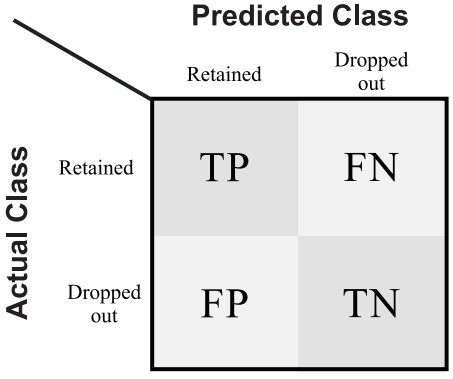

Figure 3: Confusion matrix for our experiments

Given the binary nature of this specific classification problem, the corresponding confusion matrix reports four values: True Positives (TP) – the number of retained students correctly classified, True Negatives (TN) – the number of dropout students accurately classified as such, False Positives

(FP) – The number of drop-out students mistakenly classified as *retained*, and False Negatives (FN) – retained students that were wrongfully predicted as drop-outs. Based on these labels, the individual accuracies for the negative (drop-out) and positive[2] (retained) classes, as well as the classifier's recall rates can be obtained as follows:

$$accuracy^{-} = \frac{TN}{TN + FN}$$
$$accuracy^{+} = \frac{TP}{TP + FP}$$
$$recall = \frac{TP}{TP + FN}$$

As previously mentioned, ROC curves are frequently used to summarize the performance of classifiers on imbalanced datasets. On an ROC curve, the X-axis represents the FP rate $FP/(TN + FP)$, and the Y-axis denotes the TP rate given by $TP/(TP + FN)$ at various threshold settings. The area under the ROC curve, AUROC, is also an useful metric for comparing different classifiers. The values for AUROC range from a low of 0 to a high of 1, which would represent an optimal classifier as highlighted in Figure 4.

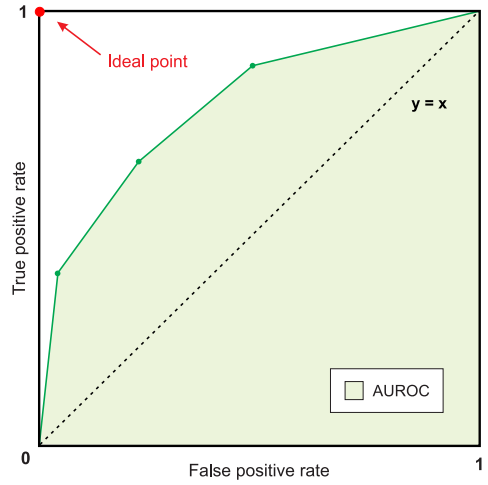

Figure 4: ROC curve illustration

Note that one shortcoming of ROC curves is that they do not explicitly show any dependence with respect to the ratio of positive and negative class instances in the dataset. A similar curve that uses precision and recall rates (Precision-Recall curve) can mitigate that issue [13].

6. EXPERIMENTAL RESULTS

To estimate how well the models generalize to future datasets, we utilized a 10-fold cross validation technique. This consists on splitting the original n data instances into 10 complementary subsets of size $n/10$, each of which preserving the original ratio of minority and majority class instances. The classifier is then given 9 of the subsets for training, and validation is performed using the remaining portion of the data. This process is repeated for 10 rounds using different partitions at each time, and an overall average of the results across each iteration is computed.

The performance of each of the five classification methods described in section 5.1 was evaluated as they were used to perform prediction on each of the four available datasets. Table 4 displays the results of each individual experiment in terms of the prediction accuracy for the negative class instances (i.e., the ratio of dropout students that were correctly labeled as *dropouts*), the prediction accuracy for the positive class instances (i.e., the ratio of retained students correctly classified as *retained*), and the overall weighted average across these two accuracy measurements. The highest accuracies achieved for each of the datasets are highlighted in bold, while the three highest overall are underlined.

Dataset	Classifier	Acc-	Acc+	Acc
all-academic	NB	**0.275**	0.902	0.842
	DT	0.083	0.930	0.833
	LR	0.104	0.900	0.803
	HT	0.083	0.884	0.792
	RF	0.000	**0.987**	**0.874**
top-academic	NB	0.167	0.954	0.864
	DT	0.042	0.949	0.845
	LR	0.000	**0.981**	**0.869**
	HT	**0.250**	0.881	0.809
	RF	0.104	0.892	0.802
all-engagement	NB	**0.833**	0.879	0.874
	DT	0.771	0.970	0.947
	LR	0.771	**0.978**	**0.955**
	HT	0.771	0.962	0.940
	RF	0.771	0.970	0.947
top-academic+ engagement	NB	**0.875**	0.892	0.890
	DT	0.792	0.962	0.945
	LR	0.750	**0.973**	**0.947**
	HT	0.771	0.965	0.943
	RF	0.750	0.965	0.940

Table 4: Prediction accuracy achieved using each of the datasets

Before analyzing these results more deeply, it is essential to consider the degree of importance that should be assigned to each of these metrics. Given our binary classification problem, two types of error could emerge. Students that ultimately remain in the program for the spring semester could be misclassified as dropouts (false negatives), and actual dropout students could be mistakenly labeled as retained (false positives). While some previous work (e.g., [15]) considered the first type of error to be more serious, we argue that the opposite is true. If these techniques are to be used in the development of an effective early warning system, failing to identify students that are at risk of dropping out can be much more costly than incorrectly labeling

[2]The prediction accuracy for the positive class can also be labeled *precision*

someone as a dropout.

In Table 4 we can see that predictions based only on academic performance and demographic data achieve a maximum *acc-* of 27.5% when the *all-academic* dataset is paired with a naive Bayes model. That corresponds to only 11 of the 48 dropout students being correctly identified. Conversely, when engagement features are utilized, that accuracy improves very noticeably to 83.3% and 87.5%, both also achieved with the previously mentioned classifier.

The naive Bayes model using the *top-academic+engagement* dataset remarkably identifies 42 of the 48 dropout students. The vast majority of those retained (331 out of 419) are also correctly classified. Note that the other four classifiers obtain higher *acc+* values under the same setup, and could potentially be the preferred choice depending on the circumstances.

With respect to *acc-*, the naive Bayes classifier outperformed the others for all but one dataset. We used its experimental results in Figure 5 to illustrate the ROC and Precision-Recall curves for each dataset. In our particular context, it seems apparent that the ePortfolio engagement features are very good predictors for student retention. The highest AUROC value (0.929) was obtained by the *top-academic+engagement* dataset, while *all-academic* performed worse with an AUROC of 0.654.

7. CONCLUSION

In this paper we investigated the feasibility of using electronic portfolio data as a means for predicting college retention. We showed that while datasets that do not contain features describing student academic engagement can often yield reasonable results, providing such features to the classification models greatly increases their ability to identify students that may ultimately drop out. Our experiments showed significant gains in accuracy when engagement features were utilized, and we believe this can be used to build early warning systems that would be able to identify at-risk students at very early stages of their academic life, giving educators the opportunity to intervene in a more timely and effective fashion.

Acknowledgments

We thank Jeffrey Yan and Peter Lefferts at Digication for the help they provided with the collection of the electronic portfolio datasets, and Catherine Pieronek for compiling the academic performance data.

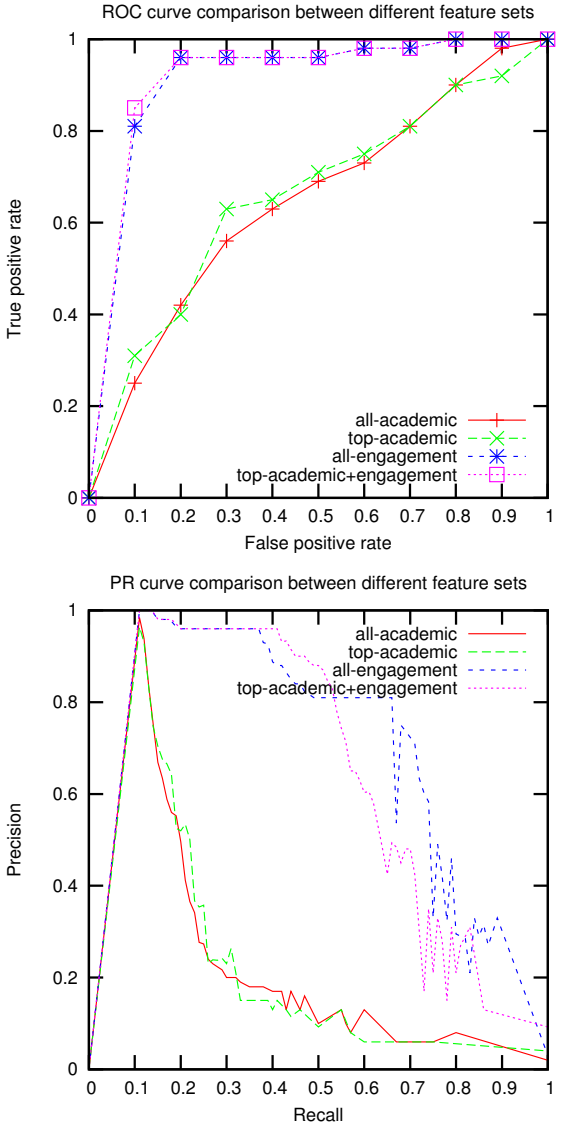

Figure 5: Naive Bayes ROC and Precision-Recall curves

8. REFERENCES

[1] ABET: Accreditation board for engineering and technology. `http://www.abet.org/`. Accessed: 2013-10-16.

[2] Digication e-Portfolios. `http://www.digication.com/`. Accessed: 2013-09-30.

[3] M. Al-Atabi, A. S. Mahdi, O. Younis, and E. Chung. An integrated portfolio and advising system for undergraduate engineering students. *Journal of Engineering Science and Technology*, 6(5):532–541, 2011.

[4] R. Alkhasawneh. *Developing a Hybrid Model to Predict Student First Year Retention and Academic Success in STEM Disciplines Using Neural Networks*. PhD thesis, Virginia Commonwealth University, July 2011.

[5] A. W. Astin and H. S. Astin. Undergraduate science education: The impact of different college environments on the educational pipeline in the sciences. final report. 1992.

[6] H. C. Barrett. Researching electronic portfolios and learner engagement: The reflect initiative. *Journal of Adolescent & Adult Literacy*, 50(6):436–449, 2007.

[7] M. Bayes and M. Price. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M.A. and F.R.S. *Philosophical Transactions (1683-1775)*, pages 370–418, 1763.

[8] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[9] J. Burtner. The use of discriminant analysis to investigate the influence of non-cognitive factors on engineering school persistence. *Journal of Engineering*

Education, 94(3):335–338, 2005.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[11] H. L. Chen and T. C. Black. Using e-portfolios to support an undergraduate learning career: An experiment with academic advising. *Educause Quarterly*, 33(4), 2010.

[12] D. A. Cieslak, T. R. Hoens, N. V. Chawla, and W. P. Kegelmeyer. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, 24(1):136–158, 2012.

[13] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[14] M. S. DeBerard, G. Spielmans, and D. Julka. Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College Student Journal*, 38(1):66–80, 2004.

[15] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers. Predicting students drop out: A case study. In *International Conference on Educational Data Mining*, pages 41–50, 2009.

[16] D. Delen. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4):498–506, 2010.

[17] D. Delen. Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory and Practice*, 13(1):17–35, 2011.

[18] B. Eynon. Making connections: the laguardia eportfolio. *Electronic portfolios*, 2(0):59–68, 2009.

[19] D. S. Fike and R. Fike. Predictors of first-year student retention in the community college. *Community College Review*, 36(2):68–88, 2008.

[20] L. A. Flowers. Effects of living on campus on african american students' educational gains in college. *NASPA Journal*, 41(2), 2004.

[21] S. Herzog. Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research*, 2006(131):17–33, 2006.

[22] L. F. Johnson, A. Levine, R. Smith, and S. Stone. The horizon report: 2010 edition. 2010.

[23] W. E. Knight, M. D. Hakel, and M. Gromko. The relationship between electronic portfolio participation and student success. *Association for Institutional Research*, 2008.

[24] G. D. Kuh, S. Hu, and N. Vesper. "They shall be known by what they do": An activities-based typology of college students. *Journal of College Student Development*, 41(2):228–44, 2000.

[25] E. J. Lauría, J. D. Baron, M. Devireddy, V. Sundararaju, and S. M. Jayaprakash. Mining academic data to improve college student retention: An open source perspective. In *International Conference on Learning Analytics and Knowledge*, pages 139–142. ACM, 2012.

[26] Q. Li, H. Swaminathan, and J. Tang. Development of a classification system for engineering student characteristics affecting college enrollment and retention. *Journal of Engineering Education*, 98(4):361–376, 2009.

[27] J. J. Lin, P. Imbrie, and K. J. Reid. Student retention modelling: An evaluation of different methods and their impact on prediction results. In *Research in Engineering Education Sysmposium*, pages 1–6, 2009.

[28] S.-H. Lin. Data mining for student retention management. *Journal of Computing Sciences in Colleges*, 27(4):92–99, 2012.

[29] O. Lopez-Fernandez and J. L. Rodriguez-Illera. Investigating university students' adaptation to a digital learner course portfolio. *Computers & Education*, 52(3):608–616, 2009.

[30] J. Luna. Predicting student retention and academic success at New Mexico Tech. Master's thesis, New Mexico Institute of Mining and Technology, 2000.

[31] J. MacGregor and B. Leigh Smith. Where are learning communities now? National leaders take stock. *About Campus*, 10(2):2–8, 2005.

[32] G. Mendez, T. D. Buskirk, S. Lohr, and S. Haag. Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. *Journal of Engineering Education*, 97(1):57–70, 2008.

[33] A. Nandeshwar, T. Menzies, and A. Nelson. Learning patterns of university student retention. *Expert Systems with Applications*, 38(12):14984–14996, 2011.

[34] E. T. Pascarella. How college affects students: Ten directions for future research. *Journal of College Student Development*, 47(5):508–520, 2006.

[35] E. T. Pascarella, P. T. Terenzini, and G. S. Blimling. The impact of residential life on students. *Realizing the educational potential of residence halls*, pages 22–52, 1994.

[36] K. Pittman. *Comparison of data mining techniques used to predict student retention*. ProQuest, 2011.

[37] M. Price. Purpose, audience, and engagement in spelman college's efolio project. *Handbook of Research on EPortfolios*, page 259, 2006.

[38] J. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[39] J. R. Quinlan. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann, 1993.

[40] G. Ring, B. Weaver, J. J. Jones Jr, et al. Electronic portfolios: engaged students create multimedia-rich artifacts. *Journal of the research center for educational technology*, 4(2):103–114, 2009.

[41] A. Seidman. *College student retention: Formula for student success*. Greenwood Publishing Group, 2005.

[42] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, 2013.

[43] V. Tinto. *Leaving college: Rethinking the causes and cures of student attrition*. University of Chicago Press, 2012.

[44] C. P. Veenstra, E. L. Dey, and G. D. Herrin. A model for freshman engineering retention. *Advances in Engineering Education*, 1(3):1–31, 2009.

[45] M. Xenos, C. Pierrakeas, and P. Pintelas. A survey on

student dropout rates and dropout causes concerning the students in the course of informatics of the hellenic open university. *Computers & Education*, 39(4):361–377, 2002.

[46] S. K. Yadav, B. Bharadwaj, and S. Pal. Data mining applications: A comparative study for predicting student's performance. *arXiv preprint arXiv:1202.4815*, 2012.

[47] K. B. Yancey. Reflection and electronic portfolios. *BL Cambridge, S. Kahn, DP Tompkins, & KB Yancey Electronic portfolios*, 2:5–16, 2009.

[48] C. H. Yu, S. A. DiGangi, A. Jannasch-Pennell, W. Lo, and C. Kaprolet. A data-mining approach to differentiate predictors of retention. *Online Submission*, 2007.

[49] G. Zhang, T. J. Anderson, M. W. Ohland, and B. R. Thorndyke. Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. *Journal of Engineering Education*, 93(4):313–320, 2004.

[50] Y. Zhang, S. Oussena, T. Clark, and K. Hyensook. Using data mining to improve student retention in higher education: a case study. In *International Conerence on Enterprise Information Systems*, June 2010.