IDENTIFYING STUDENTS AT RISK 
AND BEYOND: A MACHINE LEARNING APPROACH

A Dissertation

Submitted to the Graduate School 
of the University of Notre Dame 
in Partial Fulfillment of the Requirements 
for the Degree of 

Doctor of Philosophy

by

Everaldo Aguiar

______________________________
Nitesh V. Chawla, Co-Director

______________________________
Jay B. Brockman, Co-Director

Graduate Program in Computer Science and Engineering
Notre Dame, Indiana

July 2015
IDENTIFYING STUDENTS AT RISK
AND BEYOND: A MACHINE LEARNING APPROACH

Abstract
by
Everaldo Aguiar

A common denominator across different cultures and societies is that people are on a continuous quest to move forward in life – each in their own personal way. Though success can be achieved by many routes, it is also a well-known fact that a person’s level of education can be a strong indicator of their achievements, professional or otherwise. Yet, persevering through the educational ladder can still be a major challenge for some students, regardless of the benefits it confers. As the first step in providing each student with the attention and support they need, a perennial challenge for educators has been to identify those students who may ultimately struggle. Unfortunately, the current systems used to identify struggling students are often inadequate, precipitating dropout rates that that are a focal concern for both secondary and post-secondary institutions alike. To stem the tide of student attrition, these institutions continue to invest increasing amounts of resources to foster the early identification of students who may drop out. In this work, we explore the application of various machine learning techniques to the problem of predicting which students may be at risk of some adverse outcome. We compare the performance of such methods to that of more traditional approaches, discuss what specific student-level data features can be used to measure student risk, and delineate strategies that can be followed by academic institutions that wish to deploy and evaluate these predictive models.
DEDICATION

To my parents, Veralúcia and Everaldo Aguiar
FIGURES

3.1 Support Vector Machine transformation illustrated. .......................... 15
3.2 Confusion matrix for our post-secondary education analysis. ............... 18
3.3 Illustration of an ROC curve, highlighting the ideal point, AUROC concept, and the random predictor baseline. ................................. 19
3.4 Precision at top 10% metric explained for a cohort of 100 students. Only the 10 students with highest risk scores are shown. Those labeled in red did, in fact, not graduate on time. Conversely, the green shades represent students that graduated on time and who should likely not have been placed at the "at-risk" category. Model 2 in this example outputs slightly more reliable risk scores. ................................. 21

4.1 Flow of student enrollment at school district A over time. The larger (blue) portion highlights students that are currently enrolled. The green subset indicates students that are flowing into the system at each time stamp, whereas the red group represents students that are leaving the school district. ................................. 24
4.2 Student mobility values vs. on-time graduation rates. Mobility is the number of times a student unexpectedly switched schools. The shaded region corresponds to 95% CIs. ................................. 28
4.3 Student GPAs (binned) vs. on-time graduation rates. The shaded region corresponds to 95% CIs. ................................. 28
4.4 Precision at top 10% for models evaluated at the end of each grade level. 30
4.5 Student mobility vs. high-risk rates. Mobility is the number of times a student unexpectedly switched schools. The shaded region corresponds to 95% CIs. ................................. 31
4.6 Student GPAs (binned) vs. high-risk rates. The shaded region corresponds to 95% CIs. ................................. 31
4.7 Risk category assignments over time. Following a single student from 6th to 11th grade, this figure illustrates how often our models place them at different risk categories. ................................. 32
4.8 Correlation between risk scores & time to off-track. ........................ 40
4.9 Accuracy of predicting time to off-track. ................................. 40
4.10 Mean Absolute Error predicting time to off-track. ........................ 40
4.11 Student Report Card Dashboard depicting risk score, GPA, and absence rate trajectories.

5.1 ROC Curves for the task of predicting on-time graduation at both districts A and B. Models were trained using an earlier cohort and tested on the cohort for the following year.

5.2 Empirical Risk Curves. The ranking quality of an algorithm is good if this curve is monotonically non-decreasing.

5.3 Precision across instances belonging to the top $k$ highest risk score group (for varying values of $k$).

5.4 Recall across instances belonging to the top $k$ highest risk score group (for varying values of $k$).

5.5 Feature ranking by various algorithms. GPA.08 is ranked consistently high in both the districts by most of the models.

5.6 Jaccard Similarity of students at-risk for various algorithms.

5.7 Precision by grade at top K.

5.8 Identification before Off-track.

6.1 Screenshots of electronic portfolios illustrating the three main sections: engineering advising, project updates, and engineering exploration.

6.2 The effect of ePortfolio engagement on first semester retention.

6.3 First-year engineering student academic performance breakdown by retention outcome.

6.4 Average breakdown of ePortfolio engagement activity by cohort of first-year engineering students.

7.1 While dropout rates are often more prominent across low-performing students, and students with high performance tend to be retained, a couple of special cases exist and are highlighted here. Low performing/highly engaged students (quadrant I) are often retained. High performing/disengaged students (quadrant IV) may drop out.

7.2 Naive Bayes ROC and Precision-Recall curves for each subset of academic performance and engagement features.

8.1 ePortfolio data exploration and reporting dashboard. This figure shows an R Shiny app that we developed for a medium-sized university in the Mid-Atlantic region. A variety of metrics collected to describe student activity can be highlighted using this interactive tool.

8.2 Web-based classification tool. This figure illustrates the multi-step pipeline component wherein users can select their data, pre-processing tasks, models and parameters.
8.3 Web-based classification tool. This figure illustrates the interactive ROC and PR curves that are embedded into the analysis report generated for each experiment. ............................................ 112

8.4 Web-based classification tool. This figure illustrates the second half of the report page where the user has access to a confusion matrix, various performance metrics, as well as the detailed prediction output for each instance analyzed. ............................................ 113
TABLES

4.1 DESCRIPTION OF DATASET STUDENT-LEVEL FEATURES OBTAINED FROM DISTRICT A ................................. 25
4.2 FEATURE RANKINGS ILLUSTRATING THE DEGREE OF IMPORTANCE ASSIGNED TO EACH INDIVIDUAL STUDENT-LEVEL FEATURE BY EACH METHOD TESTED ............................... 27
4.3 PERCENTAGE OF STUDENTS MOVING ACROSS DIFFERENT RISK GROUPS OVER TIME .............................. 33
5.1 LIST OF STUDENT ATTRIBUTES AND THEIR AVAILABILITY IN DISTRICTS A AND B ........................................ 50
5.2 TOP 5 FEATURES IN DISTRICTS A AND B .................... 61
5.3 CLASSIFIER MISTAKE PATTERNS ........................................ 65
6.1 P-VALUES OF OBSERVED DISTRIBUTIONS FOR THREE EPORTFOLIO FEATURES WITH RESPECT TO RETENTION OUTCOMES ................................. 82
6.2 RUBRIC FOR STUDENT INTEREST IN ENGINEERING BASED ON ENGINEERING EXPLORATION REFLECTIONS .................. 83
6.3 RUBRIC FOR STUDENT INTEREST IN ENGINEERING BASED ON ENGINEERING EXPLORATION REFLECTIONS ............... 85
6.4 RATINGS OF STUDENT ENGAGEMENT BASED ON INSTRUCTOR EVALUATION ................................. 86
7.1 DATASET FEATURES DESCRIBED ........................................ 93
7.2 ACADEMIC PERFORMANCE AND EPORTFOLIO-BASED ENGAGEMENT DATASET FEATURES CATEGORIZED ................................. 95
7.3 FEATURE RANKINGS PROVIDED BY MULTIPLE METHODS %. 96
7.4 PREDICTION ACCURACY ACHIEVED USING EACH OF THE DATASETS ........................................ 100
7.5 CLASSIFICATION MODELS’ AUROC WITH DATA OVERSAMPLING ........................................ 102
7.6 PRECISION ON THE TOP 5 AND 10% ........................................ 105
ACKNOWLEDGMENTS

An almost countless number of people have been pivotal to the realization of this work. First, I thank my advisors Nitesh Chawla and Jay Brockman, for providing me the opportunity to research problems that were not only challenging from an academic perspective, but that also had the potential to steer meaningful initiatives in the education space. I am also immensely grateful for the insights provided by my committee members Kevin Bowyer, Sidney D’Mello and Alex Ambrose, who has been a key driver of our learning analytics initiative at Notre Dame since day one.

My gratitude is also extended to the National Science Foundation and the Eric and Wendy Schmidt Data Science for Social Good Fellowship, for making all of this possible. My time as a DSSG fellow allowed me to see interesting ways through which data science can be used to tackle some of the challenges we face as a society. For that, I am forever indebted to the organizers, mentors and fellows.

The University of Notre Dame has shown me the importance and effectiveness of interdisciplinary collaborative work, an aspect that is perfectly illustrated by the DIAL research group. I would like to thank all of my friends and colleagues at iCeNSA (extra points to those at Nieuwland 384L) for all of the fruitful discussions over the years, work-related or otherwise.

It goes without saying that I would not have reached this milestone without the unconditional support of my parents Veralúcia and Everaldo, my sisters Iaryn and Nayara, and my partner Nali. Your encouragement and care are the most valuable awards I have ever been given.
CHAPTER 1

INTRODUCTION

Student attrition is perhaps as old as the art of teaching itself. Everyone learns differently and in different ways, and some students – for any variety of reasons – need a more personalized system of support than others in order for them to succeed. This deeply intertwined relationship between the challenges students face, the solutions provided by academic institutions, and the observed outcomes that result, is an important piece of the current education establishment. Failure to identify students who are in need of individual attention early can reduce the number of viable intervention options and, in the grand scheme of things, may ultimately result in high dropout (and late graduation) rates – a bellwether for a number of undesired consequences.

As a result of various initiatives put forward by school districts and policy makers across the country, high schools in the United States have been observing a steady decrease in dropout rates for the past few years [60]. Nonetheless, the overall figures are still alarming. More than half a million students fail to graduate on time each year, with even more pronounced figures seen across students of African, Hispanic or Native American decent [60]. In addition to severely limiting their career choices, students who do not obtain a high school diploma experience a range of negative life outcomes such as increased prevalences of incarceration and poverty [20, 70].

In an attempt to reduce the number of students who go through this grim scenario, a considerable amount of research effort has been directed at understanding the factors that contribute to a student’s decision to drop out (e.g., [9, 12, 19, 22]).
As these factors can be related to a variety of aspects such as student behavior, coursework, performance, and others, more and more schools are beginning to utilize student-level data to identify early signs of possible undesired outcomes.

While still in their infancy, early warning indicator (EWI) systems that harness student-level data to provide educators with early flags for students who may be “at risk” are becoming increasingly popular [39]. Also noteworthy is the fact that these EWIs vary vastly in how they combine and interpret each of the contributing factors that are used to measure student risk [28]. Yet, as these approaches become more refined, and as districts and governments develop more robust strategies for operationalizing them [63], the overall benefits of such systems are becoming progressively evident.

In the higher education space, students and educators often face strikingly similar predicaments. While colleges and universities are certainly more likely to have well-established academic support services, these are typically static initiatives and their capacity to proactively identify students at risk is limited. Despite being present across all disciplines, the issues of attrition and retention can be most noticeably seen in the areas of Science, Technology, Engineering, and Mathematics (STEM) [34, 35, 45, 64, 80, 97].

With that in mind, some post-secondary institutions have begun to embed EWI systems into their traditional student management applications. For instance, Marist College is currently evaluating a predictive model that mines student performance, learning management system event-log, and electronic gradebook data to generate flags that highlight students that may be at risk of dropping out [58]. A similar program developed by Purdue University, known as Course Signals, leverages course-level data to predict which students may fail to succeed in a particular class [16].

Nevertheless, the vast majority of the current efforts rely only on subsets of historical data that describe student academic performance and other traditional
facets such as demographics and financial aid. It is well known, however, that student success in college can be tied to a variety of other factors outside that scope \cite{18, 48, 71, 79, 92, 93, 117}. In this dissertation we explore the idea of quantifying student engagement and using it to identify students that may be at risk of leaving a STEM discipline, even when they are not displaying poor academic performance.

1.1 Contributions

While there is some overlap in the approaches we utilized when working in the secondary and post-secondary spaces, for the remainder of this dissertation, we present these works mostly in isolation. This deliberate choice should allow for better readability, and it should also emphasize the efforts and challenges that were unique to each of the two focuses of our work.

To build more robust and comprehensive early warning systems, we partnered with two large U.S. school districts with a combined enrollment of approximately of 200,000 students. As a result of this partnership and with respect to our work in the secondary education space, our primary contributions can be categorized as the following:

- We developed and evaluated machine learning-based models that can detect students at risk of not graduating from high school on time earlier and more accurately than the rule-based methods that were in routine use.

- Taking into account the limited amount of resources available to school districts, we extended the functionality of our models to allow for a careful prioritization of the students deemed to be at risk. This extra layer of information allows educators to assign an urgency score to each of these students, which in turn can be used to guide the order in which interventions are provided.

- We carefully evaluate the efficacy of these models using a set of metrics that is informative to educators and school administrators.

- To thoroughly estimate the performance of our models in a real-world scenario, we test them on new, disjoint, and later sets of students, demonstrating that they are viable for deployment.
Our effort in the higher education space was focused on better understanding the patterns of students who leave the STEM field following their first or second semester of enrollment. Further, we created models that could help in the early identification of students who may be at risk of dropping out, or who may be considering other disciplines outside STEM. Through a partnership with the College of Engineering at the University of Notre Dame, we analyzed student-level data for incoming students.

Some of our key findings and contributions were:

- We replicated the traditional early warning approach that utilizes academic performance data to predict students at risk, and showed that the performance of models that focus solely on this aspect is relatively poor.

- To improve upon our ability to identify these students early on, we proposed the use of data extracted from electronic portfolios as a proxy for student engagement, and through a careful feature analysis exercise, we showed that this category of features was strongly informative of the outcomes we were interested in anticipating.

- We used classification models trained on a combination of academic performance, student engagement, and demographics data to show that exploring these three facets ultimately resulted in early predictions that were far more accurate and actionable than those of models that only consider performance.

1.2 Organization

The content presented in this dissertation is organized as follows. Chapter 2 provides a comprehensive review of the current literature on the topic of early warning systems, and the most recent advances in that area as it pertains to both the secondary and post-secondary education spaces. Following, in Chapter 3 we provide a brief description of the techniques that were used in the development of our models, as well as the evaluation metrics we adopted. Chapter 4 details the work done in collaboration with a large school district in the development and evaluation of a system-wide, machine learning-based early warning system, which, in Chapter 5, we show can be easily extended to other cohorts and districts. In Chapter 6 we introduce
the reader to electronic portfolios and their integration with the engineering curriculum at the University of Notre Dame. Later, in Chapter 7, we elaborate on our use of student engagement and performance data as a means of predicting engineering student persistence. That is followed by a general discussion on some of the current and previous extensions and collaborations of this work in Chapter 8. Finally, in Chapter 9 we present some concluding remarks.
CHAPTER 2

RELATED WORKS

Over the years, a number of studies have been carried out with the goal of understanding the root causes behind student dropouts. More recently, a variety of efforts have attempted to anticipate that outcome so as to provide educators with a window of opportunity to implement academic interventions. Below we describe some of these studies and highlight their relevance to the work presented in this dissertation.

2.1 Secondary Education

Decades of research have been dedicated to investigating the various causes and predictors of on-time high school graduation [9, 105]. This comprehensive body of knowledge has unveiled a number of factors that can influence a student’s decision to drop out of high school. Among these precursors, Alexander et al. [9] highlights the importance of first-grade experiences and family circumstances. Allensworth et al. [11-13] provide a detailed analysis of influencing factors for student dropouts in the Chicago Public School system, where nearly half of the student population fails to graduate. In that study, some of the highlighted factors include chronic absence, course failures in middle school, and failure during the first high school year. Dalton et al. [38] suggests that, in addition to socio-economical factors, some students may leave high school simply because they feel that it would be easier and more convenient to obtain a GED. Other studies focus on school-centric and external employment factors that could be related to these undesired outcomes [36, 69, 108].
Many scholars, especially in the most recent years, have proposed that school districts should take advantage of this research knowledge by implementing EWI systems. Bowers et al. [28] systematically reviewed this vast literature on predicting high school graduation by quantitatively synthesizing the predictive accuracy of 110 early warning flags presented across 36 studies. These studies found that many factors can indeed serve as effective EWIs, including those routinely recorded by schools such as having high absence rates or low grades.

Further, Bowers et al.’s review [28] found that these various initiatives used EWIs in a variety of ways. For instance, some studies created composite at-risk indicators based on the intersection of EWIs (e.g., having BOTH high absence rates and low grades), while others used the union of EWIs (e.g., having high absence rates AND/OR low grades). Consequently, these studies widely vary in performance. Rather than using the simple intersection or union of EWIs, we used machine learning methods such as random forests [29] to optimally combine these EWIs when forming composite indicators.

It is important to highlight that although these EWIs may accurately predict high school drop-out, they do not always reveal the root causes of students’ needs [27, 49, 50, 91, 105]. For example, while low grades can be highly predictive of drop-out, students may earn low grades for a variety of academic and non-academic reasons (e.g., lack of mathematical proficiency, conflicts at home or with peers). Nevertheless, these indicators could help teachers and administrators form initial hypotheses about the needs of particular students. To illustrate that aspect, a couple of studies [12, 26] showed that students whose grades sharply declined between middle and high school, were also struggling to adapt to their new social and academic environment. We therefore prototyped a data-driven dashboard that could help teachers and administrators form these insights when reviewing student records. In turn, these insights could help match at-risk students with appropriate intervention programs.
Prior implementation guides on EWIs [112] have detailed how similar integrated analytic approaches can lead to a comprehensive set of actions for addressing particular students needs.

2.1.1 District-level Work

In 2013, one of our partner school districts investigated linkages between grades 1, 3, 6, 9, and final graduation status for two cohorts of students [85, 115]. The results of the study showed that the characteristics of those who drop out of school fell under three categories: attendance, behavior, and coursework (ABC’s). Essentially, with respect to these, students who drop out of school were determined to exhibit a pattern of behaviors that were generally identifiable in advance of them dropping out of school completely.

Following the findings of West [115], the school district began to develop an EWI system for students across all grade levels (elementary through high school). However, rather than focusing only on dropouts, the district focuses on the identification of students who stay in school, but need extra support. For example, academic performance and attendance have been significantly correlated to the amount of time until achieving a bachelor’s degree for one cohort of students [123]. Furthermore, the progressive EWI system goes beyond the ABC’s and includes other factors (e.g. mobility; new to the district) reported in the literature to also impact student performance [75, 101].

2.2 Post-secondary Education

From a sociological standpoint, student attrition in higher education has been studied in great detail. Seidman [107] and Tinto [113] 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 dissertation. Following, we
provide a more elaborate description of the most recent works that utilize student
data to create and evaluate predictive models for student attrition.

Early work by DeBerard et al. [41] 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. [121] 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. [84].

A key premise of our work is highlighted by Burtner [30]. 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. [73] validate that idea by showing that the accuracy of their classification models improves after the inclusion of non-cognitive features extracted from a survey. Also similar to our initiative, work by Manhães et al. [81][83] focused on STEM students and investigated how a wide range of student-level data features could be used to predict student attrition.

Yu et al. [120] 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 [43][44] 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. [122] tested three different classifiers and observed that the best prediction accuracy for student retention was yield by naive Bayes. Alkhasawneh [10] 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. [111], 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.
3.1 Methods

Below we provide brief descriptions for each of the methods that are utilized throughout this dissertation for feature selection (ranking), classification, and survival analysis.

3.1.1 Feature Selection

To enhance a predictive model’s ability to generalize to new sets of data, as well as its efficiency and interpretability, reducing the size of the training feature set is often desirable. A variety of techniques to rank and select relevant feature subsets exist \[52\], and below we describe those that were used in this work.

**Information Gain (IG).** This feature selection method \[99\] measures the amount of information about the class being predicted, when only a single feature and the corresponding class distribution are available. More specifically, it quantifies the expected reduction in entropy (uncertainty associated with a random feature). For example, the information gain for \( Y \) (class labels) conditioned on a specific feature \( x_k \) is given by:

\[
IG(Y|x_k) = H(Y) - H(Y|x_k)
\]
Where $H(Y|x_k)$ denotes the conditional entropy of that same $Y$ with respect to the $m$ feature values taken for $x_k$:

$$H(Y|x_k) = \sum_{j=1}^{m} P(x_k = v_j)H(Y|x_k = v_j)$$

**Gain Ratio (GR).** A known drawback of information gain is that it places higher importance to features that are made up of many values. To address that issue, gain ratio [100] performs a similar measurement while attempting to reduce that bias by taking into account the number and size of the subsets that would be created when using each feature to partition the dataset. It does so by normalizing information gain by the “intrinsic information” of a split, which is defined as the information needed to determine the branch to which an instance belongs.

**Gini Index (GI).** This impurity measure can be used to quantify how often a randomly chosen instance would be incorrectly classified if this classification were to be performed at random according to the distribution of class labels. It can be computed as the sum of the probabilities of each instance being picked, times the probability of mislabeling that instance. The Gini Index for a set of items $i \in \{1, 2, \ldots, m\}$ can be computed as:

$$GI(f) = 1 - \sum_{i=1}^{m} f_i^2$$

Where $f$ denotes the fraction of instances belonging to each class in $i$.

**Chi-Squared (CS).** The chi-squared ($\chi^2$) method measures how much deviation is seen in the observed data when compared to the case where the class and feature
values are independent of each other. It evaluates whether the feature and class co-occurrences are statistically significantly related.

**Pearson Correlation (CR).** This simple metric can be utilized to measure the linear correlation between two variables. It can assume values in the range [-1,1], with the 0 midpoint representing a scenario with no correlation. In some of our experiments, we used this method to rank features based on how strongly correlated they are to the given outcome.

### 3.1.2 Classification

The supervised task of classification consists of the automated process of deriving a descriptive function $f : X \to Y$ through the use of labeled (training) data. While there are numerous approaches to this process, below we cover those that were used in this and other related works.

**Naive Bayes (NB).** Among the simplest and most primitive classification algorithms, this probabilistic method is based on the Bayes Theorem 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 [5, 86, 96, 122]. Notably, the best results reported in [122] were achieved via this method.

**C4.5 Decision Trees (DT).** Another very popular classification method, C4.5 decision trees [100] have been used to predict student retention multiple times in recent literature (e.g., [5, 68, 74, 86, 118]). 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.

**Logistic Regression (LR).** 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 [5, 7, 47, 55, 67, 68, 73, 78, 114, 121], and it often achieved highly accurate results.

**Hellinger distance decision trees (HT).** 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 [37] 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.

**Random Forests (RF).** Random forests [29] 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. [84] used this method to predict science and engineering student persistence.

**Adaboost (AB).** This popular boosting algorithm [106] works by iteratively re-weighting instances based on how difficult they are to classify. Initially, each instance receives the same weight and, based on those values, bootstrap sampling is performed to select the training data. At each iteration $i$, a classifier $C_i$ is learned, evaluated,
and the instance weights are re-adjusted so as to allow the subsequent classifier $C_{i+1}$ to have a higher chance of selecting the previously misclassified instances during the training step. The final classifier $C^*$ combines the votes obtained from each learned classifier, while also taking into account each of their reliabilities.

**Support Vector Machine (SVM).** While this technique can be vastly extended and adapted to a variety of scenarios, SVMs \cite{5} are based on the concept of decision planes. These are defined as functions that can be used to separate a set of instances belonging to different classes into disjointed subgroups. Figure 3.1 illustrates the overall idea behind this technique. On the left-hand side the original instances are displayed along with the corresponding separating boundary that best isolates the different classes. These instances are rearranged using a set of functions (known as kernels) into the new representation shown on the right-hand side. This mapping creates a new feature space wherein these instances are now linearly separable, and that final setting is used in the classification of unlabeled instances.

Figure 3.1. Support Vector Machine transformation illustrated.
3.1.3 Survival Analysis

Survival analysis methods are designed to estimate the duration of time until a particular event takes place. In our specific context, these techniques are used to provide an expected value for how much time one should expect to go by before a student drops out (time to off-track). While there are multiple approaches to generating these estimates, we made use of cox regression models for doing so.

**Cox Regression.** Also known as proportional hazards regression [90], cox regression models allow for the concurrent analysis of multiple risk factors and their effect on a given outcome through the *hazard function*. This method computes a coefficient for each available feature, indicating the degree and direction of the effect that this feature values have on an estimated baseline survival curve. A value of zero indicates that this feature is not an outcome predictor at all, whereas positive values infer that larger values for this feature can be associated with a greater outcome probability. This set of coefficients form a modified survival curve that can be used to estimate the outcomes of new and unseen instances.

3.1.4 Ordinal Classification

We note that the previously mentioned metric, *time to off-track*, is an ordinal (rather than categorical) variable. This means that there is an inherent ordering on the values that time to off-track takes, and values that are closer to each other are more similar than those that lay further apart. For example, time to off-track values of 1 and 2 are more similar to one another than are 1 and 5. For this reason, classification frameworks that treat labels as categorical variables might not be the optimal choice in this context. Therefore, we also evaluate the use of ordinal classification methods which assume that the outcome labels are ordered, such as ordinal regression trees.
Ordinal Regression Trees. In their simplest form, ordinal regression tree approaches consist of adaptations made to the traditional algorithms so as to allow for the prediction of ordinal classes [65]. The most straightforward of these is to simply treat each ordinal (integer) value as a real number, and as part of post-processing, map each result back to a corresponding ordinal value through rounding. Similarly, Kramer et al. [65] also describe a variety of transformations that can be made to the feature values in a pre-processing step with the same goal of achieving ordinal output values.

3.2 Evaluation Metrics

In order to compare the results we obtained, a variety of traditional machine learning metrics were used. Further, we also provide evaluations that are based on metrics that illustrate the real-world context and constraints to which these models are subjected when deployed.

A very popular standard used to evaluate classifiers is the predictive accuracy (i.e., the overall percentage of instances that is correctly labeled). Note, however, that utilizing this metric to evaluate classification that is based on imbalanced datasets (as is the case in most of our scenarios) can be extremely misleading. To illustrate this issue, suppose that upon being given our post-secondary dataset, an imaginary classifier predicts that all students will re-enroll to the engineering program following their first semester. This will result in a remarkable 88.5% accuracy\(^1\). It is obvious, however, that such classifier should not be awarded any merit since it fails to identify all students that should have been labeled as being \textit{at risk}.

Instead, it is more appropriate to analyze the prediction accuracy for each individ-

\(^1\)Since only 11.5% of the students in that particular dataset dropped out
ual class, or to use ROC curves to summarize the classifier performance. These and other metrics can be calculated, or somewhat illustrated, using confusion matrices (see Figure 3.2).

![Confusion Matrix Diagram](image)

Figure 3.2. Confusion matrix for our post-secondary education analysis.

Given the binary nature of our classification problem, the corresponding confusion matrix reports four values: True Positives (TP) – the number of students who were not retained correctly classified, True Negatives (TN) – the number of retained students accurately classified as such, False Positives (FP) – The number of retained students mistakenly classified as not retained, and False Negatives (FN) – not retained students that were wrongfully predicted as retained. Based on these labels, the individual accuracies for the negative (retained) and positive (not retained) classes, as well as the classifier’s recall rates can be obtained as follows:

\[^{2}\text{The prediction accuracy for the positive class can also be labeled precision}\]
\[
\text{accuracy}^+ = \frac{TP}{TP + FP}
\]
\[
\text{accuracy}^- = \frac{TN}{TN + FN}
\]
\[
\text{recall} = \frac{TP}{TP + FN}
\]

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 prediction thresholds. The area under the ROC curve, AUROC, is another 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 3.3.

![Figure 3.3](image-url)

Figure 3.3. Illustration of an ROC curve, highlighting the ideal point, AUROC concept, and the random predictor baseline.
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 [40].

Since our ultimate goal is to devise models that can be readily converted into functional early warning systems, we further evaluated the performance of our models with respect to their precision at top k%, which is the predictive accuracy within the models’ top k% most confident predictions.

Because school districts are constrained by tightening budgets, they are often limited in the number of students they can intervene with. As such, evaluating our models with this metric not only ensures that we are generating predictions that are accurate, but it also allows us to measure how well each model works as it pertains to the allocation of resources by school districts. As illustrated in Figure 3.4, we determined what percentage of students within the top 10% (i.e., those with the highest risk scores) ultimately did not graduate on time. The choice of this metric reflected the reality that many schools can only intervene on a small percentage of their entire student body

---

3In that scenario, we used $k = 10$ somewhat arbitrarily. However, based on interviews we conducted with school representatives, that seems to be an adequate estimate of how many students schools can typically place on intervention programs.
Figure 3.4. Precision at top 10% metric explained for a cohort of 100 students. Only the 10 students with highest risk scores are shown. Those labeled in red did, in fact, not graduate on time. Conversely, the green shades represent students that graduated on time and who should likely not have been placed at the “at-risk” category. Model 2 in this example outputs slightly more reliable risk scores.
CHAPTER 4
IDENTIFYING AND PRIORITIZING STUDENTS AT RISK OF NOT GRADUATING HIGH SCHOOL ON TIME

4.1 Introduction

Although high school graduation rates in the United States have been slowly but steadily increasing during the past decade, over 700,000 students per year still do not graduate high school within four years of entering it [60]. Students who take longer than four years to graduate can create large economic strains on school districts; moreover, high school drop-outs have markedly higher rates of incarceration and poverty compared to high school graduates [20, 70]. The causes for not graduating high school on time are diverse, including poor academic performance, conflicts with peers and family members, lack of interest, and unexpected life disturbances [27, 49, 105].

Recent research has found these diverse causes often manifest themselves through common sets of indicators routinely recorded by schools such as students’ attendance, behavior, and coursework [12, 19]. Consequently, school districts have been increasingly using these readily available indicators to help identify at-risk students. For instance, school districts from 31 states used some type of early warning system in 2012, up from 18 states in 2011 [39]. These Early Warning Indicator (EWI) systems are currently built in vastly different ways, as indicated by Bowers et al.’s recent review of 36 studies on this topic [28]. However, a common theme among these studies is that they often make binary, rather than continuous or ranked, predictions: a student is either predicted as likely to graduate or not.
We built upon this considerable prior literature on EWIIs and aimed to extend it in three substantial ways. First, we developed predictive models of high school graduation using machine learning approaches that can identify at-risk students more accurately than prior analytic approaches. Second, we developed urgency score models that can rank at-risk students based on when those students are most likely to go off track. Third, we developed a prototype user-interface to help teachers, counselors and administrators better understand the various factors contributing to our models’ predictions. This three-step approach could help schools more efficiently allocate limited resources by prioritizing which students are most in need of help, and target intervention programs to match those students’ particular needs.

We partnered with a large school district in the mid-Atlantic region, enrolling over 150,000 students, and were provided de-identified, student-level longitudinal data for a cohort of approximately 11,000 students who were tracked from 6th to 12th grade. Using this data, this chapter describes our approach to early identification of the students who are at risk of dropping out, prioritizing these students based on the level of that risk, predicting the urgency and timing of the dropout, and informing schools with the particular indicators for each prediction.

4.2 Dataset Description

We received a dataset containing de-identified information for nearly 11,000 students who were expected to graduate in 2013. Most students were tracked from 6th to 12th grade, while some arrived throughout the study and had missing data for any years prior to enrollment. Figure 4.1 illustrates the student flow in and out of the school district over time.

The vast majority of the students (highlighted in blue) flow in and out of the system according to the expected schedule. In green, we see that small subsets of students enter the district at the beginning of each academic year, with a noticeably
Figure 4.1. Flow of student enrollment at school district A over time. The larger (blue) portion highlights students that are currently enrolled. The green subset indicates students that are flowing into the system at each time stamp, whereas the red group represents students that are leaving the school district.

The larger group of incoming students being observed in 9th grade. Lastly, in red, we show the subset of students who did not graduate on time, either by dropping out or continuing in school beyond the fourth high school year. Overall, this lastly described subset accounts for a total of 12.4% of the student population.
<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Description</th>
<th>Available On</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>Binary</td>
<td>Student’s gender</td>
<td>6th - 12th</td>
</tr>
<tr>
<td>birth</td>
<td>Numeric</td>
<td>Student’s birth year and month</td>
<td>6th - 12th</td>
</tr>
<tr>
<td>sname</td>
<td>Nominal</td>
<td>Name of school where student is enrolled</td>
<td>6th - 12th</td>
</tr>
<tr>
<td>absrate</td>
<td>Numeric</td>
<td>Student’s absence rate for a given academic year</td>
<td>6th - 12th</td>
</tr>
<tr>
<td>tardyr</td>
<td>Numeric</td>
<td>Student’s tardiness rate for the academic year</td>
<td>6th - 8th</td>
</tr>
<tr>
<td>nsusp</td>
<td>Numeric</td>
<td>Number of suspension incidents</td>
<td>6th - 12th</td>
</tr>
<tr>
<td>mobility</td>
<td>Numeric</td>
<td>Cumulative to date number of unexpected entries or withdrawals</td>
<td>6th - 12th</td>
</tr>
<tr>
<td>new</td>
<td>Binary</td>
<td>Flag to denote if the student is new to the school district</td>
<td>9th - 12th</td>
</tr>
<tr>
<td>q1-4gpa</td>
<td>Numeric</td>
<td>Quarter Marking Period Averages</td>
<td>6th - 12th</td>
</tr>
<tr>
<td>gmapr</td>
<td>Numeric</td>
<td>MAP-R National Percentile Ranks</td>
<td>6th - 8th</td>
</tr>
<tr>
<td>gmsam</td>
<td>Numeric</td>
<td>Maryland School Assessment in Math Proficiency Level</td>
<td>6th - 8th</td>
</tr>
<tr>
<td>psatv</td>
<td>Numeric</td>
<td>PSAT Critical Reading</td>
<td></td>
</tr>
<tr>
<td>psatm</td>
<td>Numeric</td>
<td>PSAT Math</td>
<td>10th</td>
</tr>
<tr>
<td>retained</td>
<td>Binary</td>
<td>Flag to denote if the student was ever retained to date in a grade</td>
<td>9th - 12th</td>
</tr>
<tr>
<td>grad</td>
<td>Binary</td>
<td>Flag to denote if the student graduated from high school on time</td>
<td>12th</td>
</tr>
</tbody>
</table>
End of year data were captured as students progressed towards their high school graduation. Given that the district already had an EWI system in place, the features we chose to focus most of our attention on were attendance, academic performance, behavior, mobility, and a few aspects of demographics. The features were selected based on the examination of previous studies from other school districts as well as factors found to impact students in specific schools served by our partners. It is worth noting that the techniques described in this chapter are not limited to only using these types of features, and future work includes expanding to other forms of data about students, their behaviors, and their learning environments.

Table 4.1 contains a brief description of the data we utilized. Note that most information is available on a yearly basis starting at 6th grade, with the exception of a few features such as PSAT scores and Measures of Academic Progress (MAP) percentile ranks that only exist for a subset of the grades. Additionally, as performance marks were available quarterly, each was represented by four distinct features per student for each academic year in the dataset.

4.2.1 Preliminary Analysis

Given that one of our goals was to design and evaluate predictive models that would aid in the early detection of students that are at risk of not graduating high school on time, one of our first tasks was to do a preliminary analysis to see if any aspects of the available data were particularly important in predicting risk. Table 4.2 summarizes the results of that experiment, wherein we ranked features in the order of their importance, as we attempted to infer on-time graduation – grad flag in Table 4.1 – with respect to four metrics: information gain (IG), gini impurity (GI), stepwise regression (SR) and single feature performance (SFP).

\[1\] This consisted of evaluating a logistic regression trained with each feature individually.
TABLE 4.2

FEATURE RANKINGS ILLUSTRATING THE DEGREE OF IMPORTANCE ASSIGNED TO EACH INDIVIDUAL STUDENT-LEVEL FEATURE BY EACH METHOD TESTED

<table>
<thead>
<tr>
<th>Rank</th>
<th>IG</th>
<th>GI</th>
<th>SR</th>
<th>SFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>q1-4gpa</td>
<td>retained</td>
<td>q1-4gpa</td>
<td>q1-4gpa</td>
</tr>
<tr>
<td>2</td>
<td>gmapr</td>
<td>q1-4gpa</td>
<td>gmsam</td>
<td>mobility</td>
</tr>
<tr>
<td>3</td>
<td>retained</td>
<td>gmapr</td>
<td>gmapr</td>
<td>nsusp</td>
</tr>
<tr>
<td>4</td>
<td>birth</td>
<td>absrate</td>
<td>retained</td>
<td>absrate</td>
</tr>
<tr>
<td>5</td>
<td>absrate</td>
<td>birth</td>
<td>birth</td>
<td>ssname</td>
</tr>
</tbody>
</table>

Aggregating students based on some of the features provided interesting insights that were previously not known. For example, as illustrated in Figure 4.2, students who did not move across different schools (or who moved a small number of times) appear to be more likely to graduate from high school on time, whereas the opposite seems to be true for students with higher mobility values. We also found evidence supporting more obvious relationships such as the correlation between GPA and likelihood of graduating on time (Figure 4.3).

4.3 Who? – Identifying Students at Risk

In an ideal scenario, school districts would have sufficient resources to provide each and every student with individualized mentoring programs. However, given notable budget constraints, many schools must instead use these programs to focus on students who are most at risk of not graduating on time. As discussed by Gleason and Dynarski [50], secondary schools may miss ideal opportunities to help at risk
Figure 4.2. Student mobility values vs. on-time graduation rates. Mobility is the number of times a student unexpectedly switched schools. The shaded region corresponds to 95% CIs.

Figure 4.3. Student GPAs (binned) vs. on-time graduation rates. The shaded region corresponds to 95% CIs.

students when students who are on track to graduation are mislabeled.

To help prioritize such efforts, our district partners developed an early warning
system to flag potentially at risk students. These flags can be generated yearly\textsuperscript{2} so that schools can monitor their students and take immediate action when necessary.

We compared this existing system with more sophisticated machine learning models for estimating risk scores (i.e., predicted probabilities of not graduating high school on time). To assign risk scores, these models used all available information about students at the end of each academic year. One key difference between our models and those used by our partner is that we used all the available historical data. For instance, the early warning model used at the end of 9\textsuperscript{th} grade used all available data from 6\textsuperscript{th} to 9\textsuperscript{th} grade.

We evaluated these models’ performance by computing the precision at top k\%, which is the predictive accuracy within the models’ top k\% most confident predictions. As illustrated in Figure 3.4, we determined what percentage of students within the top 10\% (i.e., those with the highest risk scores) ultimately did not graduate on time.

Risk scores were estimated using 10-fold cross validation to ensure that the student instances being evaluated at each iteration were held out of the training set. Though we tested a variety of classification methods, the results presented here concern random forest models\textsuperscript{29} which were most precise, and logistic regression models which were more interpretable while still performing well.

4.3.1 Results

Figure 4.4 shows that the machine learning models performed noticeably better than our partner’s existing rule-based model. For example, in the 10\textsuperscript{th} grade models, the random forest and logistic regression models performed at 75\% and 74\% respectively, whereas the existing model performed at 38\%.

Our models, which used students’ historical data, consistently improved over time. 

\textsuperscript{2}More specifically, these flags are generated quarterly but can be consulted in an yearly basis.
as the models continued to learn more information about each student. However, this trend was not consistently observed for the existing model (blue line), which did not use historical data (e.g., the 9th grade model only used 9th grade GPAs). This model’s performance notably dropped at the transition from 9th to 10th grade. These results suggest the value of using all available historical information about each student.

Figures 4.5 and 4.6 help show what kinds of students are ranked higher by our models. These figures mirror Figures 4.2 and 4.3 but are based on the top decile of predicted risk scores (based on random forest models), rather than the observed outcome of not graduating on time. As shown, student-level features (e.g., mobility, GPA) have a similar relationship with our models’ risk scores (Figures 4.5 and 4.6) as with observed outcomes of not graduating on time. This result was expected and helps confirm that the models are learning important relationships that exist in the data.
Figure 4.5. Student mobility vs. high-risk rates. Mobility is the number of times a student unexpectedly switched schools. The shaded region corresponds to 95% CIs.

Figure 4.6. Student GPAs (binned) vs. high-risk rates. The shaded region corresponds to 95% CIs.
4.3.2 Stability of Models

We also investigated how stable our models were across time (i.e., how much risk scores for a given student varied from grade to grade). A more stable model could allow schools to more consistently focus efforts on students identified by our system, without worrying about continuously switching their focus. To examine stability, we counted the number of transitions that students made between three different risk categories (high, moderate, and low) from year to year. For example, the path illustrated in Figure 4.7 includes two transitions.

We defined *high risk* as the top 10% of risk scores, *moderate risk* as the following 20%, and all other students were placed in the *low risk category*. We found that in our partner’s EWI model, most students (63%) made at least one transition between different risk categories over time, and 25% of students made three or more transitions. In contrast, our random forest model was more stable, with 40% of the students making at least one transition and 10% making three or more. Hence, the random forest models were likely more robust to small fluctuations in data items such as GPA. In the future, we plan to validate the model transitions with empirical evidence from teachers and parents. These figures are summarized in Table 4.3.

\[\begin{array}{cccccc}
6^{th} & 7^{th} & 8^{th} & 9^{th} & 10^{th} & 11^{th} \\
\text{high} & \text{high} & \text{moderate} & \text{moderate} & \text{moderate} & \text{high} \\
\end{array}\]

---

These cutoffs were chosen to remain consistent with those already in use by our partner’s existing system.
TABLE 4.3

PERCENTAGE OF STUDENTS MOVING ACROSS DIFFERENT RISK GROUPS OVER TIME

<table>
<thead>
<tr>
<th>Number of Transitions</th>
<th>Current EWI</th>
<th>Random Forest Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>37%</td>
<td>60%</td>
</tr>
<tr>
<td>1</td>
<td>19%</td>
<td>17%</td>
</tr>
<tr>
<td>2</td>
<td>19%</td>
<td>14%</td>
</tr>
<tr>
<td>3</td>
<td>15%</td>
<td>6%</td>
</tr>
<tr>
<td>4</td>
<td>7%</td>
<td>2.7%</td>
</tr>
<tr>
<td>5+</td>
<td>3%</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

4.3.3 Identifying Dropout Students in Isolation

Thus far, we have described a machine learning approach to identifying students who are at risk of not graduating high school on time (combined group), either because they drop out (drop out group) or remain in the high school system for more than four years (remain group). Certain schools may wish to place special emphasis on identifying drop outs, since they lose contact with those students. For this, we used the same approach described above, but attempted to classify only those students who drop out. From a machine learning perspective, this classification may be more difficult because there is a smaller group of students to identify, but on the other hand, it could be easier because this distinction may increase the homogeneity of the groups. For comparison, we also used this approach to classify the remain group vs. all other students. According to the area under the curve of the receiver operating characteristic (AUROC – a normalized metric for comparing classifier performance across different scenarios; for details on using this metric in the context of early
warning identification see [28]), this approach was similar in accuracy for classifying the combined group (0.89) and the drop out group (0.89), but was slightly worse for the remain group (0.83), though still far better than random guessing (0.5). Thus, depending on the interests and needs of the school district, this approach shows promise for identifying students who are likely to have various undesirable outcomes, so that help can be provided as early as possible.

4.4 When? – Prioritizing Interventions

In the previous section, we discussed in detail the methods that we employ to identify those students who are at risk of not graduating high school on time. In addition, we also demonstrated that such models can outperform more simplistic rule-based methods that are often the first choice among school districts implementing EWIs. While it is important to identify those students who are at risk, the problem does not end there. Often, the volume of students at risk is such that schools may have difficulty estimating who may be in most immediate need. We mentioned in the previous sections that our dataset was comprised of approximately 11,000 students. As we subsequently ranked 10% of these students as being at highest risk, the group reduced to having 1,100 students. Providing mentoring and interventions to such sizable set of students is a non-trivial task, and automating a prioritization process can greatly improve efficiency.

4.4.1 Understanding Priority

An interesting question that arises in this context is How do we further prioritize students at risk? A straightforward way to do that is to rank students based on their risk scores (discussed in the previous section). Though this is an adequate measure, it would be useful to understand if we could incorporate other notions (and not just risk of not graduating on-time) into the picture. For that purpose, we decided to use a
metric that we will call *time to off-track*. To illustrate this concept, let us first define
the term *off-track*. A student can be categorized as *off-track* if he or she is retained
(or drops out) at the end of some grade level. It is ideal to provide interventions
to students before either of these undesired outcomes, as opposed to taking a more
reactive approach. In other words, access to the appropriate interventions should
be provided to at-risk students before they go *off-track*. We define *time to off-track*
as the interval between the current instant and the point at which a student goes
off-track for the first time (assuming the student has not yet gone off-track). The
smaller the time to off-track, the sooner the student is likely to be at risk.

To understand the motivation behind this metric, let us consider a scenario where
two students are classified as being at-risk by our predictive models and let us assume
that there is scope to provide help to only one student. In this case, it makes more
sense to provide help to that student who is likely to go *off-track* sooner.

The formulation of this problem boils down to predicting time to off-track for
at-risk students at the end of every grade. For instance, at the end of grade 8, for
each student at risk of not graduating on time, we need to assign a label between 1
to 5, where 1 indicates that a student is likely to go off-track by the end of next year
(at the end of grade 9), 2 indicates that a student is likely to go off-track at the end
of grade 10, and so on.

It should be noted that, since in the previous example we start at the end of
grade 8, we should only have 4 labels, one for each subsequent grade level. However,
in that context, we use a label of 5 to indicate that a student is not off-track until
the end of high school (likely indicating a possible retention in grade 12). We have
analogous prediction task formulations for every grade. We further highlight that
our dataset does not contain any students who go off-track before the end of grade
9. It is important to bear in mind that this prioritization task is carried out only for
those students that were previously classified as being at high risk (top 10%) by the
predictive models discussed in the previous section.

4.4.2 Risk Scores as a Proxy for Time to Off-track

Now that we have defined a metric for prioritizing students, it is still not out-of-place to ask the question *Does a higher risk score imply a shorter time to off-track?* That is, can we assume that students who were given high risk scores by our predictive models are more likely to go off-track sooner than those with lower risk scores? If that is the case, we do not need any other models for predicting time to off-track, and we can use the setup from the previous section to guide the prioritization for interventions. In order to validate this idea, we carried out multiple experiments with the objective of computing the correlation between risk scores and time to off-track using multiple metrics such as Pearson’s and Spearman’s rank correlation coefficients.

We note that risk scores are continuous values by definition, while time to off-track assumes discrete values starting with a minimum of 1. For that reason we computed correlations twice. The first experiments kept these scales unchanged, whereas for the second iteration we discretized the risk scores.

Figure 4.8 shows the results of these experiments with Pearson’s correlation coefficient. There, it can be seen that there does not exist a high correlation between risk scores and time to off-track from grades 6 through 10. However, it can be seen that using data up until grade 11 and predicting time to off-track results in a higher correlation with risk scores ($\sim 0.5$). This behavior can be explained by the fact that when we consider all the data collected before the end of grade 11, predicting time to off-track is tantamount to predicting if a student will be retained in grade 12 (and thus not graduate on time) or not. In other words, this is equivalent to the task of predicting if a student is at risk of not graduating on time. For all the other grades except grade 11, the correlation coefficient values are quite small. This analysis reveals that risk scores do not serve as a reliable proxy for time to off-track,
thus motivating the need for building new predictive models for the task at hand.

4.4.3 Predicting Time to Off-track

Recall that the idea of this second-step prioritization is to take the set of students classified as being at high risk by our risk prediction models (discussed in the previous section) at the end of each year, and further estimate the time to off-track for each of them. This essentially means that this task will be carried out at 5 different time stamps, one for the end of each grade starting from grade 6 to grade 11. Each of these tasks predicts if an at-risk student is likely to go off-track after one, two, three, or more years. The variable \( \text{time to off-track} \) takes a value between 1 and \((12 - \text{current grade}) + 1\) for each at-risk student, 1 denoting that the student is likely to go off-track by the end of next year and \((12 - \text{current grade})\) denoting that the student is likely to go off-track at the end of 12th grade. The value \((12 - \text{current grade}) + 1\) indicates that the student is unlikely to go off-track.

Now that we have established the background for the prediction tasks, let us consider the technical aspects. Since this is a prediction task, we can use a classification algorithm such as logistic regression, decision trees, or random forests for solving this problem. However, another interesting detail to note here is that time to off-track is an ordinal (rather than categorical) variable. This means that there is an inherent ordering on the values that time to off-track takes, and values that are closer to each other are more similar than others. For example, time to off-track values of 1 and 2 are more similar to one another than are 1 and 5. For this reason, classification frameworks that treat labels as categorical variables might not be the optimal choice in this context. Therefore, we also consider ordinal classification methods which assume that the outcome labels are ordered for our analysis. In addition to these two classes of techniques, we also investigate the usefulness of models from the survival analysis literature such as Cox regression, which can be readily applied to this scenario.
In order to evaluate these various classes of models, we rely on two metrics:

- **Accuracy**: This metric is a statistical measure for quantifying the degree of correctness with which a prediction model is able to label the data points. Let $a_i$ be the actual ground truth label for a student $i$ and let $p_i$ be the prediction. Assuming that there are $N$ at-risk students, accuracy can be written as:

$$\text{Accuracy} = \frac{\sum_i I(a_i = p_i)}{N}$$

where $I()$ is an indicator function that results in 1 if the condition is met and a 0 otherwise. The higher the accuracy, the better the prediction model. Though accuracy is a very widely used metric and is very useful in practice, it is also a very conservative metric in this context. To illustrate, let us consider a student $i$ with a time to off-track value of 2 indicating that the student actually dropped out 2 years from the current year under consideration. If the predicted value for this student turns out to be 3 instead of 2, the accuracy metric penalizes this because the predicted value is not equal to the actual outcome. Further, the metric does not distinguish between the magnitude of errors. In the case of a student $i$, a predicted value of 3 and a predicted value of 5 are both penalized. However, since we are dealing with an ordinal scale, a predicted value of 5 is much worse than a predicted value of 3, as 3 is closer to the ground truth label of 2 than 5 is.

- **Mean Absolute Error (MAE)**: This metric is a statistical measure of the degree of closeness between the actual outcome and the predicted outcome. With the notation defined above, MAE can be defined as:

$$\text{MAE} = \frac{1}{N} \sum_i |a_i - p_i|$$

The lower the value of MAE, the better the prediction model. It can be seen that this metric incorporates the magnitude of difference when penalizing prediction errors. For example, this metric penalizes the predicted value of 5 much more than the predicted value of 3 when the ground truth label is 2.

The results of this prediction task using classification frameworks (logistic regression), survival analysis techniques (Cox regression) and ordinal regression methods (ordinal regression trees) are shown in Figures 4.9 and 4.10. In addition, we also present the results from using the discretized risk scores as a proxy for time to off-track. Figure 4.9 encapsulates the accuracy metric, and Figure 4.10 presents the MAE metric. It can be seen that ordinal regression-tree based models outperform
traditional classification, survival analysis techniques, and the risk score baseline. The baseline exhibits inferior performance both in terms of accuracy and MAE. Lastly, we also see that survival analysis techniques slightly outperform traditional classification.

4.5 Why? – Understanding Individual Risk Factors

In the previous two sections we outlined methods that can be used by high schools to identify which students are at high academic risk, and from that subset, who may need attention most immediately. Knowing that information is extremely valuable and it helps schools to not only make better use of their resources, but it also provides a means to prioritize intervention efforts. This section will address the last step of our overall methodology: suggesting the appropriate context for interventions.

A variety of factors may contribute to a student’s decision to drop out of high school, and as shown by Alexander et al. [9], these factors may independently affect a student’s trajectory. Hence, knowing which features contributed towards a student’s high risk scores can alert counselors of what would potentially be the most beneficial interventions to suggest.

With that in mind, we developed a web-based dashboard application, illustrated in Figure 4.11 that helps educators dive into detailed breakdowns of their students’ reported risk scores. While these measurements are not informative in isolation, being able to see a student’s risk score trajectory over time, as well as his or her grade, absence, and mobility history can be of great help when attempting to define what form of intervention may be most appropriate for each case.

It is important to note that these indicators are not reasons for dropping out but have been found as leading indicators (and predictors) for the predictions made by our models. Our future work includes using these initial hypotheses to design experiments to see which of these could be causal factors leading to dropout and
Figure 4.8. Correlation between risk scores & time to off-track.

Figure 4.9. Accuracy of predicting time to off-track.

Figure 4.10. Mean Absolute Error predicting time to off-track.
working with schools to design effective interventions targeting these factors.

A live demo of our dashboard created with artificial data is accessible at [2]. There, as well as in the few static screenshots seen in Figure 4.11, we highlight that the major intent was to ensure the simplicity and user friendliness of this interface. Students can be quickly selected from a list that is sorted based on risk scores, and upon selection, a categorized breakdown of that student’s historic data is displayed with each category being dynamically color-coded to indicate how many standard deviations that student’s data values are from the overall mean. The entire code used to create this dashboard has been made open source and can be found at [3].

4.6 Overall Impact and Future Work

We have shown two key ideas in this chapter that can help schools graduate more students on-time. The first one is to produce a ranked list that orders students according to their risk of not graduating on time. The second one is to predict when they’ll go off track, to help schools plan the urgency of the interventions. Both of these predictions are useful in the identification and prioritization of students at risk, and allow the schools to target interventions. The eventual goal of these efforts is to focus the limited resources of schools to increase graduation rates. In order to achieve that goal, it is important to consider the impact of interventions and match them with students effectively. One of the key next steps in our effort is to build “persuasion” models that can be used to rank students in terms of how much they will be impacted by a given intervention, allowing schools to identify students who are most likely to respond to specific programs. This will require experiments to be conducted testing interventions and the use of machine learning approaches to build the persuasion models.

Another important future direction is to define a collection of interventions (i.e. what types, how to deliver them, who is involved, how often are they delivered, etc.)
and use student data to predict the optimal set of interventions for each student. This will allow schools to *personalize* the interventions and increase the chances of improving outcomes for more students who are at risk of dropping out.

### 4.6.1 Benefits to the School District

The results of this study have helped the school district systematically adjust analytical methods as they continue to build a universal EWI system. Based on the findings of our work, the school district moved from a rule-based model to applying a logistic regression model for the second prototype of their EWI system. The revisions to the model were applied to all students in the district, with a fully operational roll out of this EWI system taking place in the Fall of 2014. Additionally, given the high performance associated with the Random Forest model, the district is currently planning to investigate its application for future work.

In addition to modifying the analytical methods currently employed, the ability to predict time to off-track would serve the school district well, as it relates to allocating resources. With its preliminary EWI prototype, the school district was asked to assist in identifying those students who were deemed as priority for providing an intervention. While the focus of the district-developed EWI system is to provide support for all students, it is recognized that there may be a need to prioritize support to students who are identified at higher risk and/or more urgent. The district intends to further investigate the application of this metric to the student population.

Finally, the district is also highly interested in the web-based dashboard application that was developed. Graphic portrayal of the data not only helped to concisely summarize the data, but it also drew attention to certain nuances captured in the dataset that had previously gone unnoticed. The options for visually displaying data related to an EWI system revealed the potential for creating a dynamic interface that allows for elucidating, interpreting, and analyzing the information from a new
perspective. The school district recognizes that a web-based system that connects multiple data points for all students can serve as a valuable resource to both school-based staff and district leaders.

4.7 Acknowledgments

The work described in this chapter was done as part of (and partially supported by) the 2014 Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship at the University of Chicago, with contributions by all following peers: Himabindu Lakkaraju, Nasir Bhanpuri, David Miller, Ben Yuhas, Kecia L. Addison, Shihching Liu, Marilyn Powell, and Rayid Ghani.
Figure 4.11. Student Report Card Dashboard depicting risk score, GPA, and absence rate trajectories.
CHAPTER 5

A MACHINE LEARNING FRAMEWORK TO IDENTIFY STUDENTS AT RISK OF ADVERSE ACADEMIC OUTCOMES

5.1 Introduction

One of the perennial challenges faced by school districts is to improve student graduation rates. Though the magnitude of this problem has reduced due to a steady rise in high school graduation rates over the past few years, nearly 730,000 students in the United States (U.S.) do not finish high school on time every year [60]. A myriad of reasons ranging from economic problems, lack of motivation, and unexpected life changes can delay students’ graduation or cause them to drop out [27, 49, 105]. Studies have shown that not graduating high school on time impacts a student’s future career prospects immensely [20, 70]. In addition, students who do not graduate on time can strain school districts’ resources. To address this issue, school districts have been heavily investing in the construction and deployment of intervention programs to better support at risk students and their individual needs.

The success of these individualized intervention programs depends on schools’ ability to accurately identify and prioritize students who need help. Traditionally, schools relied on feedback from instructors, and used heuristic rules based on metrics such as GPAs, absence rates, and tardiness to identify at-risk students [28]. Though human judgment and heuristics can often be accurate, they serve as rules of thumb, are static, expensive to maintain, and often error prone [109]. Further, the set of heuristics which might help in identifying at-risk students for a particular cohort of
students within one school district might not generalize or transfer to other cohorts and schools.

As alternatives to manually created rule-based systems, recent research has indicated the potential value of machine learning approaches such as Logistic Regression, Decision Trees, and Random Forests \cite{7, 28, 110}. Trained using traditional academic data, these machine learning approaches can often identify at risk students earlier and more accurately than prior rule-based approaches \cite{7}.

Nevertheless, the application of such methods to this particular context is still in its early stages, even for schools with state-of-art technology and analytics teams. To build more robust and comprehensive early warning systems, we partnered with two large U.S. school districts with a combined enrollment of approximately of 200,000 students. Following a number of discussions with the district officials who oversee the implementation of early warning systems, we developed an outline of their expectations:

- **Using historical data**: Schools have historical data that describe current and past student performances, and would like to use that to identify students at risk in future cohorts.

- **Ranking students using risk estimates**: School districts have limited resources for intervention programs, and their exact allocation can fluctuate over time, directly affecting the number of students that can be enrolled to such programs. For that reason, school officials need the ability to pick the top k% students who are at risk at any given point (where k is a variable). This requirement calls for the use of algorithms which can rank students according to their probability of not graduating on time.

- **Interpretability**: It is important to understand the student-level features and how they are being used by each algorithm. In fact, school officials consistently ranked interpretability as a very important factor for any approach. Frequently, simple rule based systems are preferred to *intelligent* algorithms mainly because they can be easily understood and acted upon.

- **Early predictions**: Students who are at risk of not graduating should be identified as early as possible so that appropriate help can reach them in a timely manner. This requirement favors algorithms which can identify at-risk students early on.
• **Identifying risk before off-track:** It is ideal to identify students who are at risk even before they start failing or repeating grades. School officials acknowledge that it considerably more difficult to help a student who is already off-track.

• **Visualizing risk scores for each student:** All of the above information needs to presented in a way that is clear and understandable by professional educators who are not familiar with machine learning. We have previously developed web-based software that can display model predictions on each student, helping teachers and administrators gauge how much support each student needs.

Our interactions with educators revealed that there were several deeper and interesting challenges in this setting, and helped us quickly understand that evaluating algorithms simply using AUC and precision/recall metrics would not be sufficient. In this work, our goal is to investigate how to evaluate the suitability of any given algorithm for the problem at hand so as to ensure that it meets the expectations of educators and school officials. To this end, we apply several off-the-shelf machine learning algorithms to identify at-risk students and analyze their behavior according to several evaluation techniques. To summarize, our major contributions are:

• We present a novel framework for evaluating algorithms which identify students at risk of not graduating high school on time. The evaluation process is designed to cater to the needs of educators rather than focusing on commonly used machine learning metrics alone.

• We present a rigorous qualitative and quantitative comparison of several well known machine learning algorithms using the proposed evaluation process.

• We carry out all our experimentation using data from multiple student cohorts in collaboration with two major school districts in the United States. Unlike the work described in Chapter [4] where evaluation is carried out via cross validation on a single cohort, we use disjoint cohorts for training and testing, thus validating the system’s performance in a more realistic manner and making it directly applicable for deployment in all the school districts in the US.
5.2 Dataset Description

The work we describe in this chapter is being done in collaboration with two school districts in the United States, one of which (District A) is among the largest districts in the mid-Atlantic region with over 150,000 students enrolled across 40 schools. The other is a medium-sized district on the east coast with an enrollment of approximately 30,000 students across 39 schools (District B). Both of these districts are instituting several measures to help students, and have recognized the importance of early warning indicator systems for identifying at-risk students. District A (as illustrated in Chapter 4) had a rule-based early warning indicator system in place, using several important indicators such as academic performance, behavior, mobility, and a few demographic attributes. Our partnership with these school districts has been critical in developing a machine learning system that is not only based on real data but also designed for the needs and priorities of educators.

We obtained data from each of these school districts. The dataset provided by District A comprises two cohorts of 10884 and 10829 students, expected to graduate in 2012 and 2013 respectively. Most of the students in these cohorts were tracked from 6th to 12th grade, while some arrived throughout the study. Students belonging to the latter group have missing data fields for all the years prior to their enrollment in the school district (which is normal since the school only starts collecting data when students enroll). The data contains several attributes for each of these students such as their GPAs, absence rates, tardiness, gender etc. About 90% of the students in each of these cohorts graduated high school within 4 years of enrollment. A vast majority of the students in the dataset were enrolled in the school district right from 6th grade and graduated within the stipulated time.

The dataset obtained from District B represents two cohorts of 1499 and 1575 students, with expected graduation dates in 2012 and 2013 respectively. In this dataset, most of the students were tracked from 8th - 12th grade and several academic
and behavioral attributes of these students were recorded. However, some arrived throughout the study and subsequently have missing data fields for years prior to their enrollment. About 95% of the students in each of these cohorts completed high school on time. The remaining 5% of the students either dropped out of school or took more than 4 years to graduate high school.

While there could be a variety of reasons for academic difficulties ranging from lack of motivation to economic concerns, recent research has demonstrated that these diverse causes often manifest themselves through a common set of indicators such as academic performance, behavior, and attendance\cite{12, 19}. The data used for this analysis captured many of these indicators. Table 5.1 provides an exhaustive list of all features that we used in the analysis. The availability of each of these attributes in a given dataset is indicated by the two rightmost columns of the table. It can be seen that there are minor variations in the ways data is recorded for the two districts. For instance, GPA is recorded on a quarterly basis for District A, while District B records that information yearly. Our analysis is not sensitive to such representational variations. In fact, the framework proposed in this chapter is generic enough to be applicable to any given set of features.

5.3 Framework Overview

In this section, we present an overview of the models that we will be using throughout this study. In addition, we also describe in detail the experimental setup that we use for all our prediction tasks and our evaluation choices that are designed to match the real world setting as closely as possible.

**Problem Setting:** In order to provide assistance to students who are at risk of not graduating on time, we first need to accurately identify such students. This can be
TABLE 5.1

LIST OF STUDENT ATTRIBUTES AND THEIR AVAILABILITY IN
DISTRICTS A AND B

<table>
<thead>
<tr>
<th>Student Attributes</th>
<th>District A</th>
<th>District B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Age</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>City</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Street</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>School Code</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Absence Rates</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Tardiness Rates</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td># of Suspensions</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td># of Unexpected Entries/Withdrawals</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Quarterly GPA</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>Cumulative GPA</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Cumulative Math GPA</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Cumulative Science GPA</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Cumulative Social Science GPA</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Cumulative English GPA</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>MAP-R National Percentile Ranks</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>Math Proficiency Scores (MPS)</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>PSAT Critical Reading</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>PSAT Math</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>Limited English Proficiency</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Economically Disadvantaged (EDS)</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Is student new to the school district?</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Is student disabled?</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Was student ever retained?</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Did student graduate on time?</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
achieved by using algorithms that can learn from the outcomes of students in the earlier cohorts. Schools have records on which of the students from prior cohorts failed to graduate high school within 4 years. From Table 4.1 it can be seen that the flag *Did the student graduate high school on time?* captures this aspect and hence can serve as the outcome variable. We compute the complement of this flag which takes the value 1 if the student failed to graduate on time, and 0 otherwise. We use the term **no_grad** to refer to this complement variable and use it as the response variable for all our prediction tasks. The problem of identifying students who are at risk of not graduating on time can thus be formulated as a binary classification task with **no_grad** as the outcome variable. All the other variables in Table 4.1 can be used as predictors.

**Models:** To predict if a student is at risk of not graduating on time, we experiment with Random Forests [29] (RF), Adaboost [106] (AB), Logistic Regression [56] (LR), Support Vector Machines [25] (SVM), and Decision Trees [99] (DT). We use scikit-learn [94] implementations of all these models.

**Experimental Setup:** Our datasets included cohorts of students graduating in 2012 and 2013. Recent research that deals with the problem of predicting student performance evaluated the models via cross validation on a single cohort [7 110]. Though this is an acceptable way of estimating any algorithm’s performance in general, it is not ideal for the current setting. However, school districts often have access to outcomes and other features from previous cohorts, which allows us to evaluate our models more strictly by predicting future outcomes based on training data from previous cohorts. We carry out all of the evaluations in this manner, using the cohort of students graduating in 2012 as the training set and the later cohort of students graduating in 2013 as the test set.
Some of the models that we employ such as Random Forests involve sampling random subsets of data. This creates a certain degree of non-determinism in the estimated outcomes. In order to account for this, we perform 100 runs with each of these models and average the predictions (and/or probabilities) to compute the final estimates. During our analysis, we also experimented with the leave-k-out strategy. As a part of this approach, we executed 100 iterations for each classification model, training them on \((N - k)\) randomly chosen data points each time, where \(N\) is the size of the entire dataset and \(k = 0.01 \times N\).

With this framework in place, we now proceed to present how we evaluate each of the models while taking into account educators’ requirements.

5.4 Analysis of Predictive Models

School districts are interested in identifying those students who are at risk of not graduating high school on time before they reach the end of middle school. This helps them plan their resource allocation ahead of time. In this section, we focus on this setting by predicting if a student is at risk of not graduating high school on time using the data available prior to the end of middle school. More specifically, to predict the outcome variable \texttt{no_grad}, we use GPAs, absence rates, tardiness, demographics attributes, and other scores and flags listed in Table 4.1 from the time of enrollment up until grade 8. In this section, we address the following questions:

- How well do each of the models perform when evaluated using traditional metrics such as precision, recall, and AUC?

- How do we ensure that the probabilities / confidence score estimates produced by various algorithms are good in order for schools to reliably deploy interventions?

- How do we compare the goodness of such estimates and show robustness of the results?
5.4.1 Evaluation Using Traditional Metrics

Our goal here is to evaluate the performance of various models on the task of predicting if a student is likely to graduate high school on time. Since we are dealing with the prediction of a binary outcome, several standard metrics such as accuracy, precision, recall, and AUC can be readily used. We evaluate the performance of all the models using these standard metrics. Figure 5.1 shows the ROC curves corresponding to various classification models for districts A and B. It can be seen that the Random Forest model outperforms all the other models for both school districts, with AdaBoost and Logistic Regression being the next best performing solutions for both datasets. SVMs and Decision Trees exhibit varying performance across the two datasets. While SVM performs on par with Logistic Regression and AdaBoost models on District A, it performs much more poorly when applied to District B.

![Figure 5.1. ROC Curves for the task of predicting on-time graduation at both districts A and B. Models were trained using an earlier cohort and tested on the cohort for the following year.](image-url)
The usage of metrics such as AUC for a binary classification task is relatively common in machine learning. Educators, on the other hand, think about the performance of an algorithm in this context slightly differently. Their perspective stems from the fact that school districts often have limited resources for assisting students. Furthermore, the availability of these resources varies with time. Due to factors such as the number of students enrolled and budget allocated, the availability of these resources widely varies across districts. For example, while District A might have the resources to support 100 students in 2012, that number could be reduced to 75 students a year later. Building algorithms that can cater to these settings is extremely crucial when addressing the problem at hand.

After various discussions with our school district partners, we understood that an algorithm that can cater to their needs must provide them with a list of students ranked according to some measure of risk such that students at the top of the list are verifiably at higher risk. Once educators have such ranked list available, they can then simply choose the top k students from it and provide assistance to them. For instance, if District A can only support 75 students in 2013, educators in that district can simply choose the top 75 students from this rank ordered list and assist them. Furthermore, as more resources become available, they can choose more students from this list according to the rank ordering, and provide support to those students too.

The challenge associated with ranking students is that the data available to school districts only has binary ground truth labels (i.e., graduated/not-graduated). This effectively means that we are restricted to using binary classification models because other powerful learning to ranking techniques require ground truth that captures the notion of ranking. Fortunately, most of the classification models assign confidence/probability estimates to each of the data points and we can use these estimates to rank students. However, before we begin using these estimates to rank students,
we need to ensure that these estimates are indeed correct.

5.4.2 Ensuring the Quality of Risk Estimates

We begin this section by understanding how to use the confidence scores or probability estimates output by algorithms in order to rank order students. Then, we discuss how to evaluate the goodness of such estimates produced by various algorithms.

**From models to risk estimates:** Binary classification approaches output a 0/1 value for each data point. However, most of the classification algorithms involve computation of some form of confidence scores for each data point before the algorithm even assigns a label to it. In this work, we use the probability of not graduating on time as a proxy for estimating risk. While Logistic Regression estimates these probabilities as a part of its functional form, all the other algorithms output proxies to these probabilities. We obtain these proxy scores and convert them into probabilities.

Decision trees assign each data point to one of its leaf nodes and the probability of not graduating on time for any given data point is equivalent to the fraction of those students assigned to the corresponding leaf node who do not graduate on time\[31\]. Random Forests train a collection of trees, and the probability of not graduating on time for a particular data point is computed as the mean of the predicted class probabilities for each tree in the “forest” \[29\]. The class probability assigned by any single tree is computed in the same manner as that of a decision tree. Similarly, in the case of AdaBoost, which combines multiple learners, the probability assigned to a particular student is computed as the weighted mean of the predicted class probabilities of the classifiers in the ensemble\[87\]. Support Vector Machines, on the other hand, estimate the signed distance of a data point from the nearest hyperplane, and Platt scaling can be used to convert these distances into probability estimates\[72\].
Next, we describe the process of evaluating the goodness of these probabilistic estimates of risk. We use the term risk scores to refer to these probabilities from here on.

**Measuring the goodness of risk scores:** In order to understand the accuracy of the risk scores estimated by various algorithms for ranking students, we propose a simple solution. We first rank students in descending order of their estimated risk scores. We then group students into bins based on the percentiles they fall into when categorized using risk scores. For example, if we choose to create 10 bins, the bottom 10% of students who have the least risk are grouped into a single bin. Students who rank between 10\textsuperscript{th} and 20\textsuperscript{th} percentile are grouped into the next bin and so on. For each such bin, we compute the mean empirical risk, which is the fraction of the students from that bin who actually (as per ground truth) failed to graduate on time. We then plot a curve where values on the X-axis denote the upper percentile limit of a bin, and values on the Y-axis correspond to the mean empirical risk of the corresponding bins. We call this curve an empirical risk curve.

An algorithm is said to produce good risk scores and, consequently, be effective in ranking students based on those values, if and only if the empirical risk curve is monotonically non-decreasing. If the empirical risk curve is non-monotonic for some algorithm, then ranking using the algorithm’s risk scores may result in scenarios where students with lower risk scores are more likely to not graduate on time when compared to students with higher risk scores. Figure 5.2 shows these curves with 10 student bins for districts A and B respectively. It can be seen that most algorithms exhibit monotonically non-decreasing empirical curves in the case of District A. However, decision tree exhibits some degree of non-monotonicity. On the other hand, for District B, all the models except for Random Forest exhibit non-monotonicity consistently. Therefore, students should be ranked using the scores provided by Random Forest model for District B.
Figure 5.2. Empirical Risk Curves. The ranking quality of an algorithm is good if this curve is monotonically non-decreasing.

5.4.3 Comparative Evaluation of Risk Estimates

In the previous section, we discussed how to evaluate the goodness of rankings produced by various models. Here, we continue the discussion and present two metrics which are far more informative to educators than traditional precision recall curves. We previously emphasized the fact that school districts have limited resources and can assist only a certain number of students every year. Consequently, there is a strong need for algorithms which can produce good probability estimates / risk scores to rank students. Given this setting, it would be much more informative to provide precision and recall values of various algorithms at different values of $k$. We call the curves corresponding to these metrics *precision at top k curve* and *recall at top k curve* respectively. These curves help educators in readily inferring the precision and
recall of various algorithms at a threshold $k$ of their choice.

Figure 5.3 illustrates the precision at top $k$ curves for districts A and B respectively. It can be seen that there are substantial differences in the precision of algorithms at smaller values of $k$. Note that resource constraints often force educators to set $k$ to small values. Random Forests consistently outperform their counterparts across all $k$ for both districts A and B. The precision of other algorithms, however, varies with $k$. For instance, we can observe that Logistic Regression has lower precision compared to the decision tree when $k \leq 150$ on district B. Beyond this threshold, Logistic Regression has a higher precision than the decision tree.

Figure 5.4 shows the recall at top $k$ curves for both districts. Again, Random Forests outperforms all other models for all values of $k$. It can be seen that there is a higher variation in the recall values of algorithms in district B compared to district A. Further, Support Vector Machines exhibit consistently low recall in district B. The performance of all other algorithms seem to depend on the threshold $k$.

5.5 Interpreting Classifier Output

While the construction of models that can precisely identify students at risk is an important step to the design of early warning systems, it is equally important to analyze the output produced by these algorithms to make sure it aligns with the prior knowledge and/or findings of educators. In this section, we study in detail:

- How to identify features which are heavily used by algorithms?
- How to characterize patterns of mistakes made by algorithms?
- How can we compare and contrast algorithms based on the risk score estimates they produce?

Each of these aspects allow us to obtain a better understanding of the model behavior.
Figure 5.3. Precision across instances belonging to the top \( k \) highest risk score group (for varying values of \( k \)).

Figure 5.4. Recall across instances belonging to the top \( k \) highest risk score group (for varying values of \( k \)).
5.5.1 Feature Importances

In order to ensure that the output of the prediction models can be converted into actionable insights, it is essential to understand which factors contribute most heavily to the predictions. To answer this question, we make use of a variety of feature selection techniques that can be used to rank features according to their level of importance.

The setup we use to evaluate feature importances is similar to that previously described in section 5.4. We specifically chose to threshold our datasets at the end of 8th grade, as that time stamp marks the students’ transition into high school, and has been shown to be an especially opportune moment for targeted interventions [13].

The approaches that we use to compute feature importances are strictly dependent on the algorithms being used. We compute feature importances using both Gini Index (GI) and Information Gain (IG) for Decision Trees[99]. In the case of Random Forest, we consider feature importance to be the ratio of the number of instances routed to any decision tree in the ensemble that contains that feature, over the total number of instances in the training set. AdaBoost simply averages the feature importances provided by its base-level classifier – CART decision tree with maximum depth of 1 – over all iterations. For our Logistic Regression and SVM models, feature importances were simply considered to be the absolute values of each feature’s coefficient.

As before, we ran each classification model 100 times and subsequently averaged the importance scores given for each feature at each iteration. Based on the absolute values of these final importance scores, we then ranked all n features such that a rank of 1 corresponded to the feature with highest importance based on that particular classifier or metric. Figure 5.5 illustrates top ranked features by various algorithms. Table 5.2 lists the top 5 features used by each of the algorithms.
### Table 5.2

**TOP 5 FEATURES IN DISTRICTS A AND B**

<table>
<thead>
<tr>
<th>Rank</th>
<th>RF</th>
<th>AB</th>
<th>LR</th>
<th>SVM</th>
<th>GI</th>
<th>IG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Q4GPA_08</td>
<td>Q4GPA_08</td>
<td>Gender_07=Male</td>
<td>Gender_07=Male</td>
<td>Q4GPA_08</td>
<td>Q4GPA_08</td>
</tr>
<tr>
<td>2</td>
<td>Q3GPA_08</td>
<td>MPS_08</td>
<td>Gender_07=Female</td>
<td>Gender_07=Male</td>
<td>MAPR_08</td>
<td>Abs_Rate_08</td>
</tr>
<tr>
<td>3</td>
<td>Q1GPA_08</td>
<td>MAPR_08</td>
<td>Gender_06=Female</td>
<td>Gender_06=Female</td>
<td>Abs_Rate_08</td>
<td>MAPR_08</td>
</tr>
<tr>
<td>4</td>
<td>MAPR_08</td>
<td>Abs_Rate_08</td>
<td>Abs_Rate_08</td>
<td>Abs_Rate_08</td>
<td>Q1GPA_08</td>
<td>Abs_Rate_07</td>
</tr>
<tr>
<td>5</td>
<td>MPS_08</td>
<td>Q4GPA_06</td>
<td>Gender_06=Male</td>
<td>MPS_08</td>
<td>MAPR_06</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>RF</th>
<th>AB</th>
<th>LR</th>
<th>SVM</th>
<th>GI</th>
<th>IG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GPA_08</td>
<td>GPA_08</td>
<td>GPA_Science_08</td>
<td>School_Code=317</td>
<td>GPA_08</td>
<td>GPA_08</td>
</tr>
<tr>
<td>2</td>
<td>GPA_ENG_08</td>
<td>Days_Abs_08</td>
<td>Math_Credits_08</td>
<td>GPA_SocSci_08</td>
<td>GPA_Science_08</td>
<td>GPA_SocSci_08</td>
</tr>
<tr>
<td>3</td>
<td>GPA_SocSci_08</td>
<td>Num_Marks_08</td>
<td>GPA_SocSci_08</td>
<td>GPA_Math_08</td>
<td>Exc_Abs_08</td>
<td>Exc_Abs_08</td>
</tr>
<tr>
<td>4</td>
<td>GPA_Math_08</td>
<td>GPA_Science_08</td>
<td>School_Code=317</td>
<td>GPA_Science_08</td>
<td>Num_Marks_08</td>
<td>GPA_ENG_08</td>
</tr>
<tr>
<td>5</td>
<td>GPA_Science_08</td>
<td>Has_Disability_08=Y</td>
<td>Has_Disability_08=N</td>
<td>School_Code_08=315</td>
<td>EDS_08=T</td>
<td>School_Code_08=320</td>
</tr>
</tbody>
</table>

Here, GI stands for gini index and IG corresponds to information gain.
It can be inferred from Table 5.2 and Figure 5.5 that GPA at 8th grade is highly ranked across the majority of the approaches, indicating that academic performance at that particular time stamp is predictive of on-time high school graduation. Curiously, gender was highly ranked by our Logistic Regression and SVM methods for one of the cohorts. In addition to GPA, absence rates at 8th grade also show up as predominant features for both districts A and B. It is also interesting to note that some of the algorithms rank features such as economically disadvantaged (EDS) and disability flags highly.

Figure 5.5. Feature ranking by various algorithms. GPA_08 is ranked consistently high in both the districts by most of the models.
5.5.2 Characterizing Prediction Mistakes

School district administrators and educators are often interested in understanding the patterns of mistakes made by algorithms, which in turn helps them decide whether to use that model. For instance, if an algorithm is misclassifying certain kinds of students, and educators consider such patterns of misclassifications unacceptable, then they can choose not to use it in spite of the fact that the algorithm might be achieving a high precision and recall.

In order to identify such patterns for any given classification model, we use a simple technique involving frequent itemset extraction. Below is a description of the technique:

1. Identify all frequent patterns in the data using the FP-growth technique. A frequent pattern is a combination of (attribute, relation, value) tuples which occur very frequently in the entire dataset. For example, if the pattern GPA > 2.0 and Abs_Rate <= 0.1 holds true for about 80% of the students, then it can be considered a frequent pattern.

2. Rank students based on risk score estimates from the classification model. The predicted value of no_grad is 1 for the top k students from this list and 0 for others.

3. Create a new field called mistake. Set the value of this field to 1 for those data points where the prediction of the classification model does not match ground truth, otherwise set it to 0.

4. For each frequent pattern detected in Step 1, compute the probability of mistake. This can be done by iterating over all datapoints for which the pattern holds true, and computing the fraction of these datapoints where the mistake field is set to 1.

5. Sort the patterns based on their probability of mistake (high to low) and pick the top R patterns as mistake patterns.

The above procedure helped us identify several interesting mistake patterns for various algorithms. To illustrate that outcome, we present the patterns for two of these - Random Forest and Decision Trees – in Table 5.3. It can be seen that the models are making mistakes when a student has a high GPA and a high absence rate/tardiness, or when a student has a low GPA and low absence rate/tardiness. It
is also interesting to note that the Adaboost model is less accurate with respect to students who are economically disadvantaged but do well in Math and Science. This demonstrates that classification models are prone to making mistakes particularly on those data points where certain aspects of students are positive and others are negative. We found similar patterns with most other algorithms.

5.5.3 Comparing Classifier Predictions

Our discussions with school districts revealed that educators place noticeable importance on the exploratory aspects of predictive models. When we present educators with a suite of algorithms, they are keen on understanding the differences in rank orderings produced by each. Here, we address the question: How similar or dissimilar are the rank orderings produced by any two given models? This question can be answered by computing rank correlation metrics such as Spearman rank correlation coefficient, Kendall’s Tau, and Goodman and Kruskal’s gamma [61] for every pair of algorithms. While this is a perfectly reasonable strategy, recall that educators are typically interested in understanding all the metrics as a function of $k$ (the number of students that can be targeted using the available resources).

In order to measure the similarity of rank orderings for various values of $k$, we use Jaccard similarity metric. Given two sets $A$ and $B$, Jaccard similarity is the ratio of the number of elements in the intersection of $A$ and $B$ to the number of elements in the union of $A$ and $B$. The higher the value of Jaccard similarity, the more similar the sets. For a given $k$, all the algorithms return a set of $k$ students who are likely to not graduate on time based on the risk scores they produce. Similarity between rank orderings of algorithms can now be estimated by computing the Jaccard similarity metric between the set of $k$ students returned by various algorithms (for multiple values of $k$).
### TABLE 5.3

**CLASSIFIER MISTAKE PATTERNS**

<table>
<thead>
<tr>
<th>Model</th>
<th>Mistake Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forests</td>
<td>If Q4GPA_08 &gt; 3.0 and Abs_Rate_08 &gt; 0.3 and Tardy_08 &gt; 0.4, then Mistake&lt;br&gt; If Q4GPA_07 &gt; 3.0 and Abs_Rate_07 &gt; 0.4 and Tardy_07 &gt; 0.3, then Mistake</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>If Q4GPA_08 ≤ 2.0 and Q1GPA_08 ≤ 2.0 and Abs_Rate_08 ≤ -0.2 and Abs_Rate_07 ≤ -0.1, then Mistake&lt;br&gt; If Gender_07 = Female and Q4GPA_08 ≤ 2.0 and Abs_Rate_08 ≤ -0.1, then Mistake</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>If GPA_08 ≤ 2.0 and Tardy_08 ≤ -0.1 and Days_Abs_08 ≤ -0.2, then Mistake&lt;br&gt; If EDS_08 = True and GPA_Math_08 &gt; 3.0 and GPA_Science_08 &gt; 2.0, then Mistake</td>
</tr>
<tr>
<td>Adaboost</td>
<td>If Days_Abs_08 &gt; 0.5 and GPA_Science_08 &gt; 3.0, then Mistake&lt;br&gt; If Has_Disability_08 = True and GPA_Science_08 &gt; 3.0 and Days_Abs_08 &gt; 0.2, then Mistake</td>
</tr>
</tbody>
</table>

All the continuous variables except gpas are standardized to unit normal distribution. A positive value for such variables indicates above average and a negative value indicates below average.
Figure 5.6 shows the Jaccard similarity values that we computed for every pair of algorithms at various values of \( k \) for districts A and B respectively. It can be seen that Logistic Regression and SVM are highly similar for all values of \( k \) in District A. Furthermore, the ranking produced by most models is not similar to Decision Trees (RF-DT, AB-DT, LR-DT, SVM-DT curves). In the case of District B, there are interesting variations in the similarity between algorithms as \( k \) changes. For small values of \( k \), AdaBoost and Decision Trees produce similar sets. However, as \( k \) increases, Random Forest and AdaBoost appear to be the most similar algorithms. Random Forest and Logistic Regression also produce similar sets of students for various values of \( k \). Lastly, we see that SVM is the most dissimilar algorithm in District B (RF-SVM, AB-SVM, LR-SVM, SVM-DT curves).

This analysis helps school districts in understanding which algorithms to retain in their suite and which ones to discard. For instance, if they find that two algorithms are consistently similar in the rankings they produce, they may choose to retain just one of these algorithms based on parameters such as ease-of-use, computational efficiency, etc.

Next, we focus on the importance of predicting risk at early stages. We describe in detail an evaluation procedure that helps us determine if an algorithm is able to predict student risk at early stages.

5.6 Evaluating Classifiers for Early Predictions

Beyond being able to accurately identify students who are at risk of not graduating on time, it is important to make these predictions early so that educators and administrators have enough time to intervene and guide students back on track. In addition, our interactions with school districts revealed that once a student is retained in a grade, it becomes much harder to ensure timely graduation. Therefore, it is important to identify a student who is at risk before he/she starts failing grades
and/or drops out. In this section, we discuss evaluation procedures which help us determine if an algorithm is making timely predictions.

**Predicting risk early:** Here, we address the questions: *How precise is any given model at the earliest grade for which we have data? How does this performance change over time?* These questions can be answered by examining the performance of the models across all grades. The metrics that we use to evaluate the performance of our models are: `precision_at_top_k` and `recall_at_top_k`. Working with our school district partners revealed that a majority of school districts can afford resources to assist at least 5% of their student population. Therefore, we set $k$ to 5% of the student population.

![Diagram](image.png)

Figure 5.6. Jaccard Similarity of students at-risk for various algorithms
population for each of the districts.

We evaluate the performance of the models across all grades. Figure 5.7 depicts the precision at top 5% for each of the algorithms on districts A and B respectively. Random Forest consistently outperforms all other models in both the districts. In the case of District A, the performance improves steadily from 6th to 11th grade and then plateaus. AdaBoost and Decision Tree algorithms exhibit poor performance compared to other models across all grade levels for district A. The performance of SVM is consistently poor throughout all grades for district B. The corresponding recall curves (omitted due to space constraints) show similar patterns.

**Identifying risk before off-track:** Another important requirement in this setting is for a model to be able to identify students who are at risk of not graduating on time even before the student begins to fail grades and/or drops out. It is ideal to provide interventions to students before either of these undesired outcomes materialize, as opposed to taking a more reactive approach. A student can be categorized as off-track if he or she is retained (or drops out) at the end of a given grade. An ideal algorithm should be able to predict risk even before students go off-track.

Here, we investigate if our models succeed in identifying students before they go off-track. In order to do determine this, we use a metric called identification before off-track. This metric is a ratio of the number of students who were identified to be at risk before off-track to the total number of students who failed to graduate on time. For instance, if there are 100 students in the entire dataset who failed to graduate on time, and a given algorithm identifies 70 of these students as at-risk before they fail a grade or drop out, then the value of identification before off-track is 0.7. The higher the value of this metric, the better the algorithm is at diagnosing risk before an undesirable outcome takes place. Note that we exclude all of those students who graduate in a timely manner from this calculation.
Figure 5.7. Precision by grade at top K

Figure 5.8. Identification before Off-track
Figure 5.8 shows the *identification to off-track* metric values across varying values of $k$ for districts A and B respectively. The findings here match our earlier results, with Random Forests again outperforming all other models for both districts. While Decision Tree exhibits poor performance on district A, SVMs appear to be a weaker when district B data is utilized.

5.7 Conclusion

In this chapter, we outlined an extensive framework that uses machine learning approaches to identify students who are at risk of not graduating high school on time. The work described in this chapter was done in collaboration with two school districts in the US (with combined enrollment of around 200,000 students) and is aimed at giving them (as well as other schools) proactive tools that are designed for their needs, and to help them identify and prioritize students who are at risk of adverse academic outcomes. Although the work in this chapter is limited to predicting students who are likely to not finish high school on time, we believe that the framework (problem formulation, feature extraction process, classifiers, and evaluation criteria) applies and generalizes to other adverse academic outcomes as well, such as not applying to college, or undermatching [57]. Our hope is that as school districts see examples of work such as this coming from their peer institutions, they become more knowledgeable, motivated, and trained to use data-driven approaches and are able to use their resources more effectively to improve educational outcomes for their students.

5.8 Acknowledgments

The work described in this chapter was done as part of (and partially supported by) the 2014 Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship at the University of Chicago, with contributions by all following peers: Himabindu Lakkaraju, Carl Shan, David Miller, Nasir Bhanpuri and Rayid Ghani.
6.1 Introduction

The engineering education community is well aware of the current problems in retaining students in engineering. Many reports indicate that less than half of the students who start in engineering will obtain a degree in an engineering major [17]. Therefore, many institutions have dedicated significant time to identifying students at-risk of leaving engineering and implementing retention strategies. Often, these identification methods include statistical methods centered on academic performance or demographics; however, current theories on retention include many more attributes that contribute to student attrition [24, 30, 113]. While technical knowledge may play a part in retention, studies have included attitude based variables, which increase the likelihood of identifying at-risk students. Attitude based variables may include student-student interactions, dorm life, personality type, and how the degree is meeting their expectations. Some have even claimed that understanding students attitude about engineering and the university climate provides the highest correlation to retention [95, 116]. Because it is already well documented elsewhere, this study will not consider traditional academic data in assessing student retention. Instead, we propose using electronic portfolios, or ePortfolios, as a method for assessing these more difficult to understand attitudes with a future goal of identifying at-risk students in a timely fashion.
While student portfolios are not new methods of documentation for majors such as architecture or art, they have been unexplored by the vast number of engineering schools. Although some have folded them in as an assessment of communications skills [59, 89], few have used portfolios as a means to measure student engagement and their learning process throughout their degree. Previous studies have proven the efficacy of the ePortfolio format as a method of engagement. Many universities have implemented ePortfolios to enhance engagement and measure impacts of programs [68, 103]. ePortfolios have even been shown to create shared ownership of student learning between student and instructor by Chen and Black [33], and their traditional counterpart has been used as a tool in advising [8]. We will use ePortfolios to serve as a larger space for students to record their achievements in engineering, starting with specific projects from the first-year engineering course and later being used for advising within their engineering majors. Although the course will set a general template, students are encouraged to treat the ePortfolio as their own creative space where they can collect evidence of accomplishments and connect it to their learning and goals as a student. While students are encouraged to continue the use of their ePortfolios throughout their university careers, this paper will focus solely on students interactions with their ePortfolio during their first-year.

In the two semester Introduction to Engineering course sequence, all students are expected to create an ePortfolio following a course designated template. Each of the last 3 years has included changes to the template based on instructor and student evaluation of assignments. The course instructors have reflective exercises for future advisors to use as well as exploration activities to aid students in major discernment. The current ePortfolio template included three main sections, which were each updated throughout the course sequence:

- 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). Students...
are expected to create goals for each skills area and reflect on their current progress towards those goals. Future advisors could use this material to help identify opportunities for their advisees.

- Project Updates – Required updates following the completion of each course required project. Minimally, students were asked to include a picture of their project and a reflection on skills developed through the project. This space is expected to act as evidence of how students are meeting their goals outlined in the advising reflections. For instance, successfully completing a group project may be part of the evidence used to describe their development in working well within a team.

- Engineering Exploration – Required reflections after attendance at eight engineering related events that took place outside of the course. These events were a recent addition to the course sequence, allowing students more in-class time to discern their major choice. Events included seminars, engineering student group meetings, professional development activities, etc. and were executed by various groups within the university, some during class time, but most throughout evening hours. Reflections were expected to capture why they attended the event and how that contributed to their growth as a student.

In addition, students are encouraged to add other details and information to their ePortfolio that would be appropriate for a professional setting. Students are especially urged to include other course projects, internships, or personal accomplishments that they feel appropriate.

While students are expected to complete all ePortfolio assignments as a portion of the course grade, they are graded largely for completion. Student engagement and effort towards assignments is widely variable; therefore, by exploring student responses and interaction with the ePortfolio system, student engagement may be a measurable quantity.

6.2 Setting

Our work in the higher education space was born as group effort by the College of Engineering and First Year of Studies at the University of Notre Dame. To provide the reader with some background information, the following sections describe
the cohort of students we analyzed, as well as the overall context of how electronic portfolios have been made part of the curriculum.

6.2.1 The College of Engineering

The University of Notre Dame is a medium sized, Midwestern, private institution with a traditional student composition. 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 into the university as a whole, there are no 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.

6.2.2 Electronic Portfolios

Below we provide the reader with more information on electronic portfolios, how this tool has been made part of the curriculum for first year engineering students at the University of Notre Dame, and how we began mining some of the data produced by it to better estimate how enganged these students were to the program.

*ePortfolios for Engagement*

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 [21]. ePortfolios capture and document students’ learning and engagement through their reflection, rationale building, and/or planning. Chen and Black [33], found that 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. [8] found the use of portfolios 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 [62], La Guardia Community College [46], University of Barcelona [77], Spelman College [98],
Clemson [102], Penn State and Florida State Universities [119] 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 [66].

The ePortfolio platform of our choice is supported by Digication [4] 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.

Realizing the importance of having a deep understanding of how students interact with and make use of their electronic portfolios, we worked alongside Digication to develop an automated pipeline for data collection that has since been deployed. We collect student-level data on a daily basis, and while the work described in this dissertation made use of aggregated datasets collected at the end of the 2012 Fall semester, in the future we will be able to scale our models for predicting early signs of risk to much more actionable time stamps.

**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 (illustrated in Figure 6.1), 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.

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.

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.

Goodrich et al. [51] describes in detail how electronic portfolios have been incorporated to our Introduction to Engineering curriculum, and how they can be po-
entially used to measure student engagement. We contrasted instructor-generated ratings for student engagement with metrics extracted from those students’ ePortfolios, and showed that for that particular cohort, the engagement estimates provided by the ePortfolio variables were significantly more strongly correlated to retention outcomes than were the instructor ratings.

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.

*ePortfolio Use Across the First Year*

At the time of this study, the University of Notre Dame was two years into a larger University ePortfolio initiative where half of the students on campus were using ePortfolios in at least one capacity. The College of First Year Studies had launched its own initiative for all first year students, to flip and enhance the advising process of their one-on-one sessions with the ePortfolio. Using the blended advising model [14], the Advising ePortfolio was used to pre-engage students by asking them to plan and list goals in their ePortfolio before their face-to-face advising session. This would lead to a more deeply engaged one-on-one advising interaction because the students came prepared. Then, advisors and student would have a platform to re-engage and review progress and growth over the year. For a detail overview of this University-wide ePortfolio initiative see [15].

In addition, a small cohort of 12 first generation students (from a family in which no parent or guardian has earned a baccalaureate degree) that were intended Engineering students were also enrolled in a 1 credit Independent Self Study Advising Seminar as part of an invited scholars program for underrepresentend students.

While Chapter 7 will provide more details on our datasets and describe each of the features we analyze, a preliminary case study illustrated by Figure 6.2 showed
that this select group of students who were more exposed to electronic portfolios as part of the previously mentioned enhanced academic program exhibited markedly higher levels of engagement using that tool, and more importantly, was retained in its entirety.

![Figure 6.2: The effect of ePortfolio engagement on first semester retention](image)

Figure 6.2: The effect of ePortfolio engagement on first semester retention

**ePortfolio Engagement as an Early Indicator of Retention**

As mentioned earlier, factors providing early indications that a student may be at risk can come from a variety of categories. Data that has been aggregated over time can offer insights as to which sub-populations of a student body are more likely to need closer attention, whereas individual student-level performance data can be used to flag students that are beginning to diverge from their optimal path to success.

An important task is, then, to decide which information to track for each student so that we can most efficiently generate early risk warnings when these are warranted. As we claimed earlier, placing disproportional focus on academic performance data can result in warning systems that may fail to identify students that are losing interest and disengaging from school when these shortcomings are not reflected on that
student’s performance.

To illustrate such issue, take for instance the performance data summarized in Figure 6.3. There we can see how the first semester cumulative GPAs for both retained and not retained students in our dataset were distributed. While it does appear that the retained students perform, on average, slightly better, we note that such disparity is not substantial. Furthermore, we can see that a fairly sizable number of students that were not retained concluded their first semesters with very high GPAs (e.g., > 3.0), which is a reasonable indication that they would have been likely to succeed as engineering students had they chosen to remain enrolled in the program. A similar observation can be made when we look at the course grades for Introduction to Engineering (EG 111) alone.

With that premise in mind, we investigated the viability of using ePortfolios as a proxy to measuring student engagement. Again, splitting the same cohort of students depicted in Figure 6.3 into the subgroups of retained and not retained, we
see in Figure 6.4 that on average, students on each of these two groups appear to place discernibly different amounts of time and energy when interacting with their portfolios.

Figure 6.4: Average breakdown of ePortfolio engagement activity by cohort of first-year engineering students.

While Goodrich et al. [51] presents a more elaborate description of this information and a comparison between these observed values and similar engagement estimates generated by course instructors, we highlight that the distribution of values for each of the three features illustrated in Figure 6.4 (i.e., number of logins during Fall semester, number of evidences submitted through ePortfolio and number of ePortfolio page hits received) with respect to retained and not retained students is statistically significantly different, with p-values displayed in Table 6.1.

6.3 Methods

In order to assess student engagement, a multi-part assessment was conducted. Students completed ePortfolio assignments throughout the fall and spring semesters; however, this study focused on Fall 2012 and Fall 2013 data.

Student ePortfolio assignments from the Fall 2013 semester were read and graded
TABLE 6.1

P-VALUES OF OBSERVED DISTRIBUTIONS FOR THREE EPORTFOLIO FEATURES WITH RESPECT TO RETENTION OUTCOMES

<table>
<thead>
<tr>
<th>ePortfolio Feature</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of logins</td>
<td>&lt; 0.005</td>
</tr>
<tr>
<td>Number of evidences submitted</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Number of page hits</td>
<td>&lt; 0.005</td>
</tr>
</tbody>
</table>

by instructors of the first-year course. A common rubric was shared between all instructors for three mid-semester assignments, assessing students based on perceived interested and engagement with the course. First, students were expected to attend and reflect on two events outside of the engineering course by the end of October. Both engineering exploration assignments were assessed for each student on the four point scale described in Table 6.2

Additionally, a project update assignment was assessed in two parts – students’ reflections on their project and the evidence they provided. Evidence was considered to be any item added to the ePortfolio in addition to the written reflection (i.e.: pictures, video, coding segments, and project reports). Minimally, students were expected to write a 2-3 sentence reflection and include a picture of their finalized project. The rubric used is included in Table 6.3

Finally, this chapter will explore the some initial results of using a data mining tool with the ePortfolio system to act as a predictor of student retention. While some results are presented in this chapter, we will cover the experimentation details and
### TABLE 6.2

**RUBRIC FOR STUDENT INTEREST IN ENGINEERING BASED ON ENGINEERING EXPLORATION REFLECTIONS**

<table>
<thead>
<tr>
<th>Rating</th>
<th>Interest Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No rating could be discerned by reflection.</td>
</tr>
<tr>
<td></td>
<td>Interest seems low and easily dissuaded from continuing in engineering. May state interest in major outside of engineering. Possibly mentions things unrelated to engineering as reasoning.</td>
</tr>
<tr>
<td>1</td>
<td>Does show some preference for engineering but may also mention other major possibilities. May indicate that he/she didn’t know what engineering was.</td>
</tr>
<tr>
<td></td>
<td>Relates exploration towards engineering and possibly even specific fields.</td>
</tr>
<tr>
<td>2</td>
<td>May emphasize determining what type of engineering he/she wants to major in. Student seems likely to continue on in engineering.</td>
</tr>
</tbody>
</table>

### TABLE 6.3

**RUBRIC FOR STUDENT INTEREST IN ENGINEERING BASED ON ENGINEERING EXPLORATION REFLECTIONS**

<table>
<thead>
<tr>
<th>Rating</th>
<th>Reflection</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Missing any reflection or short and shallow</td>
<td>No evidence or single required picture only</td>
</tr>
<tr>
<td>2</td>
<td>Some reflection given and apparent time spent on activity</td>
<td>Some pictures or diagrams included. Obvious effort shown</td>
</tr>
<tr>
<td>3</td>
<td>Reflection deep and detailed – effort and interest clear</td>
<td>High amount of evidence shown, often including multimedia such as videos, pictures and programming segments</td>
</tr>
</tbody>
</table>
a more strict evaluation in Chapter 7. While several other ePortfolio features and student academic features are considered in that chapter, only the most important features of the ePortfolio tool will be detailed here. For this work, we focused on the number of times a student has logged into the ePortfolio system, pieces of evidence included, and ePortfolio hits. Note that ePortfolio hits are the number of visits to a student’s ePortfolio pages which may originate from the ePortfolio owner or external visitors. At this time, we are unable to discriminate between these two types of visits; however, we suspect that ePortfolio hits are largely coming from the owner of the ePortfolio as he/she creates content, edits pages, and makes assignment corrections.

6.4 Results

6.4.1 Instructor Based Assessment

The primary purpose for starting this study was to determine if anecdotal observations by course instructors correlated to the retention of students. In many cases, student ePortfolio assignments appeared indicative of student retention. In this formalized examination, each of the eight instructors assessed their own small group sections (30-35 students in each section, 2 sections per instructor) resulting in a total of over 400 students rated. The authors note that ratings bias may be present from having instructors act as the only raters for each student. Each instructor was given the same rubric and an initial training session for rating student responses. Any bias present is believed to minimally impact the rating results and overall conclusions of this study. Three examples below depict how the rubric from Table 1 correlated to statements made in the ePortfolio. As shown in these three examples, ratings were often determined by specific phrases that indicated student major choices:

- **Rating of 3** (Attended a Society of Women Engineers Event) – Apart from discerning between the different disciplines, the girls also helped in giving advice
about what our engineering experience at Notre Dame will be like and how to make it successful. Overall, this event was extremely helpful in strengthening my confidence that chemical engineering, even just engineering in general, is a good major for me.

- **Rating of 2** (Event held for Integrated Business and Engineering Minor) – I realize there is more that I can do with an engineering major than I previously knew. Although I am not totally sure that I want to major in engineering, I am confident that engineering will not inhibit my ability to change fields, but open me to new job opportunities.

- **Rating of 1** (Attended lecture by Dean of college of Engineering) – This event initially peaked my interest due to the fact that I was unsure of whether I wanted to pursue engineering further or not. At the conclusion of the Dean’s lecture, I had a rejuvenated sense of what I want to do with my life...I want to pursue business like my father, not engineering like my brother.

Additionally, project reflections were used to gauge student engagement with the first-year engineering course. Table 6.4 below reports the mean rubric scores for all students enrolled in the Fall 2013 course as well as those that have continued into the Spring 2014 engineering course (indicated as “Retained in Engineering”) and those who left the program sequence after the fall course (indicated as “Left Engineering”).

<table>
<thead>
<tr>
<th></th>
<th>All Students</th>
<th>Retained in Engineering</th>
<th>Left Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exploration 1</td>
<td>2.37</td>
<td>2.38</td>
<td>2.06</td>
</tr>
<tr>
<td>Exploration 2</td>
<td>2.47</td>
<td>2.50</td>
<td>2.09</td>
</tr>
<tr>
<td>Project Reflection</td>
<td>2.10</td>
<td>2.11</td>
<td>1.93</td>
</tr>
<tr>
<td>Project Evidence</td>
<td>1.80</td>
<td>1.82</td>
<td>1.63</td>
</tr>
</tbody>
</table>
It should be noted that the retention rate from the first semester of this course is 85%, resulting in very similar populations for the “All Students” and “Retained in Engineering” groups. For all assignments evaluated, the full student population and the retained population scoring is nearly identical. In contrast, those who left engineering had, on average, lower scores in each of the four assignments evaluated. However, only Exploration 1 ($p < 0.1$) and Exploration 2 ($p < 0.05$) assignments indicated statistically significant differences between those who remained in engineering and those who left after the first semester.

6.4.2 Quantitative ePortfolio-based Assessment

A computational tool was used to analyze features from students’ ePortfolios following the completion of the 2012-2013 academic year. In Figure 6.4 (A), the number of times a student logged in to the ePortfolio system during the Fall 2012 semester was compared for three student groups: (1) students who remained in an engineering major into sophomore year, (2) students who dropped after the spring semester of their first-year, and (3) students who dropped after the fall semester of their first year.

There was no statistical difference between the first two student groups – those that remained in the engineering course sequence throughout their first-year. However, there was a statistically significant difference in the number of logins for a student that dropped after only one semester ($p < 0.005$).

Additionally, pieces of evidence submitted during the Fall 2012 semester were also compared. Because previous studies had revealed little difference between those students who left after the spring semester and those who were retained into sophomore year, these two groups were combined for future assessments into the group “Retained in Engineering”, as shown below in Figure 6.4 (B).

There is a statistically significant difference between the students who continue
in engineering and those who leave \((p < 0.05)\). Students who were retained in engineering had on average only one more piece of evidence than those who did not; however, requirements of ePortfolio assignments would play an important role in the pieces of evidence expected from a student.

Finally, Figure 6.4 (C) below shows the number of hits on each ePortfolio during the Fall 2012 semester. In this case, those that were retained had over five times as many hits as students who left engineering after the fall semester (670 hits compared to just 126 hits).

Again, there is a statistically significant difference between the students who remain in engineering at least through the spring semester and those who leave during or after their first semester in engineering \((p < 0.005)\). In fact, of all ePortfolio measures currently under study, the number of ePortfolio hits is the strongest indicator of student retention when looking at the end of the semester.

6.5 Discussion and Future Work

The instructor based assessment indicated that student interest in engineering can be determined through fairly simple reflective exercises. Reflections on course projects offered minimal distinguishable characteristics between students that leave and those that are retained. We believe this is a result of the group nature of these projects, where all students from a project group had the same evidence available to use in their ePortfolio at the time of the reflection. Reflective assignments that focused on out of classroom events had more distinguishable markers for retention. We believe these assignments contained better markers due to the reflection requirements. For the Engineering Exploration assignments, students were asked to include personalized reasons for attending events and the growth they experienced because of it. These types of reflections led more naturally to student responses that included their doubts, interests, and career aspirations.
With many universities experiencing a rise in class sizes, reading multiple reflections from every student in a timely manner is likely not a practical solution to increase retention, though these reflections are useful additions when advising and interacting with students. The cost in staff and faculty time would be immense and place a heavy burden on those within the course if this were the only means of identifying at-risk students. Instead, using a combination of a computational tool and instructor reading would provide with a much needed balance in the cost of faculty involvement.

Data mining has provided a few much needed indicators of student retention. Many of the features that appear to be clear markers here are tied to the course requirements. For instance, students had eight ePortfolio assignments over the course of the semester, two of which had opportunities to go back and make corrections. If students fulfilled just the course requirements, they would be expected to login 10 times where the average for students that drop is 12.69. Additionally, the course has only two assignments were pieces of evidence are required as part of the grade. Those who leave after one semester had an average of just 2.64 pieces of evidence. Perhaps unsurprisingly, being near the minimum requirements may be an indication that students are not engaged with the course, signaling a waning interest in engineering.

The strongest indicator of student retention though was number of hits on a student ePortfolio. Additionally, this large difference in number of hits cannot be directly correlated to expectations set from the course. While many of these hits may come from an instructor viewing the ePortfolio, it is not expected that instructor views would vary widely from student to student. Self-created hits, on the other hand, could vary greatly between students. Those who are more engaged with the course and their ePortfolio may spend more time documenting their projects and explorations. In addition, coming back to edit and update multiple ePortfolio pages would indicate that the student considers documenting and reflecting on their engi-
engineering experience an important part of their education. Most students also leave their ePortfolios open to other members of the university community. Therefore, if they are creating authentic, meaningful entries, other students and faculty may be more likely to visit and read their ePortfolio. Finally, students who are engaged with the course and their ePortfolio may use this to show off their work to friends, family, and possible employers. Overall, a high hit count on an ePortfolio would indicate a truly living document that is an authentic reflection of the student owner.

Collectively, we interpret the data mining results to indicate that ePortfolios can accurately measure interest and engagement; however, not all ePortfolio markers will be of equal utility. Both the course requirements and the student buy-in to the process will be important characteristics that need to be considered when identifying the relevant features at different universities. Currently, this identification has occurred after the semester ends, when it is likely too late to intervene with the students who choose to leave engineering, but future work will focus on similar analysis in a timely fashion for intervention. Using ePortfolio hits and logins, data which is easily attainable in real time, we can determine the markers of an at-risk student. By allowing the computational model to flag these students, the instructor’s time commitment for very thorough evaluation is decreased to only those students flagged by the data mining tool. This could provide faculty then with a chance to carefully assess why a student is choosing to leave engineering and help him/her determine if it is a premature choice (e.g. leaving because of one midterm grade) and intervene with strategies already in place at the university.
CHAPTER 7

ENGAGEMENT VS PERFORMANCE: USING ELECTRONIC PORTFOLIOS TO PREDICT FIRST SEMESTER ENGINEERING STUDENT PERSISTENCE

7.1 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 [107] 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 [44]. 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., [10, 30, 41, 43, 71, 73, 84, 111, 120–122]).

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 [18, 48, 71, 79, 92, 93, 117]. 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 students who despite performing below the expectations, remain enrolled. Figure 7.1 illustrates these two specific groups of students.

Figure 7.1. While dropout rates are often more prominent across low-performing students, and students with high performance tend to be retained, a couple of special cases exist and are highlighted here. Low performing/highly engaged students (quadrant I) are often retained. High performing/disengaged students (quadrant IV) may drop out.
In this chapter, we focus on remediying 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 7.1.

7.2 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.

7.2.1 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 7.1 lists and describes each feature available in our final dataset and Table 7.2 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 not retained). As described in [111], 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.

7.2.2 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) [99], gain ratio (GR) [100], 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 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 7.3 where the top 10 features ranked by each method are listed, and the highest correlated one is highlighted for each column.
<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adm Intent</td>
<td>Nominal</td>
<td>Intended college major as specified by the student in his/her application package</td>
</tr>
<tr>
<td>Adm Type</td>
<td>Nominal</td>
<td>Type of admission earned by student (e.g., early, regular, waiting list)</td>
</tr>
<tr>
<td>AP Credits</td>
<td>Numeric</td>
<td>Number of credits earned through AP courses taken prior to college enrollment</td>
</tr>
<tr>
<td>Dormitory</td>
<td>Nominal</td>
<td>Name of dorm where the student resides (note: all first-year students are required to live on campus)</td>
</tr>
<tr>
<td>EG 111 Grade</td>
<td>Nominal</td>
<td>Letter grade obtained in the introduction to engineering course</td>
</tr>
<tr>
<td>ePort Hits</td>
<td>Numeric</td>
<td>Hit count for the student’s ePortfolio pages during the fall semester</td>
</tr>
<tr>
<td>ePort Logins</td>
<td>Numeric</td>
<td>Number of times the student logged in to his/her ePortfolio account during the fall semester</td>
</tr>
<tr>
<td>ePort Subm</td>
<td>Numeric</td>
<td>Number of assignment submitted via ePortfolio during the fall semester</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Nominal</td>
<td>The student’s self-declared ethnicity</td>
</tr>
<tr>
<td>First Gen</td>
<td>Binary</td>
<td>A flag to denote first-generation college students</td>
</tr>
<tr>
<td>FY Intent</td>
<td>Nominal</td>
<td>Intended college major as specified immediately prior to the beginning of the fall semester</td>
</tr>
<tr>
<td>Gender</td>
<td>Binary</td>
<td>The student’s gender</td>
</tr>
<tr>
<td>Income Group</td>
<td>Numeric</td>
<td>A numeric value ranging from 1-21, each corresponding to a different income group segment</td>
</tr>
<tr>
<td>SAT Comb</td>
<td>Numeric</td>
<td>Combined SAT scores</td>
</tr>
<tr>
<td>SAT Math</td>
<td>Numeric</td>
<td>SAT score for the math portion of the test</td>
</tr>
<tr>
<td>SAT Verbal</td>
<td>Numeric</td>
<td>SAT score for the verbal portion of the test</td>
</tr>
<tr>
<td>Sem 1 GPA</td>
<td>Numeric</td>
<td>The student’s overall GPA at the end of the fall semester</td>
</tr>
<tr>
<td>Retained</td>
<td>Binary</td>
<td>A flag identifying students that dropped out immediately after the fall semester</td>
</tr>
</tbody>
</table>
Several interesting observations can be derived from these results. First, we emphasize that all but one method reported \textit{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
TABLE 7.3

FEATURE RANKINGS PROVIDED BY MULTIPLE METHODS

<table>
<thead>
<tr>
<th>Feature</th>
<th>IG</th>
<th>GR</th>
<th>CS</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adm Intent</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>SAT Math</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>SAT Verbal</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SAT Comb</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>AP Credits</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>FY Intent</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>EG 111 Grade</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Sem 1 GPA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Adm Type</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gender</td>
<td>10</td>
<td>9</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Income Group</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>7</td>
</tr>
<tr>
<td>First Gen</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dormitory</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>ePort Logins</td>
<td>6</td>
<td>3</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>ePort Subm</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>ePort Hits</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Highlighted for each method are the top 10 features with the most important shown in bold.
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 [84, 121], 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 below:

- **all-academic**: This subset contained all academic and demographics features listed in Table 7.2.

- **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.
7.3 Methodology

For this study, we selected a range of classification methods that have been previously utilized in this particular domain\footnote{A lengthier description of which can be found in Chapter \ref{chapter:3.1.2}} 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.

Further, 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 metrics such as those described in Chapter \ref{chapter:3.2}.

7.4 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 6.1 was evaluated as they were used to perform prediction on each of the four available datasets. Table \ref{table:7.4} displays the results of each individual experiment in terms of the prediction accuracy for the negative class instances (i.e., the ratio of retained students that were correctly labeled as retained), the prediction accuracy for the positive class instances (i.e., the ratio of not retained students correctly classified as not 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.

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 not retained (false positives), and actual not-retained students could be mistakenly labeled as retained (false negatives). While some previous work (e.g., [42]) 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 leaving can be much more costly than incorrectly labeling someone as an early leaver.

In Table 7.4 we can see that predictions based only on academic performance and demographic data achieve a maximum \( \text{acc+} \) of 27.5\% when the all-academic dataset is paired with a naive Bayes model. That corresponds to only 11 of the 48 not-retained 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 \textit{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 \( \text{acc-} \) values under the same setup, and could potentially be the preferred choice depending on the circumstances.

With respect to \( \text{acc+} \), the naive Bayes classifier outperformed the others for all but one dataset. We used its experimental results in Figure 7.2 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.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>Acc+</th>
<th>Acc-</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NB</td>
<td>0.275</td>
<td>0.902</td>
<td>0.842</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.083</td>
<td>0.930</td>
<td>0.833</td>
</tr>
<tr>
<td>all-academic</td>
<td>LR</td>
<td>0.104</td>
<td>0.900</td>
<td>0.803</td>
</tr>
<tr>
<td></td>
<td>HT</td>
<td>0.083</td>
<td>0.884</td>
<td>0.792</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.000</td>
<td>0.987</td>
<td>0.874</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.167</td>
<td>0.954</td>
<td>0.864</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.042</td>
<td>0.949</td>
<td>0.845</td>
</tr>
<tr>
<td>top-academic</td>
<td>LR</td>
<td>0.000</td>
<td>0.981</td>
<td>0.869</td>
</tr>
<tr>
<td></td>
<td>HT</td>
<td>0.250</td>
<td>0.881</td>
<td>0.809</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.104</td>
<td>0.892</td>
<td>0.802</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.833</td>
<td>0.879</td>
<td>0.874</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.771</td>
<td>0.970</td>
<td>0.947</td>
</tr>
<tr>
<td>all-engagement</td>
<td>LR</td>
<td>0.771</td>
<td>0.978</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>HT</td>
<td>0.771</td>
<td>0.962</td>
<td>0.940</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.771</td>
<td>0.970</td>
<td>0.947</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.875</td>
<td>0.892</td>
<td>0.890</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.792</td>
<td>0.962</td>
<td>0.945</td>
</tr>
<tr>
<td>top-academic+engagement</td>
<td>LR</td>
<td>0.750</td>
<td>0.973</td>
<td>0.947</td>
</tr>
<tr>
<td></td>
<td>HT</td>
<td>0.771</td>
<td>0.965</td>
<td>0.943</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.750</td>
<td>0.965</td>
<td>0.940</td>
</tr>
</tbody>
</table>
Figure 7.2. Naive Bayes ROC and Precision-Recall curves for each subset of academic performance and engagement features.

The highest AUROC value\(^2\) (0.929) was obtained by the top-academic+engagement dataset, while all-academic performed worse with an AUROC of 0.654.

7.4.1 Improving the Detection of at-risk Students with Data Oversampling

As we have previously highlighted, the particular cohort of students that we utilized for our experiments had a very high retention rate of 88.5%. With such data imbalance, classification models can often learn hypotheses that are biased towards assigning the majority class label to new instances. One way to address this issue is to introduce, during the training phase of the models, “synthetic” examples of minority class instances. This can essentially decrease the rate of imbalance and it often lessens the bias towards the majority class.

Previous work has investigated the effect of various over-sampling techniques

\(^2\)A different logistic regression implementation using L1 regularization yielded slightly higher AUROC. See Table 7.4 for details.
on models designed to predict student attrition. One such method that was found to improve the overall performance was SMOTE [32], which works by generating synthetic instances by interpolating existing minority class occurrences with their k-nearest neighbors of the same class.

So as to evaluate any benefits of data oversampling in our context, we used the top-academic + engagement dataset previously described, and repeated the experiments while introducing synthetic instances of the not-retained class. Table 7.5 illustrates the performances of our models in terms of AUROC when varying amounts of synthetic minority instances are used. There, “100 SMOTE” refers to the addition of a synthetic set of minority instances that contains the same number of samples as the original set, “200 SMOTE” creates two new samples for each observed minority instance, and so on. As we can see, most models exhibit marginal improvements when SMOTEed instances are used to decrease the class imbalance during model training.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>No SMOTE</th>
<th>100 SMOTE</th>
<th>200 SMOTE</th>
<th>300 SMOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.923</td>
<td>0.923</td>
<td>0.924</td>
<td>0.926</td>
</tr>
<tr>
<td>DT</td>
<td>0.858</td>
<td><strong>0.866</strong></td>
<td>0.855</td>
<td>0.849</td>
</tr>
<tr>
<td>LR</td>
<td>0.968</td>
<td>0.969</td>
<td>0.970</td>
<td><strong>0.971</strong></td>
</tr>
<tr>
<td>RF</td>
<td>0.953</td>
<td><strong>0.960</strong></td>
<td>0.956</td>
<td>0.941</td>
</tr>
</tbody>
</table>
7.4.2 Alternative Model Evaluation

While the previous sections provided a detailed comparison between the performances obtained by each model with respect to a variety of traditionally used evaluation metrics, when considering the real-world requirements to implementing an early warning system wherein interventions are put in place to help students that are labeled as being at risk, it is often more logical to evaluate our predictive models differently.

Suppose for instance that a school is given a certain operational budget to deploy an intervention program for STEM students at risk of not being retained. Given this particular constraint, it might not be possible for the school administration to hold individual one-on-one interventions with all students at risk. Hence, they should ideally choose a predictive model that can most accurately identify which students are at highest risk.

The models described in section 3.1.2 can, in addition to providing a binary label to a new student being evaluated, also output an associated probability value between 0 and 1 that corresponds to that model’s estimate of the likelihood of this student not being retained. Based on a set threshold value in that same range, the model then decides if the student is to be labeled as likely to be retained or otherwise. Contextually, if a predictive model using a threshold of 0.5 generates a probability of 0.6 for student A and 0.99 for student B, they will both receive a “not retained” label. However, chances are that student B is the one requiring more immediate attention.

If we generate a probability value for each student in a given cohort, we can then create an overall sorted rank with respect to these values. In this setting, a preferred model would be one that can generate high probabilities for those students who are indeed at risk and lower probabilities for those who are on track to succeeding.

Next we compare the performance of four of our models in terms of how precise they are on the top 5 and 10% of the probability scores they generate. That is, if we
isolate the 5 and 10% of the students who receive the highest scores from each model, we would like to know what fraction of that group eventually left engineering. This metric is especially helpful in the context of early warning systems because when working with time and budgetary constraints, one will often need to choose a small subset of students on which to focus the appropriate intervention.

As we can see in Table 7.6, the models that include engagement features outperform those that only consider academic performance when we look at precisions on both top 5 and 10%. Further, we highlight that the results obtained by the logistic regression model that only used engagement features were the highest observed. In our specific scenario, the top 5% accounted for the 22 students whose probability scores were highest overall. Using a logistic regression (or decision tree) model coupled with the engagement dataset to select the 22 students at “highest risk” gives us a precision of 0.954, which means that 21 of the 22 selected students were in fact not retained.

When pairing at-risk students with the appropriate interventions based on flags generated by early warning systems, we might not necessarily give preference to models that can be highly accurate, but we should rather evaluate their ability to select those students that need the most immediate attention. Though we had previously shown that the Naive Bayes model is the most accurate in the aggregate, here we showed that other models might, in fact, provide a more appropriate solution from an operational standpoint.

7.5 Conclusion and Future Work

Given a course with an enrollment of 429 first semester first year Intro to Engineering students and an average retention rate of 85%, the challenge was to develop, train and test classification models to predict whether students would persist into the
**TABLE 7.6**

**PRECISION ON THE TOP 5 AND 10%**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>Top 5%</th>
<th>Top 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>all-academic</td>
<td>NB</td>
<td>0.400</td>
<td>0.268</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.200</td>
<td>0.243</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.400</td>
<td>0.341</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.350</td>
<td>0.292</td>
</tr>
<tr>
<td>top-academic</td>
<td>NB</td>
<td>0.300</td>
<td>0.268</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.250</td>
<td>0.244</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.400</td>
<td>0.268</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.150</td>
<td>0.170</td>
</tr>
<tr>
<td>all-engagement</td>
<td>NB</td>
<td>0.800</td>
<td>0.756</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.954</td>
<td>0.756</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.954</td>
<td>0.878</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.900</td>
<td>0.829</td>
</tr>
<tr>
<td>top-academic+engagement</td>
<td>NB</td>
<td>0.650</td>
<td>0.682</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.900</td>
<td>0.756</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.900</td>
<td>0.878</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.900</td>
<td>0.829</td>
</tr>
</tbody>
</table>
second semester beyond traditional measures of performance. How do you identify not just the high interest/low ability at risk students but also the low interest/high ability students? More specifically, we wanted to compare the performances of models that were based on traditional academic data (SAT scores, GPA, demographics, etc.) and models that were based on data extracted from ePortfolios (#logins, #hits, etc.). The basic premise of this being that while early warning systems that are based on academic performance data alone can be very good at finding students that are struggling in class and alerting their advisors, early signs of a possible decision to drop out of a certain academic path can be very difficult to find when you are dealing with students that do not display any changes in their class performance.

We investigated the feasibility of using ePortfolio data as a proxy to measuring student engagement and showed that these particular variables can be highly predictive of college retention outcomes. We described that while datasets that do not contain features that quantify 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 leave. 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. Our key findings included:

• Out of a set of several academic performance, demographics and ePortfolio features, the number of ePortfolio hits displayed, based on multiple metrics, the strongest correlation values to the outcome (student was retained / not retained).

• The performance of the prediction models that used only ePortfolio data was consistently better than that of models based on academic performance data alone.

• The best prediction results were obtained by using a subset of the data containing the following features: EG 111 grade, ePort logins, and ePort hits.
• Using only academic performance data and a leave-one-out cross validation setup, we were able to identify 11 of the 48 students who were not retained past their first semester (out of total of 429 students in the course). By adding the ePortfolio features, our model’s performance dramatically improved and we are able to correctly label 42 of the 48 students while incurring very minor losses in accuracy with regards to the retained group.

While the results presented in this paper were obtained using an aggregated dataset collected at the end of the fall semester of 2012, we have since begun collecting ePortfolio usage data on a daily basis. As we have shown for this particular cohort of engineering students, low engagement with this tool can work as an indicator that a certain student might not persist to engineering coursework in the following semester. With that in mind, we plan to investigate how soon after the beginning of the semester can one effectively determine that a student is “disengaging” and what interventions can be conducted during the semester to attempt to increase engagement levels before the course ends.

Furthermore, we have begun to compare, combine, and triangulate traditional learning analytics data from Sakai, the campus learning management system, with the academic, demographic, and engagement data sets. Additional samples across the College of Science are also attempting to replicate and scale the ePortfolio design, utilization, and analytics gained from this Intro to Engineering course.

While for the most part we can measure how much time and energy a student put into his/her ePortfolio by looking at the quantitative metrics discussed in this paper, there may also be substantial knowledge to be gained from understanding the qualitative data within the ePortfolio. To that end, we have also started to explore insights that can be harvested by text mining the actual content of students’ ePortfolios. Lastly, we intend to condense this predictive analysis and students’ engagement trajectories over time into an interactive dashboards that can be used by academic advisors who will then be able to track student progress and decide who is in need of individual attention and when that intervention should take place.
8.1 Introduction

As part of our overall effort in the development and evaluation of these machine learning based early warning indicator systems, a variety of partnerships were built, and some key aspects of our work were (and are being) re-purposed so as to allow the larger education community to make use of the techniques we described. The goal of this chapter is to highlight those aspects, their impact, and other ramifications of this work that emerged from the ideas we explored.

8.2 Exploring ePortfolio Data

The work we described in Chapters 6 and 7, as well as in 5, 6, 51, 88, was fundamental in showing that, when combined with more traditional student-level information, data obtained from electronic portfolios can help in the identification of students that may be disengaging from STEM. This is especially helpful when ePortfolios are incorporated into the design of courses as a medium wherein students have the opportunity to express their current state of mind as it pertains to being in the STEM field.

Following that preliminary work, we have since developed a fully automated pipeline for ePortfolio data collection that leverages the Digication 4 data API. Through this pipeline, we retrieve, store, and subsequently analyze data describing how students interact with their ePortfolios on a daily basis. In addition to the
engagement features listed in Table 7.2, we also collect information describing the number of modules and pages students create within their portfolios, the number of comments they post, and their choice of access permission settings.

Sitting atop the datasets collected through that pipeline, we developed a variety of interactive web-based tools that help educators get a detailed understanding of how ePortfolios are being utilized, as well as in the identification of students that are particularly active (or otherwise). This data exploration dashboard was created using R Shiny [104], and in Figure 8.1 we illustrate one of its specific functionalities.

Figure 8.1. ePortfolio data exploration and reporting dashboard. This figure shows an R Shiny app that we developed for a medium-sized university in the Mid-Atlantic region. A variety of metrics collected to describe student activity can be highlighted using this interactive tool.

---

1 Students have the option of making their ePortfolios publicly accessible, private, or available only to their peers.
8.3 Web-based Classification Tool

A recurring theme throughout the work described in this dissertation is the use of classification algorithms to aid in the early identification of students at risk for some adverse outcome. As we previously emphasized in Chapters 4 and 5, school districts in the US are almost always limited, from a financial standpoint, in what they can do to help students that may need extra support. Despite the provably usefulness of our proposed EWI strategies, more often than not, districts do not have the resources or personnel to explore the adoption of these tools.

With that in mind, we have since packaged the vast majority of the functionality of the models we used in the secondary education setting, as well as the code needed to generate the dashboard shown in Figure 4.11 into a format that can be easily replicated by other researchers [3]. Further, we have also abstracted that idea to a more general setting, and developed a fully functional web-based tool for data classification. This tool leverages all supervised learning methods available through the Python package scikit-learn [94], and provides users with an organized analysis pipeline where they can:

- Select from a variety of available datasets as well as upload their own.
- Apply methods for pre-processing their data prior to classification. These include SMOTE [32], undersampling, principal component analysis, outlier detection, normalization, binarization, standardization, and multiple missing data handling techniques.
- Choose from many different classifiers (Figure 8.2) and visually set their corresponding parameter values.
- Analyze the results of their experiments through an interactive report (Figures 8.3 and 8.4) that displays ROC and PR curves, confusion matrices, instance-wise prediction details, and summarized information on precision, recall and f1-scores.
Figure 8.2. Web-based classification tool. This figure illustrates the multi-step pipeline component wherein users can select their data, pre-processing tasks, models and parameters.
Figure 8.3. Web-based classification tool. This figure illustrates the interactive ROC and PR curves that are embedded into the analysis report generated for each experiment.
Figure 8.4. Web-based classification tool. This figure illustrates the second half of the report page where the user has access to a confusion matrix, various performance metrics, as well as the detailed prediction output for each instance analyzed.
CHAPTER 9
CONCLUSION

In this chapter we provide a brief summary of the work presented in this dissertation, highlighting our most important contributions, as well as future work directions to expand and improve upon them.

9.1 Summary and Discussion

We began this work in Chapter 2 by providing a comprehensive review of the recent literature that investigates early warning systems for a variety of adverse academic outcomes. We describe the effort of education researchers that carefully analyzed and listed a range of factors that contribute to a student’s decision to leave school prematurely. A second portion of our focus there was dedicated to covering works that made use of institutional data representing these risk factors in the development and evaluation of methods that identify at risk students ahead of time. Further, in Chapter 3 we introduced the reader to the supervised learning concepts, techniques and evaluation metrics that are used throughout this dissertation.

Our first contributions are presented in Chapter 4. Through a partnership established with a large school district, we were provided access to longitudinal academic, behavioral and coursework-related student-level data. A range of feature selection methods helped us unearth, for that particular cohort of students, what factors were more likely to contribute to undesired outcomes. We further utilized this data to create models that could detect what students were at risk of not graduating high school on time, as well as rank these students based on an urgency score metric. That
work is extended in Chapter 5, where we show that our machine learning approach to creating these early warning systems can easily extend to other cohorts of students at the same or different school districts. Also in that chapter, we present a more careful framework for evaluating these models from both technical and operational standpoints.

In Chapter 6, we introduced the setting for our post-secondary education efforts, which were carried out in partnership with the College of Engineering at the University of Notre Dame. We detailed how electronic portfolios were incorporated into the Engineering curriculum, and showed that some of the features extracted from them can be used as a reliable means for estimating student engagement. We then tested a variety of predictive models with the goal of detecting students at risk of switching out of their pre-declared engineering majors in Chapter 7, emphasizing the strengths and advantages of using data that represents both academic performance and student engagement.

Lastly, in Chapter 8, we showed how some aspects of this work were re-purposed and packaged to allow other researchers to replicate our ideas in their own context.

9.2 Future Work

The use of machine learning techniques to improve the accuracy, reliability and efficiency of early warning systems shows an encouraging amount of promise. Nevertheless, throughout the course of this work, we identified several obstacles that are yet to be fully circumvented. First, we note that while our proposed models are certainly more reliable than those we initially observed in use, finding a way to make the appropriate connection between students identified to be at risk, and the best possible intervention to each particular case is still an open problem. In future iterations of this work, we hope to place a larger focus on collecting data that documents the effectiveness of intervention programs. Ideally, a robust second-generation EWI
system will, not only identify students that are at risk, but also make suggestions (based on historical data) of how to best approach each specific instance.

Secondly, through our various conversations with experienced practitioners that are now beginning to leverage these predictive models to help students succeed, we learned that while the assignment of risk scores to students identified to be off track can be extremely helpful, the scores themselves may be less so. School counselors and teachers are often less interested in learning which students are at highest risk than they are in knowing which of the students at risk are more likely to respond to interventions. As such, an interesting extension to our work would be the integration of a persuasion modeling layer that can rank students based on their likelihood of benefiting from different intervention programs.

Moreover, as more and more schools begin to utilize student-level data to better understand their student population and their struggles, we are very hopeful for the future of this category of EWIs. The increasing interest for such systems by school districts and universities will organically provide answers to the challenges of operationalizing these tools as part of the institutional work-flow. Over time, we ultimately wish that the techniques proposed here become part of a larger set of weapons that will help educators graduate more, and better, students.


80. L. M. B. Manhães, S. M. S. da Cruz, and G. Zimbrão. Investigating withdraw of STEM courses in a brazilian university with EDM. In Symposium on Knowledge Discovery, Mining and Learning (KDMILE), 2014.


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


