

# Who, When, and Why: A Machine Learning Approach to Prioritizing Students at Risk of not Graduating High School on Time\*

Everaldo Aguiar  
University of Notre Dame  
Notre Dame, IN 46556  
eaguiar@nd.edu

David Miller  
Northwestern University  
Evanston, IL 60208  
dmiller@u.northwestern.edu

Himabindu Lakkaraju  
Stanford University  
Stanford, CA 94305  
himalv@cs.stanford.edu

Ben Yuhas  
Yuhas Group  
Baltimore, MD 21210  
ben@yuhasgroup.com

Nasir Bhanpuri  
University of Chicago  
Chicago, IL 60637  
nbhanpuri@uchicago.edu

Kecia L. Addison  
Montgomery County Public  
Schools  
Rockville, Maryland 20850  
Kecia\_L\_Addison@mcpmsmd.org

## ABSTRACT

Several hundred thousand students drop out of high school every year in the United States. Interventions can help those who are falling behind in their educational goals, but given limited resources, such programs must focus on the right students, at the right time, and with the right message. In this paper, we describe an incremental approach that can be used to select and prioritize students who may be at risk of not graduating high school on time, and to suggest what may be the predictors of particular students going off-track. These predictions can then be used to inform targeted interventions for these students, hopefully leading to better outcomes.

## Categories and Subject Descriptors

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

## General Terms

Measurement, Performance

## Keywords

Secondary Education, Early Intervention, Student Retention, Learning Analytics, Predictive Analytics

\*The work described in this paper was done as part of (and supported by) the Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship at the University of Chicago [19].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [Permissions@acm.org](mailto:Permissions@acm.org).

LAK '15, March 16 - 20, 2015, Poughkeepsie, NY, USA  
Copyright 2015 ACM 978-1-4503-2664-3/14/03 ...\$15.00.  
<http://dx.doi.org/10.1145/2723576.2723619>

## 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 [20]. 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 [1, 14]. 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 [8, 12, 17].

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 [5, 6]. 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 [11]. 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 [9]. 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 EWIs 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 and administrators better understand the various factors contributing most 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 6<sup>th</sup> to 12<sup>th</sup> grade. Using this data, this paper 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.

The remainder of this paper is organized as follows. Section 2 reviews selected studies from recent related literature. Section 3 describes our dataset in detail. Sections 4, 5, and 6 describe our proposed data analysis approach, preliminary results, and a mock-up interface. Section 7 ends with our conclusions and directions for future work.

## 2. RELATED WORK

Decades of research has investigated the various causes and predictors of on-time high school graduation [17]. Many scholars, especially in the most recent years, have proposed that school districts should use this research knowledge by implementing EWI systems. Bowers et al. [9] systematically reviewed this vast literature on predicting high school graduation by quantitatively synthesizing the predictive accuracy of 110 early warning flags across 36 studies. These studies found that many factors can serve as effective EWIs, including those routinely recorded by schools such as having high absence rates or low grades.

Bowers et al.’s review [9] found these studies 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 other studies 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 [10] to optimally combine these EWIs when forming composite indicators.

Although these EWIs may accurately predict high school drop-out, they do not always reveal the root causes of students’ needs [8, 12, 17]. For example, although low grades are 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, a student whose grades sharply declined between middle school and high school might have struggled adapting to the new social and academic environment [5, 7]. We therefore created a prototype of a student dashboard that could help teachers and administrators form these insights when reviewing student records. These insights could help match at-risk students with appropriate intervention programs. Prior implementation guides on EWIs [18] have detailed how these integrated analytic approaches can lead to a comprehensive set of actions for addressing particular students needs.

### 2.1 District-level Work

In 2013, our partner school district investigated linkages between grades 1, 3, 6, 9, and final graduation status for two cohorts of students [21]. 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 [21], the school district is developing an EWI system for students across all grade levels (elementary through high school). However, rather than focusing only on dropouts, the district would also like to identify students who stay in school, but need extra support for future success. For example, academic performance and attendance were significantly correlated to the amount of time until achieving a bachelor’s degree for one cohort of students [22]. 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 impact student performance [16, 15].

## 3. DATA

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 6<sup>th</sup> to 12<sup>th</sup> grade, while some arrived throughout the study and had missing data for any years prior to enrollment. Figure 1 illustrates the student flow in and out of the school district over time.

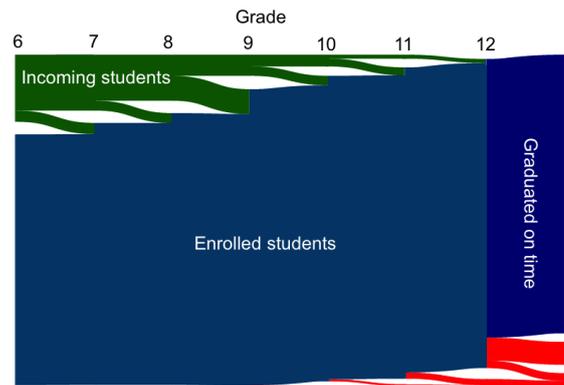


Figure 1: Student enrollment flow over time.

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 larger group of incoming students being observed in 9<sup>th</sup> 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.

End of year data were captured as students progressed towards their high school graduation. Given that the district already had an EWI system prototype, 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 districts

Name	Type	Description	Available On
gender	Binary	Student’s gender	6 <sup>th</sup> - 12 <sup>th</sup>
birth	Numeric	Student’s birth year and month	6 <sup>th</sup> - 12 <sup>th</sup>
ssname	Nominal	Name of school where student is enrolled	6 <sup>th</sup> - 12 <sup>th</sup>
absrate	Numeric	Student’s absence rate for a given academic year	6 <sup>th</sup> - 12 <sup>th</sup>
tardyr	Numeric	Student’s tardiness rate for the academic year	6 <sup>th</sup> - 12 <sup>th</sup>
nsusp	Numeric	Number of suspension incidents	6 <sup>th</sup> - 12 <sup>th</sup>
mobility	Numeric	Cumulative to date number of unexpected entries or withdrawals	6 <sup>th</sup> - 12 <sup>th</sup>
new	Binary	Flag to denote if the student is new to the school district	9 <sup>th</sup> - 12 <sup>th</sup>
q1-4gpa	Numeric	Quarter Marking Period Averages	6 <sup>th</sup> - 12 <sup>th</sup>
gmapr	Numeric	MAP-R National Percentile Ranks	6 <sup>th</sup> - 8 <sup>th</sup>
gmsam	Numeric	Maryland School Assessment in Math Proficiency Level	6 <sup>th</sup> - 8 <sup>th</sup>
psatv	Numeric	PSAT Critical Reading	10 <sup>th</sup>
psatm	Numeric	PSAT Math	10 <sup>th</sup>
retained	Binary	Flag to denote if the student was ever retained to date in a grade	9 <sup>th</sup> - 12 <sup>th</sup>
grad	Binary	Flag to denote if the student graduated from high school on time	12 <sup>th</sup>

**Table 1: Description of dataset features.**

served by our partners. It is worth noting that the techniques described in this paper 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 environments.

Table 1 contains a brief description of the data we utilized. Note that most information is available on a yearly basis starting at 6<sup>th</sup> 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.

### 3.1 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 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 1 – with respect to four metrics: information gain (IG), gini impurity (GI), stepwise regression (SR) and single feature performance (SFP<sup>1</sup>).

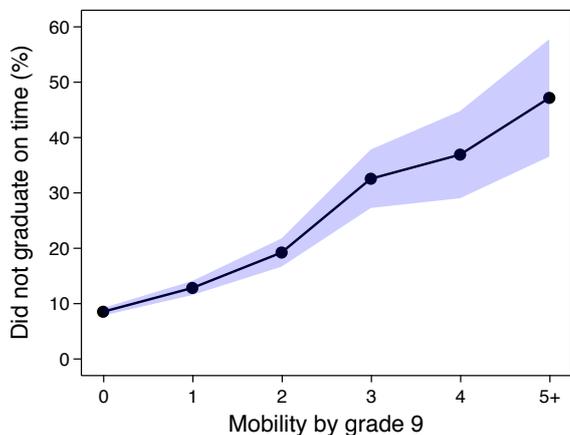
Rank	IG	GI	SR	SFP
1	q1-4gpa	retained	q1-4gpa	q1-4gpa
2	gmapr	q1-4gpa	gmsam	mobility
3	retained	gmapr	gmapr	nsusp
4	birth	absrate	retained	absrate
5	absrate	birth	birth	ssname

**Table 2: Feature importance rankings.**

Aggregating students based on some of the features provided interesting insights that were previously not known. For example, as illustrated in Figure 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

<sup>1</sup>This consisted of evaluating a logistic regression trained with each feature individually.

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 3).



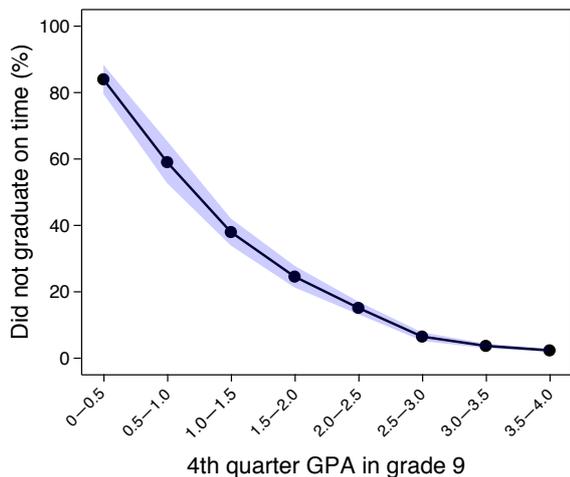
**Figure 2: Student mobility values vs. on-time graduation rates. Mobility is the number of times a student unexpectedly switched schools. The shaded region corresponds to 95% CIs.**

## 4. WHO?

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 [13], secondary schools may miss important opportunities to help at risk 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<sup>2</sup> so that schools

<sup>2</sup>More specifically, these flags are generated quarterly but can be consulted in an yearly basis.



**Figure 3: Student GPAs (binned) vs. on-time graduation rates. The shaded region corresponds to 95% CIs.**

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<sup>th</sup> grade used all available data from 6<sup>th</sup> to 9<sup>th</sup> 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 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<sup>3</sup>.

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 which were most precise [10], and logistic regression models which were more interpretable while still performing well.

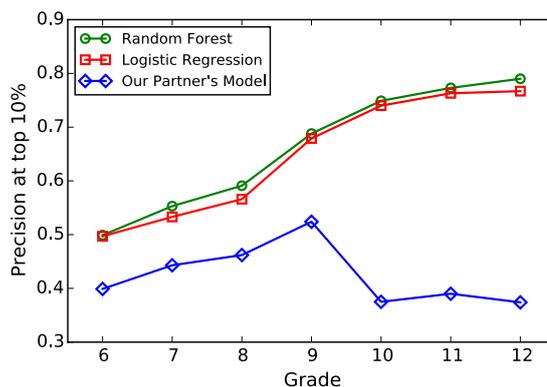
## 4.1 Results

Figure 5 shows that the machine learning models performed noticeably better than our partner’s existing rule-based model. For example, in the 10<sup>th</sup> grade models, the random forest and logistic regression models performed at 75% and 74% respectively, whereas the existing model performed at 38%.

<sup>3</sup>We used k=10 in this paper somewhat arbitrarily, though based on interviews we conducted for a parallel project, that seems to be an adequate estimate of how many students schools can typically place on intervention programs.



**Figure 4: Precision at top 10% metric explained for a cohort of 100 students.**



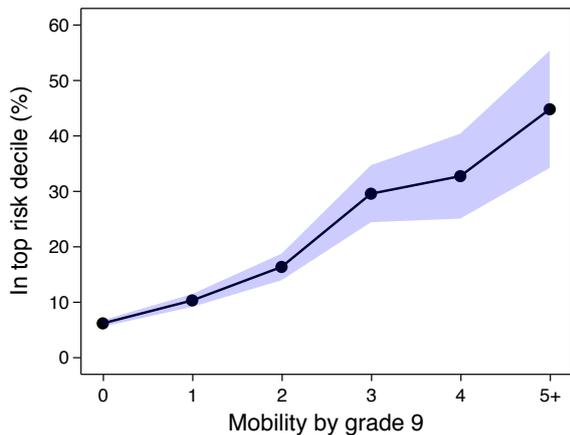
**Figure 5: Precision of models for each grade level.**

Our models, which used students’ historical data, consistently improved over time as the models continued to use 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 9<sup>th</sup> grade model only used 9<sup>th</sup> grade GPAs). This model’s performance notably dropped from 9<sup>th</sup> to 10<sup>th</sup> grade. These results suggest the value of using all available historical information about each student.

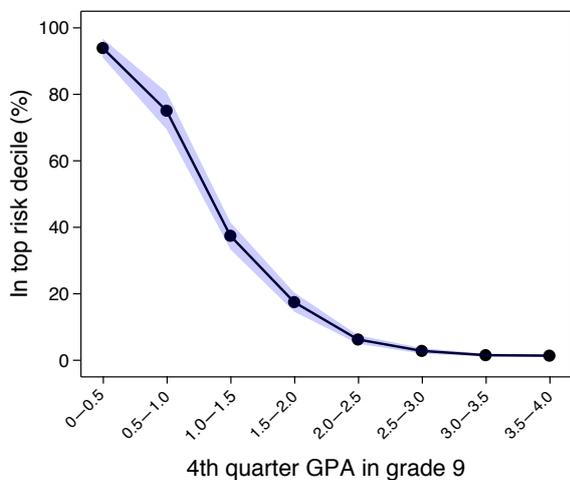
Figures 6 and 7 help show what kinds of students are ranked higher by our models. These figures mirror Figures 2 and 3 but are based on the top decile of predicted risk scores (based on random forest models), rather than the observed outcome of not graduating on time. As shown, student-level features (e.g., mobility, GPA) have a similar relationship with our models’ risk scores (Figures 6 & 7) 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.

## 4.2 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



**Figure 6: 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 7: Student GPAs (binned) vs. high-risk rates.** The shaded region corresponds to 95% CIs.

number of transitions that students made between three different risk categories (high, moderate, and low) from year to year. For example, the below path includes two transitions:

*high* → *high* → *moderate* → *moderate* → *moderate* → *high*

We defined *high risk* as the top 10% of risk scores, *moderate risk* as the following 20%, and all other students were *low risk*<sup>4</sup>. We found that in their 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.

<sup>4</sup>These cutoffs were chosen to remain consistent with categories already developed for our partner’s existing system.

Number of Transitions	Current EWI	Random Forest Model
0	37%	60%
1	19%	17%
2	19%	14%
3	15%	6%
4	7%	2.7%
5+	3%	0.3%

**Table 3: Percentage of students moving across different risk groups over time.**

### 4.3 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 [9]), 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.

## 5. WHEN?

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, however, 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.

### 5.1 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 at high risk (top 10%) by the predictive models discussed in the previous section.

## 5.2 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.

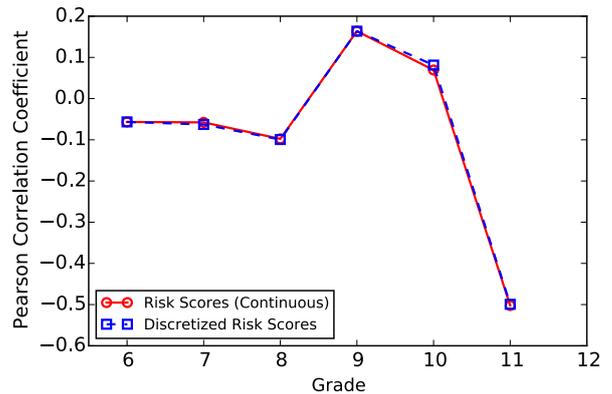


Figure 8: Correlation between risk scores & time to off-track.

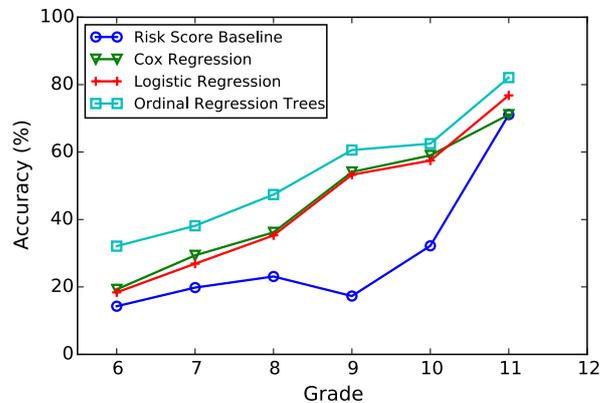


Figure 9: Accuracy predicting time to off-track.

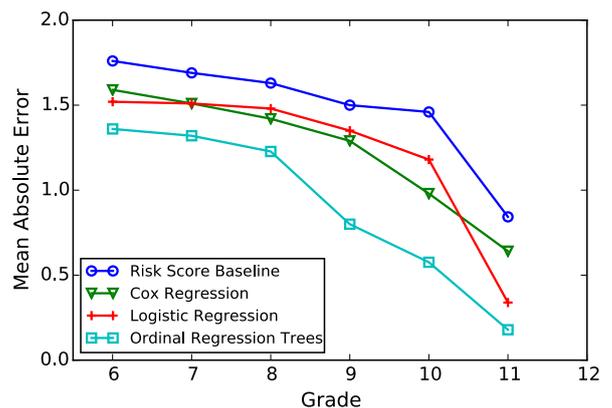


Figure 10: Mean Absolute Error predicting time to off-track.

Figure 8 shows the results of these experiments with Pearson's correlation coefficient. It can be seen from Figure 8

that there does not exist a high correlation between risk scores and time to off-track from grade 6 through to grade 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 prediction models for the task at hand.

### 5.3 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 predict 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 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 9 and 10. In addition, we also present the results from using the discretized risk scores as a proxy for time to off-track. Figure 9 encapsulates the accuracy metric, and Figure 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.

## 6. WHY?

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. [4], 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 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 working with schools to design effective interventions targeting these factors.

A live demo of our dashboard created with artificial data is accessible at [3]. There, as well as in the few static screenshots seen in Figure 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 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 [2].

## 7. SO WHAT?

We have shown two key ideas in this paper 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 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's 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 set 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.

### 7.1 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 ap-

plied to all students in the district with the intent of rolling out a fully operational EWI system 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 drew attention to the important features of a given student's record. 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 for not only school-based staff, but also district leaders.

## 8. ADDITIONAL AUTHORS

**Shihching Liu** (Montgomery County Public Schools)  
email: [Shihching\\_Liu@mcpsmd.org](mailto:Shihching_Liu@mcpsmd.org),  
**Marilyn Powell** (Montgomery County Public Schools)  
email: [Marilyn\\_Powell@mcpsmd.org](mailto:Marilyn_Powell@mcpsmd.org)  
**Rayid Ghani** (University of Chicago)  
email: [rayid@uchicago.edu](mailto:rayid@uchicago.edu).

## 9. REFERENCES

- [1] Building a Grad Nation. <http://www.americaspromise.org/sites/default/files/legacy/bodyfiles/BuildingAGradNation2012.pdf>.
- [2] Risk score Visualization - Student Dashboard Code. <https://github.com/dssg/student-early-warning>. Accessed: 2014-10-01.
- [3] Risk Score Visualization - Student Dashboard Demo. [tinyurl.com/dssgdemo](http://tinyurl.com/dssgdemo). Accessed: 2014-10-01.
- [4] K. L. Alexander, D. R. Entwisle, and C. S. Horsey. From first grade forward: Early foundations of high school dropout. *Sociology of education*, pages 87–107, 1997.
- [5] E. M. Allensworth and J. Q. Easton. What matters for staying on track and graduating in Chicago public high schools. *Chicago, IL: Consortium on Chicago school research*. Retrieved December, 17:2007, 2007.
- [6] R. Balfanz, L. Herzog, and D. J. Mac Iver. Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist*, 42(4):223–235, 2007.
- [7] A. J. Bowers and R. Spratt. Examining the multiple trajectories associated with dropping out of high

- school: A growth mixture model analysis. *The Journal of Educational Research*, 105(3):176–195, 2012.
- [8] A. J. Bowers and R. Sprott. Why tenth graders fail to finish high school: Dropout typology latent class analysis. *Journal of Education for Students Placed at Risk*, 17(3):129–148, 2012.
- [9] A. J. Bowers, R. Sprott, and S. A. Taff. Do we know who will drop out?: A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, 96(2):77–100, 2013.
- [10] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [11] Data Quality Campaign. Supporting early warning systems: Using data to keep students on track to success. <http://goo.gl/2CNw24>, note = [Online; posted 08-August-2014], August 2014.
- [12] D. C. French and J. Conrad. School dropout as predicted by peer rejection and antisocial behavior. *Journal of Research on adolescence*, 11(3):225–244, 2001.
- [13] P. Gleason and M. Dynarski. Do we know whom to serve? Issues in using risk factors to identify dropouts. *Journal of Education for Students Placed at Risk (JESPAR)*, 7(1):25–41, 2002.
- [14] H. M. Levin and C. Belfield. *The price we pay: Economic and social consequences of inadequate education*. Brookings Institution Press, Washington D.C., 2007.
- [15] S. Liu and H. Wang. An innovative tool for promoting achievement gap closing among high school students. *Maryland Assessment Group*, 2013.
- [16] V. Rethinam. Grade 9 indicators influencing high school graduation and college readiness in montgomery county public high schools. *Montgomery County Public Schools*, 2011.
- [17] R. W. Rumberger and S. A. Lim. Why students drop out of school: A review of 25 years of research. *California Dropout Research Project, Policy Brief 15*, 2008.
- [18] S. B. Therriault, M. O’ Cummings, J. Heppen, L. Yerhot, and J. Scala. High school early warning intervention monitoring system implementation guide. 2013.
- [19] University of Chicago. The Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship. <http://dssg.uchicago.edu/>. Accessed: 2014-10-01.
- [20] U.S. Department of Education, National Center for Education Statistics. The condition of education. 2014.
- [21] T. C. West. Just the right mix: Identifying potential dropouts in montgomery county public schools using an early warning indicators approach. *Montgomery County Public Schools*, 2013.
- [22] H. Zhao and S. Liu. College readiness and postsecondary educational outcomes for 2003 graduates of Montgomery County Public Schools. Rockville, MD. *Montgomery County Public Schools*.

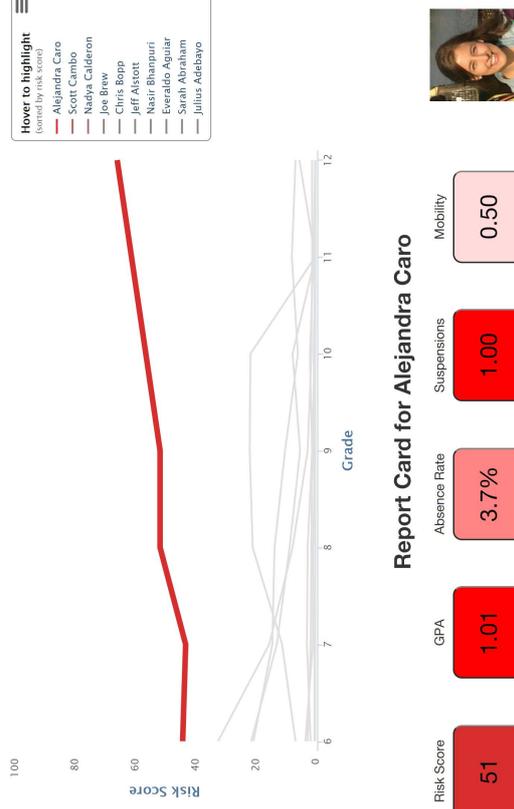
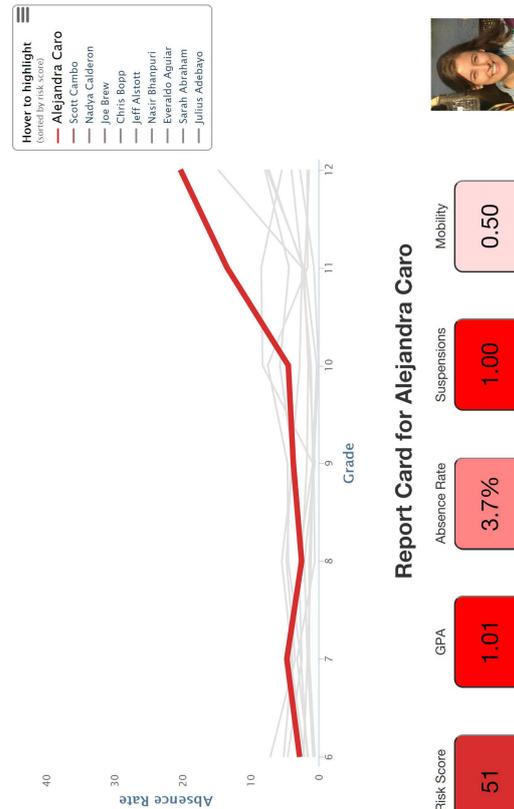
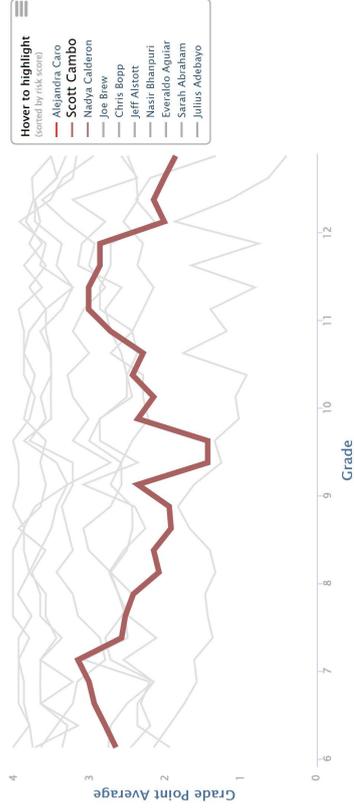
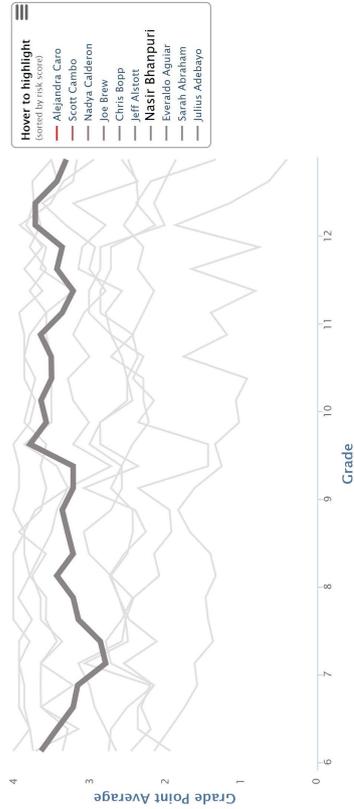


Figure 11: Student Report Card Dashboard depicting risk score, GPA, and absence rate trajectories and categorical breakdowns.