A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes*

Himabindu Lakkaraju
Stanford University
himalv@cs.stanford.edu

Everaldo Aguiar
University of Notre Dame
eaguiar@nd.edu

Carl Shan
University of Chicago
carlshan@uchicago.edu

David Miller
Northwestern University
dmiller@u.northwestern.edu

Nasir Bhanpuri
University of Chicago
nbhanpuri@uchicago.edu

Rayid Ghani
University of Chicago
rayid@uchicago.edu

Kecia L. Addison
Montgomery County Public Schools
Kecia_L_Addison@mcpsmd.org

ABSTRACT

Many school districts have developed successful intervention programs to help students graduate high school on time. However, identifying and prioritizing students who need those interventions the most remains challenging. This paper describes a machine learning framework to identify such students, describes features that are useful for this task, applies several classification algorithms, and evaluates them using metrics important to school administrators. To help test this framework and make it practically useful, we partnered with two U.S. school districts with a combined enrollment of approximately of 200,000 students. We together designed metrics to evaluate the framework's performance and tools such as interactive dashboards to help match at risk students with appropriate supports. This paper focuses on students at risk of not finishing high school on time, but our framework lays a foundation for future work on other adverse academic outcomes.

Categories and Subject Descriptors

I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems

Keywords

evaluation metrics; applications; education; risk prediction;

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 [28]. 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 [7, 13, 24]. Studies have shown that not graduating high school on time impacts a student's future career prospects immensely [1, 18]. 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 [8]. Though human judgment and heuristics can often be accurate, they serve as rules of thumb, are static, expensive to maintain, and often error prone [25]. 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 [8, 26, 3]. Trained using traditional academic data, these machine learning approaches can often identify at risk students earlier and more accurately than prior rule-based approaches [3].

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

*The work described in this paper was done as part of (and partially supported by) the Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship at the University of Chicago [27].

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

DOI: http://dx.doi.org/10.1145/2766XXX.XXXXXXX.

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 are currently developing 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 instead of only being focused on commonly used machine learning metrics.
- 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 recent work 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.

2. RELATED WORK

Our interdisciplinary work benefited from prior research at the intersection of educational research and data mining. Educational research provided a basis for selecting and understanding features important indicative of adverse academic outcomes, and data mining research helped us use these features and machine learning algorithms to develop robust early warning systems.

Educational research has found that some specific features such as students' grades and attendance are especially relevant to predicting on-time high school graduation [24, 8]. Bowers et al. [8] systematically reviewed this vast literature, finding that these commonly recorded student features can robustly predict future student outcomes. However, these studies widely varied in how they combined individual student features when developing rule based models (e.g., a rule based on the intersection or union of having low grades and low attendance). Consequently, predictive performance widely varied. Moreover, many studies focused on developing high precision rules, but at the cost of low recall. Such rule-based models are also not generalizable in the sense that they might work well for a specific district and cohort, but result in poor performance when applied elsewhere.

Recent research in data mining addresses the limitations of such rule based models by advocating the usage of automated learning methods. Several well-known machine learning algorithms such as Random Forests, Logistic Regression, Decision Trees etc. were used to predict student outcomes [26, 12, 3, 2, 11, 22, 30, 12]. These models consistently outperformed rule based models on traditional metrics such as precision, recall, and AUC. In addition, models such as Bayesian networks were employed to identify students who were likely to fail in mathematics courses [29]. Further, machine learning models were also employed to predict trajectories of future learning performance using past history [15].

Though machine learning models are very useful in practice, they are essentially black boxes from an educator's perspective. Furthermore, traditional model evaluation metrics such as AUC are both difficult to interpret for educators and not well suited to address all the issues at hand citekiri. We therefore worked closely with two school districts to identify how to best interpret and evaluate machine learning methods so that they address the districts' educational goals.

3. DATASET DESCRIPTION

The work we describe in this paper 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 had already recognized the importance of early warning indicator systems for identifying at-risk students. District A had a rule-based early warning indicator system in place, using several important indicators such as academic performance, behavior, mobility and 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 of 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 6^{th} - 12^{th} 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 6^{th} grade and graduated within the stipulated time.

The dataset obtained from District B comprises of 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 8^{th} - 12^{th} 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[4, 6]. The data used for this analysis captured most of these indicators. Table 1 provides an exhaustive list of all attributes 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 and on an yearly basis for District B. Our analysis is not sensitive to such representational variations. In fact, the framework proposed in this paper is generic enough to be applicable to any given set of features.

Student Attributes	**District A**	**District B**
Gender	✓	✓
Age	✓	X
Ethnicity	X	✓
City	X	✓
Street	X	✓
School Code	✓	✓
Absence Rates	✓	✓
Tardiness Rates	✓	✓
# of Suspensions	✓	X
# of Unexpected Entries/Withdrawals	✓	✓
Quarterly GPA	✓	X
Cumulative GPA	X	✓
Cumulative Math GPA	X	✓
Cumulative Science GPA	X	✓
Cumulative Social Science GPA	X	✓
Cumulative English GPA	X	✓
MAP-R National Percentile Ranks	✓	X
Math Proficiency Scores (MPS)	✓	X
PSAT Critical Reading	✓	X
PSAT Math	✓	X
Limited English Proficiency	X	✓
Economically Disadvantaged (EDS)	X	✓
Is student new to the school district?	✓	✓
Is student disabled?	X	✓
Was student ever retained?	✓	✓
Did student graduate on time?	✓	✓

Table 1: List of student attributes and their availability in Districts A and B.

4. FRAMEWORK OVERVIEW

In this section, we present an overview of the models that we will be using through out 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 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 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 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 (RF), Adaboost (AB), Logistic Regression (LR), Support Vector Machines (SVM), and Decision Trees (DT). We use *scikit-learn* implementations of all these models.

Experimental Setup: Our datasets comprise of 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 [26, 3]. Though this is an acceptable way of estimating any algorithm's performance in general, it is not ideal for the current setting. To illustrate, school districts often have access to outcomes and other features from previous cohorts.

The goal here is to predict the future outcomes accurately by training on data from previous cohorts. Thus, an apt way of evaluating an algorithm in this setting is to train a model using data from previous cohorts and use a later cohort as the test set. We carry out all 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 carry out 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. During each iteration, every model is trained on (N - k) randomly chosen data points, where N is the size of the entire dataset and k = 0.01*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. 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, we use GPAs, absence rates, tardiness, other scores and flags (listed in Table 1) up until grade 8 along with other demographic attributes and predict the outcome variable `no_grad`. 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 ?

5.1 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 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.

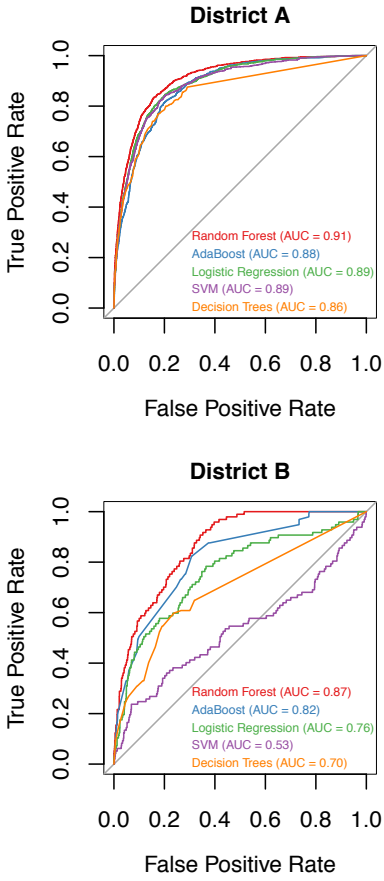

Figure 1: ROC Curves

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. The educators 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, District A might have the resources to support 100 students in 2012, however they might be able to target only 75 students in 2013. Further, District B might be able to assist 300 students in 2012 and 500 students in 2013. Building algorithms that can cater to these settings is extremely crucial to address 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 just choose the top 75 students from this rank ordered list and assist them.

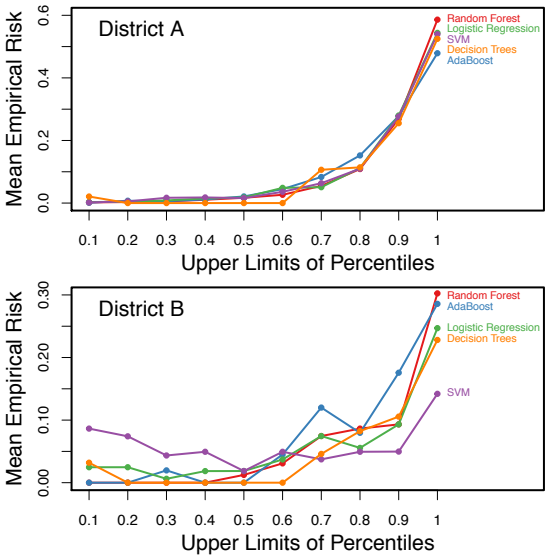

Figure 2: Empirical Risk Curves. The ranking quality of an algorithm is good if this curve is monotonically non-decreasing.

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[20] 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.

5.2 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 tree assigns 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[10]. Random Forests involve training a forest of trees on data points and the probability of not graduating on time for a particular data point is computed as the mean of the predicted class probabilities of the trees in the forest [9]. 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 involves 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[21]. Support Vector Machines 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[19].

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 $10^{th}$and 20^{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 considered to be producing good risk scores and consequently ranking students correctly if and only if the *empirical risk curve* is monotonically non-decreasing. If the empirical risk curve is non-monotonic for some algorithm, it implies that the 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 compared to students with higher risk scores. Figure 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.

5.3 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 already emphasized on 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 these curves ***precision_at_top_K curve*** and ***recall_at_top_K curve*** respectively. These curves help

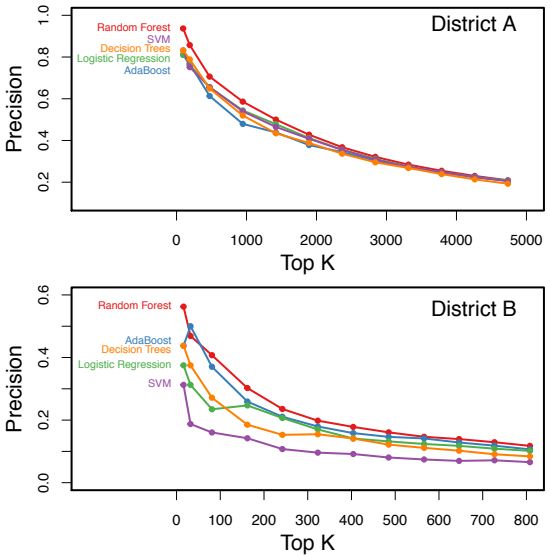

Figure 3: Precision at top K.

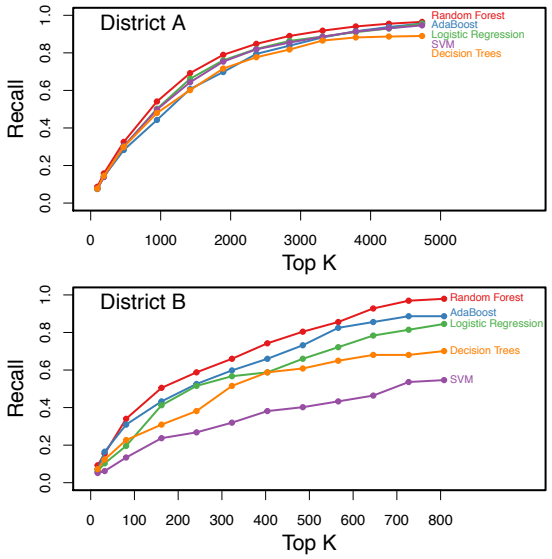

Figure 4: Recall at top K.

educators in readily inferring the precision and recall of various algorithms at a threshold K of their choice.

Figure 3 illustrates the precision_at_top_K curves for districts A and B respectively. It can be seen that there are huge 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 <= 150 on District B. Beyond this threshold, Logistic Regression has a higher precision than the decision tree.

Figure 4 shows the recall_at_top_K curves for both districts. Again, random forest outperforms all other models at 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 other algorithms depends on the threshold K.

6. 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.

6.1 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. We specifically chose to threshold our datasets at the end of 8^{th} 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 [5].

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[23]. 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 illustrates how features are ranked by various algorithms. Due to space constraints, we only present a subset of 5 features in Figure 5. Table 2 lists the top 5 features used by each of the algorithms.

It can be inferred from Table 2 and Figure 5 that GPA at 8^{th} grade is highly ranked across the majority of the approaches, indicating that academic performance at that

	Rank	RF	AB	LR	SVM	GI	IG
District A	1	Q4GPA_08	Q4GPA_08	Gender_07=Male	Gender_07=Male	Q4GPA_08	Q4GPA_08
	2	Q3GPA_08	MPS_08	Gender_07=Female	Gender_07=T	MAPR_08	Abs_Rate_08
	3	Q1GPA_08	MAPR_08	Gender_06=Female	Gender_06=Female	Abs_Rate_08	MAPR_08
	4	MAPR_08	Abs_Rate_08	Abs_Rate_08	Abs_Rate_08	Q1GPA_08	Abs_Rate_07
	5	MPS_08	Q4GPA_06	Q2GPA_06	Gender_06=Male	MPS_08	MAPR_06
District B	1	GPA_08	GPA_08	GPA_Science_08	School_Code=317	GPA_08	GPA_08
	2	GPA_ENG_08	Days_Abs_08	Math_Credits_08	GPA_SocSci_08	GPA_Science_08	GPA_SocSci_08
	3	GPA_SocSci_08	Num_Marks_08	GPA_SocSci_08	GPA_Math_08	Exc_Abs_08	Exc_Abs_08
	4	GPA_Math_08	GPA_Science_08	School_Code=317	GPA_Science_08	Num_Marks_08	GPA_ENG_08
	5	GPA_Science_08	Has_Disability_08=N	Has_Disability_08=Y	School_Code_08=315	EDS_08=T	School_Code_08=320

Table 2: List of top 5 features in districts A and B (GI stands for Gini Index and IG is Information Gain).

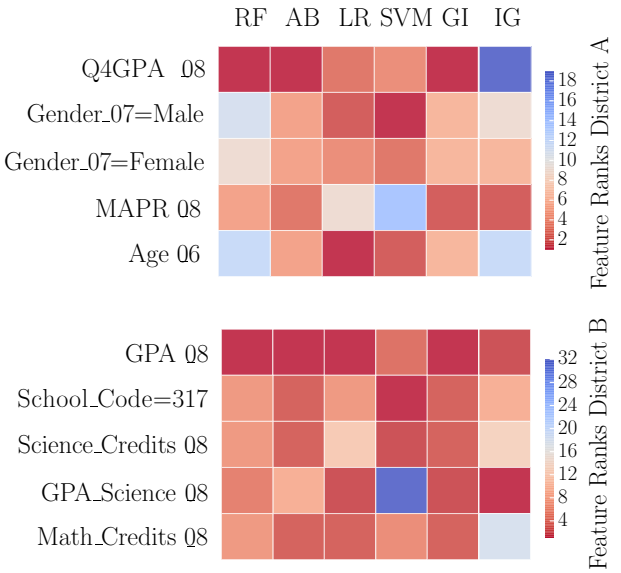

Figure 5: Feature ranking by various algorithms. GPA_08 is ranked consistently high in both the districts by most of the models.

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 8^{th} 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 high.

6.2 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[14]. 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_08 > 2.0 and Abs_Rate_08 <= 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 the datapoints for which the pattern holds true and computing the fraction of these datapoints where *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. Due to space constraints, we present the patterns for just two of these - Random Forest and Decision Trees in Table 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.

6.3 Comparing classifier predictions

Our discussions with school districts revealed that educators placed a lot of importance on exploratory aspects of models. When we present educators with a suite of algorithms, they are keen on understanding the differences in rank orderings produced by each of these algorithms. Here, we address the question: *How similar or dissimilar are the rank orderings produced by any two given models ?*

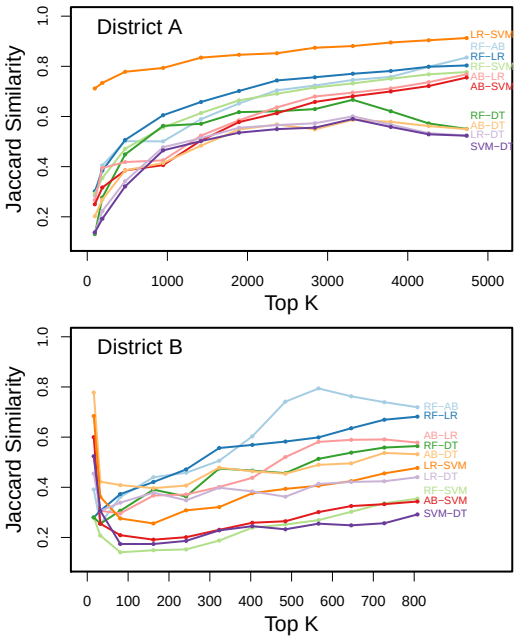

Figure 6: Jaccard Similarity of students at-risk for various algorithms

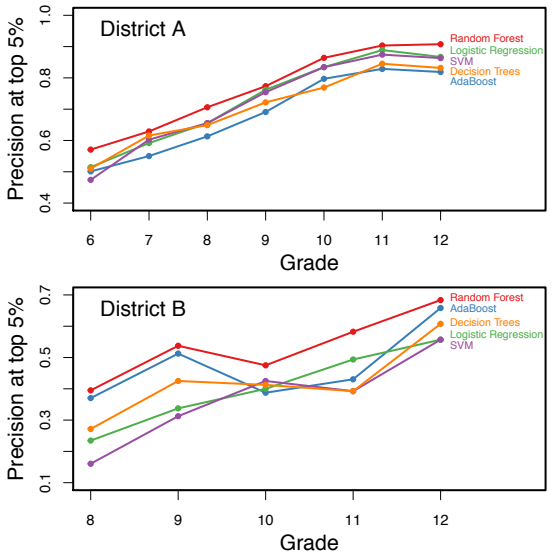

Figure 7: Precision by grade at top K

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 [17] 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).

Figure 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.

7. 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 for each of the districts.

We evaluate the performance of the models across all grades. Figure 7 depicts the precision at top 5% for each of the algorithms on districts A and B respectively. Random

Model	Mistake Patterns
Random Forests (District A)	If Q4GPA_08 > 3.0 and Abs_Rate_08 > 0.3 and Tardy_08 > 0.4, then Mistake If Q4GPA_07 > 3.0 and Abs_Rate_07 > 0.4 and Tardy_07 > 0.3, then Mistake
Logistic Regression (District A)	If Q4GPA_08 ≤ 2.0 and Q1GPA_08 ≤ 2.0 and Abs_Rate_08 ≤ -0.2 and Abs_Rate_07 ≤ -0.1, then Mistake If Gender_07 = Female and Q4GPA_08 ≤ 2.0 and Abs_Rate_08 ≤ -0.1, then Mistake
Decision Tree (District B)	If GPA_08 ≤ 2.0 and Tardy_08 ≤ -0.1 and Days_Abs_08 ≤ -0.2, then Mistake If EDS_08 = True and GPA_Math_08 > 3.0 and GPA_Science_08 > 2.0, then Mistake
Adaboost (District B)	If Days_Abs_08 > 0.5 and GPA_Science_08 > 3.0, then Mistake If Has_Disability_08 = True and GPA_Science_08 > 3.0 and Days_Abs_08 > 0.2, then Mistake

Table 3: Classifier Mistake Patterns. 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.

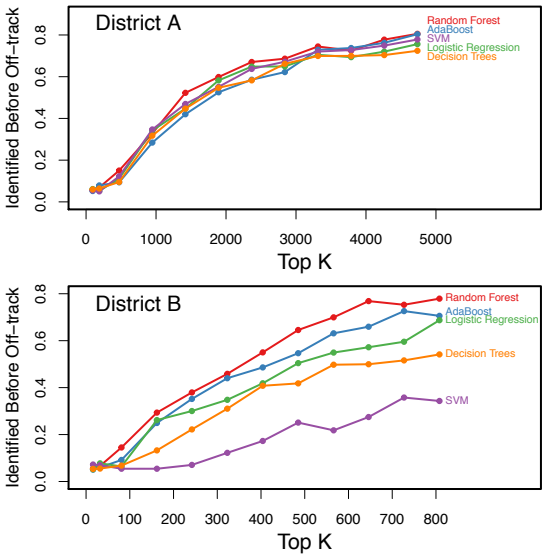

Figure 8: Identification before Off-track

Forest consistently outperforms all other models in both the districts. In the case of District A, the performance improves steadily from 6^{th} to 11^{th} 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 through out 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 if the 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 at diagnosing risk before any undesirable outcome occurs. Note that we exclude all those students who graduate in a timely manner from this calculation.

Figure 8 show the *identification to off-track* metric values across varying K for district A and B respectively. The findings here match our earlier results in that Random Forest model outperforms all the other models for both districts. While Decision Tree turns exhibits poor performance on district A, SVM turns out to the weaker model for district B.

8. CONCLUSION

In this paper, 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 paper 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 paper 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 [16]. 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.

9. ACKNOWLEDGMENTS

The authors would like to thank Amy Hawn Nelson for her help in coordinating with one of the school districts, Ben Yuhas for insightful discussions, and Shihching Liu and Marilyn Powell for assisting us with data logistics.

10. REFERENCES

[1] Building a Grad Nation. http://www.americaspromise.org/sites/default/files/legacy/bodyfiles/BuildingAGradNation2012.pdf.

[2] E. Aguiar, G. A. Ambrose, N. V. Chawla, V. Goodrich, and J. Brockman. Engagement vs performance: Using electronic portfolios to predict first semester engineering student persistence. *Journal of Learning Analytics*, 1(3):7–33, 2014.

[3] E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, and K. Addison. Who, When, and Why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In *Proceedings of the 5th Learning Analytics and Knowledge Conference*. ACM, 2015.

[4] 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.

[5] E. M. Allensworth, J. A. Gwynne, P. Moore, and M. D. L. Torre. Looking forward to high school and college: Middle grade indicators of readiness in chicago public schools. 2014.

[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. 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.

[8] 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.

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

[10] N. V. Chawla and D. A. Cieslak. Evaluating probability estimates from decision trees. In *American Association for Artificial Intelligence*, 2006.

[11] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers. Predicting students drop out: A case study. *International Working Group on Educational Data Mining*, 2009.

[12] E. Er. Identifying at-risk students using machine learning techniques: A case study with is 100. *International Journal of Machine Learning and Computing*, 2(4):279, 2012.

[13] 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.

[14] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *ACM SIGMOD Record*, volume 29, pages 1–12. ACM, 2000.

[15] A. Hershkovitz and A. T. Corbett. Predicting future learning better using quantitative analysis of moment-by-moment learning.

[16] C. Hoxby, S. Turner, et al. Expanding college opportunities for high-achieving, low income students. *Stanford Institute for Economic Policy Research Discussion Paper*, (12-014), 2013.

[17] M. Kendall. *Rank correlation methods*. Griffin, London, 1948.

[18] H. M. Levin and C. Belfield. *The price we pay: Economic and social consequences of inadequate education*. Brookings Institution Press, Washington D.C., 2007.

[19] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on platt's probabilistic outputs for support vector machines. *Machine learning*, 68(3):267–276, 2007.

[20] T.-Y. Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, Mar. 2009.

[21] A. Niculescu-Mizil and R. Caruana. Obtaining calibrated probabilities from boosting. In *UAI*, page 413, 2005.

[22] K. Pittman. *Comparison of data mining techniques used to predict student retention*. ProQuest, 2008.

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

[24] 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.

[25] J. Soland. Predicting high school graduation and college enrollment: Comparing early warning indicator data and teacher intuition. *Journal of Education for Students Placed at Risk*, 18:233–262, 2013.

[26] A. Tamhane, S. Ikbal, B. Sengupta, M. Duggirala, and J. Appleton. Predicting student risks through longitudinal analysis. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1544–1552, New York, NY, USA, 2014. ACM.

[27] University of Chicago. The Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship. http://dssg.uchicago.edu/. Accessed: 2014-10-01.

[28] U.S. Department of Education, National Center for Education Statistics. The condition of education. 2014.

[29] A. Vihavainen, M. Luukkainen, and J. Kurhila. Using students' programming behavior to predict success in an introductory mathematics course. In *The Fourth International Conference on Educational Data Mining*, 2011.

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