A Data-Driven Framework for Identifying High School Students at Risk of Not Graduating on Time

[Extended Abstract]

Reid A. Johnson
University of Notre Dame
Notre Dame, IN 46556
rjohns15@nd.edu

Ruobin Gong
Harvard University
Cambridge, MA 02138
rgong@fas.harvard.edu

Siobhan Greatorex-Voith
Harvard University
Cambridge, MA 02138
sgreatorexvoith@fas.harvard.edu

Anushka Anand
Tableau Research
Seattle, WA 98103
aanand@tableau.com

Alan Fritzler
Northwestern University
Evanston, IL 60208
alanfritzler2015@u.northwestern.edu

ABSTRACT

Some students, for a variety of factors, struggle to complete high school on time. To address this problem, school districts across the U.S. use intervention programs to help struggling students get back on track academically. Yet in order to best apply those programs, schools need to identify off-track students as early as possible and enroll them in the most appropriate intervention. Unfortunately, identifying and prioritizing students in need of intervention remains a challenging task. This paper describes work that builds on current systems by using advanced data science methods to produce an extensible and scalable predictive framework for providing partner U.S. public school districts with individual early warning indicator systems. Our framework employs machine learning techniques to identify struggling students and describe features that are useful for this task, evaluating these techniques using metrics important to school administrators. By doing so, our framework, developed with the common need of several school districts in mind, provides a common set of tools for identifying struggling students and the factors associated with their struggles. Further, by integrating data from disparate districts into a common system, our framework enables cross-district analyses to investigate common early warning indicators not just within a single school or district, but across the U.S. and beyond.

Categories and Subject Descriptors

I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems; K.3.0 [**Computer Uses in Education**]: General

Keywords

Education; Risk Prediction

Bloomberg Data for Good Exchange 2015, NY, USA

1. INTRODUCTION

A perennial challenge faced by school districts is to improve their student graduation rates. Graduation from high school is associated with relatively higher overall lifetime earnings and life expectancy, and lower rates of unemployment and incarceration [3, 2]. Yet, roughly one in five students in the U.S. does not complete high school on time, a rate of over 700,000 students each year [10]. To help more students graduate on time, school districts across the country use intervention programs to help struggling students get back on track academically. However, in order to best apply those programs, schools need to identify off-track students as early as possible for enrolling them in the most appropriate interventions. Further, schools need to know what factors contribute to students being off-track to provide interventions that are focused on individual student needs.

2. CURRENT APPROACH

Traditionally, school districts, administrators, and counselors have been tasked with identifying students likely in need of support. While intellectual ability or academic capability has not been found to vary by demographic factors such as race, sex, or socioeconomic status, failing to complete high school has been associated with such characteristics in research since the 1970s [14], largely indicating social or environmental factors contribute to high school dropout. Lacking other indicators, struggling students are frequently identified by demographic factors combined with simple heuristics, such as attendance and grades [4].

However, these heuristics are not without problems. Due to characteristics that vary across distance and time, the set of heuristics which might help in identifying at-risk students for a particular cohort of students within one school district may not generalize or transfer to other cohorts or schools. Using these heuristics alone also lacks a means of prioritizing which students are most at need for, or most likely to benefit from, intervention. Further, the use of even the most powerful heuristics have been found not to be very predictive by themselves [9]. These issues highlight a need for more generalizable alternatives to these manually created, rule-based systems that can be used to discern predictive indicators.

As an alternative, forward-looking school districts are in-

creasingly exploring data-driven "early warning indicator" (EWI) systems that can help schools find students in need of extra support. Rather than using simple rules, these systems can employ sophisticated analytical methods to combine EWIs into composite factors, with recent work employing methods such as machine learning models [1, 11] and survival analysis [13]. These initial systems have been developed with a focus on individual schools or districts for incorporation into particular district software. Thus far, however, these systems lack the ability to generalize effectively across a broad spectrum of school districts.

3. OUR APPROACH

We are building a framework that delivers accurate and interpretable predictive models of on-time graduation for school districts across the U.S. in order to facilitate focused interventions. This framework is being built in collaboration with several partner school districts with the goal of producing individualized predictions that are expected to perform well in out-of-sample (i.e., testing data) validations.

3.1 Predictive Framework

The framework we are developing is generalizable, and can be adopted by additional schools around the country interested in identifying struggling students. It allows for a flexible set of features at arbitrary granularity, as is commonly observed with multi-source data. Within this framework, we are building a flexible pipeline that enables a variety of data handling and modeling possibilities, including the utilization of heterogeneous sets of cohort data as well as a wide choice of machine learning algorithms. This structure accommodates local school and district effects that may not be able to be captured at the national level. In doing so, we can expand the utility of our framework from one school to many, empowering individual districts to build models customized to their own schools.

3.2 Feature Engineering

In addition to identifying struggling students reliably, it is important to provide information that can be used to understand why an individual student is struggling. To supplement machine learning techniques with this interpretability, we are constructing a structured, potentially hierarchical class of features that naturally map to intuitive risk categories. Each feature will be engineered to belong to one or more well-established feature categories with straightforward interpretations, such as student mobility, language, personality and motivation, in addition to family, community, and social-level factors (for examples, see [4] and [15]). Factor analysis can be used to evaluate the degree to which a feature belongs to a given feature category via factor loadings. Thus, the feature weights or importances generated by a machine learning model, along with each feature's category membership, can be used to decompose a prediction into scores associated with each risk category, providing a substantive basis for suggested focused interventions.

4. DATA

The data for this project comes out of a partnership with four school districts: Arlington Public Schools (Arlington, VA), Cabarrus County Schools (Cabarrus, NC), Vancouver Public Schools (Vancouver, WA), and Wake County Public School System (Wake County, NC). Each of these districts has already recognized the importance of EWI systems for identifying at-risk students, with rule-based early warning indicator systems in place that use several important indicators such as academic performance, behavior, mobility, and demographics. 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.

Each district provided de-identified historical data for at least two cohorts that describes current and past student performance, from elementary or middle school onward.

- Arlington Public Schools (APS): APS is situated in the mid-Atlantic region, with a current enrollment of 26,000 students across 31 schools. All of the APS high schools were rated in the top 2% of high schools nationally in the 2014 "Ranking America's High Schools" issue by The Washington Post.

- Cabarrus County Schools (CCS): situated in the southeast, CCS currently enrolls approximately 30,000 students across 39 schools.

- Vancouver Public Schools (VPS): located in the Pacific Northwest, VPS has a current enrollment of 23,000 students across 35 schools and is considered "high-mobility," with one-third of its students enrolled part-year. We use data from two recent cohorts, with records spanning from the 6th through 12th grades.

- Wake County Public School System (WCPSS): located in the south, WCPSS is the nation's 16th largest school district, currently enrolling approximately 155,000 students across 171 schools. We use a sample of 16 years worth of enrollment data, beginning with students enrolled in the 6th grade in 1999.

The datasets contain several attributes for each of these students such as their course enrollment and grades, absence rates, tardiness, and so on. The majority of students in each cohort graduated high school within four years of enrollment. However, some transferred into or out of the district during the study period and therefore have missing data fields for years prior to (or post) their enrollment in the district.

5. METHODS & EVALUATION

Our goal is to make nuanced predictions for individual schools within each of our partner districts, while also generating a flexible and generalizable codebase that can be expanded to additional districts. To this end, we have used systems that provide for extensibility and scalability.

5.1 Methods

As our analyses employ relatively large datasets that are relational in nature, we have loaded the provided data into a SQL-based database management system (DBMS), which provides a consistent, useful framework for organizing and manipulating our data. To facilitate a process that can be used across a variety of school districts, we have organized the data into a single database schema that is consistent across our district partners and that captures the relationship between students and schools within each district.

Given the data provided, we have framed our prediction problem as the following binary classification task: d years before a given student's expected graduation date (where $d = 1$, 2, or 3), we predict whether he or she will graduate on time. To generate predictions, we have elected to use a series of different classifiers, including logistic regression, naïve Bayes, random forest, and support vector machine.

Logistic regression (LRC) is a type of regression analysis used for predicting the outcome of a categorical variable. The method is used widely in many fields, including the medical and social sciences. Naïve Bayes (NB) is a simple probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumptions. In general terms, the method assumes that given the class, the presence or absence of a particular feature is unrelated to the presence or absence of any other feature. Despite this strong assumption, the method often performs quite well, particularly as it only requires a relatively small amount of training data to estimate the parameters necessary for classification [8]. Random forest (RF) is an ensemble classifier that consists of many classification trees. Each classification tree is fit to a bootstrap sample of the data, but at each node, only a small number of randomly selected variables are available for the binary partitioning of the tree. The trees are fully grown and the predicted class of an observation is calculated by the majority vote from the ensemble for that observation [5]. Support vector machine (SVM) searches for an optimal separating hyperplane capable of discriminating between classes. This is accomplished by non-linearly mapping the input features into a high-dimensional feature space, wherein a linear decision surface is constructed [6].

These methods were selected due to their predictive ability and ease of implementation. Altogether, they are fairly representative of the diverse array of classification methods traditionally employed in the machine learning domain. Our work uses the implementations provided by scikit-learn [12].

5.2 Evaluation

In developing our evaluation methodology, we were careful to appropriately account for the inherent temporal dependencies in our data. As new cohorts of students begin school every year, spontaneous fluctuations occur in the demographic composition and outcome distributions of these students, as well as in the joint behavior of any subset of these factors. Consequently, traditional cross-validation methods (including leave-k-out and k-fold) that rely solely on random splits of the data actually lead to biased (underestimated) prediction error estimates that provide an overly optimistic view of the fitted model. In essence, this means that the data must be divided temporally, where models are evaluated only on future data, in order to prevent what amounts to "cheating by peeking into the future."

As we emphasize out-of-sample (i.e., test data) predictive performance, we have adopted the approach of temporal model validation with a sliding window. Consider a "d-years-ahead" predictive model: that is, the prediction of a student's graduation outcome d years before his or her expected graduation. For each cohort k, we observe dropout outcomes $d + 1, \ldots, n$. The cohort itself, k, as well as all of its subsequent cohorts, $k + 1, \ldots, n$, may be used as the testing set. Naively, all previous cohorts, to the model fitted on its preceding cohorts, $1, \ldots, k - 1$. However, if we assume that a d-years-ahead model cannot be validated until

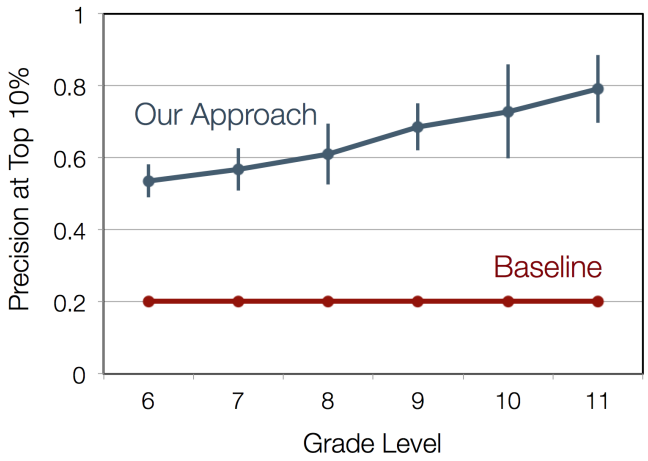

Figure 1: Performance results for random forest model predictions on VPS data. A model is generated for each grade 6–11, and each model is evaluated on the precision at the top 10%. The baseline is computed as the rate of not-on-time graduation. Error bars represent standard deviation.

d years into the future, then in practice, labels will not be available for cohorts more recent than $k-d$. Thus, provided with a sufficient number of cohorts, a d-years-ahead model should be fitted only on the preceding cohorts $1, \ldots, k - d$. This results in $n - d$ sets of cross-validation results. The final fitted model for delivery may be fitted on all cohorts from 1 through n.

We evaluated each models' performance by computing the precision at top k%, which is the predictive accuracy within the models' top k% most confident predictions (i.e., those with the highest probability scores). Within these predictions, we determined what percentage of students within the top 10% actually did not graduate on time. This metric reflects the reality that many schools can only intervene on a relatively small percentage of their entire student body.

6. RESULTS

We provide preliminary results for the data provided by VPS. For each grade level from 6 to 11, we fit and evaluate a model. As the data provided by VPS is limited to two recent cohorts, we fit all models on the first cohort and evaluate all models on the second cohort. Though we employed a variety of standard machine learning classification methods, including logistic regression, random forest, support vector machines, and variants thereof, here we summarize only the results obtained using random forest.

Figure 1 summarizes our results for VPS. For each grade level, we evaluate the predictions using precision at the top 10%. In other words, provided with all of the student predictions generated by a model for a given grade level, we determine the top 10% of students predicted most likely not to graduate on time and compute what the fraction of these students actually did not graduate on time. The final value is the precision at the top 10% for the given grade level.

We observe that our method performs well above the baseline rate of students who do not graduate on time. Even our grade 6 model, which predicts an outcome six years into

the future, produces a precision greater than 50%. We also observe that as the prediction timeframe decreases, the precision increases, with the best performance exhibited by our grade 11 model.

While our data-driven, classification-based methodology appears promising, we note that it produces only one fitted model for each grade level. Thus, while our methodology may provide deep insights at the cohort, school, and district level, it does not explicitly provide actionable insight at the student level. We leave such extensions to future work.

7. FUTURE WORK

Our current work represents only approach and a small step at addressing the problem identifying students at risk of not graduating on time. Accordingly, several straightforward avenues of further investigation exist.

First, while our current methodology does not provide explicit student-level insights, it could be easily extend to do so. For example, actionable knowledge discovery could be used to identify features that are key to a student being predicted as not graduating on-time. Indeed, existing work by Cui et al. [7] outlines a method for optimally extracting actionable knowledge from random forests, the primary method discussed in our work. Alternatively, students could be clustered to identify groups with particular risk factors.

Additionally, our methodology does not provide explicit insight into when an intervention should be applied for optimal effect. Future work could investigate the usefulness of models from the survival analysis literature. For example, Cox regression could be readily applied to provide insight into when a student is likely to be at high risk [1].

Finally, while our methodology produces a probability score for each student, further work is needed to ensure that these probability scores are well-calibrated to actual risk propensities, particularly across cohorts and districts.

8. CONCLUSION

We hope the proposed framework will enable our district partners to identify not only which students are struggling, but why they are struggling, making focused interventions possible early enough in the process to help struggling students to graduate on time. Employing a data-driven machine learning approach to the problem of identifying at risk students has a unique advantage over the current approach, in that it may identify broader patterns in high school drop out nationally. While this work is still in an early stage, we foresee a broader application of our generalized framework and modeling approach beyond our partner districts.

9. ACKNOWLEDGMENTS

This work was done as part of the Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship at the University of Chicago. We want to thank Kerstin Frailey for the time and effort she contributed to an early version of this work. We would also like to thank Scott Goldman (Coordinator of Accountability at Arlington Public Schools), Matthew Lenard (Director of Data Strategy and Analytics at Wake County Public Schools), Amy Nelson (Director of Social Research for the UNC Charlotte Urban Institute), and Paul Stern (District Enterprise Analyst at Vancouver Public Schools) for providing us with their data, experience, and time. This work would not be possible without them.

10. REFERENCES

[1] E. Aguiar, H. Lakkaraju, N. Bhanpuri, et al. 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 International Conference on Learning Analytics and Knowledge (LAK'15)*, pages 93–102. ACM, 2015.

[2] F. Alivernini and F. Lucidi. Relationship between social context, self-efficacy, motivation, academic achievement, and intention to drop out of high school: A longitudinal study. *The Journal of Educational Research*, 104(4):241–252, 2011.

[3] C. R. Belfield and H. M. Levin. *The Price We Pay: Economic and Social Consequences of Inadequate Education*. Brookings Institution Press, 2007.

[4] 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.

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

[6] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[7] Z. Cui, W. Chen, Y. He, and Y. Chen. Optimal action extraction for random forests and boosted trees. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*, pages 179–188. ACM, 2015.

[8] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2-3):103–130, 1997.

[9] C. D. Jerald. Identifying potential dropouts: Key lessons for building an early warning data system. a dual agenda of high standards and high graduation rates. *Achieve, Inc.*, 2006.

[10] G. Kena, L. Musu-Gillette, J. Robinson, et al. *The Condition of Education 2015*. (NCES 2015-144). U.S. Department of Education, National Center for Education Statistics, Washington, D.C., 2015.

[11] H. Lakkaraju, E. Aguiar, C. Shan, et al. A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*, pages 1909–1918. ACM, 2015.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] P. M. Radcliffe, R. L. Huesman Jr., and J. P. Kellogg. Modeling the incidence and timing of student attrition: A survival analysis approach to retention analysis. 2006.

[14] R. W. Rumberger. *Dropping Out: Why Students Drop Out of School and What Can be Done*. 2011.

[15] G. T. Wodtke, D. J. Harding, and F. Elwert. Neighborhood effects in temporal perspective the impact of long-term exposure to concentrated disadvantage on high school graduation. *American Sociological Review*, 76(5):713–736, 2011.