Machine-Learned School Dropout Early Warning at Scale

S. Thomas Christie*
thomas.christie†

Daniel C. Jarratt*
daniel.jarratt†

Lukas A. Olson*
lukas.olson†

Taavi T. Taijala*
taavi.taijala†

ABSTRACT

Schools across the United States suffer from low on-time graduation rates. Targeted interventions help at-risk students meet graduation requirements in a timely manner, but identifying these students takes time and practice, as warning signs are often context-specific and reflected in a combination of attendance, social, and academic signals scattered across data sources. Extremely high caseloads for counselors compound the problem. At Infinite Campus, a large student information system provider, we modeled statistical relationships between student educational records and enrollment outcomes, using de-identified records and in-system analysis to guarantee student data privacy. The resulting risk scores are highly predictive, context-sensitive, nationally available, integrated into the existing student information system, and updated daily.

1. INTRODUCTION

Approximately 15% of American students do not graduate high school on time [16]. States and districts frequently employ interventions designed to improve educational outcomes, including reducing dropout rates. A key role of school counselors is to direct the application of these interventions to the students who need them most. Counselors first need to identify these students but are faced with information overload. Each student's data is distributed across a student information system (SIS) and often other systems or people, making it difficult to synthesize into an accurate, comprehensive portrait of a student's risk. Compounding the problem are extremely large counselor caseloads—the national average is 430 students per counselor—with higher numbers typical in schools serving children with other structural disadvantages [11].

Early warning systems function as "automated attention" for overworked counselors by automatically identifying students who might benefit from additional institutional resources. They automate the more tedious data analysis and summarization tasks so that counselors can focus on what humans do best: building relationships.

*Infinite Campus, Blaine, Minnesota, USA
†@infinitecampus.com

S. Thomas Christie, Daniel Jarratt, Lukas Olson and Taavi Taijala "Machine-Learned School Dropout Early Warning at Scale" In: *The 12th International Conference on Educational Data Mining,* Michel Desmarais, Collin F. Lynch, Agathe Merceron, & Roger Nkambou (eds.) 2019, pp. 726 - 731

1.1 Alternate approaches

An effective dropout prevention system requires developing people, processes, and technology [7], identifying valid predictors, managing data and reports, assigning interventions, and monitoring student progress [10, 15]. In this paper, we focus just on the technology that identifies risk factors and estimates student dropout risk, which can then be embedded in a larger dropout prevention system.

Quantitative and qualitative determination of school dropout risk factors is a decades-old area of research [19], though the mid-2000s were a particularly important inflection point. High-profile studies of dropouts in the Chicago [3] and Philadelphia [17] urban districts led to the development of statistical methods for determining risk factors and their incorporation into early warning systems. Several organizations, often working together, have been instrumental in encouraging American schools to adopt research-based best practices [10], including the U.S. Department of Education's Regional Education Laboratories [18], the Consortium on School Research [2] and NORC [7] at the University of Chicago, the American Institutes for Research [8], and the Everyone Graduates Center at Johns Hopkins University [9]. Most states now make an early warning system available to their school districts [6].

Until the mid-2010s, all widely-used dropout early warning systems used threshold-based models, characterized by a few easily comprehensible predictors with associated risk thresholds (e.g., failing at least one course or being absent at least 20 days). The simplicity and auditability of these models, and the associated ease of implementation using common software and spreadsheet skills, is their key advantage over machine-learned systems. Students are measured on each predictor and flagged as "on-track" or "off-track" based on which side of a preset threshold their data point falls. Staff can intervene with students who have the most "off-track" risk flags, or whose risk areas correlate with particular intervention domains. A particularly influential approach is Balfanz's "ABC" taxonomy, in which students are measured on **a**ttendance, **b**ehavior, and **c**ourse performance metrics [4], optionally with different thresholds for different student subpopulations [12]. While some educational institutions

implemented an ABC-style system themselves [10], others used spreadsheets or data tools made available by organizations like American Institutes for Research [8]. An active area of research involves determining which predictors and thresholds are appropriate for each school, or whether single thresholds are appropriate at all.

To overcome the limitations inherent in threshold-based systems, we and other organizations created machine-learned dropout risk identification systems in the mid-2010s. A number of researchers describe their systems in the academic literature (e.g., [1]) or industry white papers (e.g., [20]). Several organizations serve machine-learned dropout predictions at the scale of hundreds of thousands of students in many school districts. The Wisconsin Department of Public Instruction's Dropout Early Warning System uses data reported to the state every few months by the districts' SISs to build machine-learned models. The system produces two predictions per year and is available for students in grades 6–9 [14]. Mazin Education (through BrightBytes) and Hoonuit both sell machine-learned early warning and intervention monitoring systems for all grade levels.

Machine-learned systems have two major advantages over threshold-based models. First, the additional model complexity affords more accurate predictions and allows system designers to infer which risk factors are predictive in the presence of other factors or for different populations. Second, the variety of model architectures allows for more than just inferring overall risk. Designers can choose, for instance, to model time until dropout so that staff can intervene according to acuteness of risk, or to model uncertainty.

1.2 Our contribution

Two key obstacles prevent machine-learned early warning systems from being deployed nationwide. First, the predictive quality of these models is chiefly a function of data availability, as models must be trained on a large dataset—including a variety of educational contexts and outcomes—to ensure they perform well for students they haven't seen before. Models built on a single district's or even state's data may not generalize well to other populations. However, building a model on multiple states' data requires the data to be standardized, and without a common SIS to enforce uniformity, manipulating data into a common format is costly and time consuming.

Second, predictions must be surfaced to educators in a frequent and easy-to-use manner, which means the most successful systems will be closest to existing daily workflows. In some existing systems, staff members must log into a separate software program to access risk scores. In others, there are months between score updates. Timely prediction is important; the sooner a school is aware of a student at risk, the more time it has to intervene.

Infinite Campus provides a large student information system and has made significant investment in education data localization and standardization, reporting and warehousing, and user workflows for American K–12 education. Our role in the industry positions us to address the key remaining gaps in early warning systems using centralized data warehouses, standardized data, and placement of the early warning application into the existing SIS.

Contributions
Available nationally in 32 states yet contextual to each child's educational environment
Useful predictions: highly predictive, daily updates, four risk category scores, with consistent of predictive quality between protected student groups
Integrated into the student information system with no imports, exports or synchronization necessary

In the following section we describe our implementation of a dropout early warning system on more than 6 million student-years of educational records across 32 states. Our overall dropout risk score has excellent predictive quality with an AUC of 0.941 (see table 1 for additional quality metrics), and includes additional machine-learned scores that help counselors understand the source of a student's risk to guide which interventions may be appropriate. Risk scores, delivered automatically and updated daily, are integrated into our existing SIS and available to counselors as an enhancement to their existing workflows.

2. IMPLEMENTATION

System design. Student data is stored in a number of relational SQL Server databases for school staff to create, read, update, and delete educational records. These databases exist in Infinite Campus's fully owned and operated Tier 4 data centers that fulfill security requirements of the U.S. Department of the Interior. Student records exist in a variety of data structures and are recorded with varied time granularity for 45 states and the U.S. federal government, a portion of which we use for early warning. Because our model architecture requires a common data structure, we aggregate student records into a fixed format with one row per student-year, where 'year' corresponds to an academic year. Aggregated data is periodically transferred to a central repository in the same data center. Data for past students whose educational outcomes are known (e.g., graduation or dropout) is then used to build a machine-learned model relating summarized educational records to student enrollment outcomes. The model, along with summarized properties of each district, school, and geographic area present in the dataset, is deployed behind an API. Each day, records for currently enrolled students are aggregated, de-identified, and sent to the API, which returns risk scores for each student. Figure 1 illustrates this architecture. The returned risk scores and a score history is made available to counselors via the SIS user interface. By integrating and automating the process of generating and updating risk scores into the SIS, we relieve dropout prevention teams from the burden of collecting, storing, and analyzing the source data themselves.

What we predict. Each student-year of aggregated educational records is tagged with one of three labels: "needs early warning", "does not need early warning", or "ignore for early warning". A student-year is labeled with "needs early warning" if the student's records include known undesirable outcomes during the year in question or future years. For example, records for an 8th grader in 2014 would be labeled

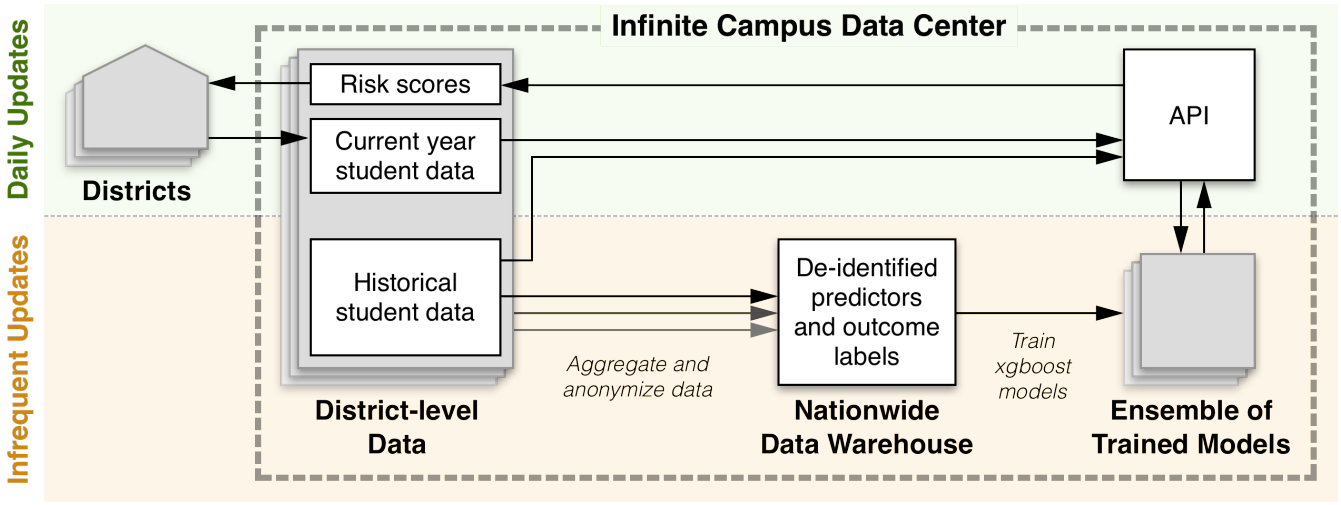

Figure 1: Schematic of architecture describing data flow for model training and prediction. Model training is performed infrequently, while risk scores are recalculated daily to incorporate new information. All data remains in the Infinite Campus data center and is not transferred to third-party servers. Identifying information is removed before transfer between systems.

"needs early warning" if they dropped out in 2016. We define undesirable outcomes based on enrollment end status codes identified by states as indicating school dropout, and expand the definition to include unsatisfactory academic progress (i.e., retention or demotion) and expulsion or other removal. A student-year is labeled with "does not need early warning" if the student had no undesirable enrollment outcomes during the year in question or future years, *and* we can confirm that the student persisted to graduation. If we do not know whether the student persisted to graduation—the student transferred outside of our system or is still enrolled—then the student's data is censored and we lack ground truth labels for it. Each of these students' years are labeled with "ignore for early warning" and excluded from training and evaluation. To translate state-specific enrollment status codes to outcome categories, codes were mapped by three independent raters, then differences were reconciled and validated by comparing resulting outcome rates across districts. Data from a school district are removed from training if that district has abnormally high "needs early warning" or "ignore for early warning" rates, as these conditions may indicate underlying inconsistencies in record-keeping that warrant further investigation. In addition, student-years are removed from training if the student's cohort is not scheduled to have graduated yet, in order to remove label bias in earlier grades.

Our data collection and labeling process produces approximately 6.4 million rows of labeled data. We use 45% of rows for model training, 5% for model validation (to determine training stopping points), and 50% for final quality evaluation, split by student. Roughly 16% of training rows are labeled as "needs early warning".

Predictors. Our training set is produced by collecting and summarizing educational information from the core SIS database. This summarized information relates to attendance, academic performance, behavior, household and enrollment stability, and other items. We chose predictors that are consistent across states and districts and that are supported by the dropout prevention literature (e.g., [19]). Where data is localized, we employ experts who communicate directly with stakeholders in districts and states to ensure we understand the unique characteristics of local use and law. Attendance information includes the proportion of class time a student was actually present, as well as absences grouped by type of excuse. Academic performance information includes the proportion of course grades attributed to each letter grade, overall high school GPA, and the proportion of attempted credits successfully earned. Behavior information includes the number of behavior infractions and resolutions, as well as whether weapons, drugs, or harassment were involved. Household and enrollment stability information includes the presence of past undesirable enrollment outcomes and how often the student changes home addresses, schools, or districts in the middle of school years. Finally, we include contextual information such as age and grade level. In total, we have approximately 70 distinct predictors per year, and each student-year row includes the current and previous year's data.

One core design goal for our system was to have a nationwide statistical model that is sensitive to local and contextual factors. We achieved context-specific performance with two types of feature engineering: including subpopulation aggregation features and calculating interactions among a student's personal features as well as the subpopulation aggregates that apply to them. For example, a student's attendance or academic data *relative to their peers* in a given group may carry information about risk. To allow for this possibility, we calculated two types of summary statistics for each school, district, and ZIP code in our dataset. For numeric predictors, we calculated the mean value per group. For categorical predictors, we calculated the proportion-per-group of each category. These group-level contextual features allow us to capture signals about students' environ-

ments that are more informative than simple group-membership variables. The group-level information about each student-year's specific district, school, and ZIP code is joined on each student-year row for model building and prediction.

By including a wide range of features characterizing individual students and their educational environments, we allow the machine learning model to determine the significance of each feature and the relationship between features as they relate to risk. The wide range of potentially predictive features is simply not available to most non-SIS vendors. By using a relatively complex feature set and modeling architecture, we are able to capture complex contextual relationships between students and their environments.

Explanatory scores. While machine learning affords high predictive accuracy and the ability to capture complex relationships between predictive features, this comes at the cost of reduced model interpretability. In early conversations with customers, we were frequently asked *why* a student received a given risk score. To answer this question, we supplement our overall risk score with two types of explanatory 'category' scores that provide insight into which parts of a student's record are contributing to their overall risk.

The first set of explanatory scores is based on Balfanz's "ABC" categories [10]—attendance, behavior, and course performance—and an additional "stability" score including measures of household and enrollment stability. Each of the four scores is produced by a separate model trained only on predictors from its respective category. By partitioning the predictors according to category, we in turn are able to disentangle the impact each category has on the overall risk score. The predictive quality of category scores is necessarily lower than the overall risk score, because the category-specific models use a strict subset of the overall model's predictors. However, these scores indicate whether each category of a student's predictors, when taken by itself, is characteristic of a student with undesirable future enrollment outcomes.

In addition to scores built on subsets of predictors from each category, we also build scores for each category using a "counterfactual" approach. That is, if a student's records improved in a certain area (but the rest of the student's records stay the same), how would their risk change? To answer this question, we replace the values for the "actionable" predictors in a category with values corresponding to exemplary performance. The resulting data represents an attainable ideal for each student—if he or she attended every class, earned perfect scores on every assignment, or never behaved inappropriately. This data is used to produce a counterfactual risk score for each of the four categories, which when subtracted from the student's actual overall risk score indicates the potential "room for improvement" in each category; these are our final four explanatory scores.

While picking values corresponding to exemplary performance appears intuitive (e.g., 4.0 GPA, 100% attendance), using them to artificially modify student data has the potential to push the resulting data points outside the space of training examples, leading to unpredictable model behavior. Preliminary analysis found this to be a problem for some "obvious" exemplary values, leading us to select values experimentally instead. For each feature, we used a statistical model to find the optimal bin (range of values) corresponding to the lowest predicted risk, which we subsequently validated by checking that the proportion of actual dropouts was lowest for this bin. We chose a reasonable value from each optimal bin to represent exemplary performance.

2.1 Modeling technique

The system described here must operate at scale within an industry setting and be robust to messy and missing data. To achieve this, we use the `xgboost` package [5] for modeling, which constructs a series of simple decision trees. Unlike logistic regression or neural networks, `xgboost` is robust to the presence of outliers and appropriately handles missing values. The decision-tree structure of model components provides an integrated way to capture contextual relationships between individual predictors and group-level aggregates. `xgboost` supports parallelization of model training, so training scales well on enterprise server hardware. We use the 'binary:logistic' training objective, so that `xgboost` models produce the probability of a student-year being labeled as "needs early warning". We use the area under the receiver operating characteristic curve (AUC) as the `xgboost` evaluation metric. AUC measures the quality of sorting produced by the model, with a high value for AUC indicating that the model is correctly assigning higher probabilities to student-years labeled as "needs early warning" than to those labeled as "does not need early warning".

Our modeling strategy ensures robustness to noise in two ways. First, we heavily regularize our `xgboost` models by using a small tree depth and relatively few training rounds, reducing over-fitting and making the model more robust to small changes in student data (both during training and during prediction). Regularization makes it possible to provide high-quality predictions for unseen data, such as a student in the evaluation dataset or a new customer whose data was not included in model training. Second, for each score type we train an ensemble of 11 to 25 `xgboost` models, and use the median prediction of all models. This technique further reduces variability between model deployments.

Predictions. Predictions are refreshed daily by aggregating educational records of currently enrolled students and sending those aggregates to our API, as illustrated in Figure 1, which provides GRAD scores and category scores back to the SIS. This technique allows us to provide score updates more frequently than competitors, and eliminates the requirement for school districts to transfer or analyze data on their own. The SIS then displays risk scores to staff members that have been given access to the early warning tool by district administrators.

As described above, we train the model using aggregations from past entire student-years. However, daily predictions are made for currently enrolled students, whose current year records contain only a partial year of data. We used several methods to mitigate the mismatch with our training dataset. First, in addition to aggregates summarizing a student's data from the current academic year, we also in-

clude aggregates from the student's previous year as predictors. This allows the model to observe student data for an already-completed year and it affords analysis of year-over-year changes. Second, some data is in the form of event counts that accumulate throughout the year, such as the number of missed periods or the number of behavior resolutions. This data is converted into a rate, such as missed periods per instructional day. Rates are then directly comparable at all points in the academic year. Since rate-converted values are sensitive to small changes at the beginning of the school year, we instead use an 'estimated rate' calculated as a weighted combination of the previous and current year's rates. After about four weeks of school, the previous and current year's rates are equally weighted, with the current year's rate weighted more heavily after each additional day.

We convert the probability output from `xgboost` into a user-facing "GRAD score" that ranges from 50 to 150, where 50 indicates high likelihood of undesirable enrollment outcomes in the future and 150 indicates high likelihood of persistence to graduation. We also considered that counselors may have a greater need to distinguish between students with low and moderate risk rather than between students with high and very high risk. That is, a 0.1 change in dropout probability from 0.05 to 0.15 is more important than a change in probability from 0.75 to 0.85. We therefore transform the raw probabilities to 'spread apart' students at the low end of the probability range (low risk), while compressing probabilities at the high end of the range (high risk).

2.2 Model evaluation

We evaluated our overall and subscore models on an evaluation set containing approximately 3.2M student-years (50% of the total dataset) that were not used for model building or validation during training. Results are listed in Table 1 and represent, as far as we are aware, the highest predictive quality in the industry. In Table 2, we also list evaluation results for our overall model by protected subpopulation [13]: sex, race/ethnicity, grade level, and free/reduced meal eligibility (a proxy for socioeconomic status).

Because counselors do not see predicted probabilities, but rather ordered GRAD scores, we chose the area under the receiver operating characteristic curve (AUC) metric that measures whether students' predictions are ordered in terms of actual risk. The AUC effective range is 0.0 (perfectly inversely sorted) to 1.0 (perfectly sorted), with 0.5 indicating random predictions.

Futhermore, because counselors will give additional institutional resources to the few percent of students predicted most at-risk, we chose precision and recall metrics that measure whether how well that most at-risk prediction category actually contains at-risk students. We evaluated precision at 10% (P@10) and recall at 10% (R@10) following the literature [1]. A key limitation of precision@k and recall@k occurs when k is less than the population's condition-positive rate, and therefore the effective range of those metrics is less than 1.0. To correct for this limitation, we also measured precision=recall at 16% (PR@16) because the condition-positive ("needs early warning") rate of the training set is 0.158. For subpopulation evaluation, precision and recall at baseline (PR@b) is based on that subpopulation's own condition-positive ("needs early warning") rate. The effective range of PR@16 and PR@b is 0.0 (completely incorrect) to 1.0 (completely correct), with the random prediction rate equivalent to the baseline rate for that population.

Model	AUC	P@10	R@10	PR@16
GRAD Score	0.941	0.865	0.549	0.719
Academics	0.914	0.825	0.524	0.682
Attendance	0.852	0.648	0.411	0.547
Behavior	0.808	0.582	0.368	0.493
Stability	0.860	0.654	0.415	0.548

Table 1: Risk score quality evaluation

Subpopulation	+ rate	AUC	P@10	R@10	PR@b
Female	0.130	0.935	0.770	0.597	0.683
Male	0.185	0.941	0.921	0.498	0.742
Hispanic	0.207	0.925	0.923	0.449	0.725
Asian	0.065	0.927	0.470*	0.712*	0.620
AIAN	0.275	0.927	0.972	0.354	0.771
NHPI	0.096	0.935	0.605*	0.676*	0.641
2+ races	0.185	0.935	0.909	0.487	0.717
White	0.134	0.937	0.790	0.592	0.692
Black	0.268	0.940	0.986	0.371	0.783
Not stated	0.136	0.951	0.871	0.639	0.762
6^{th} grade	0.218	0.896	0.918	0.421	0.695
7^{th} grade	0.212	0.910	0.924	0.434	0.709
8^{th} grade	0.207	0.921	0.921	0.444	0.716
9^{th} grade	0.232	0.937	0.978	0.423	0.779
10^{th} grade	0.169	0.937	0.888	0.529	0.727
11^{th} grade	0.112	0.940	0.730	0.654	0.692
12^{th} grade	0.079	0.953	0.573*	0.728*	0.649
NSLP: Free	0.252	0.919	0.961	0.385	0.737
NSLP: Reduced	0.130	0.920	0.745	0.579	0.652
NSLP: Paid	0.097	0.942	0.683*	0.700*	0.690
NSLP: N/A	0.125	0.942	0.788	0.634	0.711

Table 2: Overall model risk score quality for subpopulations. '+ rate' refers to the baseline 'condition-positive' negative enrollment outcome rate for that subpopulation's current or future enrollments. NSLP refers to the National School Lunch Program. AIAN refers to American Indian and Alaska Native. NHPI refers to Native Hawaiian and Pacific Islander. * means that the '+ rate' value is less than 0.1 and therefore the effective maximum range of P@10 and R@10 is less than 1.0 for that subpopulation.

3. LIMITATIONS AND EXTENSIONS

The use of year-level aggregates in model training erases temporal relationships between individual event records, making the system relatively blind to *patterns* of individual events within an academic year. To address this limitation, we are exploring alternate modeling strategies capable of ingesting event-based data streams that are both more granular and of non-uniform length. A second limitation is our model's focus on grades 6–12. Interventions are most successful when they are applied early [21]. The ability to provide meaningful risk indicators for younger students could significantly improve outcomes by helping counselors target interventions toward the students who need them most, at the point they can ben-

efit from them most. To do this, we must overcome several data-consistency related obstacles, the most pressing being a lack of long-term datasets that are consistent in the type of information collected over time and the quality/reliability of the collection.

Finally, although target labels were created using inter-rater reliability methods, research on state policies, and student outcome data, the labels have not been verified by representatives from each school district who could personally attest to the accuracy of a given student's outcome. We intend to make the target calculation available to schools and to implement a system for users to provide feedback on our product's predictive accuracy to allow us to verify our labels and to continue to improve the quality of predictions.

4. CONCLUSION

We built a decision support system that provides high-quality, context-sensitive risk predictions and is integrated into an SIS that thousands of counselors already use in their workflows. In doing so, we offer daily risk assessments to millions of currently enrolled middle and high school students across the country. By automatically identifying students who may benefit from additional institutional resources in the service of timely graduation, we fulfill a key component of the dropout prevention process.

5. ACKNOWLEDGMENTS

We gratefully acknowledge all teams at Infinite Campus who have supported our work. We offer a special thanks to the Kentucky Department of Education and Kentucky educators for early partnership and feedback in this project.

6. REFERENCES

[1] E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, and K. L. 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 Fifth International Conference on Learning Analytics And Knowledge*, pages 93–102. ACM, 2015.

[2] E. Allensworth. The use of ninth-grade early warning indicators to improve Chicago schools. *Journal of Education for Students Placed at Risk (JESPAR)*, 18(1):68–83, 2013.

[3] E. M. Allensworth and J. Q. Easton. What matters for staying on-track and graduating in Chicago public high schools: A close look at course grades, failures, and attendance in the freshman year. Research report. `https://eric.ed.gov/?id=ED498350`, 2007.

[4] R. Balfanz, J. H. Fox, J. M. Bridgeland, and M. McNaught. Grad nation: A guidebook to help communities tackle the dropout crisis. `https://files.eric.ed.gov/fulltext/ED505363.pdf`, 2009.

[5] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[6] College & Career Readiness & Success Center at American Institutes for Research. State profile comparison. `https://ccrscenter.org/ccrs-landscape/state-profile/new-state-profile-comparison`.

[7] L. Cordeiro, R. Blank, L. Hansen, D. Leeds, and L. Selfa. Considering the best early warning system (EWS) model to fit your needs, May 2016.

[8] A.-M. Faria, N. Sorensen, J. Heppen, J. Bowdon, S. Taylor, R. Eisner, and S. Foster. Getting students on track for graduation: Impacts of the early warning intervention and monitoring system after one year. `https://eric.ed.gov/?id=ED573814`, 2017.

[9] J. H. Fox, E. S. Ingram, and J. L. Depaoli. For all kids: How Kentucky is closing the high school graduation gap for low-income students. `https://eric.ed.gov/?id=ED572766`, 2016.

[10] S. Frazelle and A. Nagel. A practitioner's guide to implementing early warning systems. *Regional Educational Laboratory at Education Northwest*, 2015.

[11] D. J. Gagnon and M. J. Mattingly. Most US school districts have low access to school counselors: Poor, diverse, and city school districts exhibit particularly high student-to-counselor ratios. `https://scholars.unh.edu/cgi/viewcontent.cgi?article=1285&context=carsey`, 2016.

[12] Key data for an "early warning system" with on- and off-track indicators that become the basis for tiered interventions. `http://guidebook.americaspromise.org/tools-directory`.

[13] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 100–109, New York, NY, USA, 2019. ACM.

[14] J. E. Knowles. Of Needles and Haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *JEDM | Journal of Educational Data Mining*, 7(3):18–67, July 2015.

[15] Minnesota Dept. of Education. MEIRS 2.0 Minnesota early indicator and response system guide, July 2018.

[16] National Center for Education Statistics. High school graduation rates. `https://nces.ed.gov/fastfacts/display.asp?id=805`.

[17] R. C. Neild and R. Balfanz. Unfulfilled promise: The dimensions and characteristics of Philadelphia's dropout crisis, 2000-2005. `https://eric.ed.gov/?id=ED538341`, 2006.

[18] I. R. E. L. Program. Early warning systems. `https://ies.ed.gov/ncee/edlabs/projects/ews.asp`.

[19] R. W. Rumberger and S. Lim. Why students drop out of school: A review of 25 years of research, Oct 2008.

[20] M. Technet. ML predicts school dropout risk & boosts graduation rates, June 2015.

[21] J. A. Temple, A. J. Reynolds, and W. T. Miedel. Can early intervention prevent high school dropout? evidence from the Chicago child-parent centers. *Urban Education*, 35(1):31–56, 2000.