Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin

Jared E. Knowles *
Wisconsin Department of Public Instruction

March 3, 2015

The state of Wisconsin has one of the highest four year graduation rates in the nation, but deep disparities among student subgroups remain. To address these disparities the state has created the Wisconsin Dropout Early Warning System (DEWS), a predictive model of student dropout risk for students in grades six through nine. The Wisconsin DEWS is in use statewide and currently provides predictions on the likelihood of graduation for over 225,000 students. DEWS represents a novel machine learning approach to the challenge of assessing the risk of non-graduation for students and provides highly accurate predictions for students in the middle grades without expanding beyond mandated administrative data collection.

Similar dropout early warning systems are in place in many jurisdictions across the country. Prior research has shown that, in many cases the indicators used by such systems do a poor job of balancing the tradeoff between correct classification of likely dropouts and false alarms (Bowers et al., 2013). Building on this work, DEWS uses the receiver-operating characteristic (ROC) metric to identify the best possible set of statistical models for making predictions about individual students.

This paper describes the DEWS approach and software, which leverages the open source statistical language R. DEWS is a flexible series of software modules that can adapt to new data, new algorithms, and new outcome variables to not only predict dropout, but also impute key predictors as well. This article describes the design and implementation of each of these modules in detail, as well as describing the open source R package, `EWStools`, that serves as the core of DEWS (Knowles, 2014).

1. Introduction

The Wisconsin Department of Public Instruction (DPI) is committed to providing school districts with the information and resources necessary to carry out the statewide goal of Every Child a Graduate College and Career Ready (Evers, 2012). Through a combination of federal and state funds, the DPI has invested millions of dollars in developing a statewide longitudinal data system, the Wisconsin Information System for Education (WISE). This data system provides longitudinal data on all students in the Wisconsin public school system across a variety of

*Jared E. Knowles is a research analyst at the Wisconsin Department of Public Instruction and a PhD candidate in the University of Wisconsin - Madison Political Science Department. The views expressed in the article are his own.

measures. Until now, Wisconsin collected this data to create public reports and for school and district accountability. Now , the DPI is providing educators and administrators with analyses of their data that can help inform their work and promote best practices.

The Wisconsin Dropout Early Warning System (DEWS) was created in the spring of 2012 to provide educators a forward-looking view of student performance. DEWS assesses the individual risk of failure to graduate on time for all students in grades six through nine in Wisconsin public schools. DEWS provides an individual probability of on-time graduation for each student, as well as an overall risk rating of "low," "moderate," or "high." Additionally, students receive a risk rating in each of four sub-domains: academics, attendance, behavior, and mobility.

Entering its second full year of statewide implementation, DEWS provides predictions on more than 225,000 Wisconsin students in over 1,000 schools across the state at two time points each school year. DEWS is able to identify nearly 65% of late and non-graduates before they enter high school with a low rate of false alarm. This accuracy is on par with some of the most well-regarded systems currently in use but is achieved at a larger scale, across a more diverse set of school environments, in earlier grades, and in the context of an educational system with relatively high graduation rates. As such, the Wisconsin approach represents a step forward in the application of predictive indicators to school improvement efforts.

In this paper, I first review the academic literature surrounding Early Warning Indicators (EWIs) in general and dropout early warning systems in particular. Next, I provide details on the specifics of education in Wisconsin and the data available to DEWS. I describe the DEWS workflow from data acquisition to data transformation, model training, prediction, and finally reporting. I then show how the models used in DEWS to provide predictions are competitive with many of the most accurate models found in the research literature today. I also evaluate how the DEWS models perform compared to an alternative approach using logistic regression. Finally, I conclude with some details about how DEWS is used by practitioners and what future steps need to be taken to continue to improve the system and the accuracy of its predictions. Throughout the paper, particular attention is paid tothe practical details of implementation in the hope that other analysts will be able to use the Wisconsin approach as a starting point for developing their own system.[1]

2. Literature Review

For quite some time, EWIs for graduation have been used to provide educators with a warning that individual students need help. As a result of EWI research, a diverse set of student attributes have been associated with an elevated risk of failing to complete high school. Some examples include failure in key courses, low attendance, suspensions, and drug use (Bowers et al., 2013; Balfanz and Iver, 2007; Easton and Allensworth, 2005). Unfortunately, there is no agreed upon standard to reliably assess the accuracy, utility, or tradeoffs associated with using a particular EWI in making decisions about interventions with individual students (Bowers et al., 2013). When constructing DEWS, a review of the literature quickly showed that Wisconsin did not possess the right data to implement many of the most common EWSs and that such systems would not be able to provide a prediction for a student early enough in the student's

[1]Not discussed here are the systems and communications challenges associated with informing stakeholders about the rollout of the system and increasing their usage. For more on these challenges, see the Wisconsin DEWS website, which contains guides for practitioners to use when reviewing EWS data and a how-to for getting access to the DEWS reports (Knowles and White, 2013) (`http://www.dpi.wi.gov/dews`)

academic career to be useful on a statewide basis. In this section, I review the most common EWI approaches and weigh their advantages and disadvantages in a statewide implementation.

2.1. CHECKLIST

For many years, large urban school districts have collected a number of data elements longitudinally to provide teachers and administrators with an advanced sense of a student's progress, often toward graduation (Easton and Allensworth, 2005; Gleason and Dynarski, 2002). In many cases, using a basic set of benchmarks for student progress along these indicators, schools are able to classify students at-risk of not completing high school in middle grades and early high school in order to provide them early intervention services.

The Chicago On-Track System developed by the Consortium on Chicago School Research (CCSR) has since been used across the nation to help high schools identify 9^{th} grade students who are struggling and at risk of exiting early (Heppen and Therriault, 2008; Kennelly and Monrad, 2007). The first year of high school is a critical time for students, so this focus on 9^{th} graders makes practical sense. Such systems are highly accurate, with the Chicago model found to identify upwards of 70% of non-graduates (Easton and Allensworth, 2005; Easton and Allensworth, 2007; Roderick and Camburn, 1999; Allensworth, 2013). The Chicago system is remarkable not only for its high accuracy, but also for its simplicity in implementation; any administrator can implement the system with just a few data elements and a spreadsheet of student names (Heppen and Therriault, 2008). It also provides administrators with a clear picture of what strategies to investigate to address the student's concerns (Roderick and Camburn, 1999).

However, research has shown that dropping out of school is not a single event, but a process of disengagement with school starting in earlier grades (Roderick, 1993; Rumberger, 1995; Balfanz and Iver, 2007). Additionally, many students may not transition to high school and instead choose to exit school after the 8^{th} grade. Even students who persist into high school drop out at very different points in their high school career. In Wisconsin, roughly 3,000 students a year dropout before reaching 12^{th} grade, and roughly 1,500 of these students appear to dropout in 9^{th} and 10^{th} grade. Identifying these students in the middle of the second semester of their 9^{th} grade year leaves little time for interventions. Any interventions attempted at this point also are likely to be more expensive in terms of time, energy, and money.

Fortunately, on-track indicators have been identified for middle grades as well. Balfanz (2009) and Balfanz and Herzog (2006) have shown that course failures, low attendance, or poor marks for behavior in middle school can substantially reduce students' likelihood graduating or catching up to their peers on state assessments. These effects are more pronounced in certain contexts – particularly in high poverty schools where the indicators seem to matter more (Balfanz and Iver, 2007). Much like the CCSR on-track system, this system for middle grades finds that simple checkpoints in student outcome data can be highly predictive of future outcomes, and are easy to monitor (Balfanz and Herzog, 2006). Absent intervention, students in Philadelphia showing these warning signs graduated at a rate of 10 to 20% (Balfanz and Iver, 2007).

The downside of such checklist based systems is that they tend to overemphasize individual attributes like grades or attendance at the risk of oversimplifying the mechanisms underlying students' risk of dropping out (Gleason and Dynarski, 2002). Often, educators are not given a strong sense of how well such indicator systems perform and, thus, may not find it easy to weigh the results of such a checklist against other evidence available to them through classroom observation, student interactions, or a short interview with the student.

2.2. Regression Techniques

While both the Balfanz and CCSR models provide high predictive power and ease of use for administrators, they also provoke questions of how to improve upon their accuracy. Recent work in Milwaukee Public Schools (MPS) has shown that students often continue to accumulate credits without course failures, but still dropout, or graduate with a very minimal skill set (Carl et al., 2013). Carl et al. (2013) use regression models to estimate a student's probability of high school completion conditional on the number and type of credits earned in the freshman year. The authors' conclude in creating a new measure, Total Quality Credits (TQC), whic is a combination of course grades in the four core subjects and is highly predictive of both on-time high school completion and college enrollment. Additionally, TQC in the freshman year can be predicted using a statistical model with middle grade data, suggesting that such a system could be extended into earlier grades (Carl et al., 2013).

Both the regression and the checklist approach represent innovative ways to predict student outcomes for the purpose of planning dropout interventions. However, they are also products of individual large urban school districts – a specific educational context. Students in schools outside of this context may have different thresholds of risk based on attendance, behavior, or course failures (the ABCs) or different risk factors altogether. Indeed, research has shown that the context, such as the school poverty level or the school building itself, can matter greatly (Balfanz, 2009; Balfanz and Iver, 2006). Thus, in a statewide implementation of an early warning system, it is important to identify an approach that is flexible across contexts.

2.3. Latent Class and Growth Mixture Modeling

The regression techniques adopted by Carl et al. (2013) provide flexibility for applying separate thresholds in separate contexts. However, these techniques are limited in their accuracy by one strong assumption – that all dropouts result from the same set of risk factors. Using a latent class analysis of a population of high school dropouts Bowers and Sprott (2012b) identify a distinct typology of high school non-completers with risk factors varying across the types of dropouts. Thus, any system that assumes a constant relationship between indicators and outcomes will fail to identify whole subsets of the non-completer population.

Growth mixture modeling (GMM) is a statistical technique that can adjust for variability in the risk factors for individual students and that has proven to be highly accurate in identifying non-completers. This method was first applied to high school completion data by Muthèn (2004). Bowers and Sprott (2012a) builds on this work by extending the GMM methodology to a nationally representative data set (the Education Longitudinal Study of 2002) and tying the model to existing theories of student disengagement and dropout. Bowers and Sprott (2012a) found that even when limiting the GMM to data available to most school staff, the model provides highly accurate predictions of individual students' high school completion.The GMM methodology has also demonstrated predictive power for student engagement indicators, as measured by a student survey instrument, which can be leading indicators of student disengagement and non-completion (Janosz et al., 2008).

Though accurate, both latent class and GMM approaches are not well-suited for deployment in an operational EWS – a role they were not designed to serve. First, while GMMs are good at modeling behavior in a sample of students, there is no sense of how such models perform when making predictions for students outside of the sample. Out-of-sample performance is critical for an accurate EWS deployment and is one of the reasons for the enduring success of the

Chicago model. Second, GMMs and latent class approaches, in contrast to checklist systems, have intensive data requirements, including requiring multiple contiguous years of data for each student. As Bowers and Sprott (2012a) note, this is problematic within a single school district where students may transfer in and out. This narrows the pool of students eligible to receive a prediction, specifically excluding students with an incomplete educational record, who are already at an elevated risk of dropping out, from being identified within the system.

Despite not being deployed in an operational early warning system, the literature reviewed above is critical in informing the design of any EWS that seeks to identify large percentages of likely non-completers. Bowers and Sprott (2012b) and Muthèn (2004) underscore the fact that students drop out for reasons other than low performance in coursework or on standardized assessments and that, for some students, high performance may also be a warning sign. Failing to account for this diversity of factors in some way ensures that an EWS will only provide accurate predictions of dropout of a certain type.

2.4. Assessing Early Warning Performance

How do we know if an EWS is any good? The ideal early warning system would identify all future dropouts and would raise no false alarms – every student identified as at-risk would fail to complete high school without intervention. The system would do so whether or not the population it was predicting was composed of 5% or 50% future dropouts. Deciding on a metric of accuracy for comparing EWSs means selecting a metric that reflects this ideal.

An example of how the choice of metric can shape how we talk about and compare systems comes from an EWS in Montgomery County, Maryland, that identifies over 75% of high school dropouts in first grade – a remarkable achievement. However, to achieve this detection rate, the system identifies nearly half of all students in the first grade as being at risk of dropping out.[2] This system has been described as both 75% accurate and 50% accurate – which is it?

The question of accuracy is important not just for comparing accuracy across systems, something of an academic exercise, but also for identifying true improvement in accuracy within a system. This is critical in the case of DEWS, which seeks to build many models and identify the best. Bowers et al. (2013) suggest that, instead of simply reporting the percentage of dropouts identified by a flag, scholars should report a set of accuracy indicators derived from the signal detection literature and used in medical applications [see (Hanley and McNeil, 1982; Swets, 1988; Vivo and Franco, 2008; Zwieg and Campbell, 1993)]. Table 1 depicts these metrics in what is known as a confusion matrix, a summary of classification accuracy for binary outcomes [adapted from Bowers and Sprott (2012a)].

Table 2 describes the most important measures of accuracy and provides an example of their values using the numbers in Table 1. For example, a hypothetical early warning system that correctly identified 50 students as non-graduates may be described simultaneously as having 66% accuracy, 33% accuracy, 90% accuracy, or a 1% false alarm rate. All of these measures capture a facet of the accuracy of the system, but do so incompletely. Depending on the local implementation, only one metric may be of value to decision-makers. However, when comparing approaches in the literature, it is important to report all metrics so that others in the field can understand the tradeoffs inherent to the system. The hypothetical system in Table 2 finds students who are very likely to dropout; thus, identified students most certainly merit intervention. Unfortunately, this system fails to identify two-thirds of the dropouts. Although this may

[2] http://educationbythenumbers.org/content/can-an-algorithm-id-high-school-drop-outs-in-first-grade_386/

Table 1: Classification Metrics for Binary Classifiers

		Event		
		non-grad	**graduate**	**total**
predicted value	**non-grad**	True Positive n=50	False Negative n=25	a+b
	graduate	False Positive n=100	True Negative n=250	c+d
		a+c	b+d	total

Table 2: Measures of Accuracy for At-Risk Indicators

Measure	Calculation	Interpretation	Example
Precision (positive predictive value)	a/(a+b)	Percent of predicted dropouts that dropout	66%
Sensitivity (recall)	a/(a+c)	Percent of actual dropouts predicted correctly	33%
Specificity	d/(b+d)	Percent of actual graduates predicted correctly	90%
false alarm	b/(b+d)	1 - Specificity (False Alarm Rate)	10%

be desirable if the goal of the system is to identify the students most likely to dropout in order to make the best use of limited resources. Conversely, the system correctly labels 90% of students as graduates, but incorrectly labels 10% of graduates as non-graduates – perhaps leading to an inefficient allocation of limited intervention resources.

This is exactly the fundamental tradeoff that Gleason and Dynarski (2002) identify as the challenge facing early warning indicators – where do the indicators draw the line between false alarm and true classification of students, and is the resulting student group the group that schools should serve? In most cases from the literature, the answer is that those using the indicators simply do not know. There has been too much inconsistency in the application of accuracy metrics to make comparisons useful and to identify where such indicators draw the line between correct classification of non-graduates and false alarm rates (Gleason and Dynarski, 2002; Jerald, 2006; Bowers et al., 2013). As Bowers et al. (2013) indicate, most of the 110 at-risk flags found in the literature only include a measure of the sensitivity, or the specificity, but rarely both. The metric chosen for determining the best model is thus a pivotal decision, both for technical and practical reasons. In an effort to bring cohesion and clarity to the comparison of EWIs, Bowers et al. (2013) calculated the performance metrics for 110 separate EWIs found in the literature. Figure 1 recreates a plot from Bowers et al. (2013) depicting the sensitivity and the specificity for these early warning indicators.

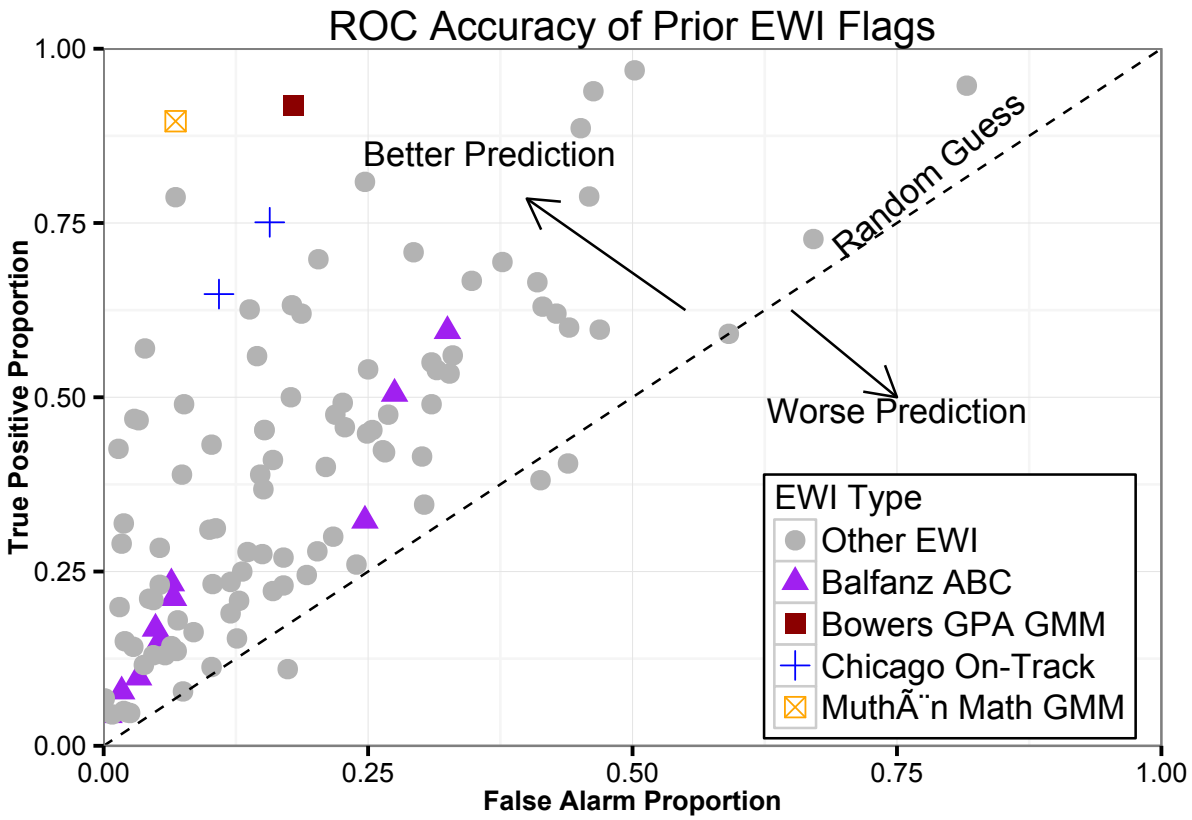

Figure 1: ROC points for 110 published early warning indicators from Bowers et al. (2013). The diagonal line represents random chance; the top left corner represents perfect prediction.

This plot shows the explicit tradeoff between false alarms and true positive classifications, which is known as the Receiver Operating Characteristic or ROC. In other words, what percentage of dropouts can we successfully identify (along the y-axis) in exchange for falsely identifying eventual graduates as dropouts (along the x-axis)? The ROC also permits comparisons of binary classification models regardless of the underlying rate of the event we are classifying, allowing comparisons to be made among indicators developed on samples with very low and very high dropout rates (Hastie et al., 2009). For example, the growth mixture models discussed above (and represented as boxes in Figure 1) are highly accurate with a high proportion of true positives and a relatively low proportion of false alarms. The Chicago On-Track indicator (represented as crosses) is also highly accurate. Dropping into the middle grades, the Balfanz ABC indicators are somewhat less informative but still capture a relatively large proportion of eventual dropouts. Thus, the ROC efficiently describes the performance of binary classifiers as the tradeoff between false alarm and true positive detection. Kuhn et al. (2013) demonstrate how using the ROC on classification models can be intuitive and easy to do. For policy making, the ROC is valuable because it depicts the tradeoff between the rate of classification of true positives and false-positives as was depicted in Table 1. Following the guidance of Bowers et al. (2013) DEWS also employs the ROC.

Accuracy is a useful way to understand an early warning system's behavior – which is essential in making decisions about how to use of predictions in the field. While accuracy is critical, accuracy is affected by many things beyond the control of any given model, and the accuracy of models will vary from context to context. This makes it important to understand the nature of the dropout problem in the context in which the early warning system is deployed in order to set realistic expectations about model accuracy and the type of accuracy such a system will seek to optimize. I now review the Wisconsin case.

3. DATA

As we have seen, EWS systems vary dramatically in their predictive performance (Bowers et al., 2013). Performance depends on the strength of the correlation between observable characteristics of students and their eventual graduation. In school systems where students with low academic performance and behavioral indicators dropout at a high level, predictive models based on the ABC indicators will be highly accurate. In situations where many of the students with these low indicators continue on to graduate, the ABC indicators will struggle.

Because prior research has predominantly focused on large urban areas like Baltimore (Balfanz, 2009), Chicago (Easton and Allensworth, 2005; Easton and Allensworth, 2007), and Milwaukee (Carl et al., 2013), before beginning to build a model on dropouts in Wisconsin, it is important to look at the relationship between observable indicators and dropout rates. The statewide view is different from models built in large urban districts because of the substantial contextual variance that exists across schools in Wisconsin. Wisconsin has middle schools serving every possible grade range; sizes ranging from a few dozen students to several hundred; and locations as diverse as urban Milwaukee, rural Butternut, and remote Washington Island.

The most important implication of this variation is that late-graduates are not evenly distributed across the state or within individual school districts. Nationally, this late-graduate clustering has been highly publicized by focusing on high schools that are so-called "dropout factories" (Balfanz and Legters, 2004). It is clear that middle schools also exhibit the same phenomenon as high schools, with a small percentage of schools accounting for a large percentage

of the late- and non-graduates statewide. Figure 2 demonstrates that a disproportionate amount of late-graduates attend just a handful of schools in the state. Forty percent of all students who graduated late or dropped out came from only 48 middle schools (7.5%). Clearly any analysis of the likelihood of graduating late must take into account the clustered structure of the data in order to accurately model the data-generation process.

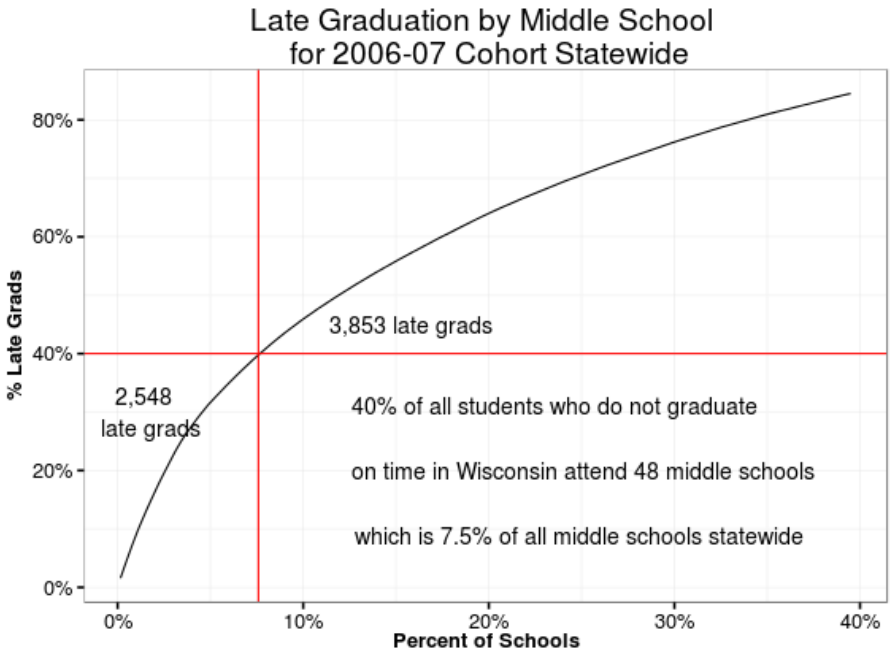

Figure 2: Late and Non-Graduation by Middle School

For illustration this paper uses the 2006-07 grade 7 cohort.[3] The Wisconsin DEWS draws on all cohorts of students who were in grades 5-8 after 2005-06 and who were expected to graduate in or before 2012-13. Details on the specific data elements used in the DEWS are available in Section 7. of the appendix. Overall, Wisconsin has one of the highest graduation rates in the country; the state's most recent four-year high school completion rate stood at 87%.[4] This puts Wisconsin well below the average dropout rate for samples used in other studies of EWI indicators, which was roughly 22% (Bowers et al., 2013).

Because of Wisconsin's high graduation rate, the Wisconsin DEWS focuses on identifying students who take more than four years of high school to graduate or do not do so at all. This serves the substantive issue the system is trying to address: providing educators with an accurate assessment of the need for intervention for particular students. Students who fall into a five- or six-year graduation track during high school are more likely to dropout, less likely to receive a regular diploma, and less likely to enroll in college. In other words, they need additional

[3]The 2006-07 cohort was chosen because it is the first cohort in the DPI data on which predictive models can be built and tested on a follow-up cohort. As the data represent all students statewide, it is, thus, a census of students for that year.

[4]http://data.dpi.state.wi.us/data/HSCompletionPage.aspx?GraphFile=HIGHSCHOOLCOMPLETION&SCounty=47&SAthleticConf=45&SCESA=05&OrgLevel=st&Qquad=performance.aspx&TmFrm=4

help. Focusing on on-time high school graduation, rather than only confirmed dropouts, has several advantages. First, student dropout has a number of different meanings, and students can dropout multiple times(sometimes still graduating on time). In contrast, four-year graduation is an unambiguous measure of student attainment [see Bowers and Sprott (2012a)]. Second, students who graduate late often dropout, and students who dropout often graduate late (when they graduate). Thus these two populations overlap. Third, from the perspective of a school system, late graduates are more expensive and require more intensive educational interventions than do dropouts.[5] Finally, from a statistical perspective, the number of late and non-graduates is much larger than the number of dropouts, making the identification of late- and non-graduates in middle grades more tractable. Only around 3,000 members of any given cohort of $\approx$ 60,000 students are confirmed dropouts, well below the rate common in the EWI literature and making identification a substantial challenge. In contrast, late and non-graduates make up around 9,000 members of any given cohort, which results in a much more manageable 15% of students to detect with statistical models.

In addition to the clustering evidenced in Figure 2, deep disparities in graduation rates exist along economic and racial lines. Table 3 shows racial disparities in the 2010-11 four-year graduation rates in Wisconsin and, Table 4 shows the graduation gap for students who are eligible for free and reduced price lunch (FRL) and those who are not.[6]

Table 3: Racial Graduation Disparities in Wisconsin for 2010-11

Group	Expected	Grads	Rate	Gap
White	54,468	49,783	91.4%	-
American Indian	1,027	737	71.7%	19.7
Asian	2,517	2,225	88.4%	3
Black	6,889	4,395	63.8%	27.6
Hispanic	4,751	3,420	72.0%	19.4

*All rates are 4 year graduation rates

Table 4: Economic Graduation Disparities in Wisconsin for 2010-11

Group	Expected	Grads	Rate	Gap
Non FRL	50,834	46,715	91.9%	-
FRL	19,542	14,481	74.1%	17.8

*All rates are 4 year graduation rates

These disparities are quite stark. In order to close these attainment gaps, DPI is focusing

[5]Note that students with diabilities may have an Individual Educational Plan (IEP) that explicitly include a five- or six-year graduation as the desired outcome. Schools are instructed to explicitly review IEPs of their students with disabilities when reviewing DEWS reports to avoid raising a false alarm about students with planned graduations later than four years.

[6]Categories for two or more races and Native Hawaiian/Pacific Islander are not shown – combined these categories account for less than 1,000 students in the cohort for this school year. All data from the Wisconsin Information System for Education dashboard (WISEdash) online reporting system: `http://www.wisedash.dpi.wi.gov/`

on strategies aimed at addressing the needs of disadvantaged groups of students (Evers, 2012). Providing educators the tools to understand early who is at risk of not completing high school is intended to allow staff more time to assess the individual needs of the student and to try different strategies to address the root cause while the student is still in the middle grades.

3.1. CHOOSING PREDICTORS

Now that the nature of on-time graduation in Wisconsin has been explored, we turn our attention to the data available to predict this outcome. Instead of starting with a theory about which mechanisms translate into high school completion, I start by describing the data that has been used in the prior literature and is available statewide. As DEWS is a production machine learning system, its goal is not to test any given theory of the causes of high school completion, but to maximize the accuracy of predictions for individual students using available data.

I begin by leveraging data available in the Wisconsin statewide longitudinal data system (SLDS), known as the Wisconsin Information System for Education (WISE). Since the 2005-06 school year, data has been collected statewide on all students in a limited number of domains. This data system covers roughly 870,000 students enrolled in public K-12 schools, including charter schools, in the state each year and makes it possible to longitudinally analyze records for students who stay enrolled in public K-12 schools.[7] I start by looking at how many data elements were available from the ABC (Attendance, Behavior, and Coursework) indicators described by Balfanz (2009).

Table 5: Data available in the DPI SLDS for EWS Predictions

Domain	**Elements**	**Years of Collection**
Attendance	% attendance; days missed	05-06 to present
Behavior	days suspended; days expelled	06-07 to present
Mobility	schools and districts attended	06-07 to present
WKCE Reading	scale score	05-06 to present
WKCE Math	scale score	05-06 to present
Demographics	race; gender; ELL; SwD; FRL	05-06 to present
Cohort	cohort aggregates of above	05-06 to present

Table 5 describes the data available in DPI's SLDS, which has much in common with the data used in the early warning systems discussed in the literature review, however, a few substantial differences exist. First, course completion data is not yet available statewide, and thus, the pivotal indicators of course failures and course-taking are outside of the scope of the current system. Another downside is that most of the data is available only as an annual summary, an issue discussed in detail below. Statewide coverage does, however, provide some information not available in other systems – namely, student mobility indicators. Highly mobile students are much more likely to graduate late than are their peers, and, as long as students remain within Wisconsin public K-12 schools, they are included in this data along with an indicator of their level of mobility.

[7] Such data systems are common in state education agencies in the U.S. and are used to fulfill federal reporting requirements.

In addition, we know from the clustered nature of the dropout problem in Wisconsin that some school and community level attributes may also be necessary to include. Unlike previous EWS systems that either used a nationally representative sample or operated in only a single school district, DEWS attempts to accurately estimate dropout risk across the diverse school environments within the state of Wisconsin. Thus, aggregating student performance, demographic, and school system information – and including these variables in the data set – is important step given that there is strong prior information that such community-level variation is an important aspect of the phenomenon we are modeling.

In the data mining literature, variable selection, sometimes called feature selection, is the process of moving from a set of N possible predictor variables to the best possible set K. In the case of a predictive modeling system like DEWS, variable selection is among the most consequential choices that analysts make. Feature selection can be quite difficult in cases with many uninformative predictors or many subsets of highly collinear predictors. In the end, the process of feature selection must combine substantive knowledge of the issue domain and data collection process with the empirical properties of these predictors. Some predictive models have methods (e.g. stepwise regression and feature elmination) to aid analysts in selecting variables with predictive power (Efroymson, 1960).[8] The final list of predictors and their impact on the student risk score are discussed in Section 5..

3.2. TIMING

One of the biggest concerns when providing early warning predictors to educators is the timeline of when results become available. This concern led to a number of a crucial design decisions in the Wisconsin implementation and will likely influence any early warning system.

In Wisconsin there is a considerable lag (i.e. anywhere from 6 to 18 months) between when the school year ends and when information on student outcomes becomes available within the data warehouse. To provide predictions using this "stale" data, it is necessary to focus on developing predictions that are based on less volatile measures, leaving the updating of student risk predictions to local educators with richer and timelier data at their fingertips. Therefore, data from the prior grade level is used to provide predictions (e.g. predictions for a student in grade six are based on data and models using fifth grade data). In order to be even more timely, the DPI releases a preliminary DEWS score at the start of the school year based on data from two years prior. This means that fall preliminary estimates of student risk for students in grade six, imputed estimates of the fifth grade data are used which are based on imputation models developed using grade fourd data for all unavailable predictors including attendance, mobility, and student discipline. [9]

The other way that timing affects DEWS is that predictors that are available for current students, but not for previous graduates, are not able to be included. This creates a substantial lag between the introduction of new data elements and their ability to be incorporated into the DEWS predictions. For example, if student courses and grades became available in 2009-10, we would need to wait until the students in grade eight in 2009-10 have graduated before course and grade information could be incorporated into DEWS. This presents a substantial drawback. One

[8]Feature selection is criticized as a violation of the assumptions necessary to interpret confidence intervals and p-values under frequentist inference (Chatfield, 1995; Chapelle et al., 2002; Vapnik, 1998). Since we are working with population data and not a sample, and we are explicitly testing model fit on test data to validate future predictive validity, these criticisms are not as concerning.

[9]The imputation techniques used in DEWS will be the focus of future work.

potential workaround that is being considered is developing intermediate models that predict the DEWS score itself – thus allowing new data elements to be incorporated into the system much more quickly at the cost of some precision. However, introduction of new data and new steps in the process must be carefully considered if the system is to be maintained moving forward. Additional steps can introduce more costs due to complexity than benefits of the additional accuracy (Sculley et al., 2014).

A last more technical difficulty arises if school-specific parameters are estimated. Schools can and do open and close between the time the training cohort is in place and current students in school today. This makes using fixed effect models or other paramaterizations of specific schools difficult. Thus, instead of individual school parameters, DEWS parameterizes the features of schools and school cohorts, to enable a smoother prediction process for current students.

4. METHODS

Clearly, some design decisions in the Wisconsin system have been driven by the data available within the state and the challenge of providing valid predictions on a statewide scale. However, the approach taken by DEWS is flexible to accommodate any number of alternatives to the Wisconsin implementation. This section describes the Wisconsin approach in detail and seeks to highlight areas where analysts may consider deviating from those decisions made in DEWS.

Overall, the approach taken in DEWS is best described as a system of applied statistical learning. In the literature, statistical learning techniques have previously been identified to hold promise for increasing the accuracy of prediction in the early warning context (Bowers et al., 2013). More practically, they appear well suited to the unique case faced by state and local education agencies developing EWS systems. The results-focused nature of statistical learning allows analysts to rapidly test statistical models and illustrate the classification power of these models to others. However, such techniques increase the complexity of the model building process and reduce the transparency of the results to stakeholders (Hastie et al., 2009; James et al., 2013). This makes decisions about every aspect of the modeling process crucial in balancing the tradeoff between accuracy and complexity.

This section focuses on DEWS as an entire system and not on the individual models it results in. This is intended to provide guidance to others about how to confront the key decisions along the way and not to highlight the specifics of the Wisconsin implementation. I believe the value of DEWS is not the specific predictive models that emerge, but rather the systematic process of searching for and identifying the best possible models given available data – an approach that can benefit analysts working in states and school districts across the country.

4.1. APPLIED STATISTICAL LEARNING

Breiman (2001b) describes two cultures of statistical modeling – the data and the algorithmic modeling cultures. The latter goes by different names – statistical learning, machine learning, predictive modeling, data mining, or applied modeling – but is defined by a goal of learning relationships between predictors and outputs and using those modeled relationships to predict outputs for new data sets (Hastie et al., 2009). This learning may occur over time in cases where new data is streaming into the system, across samples when some subset of the population is unobserved, or both. Applied models are focused on generating the most accurate prediction of the output on data sets other than the data set used to *train*, or fit, the model.

Building an applied statistical model, such as an EWS, differs from building a statistical model for the purposes of inference or theory testing in three important ways.

First, applied models are defined by their ability to make accurate out-of-sample predictions on future data. In the EWS context Gleason and Dynarski (2002) note that many EWIs fail by providing too many false-positives and too few true-negative classifications, leading to staff inappropriately allocating resources to intervene with students who are not in need. In contrast, inferential models are focused on testing the impact of particular measured variables within a specific sample, then generalizing to the appropriate population; this approach can result in overfitting to the sample data, causing unreliable out-of-sample prediction or an inability to estimate the reliability of predictions out-of-sample at all (Hastie et al., 2009; James et al., 2013).

Second, applied models are constrained not by a theory of the data generation process but by the goals of the application (Breiman, 2001b). A model may be designed to include or exclude certain factors in order to provide face validity to the users of the model.[10]

Third, applied models that will be used in an ongoing manner must be flexible to new inputs and built much like a software application with a focus on stability, reproducibility, and modularity (Sculley et al., 2014). While Bowers et al. (2013) have moved the field forward by suggesting a consistent set of accuracy metrics to measure the performance of an EWS, there are significant limitations with only including measures of accuracy on the data used to fit the model. In machine learning data is commonly divided into *training* and *test* data sets. The *training* data set is used to build the model, while the model performance isevaluated by how well the model predicts the *test* data, which is held out of the model building process (Hastie et al., 2009; James et al., 2013). It is no surprise that, when data is limited or expensive to collect, holding out data is not a commonly used practice. However, with the advent of large administrative record systems, it is no longer cost prohibitive to divide the data into both *training* and *test* data sets – indeed, computationally it may be a necessity.

Applied modeling then is concerned less with model fit to the current data than it is with correct prediction on new data. In fact, many applied modeling techniques are employed to avoid *overfitting* the model to the current data, which can lead to greater prediction error with new data (James et al., 2013). This focus on reducing predictive error for future cases makes applied modeling particularly well suited for development of a dropout early warning system. The primary goal of DEWS is to correctly classify new students each year as accurately as possible – not to understand the fundamental relationship between available predictors and student attainment.

James et al. (2013) note that the process of finding a predictive model proceeds along two dimensions. For abinomial process (such as the case here of classifying students as on-time graduates) the focus is on identifying both variables X and the function f such that:

$$Y = \begin{cases} 0, & \text{if } f(X) > 0 \\ 1, & \text{if } f(X) < 0 \end{cases}$$

Searching across algorithms is equivalent to identifying the best $f(X)$, such that:

$$Y^* = f(X) + \epsilon$$

[10]This also occurs in inferential models in an academic setting when researchers report using "the standard control variables" in their models.

This is in contrast to an inferential framework where the analyst selects the subset of X using prior literature, available data, and theory to build a model of the data generating process and then fit a statistical estimate to that model. In a case with administrative data that is both broad and rich, an analyst can instead focus on identifying the subset of X that best predicts y given the function $f(X)$. Expanding the model search across a wide array of algorithms could yield improvements in accuracy and an ability to capture more varying types of dropouts than possible using conventional binomial regression models.

In addition to the concern for accuracy, there are other considerations. For example, a model must also be accepted as valid by those who will make use of it, must apply equally well to new cases, and must be able to deliver predictions on the timeline necessary for the application. Balancing these concerns is critical to successfully transitioning a system from the academic to the applied setting, and is one of the reasons that the Chicago On-Track System has proven so successful. Despite its lower reported accuracy than more complex metrics, the Chicago System has proven robust, transparent, and easy to communicate (Kennelly and Monrad, 2007).

The Wisconsin DEWS represents just one possible resolution of the tensions between accuracy, reproducibility, transparency, and computability – but it is an implementation of an early warning indicator that produces predictions for all students in public schools in Wisconsin in grades six through nine. In the sections that follow, I explain the design of DEWS starting at a high level and moving through each component of the system from fetching the data to reporting predictions. Analysts seeking to implement an EWS of their own may depart from the methods described here at any point, and I identify the tradeoffs associated with such decisions wherever possible.

4.2. The Wisconsin Procedure

DEWS can be described as consisting of four major components defined by the verbs "get," "transform," "train," and "score." Each of these verbs represents a defined subroutine. Figure 4.2. shows a high-level view of the workflow and each of the subroutines.[11]

This design accommodates two distinct workflows – *training* accurate predictive models and *scoring* current students using those models. Both of these workflows are represented in Figure 4.2. and by design only depart at the first decision point, marked as a diamond. The workflow then adapts depending on whether the user is seeking to train, score, or do both. In the Wisconsin case, this procedure is conducted using a single year of data, and the process is repeated independently for each grade level. If imputation is necessary, imputation models, discussed below, are also trained in the same method.

The system starts with accessing the data and bringing it into the analytical environment to build models and produce predictions. This process deviates slightly depending on whether the user is training models or scoring new cases. Next, a common set of transformations is performed on the data to optimize it for analysis. Then, either models are trained or the data is scored using previously trained models. The results are then stored and the procedure is repeated for other variables of interest or other grade levels.

As much as possible, DEWS attempts to make the four major subroutines identical wherever they appear in order to promote modularity, reduce code redundancy, and allow the system to

[11] Elements of the transform, train, and score subroutines are freely available as an open source extension for the R statistical computing language. The `EWStools` package provides many of the key components of these subroutines in order to make it easier for others to implement a DEWS-like system (Knowles, 2014).

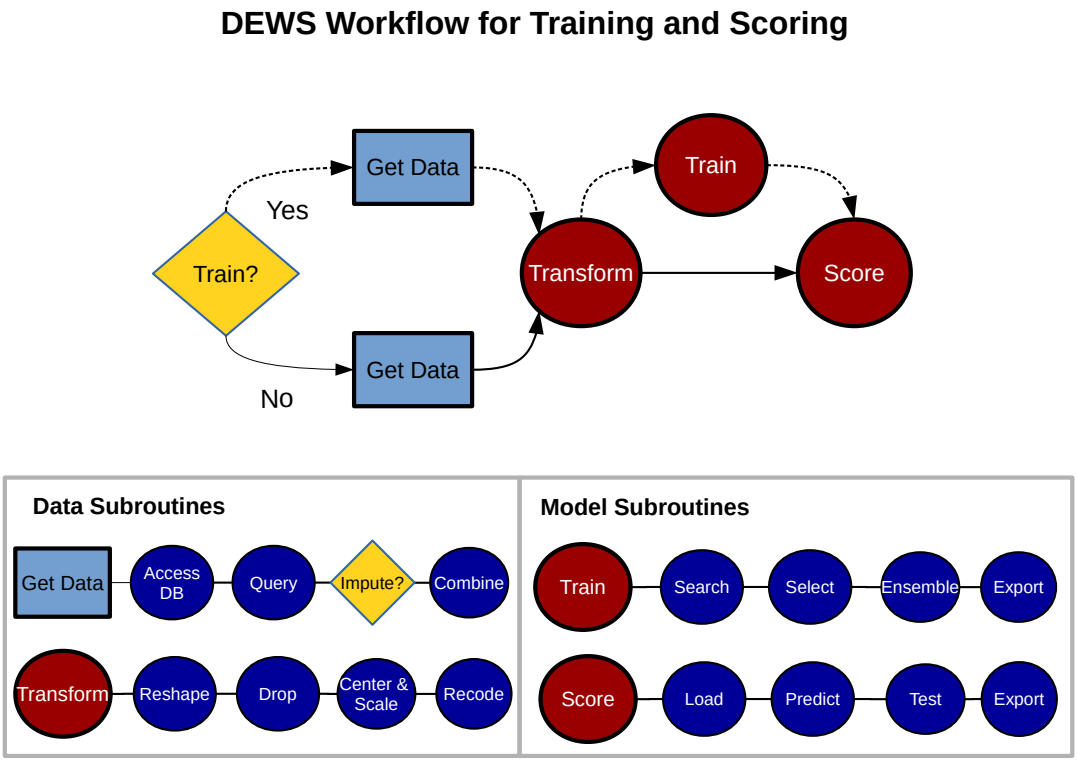

Figure 3: A Workflow of the Wisconsin DEWS Procedure

be flexible to scale to new data each school year and new predictors as new data systems come online. Additionally, this allows new models to be continuously tested to identify opportunities to improve accuracy. I now describe each of the subroutines of the DEWS workflow in detail, describing both the theoretical motivations for the design decisions and the practical implementation of those decisions. The end result is a system which provides an individual prediction, or score, for almost all students that represents the probability of on-time graduation.

4.3. Get Data

In the Wisconsin system, data comes directly from a data warehouse, which stores longitudinal data on current and past students. A data warehouse is preferable to a transactional data store, such as a student information system, because it greatly reduces the transformations needed to move data from the database into a suitable form for analysis (Inmon, 2005; Kimball and Ross, 2002). The data warehouse allows different systems to focus on their strengths – the data warehouse handles the relational mapping between data elements and the resolving of record conflicts, while the EWS is left to train models and produce accurate predictions that can then be loaded back into the data warehouse. This reduces the need to build extensive data cleaning routines to resolve conflicting records. The key issue is that the processes for retrieving the data for both the training and scoring steps are as close to identical as possible. Data sources

should be kept the same in order to ensure that the variables and their scales and meaning are consistent between the two workflows and predictions do not become biased.

A unique feature of DEWS is that it allows for imputation to occur here. As discussed previously, imputing values is necessary in order to give schools as up-to-date as possible assessment of student risk at the start of the school year. For each variable that needs to be imputed, the same transform, train, and score workflow is in place to build the imputation models. This is the approach recommended by Kuhn and Johnson (2013) – building layered predictive models tuned to provide as much accuracy as possible in estimating the missing predictors. Traditional imputation methods are not available in the predictive framework because they generally require that the dependent variable be included in the imputation equations to avoid bias (Honaker et al., 2011). For predictive models, the dependent variable is only available when training models, but, when scoring new cases, the data also needs to be imputed and there is no dependent variable available. Once the data has been accessed, queries have been run, and any imputation necessary has occurred, the data must be combined and made available to the analytical tools that will be used to train and score the models.

4.4. TRANSFORM

After acquiring the data, next comes the important process of preparing or "tidying" the data in order to make it suitable for building a predictive model (Wickham, 2014). In many cases, the choices made about cleaning, transforming, and rescaling variables may prove to be more consequential to the ultimate accuracy of the system than the exact subset of X or the functional form chosen (Dasu and Johnson, 2003). Despite this, moving from administrative data to data suitable for analysis – as most EWSs will need to do – is an exercise that has received only limited attention in the literature (US Department of Education, 2012a; US Department of Education, 2012b). This section will focus on the classes of decisions that need to be made and the consequences those decisions can have.[12]

The data transformation must be consistent between the training and scoring data. Thus, transformations must be applied in both contexts in a consistent and reliable manner to avoid prediction errors resulting from a misapplication of a transformation. There are four major types of transformations that should be considered:

- Reshape
- Define and drop missing data
- Center and scale
- Recode

The treatment of these issues here serves to remind the reader that these decisions can have substantial consequences for the performance of the model and to highlight the ways in which such decisions are different in an applied modeling context.

4.4.1. Reshape

The shape of the data set is determined in part by the software selected to train and score models and in part by the type of analysis the analyst is considering. In most cases, it is prefer-

[12] Analysts seeking a practical treatment of these challenges and some specific recommendations for education data should review the Strategic Data Project Toolkit for Effective Data Use, `http://cepr.harvard.edu/sdp/resources/toolkit.php` (The Strategic Data Project, 2012)

able to have each row represent a student and each column represent a student attribute. Multiple years of a single attribute should be represented by multiple columns. Additionally, group-level predictors may be merged onto these student-level records. Group-level predictors may be drawn from a separate data source or calculated from the individual-level predictors. In the Wisconsin implementation, I calculate cohort effects – aggregates of predictors for students in the same school and grade. This approach was chosen for the ease of incorporation into the transformation workflow and its ability to provide more variability within schools across years as the attributes of cohort members in a school may change between the training and the scoring data. An alternative would be to use school level aggregates, or to use multiple cohorts of students from a single school to calculate school-grade aggregates. [13] The result must be a single data set of predictors that can be coerced into a matrix to allow for the maximum compatibility with fitting a diverse set of models.

4.4.2. Define and Drop

Handling missing data is a particular challenge in the machine learning context (Kuhn and Johnson, 2013). The procedure described here is called "Define and Drop" in recognition of the fact that before missing data can be removed, what data constitutes missing must first be defined. Data are often missing in administrative records, and the missingness is almost never at random (US Department of Education, 2012a; US Department of Education, 2012b; The Strategic Data Project, 2012). For example, missingness can be due to a change in reporting requirements, conflicting records for a single student, or an inability to identify the appropriate student record to associate with a data element. All of these elements are much more likely to occur for students who are changing educational systems, have inconsistent enrollment records, or are at risk of academic challenges in other ways.

Due to this, it is important to consider the treatment of missing data and the impact missing data has on the predictions generated by the model. In a longitudinal framework, the constraint that students must have multiple prior years of complete (with respect to EWS elements) data and an observed graduation outcome creates a distinct subset of students – students who maintained consistent enrollment for N years. Defining and treating missing data is treated very inconsistently in the existing literature with scholars often using national data sets that have been imputed or not documenting how cases are dropped or dropouts are defined (Bowers et al., 2013). The subset of students with a complete panel of four, three, or even just two consecutive years of data will be a group of students with an elevated chance of graduation just by virtue of the fact that their educational environment was stable enough that they had consistent records in two or more consecutive years.

The result is that DEWS uses only a snapshot of a single year of records for students. This is to ensure that DEWS has the greatest ability to produce estimates for a group of students with known elevated risk – mobile students. Despite the evidence that longitudinal growth models, particularly growth mixture models, can be highly accurate (Bowers et al., 2013; Bowers and Sprott, 2012a; Muthèn, 2004), each additional year of longitudinal data included in the model would result in roughly 3,000 fewer students able to receive scores – a tradeoff that was not

[13] For model training and prediction the data needs to be in a single row per student shape. As Tables 10 and 11 show, each row consists of the predictors from the time period of interest, and of the outcome variable of interest – in this case on-time graduation.

deemed acceptable in the Wisconsin implementation.[14] The DEWS approach can be thought of as building a single link from a student record in one of the middle grades to the time of expected graduation, ignoring all factors outside of those two time points.

Missingness due to uneven reporting of some data elements we may desire to include in the EWS is another concern. For example, in Wisconsin in-school suspensions are only collected for students with disabilities. Including in-school suspensions as a predictive variable provides additional information about this subgroup of students but the parameter estimate is biased with respect to students for whom such data are not available.

A final concern comes from appropriately identifying students as dropouts or non-graduates. Careful consideration should be made about students who exit the district or state. Researcher need to know and understand the rules for classifying a student as belonging to a graduation cohort after a transfer. As Bowers and Sprott (2012a) note, studies of early warning indicators vary significantly on this point. This variation makes comparison across studies difficult, but it is not problematic to the accuracy of the predictive model itself as long as the decision rule is consistently applied across grade and year training sets. In Wisconsin, cases are only dropped under two scenarios. First, if a student has any of the key predictors missing in the grade level being used for prediction, and second, if the student later transfers out of the public K-12 system and is not included in state graduation rate calculations.

4.4.3. Center and Scale

Centering and scaling data is the process of subtracting the a mean from a continuous measure, then dividing it by a measure of spread, such as the variance, standard deviation, or a multiple thereof (Gelman and Hill, 2006). For continuous measures this method can attenuate year-to-year fluctuations in assessment scores due to form effects or specific test administration effects, and it can help reduce non-normality in measures like attendance rates. Scaling and centering the data also makes estimation methods like maximum-likelihood perform more efficiently, saving computation time and avoiding errors (Gelman and Hill, 2006; Kuhn and Johnson, 2013; James et al., 2013)

The main argument against centering and scaling continuous data is that it makes interpretation of individual coefficients more complicated. In the DEWS framework interpretability is downplayed in favor of accuracy and computational stability. The second argument against centering and scaling is that doing so introduces additional complexity to the prediction system. The most problematic question is what values the data should be scaled and centered against. If the data is scaled and centered against the test data set, then the means and standard deviations used need to be preserved, and then applied consistently to all future scoring data sets. Additionally, if students are scaled and centered in reference to their peers in the same school or district, problems arise in cases where the reference school does not exist or the size and composition of the school changes from the cohort the model is trained on to the cohort receiving predictions.

In the Wisconsin implementation, continuous variables – assessment scores and their polynomial transformations, attendance rates, cohort mean assessment scores, cohort attendance rates, and cohort demographic attributes – were centered and rescaled. These transformations were conducted on the entire training cohort, pooling all students, and then preserved to be applied to the scoring data cohorts in the future. The scaling and centering is done in conjunction with

[14]Note that imputation methods and model averaging methods may allow models to be fit for students with less than complete data. `http://www.satisfactorily.com/missing-data-resources/`

the model training process and new centers and scales are calculated each time new models are trained. This scaling and centering process also occurs during the construction of the imputation models.

4.4.4. Recode

Many of the elements contained in administrative data systems are categorical, such as student demographic characteristics, categories of discipline events, progress indicators like retention, and status indicators like English proficiency. Before conducting an analysis, special attention needs to be paid to how these variables are coded.

The key tradeoff in coding these variables is between losing information by reducing the number of categories versus losing the ability to generate estimates due to sparse counts within each category. In most cases, an EWS will have enough degrees of freedom to estimate several variables with many categories, but that does not mean there are not tradeoffs resulting from preserving many categories for several variables. An instructive example is the case of how to code students with disabilities (SwD). In most cases, using a binary indicator of SwD throws out too much information because of the substantial heterogeneity that exists in the group in terms of assessment scores and graduation rates. At the same time, using the 13 SwD categories that correspond to the federal SwD codes presents issues of sparsity. Since some of these categories have very few students, using them results in inefficient or unstable estimates of the coefficients and biases $\hat{y}$ for some students.

A related concern is when a category exists for which only one outcome of Y is observed (e.g., all the students graduate). This may happen when disaggregating subgroups within districts or schools – for example, a school that graduates all of its students who are FRL-eligible. This is formally known as model separation or quasi-separation and, depending on the statistical model and software used, leads to various forms of inefficiency and bias in the resulting model when used to predict new data. [15]

One solution to these problems is to examine categorical variables with respect to our dependent variable. Categories with a small count that are very similar to one another in terms of the group mean and distribution of the dependent variable are candidates for being collapsed into a larger category. Other variables that may be examples of this include:

- Free and reduced lunch status
- Prior year binary on-track indicators
- English proficiency level
- Heavily skewed continuous variables
- Discipline offenses
- School indicators

An additional concern is resolving group memberships that change dynamically (e.g. a student entering and exiting free and reduced price lunch status from one year to the next). The Strategic Data Project (2012) recommends using any longitudinal exposure to groups such as free and reduced lunch or disability status to smooth out inconsistency in reporting and self-identification among high school students. This needs to be balanced against the potential bias

[15] For a practical treatment of this and further resources refer to: `http://www.ats.ucla.edu/stat/mult_pkg/faq/general/complete_separation_logit_models.htm`

against students with shorter records in the data system – requiring multiple years of longitudinal indicators for individual students has drawbacks associated with record availability that were described above. In Wisconsin, only the student status in the year being trained or scored is used to avoid dropping students with only a single year of data.

4.5. TRAIN

Until now the approach described mirrors the approach of any quality data analysis and would apply equally to most studies already in the literature. Now, I depart from the prior literature and the inferential modeling framework and introduce the novel approach applied in Wisconsin – statistical learning. The model training process is built on the steps of searching through candidate models, selecting some subset of "best" models, averaging (ensembling) those models into a single predictive model, and preserving this ensembled object for later use in scoring new cases. This approach seeks to resolve the biggest shortcoming in the literature on EWIs; the lack of focus on out-of-sample model testing. In this section, I describe the details of the core DEWS functionality: the training and evaluation of statistical models with high out-of-sample predictive power.

4.5.1. Search

The training step begins with a broad test of the accuracy of many algorithms. What are we searching for? The ultimate prize is a subset of models with good out-of-sample predictive accuracy that are computationally feasible. These models become candidates for ensembling. Instead of focusing on building a single best model, this framework seeks to evaluate a large number of models in order to identify the most promising candidates. It is this systematic process of testing the accuracy of models that forms the heart of the DEWS approach. This technique, sometimes referred to as data mining, is critical in a situation where a 1% increase in correct classification can mean the correct assignment of extra attention and resources to hundreds of students otherwise at risk of not completing high school on time.

In order to find algorithms that best predict the outcome of interest, we must first specify the properties of a "best" prediction. There are a number of possible metrics available to judge the accuracy of a given algorithm including, but not limited to, Kappa, area under the curve (AUC), sensitivity, and specificity (Kuhn et al., 2013; Kuhn and Johnson, 2013; James et al., 2013; Hastie et al., 2009). As we have previously seen, in the early warning literature, Bowers et al. (2013) established a baseline of comparison using the receiver operating characteristic (ROC) measures of sensitivity and specificity. The ROC of a classifier can be summarized by the AUC; the closer the AUC is to 1, the better the model is at classification (Hanley and McNeil, 1982).[16]

Once the decision about how best to estimate the model accuracy has been made, the search can begin. DEWS employs a four-step procedure applied to the pool of models available within the `caret` package in the R statistical computing language (Kuhn et al., 2013; R Core Team, 2013). The procedure consists of identifying the available cohorts with graduation outcomes; pooling them; splitting them into test and training sets based on the total number of cases available; estimating the in-sample and out-of-sample fit for each model trained; and finally selecting the best N models.

[16] As an indicator of classification performance, the AUC has received some criticism for noisiness, particularly with small samples (Hand, 2009; Lobo et al., 2008; Hanczar et al., 2010). However, in the current case, the classifier being evaluated is fit to an entire population of students, so small sample sizes are not a concern.

In the literature, model accuracy is generally only reported based on the data used to build the models. While the DEWS training data has a substantial number of observations, we still need to evaluate the performance of our models on data outside of the training data by using a test set (Hastie et al., 2009; James et al., 2013). This provides the best estimate of the performance of DEWS on current students and ensures the models are not overfit to the data and produce biased predictions. In the EWS framework, the best test set would be an entire subsequent cohort of students. This most closely resembles the threats to accuracy a model faces, because a model must be valid not only on new cases, but on new cases drawn from a separate school year. However, using an entire subsequent cohort of students requires a substantial amount of data and may not be feasible in all implementations.

Table 6: Tradeoffs with Different Test Set Selection Methods

Method	**Data Loss**	**External Validity**
Hold 1 Cohort Out	Highest	Highest
Random Sample from 2+ Pooled Cohorts	High	Higher
Simple Random Sample Within Training Cohort	Moderate	Low
Stratified Sample within Training Cohort	Moderate	High
Repeated Fold Cross-Validation	Low	Moderate

Table 6 displays some recommendations on the tradeoffs associated with different ways of selecting a test set.[17] For initial model building purposes, repeated fold cross-validation offers an efficient use of limited data to estimate the error rate of the test set, and can be replaced with a held out cohort as more data becomes available. Attention must be paid to sampling choices such as whether to oversample the dropout class in cases where there is severe imbalance between the two groups (Kuhn and Johnson, 2013; Kuhn et al., 2013). Of further consequence is that, given the amount of computational time necessary for some algorithms or data sets, it may be necessary to use some technique in Table 6 in order to test multiple methods in a reasonable amount of time. In the case of DEWS, it takes over 48 hours per grade level to run all of the candidate algorithms using the cohort hold-out strategy, and only approximately 8 hours to use a reduced sample of $\approx$ 40,000 cases selected from pooled cohorts. Depending on the volume of data, particular algorithms chosen, and nature of the problem, an analyst will have to balance these concerns with the computational resources available.

The entire procedure is implemented using the `caret` package for the R statistical computing language augmented by (1) the `EWStools` package for making ROC comparisons and searching across model methods and (2) the `caretEnsemble` package for averaging multiple models together to improve accuracy (R Core Team, 2013; Kuhn et al., 2013; Knowles, 2014; Mayer and Knowles, 2014). [18]. The `caret` package provides access to over 149 model methods, 119 of which are suitable for binary classifciation. Many of these models are very similar

[17] While the theory of defining a test data set is outside the scope of this paper, see (Hastie et al., 2009; James et al., 2013; Kuhn and Johnson, 2013).

[18] R is particularly well suited to this task due to its open nature and the availability of high-level representations of hundreds of statistical models. Many of these are made freely and easily available through user-contributed packages (R Core Team, 2013). Additionally, the open source nature means that the "black box" of these methods is more easily unpacked for communication and explanation. Another great alternative is Python and `Scikit-learn` (Pedregosa et al., 2011). See Aguiar et al. (2015) for an example of a similar system implemented in Python

to one another, and some of them are not suitable for large data sets due to the computation time involved.

The goal of the search procedure is to evaluate models from across this space in order to introduce diversity into the predictions (James et al., 2013). Unlike traditional regression, many of these models have so called "tuning parameters", fixed user-specified values that direct the algorithm how to proceed. Currently, DEWS searches through 35 of these algorithms, which were selected for their diversity, their efficiency in training, and their compatibility with the OS and computer hardware available.[19] The search procedure splits the data into a training and a test subset. The proportions are based on the number of cases available, but the maximum size of the training set is capped at $\approx$ 40,000 cases to allow for the search to complete.[20] The data available for all cohorts with graduation outcomes is pooled and the training set is composed of roughly 40,000, selected randomly and stratified to reflect the population balance between on-time graduates and not.

Next, the training procedure is established – each algorithm is evaluated using 10 fold cross validation. This provides the estimated AUC (which is the element DEWS seeks to maximize) for the test data. Models are all tested across a series of values for the tuning parameters unique to the algorithm, a process which is automated using the `train` function in the `caret` package and the `tuneLength` parameter, which is set to eight. This means that each algorithm's parameter space is split into eight and the cross-validation procedure is repeated for the model at each of the eight values of the tuning parameters. The `caret` package reports back the tuning parameters that produce the best model fit, as well as the resulting accuracy.[21] Once the best tuning parameter is selected for each model, the model accuracy is assessed on both the training and the test data using testing functions from the `EWStools` package and the pre-defined test set. The result is a data set that describes the accuracy of the models in a format similar to Table 12 in Section 5, with the name of the model method in R, the record of whether it is applied to the test or training set, the AUC, the estimated standard deviation of the AUC, and the computational time. These results can be used to diagnose which models are the most suitable deployment and will be discussed further in Section 5.

4.5.2. Select

Using a result set like that in Table 12, from the approximately 30 models that complete the search procedure, I select a subset for inclusion in the final ensembled model. With well over 30 models to choose from, how do we go about selecting the best models? As mentioned above, the choice of test statistic depends on the application of the model – which requires consultation with the intended user[22] – but a strong candidate is the AUC. The closer the AUC is to 1, the closer the model is to perfect classification.

Figure 11 depicts the AUC for each method searched in DEWS for the 7^{th} grade cohort. Immediately we see that most of the models are clustered around an AUC of between 0.83 and 0.87. Further testing remains to be done, but two approaches stand out as possible for selection: selecting only the models with the highest AUC, or alternatively selecting from among

[19] For details, see the Technical Appendix for a reproducible script on a sample data set. For imputation, a separate set of models is tested in the case of problems with continuous variables.

[20] In practice, the accuracy on this training set very closely approximates that of the test set in the Wisconsin data.

[21] A complete script of the test procedure is available for review.

[22] Involving intended users throughout the process is critical in building a predictive model. The method and importance of doing so are beyond the scope of this essay, but cannot be emphasized strongly enough.

the models based on the diversity in their algorithms. There is some evidence that selecting diverse algorithms yields better performance than selecting the best fitting set of models to ensemble (Kuncheva and Whitaker, 2003; Sollich and Krogh, 1996; Kuhn and Johnson, 2013) – but more work needs to be done. Initial results for both methods are demonstrated in the results section below.

The reader may wonder what the benefit of implementing the complicated procedure described above is when the end result may simply be a familiar generalized linear model. The motivation is two-fold. First, the model selected was not selected because of its within-sample model fit or because it represented some a priori theory of why students fail to graduate on time. Instead, it was selected expressly for its ability to provide accurate predictions on future cohorts of students, its rate of accuracy is now known thanks to this procedure. Second, there is no guarantee that with more years of data and new patterns of student graduation the ranking of these models in terms of accuracy will be stable. By setting up a systematic testing environment that searches for the most accurate models within a constrained model space, the analyst ensures that greater confidence can be conferred on the accuracy of the model's predictions. When providing tens of thousands of predictions to thousands of educators within a state, this due diligence is absolutely necessary.

The number of models to select is a question of computing power and time. In the Wisconsin implementation, I select between 5 and 8 models depending on the grade and the number of models that successfully train. Models are selected by their performance on the test data, and some effort is made to select for algorithmic diversity. There are functions in the `EWStools` package to help users select a set of available models that are dissimilar from one another (Knowles, 2014). Future work remains to be done to explore the value in different selection procedures for ensembles.

4.5.3. Ensemble

In some applications ensembles, or averaging predictions from different models, can greatly increase accuracy. In other cases, ensembling provides a way to hedge against overfit, particularly at the fringes of observed values of predictors, by borrowing the strengths of various functional forms used in different algorithms (Kuncheva and Whitaker, 2003; Sollich and Krogh, 1996). Ensembling, or model averaging, is a general technique used in inferential as well as predictive modeling to formalize the uncertainty inherent in individual models (Burnham and Anderson, 2002; Gelman et al., 2013; Kuhn and Johnson, 2013). In predictive modeling, many individual algorithms, such as boosted tree and random forests, employ ensembling internally (Breiman, 2001a; Kuhn et al., 2013; Friedman et al., 2000; Kuhn and Johnson, 2013; Hastie et al., 2009). The practical benefits of ensembling in the DEWS case are illustrated below in the results section.

In DEWS, the N models selected in the previous step are ensembled using the `caretEnsemble` package in R. This package determines how much weight the prediction from each model should carry using an optimization algorithm (Mayer and Knowles, 2014). The package currently implements multiple algorithms for this purpose, and the Wisconsin implementation utilizes one with a stopping criterion to ensure that the ensembled model always has an AUC equal to or better than the best individual model. The resulting vector of model weights is then used to combine predictions from the composite models into a single prediction. The package can also provide the same functionality for continuous variables by optimizing the RMSE of the ensem-

bled predictions. In addition, users can also fit a meta-model to the predictions from the library of models to combine predictions using any method available to `caret` (Mayer and Knowles, 2014).

4.5.4. Export

After the model is ensembled, the model object has to be stored and preserved in order to be used to predict future cases. Unlike simple rules-based or regression models, DEWS produces complex model ensemble objects that package multiple algorithms together. As such, the model object is stored as a binary file. An alternative would be to use less complex models and store them using something like Predictive Modeling Markup Language (PMML) or convert them into database code that can be used to score cases directly in the database.[23] PMML has a substantial advantage in speed compared to saving and loading binary model objects from a hard disk. In many cases the tradeoff is between the speed at which predictions need to be supplied to the user, and the complexity of scoring procedure.

4.6. SCORE

The scoring subroutine is the crucial feature of any EWI; without this step, no risk scores would be produced to share with educators. The score subroutine serves two purposes. First, it allows model fit to be evaluated by predicting outcomes for the test set and evaluating the model accuracy. Second, it allows new cases to be scored – the ultimate goal of the predictive model. In both scenarios, the subroutine has the same basic steps: load, predict, export, and test. All early warning indicator systems have to have a method for scoring new cases, and, as we have seen, many of these simply involve setting cut points based on currently observable student attributes (Balfanz, 2009; Easton and Allensworth, 2005). A key design decision is to think about the complexity of the scoring method up-front when building a system that integrates with other data systems.

4.6.1. Load

Storing models is an important technical consideration. It is important to ensure that the correct model is loaded for scoring current cases. In the Wisconsin implementation models are stored based on two parameters – the grade level they were trained on and the variable they predict. In an imputation workflow, several models will be loaded at once. Having good metadata and model library organization is key to efficiently leveraging the multiple models created in the DEWS framework and avoiding the misapplication of models to the wrong data.

4.6.2. Predict

DEWS results in predicted probabilities of graduation ranging from 0 to 1 for each student. These probabilities are computed by the `predict` method in R, which has functions that connect to the various model objects stored during the ensemble routine. Making such predictions presents two key decisions to analysts. First, all predictions should carry with them a measure of the uncertainty surrounding the prediction. In a generalized linear model, the prediction interval can be derived from the deviation of values of x from the sample mean, the standard deviation

[23] PMML is an open standard and can be found online: `http://www.dmg.org/v4-2-1/GeneralStructure.html`

of x, the sample size, and a transformation involving the t-distribution (Gelman and Hill, 2006). Statistical packages can easily calculate confidence intervals and prediction intervals for such analyses. In the case of statistical learning models, prediction intervals are not always readily available for all algorithms (Kuhn et al., 2013). In these cases, determining the proper prediction interval is not straightforward and varies with the modeling approach taken by the analyst. In the Wisconsin case, the predictive interval is represented by the average disagreement among the predictions generated by the ensembled models, weighted by the model weights. Thus, a weighted standard deviation is calculated from the predictions of each individual model to represent the amount of disagreement among the models (Mayer and Knowles, 2014). This approach has the drawback of ignoring the uncertainty inherent in the predictions of each individual model, a drawback that has not yet been resolved. Many desirable alternatives to this approach exist in other predictive frameworks, such as bootstrapping the prediction intervals or simulation methods (Gelman and Hill, 2006).

A second concern is how to handle predictions in the case of missing data. If data is imputed, should the uncertainty of the imputation predictions be taken into account using an approach like multiple imputation (Hastie et al., 2009; Honaker et al., 2011)? In the Wisconsin case, missing values are not imputed, and students with insufficient data in the current year are given a risk score of "unknown." This risk score is considered the highest risk and schools are urged to review these cases first (Knowles and White, 2013). Future work remains to explore using models with a reduced subset of predictors (which will be less accurate) to serve as fallback models for incomplete cases. The ability layer nested models to provide predictions from a reduced set of predictors demonstrates the strength of the modularity designed into DEWS – its ability to be modified to meet future demands of users.

4.6.3. Test

The two main tests performed after models are ensembled assess (1) their AUC on a third validation set of data (2) the balance between the false alarm and true negative identification from Table 1. To perform these tests it is necessary to choose a cutpoint in the continuous probability of each student above which a student is identified as "at-risk" for late- or non-graduation. The desire to categorize students comes from feedback by practitioners; the process used in DEWS is described in some detail by Knowles and White (2013). As we will see in Figure 4, choosing a cutoff at the extremes (e.g. 0.01 or 0.99), all models generate similar classification accuracy. We can think of moving along the probability curve to select a cutpoint as an adjustment to our tolerances for false-positive and false negative classification. If false-positives and false-negatives are of equal importance identifying the best threshold is straightforward. Using the R `pROC` package (Robin et al., 2011), we can choose to either identify the threshold that is the closest to the upper-left corner of the graph, or we can choose Youden's J statistic, which maximizes the distance between the curve and the identity line (Youden, 1950). I opt for the "closest top left" method.

However, in the case of identifying potential dropouts, I do not treat false-positives and false-negatives having equal weight. Working with educators and content specialists within the Department of Public Instruction, I learned that false-positives (i.e. falsely identifying a student as a potential drop out), are considerably less problematic than a false-negative (i.e. falsely identifying a student who is likely to dropout as a potential graduate). Across many data elements available to DEWS, students who graduate late or dropout can often look very similar

to students who graduate on time – especially in early middle grades. We do not observe what influences some of these students to graduate on time and others to not. Early identification and intervention could be one such factor – thus, identifying these students is of importance to practitioners, even if it increases the false-positive rate. Additionally, schools looking to intervene in early grades to help students are likely to already identify the lowest performing students on many of the measurements included in DEWS, thus, the real benefit of the system comes from identifying the next tier of students who are at risk, even though this also increases the false-positive rate. In the ROC framework, I weigh false-negatives as much more costly, explicitly stating that we are willing to accept roughly 25 additional false-positives for 1 less false-negative. Note that this decision does not have any bearing on the final linear predictor, or $\hat{Y}$, but only on model selection and the cutpoint used to classify students into categorical descriptions of risk. In practice, schools can set their own threshold for risk and their own tolerance for false positives and false negatives based on the predicted probabilities generated from the EWS model (Knowles and White, 2013). As Figure 4 shows, unless they choose extreme values, the more complex model is likely to still provide a superior fit.

4.6.4. Export

DEWS models are used to build an export file that includes the predicted score for each student as well as the individual data elements that make up the predicted score. This file is loaded into a reporting system for dissemination to the eventual users of the system. In the export step, the Wisconsin approach applies a unique method to identify student risk "domains." In an generalized linear model framework, the large statistically significant coefficients of the model might be used to determine the risk elements for a student. An alternative approach for models that do not produce coefficients is to fit a linear model predicting students' predicted scores and scoring students based on the elements that are most significant in this model.

Instead of these approaches, in Wisconsin, individualized student risk levels are calculated using a nearest-neighbor approach. Data elements are broken down into four domains – academics attendance, behavior, and mobility (Knowles and White, 2013). Then, the individual elements that make up each of these domains is aggregated. From these aggregations, the student is compared to the distribution of students who were identified as "low," "moderate," or "high" risk. Whichever group's distribution the student falls within, that is the risk level assigned to that student for that sub-domain. This way of scoring sub-domains has a few key advantages: it makes the sub-domain scores transparent; it allows sub-domain scores to be calculated for students with incomplete records and no overall score; it avoids concerns about multicollinearity among the key predictors misrepresenting their influence on the overall outcome; and it provides a sense of direction for steps to take with investigating the results for an individual student.[24] For more details on this approach and to see examples of how scores are presented and schools are advised to interpret scores, see Knowles and White (2013). The student overall risk, predicted probability, prediction error, and sub-domain scores are then saved to an export file for loading into the data warehouse. Once in the data warehouse these records are combined with students' current records and displayed within the statewide business intelligence dashboard system known as WISEdash.

[24]The drawback to this method is that the sub-domain scores do not always align with the overall score, resulting in the most frequently asked DEWS question: "Why does a student with all 'low' sub-domain scores have an overall risk of 'high'?"

5. EXAMPLES AND RESULTS

In practice, how do DEWS models perform? I begin by comparing DEWS to some benchmark binomial regression models. Then, I compare DEWS models and the GLMs to other EWIs in the literature. Finally, I look at the specific ensembled models that form the prediction engine for DEWS currently, demonstrate their performance with Wisconsin data, and explore their properties.

5.1. LOGISTIC REGRESSION

As a benchmark, I start with as complete a set of the ABC indicators as available in the data system and use a traditional logistic model as a starting point, creating models of the form:

$$P(y_i) = \alpha + \beta[X_i] + \epsilon_i$$

In this case, X_i is a matrix of individual student characteristics. In the first model I include the following student characteristics from Table 5: math and reading assessments, attendance rate, days of suspension and expulsion, student demographics, and school mobility. These measures were chosen because they were available for all students in the SLDS and there is prior research demonstrating that they have an effect on student graduation. For demonstration purposes, I also fit simple models of assessment data, attendance data, and discipline data independently. All models are fit on the pooled grade 7 training set, which consists of a sample of roughly 40,000 7^{th} graders with graduation outcomes.

As shown in Figure 4 instead of reporting coefficients and traditional model fit statistics, I benchmark the ABC models directly against the EWIs from Figure 1 (Bowers et al., 2013). Figure 4 compares five models – (1) using only attendance and demographics, (2) using only behavior and demographics, (3) using only assessments and demographics, (4) a simple model combining all three, and finally (5) a more complex model that also includes school fixed effects. Combining the ABC indicators, as first demonstrated by Balfanz and Iver (2006), does result in a significant improvement in accuracy.[25] The figure demonstrates that the full model used in this example is roughly equivalent to the Chicago On-Track indicator and clearly superior to the Balfanz ABC flags for middle school. The two best indicators, the Muthèn and the Bowers growth mixture models (GMM) stand out as far above all the other methods.

Figure 4 shows three important things. First, for in-sample prediction on the training data, the regression strategy used here provides strong predictive power relative to other early warning indicators – the majority of which have very low rates of correctly classifying dropouts. Second, combinations of EWIs together improve the predictive power notably – both in the case of the Balfanz predictors, and in the case of the regression model fit here. Third, longitudinal growth mixture models are clearly a superior approach in terms of predictive power, and they stand out above the regression models and the indicator methods.

However, while this comparison is useful for a relative understanding of the power of various predictive indicators, it is deficient for two key reasons. First, there are inherent tradeoffs, discussed above, in using longitudinal data for individual students. While the GMM models in (Bowers and Sprott, 2012a; Muthèn, 2004) are highly predictive, they require several years of

[25] Figure 4 depicts the models fit to start the EWS building procedure as curves, while other EWIs are represented as single points. This is because the logistic regression procedure we use generates a continuous variable, a predicted probability, instead of a single binary classification.

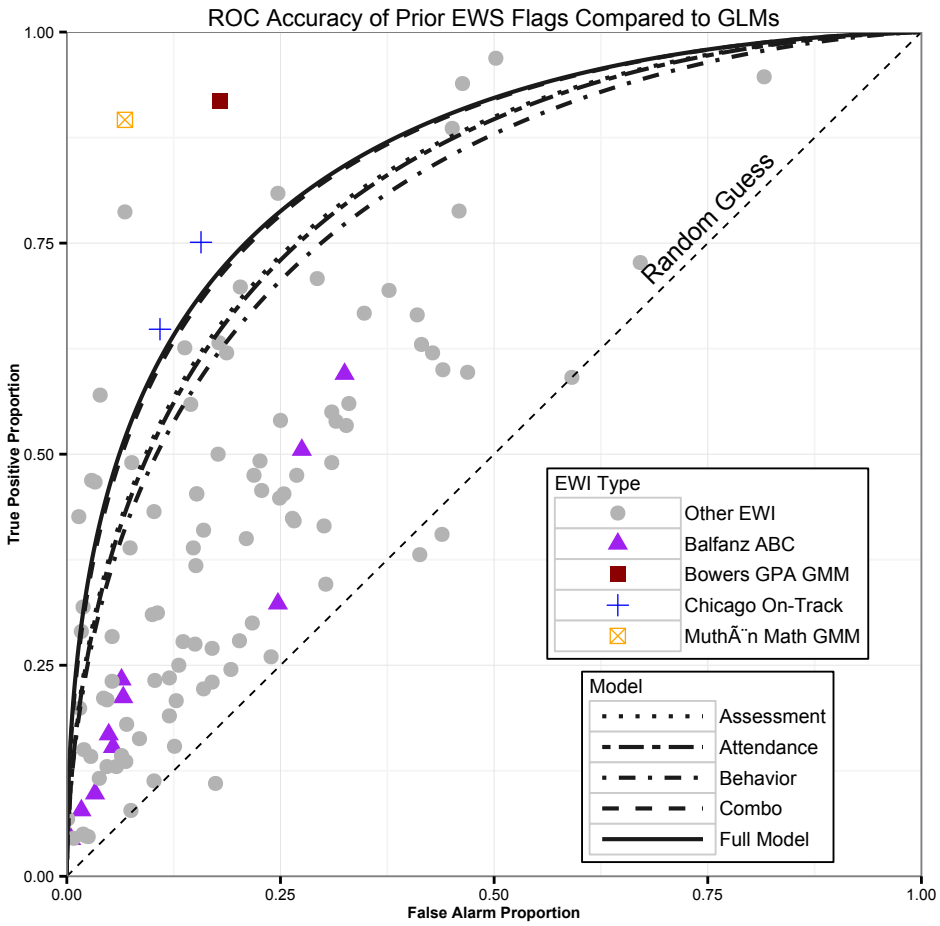

Figure 4: ROC curves of all models compared to prior EWI indicators. Each model includes demographic variables of race, economic status, gender, ELL status, and SwD status in addition to the student indicators indicated.

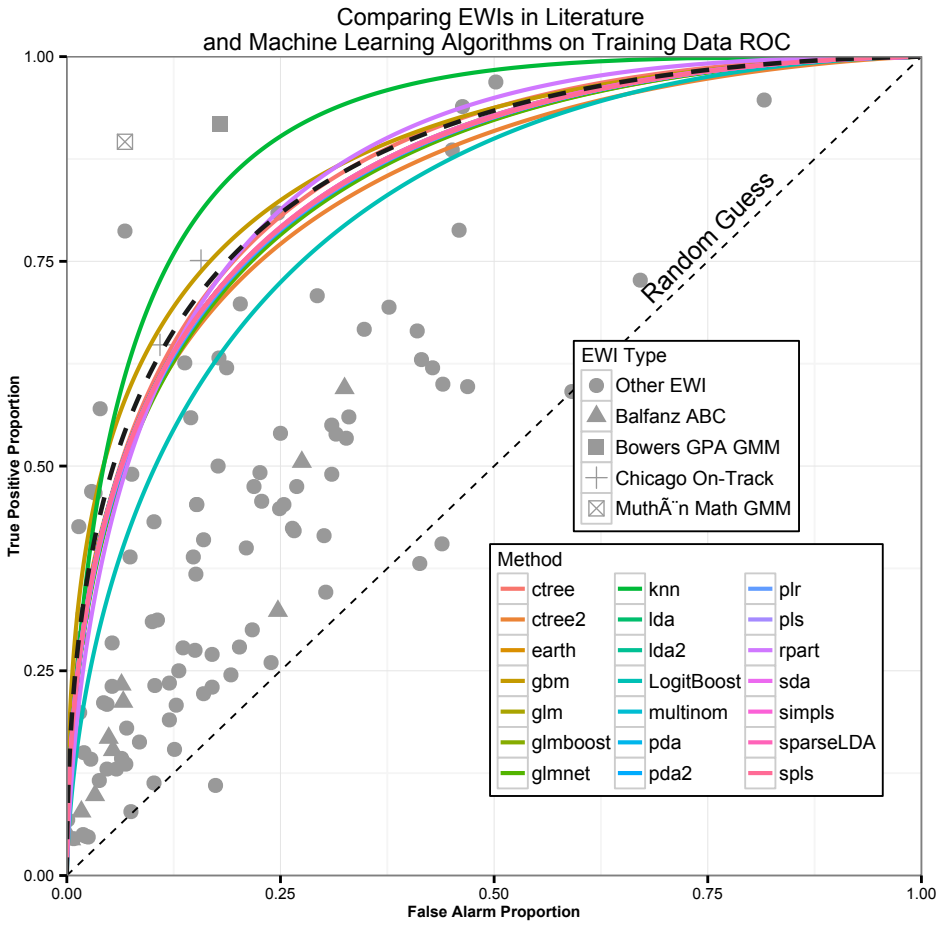

Figure 5: ROCs for Machine Learning Algorithms Implemented on Training Data

consistent records for each individual student. Second, simply comparing the accuracy of EWIs on the data sets upon which they were developed masks the major challenge of creating a strong early warning system – the need to make accurate predictions about new data on an ongoing basis. This will be examined in detail below as I demonstrate the performance of the statistical learning and ensembling approach.

5.2. Applied Models

Figure 5 shows the accuracy of the various individual models (in color) and the ensembled model (in black) on the training data. Over 30 models are included in Figure 5, and their performance on the training data has much more variation than the various flavors of logistic regression shown in Figure 4. Notable in Figure 5 is the bright green line representing the k-nearest neighbor algorithm. While this model performs very well on the training data, as shown here, it is not included in the ensemble model because the ensemble model selects models based on their superior performance on the test data. The k-nearest neighbor algorithm experiences a significant drop in performance when predicting the test set – a likely sign that the algorithm is overfit to the training set. Figure 5 also shows that a wide array of algorithms easily outperform the vast majority of EWIs found in the literature, and that , for grade 7, the ensembled model is

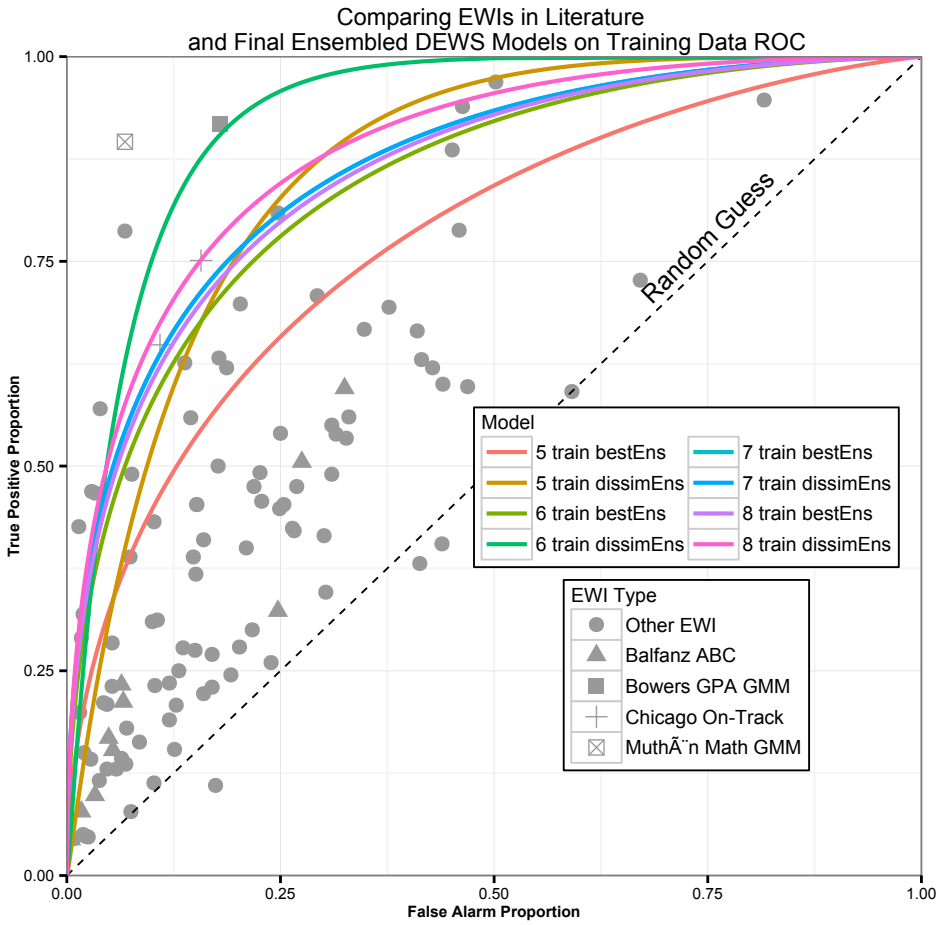

Figure 6: Ensembled models training data accuracy compared to accuracy of prior EWIs

almost competitive with the predictions made by the Chicago On-Track method for high school students.

Next, Figure 6 shows the accuracy of the ensembled models for each grade relative to the EWIs identified by Bowers and Sprott (2012a) and also compares the two alternative model selection methods described above. Overall, the performance of these models is on par with or better than the GLMs in Figure 4, and with each grade, the accuracy of the model increases – as expected. Using 5^{th} grade data, these models outperform many of the middle grade EWIs studied by Balfanz (2009). By grade 8, DEWS models are approaching the accuracy of the Chicago On-Track system but is available a full semester earlier. This early warning is crucial as schools make their planning and staffing decisions in preparation for the new school year. There is also some evidence in the DEWS results that selecting models to be ensembled based on their dissimilarity from one another may be a more effective strategy than ensembling the best fitting models. This is an area that merits further exploration.

In both Figure 6 and 5 the logic of the DEWS framework becomes clear. There are limits to how accurate a model can be if it imposes a uniform relationship between predictors and outcomes. Since we know that dropouts do not themselves stem from the same causes, employing a more flexible modeling strategy is the best chance to capture a larger portion of dropouts. This

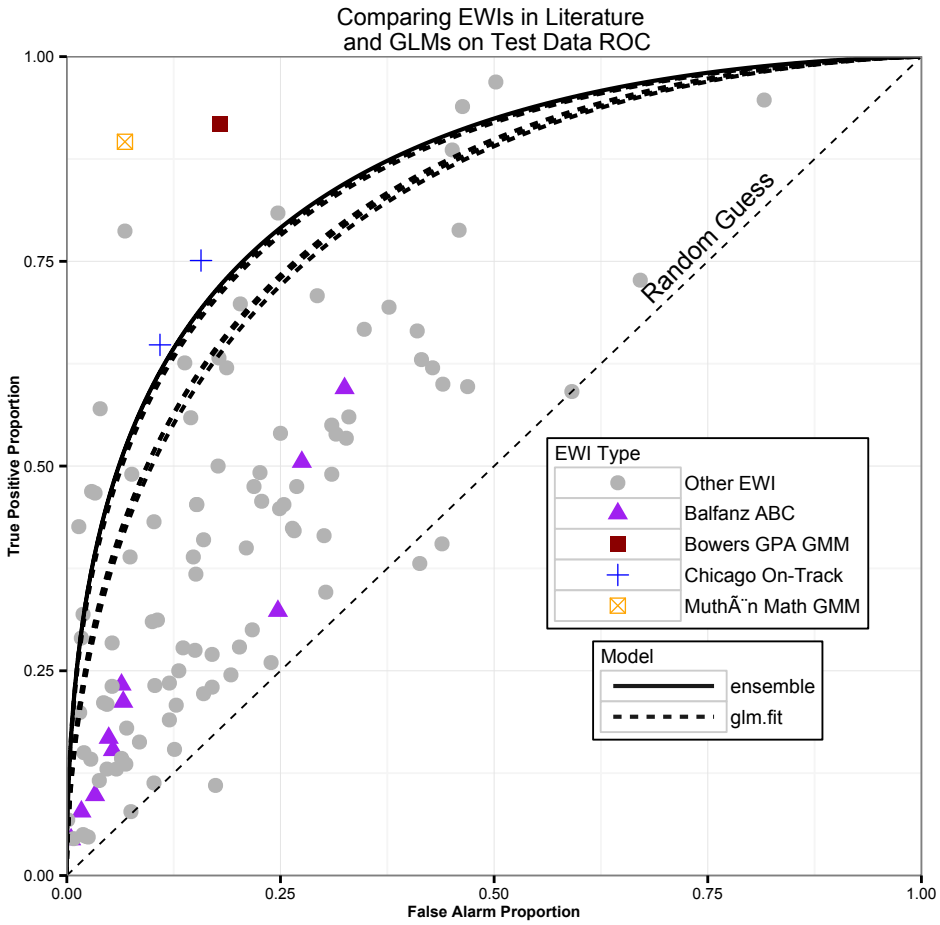

Figure 7: Ensembled model training data accuracy compared to prior EWIs

flexibility is important in environments like Wisconsin, where the problem of identifying students who will graduate late or not at all is particularly difficult given the high overall graduation rate. Yet, the real advantage of the Wisconsin approach comes when comparing the performance of the models on the hold-out test data set – where accuracy naturally falls, but which serves as an estimate of the accuracy of the implemented system.

5.3. Out-of-Sample Comparison

Until now, I have focused on the accuracy of the models on their training data. Figure 7 shows the performance of both the generalized linear models and the ensembled models with the test data. The ensembled models for grade 7 is shown in black and the GLMs are shown in gray.

Table 7 shows the confusion matrix for the ensembled model with a threshold of 0.8 dividing on-time graduates from those who fail to graduate on-time. The results are for the full grade 7 cohort including the training, test, and validation data sets. This helps trasnalte the ROC curves above into the metrics described in Table 1. The result is a sensitivity of 0.86, a specificity of 0.67, and a negative predictive value of 0.35. This means that, at the start of 8^{th} grade, DEWS has identified nearly 70% of students who will not graduate on time. Of the students identified,

approximately 65% represent a false alarm and will graduate on-time. Given the extremely high on-time graduation rate in the state and the earliness of the identification these results are encouraging. The practical communication and interpretation of these results is discussed further in Knowles and White (2013).

	Grad	Non.Grad
Grad	84,744	3,670
Non.Grad	13,718	7,454

Table 7: Confusion Matrix for combined train and test data in Grade 7: Ensembled Model

While Table 7 shows the confusion matrix for the ensembled model, Table 8 shows the performance of the best logistic regression on the same data. The ensembled method does not result in a large improvement over the logistic regression. In fact, only about 858 more non-graduates were identified in exchange for 3,540 additional false-positives. Both methods, however, are greatly superior to other middle grade methods in the literature and approach the accuracy of high school early warning tools.

	Grad	Non.Grad
Grad	88,313	4,512
Non.Grad	10,149	6,612

Table 8: Confusion Matrix for combined train and test data in Grade 7: Logistic Regression Model

So why would practitioners adopt a machine learning approach vs. logistic regression approach? For some applications the logistic regression approach may be preferable. Its advantages include relative ease of implementation, low computational costs, and more familiarity within the education research community. Additionally, for analysts with a strong background in regression modeling, the generalized linear model has helpful properties such as the ability to calculate prediction intervals for individual observations and to estimate group-level parameters (with an extension to a multi-level modeling framework) for grades, schools, or districts.

The machine learning approach has advantages as well. First, if an analyst does the logistic regression step properly and constructs validation sets, there is not much additional cost to applying machine learning algorithms and comparing performance. Second, while in the Wisconsin case the logistic regression method performed well, this analysis provides no evidence that practitioners can expect similar performance with different samples, different predictor variables, or different underlying rates of graduation and dropout. The systematic machine learning approach provides a method for continually identifying high-performing models in the face of changing data availability. Changing data availability is a particular problem for logistic models that include school fixed or random effects because estimation of those effects is not possible for new schools. Machine learning approaches can provide the accuracy of fixed effects while retaining the ability to score students attending new schools or adjust to changes in the populations served by particular schools. Furthermore, the machine learning approach prevents models from losing accuracy over time as underlying relationships between attributes and graduation change. In exchange for the added complexity of the machine learning approach, analysts gain a systematic approach that guards against parameter drift.

5.4. Inside the Black Box

A common criticism of the machine learning approach is that the models it produces are a "black box" with difficult to interpret relationships between predictors and the final probability. One way to unpack the black box is to use the `caret` method `varImp`, or variable importance, within the `train` package (Kuhn et al., 2013; Kuhn and Johnson, 2013). This method attempts to estimate and standardize the contribution of each predictor to the final outcome across a wide variety of model types. The `EWStools` package includes an extension of this method for use with ensembled models; this extension weights the predictor contribution by the model weights. The results are shown in Figure 8. As discussed in the methods section, these results should not be interpreted as identifying any causes of relationship between student attributes and eventual outcomes. For example, while this model is accurate, much of that accuracy comes from estimating multiple combinations of the same indicator, such as math and reading assessment scores and their polynomials, or categories of discipline behavior in addition to total days of disciplinary removal. Due to this overlap and saturation of predictors, the ranking in Figure 8 is best used for diagnostics on the specific models and not for making generalized statements about the antecedents of failure to graduate on-time in Wisconsin.

Variable importance for linear models is sometimes computed using the absolute value of the t-statistic for the parameters in the model [for information on this and other methods see Grömping (2006)]. For algorithmic models, the assessment of variable importance will depend on the model type (Kuhn et al., 2013). Model ensembles, like those in Figure 8 have their own method of assessing importance where the importance of each variable in each component model is calculated using the preferred method for that model and these weights are then summed by the weight of the model within the ensemble (Mayer and Knowles, 2014). A drawback to this approach is that the method of assessing importance is not consistent across models, resulting in a potential bias based on how each model computes importance. To guard against this, the weights within each method are scaled to sum to 100 and the mean of these weights is taken weighted by each model's weight within the ensemble. Thus, Figure 8 can be interpretedas a depiction of the relative importance of predictors summed across the models and weighted by the weight of each model in the ensemble.

Figure 8 provides a way to look inside the ensembled model and explore the way the algorithm works. While more work on how to visualize and interpret parameters in algorithmic machine learning approaches is needed, variable importance provides analysts with a way to check the face validity of such models and raises questions about feature choices. The component models in the ensemble in Figure 8 include a boosted generalized linear model, penalized multinomial regression model, and a partial least squares model. The results of ensembling for all grade levels across both methods of selecting component models are shown in Table 9.

Table 9 shows the methods ensembled under both the selection of the best fitting models and the best fitting, but most dissimilar models. Both approaches show some algorithms that come up regularly and flavors of penalized regression are consistently high-performing. An advantage these algorithms appear to have is their relatively low computational costs – something analysts need to keep in mind with large data sets.

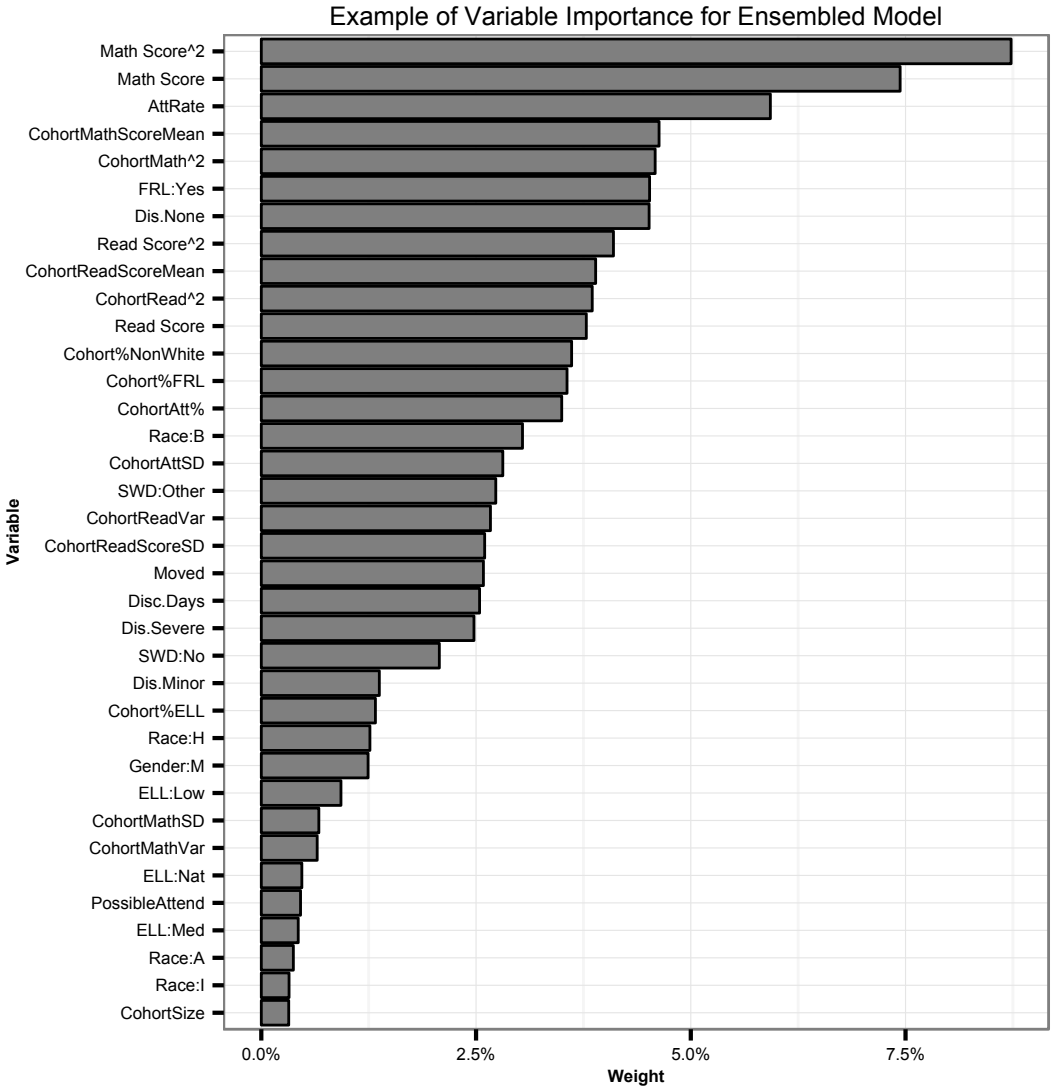

Figure 8: Variable Importance in Ensembled Model. The bars represent the proportion of total explained variance attributed to each variable in the model.

Table 9: Algorithms Selected in Wisconsin DEWS in 2013-14

Grade	**Selection**	**Algorithm**
5	Best Models	linear discriminant analysis, partial least squares, flexible discriminant analysis
6	Best Models	linear discriminant analysis, partial least squares, penalized multinomial regression
7	Best Models	linear discriminant analysis, partial least squares, GLM
8	Best Models	linear discriminant analysis, GLM, penalized multinomial regression
5	Dissimilar Models	bagged trees, partial least squares, conditional inference tree, flexible discrimminant analysis
6	Dissimilar Models	Not available at time of press
7	Dissimilar Models	boosted GLM, penalized multinomial regression, partial least squares
8	Dissimilar Models	linear discriminant analysis, stochastic gradient boosting, multivariate adaptive splines

*All models are from the caret library for R.

5.5. IMPLEMENTATION

The Wisconsin DEWS is entering its third year of deployment. Prior to year 1, in the spring of 2012, DEWS was piloted with 52 Wisconsin school districts. Districts were asked to identify if DEWS was useful, if DEWS identified students they themselves had not viewed as at-risk, and if they identified students that DEWS had not. The results of this pilot were positive so the DPI decided to conduct a full release of DEWS statewide at the start of the 2012-13 school year.[26] The release was accompanied by the publication of the Wisconsin DEWS Action Guide, a handbook explaining the predictions to educators and giving guidance about how to evaluate, interpret, and utilize the predictions provided by DEWS (Knowles and White, 2013). DEWS has been in operation ever since, producing preliminary student predictions at the start of each school year and updating those results in the spring once all data is reported. Schools and districts have secure access to the DEWS reports for their students through the WISEdash Secure statewide reporting tool.

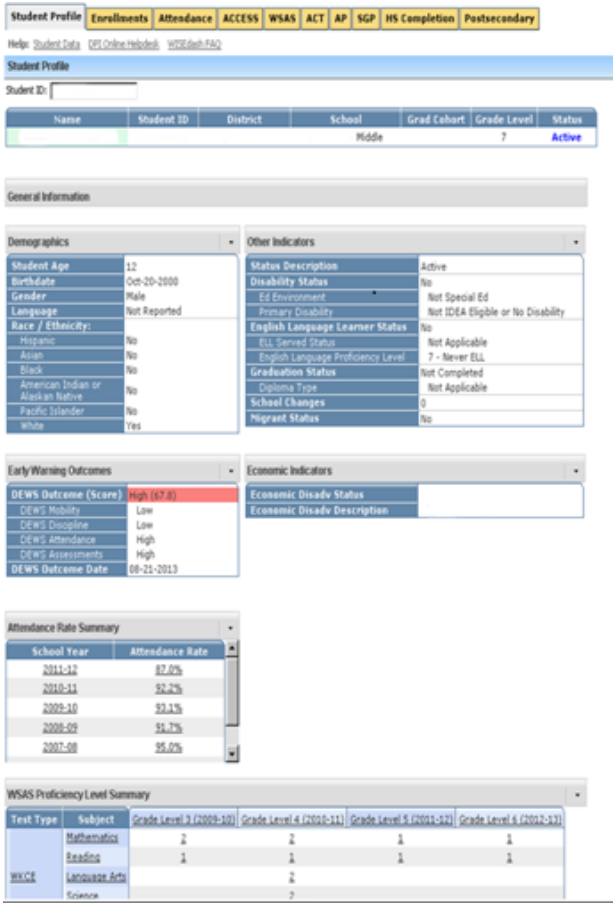

Figure 9: Example of Secure DEWS Report on the WISEDash Student Profile

Figures 9 and 10 show examples of these secure reports that authorized users see within Wisconsin's WISEdash system. Figure 9 shows the whole student profile with the EWS box

[26]For details on the implementation process and materials on the outreach to school districts, visit `http://www.jaredknowles.com/presentations`.

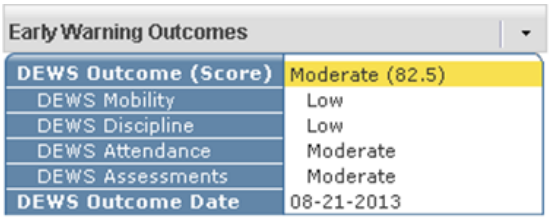

Figure 10: Detailed View of Early Warning Report on Student Profile

prominently in the middle with a red stripe for high-risk students. Figure 10 shows the detail of the DEWS section of the student profile, including indicators of the subdomains that make up a student's overall score. Users can click on these metrics to get detail about each of the data elements within the dashboard.

6. Conclusions and Future Extensions

This paper has described one approach to scaling an early warning system for high school completion to a statewide longitudinal data system. The nature of the data and the scale of the problem led to the adoption of a machine learning approach – building plausible sets of predictor variables, testing them with multiple algorithms against a test set of data, and selecting the most accurate. The tradeoffs starting with the collection and tidying the data to choosing the metrics used to select the best fitting model are described in order to provide insight into the process of constructing such a system within a particular context.

The Wisconsin DEWS represents just one approach to balancing the competing concerns posed by an early warning system. It is an approach that will constantly be revisited and reviewed in light of future developments. Currently, we are focusing on improving the accuracy of EWS models by investigating the benefits of training the models on larger training sets composed of a 50/50 balance between the classes. Development is also occurring in identifying ways to build a secondary set of models that include student growth measures for students with multiple years of data available. I am also exploring extending the group-level predictors in the model by using U.S. Census data on the neighborhoods served by schools.

The system described here is a systematic approach to optimize the predictive power of the available data and to minimize the reliance on a single model. DEWS also seeks to reduce development time and focus on exploring only improvements which will increase the accuracy of the system and, reduce the risk a student is incorrectly served or not served. In order to meet this goal, DEWS sacrifices a substantial amount of interpretability in the model – first by performing substantial transformations to the input data, next by using algorithms from non-linear families, and finally by ensembling those algorithms together.

The system presented here is the result of these tradeoffs and through careful documentation these tradeoffs can be explicitly monitored and continually discussed with internal and external stakeholders. This grounds the updating and improvement process of the EWS in data. Future extensions of the system must demonstrate their advantages empirically in order to justify the development time and additional complexity they may introduce (Sculley et al., 2014). Creating consensus around these deliberate tradeoffs and monitoring the results in a feedback loop with internal and external users is critical to the ongoing success of an EWS.

7. TECHNICAL APPENDIX

The technical appendix contains further details on the data in Wisconsin, the computing environment used to run DEWS, the details of the model search and ensemble procedures, and further notes on deploying a DEWS like system.

7.1. SUPPLEMENTAL DATA

Table 10 and Table 11 show the categorical and continuous data elements available in the Wisconsin data system.

Variable	**Levels**	**n**	%
White	0	21821	19.9
	1	87765	80.1
	all	109586	100.0
FRL Elig.	0	76704	70.0
	1	32882	30.0
	all	109586	100.0
Male	0	53759	49.1
	1	55827	50.9
	all	109586	100.0
No Discipline	0	8691	7.9
	1	100895	92.1
	all	109586	100.0
Not ELL	0	5382	4.9
	1	104204	95.1
	all	109586	100.0
Not SwD	0	13601	12.4
	1	95985	87.6
	all	109586	100.0
Moved Schools	0	107170	97.8
	1	2416	2.2
	all	109586	100.0
Minor Discipline	0	106094	96.8
	1	3492	3.2
	all	109586	100.0

Table 10: Categorical Variable Descriptive Statistics for Grade 7 Cohort

Variable	**Min**	$\mathbf{q_1}$	$\widetilde{\mathbf{x}}$	$\bar{\mathbf{x}}$	$\mathbf{q_3}$	**Max**	$\mathbf{s}$
Math	330.000	511.000	538.000	536.635	564.000	710.000	42.932
Read	310.000	489.000	519.000	515.482	546.000	780.000	46.902
Attend.	0.000	94.100	96.900	95.419	98.600	100.000	5.483
Cohort Math	404.800	528.376	537.116	535.484	546.543	606.357	18.382
Cohort Read	384.333	508.090	517.716	514.505	524.906	581.200	17.437
Cohort Attend.	55.878	94.556	95.557	95.067	96.232	100.000	2.553
Cohort % Nonwhite	0.000	0.056	0.111	0.211	0.242	1.000	0.246
Cohort % ELL	0.000	0.006	0.027	0.053	0.067	0.912	0.071
Cohort % FRPL	0.000	0.165	0.275	0.318	0.418	1.000	0.212

Table 11: Continuous Variable Descriptive Statistics of Students in Grade 7 Cohort

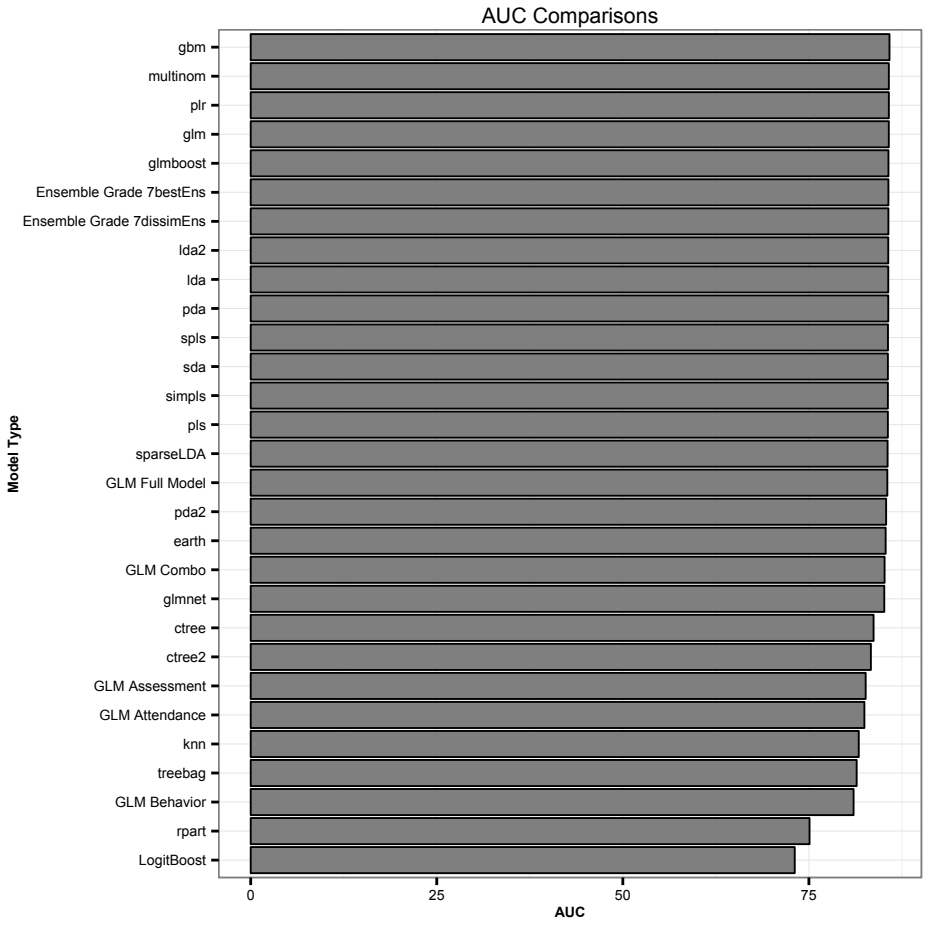

Figure 11: Comparing Test Data AUC for Grade 7 GLM and Final EWS Models from All Grades

Figure 11 depicts the AUC for all of the models successfully evaluated within DEWS for grade 7. It also includes, for comparison, the performance of the final GLM models in grade 7 and the final ensembled models. The performance on the test data is reported here.

These results are also shown in Table 12 for more precision. This table includes the time, in seconds, to fit for different methods on Wisconsin grade 7 cohort data. Model names are the `caret` method shorthand reference for R. The full algorithm descriptions can be found online at `http://caret.r-forge.r-project.org/`.

Model	Data	AUC	AUC std. err.	Time
Ensemble Grade 6dissimEns	test	0.91		
Ensemble Grade 8dissimEns	test	0.87		
Ensemble Grade 6bestEns	test	0.86		
Ensemble Grade 8bestEns	test	0.86		
gbm	test	0.86	0.01	345.28
multinom	test	0.86	0.01	125.64
plr	test	0.86	0.02	756.56
glm	test	0.86	0.01	108.19
glmboost	test	0.86	0.01	125.17
Ensemble Grade 7bestEns	test	0.86		
Ensemble Grade 7dissimEns	test	0.86		
lda2	test	0.86	0.01	99.52
lda	test	0.86	0.01	93.62
pda	test	0.86	0.01	105.28
spls	test	0.86	0.01	212.46
sda	test	0.86	0.01	105.17
pls	test	0.86	0.01	95.39
simpls	test	0.86	0.01	96.61
sparseLDA	test	0.86	0.01	261.89
GLM Full Model	test	0.86		
pda2	test	0.85	0.01	103.26
earth	test	0.85	0.02	266.46
GLM Combo	test	0.85		
glmnet	test	0.85	0.02	158.63
ctree	test	0.84	0.01	199.58
ctree2	test	0.83	0.02	191.29
GLM Assessment	test	0.83		
GLM Attendance	test	0.82		
knn	test	0.82	0.01	222.11
treebag	test	0.81	0.01	427.28
GLM Behavior	test	0.81		
Ensemble Grade 5dissimEns	test	0.80		
Ensemble Grade 5bestEns	test	0.77		
rpart	test	0.75	0.21	118.01
LogitBoost	test	0.73	0.14	113.67

Table 12: Example of AUC Results from Model Search

7.2. Computing Environment

The Wisconsin DEWS trades a lack of rich individualized information about students and timely data for a statewide comparison group and computing power. Searching through, ensembling, and predicting using dozens of statistical algorithms for hundreds of thousands of student records requires a robust and powerful computing environment. The Wisconsin DEWS runs on a Windows 2012 Server R2 computer powered by two Intel Xeon E5-2620 12 core processors running at 2.0GHz. The computer is equipped with 64 GB of RAM and has direct access to the data warehouse via ODBC connections. The system runs the 64-bit flavor of R and is accessed via Remote Desktop.

7.3. Open Source Tools

Most of the modules of the Wisconsin DEWS are available with example data and example code as R packages. Both the `EWStools` package and the `caretEnsemble` package serve as the analytical backbone for DEWS. These packages provide interfaces to a number of convenience functions to make leveraging the `caret` package even easier for model testing, selection, and ensembling. The code for these packages is available online through GitHub: `http://www.github.com/jknowles`. Additionally, readers seeking to explore these tools further are invited to install the tools in their own R environment:

```
install.packages(c("devtools", "caret"))
install_github("jknowles/EWStools")
install_github("zachmayer/caretEnsemble")
```

Alternatives to the R environment do exist. Two of the most promising are the `scikit-learn` module for the Python programming language and the Julia language for technical computing. Both provide the ability to implement a model building and prediction workflow similar to that described above.

7.4. A Reproducible Example

Below is an example of the search, ensemble, and test procedure making use of the `EWStools` package. This code leverages a simulated data set available in the `EWStools` package so that analysts can try out the model search and model test procedure on common data and share results, issues, and innovations without worrying about confidential data (Knowles, 2014).

```
library(EWStools)
library(caretEnsemble)
data(EWStestData)
data(caretMethods)
set.seed(2014)
# choose some available methods
mymethods <- c("glm", "knn", "pda", "ctree2",
               "rpart2", "gbm", "rf", "nnet")

ctrl <- trainControl(method = "cv", number = 5,
```

```
                      classProbs = TRUE,
                      savePredictions = TRUE,
                      summaryFunction = twoClassSummary)

results <- modSearch(methods = mymethods,
                      datatype = c("train", "test"),
                      traindata = modeldat$traindata,
                      testdata = modeldat$testdata,
                      metric = "ROC", fitControl = ctrl,
                      modelKeep = FALSE, length = 6)

out <- buildModels(mymethods, control = ctrl,
                    x = modeldat$traindata$preds,
                    y = modeldat$traindata$class,
                    tuneLength = 7)

out.ens <- caretEnsemble(out, optFUN = safeOptAUC, iter = 500L)
summary(out.ens)
```

```
The following models were ensembled: glm, knn, pda, ctree2, rpart2, gbm, nn
They were weighted:
0.706 0.038 0.002 0.008 0.002 0.084 0.16
The resulting AUC is: 0.9853
The fit for each individual model on the AUC is:
 method    metric    metricSD
    glm 0.9847379 0.008232874
    knn 0.9440451 0.015070482
    pda 0.9664728 0.012007210
 ctree2 0.9067024 0.026127229
 rpart2 0.8514493 0.036335499
    gbm 0.9762841 0.012794421
   nnet 0.9841620 0.027493360
```

7.5. R SESSION

The following R session was used to both author the paper and to build the EWS.

- R version 3.1.2 (2014-10-31), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=English_United States.1252`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Base packages: base, datasets, graphics, grDevices, grid, methods, parallel, splines, stats, stats4, tools, utils

- Other packages: apsrtable 0.8-8, caret 6.0-41, caretEnsemble 1.0.0, class 7.3-12, doParallel 1.0.8, e1071 1.6-4, eeptools 0.3.1, EWStools 0.9, foreach 1.4.2, gbm 2.1, ggplot2 1.0.0, iterators 1.0.7, knitr 1.9, lattice 0.20-30, MASS 7.3-39, mda 0.4-4, modeltools 0.2-21, mvtnorm 1.0-2, nnet 7.3-9, party 1.0-20, plyr 1.8.1, pROC 1.7.3, reporttools 1.1.1, rpart 4.1-9, sandwich 2.3-2, scales 0.2.4, stargazer 5.1, strucchange 1.5-0, survival 2.37-7, xtable 1.7-4, zoo 1.7-11

- Loaded via a namespace (and not attached): abind 1.4-0, arm 1.7-07, bitops 1.0-6, BradleyTerry2 1.0-6, brglm 0.5-9, car 2.0-24, caTools 1.17.1, chron 2.3-45, coda 0.16-1, codetools 0.2-10, coin 1.0-24, colorspace 1.2-4, compiler 3.0.2, data.table 1.9.4, digest 0.6.8, evaluate 0.5.5, foreign 0.8-63, formatR 1.0, gridExtra 0.9.1, gtable 0.1.2, gtools 3.4.1, highr 0.4, labeling 0.3, lme4 1.1-7, maptools 0.8-34, Matrix 1.1-5, memisc 0.96-10, mgcv 1.8-4, minqa 1.2.4, munsell 0.4.2, nlme 3.1-120, nloptr 1.0.4, pbapply 1.1-1, pbkrtest 0.4-2, proto 0.3-10, quantreg 5.11, Rcpp 0.11.4, reshape2 1.4.1, sp 1.0-17, SparseM 1.6, stringr 0.6.2

8. ACKNOWLEDGEMENTS

The author would like to acknowledge his colleagues at the Wisconsin Department of Public Instruction, without whom this work would not have been possible. In particular, he thanks Doug White, Kurt Kiefer, Mike Bormett, and the DEWS task force members for their leadership and support in moving DEWS from an idea to a statewide system. Additional thanks go to attendees at the 2012, 2013, and 2014 STATS-DC conferences and to colleagues in other states and school districts for their encouragement, suggestions and enthusiasm. This paper benefited from the generous comments of two anonymous reviewers, and also Hannah Miller who read multiple drafts and helped clarify the language. Any errors remaining are the sole responsibility of the author.

REFERENCES

AGUIAR, E., LAKKARAJU, H., BHANPURI, N., MILLER, D., YUHAS, B., AND ADDISON, K. L. 2015. Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In *Proceedings of the 2015 Learning Analytics and Knowledge Conference*.

ALLENSWORTH, E. 2013. The use of ninth-grade early warning indicators to improve chicago schools. *Journal of Education for Students Placed at Risk 1*, 68–83.

BALFANZ, R. 2009. Putting middle grades students on the graduation path: A policy and practice brief. Tech. rep., National Middle School Association, Westerville, Ohio. `http://www.amle.org/portals/0/pdf/research/research_from_the_field/policy_brief_balfanz.pdf`.

BALFANZ, R. AND HERZOG, L. 2006. Keeping middle grades students on-track to graduation: Initial analysis and implications. `http://web.jhu.edu/CSOS/graduation-gap/edweek/Balfanz_Herzog.ppt`.

BALFANZ, R. AND IVER, D. M. 2006. Closing the mathematics achievement gap in high poverty middle schools: Enablers and constraints. *Journal of Education for Students Placed At Risk 11,* 2, 143–159.

BALFANZ, R. AND IVER, D. M. 2007. Preventing student disengagement and keeping students on the graduation path in the urban middle grade schools: Early identification and effective interventions. *Educational Pyschologist 42,* 4, 223–235.

BALFANZ, R. AND LEGTERS, N. 2004. Locating the dropout crisis: Which high schools produce the nation's dropouts? Tech. Rep. 70, Center for Research on the Education of Students Placed At Risk, Baltimore, MD. `http://files.eric.ed.gov/fulltext/ED484525.pdf`.

BOWERS, A. J. AND SPROTT, R. 2012a. Examining the multiple trajectories associated with dropping out of high school: A growth mixture model analysis. *The Journal of Educational Research 105*, 176–195.

BOWERS, A. J. AND SPROTT, R. 2012b. Why tenth graders fail to finish high school: A dropout typology latent class analysis. *Journal of Education for Students Placed at Risk 17*, 129–148.

BOWERS, A. J., SPROTT, R., AND TAFF, S. A. 2013. 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*, 77–100.

BREIMAN, L. 2001a. Random forests. *Machine Learning 45,* 1, 5–32.

BREIMAN, L. 2001b. Statistical modeling: The two cultures. *Statistical Science 16*, 199–231.

BURNHAM, K. AND ANDERSON, D. 2002. *Model Selection and Multi Model Inference: A Practical Information-Theoretic Approach*, Second ed. Springer, New York. ISBN 0-387-95364-7.

CARL, B., RICHARDSON, J. T., CHENG, E., KIM, H., AND MEYER, R. H. 2013. Theory and application of early warning systems for high school and beyond. *Journal of Education for Students Placed at Risk 1*, 29–49.

CHAPELLE, O., VAPNIK, V., BOUSQUET, O., AND MUKHERJEE, S. 2002. Choosing multiple parameters for support vector machines. *Machine Learning 46,* 1-3, 131–159.

CHATFIELD, C. 1995. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society 158*, 419–466.

DAHL, D. B. 2013. *xtable: Export tables to LaTeX or HTML*. R package version 1.7-1.

DASU, T. AND JOHNSON, T. 2003. *Exploratory Data Mining and Data Cleaning*. Wiley-Interscience, New York.

DAVIS, M., HERZOG, L., AND LEGTERS, N. 2013. Organizing schools to address early warning indicators (ewis): Common practices and challenges. *Journal of Education for Students Placed at Risk 1*, 84–100.

DOWLE, M., SHORT, T., AND LIANOGLOU, S. 2013. *data.table: Extension of data.frame for fast indexing, fast ordered joins,fast assignment, fast grouping and list columns.* R package version 1.8.8.

EASTON, J. AND ALLENSWORTH, E. 2005. The on-track indicator as a predictor of high school graduation. Tech. rep., Consortium on Chicago School Research, Chicago. http://ccsr.uchicago.edu/publications/track-indicator-predictor-high-school-graduation.

EASTON, J. AND ALLENSWORTH, E. 2007. 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. Tech. rep., Consortium on Chicago School Research, Chicago. http://ccsr.uchicago.edu/publications/what-matters-staying-track-and-graduating-chicago-public-schools.

EFROYMSON, M. 1960. Multiple regression analysis. In *Mathematical Methods for Digital Computers*, A. Ralston and H. Wilf, Eds. Wiley, New York.

EVERS, A. 2012. Agenda 2017: Every child a graduate college and career ready. http://dpi.wi.gov/sprntdnt/pdf/agenda2017.pdf.

FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. 2000. Additive logistic regression: A statistical view of boosting. *Annals of Statistics 28,* 2, 337–374.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A., AND RUBIN, D. B. 2013. *Bayesian Data Analysis*, Third ed. Chapman & Hall / CRC Texts in Statistical Science, London.

GELMAN, A. AND HILL, J. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

GLEASON, P. AND DYNARSKI, M. 2002. Do we know whom to serve? issues in using risk factors to identify dropouts. *Journal of Education for Students Placed At Risk 7*, 25–41. http://www.mathematica-mpr.com/publications/PDFs/dod-risk.pdf.

GRÖMPING, U. 2006. Relative importance for linear regression in r: The package relaimpo. *The Journal of Statistical Software 17*.

HANCZAR, B., HUA, J., SIMA, C., WEINSTEIN, J., BITTNER, M., AND DOUGHERTY, E. 2010. Small-sample precision of roc-related estimates. *Bioinformatics 26*, 822–830.

HAND, D. J. 2009. Measuring classifier performance: A coherent alternative to the area under the roc curve. *Machine Learning 77*, 103–123.

HANLEY, J. AND MCNEIL, B. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology 143*, 29–36.

HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 ed. Springer New York, New York. http://books.google.com/books/about/The_Elements_of_Statistical_Learning.html?id=tVIjmNS3Ob8C.

HEPPEN, J. B. AND THERRIAULT, S. B. 2008. Developing early warning systems to identify potential high school dropouts. Tech. rep., National High School Center, Washington D.C. http://www.betterhighschools.org/pubs/documents/IssueBrief_EarlyWarningSystemsGuide.pdf.

HLAVAC, M. 2013. *stargazer: LaTeX code for well-formatted regression and summary statistics tables*. Harvard University, Cambridge, USA. R package version 3.0.1.

HONAKER, J., KING, G., AND BLACKWELL, M. 2011. Amelia II: A program for missing data. *Journal of Statistical Software 45,* 7, 1–47.

INMON, W. 2005. *Building the Data Warehouse*, Fourth ed. John Wiley and Sons, New York. ISBN 978-1265-0645-3.

JAMES, G., WITTEN, D., HASTIE, T., AND TIBSHIRANI, R. 2013. *An Introduction to Statistical Learning*, 1 ed. Springer New York, New York.

JANOSZ, M., ARCHAMBAULT, I., MORIZOT, J., AND PAGANI, L. S. 2008. School engagement trajectories and their differential predictive relations to dropout. *Journal of Social Issues 64,* 1, 21–40.

JERALD, C. D. 2006. Identifying potential dropouts: Key lessons for building an early warning data system. Tech. rep., Achieve, Inc., Washington D.C. `http://www.jff.org/sites/default/files/IdentifyingPotentialDropouts.pdf`.

KEMPLE, J. J., SEGERITZ, M. D., AND STEPHENSON, N. 2013. Building on-track indicators for high school graduation and college readiness: Evidence from new york city. *Journal of Education for Students Placed at Risk 1*, 7–28.

KENNELLY, L. AND MONRAD, M. 2007. Approaches to dropout prevention: Heeding early warning signs with appropriate interventions. Tech. rep., National High School Center, Washington D.C. `http://www.betterhighschools.org/docs/nhsc_approachestodropoutprevention.pdf`.

KIMBALL, R. AND ROSS, M. 2002. *The Data Warehouse Toolkit*, Second ed. John Wiley and Sons, New York. ISBN 0-471-20024-7.

KNOWLES, J. AND WHITE, D. 2013. The wisconsin dropout early warning system action guide. Tech. rep., Wisconsin Department of Public Instruction, Madison, WI. `http://wise.dpi.wi.gov/files/wise/pdf/wi-dews-actionguide.pdf`.

KNOWLES, J. E. 2014. *EWStools: Tools for automating the testing and evaluation of education early warning system models*. R package version 0.1.

KUHN, M. AND JOHNSON, K. 2013. *Applied Predictive Modeling*, First ed. Springer, New York. ISBN 978-1-4614-6848-6.

KUHN, M., WESTON, S., AND CODE FOR C5.0 BY R. QUINLAN, N. C. C. 2013. *C50: C5.0 Decision Trees and Rule-Based Models*. R package version 0.1.0-15.

KUHN, M., WING, J., WESTON, S., WILLIAMS, A., KEEFER, C., ENGELHARDT, A., AND COOPER, T. 2013. *caret: Classification and Regression Training*. R package version 5.15-61.

KUNCHEVA, L. AND WHITAKER, C. 2003. Measures of diversity in classifier ensembles. *Machine Learning 51*, 181–207.

LOBO, J. M., JIMNEZ-VALVERDE, A., AND REAL, R. 2008. Auc: A misleading mesure of the performance of predictive distribution models. *Global Ecology and Biogeography 17*, 145–151.

MAYER, Z. AND KNOWLES, J. 2014. *caretEnsemble: Framework for combining caret models into ensembles*. R package version 1.0.

MUTHÈN, B. 2004. Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In *The SAGE Handbook of Quantitative Methodology for the Social Sciences*, D. Kaplan, Ed. Sage, Thousand Oaks, CA, 345–370.

NEILD, R. C., STONER-EBY, S., AND FURSTENBERG, F. 2008. Connecting entrance and departure: the transition to ninth grade and high school dropout. *Education and Urban Society 50*, 543–569.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

R CORE TEAM. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RIPLEY, B. AND LAPSLEY, M. 2012. *RODBC: ODBC Database Access*. R package version 1.3-6.

ROBIN, X., TURCK, N., HAINARD, A., TIBERTI, N., LISACEK, F., SANCHEZ, J.-C., AND MLLER, M. 2011. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics 12*, 77.

RODERICK, M. 1993. *The path to dropping out: Evidence for Intervention*. Auburn House, Westport, CT.

RODERICK, M. AND CAMBURN, E. 1999. Risk and recovery from course failure in the early years of high school. *American Educational Research Journal 36*, 303–344. http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=EJ600524&ERICExtSearch_SearchType_0=no&accno=EJ600524.

RUMBERGER, R. W. 1995. Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal 32*, 583–625. http://www.education.ucsb.edu/rumberger/internet%20pages/Papers/Rumberger--Droputs%20from%20middle%20school%20(AERJ%201995).pdf.

SCULLEY, D., HOLT, G., GOLOVIN, D., DAVYDOV, E., PHILLIPS, T., EBNER, D., CHAUDHARY, V., AND YOUNG, M. 2014. Machine learning: The high interest credit card of technical debt. In *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*.

SINCLAIR, M., CHRISTENSON, S., AND THURLOW, M. 2005. Promoting school completion of urban secondary youth with emotional or behavioral disabilities. *Exceptional Children 71*, 465–482. http://www.iod.unh.edu/APEX%20Trainings/Tier%202%20Manual/Additional%20Reading/3.%20Check%20and%20Connect.pdf.

SOLLICH, P. AND KROGH, A. 1996. Learning with ensembles: how overfitting can be useful. *Advances in Neural Information Processing Systems 8*, 190–196.

SWETS, J. 1988. Measuring the accuracy of diagnostic systems. *Science 240*, 1285–1293.

THE STRATEGIC DATA PROJECT. 2012. The strategic data project toolkit version 1.1. http://www.gse.harvard.edu/~pfpie/index.php/sdp/tools.

US DEPARTMENT OF EDUCATION. 2012a. Slds data use issue brief iii: Turning administrative data into research-ready longitudinal datasets. Tech. rep., US Department of Education, Washington, D.C. http://nces.ed.gov/programs/slds/pdf/Data-Use-Issue-Brief-3_Research-Ready-Datasets.pdf.

US DEPARTMENT OF EDUCATION. 2012b. Slds data use issue brief iv: Techniques for analyzing longitudinal administrative data. Tech. rep., US Department of Education, Washington, D.C. http://nces.ed.gov/programs/slds/pdf/Data-Use-Issue-Brief-4_Analysis-Techniques.pdf.

VAPNIK, V. 1998. *Statistical Learning Theory*. John Wiley and Sons, New York NY.

VENABLES, W. N. AND RIPLEY, B. D. 2002. *Modern Applied Statistics with S*, Fourth ed. Springer, New York. ISBN 0-387-95457-0.

VIVO, J. AND FRANCO, M. 2008. How does one assess the accuracy of academic success predictors? roc analysis applied to university entrance factors. *International Journal of Mathematical Education in Science and Technology 39*, 325–340.

WICKHAM, H. 2009. *ggplot2: elegant graphics for data analysis*. Springer New York. `http://had.co.nz/ggplot2/book`.

WICKHAM, H. 2014. Tidy data. *The Journal of Statistical Software 59*.

XIE, Y. 2013. *knitr: A general-purpose package for dynamic report generation in R*. R package version 1.1.

YOUDEN, W. 1950. Index for rating diagnostic tests. *Cancer 3*, 32–35.

ZWIEG, M. AND CAMPBELL, G. 1993. Receiver-operating characteristic (roc) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry 39*, 561–577.