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

Jared E. Knowles *
Wisconsin Department of Public Instruction
January 28, 2014

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Introduction</td>
<td>2</td>
</tr>
<tr>
<td>2 Literature Review</td>
<td>3</td>
</tr>
<tr>
<td>2.1 Checklist</td>
<td>3</td>
</tr>
<tr>
<td>2.2 Regression Techniques</td>
<td>3</td>
</tr>
<tr>
<td>2.3 Growth Mixture Modeling</td>
<td>4</td>
</tr>
<tr>
<td>2.4 Other Extensions</td>
<td>5</td>
</tr>
<tr>
<td>3 Data</td>
<td>8</td>
</tr>
<tr>
<td>3.1 Exploring Wisconsin’s Data</td>
<td>10</td>
</tr>
<tr>
<td>4 Methods</td>
<td>13</td>
</tr>
<tr>
<td>4.1 Applied Modeling</td>
<td>13</td>
</tr>
<tr>
<td>4.2 From Theory to Model</td>
<td>14</td>
</tr>
<tr>
<td>4.3 Variable Selection</td>
<td>15</td>
</tr>
<tr>
<td>4.4 Preparing the Data</td>
<td>16</td>
</tr>
<tr>
<td>4.4.1 Recoding Categorical Data</td>
<td>16</td>
</tr>
<tr>
<td>4.4.2 Scaling and Centering Continuous Data</td>
<td>17</td>
</tr>
<tr>
<td>4.4.3 Missing Data</td>
<td>17</td>
</tr>
<tr>
<td>4.5 Evaluating Model Fit to Select Functional Form</td>
<td>18</td>
</tr>
<tr>
<td>4.6 Testing Procedure</td>
<td>23</td>
</tr>
<tr>
<td>5 Examples and Results</td>
<td>24</td>
</tr>
<tr>
<td>5.1 Choosing the Best Model</td>
<td>25</td>
</tr>
<tr>
<td>6 Conclusions and Future Extensions</td>
<td>28</td>
</tr>
<tr>
<td>7 References</td>
<td>29</td>
</tr>
<tr>
<td>8 Appendix</td>
<td>32</td>
</tr>
</tbody>
</table>

---

*Jared Knowles is a research analyst with the Wisconsin Department of Public Instruction and PhD Candidate in the University of Wisconsin - Madison Political Science department. The views expressed in the article are his own.
1 Introduction

The Wisconsin Department of Public Instruction (DPI) is committed to providing school districts with 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 tens of millions of dollars in a statewide longitudinal data system, the Wisconsin Information System for Education (WISE). This data system provides longitudinal measures of all students in the Wisconsin public school system across a variety of measures. Until now, these data have been collected with the purposes of creating valuable public reports and for school and district accountability. Now DPI is turning to provide educators and administrators analyses of their data that can help inform their work and promote best practices.

In particular, WISE provides is a predictive indicator system that offers educators a forward-looking view of student performance. Any such system must be accurate, reproducible, stable, and transparent. The discussion below reviews the prior work on EWS’s, discusses the case of Wisconsin, and then describes the methods used to ensure the development of an accurate and stable EWS in a statewide framework. Not discussed here are the systems and communications challenges associated with increasing usage and informing stakeholders about the rollout of the system. For more on this see the Wisconsin Dropout Early Warning System (DEWS) website which contains guides for practitioners to use when reviewing EWS data and a how-to for getting access to the DEWS reports[1]

2 Literature Review

The literature on early warning indicators (EWIs) of high school dropout is full of ideas about specific indicators, but lacks coherence. There is no standard reporting on how to assess the accuracy, utility, or the tradeoffs associated with using a particular EWI in making decisions about interventions with individual students (Bowers and Sprott 2012a).

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 were 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 Chicago Consortium on School Research (CCSR) has since been used across the nation to help high schools identify 9th grade students who are struggling and at risk of exiting early (Heppen and Therriault 2008; Kennelly and Monrad 2007). These systems have shown high predictive power historically been focusing on the first year of high school (Easton and Allensworth 2005; Roderick and Camburn 1999; Allensworth 2013). The first year of high school is a critical time for students, and this focus makes much practical sense—indeed such systems are highly accurate with the system used in Chicago being found to identify upwards of 70% non-graduates. This system is remarkable not only for its high accuracy, but also for the 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 and disassociation with school with roots 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 K-12 schools after the 8th 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 drop out before reaching 12th grade, and roughly 2,000 of these dropout in 9th and 10th grade. Identifying these students in the middle of the second semester of their 9th 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); Balfanz and Herzog (2006) have shown that course failures, low attendance, or poor marks for behavior in middle school can substantially reduce student likelihood to graduate or to catch 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 10% to 20% rate (Balfanz and Iver 2007).
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 we might improve their accuracy. Recent work in Milwaukee Public Schools (MPS) has shown that students may often continue to accumulate credits without course failures, but still dropout, or graduate with a very minimal skillset (Carl et al., 2013). Carl et al. (2013) use several regression models estimating a student’s probability of high school completion built on the concept of Total Quality Credits (TQC) in the freshman year. The authors’ TQC measure is a combination of course grades in the four core subjects and is shown to be 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 predictors as well, suggesting that such a system could be extended into earlier grades. Indeed, the authors envision a continuous early warning system from grade 6 to grade 12 is which provide educators with semi-annual assessments of the progress of their students (Carl et al., 2013).

<table>
<thead>
<tr>
<th>Model</th>
<th>Indicators</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCSR</td>
<td>Course failures</td>
<td>checklist</td>
</tr>
<tr>
<td>Balfanz</td>
<td>Attendance, behavior, course failures</td>
<td>checklist</td>
</tr>
<tr>
<td>Carl et al.</td>
<td>TQC</td>
<td>regression</td>
</tr>
</tbody>
</table>

All of these systems 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 all together than their urban peers. Indeed, research has shown that the context, such as the school level of poverty and the school building itself can matter greatly (Balfanz, 2009; Balfanz and Iver, 2006).

2.3 Growth Mixture Modeling

The regression technique adopted by Carl et al. (2013) provides flexibility for applying separate thresholds in separate contexts. However, these techniques are limited in their accuracy by one strong assumption – that all dropouts are the same in experiencing risk factors. To address this, a promising technique that has shown itself to be highly accurate involves estimating growth trajectories of latent classes of students to identify dropouts with different academic trajectories through high school (Bowers and Sprott, 2012a; Muthén, 2004). This work, first demonstrated by Muthén (2004) in a tutorial on growth mixture modeling (GMM), allows for a more accurate classification of student types using data such as a repeated academic outcome - math achievement - and student demographic and engagement indicators. Bowers and Sprott (2012a) continues where Muthén (2004) leaves off by extending this tutorial of the GMM methodology to the ELS:2002 dataset and tying the model to existing theories of student disengagement and dropout. The GMM methodology has also been used in relation to student engagement indicators, as measured by a student survey instrument, and shown predictive power (Janosz et al., 2008). Bowers and Sprott (2012a) choose to focus on data available to most school staff, and argue that this more effective.

Modeling the process of dropout has created a more nuanced picture, particularly of dropout among high schoolers (Bowers and Sprott, 2012a; B). Most importantly, Bowers and Sprott (2012a) has re-emphasized that students may drop out for reasons other than low performance in coursework or on standardized
assessments, and in some cases higher performance may also be a warning sign.

However, even the [Bowers and Sprott (2012a)] study makes some decisions that show the difficulty of GMMs for use in an operational EWS used on an ongoing basis. First, while GMMs are good at modeling behavior in a sample of students, there is no sense of how the model performs at making predictions for students outside of the sample. Second, GMMs in contrast to checklist systems, have fairly 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. It also is problematic even in a statewide data system where students along the state borders may transfer in and out.

2.4 Other Extensions

Beyond the regression technique recent developments have included assessing a variety of machine learning techniques for increasing the accuracy of prediction [Bowers et al. (2013)]. These machine learning techniques show much promise, especially in the unique case faced by state education agencies developing EWS systems - an abundance of data and predictors and an uneven structure to the data on students nested within schools and districts. The results focused nature of machine learning also allows for possible increases in accuracy by extending modeling choices beyond linear regression. However, such techniques increase the complexity of the model building process and reduce the transparency and explainability of the results to stakeholders - the black box problem.

As [Bowers et al. (2013)] finds, most of the existing literature on dropout flags suffers from failing to properly specify the accuracy of the predictions. The authors suggest that instead of simply reporting the percentage of dropouts identified by a flag, scholars should turn to reporting a broad 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 2: Classification Metrics for Binary Classifiers

<table>
<thead>
<tr>
<th>Event</th>
<th>dropout</th>
<th>graduate</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>predicted</td>
<td>True Positive</td>
<td>False Negative</td>
<td>a+b</td>
</tr>
<tr>
<td>dropout</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>graduate</td>
<td>False Positive</td>
<td>True Negative</td>
<td>c+d</td>
</tr>
<tr>
<td>a+c</td>
<td>b+d</td>
<td>total</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 depicts a confusion matrix, a display of prediction that is traditionally used to show the tradeoffs associated with developing predictive indicators of binary outcomes (adapted from [Bowers and Sprott 2012a]). Table 3 describes the most important measures of accuracy, derived from the confusion matrix in Table 2.

In developing an EWS system, many competing factors must be balanced. The accuracy of various at-risk indicators has been found to vary widely and the appropriate measure of accuracy has been inconsistently
Table 3: Measures of Accuracy for At-Risk Indicators

<table>
<thead>
<tr>
<th>Measure</th>
<th>Calculation</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision (positive predictive value)</td>
<td>( \frac{a}{a+b} )</td>
<td>Percent of predicted dropouts that dropout</td>
</tr>
<tr>
<td>Sensitivity (recall)</td>
<td>( \frac{a}{a+c} )</td>
<td>Percent of actual dropouts predicted correctly</td>
</tr>
<tr>
<td>Specificity</td>
<td>( \frac{d}{b+d} )</td>
<td>Percent of actual graduates predicted correctly</td>
</tr>
<tr>
<td>False-alarm</td>
<td>( \frac{b}{b+d} )</td>
<td>1 - Specificity (False Alarm Rate)</td>
</tr>
</tbody>
</table>

applied (Gleason and Dynarski 2002; Jerald 2006; Bowers et al. 2013). As Bowers et al. (2013) indicates - most of the 110 at-risk flags found in the literature only include a measure of the sensitivity, or the specificity, but rarely both and even more rarely the false-alarm rate. This is problematic, because this thorough view of accuracy is needed to avoid costly interventions applied to the wrong subset of students (Jerald 2006; Gleason and Dynarski 2002). There is also a need for clarity and understanding – the determinations made by the system must be easily processed by educators and administrators into decision making. Most importantly there must be trust between those providing the source data, those doing the calculation of the determination, and those taking action on the reports.

In a binary classification framework, we are concerned with identifying as many students correctly as possible while reducing the false-alarm rate as much as possible. We have previously discussed that the failure to do this is a strong critique of many existing EWS indicators and flags used in the field (Bowers et al. 2013; Gleason and Dynarski 2002). For example, a system in Montgomery County Maryland identifies over 75% of high school dropouts in first grade, but has the consequence of identifying nearly half of all students in the first grade as being at risk of dropping out. Figure 1 recreates a plot from Bowers et al. (2013) depicting the accuracy of 110 separate early warning indicators found in the literature.

The metric chosen for determining the best model is thus a pivotal design decision. In this case, following the guidance of Bowers et al. (2013), we employ the Receiver Operating Characteristic or ROC (Hastie et al. 2009) depicted in Figure 1. This plot shows the explicit tradeoff between false alarms and true positive classifications. In other words, what percentage of dropouts can we successfully identify (along the y axis) in return for falsely identifying eventual graduates as dropouts (along the x axis). ROC also allows comparisons of binary-classification models to be made 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 (Bowers et al. 2013).

For example, the growth mixture models discussed above, represented as boxes in Figure 1, are highly predictive with a high proportion of true positives and a relatively low proportion of false alarms. The Chicago On-Track indicator is also highly predictive as well, indicated by the crosses. Finally, the Balfanz ABC indicators are somewhat less informative, but still capture a relatively large proportion of eventual dropouts – and are again being applied in middle grades as opposed to high school like the GMM and Chicago On-Track indicators.

It is clear that ROC efficiently describes the performance of binary classifiers as the tradeoff between false-alarm and true positive detection. Kuhn et al. (2013) demonstrate how use of the ROC on classification models can be intuitive and easy to do. For policy making, the ROC is extremely valuable because it depicts the tradeoff between the rate of classification of true positives and false positives as was depicted in Table 2.

In this paper we discuss the decisions made in developing the Wisconsin DEWS and how they address some of the challenges outlined above, as well as the limitations of this approach.

---

2 http://educationbythenumbers.org/content/can-an-algorithm-id-high-school-drop-outs-in-first-grade_386/
Figure 1: ROC points for 110 published early warning indicators from Bowers & Sprott (2010). The diagonal line represents random chance, the top left corner represents perfect prediction.
3 Data

Wisconsin ranks near the top in the national ranking of graduation rates. The state’s most recent four year high school completion rate stood at 87%. 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).

Despite this success, too many students in Wisconsin fail to graduate or graduate late - in 2010-11 over 9,000 students failed to graduate or graduated late. Importantly, deep disparities in graduation rates exist along economic and racial lines. Table 4 represents the most recent publicly reported figures for the four year graduation rates in Wisconsin and Table 5 shows the graduation gap for students who are eligible for free and reduced price lunch (FRL) and those who are not.

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

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

*All rates are 4 year graduation rates

Table 5: Economic Graduation Disparities in Wisconsin for 2010-11

<table>
<thead>
<tr>
<th>Group</th>
<th>Expected</th>
<th>Grads</th>
<th>Rate</th>
<th>Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non FRL</td>
<td>50,834</td>
<td>46,715</td>
<td>91.9%</td>
<td>-</td>
</tr>
<tr>
<td>FRL</td>
<td>19,542</td>
<td>14,481</td>
<td>74.1%</td>
<td>17.8</td>
</tr>
</tbody>
</table>

*All rates are 4 year graduation rates

These disparities are quite stark. In order to close these attainment gaps, DPI is focusing on strategies aimed at addressing the needs of these disadvantaged groups of students (Evers, 2012). Increasing high school graduation rates provides a clear shared purpose, but requires a diverse set of strategies to achieve. 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.

Focusing on on-time high school graduation has several advantages. First, student dropout has a number of different meanings, and students can dropout multiple times sometimes graduating on time after doing so. In contrast, four-year graduation is an unambiguous measure of student attainment (see Bowers and Sprott, 2012a, for an example). Second, students who graduate late often dropout, and students who dropout – when they graduate – often graduate late. Thus these two populations are very similar. Third, from the perspective of a school system, late graduates are more expensive and require more intensive educational


http://data.dpi.state.wi.us/data/HSCompletionPage.aspx?GraphFile=HIGHSCHOOLCOMPLETION&County=47&AthleticConf=4#&SCCESA=08&GeoLevel=st&Quad=performance.aspx#rFrm=4

The race categories for two or more races and Native Hawaiian Pacific Islander are not shown - combined this accounts for less than 1,000 students in the cohort for this school year. All data from the Wisconsin Information Network for Successful Schools (WINSS) online reporting system: http://winss.dpi.wi.gov/
Finally, from a statistical perspective, the number of late-graduates is much larger than the number of dropouts in Wisconsin, making the identification of late-graduates in middle grades more tractable with a statistical model. Confirmed dropouts make up only around 3,000 members of any given cohort out of 60,000+ students, making a statistical prediction more challenging.

Figure 2 shows that late graduates differ from on-time graduates, on standardized math scores in 8th grade. It also shows that transfer students, who comprise a mix of students who will graduate late, on-time, or not at all, are right between these two groups. This provides further evidence that focusing on predicting late graduation instead of strictly dropout will produce results that have practical value to educators and can be accurate and reliable.

This also 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. Additionally, there are continuing costs associated with providing educational services beyond the student’s expected year of exit.

To build the predictive model, we leverage the Wisconsin Statewide Longitudinal Data System, now known as the Wisconsin Information System for Education (WISE). In response to federal reporting and accountability requirements, Wisconsin developed a statewide longitudinal data system (SLDS). Beginning

---

Note that the large exception to this is students with disabilities for whom the Individual Educational Plan (IEP) explicitly includes a five or six-year graduation as the desired outcome.
in the 2005-06 school year, data has been collected statewide on all students in a limited number of domains. This datasets covers roughly 870,000 students enrolled in public K-12 schools in the state each year and makes it possible to longitudinally analyze student records for students who stay enrolled in public K-12 schools.

Taking the ABCs (indicators of Attendance, Behavior, and Coursework) described by Balfanz (2009) as a starting point as well as other indicators found in the literature by Bowers et al. (2013), the DPI SLDS has the elements listed below:

<table>
<thead>
<tr>
<th>Domain</th>
<th>Elements</th>
<th>Years of Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendance</td>
<td>% Attendance; days missed</td>
<td>05-06 to present</td>
</tr>
<tr>
<td>Behavior</td>
<td>Days suspended; days expelled</td>
<td>06-07 to present</td>
</tr>
<tr>
<td>Mobility</td>
<td>Schools and districts attended</td>
<td>06-07 to present</td>
</tr>
<tr>
<td>WKCE Reading</td>
<td>Scale Score</td>
<td>05-06 to present</td>
</tr>
<tr>
<td>WKCE Math</td>
<td>Scale Score</td>
<td>05-06 to present</td>
</tr>
<tr>
<td>Demographics</td>
<td>Race; gender; ELL; SwD; FRL</td>
<td>05-06 to present</td>
</tr>
<tr>
<td>School</td>
<td>school of attendance</td>
<td>05-06 to present</td>
</tr>
</tbody>
</table>

The data in Table 6 has much in common with the data used in early warning systems discussed above in the literature review, but a few substantial differences exist. First, a downside of the statewide system in Wisconsin is that course completion data is not yet available statewide. However, the statewide coverage does provide some information not available in other systems – namely student mobility indicators. Highly mobile students are much more likely to graduate late than their peers, and as long as the students are mobile within Wisconsin public K-12 schools, they are included in this dataset along with an indicator of their level of mobility.

### 3.1 Exploring Wisconsin’s Data

EWS systems vary dramatically in their predictive performance and their utility (Bowers et al., 2013). The 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.

As prior research has predominantly focused on large urban areas like Baltimore (Balfanz, 2009), Chicago (Easton and Allensworth, 2005, 2007), and Milwaukee (Carl et al., 2013), before beginning building a model on dropouts in Wisconsin, it is important to look at the relationship between observable indicators and dropout rates to inform model building. A statewide system is different than these models built in large urban districts because of the substantial contextual variance that exists across clusters, or schools, within the data. Wisconsin has middle schools with grade configurations ranging from K-8 to 7-9, sizes ranging from a few dozen students to several hundred, and locations as diverse as urban Milwaukee, rural Ashland, 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 clustering has been highly publicized by focusing on high schools that are so called "dropout factories". It is clear, not surprisingly, 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-graduates statewide. Figure 3 demonstrate that a disproportionate amount of late graduates attend just a handful of schools in the state. 40% of all students who graduated late started out in only 48 middle schools (7.5%). This makes it clear that 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-generating process.

![Late Graduation by Middle School for 2006-07 Cohort Statewide](image)

Figure 3: Late Graduation by Middle School

For this paper we take the 2006-07 grade 7 cohort as an exemplar for illustrating the estimation problem. Descriptive statistics for the data available in the DPI SLDS are shown in Table 7 and Table 8 in the Appendix. These tables display descriptive statistics for the continuous and nominal variables in the dataset, respectively.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>q1</th>
<th>x̄</th>
<th>q3</th>
<th>Max</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>math7</td>
<td>330</td>
<td>513.000</td>
<td>539.000</td>
<td>536.932</td>
<td>710</td>
<td>42.685</td>
</tr>
<tr>
<td>read7</td>
<td>310</td>
<td>488.000</td>
<td>519.000</td>
<td>514.759</td>
<td>780</td>
<td>46.884</td>
</tr>
<tr>
<td>att_rate_g7</td>
<td>0</td>
<td>94.200</td>
<td>96.900</td>
<td>95.377</td>
<td>100</td>
<td>5.898</td>
</tr>
</tbody>
</table>

Table 7: Continuous Variable Descriptive Statistics of Students in Cohort

The 2006-07 cohort was chosen as 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. With large censuses there is little year to year variation in the pattern depicted in these tables and figures. The issues with year to year variation in cohorts will be discussed in the methods section.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Levels</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ontime_grad</td>
<td>0</td>
<td>6647</td>
<td>11.6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>50490</td>
<td>88.4</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>57137</td>
<td>100.0</td>
</tr>
<tr>
<td>race</td>
<td>A</td>
<td>1976</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>5431</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>3444</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>781</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>45505</td>
<td>79.6</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>57137</td>
<td>100.0</td>
</tr>
<tr>
<td>frpl2</td>
<td>N</td>
<td>39812</td>
<td>69.7</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>17325</td>
<td>30.3</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>57137</td>
<td>100.0</td>
</tr>
<tr>
<td>gender</td>
<td>F</td>
<td>27914</td>
<td>48.9</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>29223</td>
<td>51.1</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>57137</td>
<td>100.0</td>
</tr>
<tr>
<td>disab_code</td>
<td>A</td>
<td>416</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>CD</td>
<td>753</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>EBD</td>
<td>1228</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>111</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>LD</td>
<td>3727</td>
<td>6.5</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>49079</td>
<td>85.9</td>
</tr>
<tr>
<td></td>
<td>OHI</td>
<td>1364</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>OI</td>
<td>65</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>SL</td>
<td>339</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>TBI</td>
<td>27</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>28</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>57137</td>
<td>100.0</td>
</tr>
<tr>
<td>ell_comp_med2</td>
<td>hi</td>
<td>977</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>lo</td>
<td>408</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>med</td>
<td>1380</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>nat</td>
<td>54372</td>
<td>95.2</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>57137</td>
<td>100.0</td>
</tr>
<tr>
<td>mobility_sch_yr6</td>
<td>0</td>
<td>2697</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>53120</td>
<td>93.0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1238</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>71</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>11</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>57137</td>
<td>100.0</td>
</tr>
<tr>
<td>mobility_dist_yr8</td>
<td>0</td>
<td>780</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>55410</td>
<td>97.0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>877</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>66</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>57137</td>
<td>100.0</td>
</tr>
<tr>
<td>disflag_g7</td>
<td>Major</td>
<td>2171</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td>Minor</td>
<td>1883</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>52507</td>
<td>91.9</td>
</tr>
<tr>
<td></td>
<td>Severe</td>
<td>576</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>57137</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 8: Nominal Descriptive Statistics of Students in Cohort
Next we turn our attention to the method used to generate the best possible predictions for the data at hand.

4 Methods

In reading the previous literature on EWIs, the biggest missing component is the lack of focus on out-of-sample model checking. In this section we detail why out-of-sample model testing is critical to a successful EWS and discuss one approach for balancing the tradeoffs between model flexibility and model interpretability when developing an EWS to be used operationally at a large scale by decision makers in schools. First, we discuss the difference between applied modeling and inferential statistics. Second, we discuss the concerns of feature selection, or how to select variables for inclusion in the predictive model. Third, we depart from the literature and discuss in-depth the data preparation steps necessary to use administrative data in the development of predictive models. Finally, we discuss the out-of-sample testing procedure and model comparison across disparate model types.

4.1 Applied Modeling

Breiman (2001) describes two statistical modeling cultures that exist within statistics – the data modeling and the algorithmic modeling cultures. The latter, often described as applied statistical modeling, goes by different names – statistical learning, machine learning, predictive modeling, applied modeling – but it is defined by a goal of learning relationships between predictors and outputs and using those modeled relationships to predict outputs for new datasets (Hastie et al., 2009). This learning may occur over time in cases where new data is streaming into the system, or 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 datasets other than the dataset 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, and particularly predictive models, are defined by their ability to make accurate out-of-sample predictions on future data. In contrast, inferential models are focused on testing the impact of particular measured variables within a particular sample, and then generalizing to the appropriate population. Second, applied models are constrained not by a theory of the data generation process but by the goals of the application. A model may be designed to include or exclude certain factors in order to provide face validity to the users of the model. Inefficient allocation of resources when the models are applied to support decision making (Breiman, 2001). In the EWS Gleason and Dynarski (2002) note that many EWS indicators fail by providing too many false positives and too few true-negative classifications leading to 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.

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 by, there are significant limitations with only including measures of accuracy on the sample. In machine learning data is commonly divided into “training” and “test” datasets. The “training” dataset is used to build the model, while the model performance is evaluated by how well

---

8 Note that this is the case for supervised machine learning or applied modeling. Unsupervised machine learning only focuses on exploring relations between a set of variables without any regard to inputs or outcomes.
9 Though this often occurs in inferential models in an academic setting when researchers report using “the standard control variables” in their models.
it predicts the “test” data, which was 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 that holding out data is not a commonly used practice, but with the advent of large administrative record systems like the Wisconsin WISEdash the it is no longer cost prdivide data into “training” and “test” datasets is

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. Applied modeling seeks to minimize an error rate not in the training data, but an estimated error rate on future data. Thus in an applied model setting the analyst is seeking to estimate the error rate of the model on new data and to minimize this error rate, instead of the error rate on “training” data upon which the model is built (James et al., 2013).

This focus on reducing predictive error for future cases makes applied modeling particularly well suited to the case of developing a dropout early warning system. The primary goal of DEWS is to correctly classify new students each year as accurately as possible on new data each year. That is, any statistical model derived must perform well in predicting the likely on-time graduation of students whose data were not included in developing the model. The model ultimately will be used to predict the likelihood of current middle grade students completing high school many years later and, thus, it must be designed to have predictive power outside of the cohort of students on which it is fit.

In developing a statistical model to be used in an applied setting, a different set of considerations are taken into account than in an inferential setting. The model must be accurate, it must be stable, it must scale to new cases, and it must be presented transparently. These challenges are unique when taking a system from the academic to the applied setting and one of the reasons that the Chicago On-Track Indicator has proven so successful; despite it’s lower reported accuracy than some other metrics, it has proven stable, scalable, transparent, and easy to communicate (Kennelly and Monrad, 2007). These concerns must be balanced with the goal of accuracy in a setting where the statistical model is to be used to make decisions with varying stakes.

4.2 From Theory to Model

The process of searching for the most accurate model is discussed in detail. The final results of this project are not the ultimate model that results. Instead, it is the systematic process of searching for the model and evaluating among all possible models for the most accurate. This technique, sometimes referred to (perhaps pejoratively) 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 over 100 students otherwise at risk of not completing high school on time.

Model identification proceeds along two axes. For the binomial process described below, analysts must focus on identifying both variables \( X \) and the function \( f \) such that:

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

Second the analyst must find the functional form, \( f(X) \), such that:

\[
Y^* = f(X) + \epsilon
\]

For example, in a binomial regression model, \( f(X) \) takes the form:
In an inferential framework 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 and transactional data that is both broad and rich, as well as a need for increased accuracy, an analyst can instead focus on identifying the subset of \( X \) that best predicts \( y \) given the function \( f(X) \). This is the situation in which many education agencies find themselves – with a need to predict who will complete high school, a wealth of data, and a strong preference for high accuracy to efficiently allocate scarce intervention resources. In the following sections both the methods used to identify the appropriate subset \( X \) from the space of available data elements, as well as methods for testing and comparing between various values of \( f(X) \).

4.3 Variable Selection

Unlike in the inferential modeling case, we do not presuppose the exact mechanisms through which student behaviors and attributes translate into high school success or not (Breiman, 2001). Instead, we start with a goal of maximizing the accuracy of our predictions of student success based on available attributes at an earlier period in time. Thus, variables may be included in an EWS for many reasons, including:

- They improve test data prediction accuracy
- They provide face validity to the model
- They control for important confounders
- They have been used previously in the 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 many cases feature selection can be quite difficult such as cases with many uninformative predictors and a few informative predictors, or cases where there are many subsets of highly collinear predictors. Stepwise regression is the canonical and controversial example, but many other methods exist including recursive feature elimination (Efroymson, 1960)\(^{10}\).

In our case we take the data available in the Wisconsin SLDS, listed in Table 6 as our starting point for the set of \( N \) possible predictors. From here, we consult with the previous research on strong predictors of student dropout listed in Table 1 and develop some informed subsets of predictors to investigate\(^{11}\).

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 in the literature that either used a nationally representative sample or operated in only a single school district, we are attempting to construct a model that accurately reflects dropout risk across the diverse school

---

\(^{10}\)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 applicable.

\(^{11}\)For analyses where many more predictors are available, a more robust and automated technique may be necessary. An alternative approach in the current analysis would be to create a large feature set from the variables listed in Table 6 by including multiple years or multiple transformations of the same variable for each case. The result would be a set of variables with high correlations among them instead of multiple orthogonal groups of measures, which may prove to be a better approach in some applications, but not one taken up here.
environments within the state of Wisconsin. Thus, aggregating student performance, demographic, and school system information and including these variables in the dataset is an important step as we have strong prior information that such community level variation is an important aspect of the phenomenon we are modeling.

Taking this data as the starting point we depart from common practice and discuss the often overlooked but important process of preparing or "tidying" the data in order to make it suitable for building a predictive model (Wickham 2016). In many cases, the choices made about cleaning, transforming, and rescaling variables may prove to be as consequential to the ultimate accuracy of the system than the exact subset of X or the functional form chosen (Dasu and Johnson 2003).

4.4 Preparing the Data

Moving from administrative data to data suitable for analysis - as most EWS’s will need to do - is an exercise that has received only limited attention in the literature, but is full of consequential decisions (US Department of Education 2012a,b). This section will focus on the classes of decisions that need to be made and the consequences those decisions can have. There are three major types of decisions that must be considered:

- Recoding data
- Scaling and centering data
- Handling missing data

Each of these issues is the subject of much academic debate with respect to statistical modeling and the following is not meant to be a full treatment of these debates. Rather, the treatment of these issues here serves to remind the reader that these decisions can have substantial consequences for the performance of the model. Therefore we discuss the practical implications of each group of transformations in the following sections.

4.4.1 Recoding Categorical Data

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

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 that there are not tradeoffs that result from preserving many categories for several variables. An instructive example is the case of how to code students with disabilities (SwD). In most cases solely using a binary indicator of SwD throws out too much information because of the substantial heterogeneity that exists within the SwD group in terms of assessment scores and graduation rates. At the same time, using the 13 individual SwD categories that correspond to the federal SwD codes presents some problems because some of these categories are very sparse. Since these categories have very few students in them, using them results in inefficient estimates of the coefficients and potentially 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 software used leads to various forms of inefficiency and bias in the resulting model when used to predict new data.  

One solution to these problems is to examine categorical variables with respect to our dependent variable. Categories which are very similar to one another in terms of the group mean and distribution of the dependent variable and which a small count 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 for recoding categorical variables is whether or not to treat them as ordered. For non-regression frameworks this distinction may not be as important depending on the estimation technique employed, but in the regression framework the ordered variables put linear or polynomial constraints on the estimation of their coefficients that may or may not be appropriate depending on the sample size and the observed relationship between levels of the variable and other student outcomes.

4.4.2 Scaling and Centering Continuous Data

For continuous measures such as attendance rates or assessment scores, a different set of concerns arise. In many cases, these data may exhibit non-normality or scores may be inconsistent from year to year. Rescaling and centering the data can help preserve the continuity of assessment score measures year to year and reduce some of the non-normality.

The only real argument against re-scaling and centering continuous data is that it makes interpretation of individual coefficients slightly more complicated. However, re-scaling and centering the data makes estimation methods like maximum-likelihood and Markov-Chain Monte Carlo sampling perform more efficiently, saving computation time and avoiding errors (Gelman and Hill [2006]). Re-scaling and centering can also help with the interpretation of parameters of interest by allowing binary indicator variables and continuous variables to be interpreted on a roughly similar scale (Gelman and Hill [2006]).

4.4.3 Missing Data

Data are often missing in administrative records, and the missingness is almost never at random. 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. Missing data can be introduced into an EWS in a number of ways. Three

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](http://www.ats.ucla.edu/stat/mult_pkg/faq/general/complete_separation_logit_models.htm)
ways are important to consider: biasing results by building models on students with complete longitudinal histories, the inability to generate predictions for current students with abbreviated longitudinal records, and including elements in the model that are not universally reported.

In an SLDS, or in a district longitudinal data system, the constraint that students must have multiple prior years of complete (with respect to EWS elements) data and an observed graduation subset outcome creates a distinct subset of students - students who maintained consistent enrollment for N years. This concern is treated very inconsistently in the existing literature with scholars focusing on national datasets that have been imputed or discarding cohort members who exited the district \cite{Bowers et al. 2013}. This concern is extremely important when seeking to develop a predictive model that is of use to practitioners and must be considered even in the case of a statewide data system where more mobility can be captured than at a single school district.

The extent to which this subset of students with complete data differ from those without complete data will vary. However, the constraint of requiring multiple years of data negatively affects the ability of an EWS to produce estimates for a group of students with known elevated risk – mobile students. 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 consecutive years. This means that missingness due to mobility results in two kinds of issues - a limited ability to provide predictions on new cases by requiring a student to have two years of data before a prediction can be made, and a potential bias in the model used to make predictions that results from the differences between students with a single year of data versus having multiple years of data.

In Wisconsin due to concerns about losing the ability to provide estimates for mobile students, we chose to adopt a single year cohort. Despite the evidence that longitudinal growth models, particularly growth mixture models, can be highly accurate \cite{Bowers et al. 2013, Bowers and Sprott 2012a}, each additional year of longitudinal data would result in roughly 3,000 fewer students receiving scores – a penalty that the increased accuracy a longitudinal model provided could not overcome.\footnote{Note that imputation methods and model averaging methods may allow models to be fit for students with less than complete data. http://www.theanalysisfactor.com/missing-data-resources/}

Uneven reporting of some data elements we may desire to include in the EWS is another concern. For example, in Wisconsin in-school suspension events 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. Know and understand the rules for classifying a student as belonging to a graduation cohort after a transfer. As \cite{Bowers and Sprott 2012a} note, there exist substantial variation across the studies of early warning indicators. This variation makes comparison of varying indicators difficult, but is not problematic to the predictive model itself as long as the decision rule is clear, consistently applied, and documented and communicated.

\section{4.5 Evaluating Model Fit to Select Functional Form}

As previously discussed, we have focused on the use of the ROC as the accuracy metric most appropriate for evaluating the utility of EWIs in an early warning system. However, after selecting ROC, it is still important to consider a number of decision points around that measure.
The procedure we adopt is a five step procedure:

1. Identify best in-sample fit by repeated cross-fold validation
2. Set test conditions for model accuracy
3. Build test set from out-of-sample data
4. Loop through available model types, retain those that exceed step 1
5. Decide!

First, a series of generalized linear models (GLMs) are fit to one cohort of students and used to generate predictions on another cohort. The best performing of these parametric models is taken as the baseline. Generalized linear models are a strong starting point for identifying promising features because they provide a nice balance between time to compute and flexibility in fit. Figure 4 adapts a plot from James et al. (2013) that describes the approximate tradeoffs between flexibility and interpretability of different functional forms. In a search for accuracy it is important that we test across this model space. In order to inform that search, starting with GLMs has the of being a) familiar to most quantitative researchers, and b) being relatively easy to interpret.

In the current case, we recommend taking as complete a set of the ABC indicators as available in the data system and using a traditional logistic model as a starting point, creating model 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 our first model we will include the following student characteristics from Table 6: 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 we also fit simple models of assessment data, attendance data, and discipline data independently.

Instead of reporting coefficients and traditional model fit statistics, we can benchmark the ABC models directly against the EWIs from Figure 1, as shown in Figure 5. This figure compares five models - one using only attendance and demographics, only using behavior and demographics, only using assessments and demographics, a simple model combining all three and finally a more complex model that also includes school fixed effects. The full model shown in Figure 5 is a considerable improvement.

Figure 5 depicts the models we have 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. It is up to us to decide how to assign students into the graduate and dropout categories based on that measure. 14

This figure also displays these models compared to all other EWIs published in the literature as reported by Bowers et al. (2013). This clearly shows 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 Muthén and the Bowers growth mixture models (GMM) stand out as far above all other methods.

Figure 5 tells us three important things. First, for in-sample prediction on the training data, the regression strategy used here provides very strong predictive power relative to other early warning indicators – the majority of which have very low rates of 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 models we 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 in creating 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); Muthén (2004) are highly predictive, they require several years of consistent records for each individual student. In the future, new students without several years of data would not be able to receive a prediction, or the accuracy of their prediction would be potentially compromised by the reliance on assessment of student trends over time. This is a practical tradeoff with real significance.

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 and to avoid overfitting the model to the particular sample available for the analysis. Figure 6 shows the different ROC curves for our models when fit to their training data and then refit to a test set of data. In this case, the variability between test and training samples is not very large, but curiously the ROC is higher on the test data than the training data. This variance in performance between the test and training data is why more sophisticated methods are necessary.

Identifying a model search procedure that minimizes the variance in the test set performance of our

14Note that we may calculate the probability of dropping out or the probability of graduation depending on how we have coded the dependent variable.
Figure 5: 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.
model and increases accuracy is the challenge posed by building an applied model. The next section discusses specific strategies to guard against model overfit and to evaluate model performance that provide for accurate assessment of the test data prediction rate.

4.6 Testing Procedure

To address these shortcoming, next we decide on a set of accuracy criteria. In an ROC framework the most important accuracy criteria to set is the weight of false-positives relative to false-negatives, which helps us compare models at their optimal thresholds of binary classification.

Using the information in Figure 5 and 6 to make a decision about the best model requires both selecting a metric and determining our values for tradeoffs between false-positives and false-negatives. As Bowers et al. (2013) describes, we can use the area under the curve (AUC) metric to compare models without deciding in advance where we want to set our binary classification threshold. The closer AUC is to 1, the better our model is at classification (Hanley and McNeil, 1982).

Figure 5 also shows that at the extremes all of the models converge in their accuracy. We can think of moving along the curve as an adjustment to our tolerances for false positives 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). We opt for the “closest topleft” method.

However, in the case of identifying potential dropouts, we do not treat false-positives and false-negatives as being of equal weight. Working with educators and content specialists within the Department of Public Instruction, we identified that false-positives, falsely identifying a student as a potential drop out, as considerably less problematic than a false-negative, or, falsely identifying a student who is likely to dropout as a potential graduate. As we saw in Figure 2, students who graduate late or dropout can often look very similar to students who graduate on time across many of the data elements we have available to us. 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, and thus the real benefit of the system comes from identifying the next tier of students who are at risk, even though this also increase the false-positive rate. In the ROC framework, we thus 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 our final linear predictor, or Ŷ, 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. As Figure 5 shows, unless they choose extreme values, the more complex model is likely to still provide a superior fit.

So far we have only talked about model fit and correct classification based on the data we have at hand. However, while we are interested in the performance of DEWS on the data it is fit to, and while our training data has a substantial number of observations, we need to evaluate the performance of our models on data...
Figure 6: ROC curves of all models compared to prior EWI indicators. The dashed line represents the ROC performance of the model on the test dataset. In this case the test set is composed of half of a grade 8 cohort, the training data was composed of the other half, roughly 30,000 observations each.
outside of the training data using a test set (Hastie et al., 2009; James et al., 2013). This is to ensure that the models we develop for prediction are not overfit to the data they are built upon. This is also known as the bias-variance tradeoff.

In the EWS framework the best test set would be to use a second (subsequent) cohort of students. This most closely resembles the EWS workflow in which a model is built on one cohort of students who graduated (or dropped out) and used to generate predictions on current students in an early grade. However, this requires a substantial amount of data and may not be feasible. While the exact choice of a test data set is out of the scope of this paper, (see instead: Hastie et al., 2009; James et al., 2013) or for a Bayesian interpretation (Gelman et al., 2013), some recommendations can be found in Table 9.

![Table 9](image)

<table>
<thead>
<tr>
<th>Method</th>
<th>Data Loss</th>
<th>External Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hold 1 Cohort Out</td>
<td>Highest</td>
<td>Highest</td>
</tr>
<tr>
<td>Random Sample From 2+ Pooled Cohorts</td>
<td>High</td>
<td>Higher</td>
</tr>
<tr>
<td>Simple Random Sample Within Training Cohort</td>
<td>Moderate</td>
<td>Low</td>
</tr>
<tr>
<td>Stratified Sample Within Training Cohort</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td>Repeated Fold Cross-validation</td>
<td>Low</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

As a rule of thumb it is best to take advantage of the cohort structure of the data as the applied model continues to be used. For initial model building purposes, repeated fold cross-validation is 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. Analysts comfortable with sampling theory should take advantage of their understanding of the potentials for bias in various sampling techniques when considering a sampling strategy.

Of further consequence is that given the amount of computational time necessary for some algorithms or datasets, it may be necessary to use some technique in Table 9 in order to test multiple methods. In the case of the EWS, 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 repeated fold cross-validation strategy on a single cohort. Depending on the volume of data, particular algorithms chosen, and the nature of the problem an analyst will have to balance these concerns within the computational resources available for the project.

5 Examples and Results

Next, a series of non-parametric models are fit to one cohort of students and used to generate predictions about another cohort of students. The models that prove more accurate than the parametric models are retained, and those that fail to achieve better classification than the parametric models will not be considered.

In order to demonstrate the utility of testing across functional forms, we build a testing procedure to evaluate a number of algorithms simultaneously. To do this, we make use of the caret package in the open source statistical programming language R, which contains a number of functions to simplify the process of evaluating statistical models using a common metric (Kuhn et al., 2013; R Core Team, 2013). For the results in Figure 7, we build a training set using the 2006-07 grade 8 cohort, and a test set using the 2007-08

---

16 Non-parametric models is used here broadly to represent several classes of models that are not in the generalized linear family and are available in the caret package for the R Statistical package (R Core Team, 2013) and are depicted in Figure 4. 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 easily unpacked for communication and explanation.
grade 8 cohort. Within the training set we randomly select 40,000 cases in order to reduce computational time.\footnote{For this case, using the full 55,000 + cases results in a slowdown of approximately 2x across the procedure, which at the time of press could not produce results.} Next, the training procedure was established – each algorithm would be evaluated using 10 folds, repeated 3 times. This would provide the estimated test accuracy, which is the element we are seeking to maximize. Next, we choose to test the models across a series of tuning parameters unique to the algorithm, a process which is automated using the \texttt{caret} package and the \texttt{tuneLength} parameter, which was set to six. This means that each algorithm’s parameter space was split into 6 and the procedure was repeated across each of the six with the \texttt{caret} package reporting back the tuning parameters that produce the best model fit.\footnote{A complete script of the test procedure is available for review.}

Next, after each model is trained on the data, we extract several fit statistics and store them: training AUC, test AUC, sensitivity, specificity, and the time elapsed to fit the model. The sensitivity and specificity are recorded at various thresholds of the predicted probability to allow for the drawing of ROC curves to compare the models directly. Finally, these results are plotted against the results from Figure 5 which is shown in Figure 7.

Figure 7 reports the results of this procedure by comparing the training set accuracy for all of the algorithms and comparing them to the prior EWIs. Here we do not show the test set accuracy because in the prior literature test set accuracy is not directly reported. The results show that while most methods cluster around the GLM models fit in Figure 5, a few models are much more highly accurate and reflect the performance of the growth mixture models. It is important to note that these models all rely on only one year of data for students in eighth grade. Additionally, the data requirements are very modest relying on data commonly found in administrative records.

The improvement between the best model and the best GLM model we fit equates to an additional 20% of potential dropouts correctly classified for the same level of false-alarm – or an additional 1,300 correctly identified students.

### 5.1 Choosing the Best Model

With well over 30 models to choose from, how do we go about selecting the best model? As mentioned above, the choice of test statistic will depend on the application of the model – a process that can only be conducted after discussion with the intended users of the model.\footnote{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.} As mentioned above, a strong metric is AUC. The closer AUC is to 1, the closer the model is to perfect classification. Figure 8 depicts the AUC for each method shown in Figure 7. Immediately we see that most of the models are clustered around an AUC of between 0.83 and 0.87. Two models stand out as particularly terrible – neural networks and stepwise quadratic discriminant analysis (Venables and Ripley 2002). Two models stand out as clearly superior – a bagged tree algorithm (a method pioneered by Leo Breiman) and C5.0 (Kuhn et al. 2013).
Figure 7: 26+ Machine Learning Algorithms’ training data accuracy compared to prior EWIs and GLMs
Figure 8: Choosing the best model using test fit statistics
warning indicators, communicate that information efficiently, and incorporate that information into decision making processes about interventions and educational supports for students identified.

The reader may wonder what the benefit of implementing the complicated procedure described above is when the end result of this balancing 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 characteristics such as the AIC or other measures of model fit. Instead it was selected expressly for its ability to provide accurate predictions on future cohorts of students and this 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 that the ranking of these models in terms of accuracy will be stable. By setting up a systematic testing environment for statistical models for the EWS 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 prediction and thus in the action take in response to those predictions. When providing tens of thousands of predictions to thousands of educators within a state, this due diligence is absolutely necessary.

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, building on the work done in large urban school districts across the country. The nature of the data and the scale of the problem have resulted in the adoption of a constrained form of an applied modeling approach to the problem – building plausible sets of predictor variables and testing them within a number of functional forms against a test set of data, choosing the most accurate. The tradeoffs and decision points beginning from collecting and tidying the data up 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 system presented here is the result of these tradeoffs and by documenting them 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 and around decision points which can be tested and explored. Future extensions of the system, for example incorporating a multiple model approach using Bayesian model averaging, must demonstrate their advantages against these tradeoffs in order to justify the development time and additional complexity they may introduce. Creating this consensus around these deliberate tradeoffs and monitoring the results in a feedback loop with internal and external users is critical to the success of the EWS.
7 References

References


8 Appendix

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

- R version 3.0.2 (2013-09-25), 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, stats, tools, utils
- Other packages: apsrtable 0.8-8, eceptr 0.3, ggplot2 0.9.3.1, highlight 0.4.4, highr 0.3, knitr 1.5,
  MASS 7.3-29, pROC 1.6.0.1, reporttools 1.1.1, scales 0.2.3, stargazer 4.5.3, xtable 1.7-1
- Loaded via a namespace (and not attached): abind 1.4-0, arm 1.6-10, car 2.0-19, coda 0.16-1,
  colorspace 1.2-4, data.table 1.8.10, dichromat 2.0-0, digest 0.6.4, evaluate 0.5.1, foreign 0.8-58,
  formatR 0.10, gtable 0.1.2, labeling 0.2, lattice 0.20-24, lme4 1.0-5, maptools 0.8-27, Matrix 1.1-11,
  memisc 0.96-9, minqa 1.2.2, munsell 0.4.2, nlme 3.1-13, nnet 7.3-7, plyr 1.8, proto 0.3-10,
  RColorBrewer 1.0-5, Rcpp 0.10.6, reshape2 1.2.2, sp 1.0-14, splines 3.0.2, stringr 0.6.2

This is an unedited sample R script for constructing the model test procedure discussed above. This
script and the supporting functions will be open sourced on GitHub in the future.

```r
# The EWS Testing Routine
#
# 1. Get data set up into test and training sets
# 2. Set test parameters
# 3. Set up functions to get outputs from the model fit tests
# 4. Set tolerances and run times
# 5. Import models to test
# 6. Execute test and capture results
```

```r
# Set Test Parameters

#MODE <- "PROD"
MODE <- "DEV"
CORES <- 10
myOS <- Sys.info()$"sysname"
```

# Set up parallel processing for speed

if(myOS!="Windows"){
    library(doMC)
    if(MODE=="DEV"){
        CORES <- CORES / 2
        registerDoMC(CORES)
    } else{
        registerDoMC(CORES)
    }
} else {
    message("Not running in parallel /n I pity you and your single thread...")
    CORES <- 1
}

# Sample sizes for training sets
if(MODE == "PROD"){
    SAMPLE <- TRUE
    N <- 40000
    TIMEOUT <- 15000
} else {
    TIMEOUT <- 20
    SAMPLE <- TRUE
    N <- 6000
}

# CV settings
folds <- ifelse(MODE == "DEV", 5, 10)
reps <- ifelse(MODE == "DEV", 1, 3)

# Train methods for caret package to estimate the test error
if(MODE == "PROD"){
    fitControl <- trainControl(method='cv', number=folds, repeats=reps,
                               classProbs=TRUE, summaryFunction=twoClassSummary,
                               savePredictions = FALSE)
} else if(MODE == "DEV"){
    fitControl <- trainControl(method='boot', savePredictions = FALSE,
                               classProbs=TRUE)
}

# Show the user what they have selected
print("The following settings have been chosen:")
print(paste0("Running in ", myOS))
print(paste0("Mode: ", MODE))
print(paste0("Timeout: ", TIMEOUT, " seconds"))
print(paste0("CPU Cores: ", CORES))
print(paste0("Sample size: ", N))

################################################################################
## Load packages
################################################################################
library(caret)
library(pROC)
library(R.utils)

################################################################################
## Get data
################################################################################
# Load data and model testing functions
setwd(" ../DEWS2")
source(" ../DEWS_Paper/ROCfunctions.R")
mods <- read.csv(" ../DEWS_Paper/data/ewsModels.csv", stringsAsFactors=FALSE)

load("cache/training/trainingsetG8Y2006.rda")
train <- merge(studata, sch8, by=c("schg8"))
load("cache/training/trainingsetG8Y2007.rda")
test <- merge(studata, sch8, by=c("schg8"))
rm(mdat, a_model, b_model, c_model, studata, sch8)

train <- subset(train, transfer_out<1)
train$ontime_grad2<="Non.grad"
train$ontime_grad2[train$ontime_grad==1]<="Grad"

test <- subset(test, transfer_out<1)
test$ontime_grad2<-"Non.grad"
test$ontime_grad2[test$ontime_grad==1]<="Grad"

# Subset predictors
predsTRAIN <- train[,c("ontime_grad2",'math8','read8','race','frpl2','gender',
                      'disab_code', 'disflag_g8', 'disdays_g8', 'disab_flag',
                      'ell_comp_med2','att_rate_g8','mobility_dist_yr8',
...
'attendance', 'attend_var', 'mean_mathG8', 'sd_mathG8',
'mean_readG8', 'sd_read8', 'cohort_size2', 'per_nonwhite',
'per_ell', 'per_frpl')]

predsTRAIN <- na.omit(predsTRAIN)

predsTEST <- test[,c('ontime_grad2', 'math8', 'read8', 'race', 'frpl2', 'gender',
                     'disab_code', 'disflag_g8', 'disdays_g8', 'disab_flag',
                     'ell_comp_med2', 'att_rate_g8', 'mobility_dist_yr8',
                     'attendance', 'attend_var', 'mean_mathG8', 'sd_mathG8',
                     'mean_readG8', 'sd_read8', 'cohort_size2', 'per_nonwhite',
                     'per_ell', 'per_frpl')]

predsTEST <- na.omit(predsTEST)

# Sample data if necessary
if(SAMPLE==TRUE){
  sampsize <- N
  tmp2 <- predsTRAIN[sample(nrow(predsTRAIN), sampsize), ,drop=FALSE]
  tmp <- predsTRAIN[predsTRAIN$ontime_grad2 == "Non.grad", ]
  predsTRAIN <- rbind(tmp2, tmp)
  rm(tmp2, tmp)
} else{
}

# Split into classification and predictors, split into training and test data
trainCLASS <- predsTRAIN$ontime_grad2
trainCLASS <- as.factor(trainCLASS)
testCLASS <- predsTEST$ontime_grad2
testCLASS <- as.factor(testCLASS)

predsTRAIN <- predsTRAIN[, -1]
predsTEST <- predsTEST[, -1]

rm(test, train)

# Build a matrix of predictors for training
myF2 <- formula(~ 0 + math8 + read8 + I(math8^2) + I(read8^2) +
                factor(race) + factor(frpl2) + factor(gender) +
                factor(ell_comp_med2) +
factor(disflag_g8) + disdays_g8 + factor(disab_flag) + 
att_rate_g8 + mobility_dist_yr8 + attendance + attend_var + 
mean_mathG8 + sd_mathG8 + mean_readG8 + sd_readG8 + 
I(mean_mathG8^2) + I(mean_readG8^2) + I(sd_mathG8^2) + 
I(sd_readG8^2) + cohort_size2 + per_nonwhite + per_ell + 
per_frpl)

trainData <- model.matrix(myF2, data=predsTRAIN)
testData <- model.matrix(myF2, data=predsTEST)

# Read in candidate models from a list and subset
candidatemods <- mods$method.[mods$EWS."Yes"]
modeltests <- vector("list", length(candidatemods))

varnames <- c("math8", "read8", "math8_sq", "read8_sq", "raceA", "raceB", 
"raceH", "raceI", "raceW", "frpl2Y", "genderM", "ellL0", 
"ellMED", "ellNAT", "disflagMinor", "disflagNo", "disflagSevere", 
"disdays_g8", "disab_flagNone", "disab_flagOtherSevere", 
"att_rate_g8", "mobility_dist_yr8", "attendance", 
"attend_var", "mean_mathG8", "sd_mathG8", "mean_readG8", 
"sd_readG8", "mean_mathG8_sq", "mean_readG8_sq", 
"sd_mathG8_sq", "sd_readG8_sq", "cohort_size", 
"per_nonwhite", "per_ell", "per_frpl")

colnames(trainData) <- varnames
colnames(testData) <- varnames
rm(predsTEST, predsTRAIN)

# Clean up
print("Data is now prepared...")

### Load Test Functions

# Set the tune length for the train function
# algorithm parameters to evaluate
LENGTH <- ifelse(MODE == "PROD", 6, 4)
modSearch <- function(x, datatype=c("train", "test"), modelKeep=NULL, 
   length = LENGTH, timeout = NULL)
{
  datD <- c("rda", "lda2", "hda", 'mlp', 'mlpWeightDecay', 'rbf', 'rpart2', 
               "treebag", 'rf', 'plr', 'lda', 'xyf')
  if(x %in% datD){
    fit <- tryCatch({
      evalWithTimeout(
        train(trainData[, -9], trainCLASS, 
        method=x, 
        trControl=fitControl, 
        tuneLength = length, metric="ROC")
      }, timeout = timeout, elapsed = timeout),
      TimeoutException = function(ex) {
        print("Timeout. Skip");
      }, error = function(e) print(paste0("Failure of model: ", x, 
               " \n", " For: ", e)))
    } else {
      fit <- tryCatch({
        evalWithTimeout(
          train(trainData, trainCLASS, 
          method=x, 
          trControl=fitControl, 
          tuneLength = length, metric="ROC")
        }, timeout = timeout, elapsed = timeout),
        TimeoutException = function(ex) {
          print("Timeout. Skip");
        }, error = function(e) print(paste0("Failure of model: ", x, 
          " \n", " For: ", e)))
      }
    }
    if(class(fit) == "character"){
      cat(fit)
    } else if(class(fit) == "train"){
      if (length(datatype) > 1){
        ROCtr <- ROCtest.train(fit)
        ROCte <- ROCtest.train(fit, testdata=list(testData = testData, testCLASS = testCLASS),
          best.method="closest.topleft",
          best.weights=c(10, .11))
      } else if(length(datatype) < 2 & datatype=="test"){
        ROCte <- ROCtest.train(fit, testdata=list(testData = testData, testCLASS = testCLASS),
          best.method="closest.topleft",
          best.weights=c(10, .11))
      }
    }
  } else{
  }
  return(fit)
}

modSearch <- function(x, datatype=c("train", "test"), modelKeep=NULL, 
   length = LENGTH, timeout = NULL)
{
  datD <- c("rda", "lda2", "hda", 'mlp', 'mlpWeightDecay', 'rbf', 'rpart2', 
               "treebag", 'rf', 'plr', 'lda', 'xyf')
  if(x %in% datD){
    fit <- tryCatch({
      evalWithTimeout(
        train(trainData[, -9], trainCLASS, 
        method=x, 
        trControl=fitControl, 
        tuneLength = length, metric="ROC")
      }, timeout = timeout, elapsed = timeout),
      TimeoutException = function(ex) {
        print("Timeout. Skip");
      }, error = function(e) print(paste0("Failure of model: ", x, 
               " \n", " For: ", e)))
    } else {
      fit <- tryCatch({
        evalWithTimeout(
          train(trainData, trainCLASS, 
          method=x, 
          trControl=fitControl, 
          tuneLength = length, metric="ROC")
        }, timeout = timeout, elapsed = timeout),
        TimeoutException = function(ex) {
          print("Timeout. Skip");
        }, error = function(e) print(paste0("Failure of model: ", x, 
          " \n", " For: ", e)))
      }
    }
    if(class(fit) == "character"){
      cat(fit)
    } else if(class(fit) == "train"){
      if (length(datatype) > 1){
        ROCtr <- ROCtest.train(fit)
        ROCte <- ROCtest.train(fit, testdata=list(testData = testData, testCLASS = testCLASS),
          best.method="closest.topleft",
          best.weights=c(10, .11))
      } else if(length(datatype) < 2 & datatype=="test"){
        ROCte <- ROCtest.train(fit, testdata=list(testData = testData, testCLASS = testCLASS),
          best.method="closest.topleft",
          best.weights=c(10, .11))
      }
    }
  } else{
  }
  return(fit)
ROCtr <- NULL
} else if (length(datatype) < 2 & datatype=="train"){
    ROCtr <- ROCtest.train(fit)
    ROCte <- NULL
}
if(modelKeep == TRUE){
    return(list(model=fit, summaryTr=ROCtr, summaryTe = ROCte, method=fit$method,
        time = fit$times$everything[3]))
} else if(modelKeep == FALSE){
    return(list(method=fit$method, summaryTr=ROCtr, summaryTe = ROCte,
        time = fit$times$everything[3]))
}

# Define a data capture function
# Want to capture
# 1. Training ROC
# 2. Test ROC
# 3. Model fit time
# 4. Tuning parameters for each model
## MOD1 <- modSearch("lda", datatype="test")
## MOD1train <- modSearch("lda", datatype="train")
## MOD1 <- modSearch("lda", datatype=c("train", "test"))

dfExtract <- function(mod){
    #mod <- list(mod$model, mod$summaryTr, mod$summaryTe)
    suppressWarnings({
        newdatB <- data.frame(sens = smooth(mod$summaryTr@rocobj)$sensitivities,
                  spec = smooth(mod$summaryTr@rocobj)$specificities,
                  grp="train",
                  auc = mod$summaryTr@auc,
                  method = mod$method,
                  elapsedTime = mod$time)

        newdatA <- data.frame(sens = smooth(mod$summaryTe@rocobj)$sensitivities,
                  spec = smooth(mod$summaryTe@rocobj)$specificities,
                  grp="test",
                  auc = mod$summaryTe@auc,
                  method = mod$method,
                  elapsedTime = mod$time)

tmp <- rbind(newdatA, newdatB)
}
tmp$sens <- as.numeric(tmp$sens)
tmp$spec <- as.numeric(tmp$spec)
tmp$auc <- as.numeric(tmp$auc)
tmp(method) <- as.character(tmp$method)
tmp$auc <- as.numeric(tmp$auc)
tmp$grp <- as.character(tmp$grp)
tmp$elapsedTime <- as.numeric(tmp$elapsedTime)
return(tmp)
}
}

## Build Outputs
#
### Set up a dataframe to capture model fit

ModelFits <- expand.grid(sens = NA, spec = NA, grp = NA, auc = NA,
             method = rep(candidatemods, each = 1028),
             elapsedTime = NA)

# Class variables correctly to avoid errors
ModelFits$grp <- as.character(ModelFits$grp)
ModelFits$method <- as.character(ModelFits$method)
ModelFits$sens <- as.numeric(ModelFits$sens)
ModelFits$spec <- as.numeric(ModelFits$spec)
ModelFits$auc <- as.numeric(ModelFits$auc)

# set up a progress bar
pb <- txtProgressBar(min = 0, max = length(candidatemods), style = 3)

# Optional changes to parameters
# TIMEOUT <- 600
# CORES <- 8

# Loop to test models and capture results
for(i in candidatemods){
  p <- match(i, candidatemods)
  fit <- try(modSearch(i, datatype=c("test","train"), modelKeep=FALSE,
                       length = LENGTH, timeout = TIMEOUT))
tmp <- tryCatch(dfExtract(fit), error = function(e) "No Model Ran")
#
if(class(tmp) == "data.frame"){
  ModelFits[ModelFits$method == i,] <- tmp[tmp$method == i,]
} else{
  ModelFits <- ModelFits
  print(paste(tmp, "failure for model type:", i, sep=" "))
}
setTxtProgressBar(pb, p)