# Exploring a Generalizable Machine Learned Solution for Early Prediction of Student At-Risk Status

Chad J. Coleman

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

# Abstract

Exploring a Generalizable Machine Learned Solution for Early Prediction of Student At-Risk Status

Chad J. Coleman

Determining which students are at-risk of poorer outcomes -- such as dropping out, failing classes, or decreasing standardized examination scores -- has become an important area of both research and practice in K-12 education. The models produced from this type of predictive modeling research are increasingly used by high schools in Early Warning Systems to identify which students are at risk and intervene to support better outcomes. It has become common practice to re-build and validate these detectors, district-by-district, due to different data semantics and various risk factors for students in different districts. As these detectors become more widely used, however, a new challenge emerges in applying these detectors across a broad spectrum of school districts with varying availability of past student data. Some districts have insufficient high-quality past data for building an effective detector. Novel approaches that can address the complex data challenges a new district presents are critical for advancing the field. Using an ensemble-based algorithm, I develop a modeling approach that can generate a useful model for a previously unseen district. During the ensembling process, my approach, District Similarity Ensemble Extrapolation (DSEE), weights districts that are more similar to the Target district more strongly during ensembling than less similar districts. Using this approach, I can predict student-at-risk status effectively for unseen districts, across a range of grade ranges, and achieve prediction goodness but ultimately fails to perform better than the previously published Knowles (2015) and Bowers (2012) EWS models proposed for use across districts.

# Table of Contents

List of Charts, Graphs, Illustrations ..... v

List of Tables ..... viii

Acknowledgments ..... x

Dedication ..... xi

Chapter 1: Introduction & Background ..... 1

1.1 Why Predicting Dropout is Important ..... 3

1.2 Problem Statement ..... 5

1.3 Research Questions ..... 7

1.4 Expected Limitations ..... 8

Chapter 2: Review of Literature ..... 11

2.1 Selection Criteria ..... 12

2.2 Learning Analytics & Educational Leadership Data Analytics ..... 14

2.3 Early Work on Predictors of Dropout ..... 16

2.3.1 Academic Indicators ..... 17

2.3.2 Attendance Indicators ..... 19

2.3.3 Student Behavioral Indicators ..... 20

2.3.4 Student Identity ..... 23

2.4 Threshold Based Methods ..... 25

2.5 Machine Learned Early Warning Systems ..... 30

2.6 Generalizing Models ..... 35

2.7 Evaluating Early Warning Systems ..... 36

Chapter 3: Methodology ..... 38

3.1 Criterion for Building the District Similarity Ensemble Extrapolation Model ..... 38

3.1.1 Model Validation ..... 41

3.1.2 Calculating District-to-District Similarity ..... 45

3.1.3 Comparing the District Similarity Ensemble Extrapolation Model ..... 51

3.2 Data for Analysis ..... 56

3.2.1 Study Participants ..... 58

3.2.2 Data for District-to-District Similarity ..... 60

3.3 Instruments ..... 63

3.4 Preparing the Data for Modeling ..... 64

3.5 Model Parameter Tuning ..... 67

3.6 Model Fitting ..... 71

3.6.1 Aggregate Data Model ..... 72

3.6.2 Mean Model ..... 73

3.6.3 District Similarity Ensemble Extrapolation ..... 78

3.6.4 The Chicago Model ..... 87

3.6.5 The Balfanz Model ..... 88

3.6.6 The Knowles Model ..... 89

3.7 Measuring Feature Importance ..... 91

3.8 Calculating Prediction Equity ..... 92

Chapter 4: Research Findings ..... 95

4.1 Results of the Aggregate Data Model ..... 95

4.1.1 Aggregate Model Feature Importance ..... 98

4.2 Results of the Mean Model ..... 100

4.2.1 Pillar Selection ..... 101

4.2.2 Mean model Performance on New Districts ..... 104

4.2.3 Pillar Model Feature Importance ..... 109

4.3 Results of the District Similarity Ensemble Extrapolation ..... 110

4.4 Results of The Chicago Model ..... 112

4.5 Results of The Balfanz Model ..... 114

4.6 Results of The Knowles model ..... 116

4.6.1 Knowles Model Feature Importance ..... 118

4.7 Comparing Across the Generated Models (Grades 1st - 12th) ..... 120

4.8 Comparing Across the Generated Models (6th Grade Students) ..... 121

4.9 Comparing Across the Generated Models (9th Grade Students) ..... 122

4.10 Prediction Equity Results ..... 124

4.11 Summary of Findings ..... 131

Chapter 5: Conclusions & Discussion ..... 134

5.1 Common Data Standards & Open Access Algorithms ..... 135

5.2 Dissecting the Early Warning System ..... 136

5.3 Prediction-Driven Intervention Strategies ..... 139

5.4 Addressing Prediction Bias ..... 146

5.5 Limitations ..... 150

5.6 Future Work ..... 154

5.7 Concluding Remarks ..... 158

References ..... 162

Appendix A: Study Participant Descriptive Tables & Figures ..... 187

Appendix B: Similarity Data Descriptive Tables ..... 197

Appendix C: Data for Modeling ..... 203

Appendix D: Pillar & Target Model Designation ..... 205

Appendix E: Aggregate Data Model Performance ..... 209

Appendix F: District Similarity Ensemble Extrapolation ..... 213

Appendix G: Chicago Model Performance ..... 215

Appendix H: Balfanz Model Performance ..... 217

Appendix I: Knowles Model Performance ..... 219

Appendix J: AUC Performance at 95% CI by Populations ..... 222

# List of Charts, Graphs, Illustrations

Figure 1: Data Analytics Lifecycle in Education ..... 15

Figure 2: Example of Data Assignment During Hold-Out Testing Model Validation ..... 42

Figure 3: Example of 5-fold Cross-Validation Data Assignment ..... 43

Figure 4: Visual Representation of Existing EWS Model Performance ..... 54

Figure 5: Flow chart of hyperparameter search procedure ..... 70

Figure 6: Process for fitting the Aggregate Data model EWS ..... 72

Figure 7: Count of Historical Outcomes by Grade and Organization ..... 75

Figure 8: Normalized Features After Missing Data Imputation Using Z-Score Standardization in the DSEE Calculation ..... 81

Figure 9: Adjusted Similarity Between Pillar District and Target District Models Using Normalized Euclidian Distance Function ..... 84

Figure 10: Flow Diagram of Knowles EWS Modeling Process ..... 91

Figure 11: AUC Performance of Aggregate Data model within Grade Levels ..... 96

Figure 12: AUC Performance of Aggregate Data model by Organization and Grade Level 1st Through 12th ..... 98

Figure 13: AUC Performance of Pillar Models on Test Hold-Out Compared to % of Missing data in Feature Space ..... 101

Figure 14: Average AUC Performance of Pillar Organization Models on Test Hold-Out Data During Model Training ..... 102

Figure 15: Performance of Pillar Model AUC within Grade and Organization ..... 103

Figure 16: Model Performance of Individual Pillars and Mean model on Target Data ..... 104

Figure 17: Average Performance of Pillar Model and Mean Detectors on All Target District Data. ..... 106
Figure 18: AUC Performance of Pillar Model and Mean Detectors within each Target District Data ..... 107
Figure 19: Mean model Performance on Target Districts by (%) Missingness ..... 108
Figure 20: Gini Feature Importance Values of Pillar District Models ..... 109
Figure 21: DSEE AUC Performance Across All Target District Records ..... 111
Figure 22: Calculated DSEE AUC Performance within Grade Level ..... 111
Figure 23: AUC performance of the Chicago On-Track Indicator EWS by (%) of missing data across 9th grade students ..... 113
Figure 24: AUC performance of the Chicago On-Track Indicator by School District. Red line provides a reference for a 0.5 AUC. ..... 114
Figure 25: AUC performance of the Balfanz EWS by (%) of missing data across 6th grade students ..... 115
Figure 26: AUC performance of the Balfanz EWS by School District. A red reference line is provided to show the cutoff for 0.5 AUC ..... 116
Figure 27: AUC Performance of Knowles model within Grade Levels ..... 117
Figure 28: AUC performance of the Knowles EWS by Grade and School District ..... 118
Figure 29: Gini Feature Importance Values of Knowles models ..... 119
Figure 30: Pearson correlation of EWS AUC performance on 1st through 12th grade predictions and reported district graduation rates. ..... 121
Figure 31: Pearson correlation of EWS AUC performance on 6th grade predictions and reported district graduation rates. ..... 122

Figure 32: Pearson correlation of EWS AUC performance on 9th grade predictions and reported district graduation rates. ..... 124
Figure 33: Average AUC performance of EWSs with 95% confidence intervals..... 132
Figure 34: Example output of SHAP value implementation for Machine-Learning EWS for a student predicted to graduate. ..... 138
Figure 35: Example dashboard of *fairlearn* toolkit for gender-based bias analysis of Mean model performance disparity on 10,000 random sampled student predictions ..... 147
Figure 36: Example dashboard of *fairlearn* toolkit for gender-based bias analysis of Mean model prediction disparity on 10,000 random sampled student predictions ..... 147
Figure 37: Example of an annotated slice plot of ABROCA statistic ..... 149
Figure 38: Proportion of Dropout Records by Grade and School District ..... 187
Figure 39: AUC Curve Performance of Pillar Models and Mean model on Target District Data ..... 212

# List of Tables

Table 1: Student and School Characteristics Used to Derive Similarity Scores ..... 45

Table 2: Pillar Model Hyperparameters Selected During Model Tuning ..... 77

Table 3: Example of Distances Between Target and Pillar Converted to Weights ..... 86

Table 4: Example of Student Level Predictions ..... 86

Table 5: Description of Early Warning Systems Evaluated in This Research ..... 95

Table 6: Aggregate Data model Gini Feature Importance ..... 98

Table 7: Summary results of DSEE AUC performance on Target District Data ..... 112

Table 8: AUC Results Calculated Within Demographic Groups ($1^{st}$ – $12^{th}$ Grade) ..... 125

Table 9: AUC Results Calculated Within Demographic Groups ($6^{th}$ Grade Predictions) ..... 128

Table 10: AUC Results Calculated Within Demographic Groups ($9^{th}$ Grade Students) ..... 130

Table 11: List of Potential Prediction-Driven Intervention Strategies to Mitigate the Likelihood of High School Dropout ..... 140

Table 12: Student Gender Distribution Within Each District ..... 188

Table 13: Student Ethnicity Distribution Within Each District ..... 191

Table 14: Count of Recorded Student Outcomes Within Each District ..... 194

Table 15: Reported Demographic Descritpives of Current Student Populations ..... 197

Table 16: Summary Statistics of Data Used to Determine EWS Similarity Scores ..... 200

Table 17: Percent of Missing Data Across Model Features ..... 203

Table 18: Results of Data Based Pillar and Target Organization Assignment ..... 205

Table 19: Results of All Pillar Models During Training ..... 208

Table 20: Performance of Aggregate Data model Across All Districts in Data Test Set ..... 209

Table 21: AUC Performance of DSEE on Target Districts ..... 213

Table 22: Chicago model Performance on Target Districts (9th Grade Records) ..... 215

Table 23: Balfanz model Performance on Target Districts (Grades 6th – 12th) ..... 217

Table 24: Knowles model Algorithm Search Results ..... 219

Table 25: Knowles model Performance on Target Districts (Grades 6th – 12th) ..... 220

Table 26: Calculated Mean, Standard Deviation, Standard Error, and 95% Confidence Interval of EWS AUC Performance ..... 222

# Acknowledgments

First and foremost, I must thank my doctoral sponsor, Dr. Ryan Baker at the University of Pennsylvania. Without his support, assistance, and dedicated guidance throughout this entire process, this research would never have been completed. I would also like to acknowledge Dr. Alex Bowers at Teachers College, Columbia University who was the second reader of this dissertation, words cannot express the appreciation I have for his comments and edits on the many versions that preceded this final draft.

Additionally, I would like to show gratitude to my other three committee members, Dr. Gary Natriello at Teachers College, Columbia University, Dr. Russell Neuman at Steinhardt School of Culture, Education, and Human Development, New York University and Dr. Charles Lang at Teachers College, Columbia University. Without their valuable participation, this Early Warning Systems research would have never been successfully conducted. Lastly, I would like to thank Brian Gawalt for his help in validating the statistical formulas in my research.

# Dedication

Dedicated to Mrs. Lauren Lutz-Coleman

# Chapter 1: Introduction & Background

Most researchers agree there are clear benefits to completing high school education (Amos, 2008; Clark, & Martorell, 2014; Ensminger, & Slusarcick,1992; McCallumore, & Sparapani, 2010; Swanson, 2004; Upchurch, & McCarthy, 1990), so why do millions of students continue to drop out of high school every year (Rumberger, 2020; Snyder, De Brey, & Dillow, 2018)? Researchers have committed extensive efforts to try to answer this question, with the hope that once a student is at-risk of dropping out, educators and administrators can apply a preventative or remedial intervention to curb student dropout (Bowers & Sprott 2012; Bowers, 2021). However, many factors appear to lead to student dropout, including lack of social support from parents, poor motivation, low self-esteem, parental educational achievement and value, and economic factors, making it difficult to create a single intervention that works for all students (Driscoll, 1999; Legault, Green-Demers, & Pelletier 2006).

While demographic factors correlate with eventual dropout (Dunn, Chambers, & Rabren, 2004; Rumberger, 2011), these indicators are not considered actionable. Demographic factors are considered non-actionable indicators because a school district generally does not have the capacity to improve a student's economic condition. As such, the educational research community has focused on more actionable factors such as behavior, attendance, engagement, and social-emotional learning (Barfield, Hartman, & Knight, 2012; Finn 1989). The most successful interventions have attempted to address issues related to specific indicators while also attempting to improve overall student academic engagement (Christenson & Thurlow 2004). There is a range of potential interventions, and many are costly, driving a need to identify the students that could benefit most from specific forms of support. Identifying these students can be a difficult task (Bowers, Sprott, & Taff, 2012) which has led to an ongoing effort within the

educational research community to determine which students are at risk of not graduating from high school (Ensminger & Slusarcick, 1992) to apply proactive interventions that can help get students back on track (Belfield & Levin, 2007). As such, the work in this field is twofold: researchers must identify both the indicators that determine educational success and the students most in need of receiving interventions.

These goals, along with the growing availability of student data, have led to early warning systems and early warning indicators (EWS/EWI). While some researchers have begun to classify EWIs and EWSs as two distinct solutions, with EWIs focused primarily on providing an indicator for dropout risk and EWSs designed to collect insights from an EWI to enable more focused applications of educational resources to reduce risk (Allensworth et al., 2018; Davis et al., 2013; McMahon & Sembiante, 2020), there is still debate on whether this difference is meaningful, as both EWSs and EWIs often rely on statistical methods applied to historical student data to predict outcomes for new students, and ultimately serve the same purpose of providing educators actionable predictors of a student failing to graduate high school (Bowers, 2021). Early work on predicting high school graduation tended to use statistical methods in order to infer the relationship between graduation and indicators such as grades and attendance. For example, the seminal Chicago model developed an "On-Track" indicator built from first-year high school student performance indicators and then used this newly defined feature within logistic regression to model student risk (Allensworth & Easton, 2007). This method proved useful in Chicago Public Schools with 80+ percent accuracy in predicting student dropout, leading to high popularity and wide-scale implementation (Balfanz, Herzog, & Mac Iver, 2007). Despite the On-Track indicator’s promising results in Chicago, the authors of this EWI stress that it may not perform the same for different student populations. They state that this EWI does

not consider the role that school climate and structure play in whether students succeed in high school, therefore possibly reducing the likelihood it can scale (Allensworth & Easton, 2005). While this work provided states and districts a method of addressing the dropout crisis (by identifying potential at-risk students to apply proactive, positive interventions), there is still work to be done on improving the performance of these early warning systems (Balfanz & Byrnes, 2019; Bowers, 2021). Given this need for further improvement in EWSs, the focus of this research aims to address this demand for more accurate EWS solutions that can better scale across student populations.

### 1.1 Why Predicting Dropout is Important

Graduating from high school is an educational achievement that is strongly linked to gainful well-paying employment, higher personal income, better personal health, reduced risk of incarceration, and lowered reliance on social welfare programs (Amos, 2008; Hoffman, Vargas, Venezia, & Miller, 2007). Graduation rates have been rising in the United States, towards reaching 85% nationwide by the year 2020 (NCES, 2020). While this is a positive accomplishment, it leaves millions of students not completing high school, representing a continuing crisis within the American educational system. This crisis is not evenly distributed; in the USA, there are much higher dropout rates for African American, Native American, and Hispanic/Latinx students (Driscoll, 1999; Rumberger, 1987), up to four times the rate for white students, as well as for learners from low-income families and with disabilities (Stark & Noel, 2015). Research by Reardon found that historical policies of race segregation continue to produce inequalities in learning opportunities across U.S. school districts, with the early learning opportunities available strongly associated with the school districts' socioeconomic status. Reardon states "affluent families and districts are able to provide much greater opportunities than

poor ones early in children's lives" (2019). Providing an accessible system that school districts can leverage for early dropout detection enables us to move one step closer to reducing educational inequality based on community socio-economic level (Reardon, 2019).

As students progress through their education, they may learn at different rates, and those that learn slower start to lag behind (Kaznowski, 2004). This lag causes an achievement gap, which then widens year-after-year. One way to remedy this issue is to retain a student a year and provide them additional time to catch up and close the gap (Martin, 2011; West, 2012). While this solution may be simple, it ignores the fiscal burden that an additional year of education puts on schools (Chaifetz, & Kravitz, 2004). There is also research that suggests this approach may not be beneficial to improving outcomes. Eide and Showalter analyzed the impact that grade retention and high school graduation have on overall labor market outcomes. They found that students that were retained at least one grade are less likely to graduate from high school. They also find that students who are retained have a higher likelihood of achieving lower earnings once they enter the job market compared to their non-retained counterparts (2001), making early identification of risk all the more critical. A 2005 research study conducted on the students in the Chicago Public Schools analyzed the experience of students that were retained in either the 3rd or 6th grades by, over two years, examining the relationship between the students retention and the students reading achievement. They found that students who were retained continue to struggle during the retained year. For students retained in third grade, there was no evidence to conclude that achievement rates increased. For students retained in the 6th grade, they found evidence that retention was associated with lower achievement growth (Roderick, & Nagaoka, 2005).

Improving the rate of high school graduation has the potential of positively impacting our overall economy (Heckman, 2011). Taking a new approach to the analysis of dropout, Gilbert examined the impact not graduating has on employment and labor markets. To accomplish this, a target population of 18 to 20-year-olds was identified and sampled using the Canadian Family Allowance file as the sampling frame. 18,000 individuals were selected, with a total of 9,460 individuals responding to the computer-assisted survey. This survey interview obtained information regarding demographics, social and economic characteristics, school experiences, and post-school outcomes. Though the study was conducted during an economic recession, the results suggest the high school graduates are presented with greater economic employment opportunities and students who left school early were more likely to receive public assistance (Gilbert, 1993). By reducing the number of dropouts, we would, in turn, reduce the number of individuals reliant on public support as they would hopefully have better opportunities for gainful employment with the completion of their academic credentials.

## 1.2 Problem Statement

More recently, researchers have begun to employ machine learning and data mining methods, sometimes termed predictive analytics, to find complex patterns associated with future student outcomes (Kotsiantis et al., 2003; Dekker, Pechenizkiy, & Vleeshouwers, 2009; Bowers, 2021). In K-12 education, Lakkaraju et al. (2015) used this approach to predict student dropout in two districts, finding that the Random Forest algorithm outperformed several other algorithms. Some of the efforts to use machine learning in predicting student success have scaled beyond single districts to entire states (Knowles, 2015). However, these implementations are rare as it remains a challenge to deploy predictive analytics for use in schools at scale. District data often contain substantial information about its schools and students: demographic data about the

student and teacher populations, academic performance information, financial information, disciplinary actions, and attendance records (Schildkamp, Lai, & Earl, 2012). However, in many school districts, data quality is limited. Common problems that researchers encounter when working with school district data include incompatible student ID numbers, errors in data entry, and local idiosyncratic interpretations of often ambiguous data fields. Often, accessing the data mentioned above also involves integration across multiple data warehouses to compile all the available information. In some situations, even when current data is readily available, critical data from past years is often unavailable due to the absence of a formal data system or due to the use of a data system that is difficult to query. Semantics may also change; for example, the definition of "not graduated" is not stable across years and contexts (Rumberger, 1987), but these changes may not always be clearly understood when reviewing past data.

One solution is to use models that involve simple variables that are feasible for almost any districts' data. In doing so, researchers then could assume that the model will be valid in new contexts, even contexts that may be quite different from the context where the model was initially developed (e.g., Neild, Stoner-Eby, & Furstenberg, 2008). The Chicago model (Allensworth & Easton, 2007) is a common choice for this type of application. While this method has proved useful in the past, such a system has not been shown to achieve the performance of those driven by more advanced techniques of modeling, such as machine learning.

Despite the advancements made with early warning systems, there has yet to be an effective modeling method that can be applied to school districts that suffer from data quality issues, while also taking into consideration the unique heterogeneity properties of the individual school district. This presents a challenging problem, as schools that suffer from data quality

issues seem destined to use lower-performing methods of risk analysis until they can populate the data required to drive the development of a machine-learned model.

## 1.3 Research Questions

I hypothesize that by utilizing models from districts with sufficient data, researchers can create a process for generalizing models, which will produce predictions for districts lacking high-quality data, districts for which it is otherwise infeasible to generate their own unique models. As such, my objectives for this research are three-fold.

**RQ1**: First, I explore the efficacy of whether it is possible to develop a predictive modeling approach that can determine student risk of high school dropout with better accuracy than simple methods, such as the Chicago model, for school districts with low amounts of high-quality data.

**RQ2**: Second, I investigate solutions that take a separate set of features selected to describe each population's attributes into account within the modeling approach, i.e., not building separate models for each district but taking district features into account within a broader model, with the hope that including these features will enable the models to scale while improving overall model performance.

**RQ3**: Lastly, I compare the performance of this system against existing generalized EWS detectors with varying levels of complexity and interpretability, mainly a Growth Mixture Model published in 2012 (Bowers & Sprott, 2012) and replicated in 2015 (Knowles, 2015), the Knowles Machine Learning Ensemble published in 2015 (Knowles, 2015), the Balfanz logistic regression model,

published in 2007 (Balfanz, 2007), and the Chicago model, originally published in 2007 (Allensworth & Easton, 2007).

I call this alternative solution the District Similarity Ensemble Extrapolation (DSEE). The DSEE attempts to customize a model for a specific "Target" school district based on models from other school districts where full datasets are available, taking into account the degree of similarity each school district has to the Target district. I compare the effectiveness of this approach to simply averaging multiple existing models from different districts, where all existing models are given equal weight. I also compare the quality of the DSEE approach to the earlier solution of using simple generic models-- specifically, the Chicago model and the more recently published, higher-performing Growth Mixture Model (Bowers & Sprott, 2012; Knowles, 2015).

My approach differs from previous research on early warning systems in that there exists a gap of knowledge on how to generalize models across districts to develop high-quality machine learning-driven at-risk predictions for schools with access to little historical data. The data I use for this study is sufficiently large enough to be considered nationally representative, allowing me to validate this method across a wide range of unique students from various regions and backgrounds within the United States. The magnitude of this data also presents the possibility of conducting additional analysis related to identifying any algorithmic predictive bias that may occur given the inherent risks of utilizing a machine learning driven solution.

### 1.4 Expected Limitations

It is worth noting that this study may encounter several limitations. While the results of this study prove useful to educators, there likely will need to be additional analysis conducted with factors beyond the scope of this initial investigation in order to improve external validity.

Analytical dissection of how the models perform in locations with highly diverse student populations would be helpful, as such diversity is not wholly found within the data set used in this dissertation work (i.e., this dissertation uses a diverse range of settings, with diverse student populations in aggregate, but not necessarily in any one specific school).

Moreover, the data utilized within this study was gathered using an educational data management tool purchased by educators across the U.S. This specific tool provided educators with three primary functions: (a) to aggregate data from historically siloed systems (grade books, attendance records, assessment scores, etc.), (b) to flatten this aggregated data by mapping to a unified schema, and (c) to provide actionable data-driven insights to educators through the use of a dashboard. This means that this analysis is limited to school districts with the capacity to purchase such a tool and may not include districts that opted to spend the funding on other resources they deemed more necessary or districts that did not have sufficient funding to purchase this tool. However, many districts serving low-income students are included in the population being studied.

Furthermore, as this data was collected using a third-party software system not owned by the researcher, additional stakeholders (data engineers) are involved with accessing certain aspects of the data sample. This means the capacity to conduct a further, more in-depth analysis of certain areas of the modeling approach is limited by the availability of these stakeholders. Additionally, while the data for this research comprises millions of unique students collected at a national level, at the time of this research, the data set still lacks data from districts with substantial Native American populations or those located in extremely rural regions, such as northern or western Alaska.

Lastly, recent research has shown that common, widely used risk indicators are often ineffective in accurately identifying students at-risk of dropping out, which could potentially limit this model's performance. Without the ability to incorporate meaningful insight from teachers or counselors within each school, this model is unable to account for unobserved factors, such as personal home life issues, or drug use, not recorded in the data that could potentially provide better indications of risk than the current set of widely used factors (Gleason & Dynarski, 2002).

# Chapter 2: Review of Literature

A literature review was conducted over existing published evidence related to the topic of predicting student risk of high school dropout. To perform this review, peer-reviewed published literature was collected and examined beginning from the year 1980 to the present date (2020). Early Warning focused research was then grouped into two categories: simplified threshold-based methods and advanced contemporary methods.

Existing literature classifies EWS as simple threshold-based when they rely on generated threshold values that can be applied to specific education-related indicators to identify risk (Allensworth, 2013; Allensworth, Nagaoka, & Johnson, 2018; Carlson, 2018; Davis, Gleason, & Dynarski, 2002; Herzog, & Legters, 2013; Bowers, 2021). These simple threshold-based methods of EWSs require little to no implementation effort on the part of educators or districts and rely on little or no statistical modeling (Neild, Balfanz, & Herzog, 2007) to be put into practice. They are often based on research designs that would be considered standard statistical procedures for data modeling, which are then used to extract predictor level cut-points. Simple threshold-based EWSs rely on surface-level student indicators such as (non) cumulative grade point average, course pass rate, and current grade level to generate the prediction. Simple threshold-based EWSs utilize methods such as generalized linear modeling (Roderick & Camburn, 1996), growth modeling (Bowers & Sprott, 2012), maximum likelihood logistic regression (Kupersmidt & Coie, 1990) or discriminant analysis (Curtis, 1983) on these predictors to generate a series of cut points and then ultimately to determine student risk. An example of this in practice would be the Chicago On-Track indicator, where a student is considered to be on track for graduation if they meet the following criteria: (a) the number credits accumulated during the first year of high school is greater than or equal to five and (b) the number of semester

core course failures during the first year of high school is less than or equal to one; otherwise the student is considered off-track and is at risk of dropping out of high school (Allensworth & Easton, 2005).

Articles classified as advanced methods use analysis techniques that are computationally intensive and have only recently been accessible to everyday researchers using emerging methods enabled by access to aggregated big data (Sara, Halland, Igel, & Alstrup, 2015). These articles generally use techniques related to supervised (Aguiar, Lakkaraju, Bhanpuri, Miller, Yuhas, & Addison, 2015), or unsupervised (Márquez-Vera, Cano, Romero, Noaman, Mousa Fardoun, & Ventura, 2016) machine learning methods, ranging from classification algorithms (Coleman, Baker, & Stephenson, 2020) to deep learning neural networks (Kotsiantis, Pierrakeas, & Pintelas, 2003), that fit more complex functions that are often difficult to re-implement by hand or understand without sophisticated inspection methods (Nagrecha, Dillon, & Chawla, 2017).

## 2.1 Selection Criteria

The selection of literature for review was based on two key criteria. First, an analysis of several existing literature reviews on dropout prediction was conducted. Dupéré, Leventhal, Dion, Crosnoe, Archambault, and Janosz (2015) conducted a review of existing dropout literature to better understand the determinants of dropout (both long-term and immediate) with the goal of understanding why and when students drop. The result of this research was the creation of a *stress process, life-course model* of dropout. This model highlights how risk factors, proximal precipitating stressors and supports, play a role in understanding eventual student graduation outcomes (Dupéré, Leventhal, Dion, Crosnoe, Archambault, and Janosz, 2015).

A literature review conducted by Freeman & Simonsen (2015) focused on outlining and understanding how policy and practice interventions impact high school completion rates. Through their analysis, the authors found that the majority of the existing research is focused on single-component, individual, or small-group interventions at the high school level, despite there being significant evidence that successful intervention is based on multiple factors and a need for interventions at grade levels beyond high school.

A similar review was conducted by Rumberger, Addis, Allensworth, Balfanz, Bruch, Dillon, & Tuttle, C. (2017), where they completed a focused analysis of dropout literature to inform secondary educators on how to better monitor their student population in order to reduce high school dropout. They found that 1) proactive intervention is important when students show early signs of attendance, behavior, or academic problems, 2) individualized support improves graduation outcomes for students that are showing signs of risk, 3) offering curriculum that promotes the benefits of high school graduation with college and career success increases student success and 4) for students with large at-risk populations, dividing students into smaller cohorts to better monitor their performance and response to interventions improves the likelihood a student will graduate. One article, in particular, was especially informational as it not only covered related publication in this space, it also provided the reader with a systematic review of dropout system performance dating from 1980 to 2012 (Bowers, Sprott, & Taff, 2012).

While these articles provided a sound basis for initial inquiry, I attempted to improve on the existing comprehensive literature by conducting an expanded search for research in this area related to educational data analysis, educational data mining, and learning analytics. After reviewing the works in these related articles, a search was done using various combinations of the following keywords: “Early Warning Dropout Systems,” “High School Dropout,”

"Predicting High School Dropout," "Predicting High School Graduation," "Learning Analytics," "Educational Data Mining," "Educational Data Analysis," "Machine Learning," and "High School Predictive Modeling" within Google Scholar, a web search focused scholarly article aggregator. The primary databases queried within the Columbia University Library system were the American Psychological Association (PsycINFO), Eric (EBSCO), and JSTOR articles databases, which resulted in a review of 198 articles.

## 2.2 Learning Analytics & Educational Leadership Data Analytics

Understanding the role that data plays within education enables leaders and practitioners to better make decisions within schools (Bowers, 2008; Mandinach, Honey, Light, & Brunner, 2008). Traditional methods of data analysis and educational technology have become more advanced in recent years, creating new classifications of educational research such as learning analytics, academic analytics, and educational data mining (Romero & Ventura, 2010; Siemens & Long, 2011). These roles leverage a similar data model (Figure 1) in different ways to better inform the many stakeholders within our educational system.

For example, academic analytics utilizes data to gain insight at the institutional, regional, national or international levels which will better inform administrators, funders, governments, educational authorities, researchers and analysts (Agasisti & Bowers, 2017) whereas educational data mining is generally used to understand learning at the course or institution level to better inform researchers, analysts, faculty, and tutors (Agasisti & Bowers, 2017). Lastly, the learning analytics field sits between the data miners and academic analytics in that it generally utilizes data to gain insight at the course and institution level in order to inform learners and faculty with their decision making (Agasisti & Bowers, 2017). EWS research has resided within both the educational data mining and the learning analytics categories as it focuses on using data to

inform educators and administrators to better understand student dropout within their population, with the overall goal of reducing risk and improving overall student outcomes (Aguilar, Lonn, & Teasley, 2014; Krumm, Waddington, Teasley, & Lonn, 2014; Lonn, Aguilar, & Teasley, 2015)

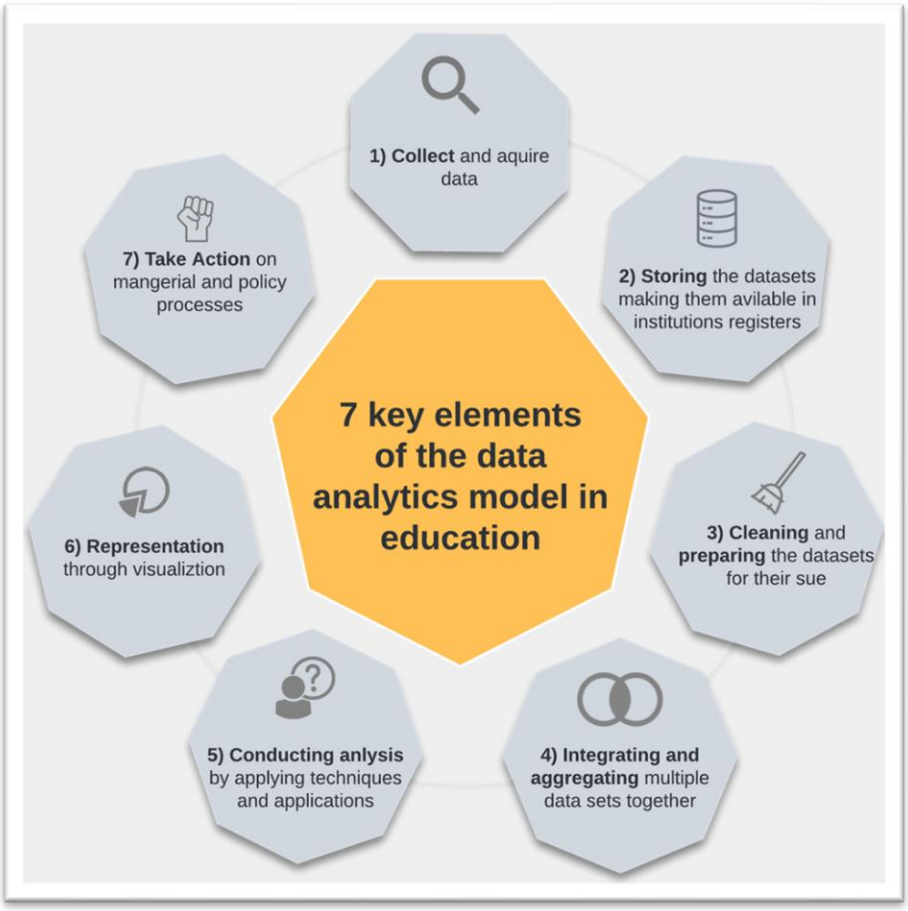



**Figure 1: Data Analytics Lifecycle in Education**[1]

Additionally, over the past several years, there has been significant progress in enabling data-driven analytics within the educational setting (Bowers, Bang, Pan, & Graves, 2019), particularly with machine learning-driven decision making (Bowers, 2017). Combining these advancements in data analysis techniques within existing Educational Data Analytics

[1] Source: Authors' elaborations, originally inspired by Siemens (2013) and adapted from Agasisti & Bowers (2017).

frameworks to identify a student at risk of dropout directly enables educational leaders to identify areas of opportunity for the educational system and student outcome improvement (Bernhardt, 2004; Bowers, 2021). Insights driven from this process can inform educational leadership in several different ways, ranging from identifying areas of opportunity for more successful applications of targeted student interventions (Pinkus, 2008; Kennelly & Monrad, 2007), identifying better resource management for dropout risk mitigation (Heppen, & Therriault, 2008), and ultimately, informing district leaders with the insights needed to implement effective whole-school reform focused on drop-out prevention that improves equity among underserved populations within their educational settings (Mac Iver & Mac Iver, 2009).

## 2.3 Early Work on Predictors of Dropout

Early research on dropout prediction analyzed several areas of student data and their relationship to high school success. These areas can be loosely summarized into four primary categories: Academic, Attendance, Behavioral, and Identity. Academic data encompasses school marks or teacher provided ratings of student achievement (Marsh & Yeung, 1997). Examples of indicators in this category are student course/semester/yearly grade point averages (GPA), summative or interim assessment scores, course-failure rates, and course credit accumulation. Attendance data reflects information regarding student absenteeism and participation within their educational system, which is often recorded as whether the student was physically present at the school on a given day or whether the student was late to report to class at a given time (tardiness). The behavioral category focuses on observed student actions and interactions, such as the perceived social status among their peers rated by school counselors and educators (e.g., popularity, friendliness, involvement in social groups, etc.), aggression, or anti-social activities.

Lastly, the category Identity is related to student demographics and covers factors such as age, sex, ethnicity, or urbanicity.

### 2.3.1 Academic Indicators

Initial dropout analysis research focused on data related to student academic performance indicators such as course GPA, test scores, or class failure rates (Barrington & Hendricks, 1989; Bearden & Spencer, 1989; Finn, 1987; Pallas, 1985; Rumberger, 1987). Early research on identifying academic predictors for high school dropout was published in 1983 with the presentation of Curtis' paper titled Dropout Prediction, at the Annual Meeting of American Educational Research Association (1983). Using high school student data collected from 1977-1981 from Austin, TX public schools (n=5,039), Curtis developed a dropout prediction model using discriminant analysis on 60% of the data and evaluated on the remaining 40%. Variables utilized within the model were collected from the school district's student information system and consisted of five specific items: student GPA, grade placement (grade in which the student was enrolled), sex, ethnicity, and the number of serious discipline problems. Student outcomes were classified into four groups: non-leavers, transfers, dropouts, and other/unknown. Special education students were omitted from the analysis. The resulting model was able to accurately predict 78% of the students that did not graduate high school. The initial results from this analysis revealed that "=students who have low GPA's, who are behind in grade for their age, who have been involved in serious discipline incidents, who are female, and who are non-Black have a higher than average probability of dropping out" (Curtis, J., 1983). Curtis found these findings to be puzzling, as the data showed that male students had a higher rate of dropout. After additional analysis, he concluded that overall, males had a higher probability of dropping out but

females that exhibit certain characteristics not specifically recorded in the data (such as teenage pregnancy) were more likely than males to leave school.

A study conducted three years later by David Doss expanded on this analysis using a similar approach (1986). He conducted a discriminant analysis on GPA, grade placement, sex, ethnicity, and the number of serious discipline problems to identify students (n=649) who were at the greatest risk of dropping out within the study sample (n=3028). Once these students were identified, a second analysis of their course registration was conducted. This analysis revealed that classes could be classified as either "above" or "below" holding power (i.e., the likelihood a student will stay in school). Classes with an above-average holding power included Spanish, introductory algebra, world history, dance, photography, biology, drawing and painting, and varsity sports. Courses with below-average holding power included drama, Spanish for native speakers, fundamentals of mathematics, field sports, and electronics. On the surface, these results suggest that the subject area of a student's course enrollment can be predictive of whether she is on track to graduate.

In 1986, using student (n=3,000) surveys, subject-specific achievement test scores, and demographic variables, Ekstrom, Goertz, Pollack & Rock constructed a path analysis model to investigate causal reasons as to why students drop out of high school (1986). Estimates derived from the path analysis were compared to estimates produced by a second propensity score analysis to verify results. Lastly, the authors conducted a third value-added analysis on the impact that test achievement gain has on student outcomes. The findings from the path analysis suggest that school grades and student behavior are more explanatory for dropout behavior than other variables used in their analysis. The value-added analysis found that females and minorities were impacted the most from unrealized achievement due to dropping out of high school, with

these two groups “falling the furthest behind in language development, vocabulary, reading, and writing when they leave school early” (Ekstrom, Goertz, Pollack & Rock, 1986).

In addition to individual course performance, information regarding student grade retention has shown promising results in identifying dropout risk. Using Logistic Regression and Survival Analysis, Melissa Roderick examined the relationship between grade retention and the likelihood of graduating from high school. Her research suggests that even after controlling for external factors such as student background and school performance, students who repeated a prior grade were substantially more likely to never graduate high school, with students over the age of 16 at over double the risk of dropping out after repeating a grade. The influence of repeating a grade has on high school graduation is reduced at lower grades, students that were held back a year in kindergarten through third grade, not any more likely than their non-retained counterparts to drop out of high school (Roderick, 1994).

### 2.3.2 Attendance Indicators

While it’s clear that academic performance is an important metric in evaluating a student's overall achievement in a course, term, or year, unsurprisingly, this information is only reliable if the student is physically present in the school to be evaluated. This presents several problems as student attendance can fluctuate for many different reasons. For example, home life issues such as lack of residence or chronic homelessness (Epstein & Sheldon, 2002; Mawhinney-Rhoads & Stahler, 2006), medical illness, negative peer influence (Hartnett, 2007), lack of student interest or engagement (Legault, Green-Demers, & Pelletier, 2006) and student mobility (Dunn, Kadane, & Garrow, 2003) can all contribute to a reduction in student participation (Hocking, 2008). Various studies have been conducted that suggest this data can provide useful

insight into the trajectory of a student's likelihood of graduating high school using data from grades as early as elementary or middle school. (McKee & Caldarella, 2016).

Recent studies focusing on attendance patterns and their potential to impact a student's long-term academic high school outcomes have suggested these indicators to be significant to early identification of student at-risk status. Research completed in 2012 by Schoeneberger using a group-based trajectory structural equation model analyzed twelve years of student' records (n=286,529) within a large urban school district in the southeastern United States. The results of this research found students could be grouped into four distinct groups: (a) Constant Attendee which represented students who consistently attended school, (b) Developing Truants, representing students who historically had constant attendance but had recently began to show indications of truancy, (c) Early Truants which consisted of students that were once Constant Attendees but were now consistently truant in school attendance, and (d) Chronic Truants which represented students that have historically and currently been absent from their school setting. These findings suggest that these four attendance related groups differ in terms of eventual high school dropout (Schoeneberger, 2012).

### 2.3.3 Student Behavioral Indicators

A substantial amount of dropout research has focused on student academic performance, attendance, and demographic indicators. While these data points have shown to be important factors in identifying dropout (Suh, Suh, & Houston, 2007), little research has been completed on understanding how social contexts can impact student educational outcomes (Hartnett, 2007). Barbara S. Mensch and Denise B. Kandel explored the relationship between substance abuse and high school dropout (1988). To conduct this research, they built a discrete-time logistic regression on variables related to the use and abuse of specific substances. These included the

age of initiation at which a substance or behavior (cigarette use, marijuana use, other illicit drug use, alcohol use, and did not use) was first exhibited, the age of initiation for each individual substance or behavior if students used various substances, and whether or not the student eventually dropped out or completed high school. Due to computational cost limitations, the study sample was downsampled to represent 30% of the original dataset. The analysis was based on a youth cohort sample of the National Longitudinal Survey of Youth-1997 (US Bureau of Labor Statistics, 2002) (NLSY), representative of individuals born in 1957-1964. This cohort was interviewed manually in 1984 regarding various aspects of their life, including sexual activity, alcohol consumption, and pregnancy, and exposure to violence (number of school fights) history. The results of this research found that substance abuse and deviant behavior increased the probability of not graduating high school. The researchers conclude that if these influential factors can be mitigated, the possibility exists for an improvement in overall student achievement outcomes in the form of successful high school completion.

While initial research into the relationship between behavioral data and dropout focused on negative substance use, researchers soon began to expand the breadth of their analysis to include social, teacher, and peer reported behavioral data. Research by Kupersmidt and Coie investigated the role of peer status, aggressive behavior, and school adjustment that influences a student's likelihood of achieving a high school education (1990). To accomplish this, the researchers selected a (n=112) cohort of 5th graders and followed them for 7 years. They then collected data related to SES, aggressive behavior, and school adjustment as well as the student's high school academic outcomes. They built a series of logistic regression models to test their hypothesis. Results from this analysis found two significant predictors of dropout: peer-perceived aggression and an excessive number of school absences. Students that were

excessively absent were found to be 27% more likely to drop out, students that were aggressive were 45% more likely to drop out, and students that were both aggressive and frequently absent were 73.7% more likely to not complete high school compared to the reference group (Kupersmidt & Coie, 1990). Student social behavior, in particular, has proven to be useful in identifying students at risk of dropping out of high school. A 2001 study conducted on (n= 516) 8th-grade students and (n=1157) 10th-grade students looked at the impact that anti-social behavior and peer rejection have on a student's likelihood of dropping out of school. Using logistic regression analysis, the researchers found evidence that suggests antisocial behavior and rejection may lead to heightened levels of student risk (French & Conrad, 2001).

Analysis of student social activities has continued to be a topic of research in the field of high school retention. Using longitudinal cluster analysis, Joseph L. Mahoney investigated the impact of social, extracurricular activity participation on a student's development of anti-social patterns and eventual academic and life outcomes. To accomplish this study, students were interviewed in the fourth or seventh grade and tracked until twelfth grade to determine their academic outcome. Participants were then interviewed twice at both 20 and 24 years of age. The interview questions covered items related to the interpersonal Competence Scale, physical maturation, extracurricular activity involvement, socioeconomic and demographic information, social networks, early school dropout, and criminal offending. Cluster analysis was then used to identify patterns within the cohorts of study, which were then compared across groups based on gender, educational outcome, and criminal involvement. Results from this study found that a student's involvement in extracurricular activities was correlated with lower rates of dropping out of school or becoming involved with criminal activity as adults. Additionally, Mahoney also

found evidence that suggests student risk of antisocial behavior for both boys and girls was reduced by peer social network interactions and in school activities (Mahoney, 2000).

### 2.3.4 Student Identity

Lastly, there is evidence to suggest that a relationship exists between a student's demographic characteristic, such as urbanicity, and their eventual educational outcomes (Adelman, 2002), with school-level variables such as socioeconomic status or school size showing significant results in a student's eventual educational outcome (Wood, Kiperman, Esch, Leroux & Truscott, 2017). Research has also suggested that rural student dropouts may differ statistically from dropouts in suburban/urban schools in several ways. For example, rural dropouts are often cited as leaving for reasons such as pregnancy or marriage, whereas urban students are cited as dropping to enter the workforce so they can better support their current family. They are also cited as dropping out of high school because their peers are leaving the educational environment. Additionally, when conducting analysis on a student's attitude towards the general school conditions, urban students were more likely to rate their school higher than rural students. This analysis suggests that rural students were less likely to get along with their instructors compared to their urban student counterparts, which could be an influencing factor in their decision to drop out (McCaul, 1989).

Ensminger and Slusarcick conducted a longitudinal analysis of black first-graders over the course of 12 years. They selected 1,242 first grade students from an urban community who were classified as high risk for dropping out of school and collected several measures around their family background, school behavior, academic performance, and parent-child interactions concerning school, educational values, and expectations. At the conclusion of the 12 years, a final measurement was made on the sample that collected data on whether the student

successfully graduated or dropped out of school. Using this information, the authors built a logistic regression model to determine the likelihood each coefficient has on whether the student will drop out or graduate. Their results from this analysis suggest that student poverty played a crucial role in student risk, with the link between early school academic performance and high school graduation decreasing for students who were not considered poor. Their findings also imply that there is a generational link between parental academic achievement and the likelihood a student will graduate (Ensminger & Slusarcick, 1992).

Factors that are external and often unreported (at least, to the schools) can influence high school dropout. For example, research by McNeal examined the relationship that student employment has on dropping out of high school. Students sampled from a high school in 1980 were surveyed regarding their employment, the field of employment, hours employed, and academic performance. These students were then followed for 2 years to determine if they dropped out or successfully graduated from high school. Logistic regression was then utilized on the variables of interest to determine the odds of a student dropping out versus graduating. Results from this analysis suggest that the type of student employment and the intensity at which they are employed significantly impacted their trajectory in high school. McNeal Jr. also found that the effects of employment were contingent on the student's gender (1997). While results showing an association between teenage employment and graduation outcome have generally been replicated by educational researchers, there remains the question of whether employment is truly causal in determining high school outcomes or whether these are spurious findings resulting from other non-observed factors such as student socioeconomic status or general aspirations. Attempts to answer this question have produced evidence to support the latter. A study that utilized a propensity score matching design on nationally representative longitudinal student

survey data to model the effects of after-school paid work intensity on the probability of dropping out found that there was no significant correlation between the number of hours worked by a student and their likelihood of not graduating (Lee & Staff, 2007). This study suggests that modeling teenage employment intensity as a single factor producing high school dropout is insufficient to explain dropout, and that researchers need to account for possible external effects on employment intensity by identifying student factors such as socioeconomic status. These early studies on identifying predictors of dropout would go on to provide the foundation for the creation of the Early Warning Systems in use today and enable researchers to identify the relevant data points that are included in the design of previous and current systems.

## 2.4 Threshold Based Methods

Traditional research into high school dropout has provided a wealth of information about the many factors that can impact student success. While these studies are useful in interpreting the relationship of specific academic, behavioral, attendance, or identity variables, they also provided researchers the opportunity to develop predictive systems based on the findings of this work. These systems are designed with the intention to identify students prior to their dropout event occurring, allowing educators the ability to apply prediction- driven interventions rather than traditional prescriptive interventions. The foundation of these systems is built upon the traditional research conducted over the past several decades, with the first of these systems relying solely on insights generated from these early analyses.

Deploying dropout identification systems can require a significant amount of resources to test, build, and deploy a predictive model within a school district (Frazelle & Nagel, 2015). Districts that face resource and funding constraints often have to rely on simpler methods of early at-risk detection (Balfanz & Byrnes, 2019). These methods rely less on statistical rigor, and

more on the ease of implementation and on how understandable they are for school districts' employees (i.e., administrators and educators). While the approach to these detectors may seem facile in comparison to more advanced methods of statistical data modeling (which will be discussed later), their simplicity enables non-technical users, such as educators and guidance counselors, to understand the inner workings of the detector which is the primary reason why these types of EWSs still remain widely popular and in use today, despite the often heuristic approach to their design.

In some cases, simple heuristic early-warning systems have been mandated by state legislatures. House Bill (H.B.) 1010, passed by the Texas State Legislature in 1986, attempted to reduce the number of dropout students within the state by providing educators with indicators that can be used to classify students (Frazer, 1991; Supik & Johnson, 1999). The bill was specific to students within grades 7 through 12, with earlier grade students omitted from risk classification. In order for a student to be flagged as high-risk, they must either 1) not have advanced from one grade level to the next in two or more school years, 2) have mathematics or reading skills that are two or more years below grade level, 3) not maintain an average of 70% in two or more registered courses, and 4) not obtain a satisfactory score on the state-mandated end of year exams. This EWS was the first to be mandated at the state level and scaled across all relevant schools throughout Texas, with the eventual performance providing mixed results on the capacity of this generalized EWS's effects on reducing the number of high school dropouts.

While such systems are easy to implement and understand, they can be inaccurate at identifying students who are at-risk. A study to evaluate the performance of this system was conducted by the Austin (Texas) Independent School District (AISD). This research focused on 25,587 students from 1987-88, 25,292 from 1988-89, and 25,998 students from 1989-90 who

were in grades 7-12. Using the state-mandated Texas at-risk definition, AISD assigned these students to relevant risk groups, and then evaluated their performance three years later after a true outcome (whether the student graduated or dropped out) was recorded. The results of this research found that; (a) the classification accuracy of the Texas legislative at-risk system grossly over labeled students as at-risk who did not eventually drop out of high school with approximately 87% of students across all 3 years of study classified as at-risk of not completing their high school education, (b) students with lower risk in year one were nonetheless more likely to graduate than high-risk students, (c) and that students who are in the high-risk category in their first year are more likely to grow in risk throughout the subsequent years (Frazer, 1991). Frazer's research reveals that while threshold-based EWSs are easily deployed within a school district, they are prone to significant classification errors.

The implementation of Texas H.B., while not as successful as one would have hoped, did reveal civic, legislative interests in adopting some form of at-risk dropout detection. Addressing this need led to the eventual design of more advanced methods of threshold-based systems, such as the Chicago model, mentioned earlier in the introduction of this proposal. The Chicago model is similar to Texas H.B. 1010 in that it relies on simple cut points to determine the student's risk status; where it differs is how those cut points were generated. The Chicago on-track indicator, developed by Allensworth, utilized two primary indicators that focused on a student transition through 9th grade, an important milestone in a student's high school career (Easton, Johnson, & Sartain, 2017). The first indicator is the accumulation of course credits, and the second is whether or not the student has failed at least one core course in their ninth-grade year (Allensworth, 2013; Allensworth & Easton, 2007). The cut-point values used to determine student at-risk status was based on several studies conducted by the consortium beginning in the

1990s (Roderick & Camburn, 1996; Miller et al., 1999) that suggested there was a strong correlation between course failures and credits earned to the likelihood a student will graduate high school (Miller, Allensworth & Kochanek, 2002).

On the surface, the performance of the on-track indicator proved to be widely adopted, with a significant amount of schools nationally utilizing the on-track indicator as an accountability measure with varying levels of success. Several researchers have reviewed the performance of the indicators used in the on-track metric and compared to other commonly used drop-out indicators and found that the on-track EWS outperforms many of its competitors (Bowers et al., 2012; Bowers & Zhou, 2019a; Hoff, 2019). While these findings, coupled with the on-track indicator's high adoption rate in schools, suggested promising results, recent research suggests there still exist several limitations in its performance. A 2019 study found that implementing an Early Warning Intervention (EWI) model, used to monitor ninth-grade indicators in an attempt to modify student behavior and based off of the On-Track EWS in 41 geographically and demographically diverse high schools, showed no statistically significant impact on overall student performance for 9th-grade students in regards to either attendance or credit accumulation (Mac Iver, Stein, Davis, Balfanz, & Fox, 2019). The authors believe that this lack of significance was due to the research and best practices for ninth-grade interventions already having been disseminated; that is to say, the EWI processes and procedures for intervention had become common knowledge among educators, regardless of whether the school had a designated program in place. At this point then, the On-Track indicator may not capture enough indicators to make a difference compared to the knowledge that now exists among teachers and administrators. Alternatively, the On-Track indicator may not be effective once it is taken out of its initial setting of development and closely related schools.

Building on top of the University of Chicago Consortium's On-Track indicator (The Chicago model) work, the American Institute of Research (AIR) launched a threshold-based EWS spreadsheet that districts could use to identify students at-risk of dropping out (Heppen & Therriault, 2008). This tool, similarly to the On-Track indicator, focused on the performance data of students in the 9th grade. In addition to the course credits and course failure indicators, they also looked at student attendance and overall student GPA. While AIR's work expanded the range of indicators used in a threshold-based system, validity analysis of this EWS suggests that it largely performs the same as the On-Track indicator when it comes to identifying students at risk of dropping out (Bowers et al., 2012; Bowers & Zhou, 2019a; Johnson & Semmelroth, 2010).

While the continued use of these threshold-based Early Warning Systems suggests that there is a demand for simple methods of detecting student risk, their inability to take localized trends into consideration when making a risk prediction diminishes their ability to make meaningful predictions that identify not only students at-risk, but identify the areas most susceptible to positive early intervention. Determining which predictors are important for each school or district is still an active area of research as we begin to consider both the regionality and population diversity within school districts (Bowers, 2010). Threshold-based approaches lack the ability to account for these identifying factors when determining relevant predictors and thresholds, which presents a serious flaw in their design. Implicitly, these models lead individuals viewing the results to make decisions based on the inaccurate assumption that all students are the same, regardless of external factors or regional localities.

## 2.5 Machine Learned Early Warning Systems

Until the 2010s, if a school district employed an EWS, it was derived from these threshold-based methods. It wasn't until recently that emerging methods of dropout detection began to utilize advanced methods of data modeling to detect student risk of high school dropout. These modeling solutions rely on the aggregation of multiple student data sources, which were once inaccessible due to siloing. These new solutions largely have become available through school districts' recent adoption of data management systems, coined Student Information Systems (SIS), which are specifically created to store student records (Halverson, & Smith, 2009). These systems store both past and present student academic performance records, attendance records, behavioral data, demographics, attendance, and test scores all in one location. With access to this wealth of information in one place, researchers are now capable of utilizing modeling methods that require significant amounts of historical data, such as machine learning, to create more accurate risk detectors.

The state of Wisconsin was one of the first major adopters of such a system. In 2012, they created and deployed to all schools The Wisconsin Dropout Early Warning System (DEWS). This EWS provides over 225,000 at-risk predictions and is focused on identifying sixth through ninth grade students at risk of failure to graduate on time (Clune & Knowles, 2016). This EWS utilized an advanced statistical method that scans through 35 different analytical techniques and selects the best models by building and evaluating performance with each solution. It then takes an ensemble approach and combines the best models to generate the final detector. This approach performs better than previously developed solutions, with a dropout detection accuracy of 65% on students before they enter high school (Knowles, 2015). The performance of this

EWS at such a large-scale proved promising for research to continue developing such advanced solutions that are not only accurate but can also generalize.

While there is value in the ability to analyze diverse amounts of aggregate student data, regardless of whether or not this data can be modified through intervention, there still remain opportunities to leverage a small subset solely consisting of actionable student data to build advanced EWS solutions. For example, instructor-assigned academic course achievement measured over time has proven to be a valuable indicator for early identification of high school dropouts. This is evident in one of the more recent studies conducted by Bowers and Sprott (2012a). This now pivotal study utilized a Structural Equation Modeling approach (Anderson & Gerbing, 1988; Hoyle, 1995; Russell, Kahn, Spoth, & Altmaier, 1998) known as Growth Mixture Modeling (Muthén, 2001; Wang & Bodner, 2007) on a 2002 nationally representative data set (n=5400) (Ingels, Pratt, Rogers, Siegel, & Stutts, 2004) to identify students at risk of dropping out. The researchers focused on two primary questions; a) measuring the influence that non-cumulative GPA for 9th-grade students has on their overall likelihood of dropping out of high school and b) dissecting definitions of student dropout classifications (dropout typologies). Bowers and Sprott found that they were able to identify 91.8% of the dropouts using only the non-cumulative GPA indicator (measured over 3 semesters). They also found evidence to support that rather than one binary category of either graduation or dropout; there are four latent levels of dropout trajectory (the four trajectories are Mid-Decreasing, Low-Increasing, Mid-Achieving, and High-Achieving). The researchers found that the variables impacted the dropout trajectory differently for each typology, leading the researchers to conclude that understanding the different types of dropout typologies could better enable schools to provide better, more personalized interventions for students. This research still remains one of the best performing

models of high school dropout risk identification within the literature (Bowers & Zhou, 2019). A follow-up study conducted by the same authors, utilizing a latent class analysis method, was able to identify the remaining 9% of student dropouts as "lost at the last minute" or "involved." *Lost at the last minute,* encompassed students with decreasing GPA trajectories and *involved* consisted of students that were more similar to graduates but ended up not completing their high school graduation due to a mistake in their transcript, not knowing they needed to take a class, or a major life event, such as pregnancy or a sudden move or life change (Bowers & Sprott, 2012a; Bowers & Sprott, 2012b).

Utilizing machine learning approaches allows the researcher to let the algorithm determine the value of model variables within the detector. This enables EWS design to be deployed across a wide range of variables, making use of any data available rather than relying on a limited set of specific predictors. This is especially important if the intention is to produce predictions at earlier grades as teacher-reported academic performance becomes less available and standard (GPA not recorded, credit system not implemented, etc.). For example, researchers in the State of Florida were able to build a dropout detector for 1st and 2nd-grade students using interim and summative assessment scores when GPA data was not available (Koon & Petscher, 2015).

Not only are methods of model building enabled through advanced statistical methods, but opportunities exist for addressing data issues used to train these models. For example, a recent study aimed to address class imbalance (i,e., the number of dropouts and graduates are not close to equal within the data) by applying an advanced technique called synthetic minority oversampling techniques (SMOTE), which generates new data records using existing records (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). This can be especially beneficial for school

districts that have low dropout rates or lack dropout records and wish to build their own EWS. Recent research using the SMOTE technique on educational data has shown mixed results with improving classification accuracy of the (often dropout) minority (less common) class. This is due to a phenomenon where classifiers trained on class-imbalanced datasets tend to show a poor sensitivity of predicting minority classes because classification algorithms tend to weight the misclassification of minority classes lower than the misclassification of the majority class. Building off of these findings, researchers in 2019 attempted to use the SMOTE technique on a large number of student historical records (n=165,715) and then built detectors with machine learning tree-based algorithms. They found that implementing SMOTE impacted the detectors' true positive rate (ability to classify dropouts) most positively and true negative rate (ability to classify graduates) most negatively, with ROC AUC values dropping for models that utilized SMOTE (S. Lee & Chung, 2019). These findings suggest that even advanced statistical applications within an EWS can present researchers with similar challenges faced by implementations of simplified threshold-based systems and that striking a balance between over classifying students not at-risk or under classifying students at-risk is a continuing area of active improvement in the community.

While advanced statistical models have shown promising results with improving dropout detector accuracy (Bowers et al., 2012; Bowers, 2021), there exist limitations in their interpretability for stakeholders attempting to utilize these insights. The inherently complex nature of machine learning algorithms makes it difficult for educators to interpret results into actionable interventions. A researcher may classify a student as at-risk, but understanding that the risk is associated with specific indicators is essential for determining the appropriate interventions. Fortunately, in addition to the recent uptick in advanced statistical EWS research,

there is a growing number of companies, non-profits, and researchers focused on providing EWS tools to schools within the educational technology industry that assist educators with better understanding and interpreting these models (McIntire, 2004). For example, Infinite Campus, an educational technology company, recently published results from their EWS dashboard product where they provide users with domain level insights in addition to an overall student risk prediction. They found that building individual domain-specific (academic, attendance, behavior, etc.) machine-learned models produced highly predictive context-specific results, achieving an average AUC score above 0.86 for $6^{th}$ – $12^{th}$ grade student predictions (Christie, Jarratt, Olson, & Taijala, n.d.) using a combination of four separate educational domain models, resulting in improving the actionable outcomes of these insights by educators and relevant user stakeholders.

Organizations such as the American Institute of Research (AIR) have partnered with these educational technology companies to help districts better support students that are showing early indications of dropout (O'Cummings, & Therriault, 2015). Through a systematic process, they provide educators with the resources to 1) establish roles and responsibilities, 2) review early warning data, 3) correctly interpret this data, 4) assign appropriate interventions, 5) monitor the students' intervention progress, and 6) evaluate the effectiveness of these interventions to better refine future processes (Therriault, O'Cummings, Heppen, Yerhot & Scala, 2013). While this partnership has improved the design of the tools offered by the educational technology companies, it is limited to school districts that have the financial capacity to purchase these tools and resources. This has led to a division among school districts, where some districts are able to provide educators with the necessary support to better understand the data, through both technology (EWS, Dashboards, visualizations, etc.) and professional development and some districts are not. This presents a significant challenge as these resources are paramount to help

mitigate and reduce dropout risk within our educational communities (Dwyer, Osher, & Warger, 1998; Dwyer, Osher, & Hoffman, 2000).

## 2.6 Generalizing Models

One of the primary reasons threshold-based EWS remains so prominent in use throughout schools today is their ability to generalize across schools due to their overall design simplicity. The same cannot be said for models that rely on advanced statistical techniques. Generalizing machined learned models across domains or populations has shown to be a challenging accomplishment (Ocumpaugh, Baker, Gowda, Heffernan, & Heffernan, 2014). This presents limitations for those districts that want to leverage these more accurate detectors but lack the ability to build their own context-specific models. This challenge of implementing advanced solutions within education is a commonly discussed barrier within the field of educational data mining and learning analytics (Baker & Koedinger, 2018; Niemi, Pea, Saxberg, & Clark, 2018).

Overcoming these challenges may seem daunting, but researchers have made positive gains in understanding methods for replicating a standard predictive modeling method across populations (Gardner & Brooks, 2017). A recent systematic review of research found that a primary way to mitigate these issues and improve generalizability across populations is by taking contextual factors into consideration (Joksimović et al., 2018). These findings suggest that a student's individual identity and external factors of their environment play an important role in determining their eventual educational outcome and that in order to move forward with developing successful EWSs, we need to better evaluate and understand their performance as they are deployed across an ever-widening range of student populations.

## 2.7 Evaluating Early Warning Systems

Auditing the performance of these detectors is an important step in validating and measuring their success in real-world contexts (Sullivan, 2017). Prediction can sometimes fail, with students still slipping between the gaps of detection. Previous research has shown the machine learning-based systems can bias their predictions in unforeseen ways, which can cause more harm than good (Sansone, 2019). A high-performing EWS may be performing exceptionally well on the surface, but when researchers evaluate the performance within specific subgroups, they have, in some cases, found that their model was biasing dropout towards students with specific demographic characteristics, such as gender (Pagani et al., 2008). Despite their decision to expressly exclude this feature in the initial model, the data still contained latent information that caused the model to skew towards this specific group.

Prediction bias can manifest in different ways presenting challenges to researchers working to improve the trust and accuracy of the EWS. For example, over-prediction of certain groups (ethnicity, gender, etc.) can highlight or propagate already existing discriminatory practices within the school environment (Catterall, J. S., 1998; Huysamen, J. E.,1999). Differential model accuracy between certain groups where the model performs well overall but underperforms when evaluating at the subgroup level can reduce value and benefit for already underserved populations (Soland, J., 2013). Bias towards features that are obtained at different times within education, such as college placement exams, could over classify students at risk who are interested in vocational or non-traditional post-secondary pathways (Patrick, L., Care, E., & Ainley, M., 2011). This model biasing presents a severe obstacle to overcome if these systems are to be implemented within our educational system. Addressing these challenges is

paramount to generating an effective generalizable EWS. Therefore, it is essential to select the appropriate evaluation metric when validating EWSs.

# Chapter 3: Methodology

In the following sections, I discuss my method for making at-risk student predictions for school districts with insufficient data. These predictions incorporate the unique demographic and local attributes of each district to generate the final student risk values. This model is termed District Similarity Ensemble Extrapolation (DSEE). In addition to the design of the DSEE, I discuss my methods of validating the performance of my modeling solution both internally and externally, which occurs by calculating metrics of performance during model creation and by comparing my model against three well established existing methods through the replication of their design.

## 3.1 Criterion for Building the District Similarity Ensemble Extrapolation Model

Building an EWS driven by machine-learned methods presents several challenges related to data. A sufficient quantity of historical labeled data must be available in order to create a machine-learned model (Byrd, Chin, Nocedal, & Wu, 2012; Stockwell & Peterson, 2002). Research has shown that these types of models perform better when built using rich datasets (i.e., the data contains enough representative qualitative and quantitative data to reveal the complexities of what is being studied) and perform worse when quality issues exist in the data (Cortes, Jackel, & Chiang, 1995). Additionally, in order to build an EWS that can provide risk predictions down to the first-grade student level, a sufficient number of historical records (12 years) are required to not only build the model but validate its performance at the many grade levels it will be utilized (Žliobaitė, Bifet, Read, Pfahringer, & Holmes, 2015). Several issues exist without this historical data: we are unable to accurately determine how the EWS is performing for these lower-grade students; interpretability is reduced, limiting opportunities for

intervention; and the ability to mitigate latent biases that the model may be producing (Cawley & Talbot, 2010).

A study conducted by Coleman, Baker & Stephenson (2020) focused on building an EWS model for students in the 1st through 12th-grade levels. Using academic, attendance, behavior, and assessment data collected from a nationally representative (n =3,575,724) sample from 34 diverse U.S. K-12 educational systems (one large educational agency with decision-making power over a large geographical region, and 33 individual school districts) found data quality to be a significant barrier for building high performing machine-learning driven district (or education agency) level EWS models. Their research concluded that of the 34 systems in their sample, only four had nearly complete data (with only small numbers of variables). The remaining 30 districts suffered from three major data quality issues related to missingness, 1) a high degree of missingness in the feature data, 2) a deficiency in the number of records available that span 12 years down to first grade, and 3) a low number of unique student records. The percent of feature data missingness within these agencies and districts was as high as 60% in some cases ($M$= 41.65, $SD$=7.498). The majority were also missing 100% of the data for students at lower grade levels (primarily elementary), and most of these systems contained a low number of total records across all students (less than 20,000 records), with some having as few as 271 total historical student records available for analysis. When the authors attempted to fit district/agency level models on the districts that suffer from these data quality issues, the performance was suboptimal, with no model obtaining an AUC above .70, further supporting the importance of data quality when building an advanced EWS solution (Coleman, Baker, & Stephenson, 2020).

Recently, the educational community has made progress in attempting to address these educational data quality issues. One such initiative is the U.S. Department of Education's creation of the Common Education Data Standards (CEDS), a collaborative effort to develop common data standards for key sets of educationally related data elements which include standard definitions, option sets, and technical specifications to assist educators with sharing, analyzing, and comparing information within their system (Common Education Data Standards, 2019). In addition to the data standards, the CEDS initiative has partnered with other organizations to provide education stakeholders with the tools they need to understand their data. One particular organization, the Ed-Fi Alliance, is a nonprofit focused on assisting school districts and states reach data interoperability by aggregating disparate educational data collected on students into one standards-aligned unified datastore (Alliance, 2015), enabling educational stakeholders with the ability to conduct robust, in-depth analysis on their student population. While these advancements of standards and tools greatly improve opportunities for data-driven EWS analytic enablement within schools and districts, there remain challenges.

With these standards just now coming online, educators will need to dedicate a significant amount of time and resources to adopt these changes, which could encourage them to only apply these standards to current and future students, rendering their historical data incompatible for modeling future student risk. Additionally, districts that simply do not have 12 years of digitally stored historical data on their past students, and lack the resources to sift through and manually store analog paper files and records, would be faced with the challenge of first having to implement CEDS, and then waiting several years (potentially as many as 11) before they would have a rich enough data set to build their own EWS, putting them in the position to miss the opportunity of leveraging EWSs for early identification of at-risk students

reducing intervention success and overall student outcomes during this data collection period. The research in this document hopes to address some of these identified challenges and issues.

The design of my model solution involves first developing and validating a series of predictive analytics models for school districts with enough historical data (coined Pillar Models). These models predict each student's probability of graduating (or risk of not graduating). I then develop a distance metric capturing the degree of similarity between these validated models to school districts that lack sufficient historical data (coined Target districts) to build their own district-level models. Building on the research of Coleman, Baker and Stephenson (2020), districts are categorized as data deficient if they suffer from one or more of the following data quality issues; 1) Contain less than 20,000 records across all students (regardless of the total number of unique students in the district), 2) have over 40% of their feature space data missing, or 3) are missing historical records for all 1-12 grades . I then ensemble each of the existing models, weighting them by the similarity to the Target district, producing a single ensemble prediction for each student. I test the quality of this approach by applying this method on all the records that exist for the district, treating these records as the hold-out test set.

### 3.1.1 Model Validation

Data hold-out testing is a common method of validation implemented within the machine learning data science community (Schaffer, 1993). The advantages gained by utilizing this technique enable the scientist to better understand how the model will perform when generalized to the target population. This is accomplished by training the model on a subset (usually between 70% to 85%) of the original labeled data (i.e., the historical records of dropouts or students). The model is then applied to the remaining unseen data to generate the at-risk student predictions,

which can then be compared to the actual historical outcome of that student (the label) (Forman & Scholz, 2010). Model performance metrics can then be calculated to provide insight into how the model will perform when generalized to current attending students (Wong, 2015). Without conducting hold-out testing or cross-validation, the calculated performance could be misleading if the model is over-fit to the training data (Hawkins, 2004). Figure 2 below provides an example of an 80% train and 20% test data assignment during hold-out testing.

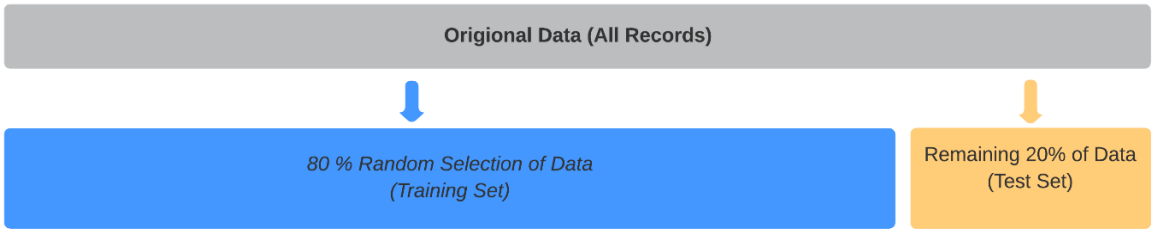



**Figure 2: Example of Data Assignment During Hold-Out Testing Model Validation**

Cross-validation is another method of data preparation that can be implemented to validate a model's performance. Like hold-out testing, cross-validation utilizes subsets of the data to build a model and then test the performance (Forman & Scholz, 2010). It differs from hold-out testing by using all the data available; this is accomplished by randomly assigning records in entire data set into a specified number of partitions (k-fold assignments), and then systematically training and testing across the data, utilizing a different partition for testing each time (Wong, 2015). While some researchers argue that cross-validation is more effective for model validation (Blum, Kalai, & Langford, 1999), there is debate on the number of k-fold assignments needed with the validation, with some researchers stating that 3-fold or 5-fold cross-validation is as effective as 10-fold cross-validation (Wiens, Dale, Boyce & Kershaw, 2008), while other suggest the higher number of folds the better the validation (Moreno-Torres, Sáez, &

Herrera, 2012). Cross-validation also presents a significant challenge to researchers working on large data sets, as the time and resources needed to validate the model increases with each additional k-fold assignment, requiring additional computational resources (Yadav & Shukla, 2016). Figure 3 below provides a visual representation of cross-validation.


**Figure 3: Example of 5-fold Cross-Validation Data Assignment**

Conducting validation tests using either method (hold-out testing or k-fold cross validation) not only makes it possible to measure model performance, it is also a good way to check for issues related to model over-fitting, area of concern when working with supervised machine learning algorithms (Hastie, Tibshirani, & Friedman, 2008). Model overfitting occurs when the model procedure violates the principle of parsimony, which states that the model should only use the information that is necessary to produce the prediction, and nothing more (Hawkins, 2004). Since machine learning algorithms generally employ methods that search for an optimal function that fits the provided training data (Dietterich, 1995), researchers are at risk of building a model that not only fits the relationship between the features but has also fit the

meaningless information (noise) in the training data by utilizing features in the data that do not correspond to underlying general patterns (Lever, Krzywinski & Altman, 2016).

Several techniques can be used to reduce the likelihood a model is overfitting. Increasing the size of the data can introduce new information, allowing the algorithm to better separate the useful features from the noise (Jabbar, & Khan, 2015). Tuning the model parameters to limit the maximum number of features the model can use during training can also lead to better generalization as it reduces the number of optimum functions possible in the search space (Sarle, 1996). Lastly, utilizing ensemble methods (where multiple models are trained and combined) such as boosting or bagging can help reduce overfitting (Ghojogh & Crowley, 2019). Boosting is where the model trains a series of weak constrained models, each one learning from the error of the model before it, and then combines them to create one final strong predictor (Vezhnevets & Barinova, 2007). Bagging is similar to boosting in that it builds a series of models and combines them; it differs in that it builds a series of unconstrained models (sometimes using different algorithms, sometimes using different subsets of the training data) to combine together with the hope of smoothing out the prediction error (Quinlan, 1996).

Reducing the likelihood of model overfitting is an important step when creating at-risk student prediction using an EWS. Without addressing over-fitting during the model building stage, issues can surface when attempting to generalize the model to any new data as over-fit models generally perform worse than a correctly-fit model, leading to error or bias in prediction (Kuhn & Johnson, 2013). This is caused by the model including predictors (learned during the training) that perform no useful function, which adds noise to the model, leading to misclassification prediction errors (Bramer, (2007).

3.1.2 Calculating District-to-District Similarity

The district-to-district similarity is calculated based on several properties (not utilized within the predictive model) that capture some of the key differences between school districts and the students within them. These indicators include information such as general student demographic ratios, grade enrollment distributions, and district graduation rates. The list of features is as follows:

**Table 1: Student and School Characteristics Used to Derive Similarity Scores**

| Student Demographics | Local Attributes |
|---|---|
| % of Students Classified as Pacific Islander | % 1st to 4th Grade |
| % of Students Classified as Native American | % 5th to 8th Grade |
| % of Students Classified as Multiracial | % 9th to 12th Grade |
| % of Students Classified as White | Total Students |
| % of Students Classified as Black | Avg Graduation Rate |
| % of Students Classified as Asian | Local Total Population |
| % of Students Classified as Hispanic | Urbanicity (rural, urban and suburban) |
| % of Students who District did not have Race/Ethnicity Data | Local Population Economics (employment rates, median income, etc) |

The selection of these indicators is based on two primary factors; 1) the availability of the data (i.e. the data was either collected and stored in the Clarity platform, or the data was publicly available from a secondary source such as the U.S. Census Bureau) and 2) evidence provided by existing research in this space. The selection process first involved reviewing the existing literature on high-school dropout and Early Warning Systems (see literature review above). After reviewing the literature, a list of potential population descriptive indicators was created. The final indicators were then isolated and selected for use in the DSEE model based on whether the data was available either from the Clarity system or from a reputable public source.

The district-to-district similarity is determined by a similarity score distance calculation based on Euclidean distance (Cha, 2007). This score is derived by computing how similar a district is to each of the districts for which a predictive analytics model is available. The higher the similarity between a new and Pillar district, the smaller the distance. Selecting the appropriate distance measure is an important step in calculating similarity, as it has a strong influence on the clustering results (Tan, Kumar, & Srivastava, 2002; Tan, Kumar, & Srivastava, 2004). Correlated-based distance measures such as Pearson's correlation assume that two separate feature sets share a linear relationship. While this measure can be advantageous when comparing data gathered on different scales across the feature set, it can be highly sensitive to outliers in the data, producing non-optimal results (Kim, Kim, & Ergün, 2015). To adjust for outlier concerns, Spearman and Kendall correlation distances can be used as an alternative to Pearsons's correlation as they are non-parametric metrics that perform rank-based analysis (Gideon, & Hollister, 1987). As we are comparing a scale-normalized feature set of a model built in one district directly to another district in order to better utilize that model's predictions, we are not concerned with differences in scale (Zhang, Kwok, & Yeung, 2003). Given the nature of our data, we need to consider distance measures that can best identify the nearest neighbor using our normalized identity feature set.

One alternative to using a pure distance measure to calculate similarity would be the implementation of a recommender system approach. Similar to predictive models, these systems rely on historical data to match, or recommend, an outcome based on similarities within this data (Resnick & Varian, 1997). Methods such as content-based filtering (Basilico & Hofmann, 2004), collaborative filtering (Schafer, Frankowski, Herlocker & Sen, 2007), K-Nearest Neighbors (K-NN) (Wang, Liao & Zhang, 2013), Latent-factor (Koren, 2011), or a combination of some or all

of these methods have been used to build recommender systems (Li & Kim, 2003). While extensive research exists on building recommender systems (Amatriain, Jaimes, Oliver, & Pujol, 2011), the choice to implement a distance-based weighting measure for this research is driven by three factors: 1) implementing a recommender style system further reduces the interpretability of the final dropout results to the stakeholder (Gedikli, Jannach & Ge, 2014), 2) building a recommender layer on-top of a machine-learning driven EWS presents a significant technical challenge, significantly increasing the resources required to generate the EWS predictions and also compounding the time it would take to develop models in practice (Manouselis, Drachsler, Verbert & Santos, 2014), and 3) these systems rely on historical data to build the recommendation, which presents a problem for districts with high levels of missingness within their feature set (Marlin, Zemel, Roweis & Slaney, 2011).

The Euclidian distance measure was selected for this research method for several key reasons: 1) It is the basis of many measures of similarity and dissimilarity (Krislock & Wolkowicz, 2012), and one of the most commonly utilized measures utilized within clustering software (De Hoon, Imoto, Nolan & Miyano, 2004), 2) It provides us the opportunity to leverage some aspects of a k-NN recommender system approach within our approach by determining the similarity for the nearest neighbor within the district characteristic feature set (Hu, Huang, Ke & Tsai, 2016), and 3) the results are highly interpretable compared to other, more technical, measures (D'Agostino & Dardanoni, 2009). While the Euclidian distance can become more sensitive to noise in the data with high-dimensional feature spaces due to the squared terms (as the number of features grows, the relative distance between points can change in non-obvious ways), the n size of our distance features is sufficiently small for this limitation to be a mitigating factor for our choice to use this metric (Hassanat, 2014).

Using these similarity values, an algorithmic solution is developed to weight the importance of Pillar model probabilities generated for Target districts with the hope of improved performance when generalizing across districts. Specifically, to generate this similarity score, I take similarity features *F,* with the number of similarity features $F_s$. I then Z score each of these features across all districts to ensure equal weighting during distance calculation. For each of the *I* Pillar districts, I calculate the Euclidian distance *Eai,* between each Pillar district $A_i$ and the Target district *S* using feasible features *F.* I then find the average distance per feature $Eai_s$ by dividing by $F_{s.}$ The resultant values of $Eai_s$ scale between 0 (identical district properties) and infinity (most different district possible). The next step is to scale the values $Eai_s$ to be between 0 and 1, for easier calculation. I do this by using a re-scaling function $Eai_{sb} = (1/Eai_s) / ((1/Eai_s) + Q)$; where *Q* is a static value used to increase or reduce the severity of distance. This provides me with the distance between each pillar district and the target district ($Eai_{sb}$). The goal is to have all of the district predictions sum up to 1, in which case I can make a prediction for a given student that are scaled between 0 and 1 by simply summing the predictions from each Pillar model, multiplied by each Pillar district's distance. However, the values of $Eai_{sb}$ do not yet add up to 1. To re-scale these values so that they add to 1 across all Pilar districts, I use iterative gradient descent to find the value *M* such that the sum of all ($Eai_{sb}*M$) values together is 1. Note that the value of *M* needs to be calculated once for each Target district. Lastly, the predictions are then be taken for each student in the Target district, from each model $P_{i,}$ for all Pillar models $P_1 \ldots P_n$, *and multiple each prediction* $P_i = P_i*Eai_{sb}*M$. Finally, I sum all the $P_i*Eai_{sb}*M$ together; the final result is a prediction for that student, scaled between 0 and 1.

Both the fitted Pillar models used within the Pillar Pool and the performance of the Target district predictions generated by the DSEE are evaluated using the Area Under the Curve

for the Receiver Operator Characteristic graph (Hanley & McNeil, 1982). Models developed for specific districts as potential candidates to be Pillar models are fit and evaluated using held-out test sets from that district's own data. Districts for which we are able to produce a model with AUC higher than 0.7, averaged across all student class years, are designated as Pillar districts/models and used to create predictions for those districts which models could not be generated for all grade levels, or for which models were insufficient in quality (Targets).

Using alternative metrics such as Accuracy, Precision, Recall, or Kappa presents challenges for interpreting model results as schools are interested in not just a binary predicted outcome, but also the level of risk associated with the outcome (ex: dropout classified as High, Medium and Low) (Suh, & Suh, 2007). Accuracy is measured as the proportion of true positives observed among the total number of predictions made; using this metric of evaluation could over inflate our model results when the dropout records are highly imbalanced (ex: if 97% of our records are that a student graduated, and 3% of our records are dropout, the model can predict *graduate* for all the records and still achieve 97% accuracy) (Sidiroglou-Douskos, Misailovic, Hoffmann & Rinard, 2011). Precision and Recall involve a single threshold and are only concerned with the model's success at predicting true positives (in this case the *dropouts*, ignoring how the model performs on *graduates*) limiting interpretation to a binary outcome rather than a relative level of risk as Precision is focused on evaluating what proportion of predicted dropouts are actual dropouts, and Recall provides metrics on what proportion of historical dropouts were actually predicted to dropout (Buckland, & Gey, 1994). Cohen's Kappa is another common performance metric but fails to address the concerns presented by Accuracy, Precision, and Recall as it provides a metric of model performance for how the classifier performs over the unconditioned class probabilities, known as the base rate (Kvålseth, 1989).

The limitation of Kappa is that it is not perfect at controlling for the base rate (Delgado & Tibau, 2019) and there is no standardized way to interpret the results (Landis & Koch, 1977). Given the limitations of the metrics mentioned above, the Area Under the Curve (AUC) for the Receiver Operator Characteristic (ROC) was selected as our primary evaluation statistic due to its interpretability and validity for highl y-imbalanced test sets (Jeni, Cohn, & De La Torre, 2013). AUC ROC calculates the tradeoff between true positive and false negative for every possible threshold used for labeling data points as positive and negative; as such, it is well-suited for evaluating how well an algorithm ranks students relative to their risk (Bowers et al., 2012; Bowers & Zhou, 2019a).

Given that AUC values are reported from a scale of 0 to 1, where 1 is 100% perfect at classifying both outcomes, and 0 is 100% perfect at miss-classifying both outcomes, and 0.5 is a random guess, one could suggest that it might be effective to invert the AUC if it falls below 0.5 to turn a suboptimal model into an optimal model (Flach, Hernández-Orallo & Ramirez, 2011). This is accomplished by subtracting the probability produced by the model from 1, which produces the opposite predicted classification outcome. For example, if a trained model produces an AUC of 0.3, then we can assume it has a worse than random miss classification error. To correct this, one would simply invert the prediction so that a student predicted as dropout is now predicted as graduated, and a student predicted as dropout is now predicted as dropout. This would result in an AUC of 0.7.

While this strategy seems appropriate at face value, it is not used for this research for several reasons. A model can bias or underperform for non-obvious reasons; flipping the AUC and assuming the model is performing well provides a false sense of security in model accuracy, without understanding the root cause of why a model is performing so poorly. Additionally, early

warning systems are effective when they are both accurate and interpretable (Bowers, 2021). Inverting the AUC may produce the appearance of reasonably high model quality when it instead represents significant over-fitting to a training set with very limited signal in the data (Jamalabadi, Hamidreza, et al, 2016; Snoek, Miletić & Scholte, 2019). This could also lead to errors in interpretation and applications of interventions as there would be no evidence-based approach to understanding the reason *why* a student was predicted as at-risk. Given these concerns, this dissertation reports AUC values under 0.5 rather than inverting them.

### 3.1.3 Comparing the District Similarity Ensemble Extrapolation Model

I validate the new approach using several different methods. The first evaluates DSEE's performance against previously-published dropout detectors used at scale (though not all these models have been validated to generalize): the widely-used Chicago model (Allensworth & Easton, 2007), the Philadelphia logistic regression model (Balfanz, 2007), the Wisconsin machine learning EWS (Knowles, 2012), and the high-performing Bowers & Sprott Growth Mixture Model (Bowers & Sprott, 2012)

As mentioned previously, the Chicago model is a well-known and popular method used to identify students who are not on track for graduating from high school (Balfanz et al., 2007) and can be used for entirely new districts with no re-training. The Chicago model utilizes freshman-year GPA, the number of semester course failures, and freshman-year absences to determine the risk of the student not meeting the milestone of high school graduation (Allensworth et al., 2005). Since this traditional model relies on data collected within the first year of high school, I will only be able to compare the performance of the DSEE to the Chicago model for high school students that have freshmen year GPA, course failures, and absences data available in their records.

A similar approach is taken to compare the DSEE to the threshold-based Philadelphia-based Balfanz (2007) model. The Blafanz model looks at whether a student obtained a 1) final grade of *F* in mathematics, 2) a final grade of *F* in English/Language Arts, 3) attendance below 80 percent for the year, and 4) a final "unsatisfactory" behavior mark in at least one class. Utilizing any of these four signals, a student ($6^{th}$ grade or higher) is marked as at-risk (having a 75 percent or higher probability of dropping out of high school) if they meet at least ONE of the conditions (risk increased about 75 percent for students meeting more than one of *signal* conditions) (Neild, Balfanz, & Herzog, 2007). Like the Chicago model approach, I will only be able to compare the performance of the DSEE to the Balfanz model for $6^{th}$ to $12^{th}$ grade students that contain course grade and attendance data available in their records, as the "unsatisfactory" behavior mark is not universally collected by schools within my research data.

Comparing against the Knowles model requires me to subset my data to $6^{th}$-12 grade records and then calculate the DSEE performance against a replicated Knowles model (Knowles, 2012) built using the same $6^{th}$ – $12^{th}$ grade population of my research data. As discussed earlier in the Literature Review, the Wisconsin Dropout Early Warning System developed by Knowles utilizes a method that scans through many different machine-learning algorithms, selects the best performing models, and then ensembles them together into one predictor to generate a student's dropout risk. While Knowles did originally publish a publicly available code library with his original paper, this library will not be used when I replicate this model on my data due to insurmountable technical limitations. Specifically, these limitations involve issues with deprecated code, calls to obsolete libraries, and dependencies on non-publicly available code, making the publicly released code library no longer functional. Given these limitations, my approach to replicating the Knowles model will focus primarily on the methods outlined in his

publications, which involves writing new code to scan through a list of potential machine learned algorithms and then combining the best performing models together via an ensemble to create the detector for comparison.

Lastly, to compare against the Bower's GMM, I take the results of the calculated DSEE's AUC performance across all grades and conduct a direct comparison against the reported AUC of the GMM, published in 2010 (Bowers, 2010) and reported in 2012 (Bowers & Sprott, 2012). The decision to compare against the published results, rather than attempt to replicate the method on my data to create a new model, is primarily due to the GMM's structural equation modeling approach on one single indicator (GPA) over three semesters. Structural equation models are traditionally built using proprietary software (ex. MPLUS) and used for theory testing, making them difficult to implement in a system used to generate dynamic, on demand risk predictions (Evermann & Tate, 2016). While there are ongoing attempts to address this issue, there still exists a large gap of knowledge in how to replicate and productionize these methods using open-source coding languages such as R or Python (Wardenaar, 2020). Despite my inability to recreate the GMM model, providing a comparison is still valuable as the Bowers' published EWS leveraged a nationally representative dataset and achieved performance similar or better than other previous GMM-driven dropout research (Bowers, Sprott, & Taff, 2012). Unlike the previous comparisons made in this research, there is no requirement to subset my DSEE results data to match the GMM's results, as they both utilize an identical approach (generate predictions down to the 1$^{st}$ grade level and include records with missing feature data). Figure 4 below provides a visual representation of the model performance of existing EWSs I compare the DSEE approach to.

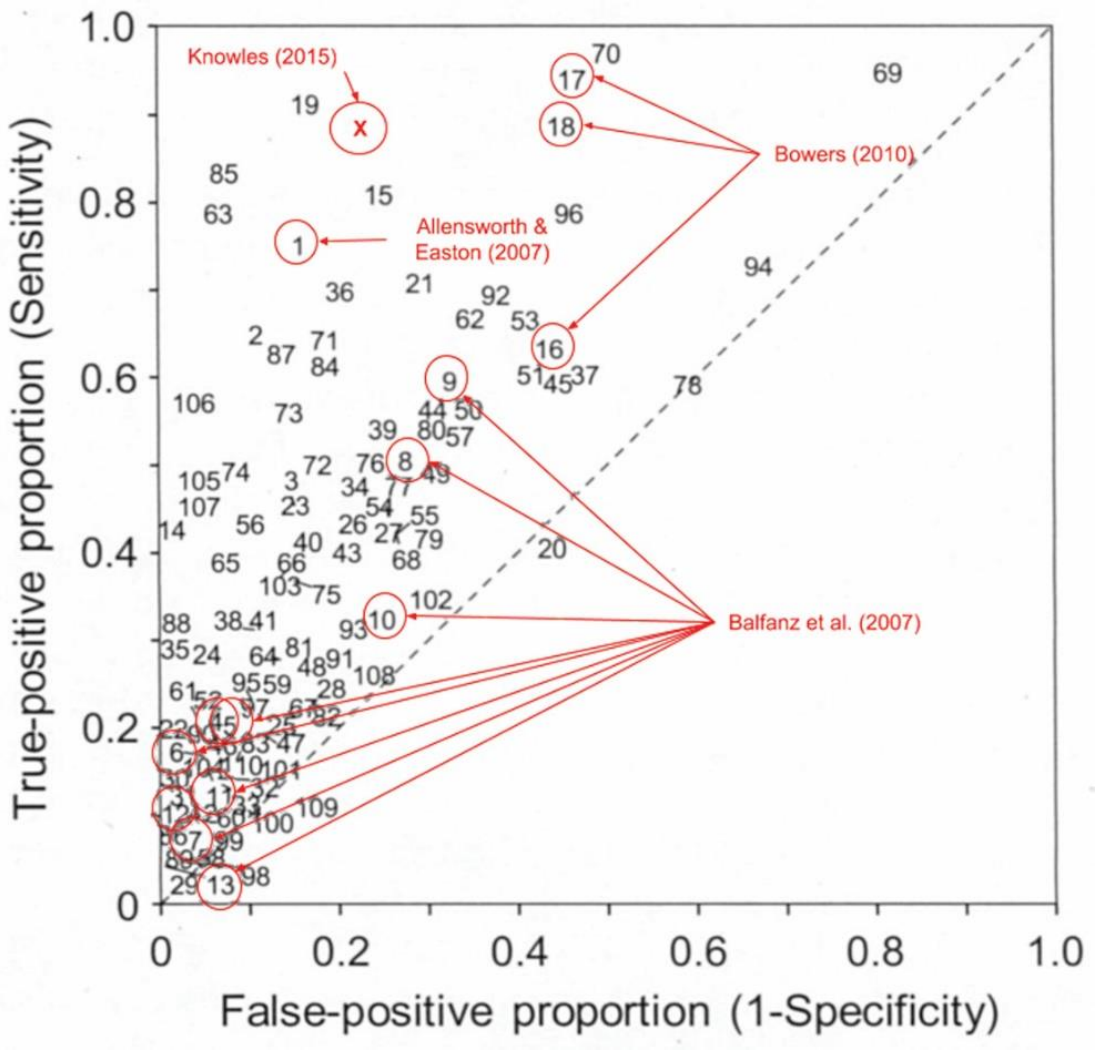



**Figure 4: Visual Representation of Existing EWS Model Performance[2]**

[2] Source: Adapted from 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, 77-100. Reprinted with permission.

An additional method of analysis is conducted which focuses on measuring 1) the performance of a base model (Aggregate Data model) comprised of all the records across all districts compared to an ensemble of multiple predictions from models created at the district level, and 2) the impact of using the distance calculation in the DSEE to weight the predictions within the ensemble. To accomplish this, I evaluate performance at all possible grades ($1^{st}$ – $12^{th}$) using two more approaches. The first method (Aggregate Data model) of comparison involves the creation of a single new model generated from an aggregate of all student records across all districts. Using the same 30% hold-out data for the DSEE and the aggregate model, predictions are created and evaluated using AUC ROC values. The results of this comparison allow me to identify if there is a significant improvement from building district/organization level models and then pooling them together rather than building one unified model with all the data.

The second and final method (Mean vs DSEE) of comparison takes a simple average of the predictions generated by the pillar models (coined the Mean model) by not utilizing the weights generated within the similarity function. Comparing a simple average against the weighted predictions allows me to determine if using a similarity function generated from descriptive features does indeed improve the performance of the model. The AUC ROC metric is used as the performance measure for this approach as well. To determine significance, I conduct a Delong Mann-Whitney-Wilcoxon test using the R pROC library (Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez & Müller, 2011) on the resulting AUC performances, a commonly used method for determining which model produces a better AUC when comparing performance across multiple detectors (Bamber, 1975; DeLong et al., 1988; Bowers & Zhou 2019).

To compare the effect of the Chicago model, Balfanz model, Knowles, Aggregated Data Model, the Mean model, and the DSEE model, a set of Delong Mann-Whitney-Wilcoxon tests

was conducted using the R programming language and the R pROC library (Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez & Müller, 2011). These tests allows me to determine if there is a statistically significant difference between each of the Early Warning Systems created above, based on the reported AUC performance within each district (Hsu & Peter, 2005). Districts for which I was unable to calculate an AUC (due to data limitation etc.) will be removed from the analysis, resulting in a maximum of 64 AUC pairings out of the original 70 Target districts. A total of eight tests are conducted, based on the target populations of the EWS. For comparing performance of predictions generated down to the first-grade level, the following *DeLong-tests* were made: the Aggregate Model compared to the Mean model, the Aggregate Model compared to the DSEE model and the Mean model compared to the DSEE model. To compare EWS performance for predictions generated for $6^{th}$ grade students, the following *DeLong-tests* conducted: DSEE compared to the Balfanz model and the DSEE compared to the Knowles model. Lastly, to compare the performance of the DSEE to $9^{th}$ grade student populations, two *DeLong-tests* were conducted between the DSEE compared to the Chicago model, the DSEE compared to the Knowles model. Note that given that the Balfanz, Chicago and Knowles models were not built for all student grade populations, the AUC performance will be recalculated for the DSEE model using the same populations defined in the Chicago, Balfanz, and Knowles methodologies.

## 3.2 Data for Analysis

Data for this research originate from the BrightBytes data analytics and visualization platform, Clarity®. The Clarity® platform ingests disparate datasets, transforms them into a standardized format by mapping district-specific variables to a common schema, prepares the data for analysis, and then visualizes the data in a meaningful, easy-to-understand way. The

Clarity® platform is used by 1 in 5 schools across 47 states to empower educational leaders to use data for decision making. The value derived from the Clarity® platform comes from using data to drive change within an organization (Strudler & Schrader, 2016; Chute, 2019). The anonymized dataset used to support the DSEE research represents a large spectrum of K-12 students in terms of free/reduced lunch eligibility, school urbanicity, and school demographic makeup consisting of almost 3 million individual student records.

The set of predictor variables was selected in partnership with the American Institutes for Research (AIR) Early Warning Systems in Education team (Heppen & Therriault, 2008), researchers and developers at BrightBytes (including this dissertation's author), and a researcher at the University of Pennsylvania. This collaboration resulted in a theory-based (Bernhardt, & Bernhardt, 2013) framework of success indicators, along with definitions of those success indicators that are used to map and align district data. Due to the data ingestion and transformation process, the same data features can be used across all districts. Below is a distillation of the broad range of potential variables into a small set of meaningful buckets:

- **General Coursework:** indicators related to student academic performance such as total credits earned or student grade point performance within course type (math, science, reading, social sciences, etc.), non-cumulative grade point average, and grade point averages within course category (core courses, elective courses, etc.) (Bowers, 2019; Bowers, 2010; Bowers, 2011; Bowers & Sprott, 2012; Kemple, Segeritz, & Stephenson, 2013; Allensworth, Gywnne, Moore, & de la Torre, 2014; Balfanz, Bridgeland, Bruce, & Fox, 2013; Brookhart et al., 2016; Stuit et al., 2016; Balfanz, DePaoli, Ingram, Bridgeland, & Fox, 2016).

- **Student Assessments**: interim or summative assessments related to math, science, reading and social studies performance (Koon, & Petscher, 2016; Cumpton, Schexnayder, & King, 2012; Bowers & Zhou, 2019).
- **Student Attendance**: recorded absences, tardies, and flags of chronic absenteeism (Balfanz, & Byrnes, 2012; Rafa, A., 2017; Caldarella, Christensen, Young, & Densley, 2011; Hein, Smerdon, & Sambolt, 2013; Bowers & Sprott, 2012).
- **Student Behavior:** data related to the number and type of recorded disciplinary incidents the student has on file (Balfanz, Byrnes, & Fox, 2015; Bowers, & Sprott, 2012; Owens, J., 2016; Office of the State Superintendent of Education, District of Columbia, 2014; Landers, Courtade, & Ryndak, 2012).

3.2.1 Study Participants

The data used in this research consists of 326,533 unique students from 88 school districts (for the purposes of this research, models will be built at the district level), all with varying levels of dropout rates, diversity, and locality. The distribution of gender was largely equal, with 158,590 female students and 159,641 male students (data on gender was unavailable for 8,302 students).

The ethnic distribution of these students consisted of 7,096 (2.17%) Asian, 67,900 (20.82%) African American, 34,592 (10.59%) Hispanic, 1,319 (0.40%) Native American, 4,896 (1.5%) multi-ethnic, 1,195 (0.37%) Pacific Islander, 193,300 (59.20%) White and 7,863 (2.41%) undefined. 8,302 student records did not contain any ethnicity data. According to July 1$^{st}$ 2019 population estimates provided by the U.S. Census Bureau, the United States has an ethnicity distribution of 5.9% Asian, 13.4% African-American, 18.5% Hispanic/Latino, 0.2% Native American or Pacific Islander, and 60% White alone (not Hispanic or Latino) (2019) suggesting

that our study sample is over-representative of some groups (African-American, Native American, and Pacific Islander) and largely under-representative, by a factor of up to 3, of other groups (Asian, Hispanic/Latino, multi-ethnic, and White). These differences are likely caused by variances in the local population such as urbanicity, the funding available to purchase an educational technology tool (Title 1), and the factors motivating the decision to choose the BrightBytes ed-tech solution used to collect the data within this analysis.

Within this population of study, 35,151 students were flagged as dropping out with 288,317 students graduating high school, showing an 11.40% dropout rate across all school districts. While this number may seem relatively high overall, the dropout rate varies significantly within each school district. Additionally, there was significant heterogeneity across districts for when the dropout event took place, with some districts observing students dropping out in earlier grades (as early as $6^{th}$ grade) and others recording the highest proportion of dropouts in higher grades (see Figure 38 in Appendix A for full dropout distribution across districts). Each student provides a record for each historical grade they attended, with one unique student having a possible max number of 12 total records in the data, one for each grade (1-12), creating a total size of 2,362,621 records for examination. See Appendix A for tables that provide a descriptive summary calculated on the data used within this research.

A primary concern about the data, and the motivating factor of this research, is that there exist large gaps of recorded data within many of these districts. Table 14 found in Appendix A provides a good summary of this issue, with 35 districts containing less than 100 historical dropout records, and 6 districts containing less than 100 historically recorded graduation records. These results are especially concerning, as the data collected from this research originates from school districts with relatively large (more than 100) current student populations. As mentioned

earlier, districts that lack historically recorded graduate or dropout records are hindered in their ability to generate a district specific machine learned EWS. Traditionally, these districts would have to rely on threshold based EWSs as they do not have the data required to implement an accurate machined learned method of dropout detection. For the purposes of this research, districts with no historical records were removed from the analysis as I am unable to conduct any cross-validation to measure prediction performance. In practice, these districts would leverage current student records to generate predictions driven by the DSEE EWS method further discussed in this paper.

### 3.2.2 Data for District-to-District Similarity

Despite the data quality issues present in the historical student data, they do not impact the current student population records. This is largely due to the school districts' decision to partner with an educational technology company (in this case, the BrightBytes company), which provides them the capability to accurately record and track current student educational data aligned to a common unified schema. Having clean, accurate, and common current student data is a requirement to generate risk predictions for students in one district using a model from another (the core approach used in the DSEE). It also makes it possible for me to create additional features that can be utilized within the similarity weighting function in the DSEE solution.

Current student populations vary across school districts; the average number of currently enrolled students across all 88 districts at is 7,936.16 (SD=13,631.62). Ethnic/racial distributions vary widely as well, with 6 districts consisting predominately Asian students (Orgs 13, 14, 15, 43, 44, 46), 8 containing mostly Hispanic (Orgs 0, 7, 8, 9, 26, 37, 38, 39) students, 1 containing mostly Pacific Islanders (Org 106), 1 containing mostly Indigenous students (Org 48), and 65

containing a majority of White students. The remaining 7 districts contain a diverse distribution of ethnicities. The extreme racial heterogeneity across districts highlights the issues that continue to persist from segregation policies enacted decades ago (Reardon, 2019). Appendix B provides summary statistics (provided by BrightBytes) of the current student populations of study.

As mentioned previously, the data used to calculate the district-to-district similarity used in the DSEE falls within two categories: (current) student demographics such as ethnic/racial distribution or average graduation rates and local population attributes such as population economics, employment rates, median income, etc. The current student demographic data is captured within the BrightBytes Clarity® platform, which is calculated at the district level using the information provided by the district. To obtain local population statistics, additional data from a reliable, publicly available source was required. The local population attributes were collected using American Community Survey 5 Year Estimates, accessible through the U.S. Census Bureau (2020) API. The American Community Survey 5-Year Estimates contains data down to the block-group level, and covers a large of topics such as social, economic, demographic, and housing characteristics of the U.S. population, with over 20,000 unique variables. To extract the data utilized in this research, zip codes were collected on every school within each school district. These zip codes were then used to query the publicly available Census Data API to obtain these additional local population features. Specifically, the following variables were extracted:

- B01003: Estimated total population.
- S2301_C01_001E: Estimated employment status of population 16 years and over.
- S1501_C01_008E: Estimated population of 25 years or older with no high school diploma.
- B19013_001E: Estimated median household income in the past 12 months (in 2018, inflation-adjusted dollars)

Urbanicity data was also collected from the Census Bureau using provided definitions. The US Census defines two types of urban zones within the United States: Urbanized Area (UAs) which consist of 50,000 or more people, and Urban Clusters (UCs) which consist of at least 2,500 people and less than 50,000 people. Communities designates as "Rural" encompass the population not included within an urban area (Census Bureau, 2020). As such, a total of four urban-rural variables were calculated consisting of a total count of persons residing in each urban-rural type at the zip code level. These four variables are defined as: UA representing the count value of Urbanized Area. UC representing the count value of Urbanized Cluster, Urban, consisting of the combined count total of Urbanized Areas and Urban Clusters, and Rural, consisting of the count of persons residing in rural areas. This data was then converted to a percent value using the total population to create a ratio representing each urban-rural zone feature.

After acquiring this data at the zip code level, an average was then calculated for each school district by taking the mean values of all the school zip code within this district. This produced an average population of 93,623 (SD=168,274), median income of $50,445 (SD=$11,145.32), ratio of peoples in combined urban area and urban clusters of 0.546 (SD=0.31), ratio of peoples urban areas of 0.365 (SD=0.387), ratio of peoples in urban clusters of 0.181 (SD=0.245), the ratio of peoples in rural areas of 0.455 (SD=0.314), employment rate of 0.81 (SD=0.04) and high school educated rate of 0.94 (SD=0.02) across all 88 school districts. After computing these averages, the data was joined with the collected current student population demographics. This combined data is then utilized within the similarity calculation of the DSEE modeling approach, discussed in further detail later in this document. Appendix B provides a summary of the local population statistics similarity data used in the analysis.

## 3.3 Instruments

Several instruments were used throughout this study. As mentioned earlier, the initial data originates from the BrightBytes data analytics and visualization platform, Clarity®. The Clarity® platform ingests disparate datasets, transforms them to a standardized format by mapping district-specific variables to a common schema, prepares the data for analysis, and then visualizes the data. The Clarity platform works by ingesting all available data from the various tools (attendance trackers, grade books, intervention management systems, etc.) used within the school district using a series of Application programming interfaces (API's). This data is then mapped and aggregated at the student level which is then stored in an Amazon Redshift Database.

To generate the feature set used within this dissertation's analyses, SQL queries were made to the Redshift database using the Psequel integrated development environment (IDE). These queries created a series of tables that contained the base features, the generated additional features, and the population descriptive features within each educational organization. Once these tables were made, they were unloaded to the Amazon S3 data lake service and downloaded using a command line interface (CLI) to a local computer to be analyzed and modeled.

Predictive models were created using the Anaconda Python programming language distribution. This Python package includes several IDEs and all the scientific computing libraries needed to manipulate and build the machine learned dropout risk prediction model (see Appendix N for full list of packages and libraries). Traditionally, the models would be productionized by integrating the python code within an automation engine and deployed using a distributed computing cloud service in user defined frequencies (i.e. the user decides how often the model updates) in order to create updated predictions as the student data changes over time. I

did not have access to such a system, and instead relied on the Spyder 4 IDE and the computing power of a local desktop computer to complete this research. This desktop computer ran on the Windows 10 64-bit operating system and contained an AMD Ryzen 7 1700 Eight-Core 3.0 GHz Processor with 32 gigabytes of DDR4 Rapid Access Memory (RAM).

## 3.4 Preparing the Data for Modeling

Identical data preparation was conducted for all three models' methods created in this research. For every unique student, their end of year records (reported values in the system as of July 31st) were collected and extracted for each grade a record was present. This resulted in a long-form data set containing multiple rows for each unique student, with each row representing their academic, assessment, behavior and attendance data for each year they attended school within the district (historical records for transfer students were added when available). Once the core feature set (attendance, assessment scores, academic performance, and behavioral Incidents) was identified for each student, it was then manipulated to generate additional insights by creating new features and to remove any data anomalies. The additional features generated consisted of several combined and computed features built using the base feature set. An example of this in practice would be taking the average of recorded grade point average (GPA) of the science, language arts, history, and math courses to create a new core courses GPA. This process resulted in a total of 56 unique features that were used to predict likelihood of high school graduation.

Given the nature of the discrepancies mentioned earlier of how individual districts record data, the feature set was then normalized within district and within class number. Normalizing this data accomplished two things: first, it allows me to account for large variations in the recorded data as some districts recorded the data in different ways and scales (ex: one district

recording academic performance on a 4 point scale vs another recording academic performance on a 100 point scale). Without normalizing this data, model performance would be heavily impacted. The second reason to normalize this data is to try to identify and reduce any unforeseen predictive bias created by the model. Research has shown that model performance can heavily bias within various populations. This bias can produce a highly inequitable environment for underserved populations. To ensure fair treatment among all groups, data normalization is completed by converting values to standard scores (z-score). This allows me to examine how far any given value falls from the population mean on a normal distribution and is used to identify data imbalances and reduce unfair treatment effects (Actionable Intelligence for Social Policy, 2020). Once these features are calculated, the next step is cleaning the data.

Data cleaning involved extracting the data from the database and then stripping out any white space within the feature space. Records before 1$^{st}$ grade and after 12$^{th}$ grade were removed from the data set. I then created a new column called “dropped” in the data containing a Boolean label where student records received a value of 1 if they were a high school dropout and a value of 0 if they were a high school graduate. Only these students were used for building the models; all other outcomes such as transferring to another school district, current students, and records that did not have a historical outcome values were removed from the filtered dataset. Student metadata (identifying keys) were then removed from the data set. The remaining features were then converted to a number data type, coercing all non-number values to Not a Number (represented as NaN in python) missing values. Converting the missing values to NaN allows me to both visualize and address any missingness within the data.

Results of the missingness analysis show that some features suffer from high numbers of missingness, with interim assessments having the highest levels of missingness (M = 0.985%,

SD = 0.134%), summative assessments containing an average of 0.611% missingness (SD = 0.321%), credit based features having an average missingness of 0.518% (SD = 0.197%), GPA based features having an average of 0.240% (SD=0.229%), and behavior based data showing the least amount of missingness at 0% (SD=0%) (see Appendix C). Given the nature of educational data, identifying, and addressing the cause of this missingness is not simple. While the features selected for this analysis attempt to be general enough that most districts would be able to populate these values, there are features that by their nature will always be missing for some students. Data collected and recorded in later stages of academic progress such as GPA, or Advanced Placement course participation, will not be present for early grade students. Additionally, some students may not have access to these initiatives and do not have an opportunity to be exposed to these programs. Simply omitting these features would address the missingness but would likely reduce the accuracy of the model as these feature types have been shown to be predictive of dropout. Taking all of this into consideration, implementing a strategy to address the missingness while also maintain the maximum amount of information on the student is required.

The severity of the missingness within the data impacts the algorithms that can be used to build the models (Marlin, 2008) and the process used to generate the predictions (Batista & Monard, 2003), which means that selecting the way we address this problem is an important step of the data preparation process. There are many different imputation methods that can be implemented to address this challenge, with varying levels of complexity (Lakshminarayan, Harp, Goldman & Samad, 1996). While data imputation can be a powerful tool for handling missing data (Schafer, 1999; Schafer & Olsen, 1998), it is not necessarily ideal when trying to predict the very variable that is missing. As a result, it can often be infeasible to create scalable,

locally validated models for specific districts that generalize to new unseen district populations. Given this limitation, the simple method of arbitrary value substitution was used to replace missing values with a high, out of bounds integer (a value of 2,000).

The resultant dataset was highly imbalanced, with substantially more students graduating than dropping out, as can be seen in Appendix A. To account for this imbalance, the training data was manually re-balanced (using random-over sampling) by adding duplicate copies of students who dropped out to the data set. A count of records within each grade is calculated to determine how many historical records exist within each recorded class number which was then used to inform the up sample. Specifically, duplicates were created such that every grade level (10th, 11th, 12th, etc.) of students in the training datasets had an equal number of students who dropped out as students who remained. The original data distribution was used when testing the models. This resulted in the final dataset used for the creation of the base model, coined the Aggregate Data model, the Mean model, and the DSEE model.

### 3.5 Model Parameter Tuning

Many machine learning algorithms contain specific settings that can be changed within the algorithm to better optimize the performance (Sonobe, Tani, Wang, Kobayashi, & Shimamura, 2014). These *hyperparameters* can be adjusted to increase the predictive results of the models within which they are used (Probst, Wright & Boulesteix, 2019), but there is a risk of creating under-performing models through over-fitting, selecting an inappropriate metric, or setting incorrect hyperparameter values (Feurer & Hutter, 2019). Like adjusting the faucet on a bathroom sink to achieve better water efficiency, tuning the hyperparameters on a machine learning algorithm during model training can improve the performance of the detector. In this section, I discuss my approach to tuning the hyperparameters of machine learning EWSs I have

created within my research (the Aggregate Data model, Mean model, and DSEE). The specific model hyperparameters for each of these modeling approaches are provided in their relevant sections found in the results section of this paper.

During a previously completed pilot study on this data, the random forest algorithm was determined to be the best algorithm to use given the nature of the data used in this research (Coleman, Baker & Stephenson, 2019). The random forest works by building multiple decision trees, and then pooling them together to achieve a higher performing model. As an analogy of this process, let's imagine a soon-to-be-graduated high school senior named Susie, who is having difficulty committing to one of the two universities she has recently been admitted. To help with her decision, she approaches her best friend Gary for advice. Gary asks Susie a series of questions regarding her interests; does she like an urban or rural campus setting? Does she prefer a school with a large sports team? Does she prefer a cold weather climate or moderate weather climate? Based on Susie's answers, Gary provides a response as to which university she should attend. In its simplest form, this is how a decision tree works. Gary created a list of conditions to present to Susie, who then provided an answer which ultimately led to the university recommendation provided by Gary. After speaking with Gary, Susie then begins asking additional friends for advice on which university she should attend. Some of her friends ask new questions, some ask the same questions, and some ask a combination of both new and similar questions until finally providing their recommendation. Susie collects all these responses, and finally decides to attend the university that was recommended most often to her. A random forest operates the same way, where instead of relying on one friend to determine the decision, it relies on many friends (trees) and then combines the cumulative results to generate the final outcome.

Given the decision to utilize a random forest method, the hyperparameter tuning method was implemented with this algorithm in mind. Tuning a machine-learning algorithm requires a significant amount of trial and error, as determining the best parameters to use is not possible prior to fitting the model (Bei, Yu, Zhang, Xiong, Xu, Eeckhout, & Feng, 2015). This means that in order to identify the optimal hyperparameter setting values, a test of all possible combinations (within the parameter value space) needs to be conducted, known as a hyperparameter sweep (Kostrikov, & Gall, 2014). Additionally, there exists an increased risk to overfitting the data when conducting the sweep to tune the hyperparameters of the model, addressed in this case by using a fully held-out data test set.

The Random Forest algorithm contains 6 primary hyperparameters that can be tuned when fitting the model to find the most optimal detector. They consist of the following items; 1) the number of trees or estimators to use in the forest, 2) the maximum number of features to consider when splitting a node, 3) the maximum number of levels in each decision tree within the forest, 4) the minimum number of samples within a node before the node is split, 5) the minimum number of samples for a node to be considered a leaf, and 6) whether the data sampling method will utilize sampling with replacement or sampling without replacement. (i.e. Bootstrapping) (Hesterberg, 2011). To accomplish this, a randomized parameter grid was created to sample from during the model fitting.

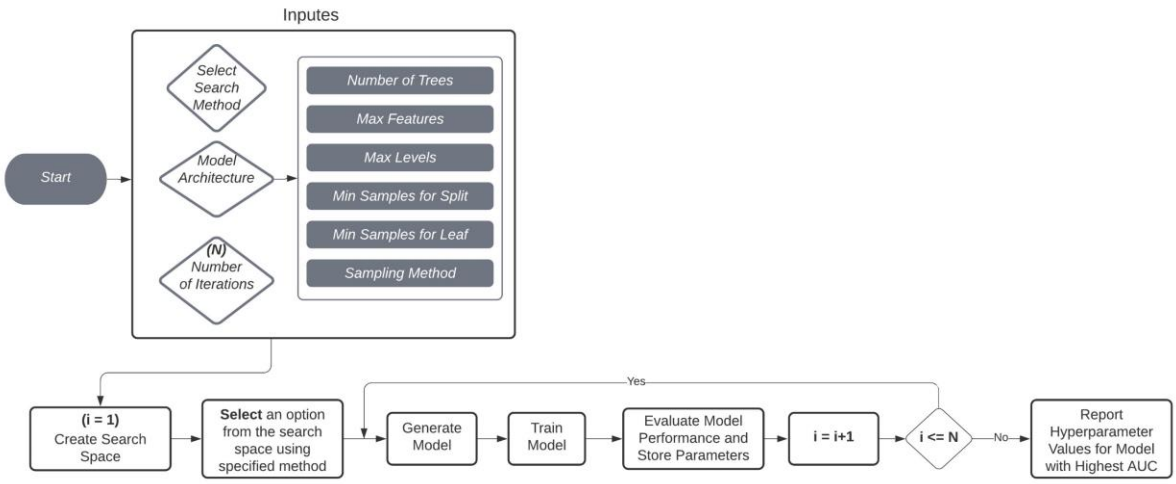



**Figure 5: Flow chart of hyperparameter search procedure**

This involves the creation of an n-dimensional vector, with each hypermeter representing a dimension and the scale of the dimension consisting of each possible value within the hyperparameter. This vector can be viewed as a catalog of all possible combinations of values that can be used (via random sampling) efficiently to train and evaluate models with various combinations of hyperparameters to identify the optimal settings. The grid was created using the following hyperparameters:

**Number of Estimators:** [5, 10, 15, 20, 25, 30, 35, 40, 45, 50]

**Maximum Features:** *sqrt*, which takes the square root of the total number of features and *auto*, which simply takes all the features into consideration.

**Maximum Depth:** [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110]

**Minimum Samples for Split:** [2, 5, 10]

**Minimum Samples for a Leaf:** [1, 2, 4]

**Bootstrapping:** *Enabled* and *Disabled*

Using this grid, 100 iterations of model training were conducted, with the algorithm randomly sampling (sweeping) from the grid during each iteration. Conducting a random search provides the benefit of not having to fit every possible combination of parameters (in this case, 3,960), but instead enables the algorithm to search through a wide range of values to identify the appropriate hyperparameters. Throughout each iteration, an AUC was calculated to evaluate the model's performance. Once the 100 iterations were completed, the parameters that produced the highest AUC was used to train the final model used to generate the student risk predictions.

To address concerns around over-fitting, 70% of the data was sampled from the total, and then 3-fold cross validation was used during the hyperparameter tuning process (Duarte & Wainer, 2017). While there is some evidence to suggest a larger number of folds (ex:10-fold) is more effective at validating model performance, there is a significant trade-off between improved performance and computational resources (Bengio & Grandvalet, 2004). Due to computational limitations, 3-fold cross validation was used as it is significantly better than simple hold-out validation and less expensive than 5 or 10-fold cross-validation (Moore, 2001). Once the optimal parameters were identified, hold-out validation was then used to fit the model, with the 70% randomly selected data set used for training and the remaining 30%, which was excluded from the hyperparameter tuning, and used to measure the predictive performance of the model. This created a total of 300 model fits during the hyperparameter tuning process, with one final fit conducted once the optimal parameters were identified.

### 3.6 Model Fitting

In the following section, I discuss my approach to fitting the generated models used to compare the performance of the proposed DSEE model in this research. As mentioned previously, an Aggregate Data model (using all records available across all the educational

organizations) and a Mean model (taking a simple average of predictions generated by the pillar model) was created.

### 3.6.1 Aggregate Data Model

By combining all the data into one, single model, I am able to better understand the performance impact of having one single model generating predictions using one national level model compared to having multiple models generating weighted and unweighted predictions at the organization/district level, this model will serve as a baseline for comparison. Creating the Aggregate Data model began with first combining all available records into one single dataset (n = 326,533) containing a total of 2,362,621 records. The data was then prepared using the method outlined above in the Preparing the Data for Modeling section. The data was then split into two partitions, one consisting of 70 percent of the records (used for training) and the remaining 30 percent used for validating the model.

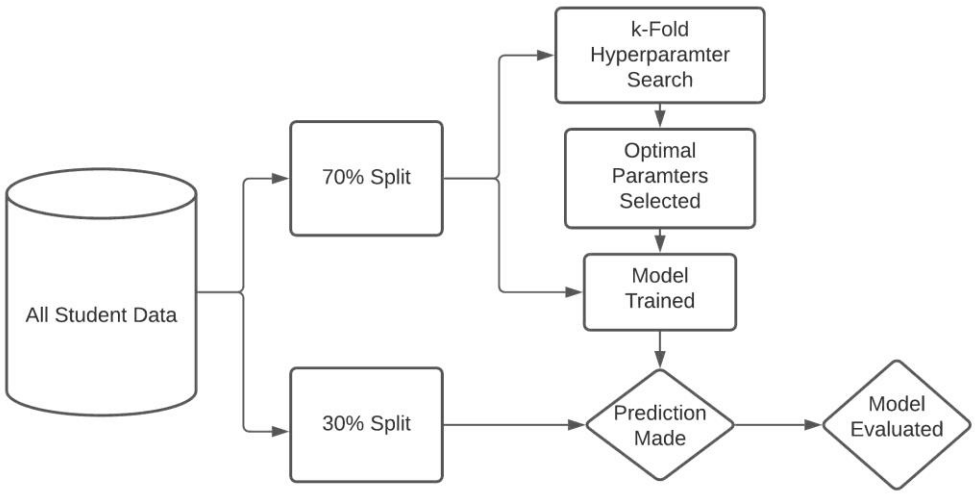



**Figure 6: Process for fitting the Aggregate Data model EWS**

Once the random forest algorithm was fit using the default base values, the hyperparameter tuning sweep (outlined in the methods section) was conducted to identify the

optimal parameters for the model based on AUC performance. This process required 8 hours of computation time (with all 16-cores utilized) and involved over 300 model fits using 3-fold cross-validation during the tuning process. After the hyperparameter tuning sweep was completed, the best performance (measured using AUC) was a model with the following hyper parameter values:

**Number of Estimators:** 40

**Maximum Features:** *sqrt*

**Maximum Depth:** 40

**Minimum Samples for Split:** 5

**Minimum Samples for a Leaf:** 2

**Bootstrapping:** *Disabled*

The final model used to create the Aggregate Data model was fit on the training using these optimal hyperparameter settings. The validation data was then scored against this model and used to evaluate the 1) overall performance of the detector across all records and 2) the performance of the detector within-district. Results of this analysis is found below in the Research Findings section of the paper.

### 3.6.2 Mean Model

The Mean model was created by developing and validating predictive models for each school district with sufficient data to create their own predictor, with these models predicting each student's probability of graduating (or risk of not graduating). These models are then used to generate predictions for students in districts that lack the data required to build their own district specific model. The predictions are then averaged together using a simple ensemble approach (taking an average of all the produced risk probabilities), to produce a single prediction

for each student. The performance of this approach is tested by using the historical records from held-out districts where data is available. This Mean model forms the basis of the proposed DSEE approach and will allow me to measure the impact the DSEE similarity calculation will have on the predictive performance compared to having no weighting function in place.

Building the Mean model was a several stage process. The first step (Stage 1) was designating districts as either Pillars or Targets based on the districts' data properties. As a first step within this stage, districts were classified as Targets that if they did not contain historical data spanning all grades 1st through 12, as districts without historical data would be less useful for modeling than districts where historical data are present. This process resulted in the selection of 79 possible Pillar models, with the remaining 9 organizations being designated Target districts.

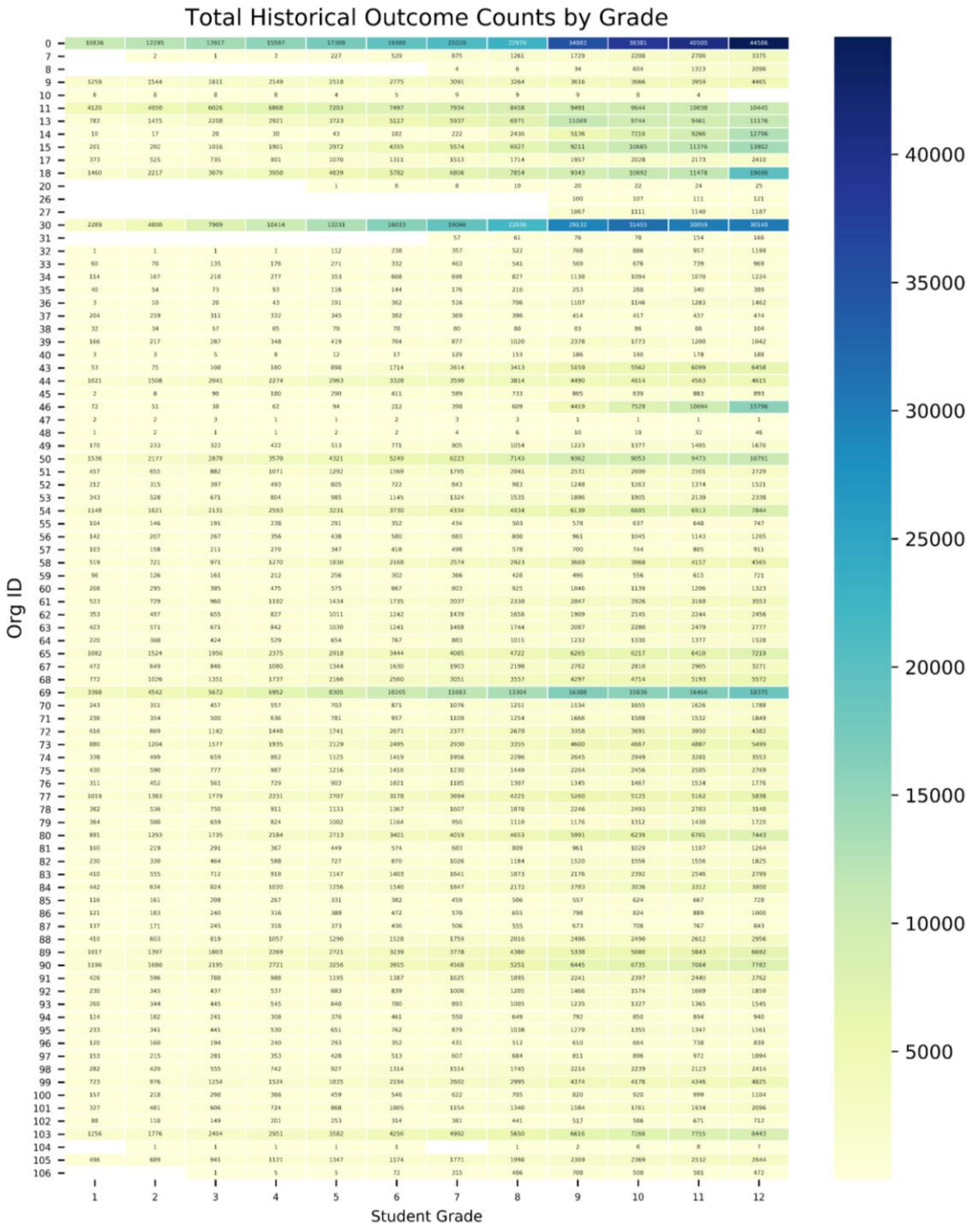



**Figure 7: Count of Historical Outcomes by Grade and Organization**

Figure 7 above provides a visual representation of the number of available outcome records across all students within each grade and organization. From this chart, we can see that most school districts contained some level of graduation recorded graduation outcome for 1st

through 12th records. Additionally, the visualization shows that that three organizations (Orgs 0, 30, and 69) contain significantly more outcome records compared to the other school districts within this analysis. Despite most districts containing records, some organizations are completely missing outcome values for a specific grade (Orgs 20, 26, 27 and 104), while some only contain single digit values (Orgs 32, 40, 47, and 48), which suggest these districts did not begin collecting student data in a digital format until recently and are unable (either through lack of resources or interest) to convert their historical data for use in a district specific machine learning driven EWS system.

I then designated districts as Targets that had low (less than 20,000 total outcome records) numbers of historical records. This resulted in the reclassification of an additional 53 districts from possible Pillar models to Targets, creating a total of 62 (M=10,934.34, SD=11,180.38) school districts identified as Targets and 26 (M=63,288.23, SD=62,776.56) identified as possible Pillar models. Lastly, I calculated the proportion of missing values within the total feature set for these Pillar candidates, and classified districts with over 40 percent of values missing across the entire feature set as Targets, as these districts would be less useful for modeling other districts where these features were present. Two of the potential Pillar candidates contained more than 40 percent missing data (M=0.562%, SD=0.089%) and were therefore reclassified as Targets. The remaining 24 potential Pillars had relatively good data completeness, with an average of 27.98 (SD=0.056%) percent missing data for all features (see Appendix D).

The next stage (Stage 2) toward building an at-risk prediction model for districts without sufficient data is to build models for districts with sufficient data. This was accomplished by attempting to fit a model for every educational organization in the data set. For each of these models, the data was prepared using the same method as the Aggregate Data model, where

records were filtered down to only the students who were flagged as ‘dropped’ and ‘graduated’, simple imputation was applied to address the missing values, with the training set having outcome records duplicated at the class number level to address any outcome imbalance issues.

Identical to the Aggregate Data model, Pillars models underwent hyperparameter tuning to identify the optimal settings for the algorithm prior to training the final detector used to generate the student risk predictions. This tuning process involved over 300 fits for each potential Pillar model, requiring significant resources and several days to complete using all computational resources available. Overall, a total of 8,937 fits were conducted to produce the optimal model for each district. The table below provides a summary of the selected hyperparameters for each potential Pillar district model.

**Table 2: Pillar Model Hyperparameters Selected During Model Tuning**

| Pillar Model Org ID | (n) Estimators | Max Features | Max Depth | Min Samples for Split | Min Samples for Leaf | Bootstrap Enabled |
|---|---|---|---|---|---|---|
| 0 | 40 | auto | 10 | 5 | 1 | TRUE |
| 11 | 40 | sqrt | 10 | 5 | 2 | FALSE |
| 13 | 50 | sqrt | 10 | 5 | 1 | TRUE |
| 18 | 45 | auto | 10 | 5 | 4 | FALSE |
| 30 | 50 | sqrt | 10 | 5 | 1 | TRUE |
| 50 | 25 | auto | 10 | 2 | 1 | FALSE |
| 51 | 40 | sqrt | 10 | 2 | 4 | TRUE |
| 54 | 40 | sqrt | 10 | 2 | 4 | TRUE |
| 58 | 40 | sqrt | 10 | 2 | 4 | TRUE |
| 61 | 40 | sqrt | 10 | 2 | 4 | TRUE |
| 65 | 40 | sqrt | 10 | 2 | 4 | TRUE |
| 67 | 10 | sqrt | 10 | 2 | 4 | TRUE |
| 68 | 40 | auto | 10 | 5 | 1 | TRUE |
| 69 | 25 | sqrt | 10 | 10 | 2 | TRUE |
| 72 | 10 | sqrt | 10 | 2 | 4 | TRUE |
| 73 | 40 | sqrt | 10 | 2 | 4 | TRUE |
| 74 | 10 | sqrt | 10 | 2 | 4 | TRUE |
| 77 | 25 | sqrt | 10 | 5 | 2 | TRUE |
| 80 | 40 | sqrt | 10 | 2 | 4 | TRUE |
| 88 | 40 | sqrt | 10 | 2 | 4 | TRUE |

| Pillar Model Org ID | (n) Estimators | Max Features | Max Depth | Min Samples for Split | Min Samples for Leaf | Bootstrap Enabled |
|---|---|---|---|---|---|---|
| 89 | 40 | sqrt | 10 | 2 | 4 | TRUE |
| 90 | 50 | sqrt | 10 | 5 | 1 | TRUE |
| 99 | 25 | sqrt | 10 | 5 | 2 | TRUE |
| 103 | 40 | sqrt | 10 | 2 | 4 | TRUE |

The goodness of each district's model was evaluated, within-district, using a train-test split method (note that models are also evaluated within entirely new districts; see below). In each case, the training set consisted of a randomly selected 70 percent of the data with label-based stratification used across grades. The test set held out to validate the model consisted of the remaining 30 percent of the data., with the Area Under the Curve for the Receiver Operator Characteristic used as the model evaluation statistic.

After attempting to fit a model for every district Pillar candidate in the dataset, the performance was reviewed for each districts predictor in order to identify the final models that will be simple ensembled (the Pillar Models), with the remaining underperforming districts used to validate the ensemble (joining the Targets). Selection of Pillar districts at this point was based model performance. As mentioned previously, models developed for specific districts as potential candidates to be Pillar models were fit and evaluated using held-out test sets from that district's own data.

### 3.6.3 District Similarity Ensemble Extrapolation

Having developed models for Pillar districts, where data are abundant, data quality is high, and where it is possible to develop a high-quality model, I next applied the DSEE (District Similarity Ensemble Extrapolation) approach. This approach combines the Pillar district models (created during the Mean model method) to obtain predictions for the Target districts, in a more

sophisticated fashion than just averaging them. This was accomplished through a several step process.

The first step to applying the Pillar models was simply to run each of them on the Target district's data and obtain predictions for each student. This was completed when the Mean model was created and provides a set of predictions for each student and for each model. Second, I calculated the similarity between the Target district and each of the Pillar districts. The district-to-district similarity is calculated based on several properties (not utilized within the predictive model) that capture some of the key differences between school districts and the students within them. These indicators include information such as general student demographic ratios, grade enrollment distributions, district graduation rates and local population data such as employment rates, median household income, and urbanicity. The last step involves converting the similarity scores into weights using a gradient descent approach, and then applying these weights to the probabilities generated by each Pillar Model. The final weighted average is then used to determine a student's risk of dropping out of high school.

I applied the DSEE model (using the same Pillar models identified in the Mean model approach) to 64 Target school districts (n = 758,379) for which data were available. These districts had considerable variation in size, graduation rate, and degree of missingness of data (and which features were missing), with values for these variables that were substantially higher or lower than the values for the Pillar districts. As such, applying models from the Pillar districts to these sixty-four Target districts represents substantial extrapolation. Note that the goal of DSEE is not just to provide models for these seventy districts, but even more for the large number of additional districts that do not have sufficient historical data available to be able to develop a model at all (for which we may not have the data to measure how well they work).

However, these 64 districts are generally representative -- in terms of their range of size, data quality, and demographics -- of the range of districts that DSEE could be applied to (see Appendix A).

To calculate the district-to-district scores, the similarity data was first cleaned and prepared. This process involved first removing any identifying metadata not used in the calculation and then coercing the entire data sate to a numeric data type to identify any missing values. There are several common strategies used to address missing data in cluster analysis (Zhang, Zhang, Zhu, Qin, & Zhang, 2008). For this analysis, simple value imputation as used where a value of 0 was imputed for all missing values. To address the differences in measurement scales (ex. population counts vs percentage rates) and ensure equal weighting during the distance calculation, the raw numeric values were converted to normalized z-scores per column for each variable included in the distance data. Figure 8 below provides a visual representation of the normalized values for each feature used within the distance calculation. From this image, we see that districts with low high school graduation rates, are often the districts with low employment rates as well. Additionally, current student population ethnic distributions shift significantly within each district. Districts containing large Hispanic or Black student rations generally see lower graduation rates, with majority White and Asian student districts seeing higher graduation rates. Urbanicity is diverse, with some districts containing mostly rural students and other's containing mostly urban student. These values are then used to calculate the final similarity distances and weights used within the DSEE EWS method.

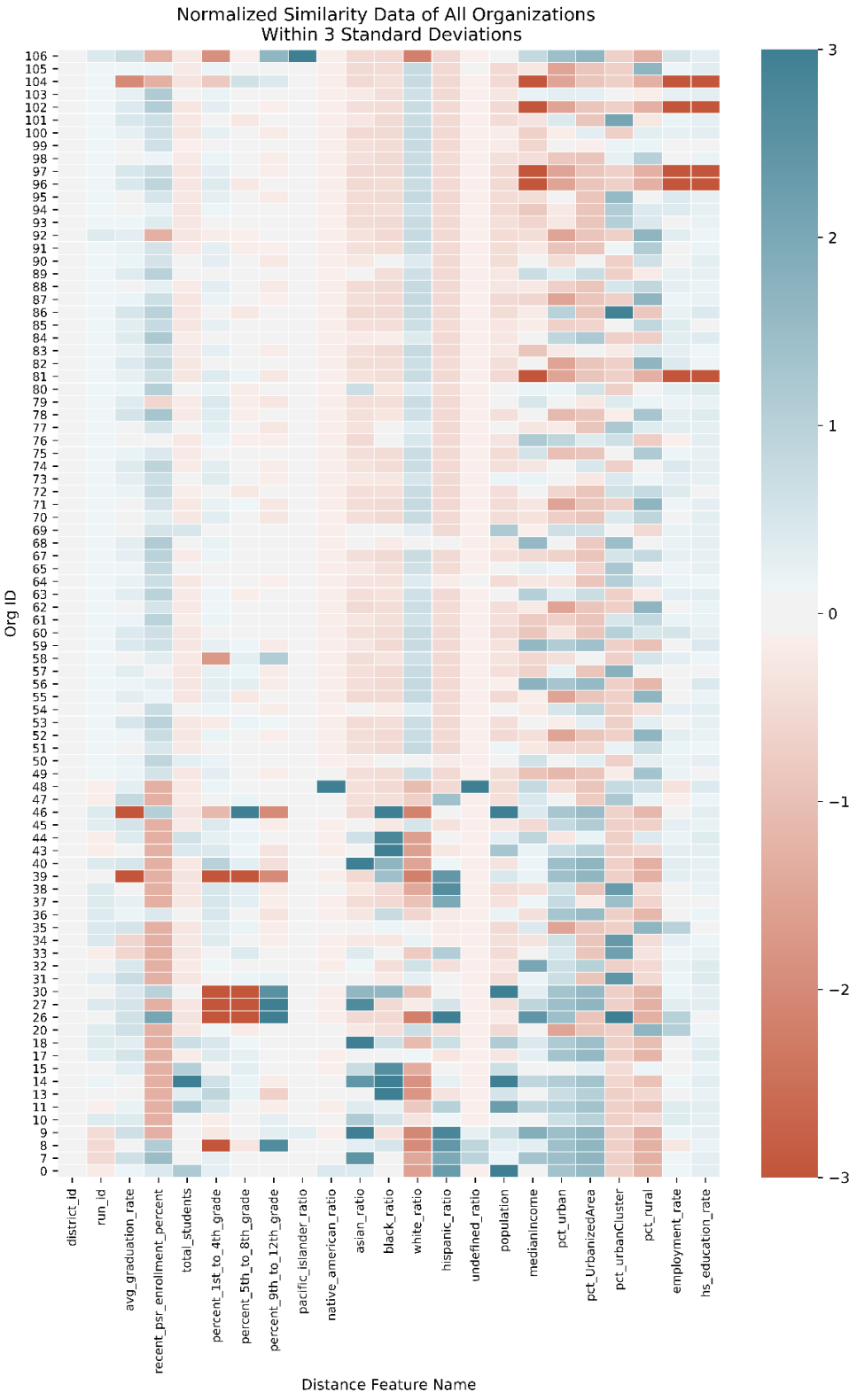



**Figure 8: Normalized Features After Missing Data Imputation Using Z-Score Standardization in the DSEE Calculation**

For each of the Pillar districts *P*, I calculated the Euclidian distance between each Pillar district and each Target district *T* using the available demographic (ethnicity distribution, district size, urbanicity, etc.) features.

$$Ea_i = dist\big(T(x_i, y_i), P(a_i, b_i)\big) = \sqrt{(x_i - a_i)^2 + (y_i - b_i)^2}$$

I then found the average distance by taking the sum of the distances by the total number of features used in the calculation.

$$\overline{Eas_\iota} = \frac{\Sigma(Ea_i)}{n}$$

The average distances were then rescaled between 0 and 1 across all Pillar and Target pairs for easier calculation.

$$Easb_i = \frac{Eas_i - min(Eas_i)}{max(Eas_i) - \min(Eas_i)}$$

To convert the distance to a similarity value, I take each distance and subtract them from 1, effectively inverting the values so that a higher value symbolizes a smaller distance between each Pillar and Target district.

$$\mathrm{NEasb_i} = 1 - \mathrm{Easb_i}$$

The goal is to have all of the similarities add up to 1, in which case I can make a prediction for a given student that will be scaled between 0 and 1 by simply summing the predictions from each Pillar model, multiplied by each Pillar district's distance. The figure 9 provides a visual representation of the similarity between each Pillar district to each Target district.

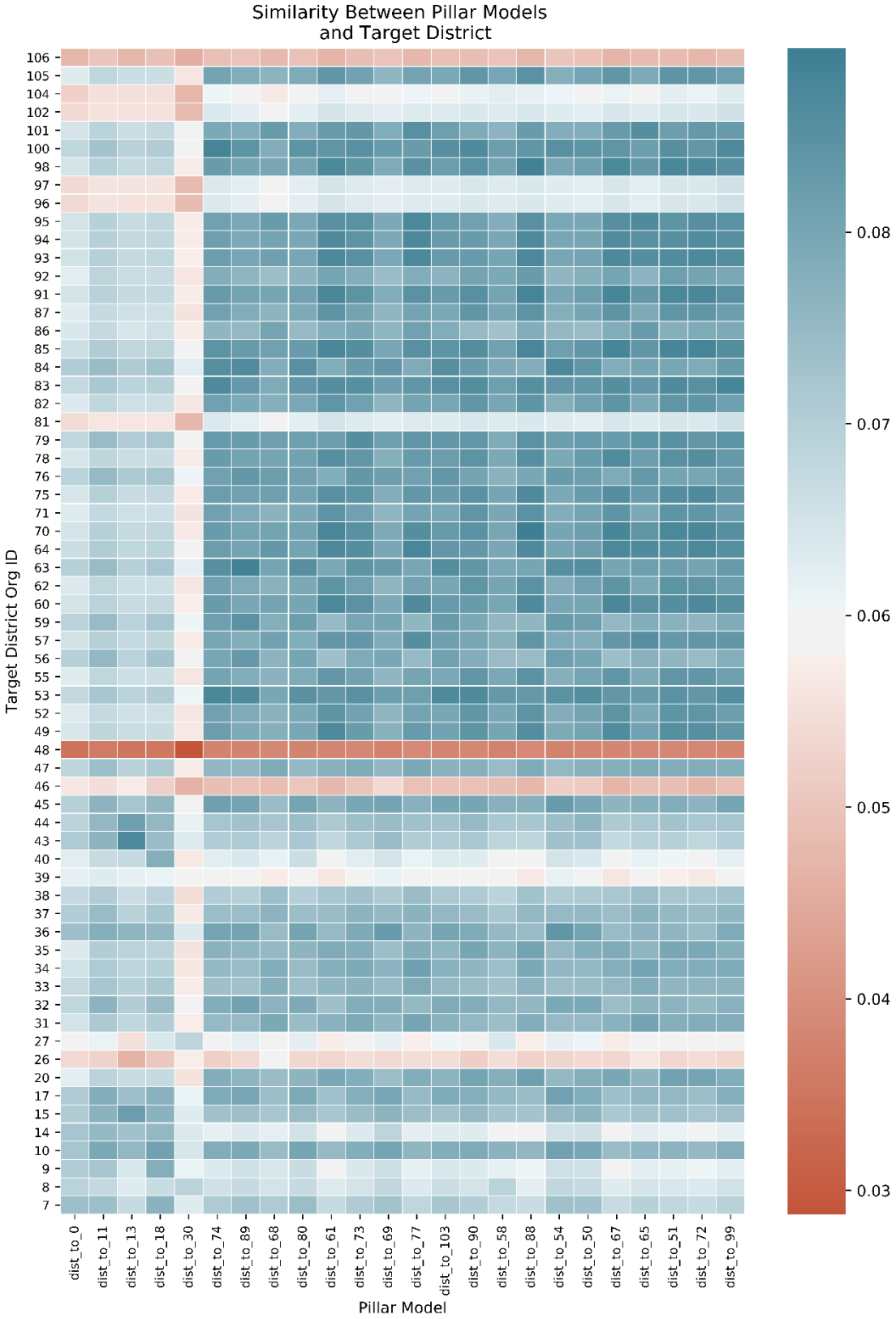

Similarity Between Pillar Models
and Target District
Target District Org ID
106
105
104
102
101
100
98
97
96
95
94
93
92
91
87
86
85
84
83
82
81
79
78
76
75
71
70
64
63
62
60
59
57
56
55
53
52
49
48
47
46
45
44
43
40
39
38
37
36
35
34
33
32
31
27
26
20
17
15
14
10
9
8
7
dist_to_0
dist_to_11
dist_to_13
dist_to_18
dist_to_30
dist_to_74
dist_to_89
dist_to_68
dist_to_80
dist_to_61
dist_to_73
dist_to_69
dist_to_77
dist_to_103
dist_to_90
dist_to_58
dist_to_88
dist_to_54
dist_to_50
dist_to_67
dist_to_65
dist_to_51
dist_to_72
dist_to_99
Pillar Model
0.08
0.07
0.06
0.05
0.04
0.03

**Figure 9: Adjusted Similarity Between Pillar District and Target District Models Using Normalized Euclidian Distance Function**

Observations from this analysis suggest that the Target 48 and 106 districts are very dissimilar to all the Pillar models and that the Pillar 30 district is dissimilar from every Target district. Target District 48 consisted of a large Native American/Indigenous population and Target 106 consisted entirely of Pacific Islander students. Pillar Model 30 contains a diverse set of students, from a large range of backgrounds. This range of diversity could be limiting the model's similarity to Target districts compared to other Pillar Models that contain a larger range of demographic feature values which could be the potential cause of these results. All other Pillar to Target similarities obtained a largely similar value, with small differences observed in the similarity values.

Predications generated by a Machine Learning model are supplied in the form of a probability, with a possible minimum value of .0 and possible maximum value of 1 (DasGupta, 2011). This presents a challenge when attempting to use the calculated similarities, as simply taking the sum of the probabilities multiplied by the similarity could potentially produce probabilities above 1 (which would mean a student is over 100% likely to drop out). To address this issue, I use iterative gradient descent (Kelley, 1999) to find a multiplier value (*M*) such that the Pillar model predictions $P_i$ multiplied by the Pillar-Target similarity N$Easb_i$ multiplied by the $M_i$ value summed together is 1.

$$f(x, M'(x)) = 1$$

$$f(x) = (NEasb_1 * M) + (NEasb_2 * M) + \ldots (NEasb_i * M)$$

Note that we need only calculate M once for each Target district as the similarity values differ between each Target and Pillar. By combining these three values, the result is a single prediction for a given student, scaled between 0 and 1, where a value of less than 0.5 is

considered a student on path to graduation and a value of 0.5 or greater is a student at-risk of dropping out of school.

$$\hat{y} = (P_1 * NEas_1 * M) + (P_2 * NEasb_2 * M) + \ldots (P_i * NEasb_i * M)$$

**Table 4: Example of Distances Between Target and Pillar Converted to Weights**

| Target District | Pillar One | Pillar Two | Pillar Three | Gradient Descent *(M)* | Pillar One Model Weight | Pillar Two Weight | Pillar Three Weight |
|---|---|---|---|---|---|---|---|
| *T1* | 0.056 | 0.063 | 0.064 | 5.433 | 31% | 34% | 35% |
| *T2* | 0.017 | 0.040 | 0.042 | 10.058 | 17% | 41% | 42% |
| *T3* | 0.026 | 0.043 | 0.044 | 8.876 | 23% | 38% | 39% |
| *T4* | 0.042 | 0.053 | 0.054 | 6.706 | 28% | 36% | 36% |
| *T…n* | $P1Eai_{sb}$ | $P2Eai_{sb}$ | $P3Eai_{sb}$ | *M* | $P1Eai_{sb}$ *(M)* | $P2Eai_{sb}$ *(M)* | $P3Eai_{sb}$ *(M)* |



**Table 3: Example of Student Level Predictions**

| Target 1 Students | Pillar One Prediction | Pillar Two Prediction | Pillar Three Prediction | Pillar One Model Weight | Pillar Two Weight | Pillar Three Weight | Final Student Prediction (> 0.5 = Dropout) |
|---|---|---|---|---|---|---|---|
| *Student A* | 0.954 | 0.888 | 0.555 | 31% | 34% | 35% | 0.792 |
| *Student B* | 0.461 | 0.624 | 0.51 | 31% | 34% | 35% | 0.534 |
| *Student C* | 0.565 | 0.985 | 0.754 | 31% | 34% | 35% | 0.776 |
| *Student D* | 0.758 | 0.113 | 0.257 | 31% | 34% | 35% | 0.361 |
| *Student E* | 0.647 | 0.554 | 0.359 | 31% | 34% | 35% | 0.514 |
| *Student…n* | *P1* | *P2* | *P3* | $P1*P1Eai_{sb}(M)$ | $P2*P2Eai_{sb}(M)$ | $P3*P3Eai_{sb}(M)$ | $\sum[P_i = P_i * P_i\, Eai_{sb}(M)]$ |

Tables 4 and 4 above shows an example of this method in practice. I applied DSEE Pillar models to Target districts, and evaluated these models using all historical records present in the data. Using all the Target student records to evaluate the model's performance was permitted as this as none of these records were used within model training (only student records from Pillar districts were used for training), effectively allowing me to treat them as the hold-out test set. As with the Pillar models, I use the AUC ROC as the metric of model goodness to evaluate the Target district student predictions.

### 3.6.4 The Chicago Model

Districts that lack enough data (Targets) to build an advanced EWS system would traditionally rely on a simpler methods of early dropout risk detection. The Chicago model On-Track indicator is a simple threshold-based EWS that relies on two freshman-year data points; the number of credits earned and the number of course (English, math, science or social science) failures within a semester. To fit this EWS to the data, a simple conditional argument can be applied using the parameters specified by the Chicago model research. The specific conditional argument used is as follows; if the student did not obtain enough credits their $9^{th}$ grade year OR if a student received at least one failure in their core courses THEN the student is off-track and at risk of dropping out of high school.

Comparing the DSEE to the Chicago model was limited by data availability, as the Chicago model relies on high school student GPA records. As such, the validation sample used to calculate the AUC was limited to students with data available in 9th grade. Due to the high missingness within the data, many of the target districts lacked data for the features outlined within the Chicago model research, for at least some students. If at least one feature was available for the Chicago model, the model was used; a student was assigned a default .5

probability of graduating if the Chicago model was missing all features and therefore incapable of producing a prediction. In practice, the Pillar models also performed more poorly for students with very high data missingness compared to students with lower missing values in their data records. Of the 126,650 unique total students across the 64 Target school districts, 124,942 contained 9th grade data records used to calculate the On-Track indicator, a reduction of 1.3 percent. Of this population, 46.6 percent did not contain credit data, 27.52 percent did not contain math course data, 27.17 percent did not contain reading course data, and 29.4 percent did not contain social science related data.

### 3.6.5 The Balfanz Model

The Balfanz model is similar to the Chicago model in that it is also a threshold based EWS that relies key data point values to identify students at risk of dropping out. The Balfanz model deviates from the Chicago model approach by leveraging four indicators to generate risk, implemented beginning at the 6th grade level compared to two indicators in the Chicago model implemented at the 9th grade level. These indicators are: student grade in mathematics, student grade in reading/Language Arts, student attendance, and student behavior. To fit this model to my data, a simple conditional statement using the parameters specified by the Balfanz model was applied on 6th grade student records. The specific conditional argument applied is as follows; IF a student obtained final grade of *F* in mathematics OR a final grade of *F* in English/Language Arts OR a attendance below 80 percent for the year OR a obtained final "unsatisfactory" behavior mark in at least one class THEN the student (6th grade or higher) is marked as at-risk.

Like the Chicago model, if at least one feature was available for the Balfanz model, the model was used; a student was assigned a default .5 probability of graduating if the Balfanz model was missing all features and therefore incapable of producing a prediction. Of the 126,650

unique total students across the 64 Target school districts, 38,715 contained 6th-grade data records used to calculate the On-Track indicator, a reduction of 69.43 percent. Of this population, 33.13 percent did not contain math course data, 32.19 percent did not contain reading course data, and 5.95 percent did not have attendance related data. All student records contained behavioral data, likely due to the system used to collect these records as it automatically defaulted to a value of 0 if the school district provided no incident records.

### 3.6.6 The Knowles Model

The Knowles model utilizes traditional machine learning techniques to build high school dropout models. It differs from previous research in that it does not rely on one single algorithm to create the detector but instead produces multiple detectors using different machine learning algorithms and then selects several (anywhere from 4 to 7) of the best performing models and averages them together to create a single detector as the final risk prediction mechanism (Knowles, 2015). As mentioned earlier, I encountered several limitations when attempting to apply the original modeling code (written in the R programming language) published by Knowles to my research data. The decision to copy the Knowles method rather than try and fit his published code to my data was due to three primary issues I encountered. The first issue is that the libraries and packages called by the Knowles published code have not been kept up to date and are significantly deprecated. The second issue is that Knowles created the codebase for the State of Wisconsin school system, with the expected features hardcoded in the provided functions. Lastly, the code relies on dependencies that are not available to the public (i.e., other libraries and packages only available to Wisconson educational researchers). Lacking the capacity to resolve these issues, I was opted to replicate his approach using newly created code using the same R language and similar statistical techniques.

While my codebase is different than the original published implementation, I follow the same 4 step process as Knowles consisting of 1) combining all the available data and building a test and training set, 2) fitting a range of models using different algorithmic approaches, and then evaluating the model's performance via cross-fold validation, 3) identifying the best performing models on the test data, and 4) selecting the N top best performing models and ensembling them together into a single EWS detector.

Preparing the data for the Knowles model began by first combining all available student records together into one large single data set. To mimic the approach by Knowles, this data was then filtered to only include students with $6^{th}$ grade records, reducing the total unique student records by 0.0067 percent from 326,533 to 324,345. Missing values were then imputed using mean value replacement for each column (i.e. an average value was calculated for each predictor, and then used as a replacement for students that had no data present for that specific variable). Lastly, the data was then randomly split into a training set, consisting of 75 percent of the data (used for training and cross-validation), and a test set, consisting of the 25% of records used for validation. The next step involved fitting many different models using a supplied list of defined algorithmic approaches to search and identify the best performing methods. This step was computationally intensive, taking several weeks to complete using the resources at my disposal. These algorithms were fit using 10-fold cross validation during training to identify the optimal parameters for each approach. Once the optimal model was fit, the test set was then scored against the model to calculate the final performance using the AUC metric.

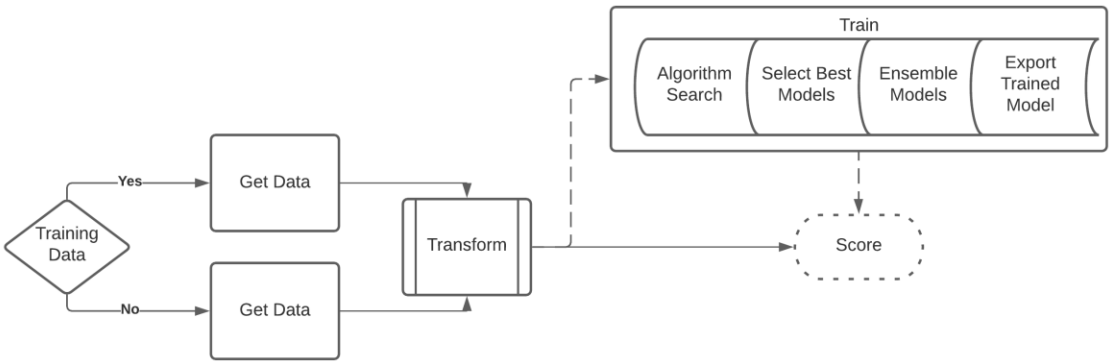



**Figure 10: Flow Diagram of Knowles EWS Modeling Process**[3]

After reviewing the results, the top four (gradient boosting machine, random forest, multivariate adaptive regression spline, and elastic-net regularized logistic regression) performing models were then selected to be combined into a single dropout risk estimator, which was then validated using the test holdout data. Interestingly, the highest performing model reported by Knowles was also the gradient boosting machine, with multivariate adaptive regression spline, and elastic-net regularized logistic regression also performing scoring high (but not in the top 4). The performance of the random forest algorithm was not reported by Knowles (Knowles, 2015). Appendix I provides a comprehensive analysis of results, with AUC performances provided for all algorithms searched during this step.

### 3.7 Measuring Feature Importance

To understand which features are particularly important to the model, a feature importance score was calculated. Feature importance was calculated using the mean decrease impurity method, sometimes referred to as the Gini importance (Breiman, & Cutler, 2007). The Gini importance measures the probability of misclassification if it was randomly classified

[3] Source: Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin. JEDM | Journal of Educational Data Mining, 7(3), 18–67.

according to the distribution of values in the features. The Gini importance value can be created by taking the sum of the probability $p(i)$ of picking a datapoint value with a true class $C$ multiplied by the probability of a mistake in the model predicting the class for this datapoint (Nembrini, König, & Wright, 2018).

$$Gini = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

The Gini importance was calculated using the *sci-kit learn* python library for each machine learning driven EWS (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, ... & Vanderplas, 2011). Creating this metric allows me to calculate how much each feature contributes to the model's eventual predictions of a student's outcomes (in this case, risk of dropout) (Breiman, 2001). The Gini importance serves two primary functions; the first is that it can provide additional insights into improving the models design through better feature selection (Katuwal & Chen, 2016) and the second is that it can serve as a basic form of model interpretability for educators who implement an advanced form of early warning system in their school district (Chung, & Lee, 2019).

### 3.8 Calculating Prediction Equity

When generating risk predictions to detect the likelihood of high school dropout, we want to make sure the EWS not only performs well overall, but also is not biased against members of specific groups (Yordanova, & Emanuilov, 2020) as there is the potential to cause unwanted harm to students when implementing any predictive risk system within an educational systems (Bird, Dudík, Edgar, Horn, Lutz, Milan, & Walker, 2020). If an EWS does not perform as expected within various groups, it could lead educators to inadvertently deny students access to services, resources, or interventions that would have improved their educational outcome

(Christenson & Thurlow, 2004). Additionally, for students that do have access and receive an intervention, there is the risk the quality of the service could be reduced based on inequitable model performance. For example, denying a student access to an after-school study program based on their gender or ethnic background harms the student and is illegal (Pinkus, 2008). Similarly, imagine that a student does get access to the after-school study program, but the EWS suggests the student is not as likely to drop out as the other students participating in the same program. This may lead the educator leading the study program sessions to decide to prioritize their time supporting the other students. The student still receives the intervention service, but the quality of that service is reduced.

To understand how fair each EWS replicated in this research is performing between student demographic groups, I subset the predictions based on either the student's ethnicity/race or gender and then calculated the AUC within group for each model. By calculating the model's performance within various demographic populations, I can better understand if some models underperform when generating student risk predictions of high school dropout. Understanding if or where the EWS is biasing predictions is the first step to implementing strategies to reduce this effect. While this dissertation does not specifically implement or test these strategies, I will outline potential methods that can be used to mitigate and address algorithmic bias further in the Conclusion & Discussion chapter of this document.

The prediction equity AUC comparison is completed in three separate steps based on the grade of target population for each EWS replicated in this research. The first comparison is made between the Aggregated Data Model, the Mean model, and the DSEE model for all 1$^{st}$ through 12$^{th}$ grade risk predictions. The second set of comparisons is made between the Aggregated Data Model, the Mean model, the DSEE model, the Balfanz model, and the Knowles model for all 6$^{th}$

through 12th grade risk predictions. Lastly, the final set of AUC comparison calculations is made for the Aggregated Data Model, the Mean model, the DSEE model, thh Balfanz model, the Knowles model and the Chicago model for all 9th through 12th grade student risk predictions.

# Chapter 4: Research Findings

In the following sections, I will discuss the results and performance of the Aggregate Date Model, Mean model, DSEE, Chicago model, Balfanz model, Kowles Model and Bowers GMM Early Warning Systems. EWS results are provided in both an overall value (across all students) and within-grade AUC for each relevant EWS population. Additionally, I provide summary results of my equity analysis, highlighting the variance of performance for each EWS created across both students reported ethnicity and student reported gender. Table 5 provides a summary of each EWS evaluated in this research for reference.

**Table 5: Description of Early Warning Systems Evaluated in This Research**

| EWS Name | Description |
|---|---|
| Aggregate Data model | Single model created by combining all student records across all available school districts. |
| Mean model | Single detector created using average prediction of multiple models from select districts (Pillar), used to generate risk for students in other districts (Targets) incapable of creating their own model. |
| DSEE | Single detector created using the weighted average of predictions using similarity features of multiple models from select districts (Pillar), used to generate risk for students in other districts (Targets) incapable of creating their own model. |
| Chicago model | Threshold based EWS that relies on two freshman-year data points |
| Balfanz model | Threshold based EWS that relies on four $6^{th}$ grade year data points |
| Knowles model | Single detector using advanced machine learning technique of stacking multiple models together. |
| Bowers GMM | Growth Mixture Model on non-cumulative GPA for 9th-grade students |

## 4.1 Results of the Aggregate Data Model

Initial results of the Aggregate Data model showed acceptable performance (Mandrekar, 2010), with an overall AUC of 0.76 across all grade levels. The performance of the detector

within grade levels varied significantly, with performance increasing over time as the students get closer to possible dropout. The lowest performance is seen when making first grade dropout predictions, which shows an AUC of 0.637. The Aggregate Data model performs the best at grade 12, with an AUC of 0.83. Additionally, there is a significant observed drop in performance before grade 5, suggesting that generating predictions for early year students could produce significantly more levels of errors compared to later year students. The figure below provides a graphical illustration of AUC performance across all grade levels, using data calculated within grade.

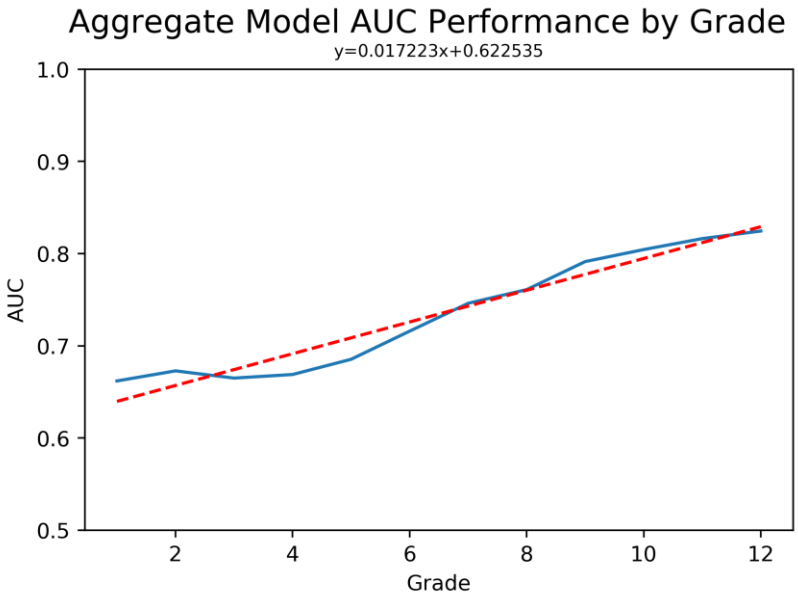



**Figure 11: AUC Performance of Aggregate Data model within Grade Levels**

While these initial results may be promising as it beats the performance of many existing EWS (Bowers & Zhou, 2019), the model performance at the district levels suggests severe issues generalizing these risk predictions across all organizations within the data. We see that despite the initial results of 0.76 of the combined test data, the true average performance is 0.696 (SD=0.06) when calculating the AUC within each organization's population, with some districts (Org 10 & 38) achieving an AUC as low as 0.49, suggesting that a coin flip would be better at

predicting dropout risk than this model method for these districts. (See Appendix E). Performance increased for districts with a larger number of records, suggesting that the fitted model was biased towards districts that provided the large rest number of records during training.

Evaluating the Aggregate Data model performance within grade produced less than optimal results and further highlighted the severity of the low data quality for several school districts within this research. Figure 12 below provides a visualization of the AUC performance within each school district calculated within each class number record, when historical outcome data was available. As the heat map shows, the performance of the model varies significantly depending on the grade of the prediction and the within which organization the prediction is made. Additionally, the amount of white space in this visual highlight areas where no historical records were available in the test set to create a prediction, indicating that many organizations do not have twelve years of historical data. This presents a significant challenge when attempting to build an organizational specific model as the ability to measure performance and bias for current student predictions is hindered for grades that do not have the historical data required to validate the model.

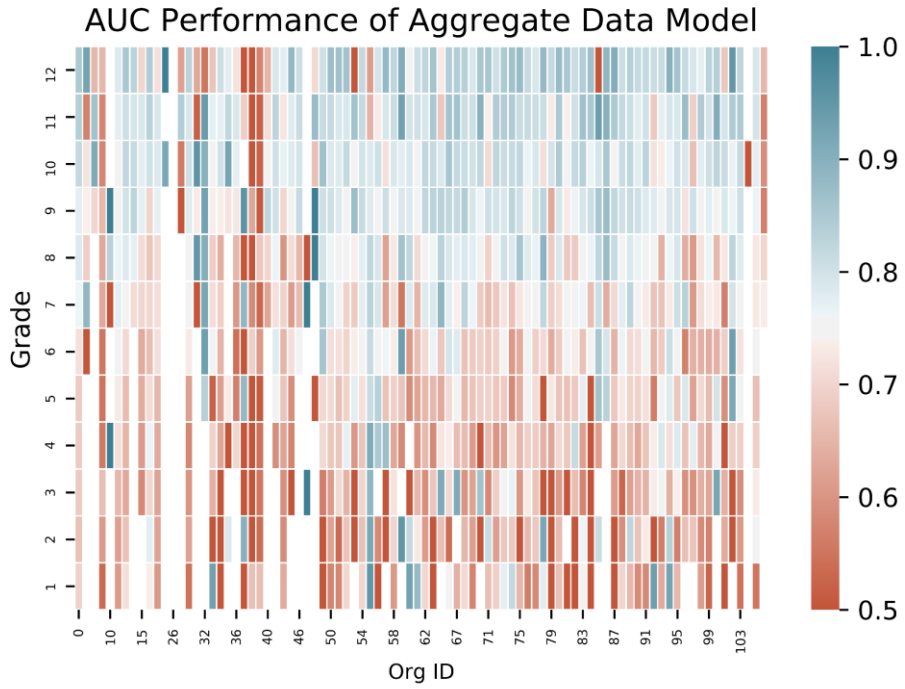



**Figure 12: AUC Performance of Aggregate Data model by Organization and Grade Level 1st Through 12th**

4.1.1 Aggregate Model Feature Importance

To understand which features are particularly important to the model, a feature importance score was calculated. Feature importance was calculated using the mean decrease impurity method, sometimes referred to as the gini importance (Breiman, & Cutler, 2007). A range of different types of features were found to be important in the Aggregate Data model.

**Table 6: Aggregate Data model Gini Feature Importance**

| Feature Name | Importance |
|---|---|
| gpa_for_grade_band | 0.15495 |
| avg_all_course_grade | 0.11080 |
| attend_ratio | 0.07724 |
| sum_absent_ratio | 0.06515 |
| pass_rate | 0.06235 |
| chronic_absent | 0.05505 |
| sum_attendance_ratio | 0.05154 |
| norm_age_for_class_number | 0.03933 |

| Feature Name | Importance |
|---|---|
| avg_core_course_grade | 0.03244 |
| count_minor | 0.03179 |
| count_major | 0.02845 |
| norm_avg_all_course_grade | 0.02666 |
| stddev_elective_grade | 0.02116 |
| avg_credits_earned | 0.01949 |
| stddev_core_course_grade | 0.01926 |
| total_absent_in_first_90 | 0.01626 |
| grad_credit_ratio | 0.01466 |
| avg_reading_grade | 0.01366 |
| avg_reading_norm_summative_score | 0.01185 |
| avg_math_norm_summative_score | 0.01127 |
| total_absent_in_first_60 | 0.01116 |
| stddev_credits_earned | 0.01050 |
| total_absent_in_first_45 | 0.00857 |
| avg_science_grade | 0.00803 |
| avg_social_science_grade | 0.00778 |
| avg_math_grade | 0.00754 |
| attnd_100 | 0.00695 |
| algebra_passed | 0.00579 |
| total_absent_in_first_30 | 0.00549 |
| norm_grad_credit_ratio | 0.00511 |
| norm_avg_math_grade | 0.00505 |
| norm_avg_elective_grade | 0.00474 |
| sum_reading_grade | 0.00462 |
| absent_in_first_90 | 0.00421 |
| count_science | 0.00398 |
| count_reading | 0.00390 |
| absent_in_first_60 | 0.00365 |
| count_social_science | 0.00354 |
| avg_science_norm_summative_score | 0.00312 |
| avg_reading_norm_interim_score | 0.00297 |
| norm_avg_reading_grade | 0.00276 |
| absent_in_first_45 | 0.00273 |
| avg_social_science_norm_summative_score | 0.00254 |
| count_math | 0.00243 |
| norm_avg_social_science_grade | 0.00228 |
| sum_tardy_ratio | 0.00213 |
| absent_in_first_30 | 0.00142 |
| norm_avg_science_grade | 0.00095 |
| stddev_science_grade | 0.00058 |
| avg_math_norm_interim_score | 0.00056 |

| Feature Name | Importance |
|---|---|
| stddev_reading_grade | 0.00051 |
| stddev_social_science_grade | 0.00045 |
| stddev_math_grade | 0.00033 |
| avg_science_norm_interim_score | 0.00030 |
| avg_social_science_norm_interim_score | 0.00000 |

Overall, the Aggregate Data model relied heavily on features related to attendance and academic credit achievement, with attendance ratio, non-cumulative GPA within grade, average course grades, and attendance ratios providing the most importance for the model. Assessment related features, such as Interim and Summative test scores, and course subject specific performance were valued the least within this model. While this model did produce a relatively acceptable model, defined by achieving an AUC between 0.7 and 0.8 (Mandrekar, 2010), the reliance on these specific features presents a challenge when generalizing to students that do not have this data available, likely leading to some districts receiving lower predictive performance (ex: organization 37 or 106).

## 4.2 Results of the Mean Model

The following section will discuss the results of the Pillar selection and final Mean model performance on Target districts. The first step to creating the Mean model is identifying potential district candidates for which a model can be created. The next step is creating and validating these models to determine which Pillar models will be pooled together and used to generate risk predictions for districts that are unable to create their own district level model. Finally, the pool of models is used to generate risk for these Target districts by taking a simple average of the predictions across each Pillar detector, creating the final Mean model prediction.

4.2.1 Pillar Selection

Districts for which I was able to produce a model with AUC higher than 0.7, averaged across all student grades, were then designated as Pillar districts/models and used to create predictions for students in districts for which models could not be generated for all grade levels, or for which models were insufficient in quality. Interestingly, the pillars that experienced the largest amount of missingness were also those that generally reported lower AUC.

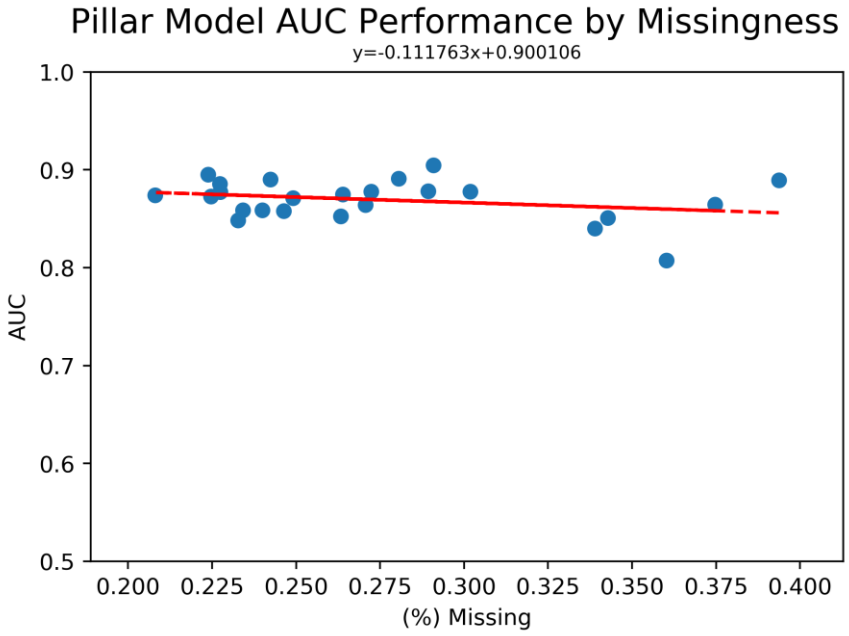



**Figure 13: AUC Performance of Pillar Models on Test Hold-Out Compared to % of Missing data in Feature Space**

All potential Pillar Candidates achieved an average AUC above 0.7 (M=0.862, SD=0.291), so no district models were reclassified at this point from Pillar to a Target, resulting in a final Pillar district model count of 24 and Target district count of 64.

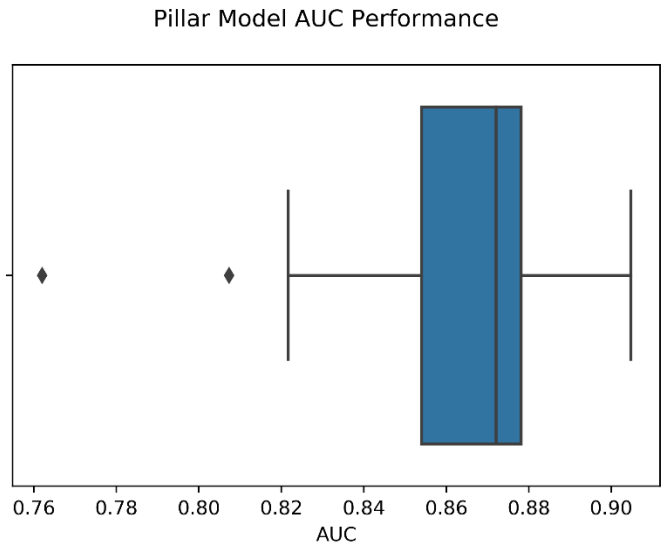



**Figure 14: Average AUC Performance of Pillar Organization Models on Test Hold-Out Data During Model Training**

The lowest Pillar model received (Org 9) a 0.762 AUC score and the highest (Org 13) received an AUC of 0.904 when validated across all grade levels within district (see Appendix D). When results were calculated within grade and within district, the variance of AUC performance shifts significantly depending on the district and the grade in which the prediction is being validated. Figure 15 below provides a breakdown of this performance.

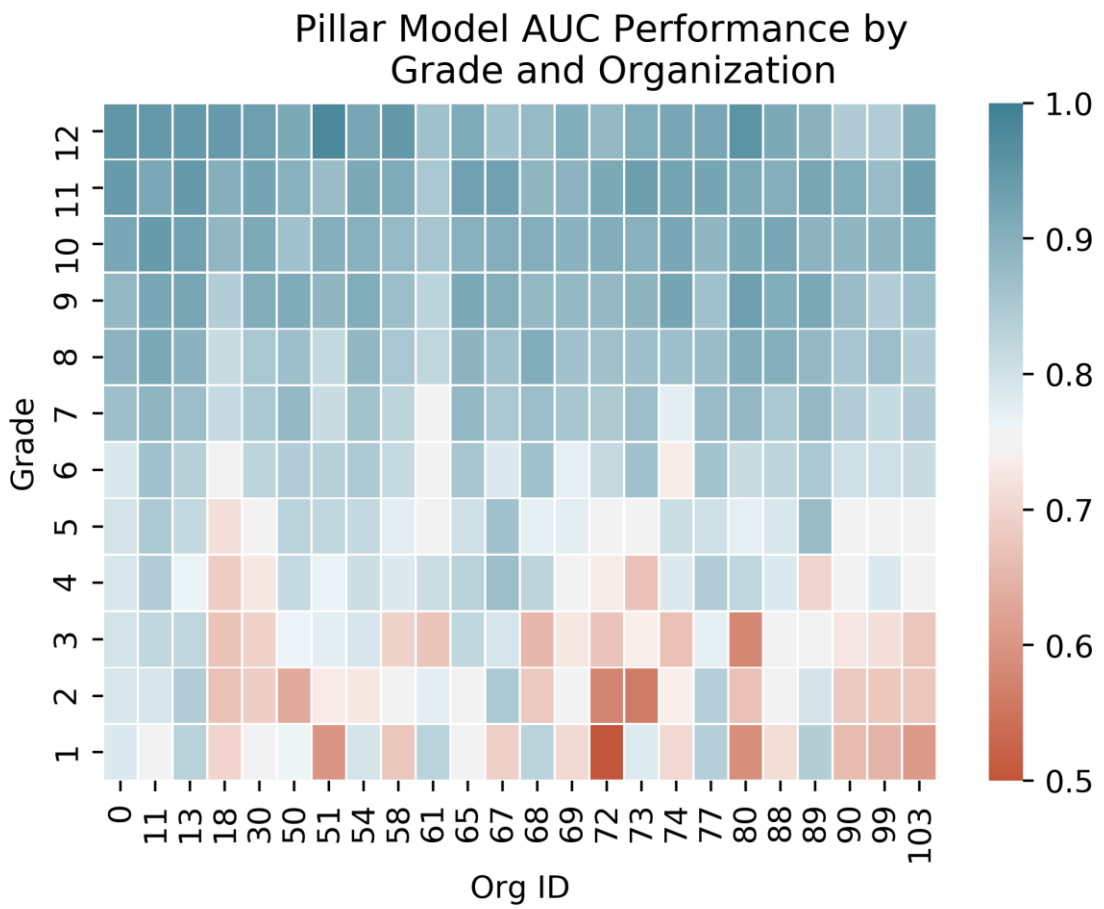



**Figure 15: Performance of Pillar Model AUC within Grade and Organization**

Calculating the AUC performance within grade and district of the Pillar models allowed me to better understand how these models will perform when generating predictions to the current student populations at the student grade level. For example, Org 80 generally performs lower when creating predictions for students in lower grades (0.588 AUC), but performs relatively better at higher grades (12$^{th}$ grade AUC of 0.898). While not explored in this research, this analysis could uncover alternative methods of designating Pillar Models, with the potential to create grade specific Pillar Model pools to generate student risk predictions using the models that performed the best within each grade.

4.2.2 Mean model Performance on New Districts

I applied the Pillar models to every record available for each student's data from the 64 Target districts (758,379 historical records) and averaged the probability across models for each student. These districts had considerable variation in size, graduation rate, and degree of missingness of data (and which features were missing), with values for these variables that were substantially higher or lower than the values for the Pillar districts. In other words, applying models from the Pillar districts to these sixty-four Target districts represents substantial extrapolation.

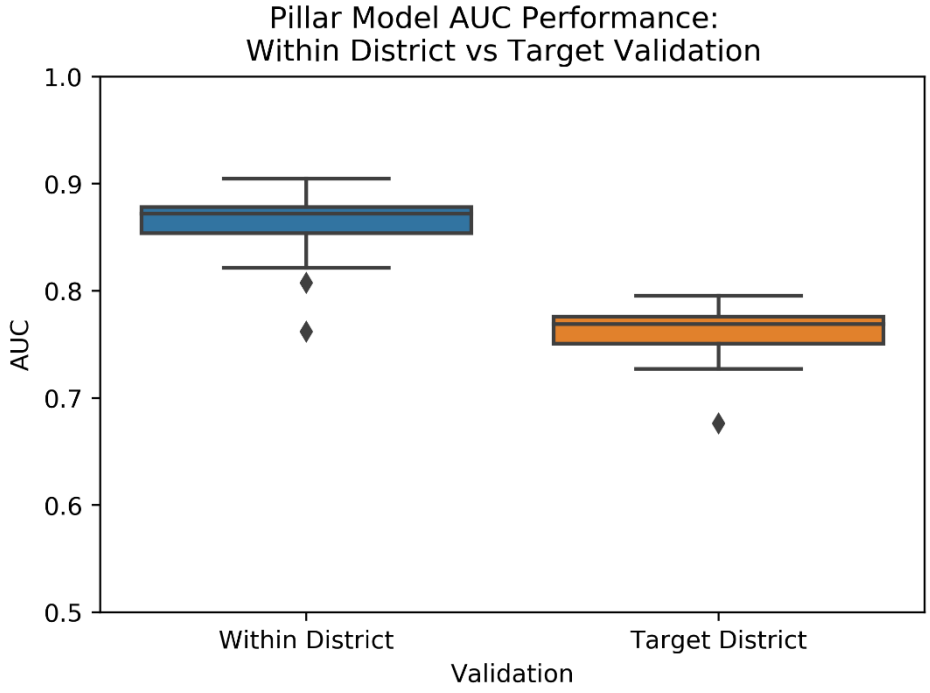
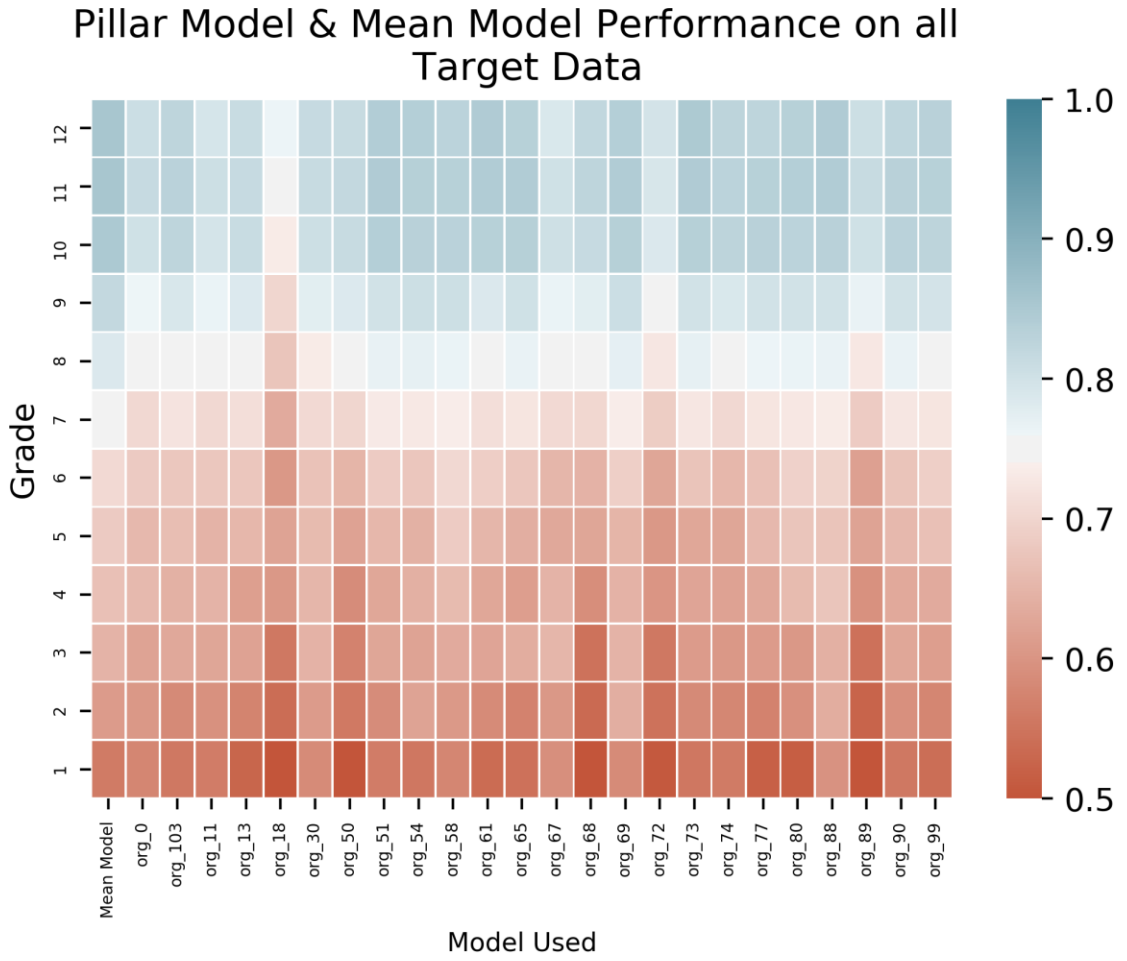



**Figure 16: Model Performance of Individual Pillars and Mean model on Target Data**

The figure above shows the average performance of each individual Pillar model detector on the Target district data compared to the Pillar model performance of the within-district validation data. The performance of these models when applied to new, unseen districts is lower than their performance on within-district data, but overall performed well despite the high degree

of extrapolation required. When taking the average probability of all detectors combined (to create the Mean model) the Mean model performed the best at 0.8 AUC. While all 25 detectors




achieved relatively good AUC scores, Org 69 appears to suggest that as the false positive rate becomes greater than 90%, the false positive rate increases more quickly than the true positive, causing the ROC curve to drop below random chance (see Appendix E). AUC results within grade produced similar results, with the models performing better within higher grades and underperforming at lower grade levels. The figure below provides a breakdown of these results.

**Figure 17: Average Performance of Pillar Model and Mean Detectors on All Target District Data.**

While the Mean model, on average, outperformed the individual Pillar district model, some Target districts did perform better when scored against the individual Pillars compared to the average scores provided by the combined Mean model. For example, Target District 47 achieved a low AUC (below .5 chance) using the Mean model but performed relatively well using the Organization 0 Pillar model (a calculated AUC of 0.77). Target District 106 shows a similar trend, where the AUC performance of the Mean model (AUC = 0.72) was lower than an individual Pillar, in this case, the Organization 88 Pillar Model, where an AUC of 0.74 was observed. This pattern is observed for several additional Target Districts within the analysis. The figure below provides a full breakdown of the AUC performance within each Target organization scored against the Pillar Models and averaged Mean model. With some Pillar models exceeding the performance of the averaged Mean model, an opportunity exists to develop a method that will better implement this modeling approach by improving the way the Pillar Models are leveraged within this simple ensemble.

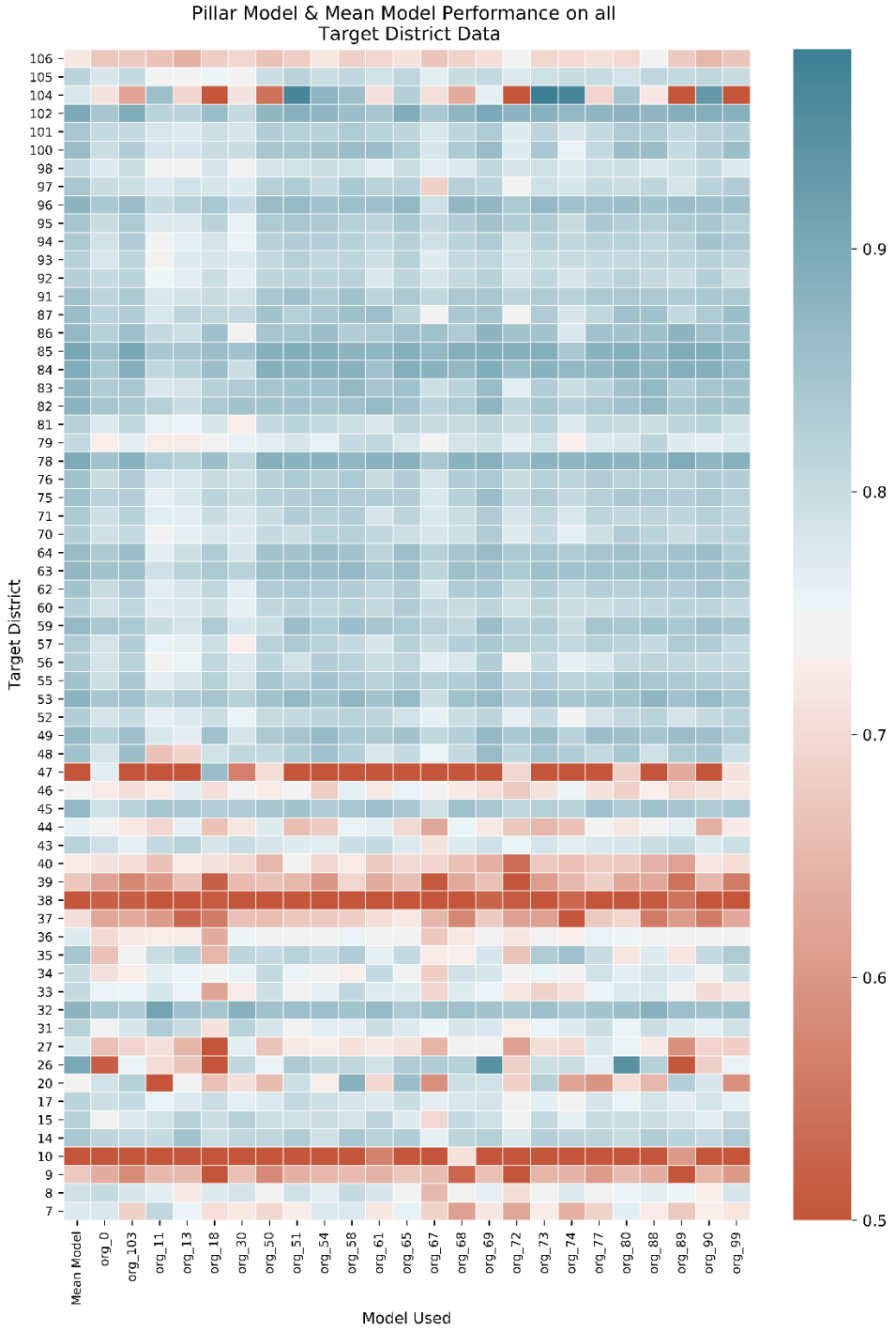



**Figure 18: AUC Performance of Pillar Model and Mean Detectors within each Target District Data[4]**

[4] White values are created when there are not enough historical records in the data to calculate an AUC score

An additional finding from these results show that some Target districts failed to produce high AUC's from any of the Pillar Models used. Target districts 38, 10 and 9 achieved AUC's close to the 0.5 chance prediction threshold for all Pillar Models. It is worth noting that these two districts had the highest rate of missing data for features that ranked most important in the Pillar models, with over 80% of students in these Target districts missing data related to coursework, over 90% of the records not containing any assessment scores, and the data for 40% of the students not containing attendance information. Overall, the districts with the highest amounts of missing data in core features were also the districts with the lowest AUC ROC values.

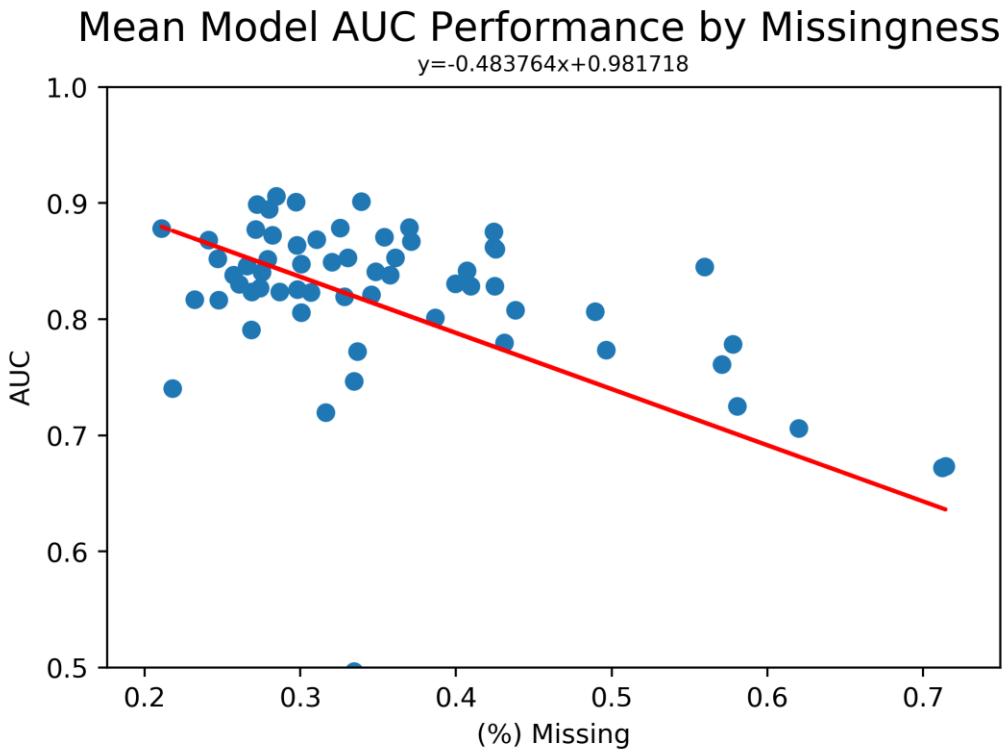



**Figure 19: Mean model Performance on Target Districts by (%) Missingness**[5]

[5] Three districts achieved AUC performance under 0.5 and are not shown in this figure. District 38 received an AUC of 0.496, District 10 received an AUC of 0.303, and District 47 received an AUC of 0.287.

4.2.3 Pillar Model Feature Importance

Like the Aggregate Data model, a feature importance was calculated using the mean decrease impurity method (gini importance) to understand which features are particularly important to each Pillar model (Breiman, & Cutler, 2007). A range of different types of features were found to be important in the twenty-four models.

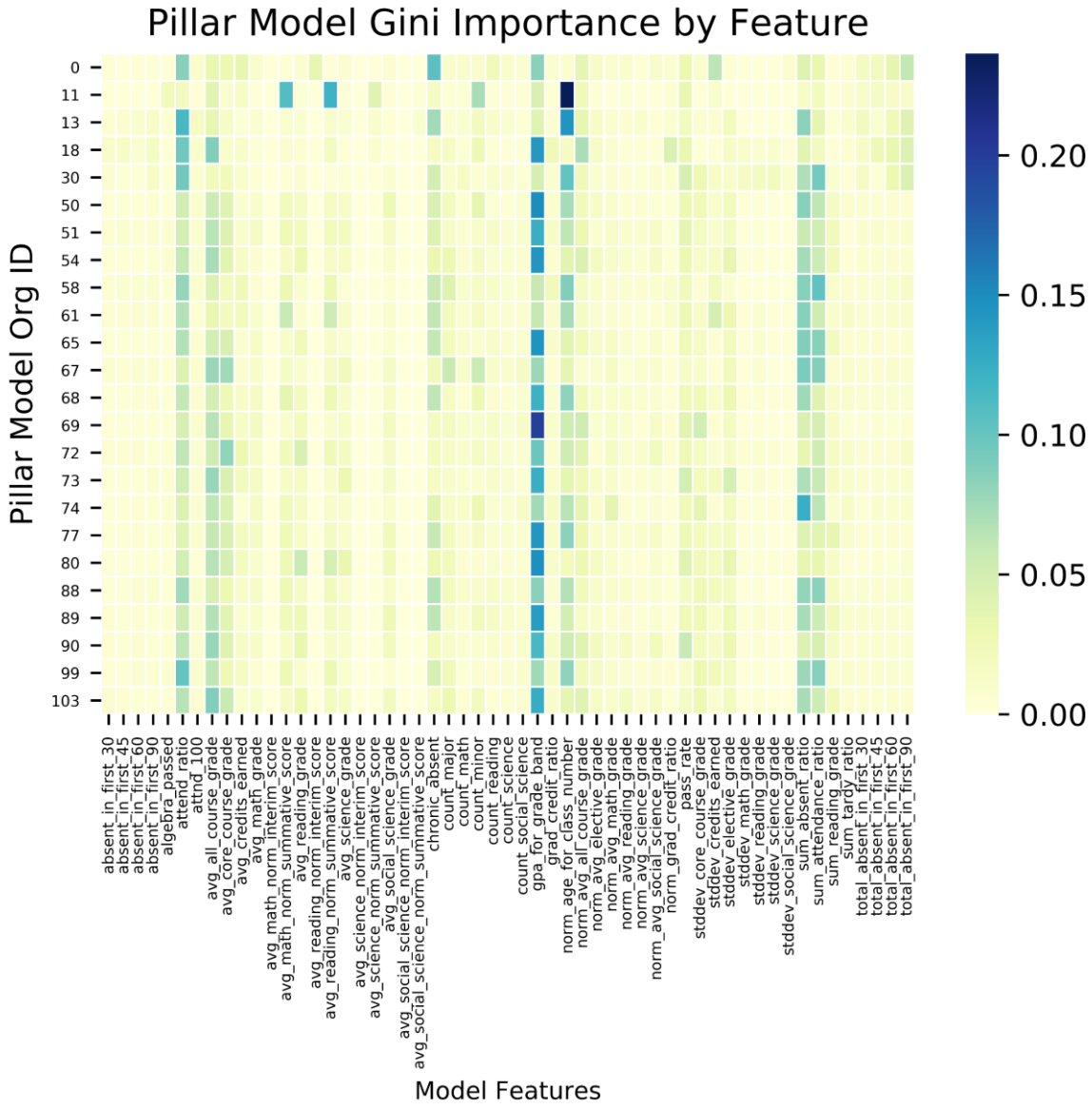



**Figure 20: Gini Feature Importance Values of Pillar District Models**

Overall, the Pillar Organization models relied heavily on features related to attendance and academic credit achievement, with attendance ratio, course grades, core course grades, GPA, and absent ratios providing the most importance for the models. While this trend was generally seen across all the models, some differences were observed. The Organization 11 model heavily valued summative assessment, student age within grade, and behavioral data (i.e., disruption, defiance, etc.) to generate risk predictions. However, student behavioral records were also important to the Organization 20 model. The Organization 67 and Organization 65 Pillar models was were most similar, with both models relying heavily on course grade data and attendance data. The differences in feature importance are likely due to a multitude of reasons. One reason could be that there were differences in the data availability of features for each district. For example, no interim assessment data was available for Organization 13, whereas Organization 0 had interim assessment data available for almost all their historical student records. Another cause could be the difference in the populations of students in each Pillar district. For example, attendance may play a larger role in graduation in urban districts (e.g. Organization 108), whereas behavioral incidents could play a larger role in the path to dropping out for students in more rural districts (e.g. Organization 67) (Jordan, Kostandini, & Mykerezi, 2012).

## 4.3 Results of the District Similarity Ensemble Extrapolation

Despite the high degree of extrapolation required, the DSEE performance was generally good, with an average AUC (across all Target districts) of 0.80, with five Target districts achieving an AUC above 0.9 across all grades.


**Figure 21: DSEE AUC Performance Across All Target District Records**

As shows in the image below, within grade AUC performance (across all Targets) achieved expected results, with the model performing worse at lower grades (0.57 in first grade predictions) and better at higher grades (0.86 in 11th and 12th grades) as the student nears the potential dropout event.

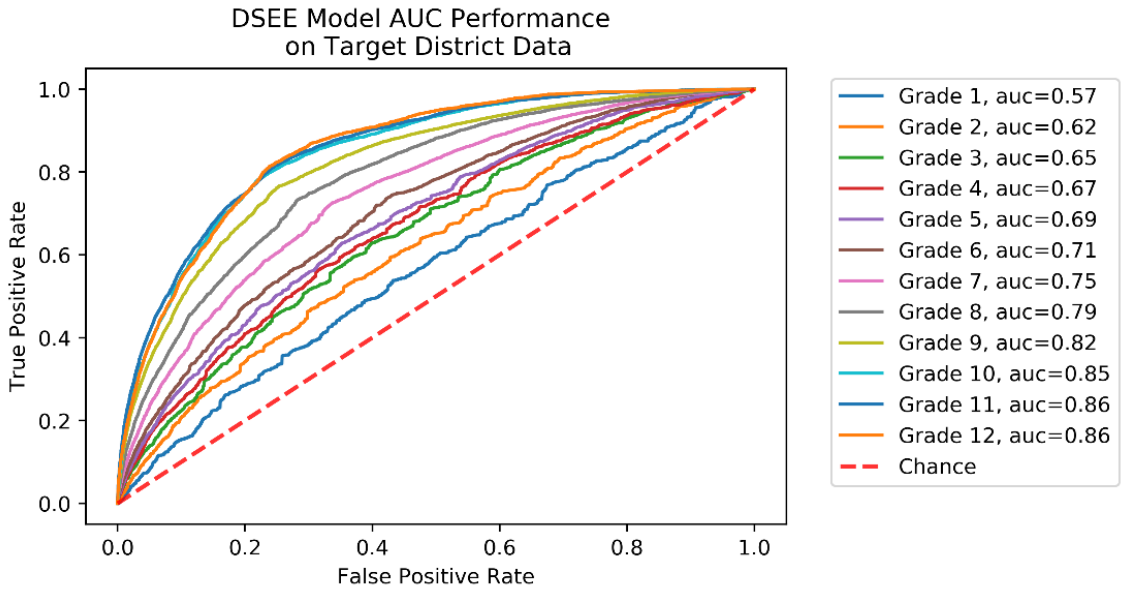



**Figure 22: Calculated DSEE AUC Performance within Grade Level**

However, three additional Target districts (districts 47, 10 & 38) achieved AUCs below 0.5. It is worth nothing that these districts had high rates of missing data, with relatively low numbers of recorded historical outcomes use to calculate the AUC. Overall, the Target districts with the highest amounts of missing data generally performed the lowest when the DSEE was applied to their data. The table below provides the AUC performance of the DSEE applied to all Target districts (see Appendix F for expanded results).

**Table 7: Summary results of DSEE AUC performance on Target District Data**

| | $\bar{X}$ | σ |
|---|---|---|
| AUC Performance | 0.805 | 0.112 |
| Count of Graduates | 10,837 | 10922 |
| Count of Dropouts | 895 | 1250 |

One potential future option for improving the modeling for these districts may be to weight the distance between districts by the degree of overlap in features available and missing, as done with the demographic features used above. This approach may become particularly useful as more Pillar models are obtained that share more feature overlap with these three districts. Another opportunity (mentioned earlier) could be using different lists of Pillar models at different the grade levels, so that only the best performing models are used to generate risk predictions for students within a specific grade.

4.4 Results of The Chicago Model

Despite the high degree of missing records in the high school Target district student records, the Chicago model On-Track indicator achieved an AUC of 0.69 across all combined 9$^{th}$ grade records, almost 0.1 points lower than the results originally reported by Allensworth and Easton (2007; Bowers & Sprott, 2012). Calculating AUC performance within-district produced slightly lower results (M=0.682, SD=0.141), with AUCs ranging from the high 0.80s to below

0.2, worse than a random coin flip. The highest AUC (0.90) found was found in Org 32, and the lowest AUC (0.0) was observed in Org 104 (see Appendix G). Due to this model's reliance on a few indicators, the performance has a linear relationship between the amount of data missing and the ability to create an accurate dropout risk prediction, shown in Figure 23.

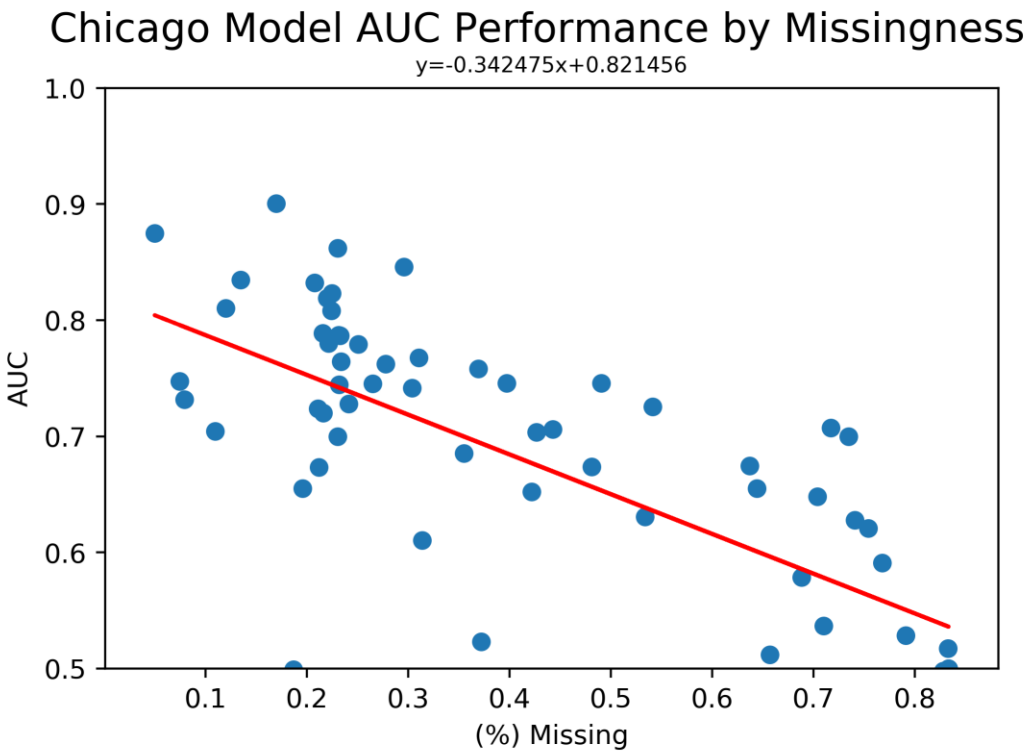



**Figure 23: AUC performance of the Chicago On-Track Indicator EWS by (%) of missing data across 9th grade students[6]**

Evaluating the Chicago model performance within district produced less than optimal results, with significant variance in AUCs observed and further highlighted the both the severity of missing data within the student records and the models capacity to generalize across populations. The figure 23 provides a visualization of the AUC performance within each school district calculated on 9th grade student records, when historical outcome data was available. As

[6] Four districts achieved an AUC below 0.5 and are removed from the figure. District 48 received an AUC of 0.499, District 9 received an AUC of 0.498, District 46 received an AUC of 0.395 and District 104 received an AUC of 0.00 (calculated on 2 outcome records).

the figure shows, the performance of the model varies significantly depending on the school district with which the prediction is made, with some school districts (Orgs 7, 9, 17, 27, 35, 40, 46, 97, & 104) receiving low AUC scores across 9th grade students. These results highlight the significant challenges of using a threshold based EWS built in one school district to create risk predictions in another, as differences in data quality and recording can severely impact the EWS's performance at detecting student dropout risk.

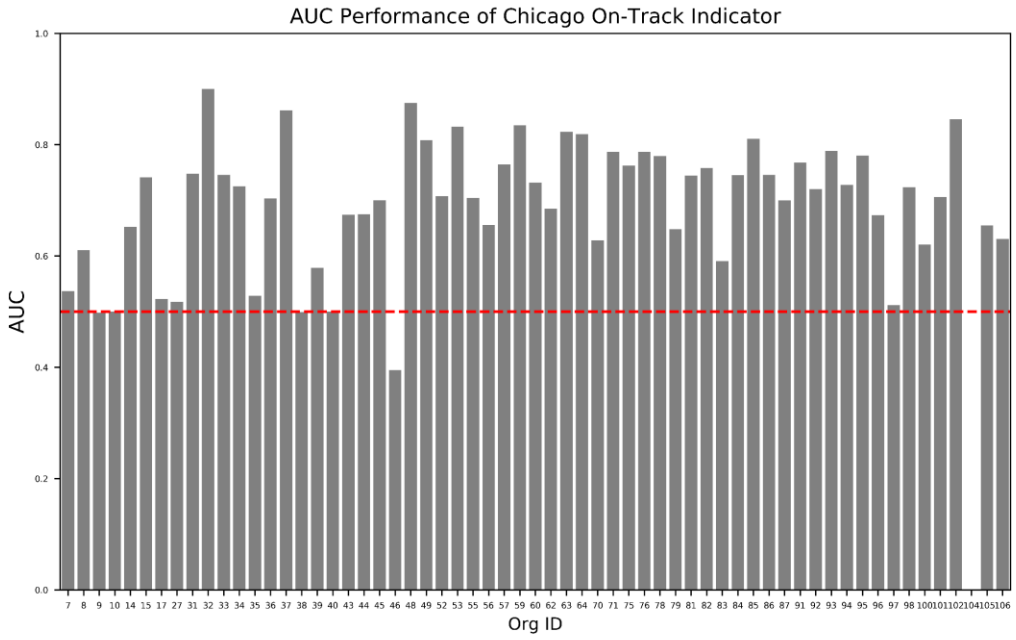



**Figure 24: AUC performance of the Chicago On-Track Indicator by School District. Red line provides a reference for a 0.5 AUC.**

4.5 Results of The Balfanz Model

The Balfanz model EWS achieved an AUC of 0.64 across all combined 6th grade records when generating risk using any of the four flags. These results are higher than the originally published performance reported by Balfanz but failed to mirror recent replicated results conducted by Bowers et. al, who reported an AUC of 0.729 (Balfanz et al.; 2007; Bowers et al.

2012; Bowers & Zhou, 2019). Calculating AUC performance within district produced slightly higher results (M=0.657, SD=0.094), with AUCs ranging from the high 0.80s to below 0.5, worse than a random coin flip. The highest AUC (0.884) found was found in Org 32, and the lowest AUC (0.463) was observed in Org 44 (see Appendix H). Like the Chicago threshold based EWS, the Balfanz model's performance has a linear relationship between the amount of data missing and the ability to create an accurate dropout risk prediction.

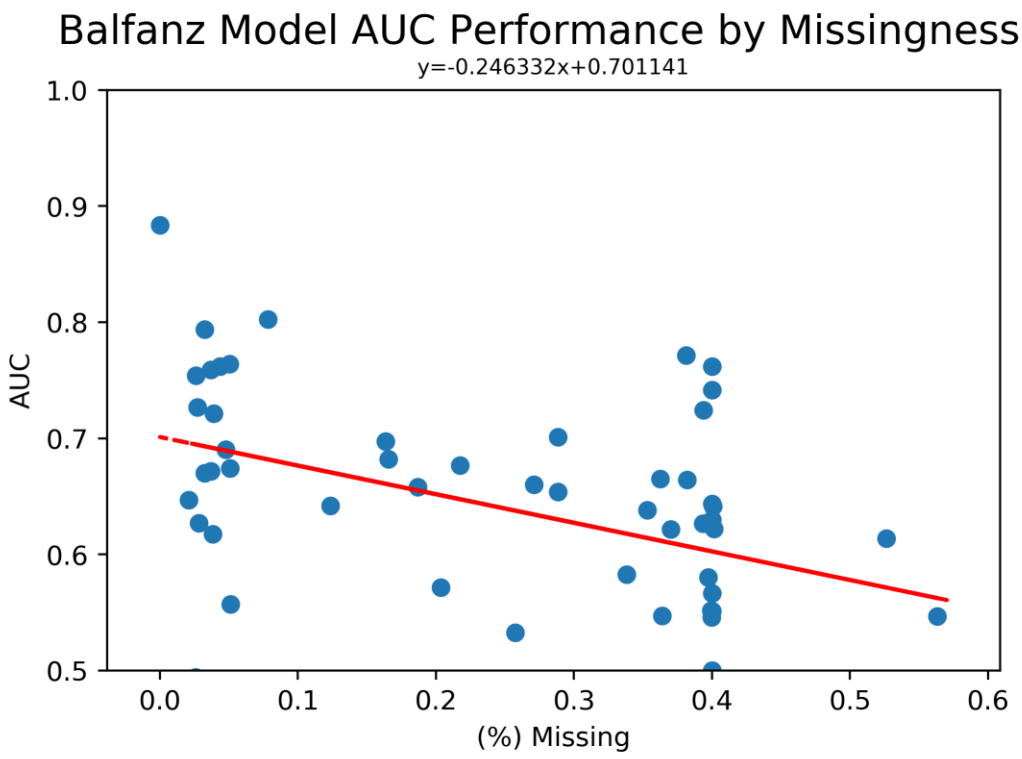



**Figure 25: AUC performance of the Balfanz EWS by (%) of missing data across 6th grade students[7]**

The similarities with the Chicago EWS continue, with the Balfanz model producing suboptimal results when evaluating the models' performance within grade and district. This finding is likely due to the severity of missing data within the student records and the model's

[7] Four districts achieved an AUC below 0.5 and are removed from the figure. District 38 received an AUC of 0.494, District 15 received an AUC of 0.478, District 96 received an AUC of 0.470 and District 44 received an AUC of 0.463.

capacity to generalize across populations. The figure 26 provides a visualization of the AUC performance within each school district calculated using 6th grade records, when historical outcome data was available. The figure suggests the performance of the model varies significantly depending on the district within which the prediction is made, with some school districts (Orgs 9, 37, 44, 47, 95 & 106) receiving low AUC scores across all 6th grade records. These results further highlight the significant challenges of using a threshold based EWS built in one school district to create risk predictions in another, as differences in data quality and recording can severely impact the EWS's performance at detecting student dropout risk.

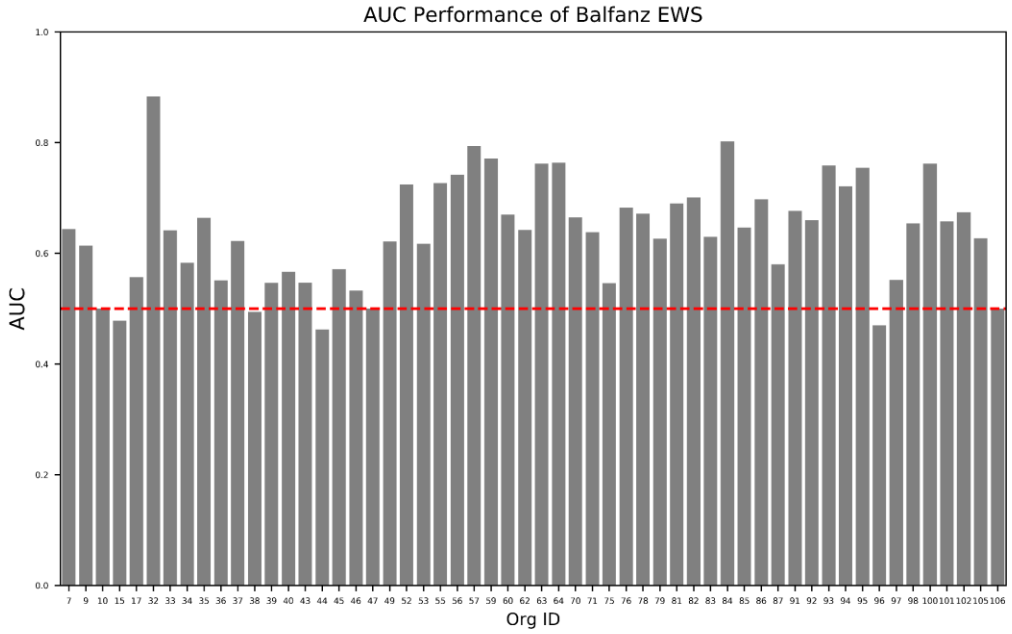



**Figure 26: AUC performance of the Balfanz EWS by School District. A red reference line is provided to show the cutoff for 0.5 AUC**

4.6 Results of The Knowles model

The individual models utilized in the final ensemble produced relatively high AUCs, with the gradient boosting machine model performance at 0.882 AUC, the random forest at 0.878 AUC, the multivariate adaptive regression spline model at 0.872 AUC, and elastic-net

regularized logistic regression model at 0.887. Combining these four models together into a single ensemble (the Knowles model) EWS achieved an AUC of 0.887 across all combined 6th through 12th grade records, performing slightly better than the original results published by Knowles, who achieved AUC between 0.83 and 0.87 (Knowles, 2015). The lowest (0.796) AUC observed for 6th grade predictions and the highest (0.899) AUC observed for 12th grade predictions.

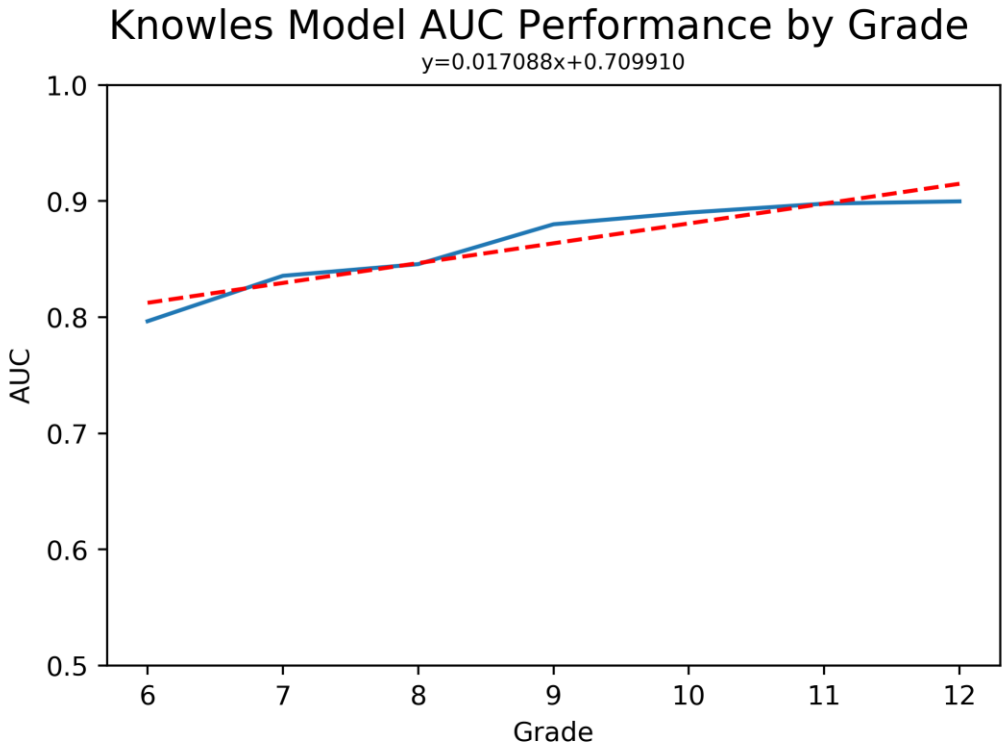



**Figure 27: AUC Performance of Knowles model within Grade Levels**

AUC performance within district produced similar results (M=0.874, SD=0.071), with AUC's ranging from 0.943 (Org 8) to a low of 0.562 (Org 38). Districts (Org 38 & Org 10) where the model underperformed (0.562 AUC & 0.569 AUC) had considerably higher dropout rates (over 50 percent) than those that generally reported higher AUC results (see Appendix I). The figure below provides a visualization of the AUC performance within each school district

calculated within each grade when historical outcome data was available. As the heat map shows, the performance of the model was generally good, but still shows some variance in performance depending on the grade of the prediction and the within which organization the prediction is made, with Org 38 receiving low AUC scores across all grades.


**Figure 28: AUC performance of the Knowles EWS by Grade and School District**

4.6.1 Knowles Model Feature Importance

Like the previously generated Machine Learning EWS models, a feature importance was calculated using the mean decrease impurity method (gini importance) to understand which features are particularly important to each individual model and the final combined ensemble

(Breiman, & Cutler, 2007). A range of different types of features were found to be important in the five total models.

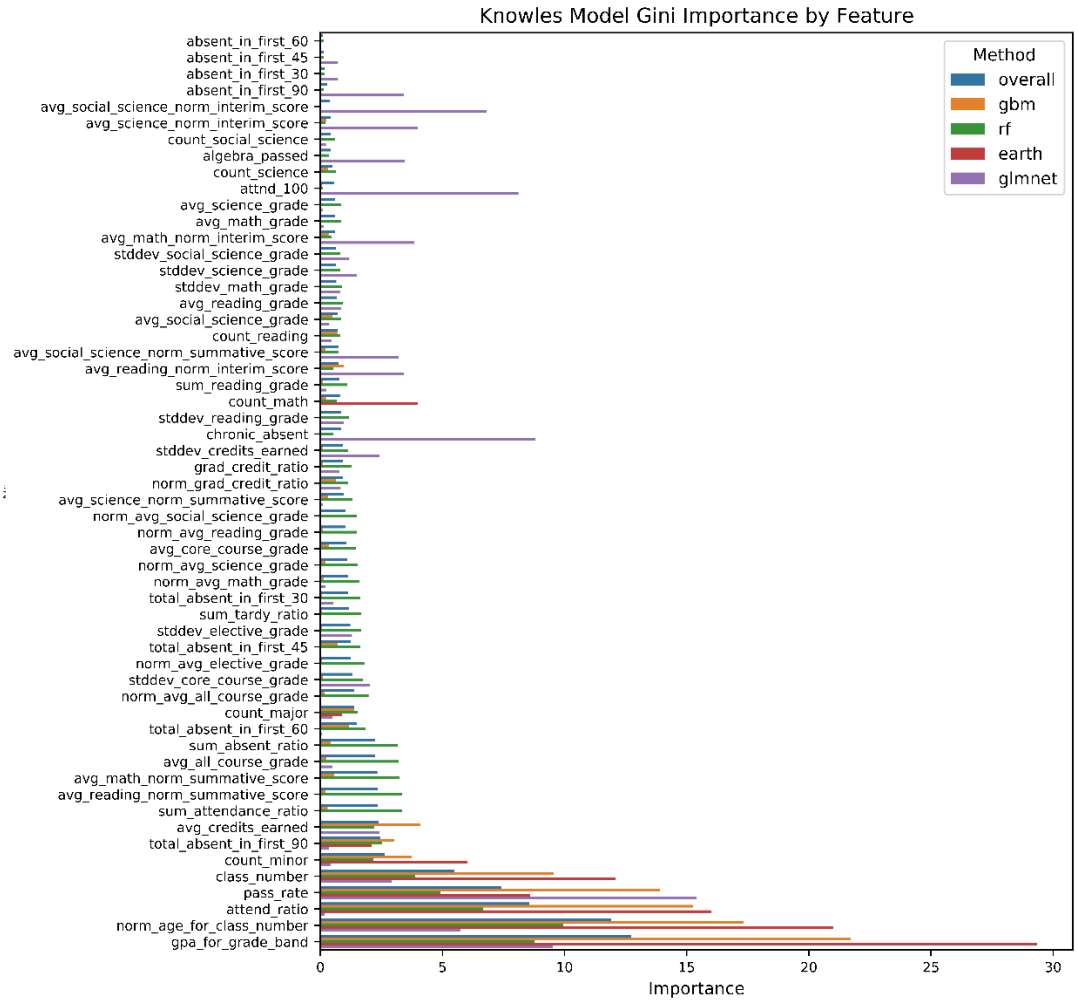



**Figure 29: Gini Feature Importance Values of Knowles models**

The Knowles models relied heavily on features related to attendance, academic achievement, student behavior, and the students age (normalized within grade) with course pass rate, attendance ratio, normalized age within grade, and GPA within grade providing the most importance for the overall ensembled model. While this trend was generally seen across all the individual models, some differences were observed. The multivariate adaptive regression spline (earth) model valued the student's count of math courses completed much higher than the other

models and the elastic-net regularized logistic regression (glmnet) model put importance on a wider set of student features related to attendance and assessments.

### 4.7 Comparing Across the Generated Models (Grades 1st - 12th)

The AUC performance for the Aggregated Data, Mean, and DSEE models are calculated across grades 1st through 12th for all districts. A DeLong test was then used to compare the model performance of each EWS pair. The first pair of tests was conducted to compare the AUC performance of the Aggregate Data model and the Mean model. There was a significant difference in AUC performance between the Aggregate Model (AUC=0.7583) and the Mean model (AUC=0.7955) EWS's; $D$= -29.759, p <0.001. These results suggest that the Mean model outperforms the Aggregate Data model when generating high school dropout predictions. Averaging the risk probabilities of individualized district level models appears to be better at detecting student at-risk status than combining all the data together to create one single, multiple district model.

The second DeLong test conducted was to compare the AUC performance of the Aggregate Data model and the DSEE model. There was a significant difference in AUC performance for the Aggregate Model (AUC=0.758) and the DSEE model (AUC=0.797) EWS's; $D$ = -31.191, p <0.001. These results suggest that the DSEE model outperforms the Aggregate Data model when generating high school dropout predictions. Using a weighted average based on similarity on the risk probabilities of individualized district level models appears to be better at detecting student at-risk status than combining all the data together to create one single, multiple district model.

The final DeLong test conducted was done to compare the AUC performance of the Mean model and the DSEE model. There was a significant difference in AUC performance

between the Mean model (AUC=0.795) and the DSEE model (AUC=0.797) EWS's; *D*=-77.18, p < 0.001. These results suggest that the DSEE model slightly outperforms the Mean model perform the same when generating high school dropout predictions. Using a weighted average based on similarity on the risk probabilities of individualized district level models produced higher results at detecting student at-risk status as taking a simple average of predictions generated by each Pillar Model. A correlation analysis of AUC performance and district graduate rates did not show significant results, suggesting the districts' dropout rate does not impact the accuracy of the Aggregate Data Model, Mean Model, and DSEE model for 1st through 12th grade predictions.

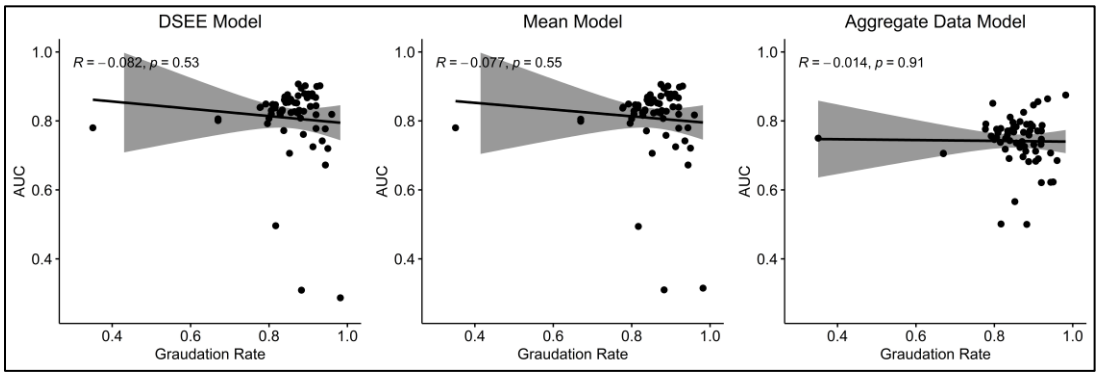



**Figure 30: Pearson correlation of EWS AUC performance on 1st through 12th grade predictions and reported district graduation rates.**

4.8 Comparing Across the Generated Models (6th Grade Students)

The reported AUC performance for the DSEE, Balfanz and Knowles models are calculated across 6th grade student records for all districts. A DeLong test was then used to compare the model performance of each EWS pair. The first test was conducted to compare the AUC performance of the Balfanz threshold based EWS and the DSEE model. There was a significant difference in AUC performance between the Balfanz (AUC=0.639) and the DSEE model (AUC=0.710) EWS's; *D*= -15.211, p <0.001. These results suggest that the DSEE model

outperforms the Balfanz model when generating high school dropout predictions using 6th grade student records. Averaging the weighted risk probabilities of individualized district level models appears to be better at detecting student at-risk status than utilizing a simplified threshold-based method.

The last DeLong test conducted was to compare the AUC performance of the Knowles model and the DSEE model. There was a significant difference in AUC performance between the Knowles model (AUC=0.801) and the DSEE model (AUC=0.710) EWS's; *D*= 20.506, p <0.001. These results suggest that the Knowles model outperforms the DSEE model when generating high school dropout predictions using 6th grade student records. Building multiple EWS models using combined data and ensembling them together into one single detector appears to be better at detecting student at-risk status than averaging the weighted risk probabilities of individualized district level models. A correlation analysis of AUC performance and district graduate rates did not show significant results, suggesting the districts' dropout rate does not impact the accuracy of the DSEE, Balfanz or Knowles models for 6th grade predictions.

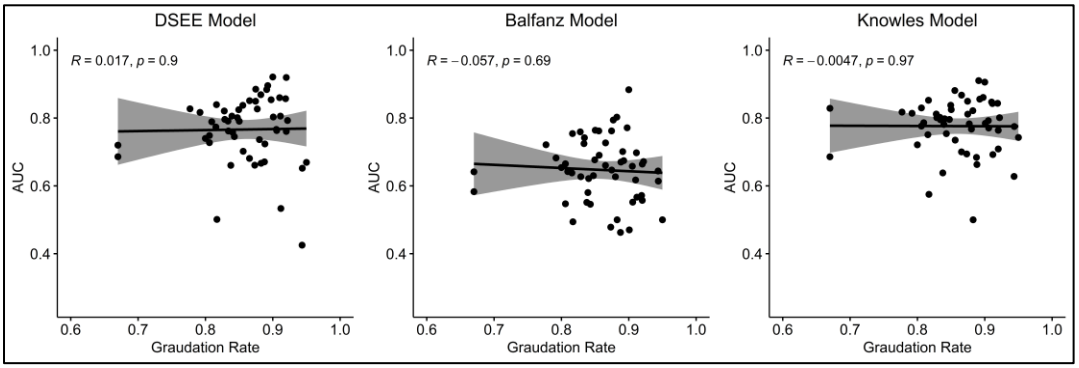



**Figure 31: Pearson correlation of EWS AUC performance on 6th grade predictions and reported district graduation rates.**

4.9 Comparing Across the Generated Models (9th Grade Students)

The reported AUC performance for the DSEE, Knowles and Chicago models are calculated across 9th grade student records for all districts. The first DeLong test conducted was

used to compare the AUC performance of the Chicago threshold based EWS and the DSEE model on high school student predictions. There was a significant difference in AUC performance between the Chicago (AUC=0.693) and the DSEE model (AUC=0.821) EWS's; *D*= -55.809, p <0.001. These results suggest that the DSEE model outperforms the Chicago model when generating high school dropout predictions. Averaging the weighted risk probabilities of individualized district level models appears to be better at detecting student at-risk status than utilizing a simplified threshold-based method.

The final DeLong test conducted was used to compare the AUC performance of the Knowles model and the DSEE model on 9$^{th}$ grade student predictions. There was a significant difference in AUC performance between the Knowles model (AUC=0.884) and the DSEE model (AUC=0.821) EWS's; *D*= 28.259, p <0.001. These results suggest that the Knowles model outperforms the DSEE model when generating high school dropout predictions. Using a weighted average based on similarity on the risk probabilities of individualized district level models does not appear to be better at detecting student at-risk status than combining multiple models built on the same data together into a single ensemble. A correlation analysis of AUC performance and district graduate rates did not show significant results, suggesting the districts' dropout rate does not impact the accuracy of the DSEE, Chicago or Knowles models for 9$^{th}$ grade predictions.

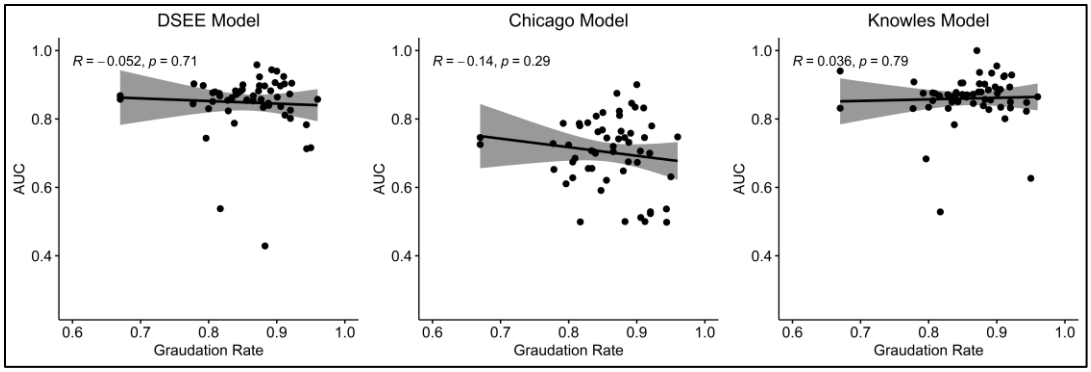



**Figure 32: Pearson correlation of EWS AUC performance on 9th grade predictions and reported district graduation rates.**

4.10 Prediction Equity Results

Regarding equity, 1st grade through 12th-grade risk predictions for the Aggregate Data model, Mean model, and DSEE model produced mixed results. Each model performed better within some demographic groups and lower in others. The Aggregate Data model performed lower than the Mean and DSEE models in all groups except the Hispanic student populations, where it achieved an AUC of 0.71 compared to the 0.693 AUC of the Mean model and the 0.695 of the DSEE model (SD=0.007). The Mean model performed slightly better than the DSEE model for predicting risk among multi-racial students (0.001 better), with the DSEE marginally performing better in every other category. AUC performance across all ethnicity groups within EWSs produced expected results, with the Mean model (M=0.744, SD=0.044) and DSEE model (M=0.745, SD=0.044) performing similarly, beating the Aggregate Data model (M=0.692, SD=0.058) which reported both a lower average AUC and higher AUC variance within groups.

When looking at all EWS's average performance within groups, White (M=0.803, SD=0.032) and Black (M=0.776, SD=0.015) students received more accurate predictions than all other populations. Overall, the models generally performed worse for students who were Hispanic (M=0.699, SD=0.007), Indigenous (M=0.713, SD=0.008), and Undefined (M= 0.690,

SD=0.060). The models achieved the lowest average AUC for Pacific Islander students, at 0.665 (SD=0.014). Curiously, Asian students performed much better in the Mean model (AUC=0.755) and DSEE (AUC=0.756) compared to the Aggregate Data model (AUC=0.629) (SD=0.059).

White students achieving the best results within all models is a notable finding. White students are typically not considered underserved populations that generally experience lower levels of dropout than other demographic groups that could benefit more from focused interventions driven by the accuracy of a high school dropout EWS (McFarland, Cui, Rathbun & Holmes, 2018). However, Black students achieving the second-highest AUC was an interesting and somewhat unexpected result, as Black students are seen as an underserved group susceptible to algorithm-driven predictive bias (Selena & Kenney, 2019). Looking at the distribution of student demographics (see Appendix A) used for these models' training, it appears that White and Black students were the average largest groups represented in the study data, with Pacific Islanders one of the smallest. This difference in population representation could be why this pattern emerges within all the EWSs AUC performance. Model performance based on Gender across all grades was relatively even. The difference between male and female students was within 0.1 percent for all models. The results of this analysis are found below.

**Table 8: AUC Results Calculated Within Demographic Groups (1st – 12th Grade)**

| Demographic | | Mean model | DSEE model | Aggregated Data Model | $\bar{X}$ | σ |
|---|---|---|---|---|---|---|
| Ethnicity | Asian | 0.755 | 0.756 | 0.629 | 0.713 | 0.059 |
| | Black | 0.787 | 0.787 | 0.755 | 0.776 | 0.015 |
| | Hispanic | 0.693 | 0.695 | 0.710 | 0.699 | 0.007 |
| | Indigenous | 0.717 | 0.719 | 0.701 | 0.713 | 0.008 |
| | Multi | 0.767 | 0.766 | 0.731 | 0.755 | 0.017 |
| | Pacific Islander | 0.707 | 0.708 | 0.581 | 0.665 | 0.060 |

| | Demographic | Mean model | DSEE model | Aggregated Data Model | $\bar{X}$ | σ |
|---|---|---|---|---|---|---|
| | Undefined | 0.700 | 0.701 | 0.670 | 0.690 | 0.014 |
| | White | 0.825 | 0.826 | 0.757 | 0.803 | 0.032 |
| | **$\bar{X}$** | **0.744** | **0.745** | **0.692** | **0.727** | **0.025** |
| | **σ** | **0.044** | **0.044** | **0.058** | **0.049** | **0.007** |
| Gender | Female | 0.799 | 0.800 | 0.748 | 0.782 | 0.025 |
| Gender | Male | 0.794 | 0.796 | 0.750 | 0.780 | 0.021 |
| Gender | **$\bar{X}$** | **0.797** | **0.798** | **0.749** | **0.781** | **0.023** |
| Gender | **σ** | **0.003** | **0.002** | **0.001** | **0.002** | **0.001** |
| | **$\bar{X}$** | **0.754** | **0.755** | **0.703** | **0.738** | **0.026** |
| | **σ** | **0.045** | **0.045** | **0.057** | **0.045** | **0.018** |

6th grade risk predictions for the Aggregate Data model, Mean model, DSEE model, Balfanz model, and Knowles model produced interesting results in terms of equity. The Knowles model more equitably performed better within all ethnic demographic groups than the other four models. On Average, the Knowles model achieved an AUC of 0.758 (SD=0.05) compared to the DSEE model's 0.626 (SD=0.063), Mean models 0.621 (0.063) Aggregate Data model's 0.569 (SD=0.142) and Balfanz model's 0.527 (SD=0.073) AUC scores across all demographic groups.

The Balfanz model produced the overall lowest AUC across all ethnic groups, with the lowest AUCs observed in Asian (0.508), Black (0.508), Hispanic (0.575), and the Pacific Islander populations (receiving the lowest score of 0.399). The Aggregate Data model was the second-lowest performing model across these groups, only performing marginally better than the Balfanz model with Asian (0.590), Pacific Islander (0.346), and Indigenous (0.350) ethnicity groups obtaining the lowest AUC scores. Interestingly, the Aggregate Data model was better at generating equitable dropout risk for Black and Multi-ethnic students than the Balfanz model, which produced lower AUCs for these populations. This result could be due to the Balfanz models threshold-based approach and reliance on a few critical indicators implemented on these

populations, which may not have been present or available to generate accurate dropout risk predictions.

When looking at the combined average performance of EWSs' within the groups, the EWSs performed similarly to the previous equity analysis. White (M=0.732, SD=0.052) and Black (M=0.64, SD=0.094) students received more accurate predictions than the other populations, with the models performing worse for students with Undefined (M=0.597, SD=0.043), Indigenous (M=0.579, SD=0.14), and Pacific Islander (M=0.553, SD=0.149) backgrounds. We observe that Hispanic students performed much better in the more advanced EWS implementations, with the Mean model (AUC=0.605), DSEE (AUC=0.611), and Knowles (AUC=0.778) model performing considerably better than the Aggregate Data model (AUC=0.596) and Balfanz model (M=0.508) (SD=0.074).

While the Knowles model shows the same pattern as the other EWSs in terms of performance within student ethnicity (White students achieving the best results and other groups receiving lower results), the overall performance is considerably higher, with the Knowles model producing an AUC above 0.70 for most of the other groups, much better than the other models. Despite these results, the Knowles model still struggled to identify dropout risk within the Pacific Islander (AUC=0.639) student population, which received an AUC of 0.035 points lower than the next lowest scoring group, students with an Undefined ethnicity (AUC=0.673). This finding is of interest, as the Pacific Islander population performs the weakest across models that use combined district data, and stronger for models that are built within district (DSEE, Mean) to identify student at-risk status. As mentioned previously, these findings could result from the low representation of Pacific Islander students within the data utilized for this analysis.

Model performance based on the reported student gender across 6th through 12th-grade students produced minimal variance across all 5 EWSs. The Knowles model had the highest AUC across both males, and female students (AUC=0.795), followed by the DSEE (AUC=0.709) and Mean (AUC=0.706) models. The Aggregated Data Model and Balfanz model produced the lowest average AUCs across groups, with the Aggregate Data model receiving 0.681 and the Balfanz model obtaining an average AUC of 0.629. The difference in model performance between male and female students was within 0.5 percent for all EWS, with the Mean, DSEE, and Knowles model slightly performing better for male students and the Aggregate Data model and Balfanz model performing marginally better for female students. The results of this gender-based equity analysis suggest the student's gender identity does not broadly impact the EWSs performance at detecting high school dropout risk, regardless of the method implemented for creating the detector. The results of this analysis are found below.

**Table 9: AUC Results Calculated Within Demographic Groups (6th Grade Predictions)**

| | Demographic | Mean model | DSEE model | Aggregated Data Model | Balfanz model | Knowles model | $\bar{X}$ | σ |
|---|---|---|---|---|---|---|---|---|
| Ethnicity | Asian | 0.565 | 0.565 | 0.590 | 0.508 | 0.797 | **0.605** | **0.100** |
| | Black | 0.604 | 0.609 | 0.692 | 0.508 | 0.787 | **0.640** | **0.094** |
| | Hispanic | 0.605 | 0.611 | 0.596 | 0.575 | 0.778 | **0.633** | **0.074** |
| | Indigenous | 0.606 | 0.621 | 0.350 | 0.534 | 0.782 | **0.579** | **0.140** |
| | Multi | 0.554 | 0.559 | 0.741 | 0.457 | 0.806 | **0.623** | **0.129** |
| | Pacific Islander | 0.688 | 0.692 | 0.346 | 0.399 | 0.639 | **0.553** | **0.149** |
| | Undefined | 0.593 | 0.593 | 0.540 | 0.585 | 0.673 | **0.597** | **0.043** |
| | White | 0.756 | 0.757 | 0.701 | 0.647 | 0.799 | **0.732** | **0.052** |
| | **$\bar{X}$** | **0.621** | **0.626** | **0.569** | **0.527** | **0.758** | **0.620** | **0.098** |
| | **σ** | **0.063** | **0.063** | **0.142** | **0.073** | **0.060** | **0.050** | **0.037** |
| Gender | Female | 0.700 | 0.703 | 0.670 | 0.605 | 0.799 | **0.695** | **0.063** |
| | Male | 0.713 | 0.716 | 0.693 | 0.654 | 0.791 | **0.713** | **0.045** |
| | **$\bar{X}$** | **0.706** | **0.709** | **0.681** | **0.629** | **0.795** | **0.704** | **0.054** |
| | **σ** | **0.006** | **0.007** | **0.011** | **0.025** | **0.004** | **0.009** | **0.009** |
| | **$\bar{X}$** | **0.638** | **0.643** | **0.592** | **0.547** | **0.765** | **0.637** | **0.089** |
| | **σ** | **0.066** | **0.065** | **0.135** | **0.078** | **0.056** | **0.056** | **0.038** |

The last prediction equity comparison is for 9th grade student risk predictions created by the Aggregate Data, Mean, DSEE, Knowles, and Chicago models. This analysis produced similar results to the previous study, with the machine learning driven EWSs outperforming the thresholds based EWS. Again, the Knowles (M=0.831, SD=0.087) model had the best-performing detector, averaged across groups, followed by the DSEE (M=0.767, SD=0.057), Mean (M=0.767, SD=0.057), and Aggregate Data (M=0.755, SD=0.077) models. The threshold-based Chicago model EWS performed the worst (M=0.626, SD=0.047), obtaining a significantly lower AUC scores across groups.

Overall performance of the EWSs within-group continues the previously seen trend of performance, with White (M=0.817, SD=0.059) and Black (M=0.801, SD=0.081) students achieving the highest scores, followed by Asian (M=0.766, SD=0.094), Indigenous (M=0.768, SD=0.086), and Multi-ethnic (M=0.784, SD=0.086) students. Undefined (M=0.681, SD=0.063) and Pacific Islander (M=0.650, SD=0.051) students continued to receive the lowest average AUC performance among all EWSs used in the high school risk detection population.

The highest performing Knowles model shows the same performance pattern within student ethnicity as the other 4 EWSs, with White and Black students achieving the best results and other groups receiving lower results. Despite this consistent pattern, the Knowles model's overall performance was considerably higher across groups for almost all ethnicities than the other models, with the Pacific Islander (AUC=0.650) and Undefined (AUC=0.681) students the only groups receiving an AUC below 0.7. As mentioned previously, these findings are of interest as the Pacific Islander population performs the weakest across all models, regardless of the

EWSs method of implementation, potentially resulting from the low representation of Pacific Islander students within the data utilized for this analysis.

Model performance based on Gender for 9th grade students produced a low average variance (SD=0.063) across all 5 EWSs. The Knowles model had the highest AUC across both male, and female students (AUC=0.884), followed by the DSEE (AUC=0.822) and Mean (AUC=0.821) models. The Aggregated Data (0.784), and Chicago (0.692) produced the lowest AUCs, with the Chicago showing the worst performance among all 5 EWSs tested. The difference in model performance between male and female students was marginal (within 0.2) percent for all EWS. The Mean, DSEE, Aggregate Data and Knowles model results show a slightly higher performance for female students. In contrast, the Chicago model perform marginally better for male students. These findings suggest that the student's gender identity does not broadly impact the EWSs performance at detecting high school dropout risk, regardless of the method implemented for creating the detector. The results of this analysis are found below.

**Table 10: AUC Results Calculated Within Demographic Groups (9th Grade Students)**

| | Demographic | Mean model | DSEE model | Aggregated Data Model | Knowles model | Chicago model | $\bar{X}$ | σ |
|---|---|---|---|---|---|---|---|---|
| Ethnicity | Asian | 0.794 | 0.794 | 0.744 | 0.894 | 0.606 | **0.766** | **0.094** |
| | Black | 0.831 | 0.832 | 0.800 | 0.891 | 0.650 | **0.801** | **0.081** |
| | Hispanic | 0.708 | 0.709 | 0.745 | 0.871 | 0.605 | **0.727** | **0.086** |
| | Indigenous | 0.724 | 0.725 | 0.885 | 0.851 | 0.656 | **0.768** | **0.086** |
| | Multi | 0.817 | 0.816 | 0.788 | 0.875 | 0.621 | **0.784** | **0.086** |
| | Pacific Islander | 0.709 | 0.709 | 0.604 | 0.641 | 0.589 | **0.650** | **0.051** |
| | Undefined | 0.708 | 0.707 | 0.694 | 0.736 | 0.558 | **0.681** | **0.063** |
| | White | 0.847 | 0.847 | 0.782 | 0.887 | 0.721 | **0.817** | **0.059** |
| | **$\bar{X}$** | **0.767** | **0.767** | **0.755** | **0.831** | **0.626** | **0.749** | **0.076** |
| | **σ** | **0.057** | **0.057** | **0.077** | **0.087** | **0.047** | **0.065** | **0.066** |
| Ge | Female | 0.829 | 0.830 | 0.792 | 0.892 | 0.686 | **0.806** | **0.068** |

| Demographic | Mean model | DSEE model | Aggregated Data Model | Knowles model | Chicago model | $\bar{X}$ | σ |
|---|---|---|---|---|---|---|---|
| Male | 0.812 | 0.814 | 0.777 | 0.876 | 0.699 | **0.796** | **0.058** |
| **$\bar{X}$** | **0.821** | **0.822** | **0.784** | **0.884** | **0.692** | **0.801** | **0.063** |
| **σ** | **0.008** | **0.008** | **0.008** | **0.008** | **0.006** | **0.005** | **0.005** |
| **$\bar{X}$** | **0.778** | **0.778** | **0.761** | **0.842** | **0.639** | **0.760** | **0.073** |
| **σ** | **0.055** | **0.055** | **0.070** | **0.080** | **0.050** | **0.053** | **0.014** |

4.11 Summary of Findings

In this dissertation, I have developed a novel approach to modeling student risk of not graduating from high school for districts where the quality, quantity, or availability of data is insufficient to produce a comprehensive student risk model. The District Similarity Ensemble Extrapolation (DSEE) approach attempts to customize a model for a specific "Target" school district based on models from other school districts where more complete data are available, taking into account the degree of similarity each school district has to the Target district. This new method achieves good predictive power for students in districts that were not used to develop the model without fitting or modifying the models or their application. Furthermore, it achieves statistically significant better results than popular alternate threshold-based approaches to predicting at-risk status in new districts (the Chicago model and the Balfanz model) and statistically significant better performance than a simply created base model using an aggregate of all records (Aggregate Data model), and slightly better than simply averaging model predictions across districts with equal weight given to each Pillar district.

However, the DSEE fails to outperform the replicated Knowles model method in both student dropout risk predictive accuracy and equitable model performance within student demographics (ethnicity and gender). Additionally, given the DSEE's performance, we can conclude that it does not beat the Bowers Growth-Mixture Model's reported results, which

reported a higher AUC than the Knowles model. These findings suggest that generalizing machine-learned district level models to new district populations for dropout risk detection is more effective than traditional threshold-based early warning systems, but ultimately fails to outperform models that implement more advanced methods of machine learning techniques.

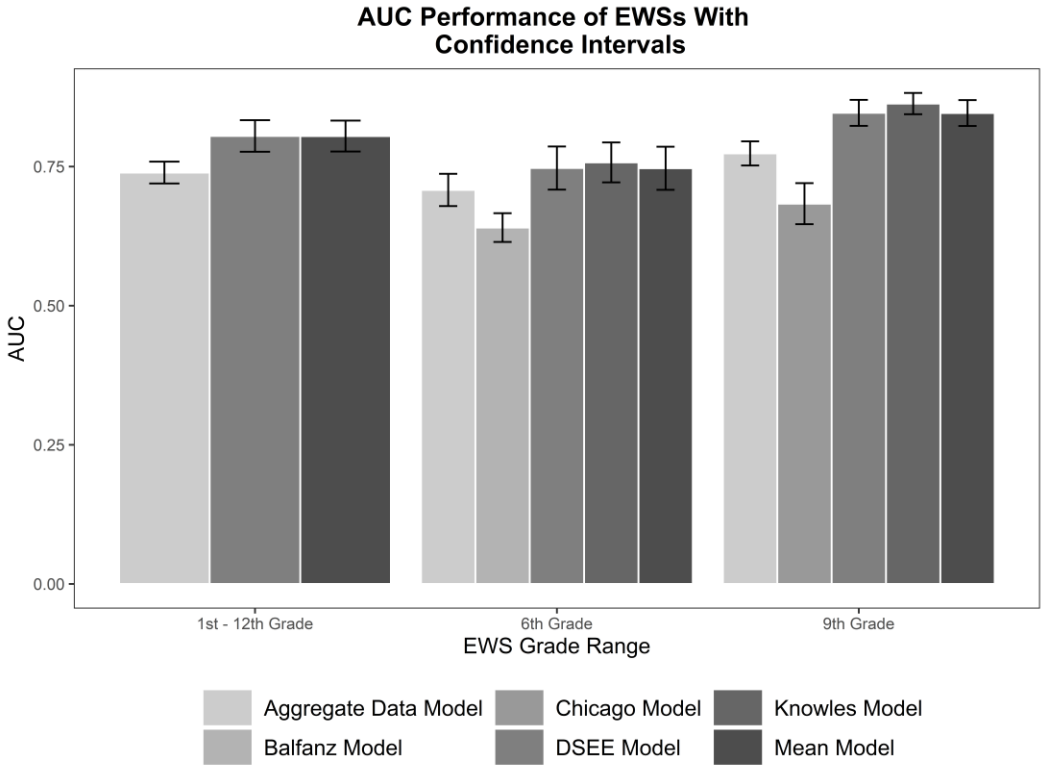



**Figure 33: Average AUC performance of EWSs with 95% confidence intervals**

Evaluating these EWS's within demographic groups to determine prediction equity suggests that a machine-learned based system produce not only better overall AUC performance but also higher levels of equity when making risk predictions within specific student ethnicity populations. While the Aggregate Data, Mean, DSEE, and Knowles models far exceeded the threshold-based Balfanz and Chicago models' performance, the Knowles model stands out as it outperformed the other machine-learned EWS's in overall predictive accuracy and reported lower levels of AUC variance within demographic groups when evaluating the EWSs prediction

equity. Moreover, the equity analysis results suggest that the EWSs tested in this research generally performed better among students that represented a larger proportion of the students contained in the research data such as White and Black students, and reported lower AUC scores for other populations with lower representation in the data such as the Pacific Islander population, which consistently received the lowest AUC scores, regardless of the grade level or EWS implementation method.

Lastly, EWS performance within gender suggests little to know modeling bias when creating risk predictions, irrespective of whether the EWS is threshold-based or machine learning-driven. Given that the gender distribution was mostly equal for both males and females in the data, this provides further evidence that the level of diverse student representation in the data is a potential driver of how the EWS performs within these populations, even after excluding demographic data when creating the EWS modes. While these findings show the Knowles model as a clear winner in terms of AUC performance, implementing this method presents many challenges to educators due to the computational power required to train, validate, and deploy this EWS.

# Chapter 5: Conclusions & Discussion

This research highlights the differences in Early Warning system performance, depending on the method of implementation. Threshold-based systems are easy to implement, often consisting of a single, simple conditional argument on a few key indicators used to dropout generate risk. While these types of EWS's may excel in their simplistic design, they fall short in their performance compared to the far more complicated and costly methods of machine learning-driven EWSs. With minimal financial and expert resources available, educators interested in deploying an EWS in their school district face the challenge of balancing complexity, interpretability, and model accuracy. Moreover, while EWS performance can be measured in binary outcomes (graduate or dropout), deploying the model into the real world is not as simple. Educators need the capacity to identify which students are at-risk, but they also require additional capabilities with understanding *why* that student is at risk to provide the best interventions and affordances (Bowers, 2021).

While machined-learning-driven EWSs are more potent at identifying dropout risk, they are often complicated in their *black-box* design, making them difficult to dissect and interpret at the student prediction level. Additionally, once a decision is made to implement an EWS, there also begins the task of mitigating any predictive bias that may propagate in the risk predictions to ensure fair and equitable distribution of resources among high-risk student populations and identifying and providing the best intervention specific to the individual needs of the student.

In the following sections, I will expand on these issues and discuss the importance of machine-learned EWS model explainability, key for delivering focused student interventions that reduce dropout risk. I will also discuss the challenges and trade-offs between EWS accuracy and interpretability and suggest potential strategies that can be implemented to minimize having to

sacrifice one for the other. Additionally, I will provide strategies for how machine-learned EWS's can be interpreted at the student prediction level, opening the black box, and allowing for more focused student interventions. Lastly, I will discuss future opportunities that build upon this research that can potentially lead to improved results for the DSEE early warning system.

## 5.1 Common Data Standards & Open Access Algorithms

As mentioned earlier in this paper, the data utilized within this study was gathered using an educational data management tool purchased by educators across the U.S. This specific tool provided educators with three primary functions: (a) to aggregate data from historically siloed systems (grade books, attendance records, assessment scores, etc.), (b) to flatten this aggregated data by mapping to a unified schema, and (c) to provide actionable data-driven insights to educators through the use of a dashboard. This tool's use provided the foundational capability to build and test the methods replicated in this research. The vendor had completed the bulk of the work standardizing the data. While I was fortunate enough to leverage such a system, many school districts interested in applying data-driven dropout detection may not have the resources to invest in such a scenario, presenting significant implementation challenges.

These challenges stem from U.S. school districts' heterogeneous nature with data collection standards set by the local or state education departments. This heterogeneity results in some school districts quantifying student data in different ways (ex: differences in GPA scale, absentee counts, formative assessments, etc.), limiting their ability to implement a generalizable EWS built in another district without significant effort to fit the data to the method. In my case, a team of four dedicated data analysts worked with each school district for several weeks to map their data to the standard schema of the BrightBytes system, representing thousands of hours of labor resources. This effort highlights educators' need to adopt a common data system across

districts that ease their capacity to test and implement Early Warning systems. While there has been a movement from the Federal Department of Education to address these challenges from the Common Education Data Standards (CEDS), these standards' adoption is not seen across the nation (Common Education Data Standards, 2019).

Districts adhering to data standards benefit from reduced preparation for EWS deployment and are enabled to take advantage of potential existing EWS algorithmic code with little to no modification. Recent work by Bowers calls for EWS researchers to make their algorithms and code open-access, available to the public (Bowers, 2021). In addition to call for open-access, Bowers proposes a “Four A” framework in the design of an EWS to ensure they are Accurate, Accessible, Actionable and Accountable. Using metrics such as ROC AUC to measure the performance of an EWS ensures the detector is *Accurate* in predicting a student outcome. Improving transparency in the algorithms design, so that it can be accessed, examined and understood, makes the EWS *Accessible.* Designing an EWS to not only create a risk prediction, but also provide insights that help educators tailor interventions based on the individual student’s data profile, ensures the EWS is *Actionable.* Lastly, implementing policies and procedures that frequently check for prediction bias created by the EWS algorithm in the communities they serve enables educators to critique and adjust the EWS to be more equitable and *Accountable* (Bowers, 2021).

## 5.2 Dissecting the Early Warning System

As this research shows, Machine learning-driven Early Warning Systems provide significant performance advantages over a traditional threshold-based EWS. While this performance increase is substantial, using these methods, introduce additional barriers for educators who utilize an EWS within their school district. The primary advantage of a threshold-based EWS is

that it is 1) easy to implement and 2) easy to explain why the EWS assigned the student an at-risk status. This explainability removes the guesswork from determining which potential interventions are required to reduce the likelihood of dropping out and enables educators to apply. For example, the Chicago model relies on two freshman-year data points to assign risk; the number of credits earned and the number of core course (English, math, science, or social science) failures within a semester. An educator in a district utilizing this EWS could review these data points for any predicted at-risk student and determine what type of course-work based intervention is required to improve that student's outcome. While highly interpretable to a non-technical educator, these data-driven interventions are limited to the few indicators utilized in the EWS, potentially reducing their effectiveness. An educator may see that the Chicago model indicates a student is off-track based on their number of core course failures and suggest the student participate in an after-school credit recovery program as an intervention. In actuality, the student was suffering from chronic absenteeism and was simply not present in school for those courses, impacting their grade and requiring a completely different set of risk-mitigating intervention strategies.

Machine learning EWSs, while better performing, are much more difficult to interpret given the complexity of their design (Sansone, 2019). Implementing these detectors often requires aggregating many different student data types from multiple areas (behavior, attendance, academic performance, etc.) and applying highly advanced statistical methods to produce the risk prediction. Additionally, this EWS prediction is given as a binary outcome (dropout or graduate). It offers no additional insight to the educator into how the model arrived at this estimate for an individual student and therefore reduces their capacity to apply focused interventions. Additionally, with resource scarcity in many school districts, educators are often faced with

prioritizing the type of intervention available based on their risk severity. The binary outcome produced by these systems is based on a (sometimes arbitrary) threshold (if the probability for dropout is above 0.5, then the student is predicted to dropout, otherwise graduate). It reduces an educator's ability to assign interventions on both the student's need and risk severity. This trade-off between accuracy and interpretability presents significant challenges in adopting and using EWSs within school districts (Knowles, 2015).

One potential opportunity for improving machine-learned EWS interpretability without sacrificing accuracy, leading to better-focused student interventions, is to leverage additional machine learning techniques that break down how the model is working for individual predictions. One such method is to utilize SHapley Additive exPlanations (SHAP) values to provide insights into a student-level risk prediction. In its simplest definition, SHAP values are created by looking at an individual prediction made by the model and analyzing and highlighting the student's data to generate the prediction. It then isolates the specific indicators that either contributed to the student's risk of dropout or contributed to against it (Ribeiro, Singh, & Guestrin, 2016). This information is then provided as an additional output to help make an informed decision around what interventions and actions educators should provide to reduce dropout risk for this student.


**Figure 34: Example output of SHAP value implementation for Machine-Learning EWS for a student predicted to graduate.**

The figure above provides a visual representation of SHAP values in practice for a single student prediction. In this example, the model produced a 0.626% probability that the student

will graduate from high school. Reviewing the SHAP values suggests that this prediction is driven by the student's relatively low number of absences in the first 30 days, low number of minor behavioral incidents, and low number of major behavior incidents. The contributing factors that cause the model's relatively low confidence in this prediction are driven by the students' low-grade performance in social science and missing data for their reading interim assessment scores. An educator reviewing this data could provide the student with additional social science learning affordances as an intervention to better improve their likelihood of graduating.

While implementing SHAP values into the EWS presents researchers with new opportunities to improve highly accurate machine learning model interpretability, limitations still exist. The resources required to build an advanced EWS are already significant compared to a simple threshold-based approach; adding a SHAP value layer on top of this solution further complicates these Early Warning Systems. Despite this complication, any district investing in an advanced Early Warning System should include some method that enables improved model explainability and interpretability to the end-user. The benefits of data-driven focused interventions are substantial for improving student graduation outcomes.

### 5.3 Prediction-Driven Intervention Strategies

Research has shown the most successful school dropout interventions "identify and track youth at risk for school failure, maintain a focus on students' progress toward educational standards across the school years, and are designed to address indicators of student engagement and to impact enrollment status" (Christenson & Thurlow, 2004). These findings suggest that educators not only have to apply focused data-driven interventions that provide academic support and enrichment, implement programs to improve student behavior and provide personalized

learning and individualized instruction, they also need to monitor and curate these strategies for the individual student over long periods (Freeman & Simonsen, 2015). This process can present significant challenges to school districts that suffer from resource scarcity, which often exhibit higher dropout levels (McPartland & Jordan, 2001).

The table below serves as an example of potential interventions that educators can implement to reduce students' drop-out risk. This table provides a set of students at varying levels of risk determined by an EWS and the factors that contributed to the detector's prediction at the individual student level. Potential interventions are categorized based on the school districts' resource levels required to implement these actions, with *low* representing relatively low levels of cost and time resources, *medium* representing moderate levels of cost or time, and *high* and conveying significant resource investment. These interventions represent only a few strategies that educators can leverage to improve graduation rates in their school district. Depending on resource availability, they may elect to apply more than one action for any given student.

**Table 11: List of Potential Prediction-Driven Intervention Strategies to Mitigate the Likelihood of High School Dropout**

| Category | Likelihood of Dropout | Triggering Event | Resource Level | Potential Intervention(s) | Intervention Description | Citation |
|---|---|---|---|---|---|---|
| General Coursework | $\hat{p} \leq 0.4$ | The student receives lower than average semester grade performance in the English language arts core course subject area but is still considered passing. | Low | Remedial Course Enrollment | The student is enrolled in a remedial ELA course | (Goldschmidt & Wang, 1999) |
| | | | | Increased responsibility | Leadership roles in the classroom (i.e., being a tutor, reteaching a lesson) | (Shernoff, Csikszentmihalyi, Schneider & Shernoff, 2014) |
| | | | | Positive Reinforcement (Rewards) | System of rewards (student points or color on color chart yields rewards), positive phone calls home, | (Nowicki, Duke, Sisney, Stricker, & Tyler 2004) |

| Category | Likelihood of Dropout | Triggering Event | Resource Level | Potential Intervention(s) | Intervention Description | Citation |
|---|---|---|---|---|---|---|
| | | | | | Honor Roll program | |
| | | | | Personalized Evaluation | Excused Deadlines, extended time on homework, change of grading scale. | (Clarke, 2013) |
| | | | | Personalized Curriculum | Chunking larger assignments into smaller deliverables and deadlines, alternative assignments (easier text in language arts, shorter essay requirements in other classes), customized lesson plans & curricula. | (Clarke, 2013) |
| | | | | Change of Instructional Delivery | The student is paired with bi-lingual native language speaking peers during class exercises | (Christenson & Thurlow, 2004) |
| | $0.6 \leq \hat{p} \leq 0.8$ | Student receives a failing grade in elective courses and lower than average grade performance in core course subject areas. | Medium | The student is enrolled in a tutoring program | Subject-specific study hall (pairing with a teacher in the same subject matter that student struggles the most), required presence at after-school tutoring program, required presence at lunch-time study sessions | (Somers, Owens & Piliawsky, 2009) |
| | | | | Personalized Curriculum | Alternative online educational resources are provided (licensed instruction, tools, and technology) | (Clarke, 2013) |
| | | | | Course substitution | Remedial course in place of study hall, a remedial course instead of an elective class | (García, Fernández & Weiss, 2013) |
| | $\hat{p} \geq 0.8$ | The student is failing two or more core | High | Credit recovery program | The student is enrolled in summer school or an online | (Rickles, Heppen, Allensworth, |

| Category | Likelihood of Dropout | Triggering Event | Resource Level | Potential Intervention(s) | Intervention Description | Citation |
|---|---|---|---|---|---|---|
| | | course subject areas; the student is not considered passing. | | | credit recovery program | Sorensen & Walters, 2018) |
| | | | | Provided In-Class Resources | Enrollment in an ICR (in-class resource) environment, which as general ed and special ed teacher present | (Betts & Shkolnik, 2000) |
| | | | | School provided technology | School-provided device/WiFi hotspot to improve assignment completion through other technology access. | (Darling-Hammond, Zielezinski & Goldman, 2014) |
| | | | | Assigned Intervention Service Professional | Educators come together throughout the year to formulate and deliver coordinated services. | (Mac Iver & Mac Iver, 2010) |
| | | | | Alternate Pathway | Enrollment in vocational programs (alternative pathways that may improve student interest) | (Tyler & Lofstrom, 2009) |
| | | | | Individualized Attention | Assignment of a paraprofessional or classroom aide | (Lane, Fletcher, Carter, Dejud & Delorenzo, 2007) |
| Student Assessments | $\hat{p} \leq 0.4$ | Students receive higher than average interim assessments and lower than average summative assessment performance in English and math course subject areas but are still considered passing. | Low | Assessment Accommodation | The student is provided additional accommodation such as extended time, has questions read aloud | (Gregg, 2009) |

| Category | Likelihood of Dropout | Triggering Event | Resource Level | Potential Intervention(s) | Intervention Description | Citation |
|---|---|---|---|---|---|---|
| | $0.6 \leq \hat{p} \leq 0.8$ | Students receive lower than average interim and summative assessment performance in English and math course subject areas but are still considered passing. | Medium | Assessment Accommodation | Provision of a scribe or language translator | (Gregg, 2009) |
| | | | | Change of Environment | Small-group testing environments to limit distractions | (Gregg, 2009) |
| | $\hat{p} \geq 0.8$ | Student receives lower than average interim assessments and failure summative assessment scores in English, math, and science course subject areas. The student is not considered passing. | High | Assigned Intervention Service Professional | Educators come together throughout the year to formulate and deliver coordinated services. | (Mac Iver & Mac Iver, 2010) |
| | | | | Individualized Attention | Assignment of a paraprofessional or classroom aide | (Lane, Fletcher, Carter, Dejud & Delorenzo, 2007) |
| Student Attendance | $\hat{p} \leq 0.4$ | The student shows more than ten absences recorded in the first 30 days of school but shows higher than average summative assessment scores in core subject areas and no core course failures. | Low | Parental Involvement | Daily parental contact when students are absent (phone call, email, text message), text messages/notifications are sent to parents in their primary language. | (Ross, 2016) |
| | | | | Positive Reinforcement (Rewards) | Perfect attendance award | (Sutphen, Ford & Flaherty, 2010) |
| | | | | Negative Reinforcement | Attendance-related punishment (losing a spot on a sports team, inability to attend prom) | (Epstein & Sheldon, 2002) |
| | $0.6 \leq \hat{p} \leq 0.8$ | The student is two standard deviations from the norm in school attendance and | Medium | Parental Involvement | Provision of a translator for parent/teacher conferences | (Ross, 2016) |

| Category | Likelihood of Dropout | Triggering Event | Resource Level | Potential Intervention(s) | Intervention Description | Citation |
|---|---|---|---|---|---|---|
| | | shows lower than average performance in core subject areas but is still considered passing. | | Transportation | Adjust the school transportation program to improve transportation access. | (Patel, Messiah, Hansen, & D'Agostino, 2020) |
| | $\hat{p} \geq 0.8$ | The student is three standard deviations from the norm in school attendance and shows three failures in core subject areas; the student is not considered passing. | High | Assigned Intervention Service Professional | Educators come together throughout the year to formulate and deliver coordinated services. | (Lane, Fletcher, Carter, Dejud & Delorenzo, 2007) |
| | | | | School provided Social Services | Before- and after-school childcare for the children of students, free and reduced breakfast and lunch access, on-site laundry services, or clothing access program | (Barnet, Arroyo, Devoe & Duggan, 2004) |
| | | | | Alternate Pathway | Enrollment in vocational programs (alternative pathways that may improve student interest) | (Tyler & Lofstrom, 2009) |
| Student Behavior | $\hat{p} \leq 0.4$ | The student shows more than the average count of minor behavior incidents and lower than average core course performance but is still considered passing. | Low | Preferential Seating | The student is provided preferential seating in the classroom to peer reduce distractions. | (Mulligan, 2001) |
| | | | | Negative Reinforcement | Behavior-related punishment (losing a spot on a sports team, inability to attend prom) | (Mayer, Sulzer & Cody, 1968) |
| | | | | Course Change | Change of instructor or course period (time) | (Sheldon & Epstein, 2002) |
| | | | | Positive Reinforcement (Rewards) | Allowing breaks to leave the classroom after x minutes of work. | (Partin, Robertson, Maggin, Oliver & Wehby, 2009) |

| Category | Likelihood of Dropout | Triggering Event | Resource Level | Potential Intervention(s) | Intervention Description | Citation |
|---|---|---|---|---|---|---|
| | $0.6 \leq \hat{p} \leq 0.8$ | The student shows a lower than average count of minor behavior incidents, higher than the average count of major behavior incidents, and lower than summative assessment performance in all subject areas but is still considered passing. | Medium | Enrollment in Behavior Program | PBSIS program (Positive Behavior Support in Schools) | (Christofferson & Callahan, 2015) |
| | $\hat{p} \geq 0.8$ | The student shows a higher than the average count of minor behavior incidents, higher than the average count of major behavior incidents, and is failing more than one core course subject; the student is not considered passing. | High | Assigned Intervention Service Professional | Educators come together throughout the year to formulate and deliver coordinated services. | (Lane, Fletcher, Carter, Dejud & Delorenzo, 2007) |
| | | | | Individualized Attention | Assignment of a paraprofessional or classroom aide | (Lane, Fletcher, Carter, Dejud & Delorenzo, 2007) |

Given that intervention strategies vary in their complexity and resource costs, educators using interpretable machine learning-driven EWS's can better equip and utilize these strategies to effect change in students based on risk. Students at lower levels of dropout risk can be provided lower resource interventions such as testing accommodations or changes in their instruction delivery, which reserves the more costly interventions like individualized, personalized learning attention for students at higher risk. While enabling educators to target better interventions based on both the binary prediction and the severity of the prediction can

improve graduation outcomes within the district, it also reinforces the need to monitor and address EWS prediction bias based on student identity.

## 5.4 Addressing Prediction Bias

As mentioned earlier in this research, reducing prediction bias is crucial when implementing predictive modeling on student populations, primarily when a model's output is used to target and apply interventions. Despite specifically excluding any demographic variables in their design, this research shows that some of the EWS model methods tested are still susceptible to bias based on student ethnicity. The EWSs show significant performance varies based on the student's identity. This fluctuation in performance can lead to unfavorable circumstances for some student populations, with the potential for the model to either over-identify or under identify student dropout risk for some student ethnicities. When educators are unaware of these risk misclassifications, they could unknowingly bias their interventions to over include or under exclude protected class students. If this occurs, the district would not observe lower levels of dropout reduction but would also be exposed to potential discriminatory civil litigation risk (Gordon, Piana, & Keleher, 2000). Fortunately, there are methods for assessing and addressing predictive model bias that can be utilized in Early Warning Systems.

The *fairlearn* open-source toolkit developed by Microsoft provides a suite of resources they can utilize to detect and mitigate machine learning model predictive bias. This toolkit, developed in the Python programming language, enables researchers to assess, visualize, and compare the disparity of performance and predictions for sub-groups by the model(s). Once any unfairness is detected, various artificial intelligence (AI) tasks and algorithms are included in the toolkit that mitigates bias and improves prediction equity (Bird, Dudík, Edgar, Horn, Lutz, Milan, & Walker, K, 2020).


**Figure 35: Example dashboard of *fairlearn* toolkit for gender-based bias analysis of Mean model performance disparity on 10,000 random sampled student predictions**

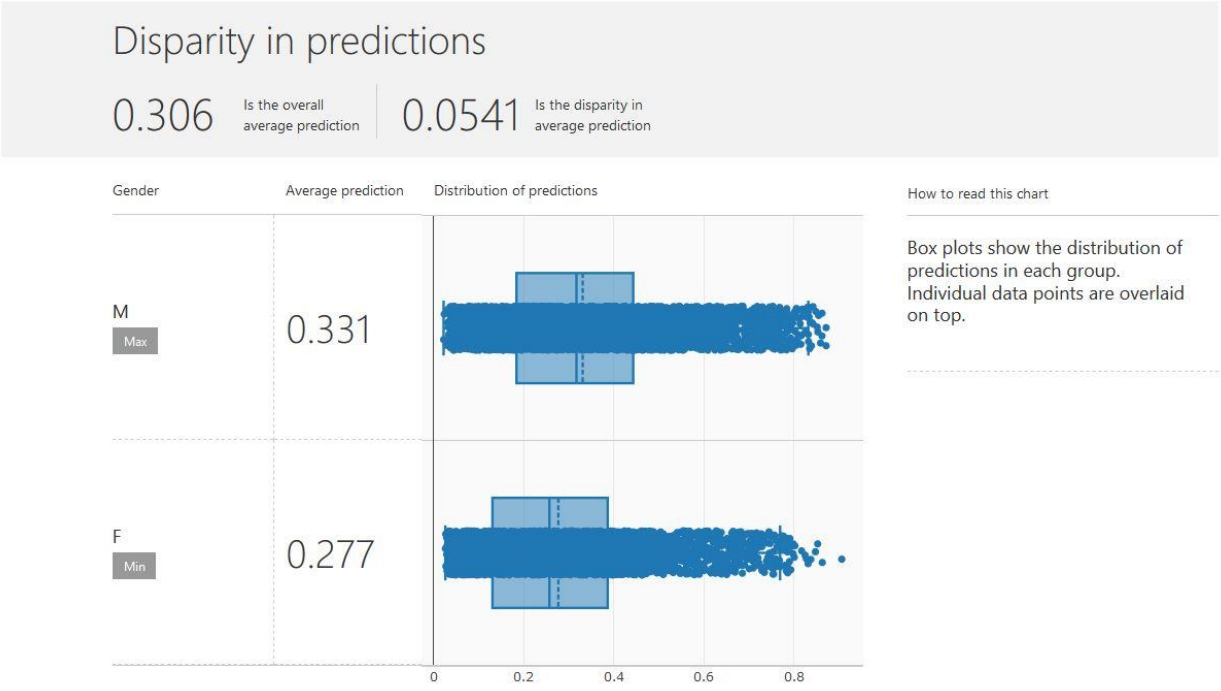



**Figure 36: Example dashboard of *fairlearn* toolkit for gender-based bias analysis of Mean model prediction disparity on 10,000 random sampled student predictions**

There are two primary functions in which these algorithms operate to improve fairness. The first is by analyzing the model's performance within a defined group characteristic and then tuning the predictions using demographic-based weights derived from the analysis. The second approach conducts a similar analysis, but rather than tuning the predictions; it attempts to identify the optimal classification probability threshold for each demographic group under investigation. Also, rather than merely adjusting the model to achieve parity across groups in AUC performance, the *fairlearn* toolkit allows researchers to select the type of bias to mitigate across several different metrics. These types include demographic parity (the selection rate of samples predicted to dropout is equal across all groups), equalized odds (true positive rate and false positive rate is similar across groups), true positive rate parity (true positive rate is equal across groups), false positive rate parity (false positive rate is equal among groups) and error rate parity (error rates are similar across groups) (Yordanova & Emanuilov, 2020). While the *fairlearn* toolkit provides EWS researchers with new capabilities to reduce bias, it is often at the trade-off of model accuracy. The algorithm attempts to *meet in the middle* across groups when achieving parity, which could lower performance from some groups while increasing others' performance.

Recent work published by Gardner, Brooks, and Baker provides an alternate method for evaluating unfairness in predictive models. Their research demonstrates that by assessing the predictive model's performance across different demographic categories in the test set (slicing) and then calculating the differential accuracy between subgroups (termed the Absolute Between-ROC Area - ABROCA), they can effectively quantify the level of unfairness present in the detector into a single value (2019). This work overcomes existing limitations of most commonly used current fairness analysis methods by 1) providing researchers the capability to evaluate

model performance across all thresholds; instead of a specific threshold set by the evaluator, 2) assess the model accuracy without strictly focusing on the positive case outcome (dropout), 3) relies solely on the predicted probabilities and predicted class, making it easy to implement, and 4) build on existing performance metrics (ROC) that are commonly used in machine learning making it easy to interpret and visualize (Gardner, Brooks & Baker, 2019).

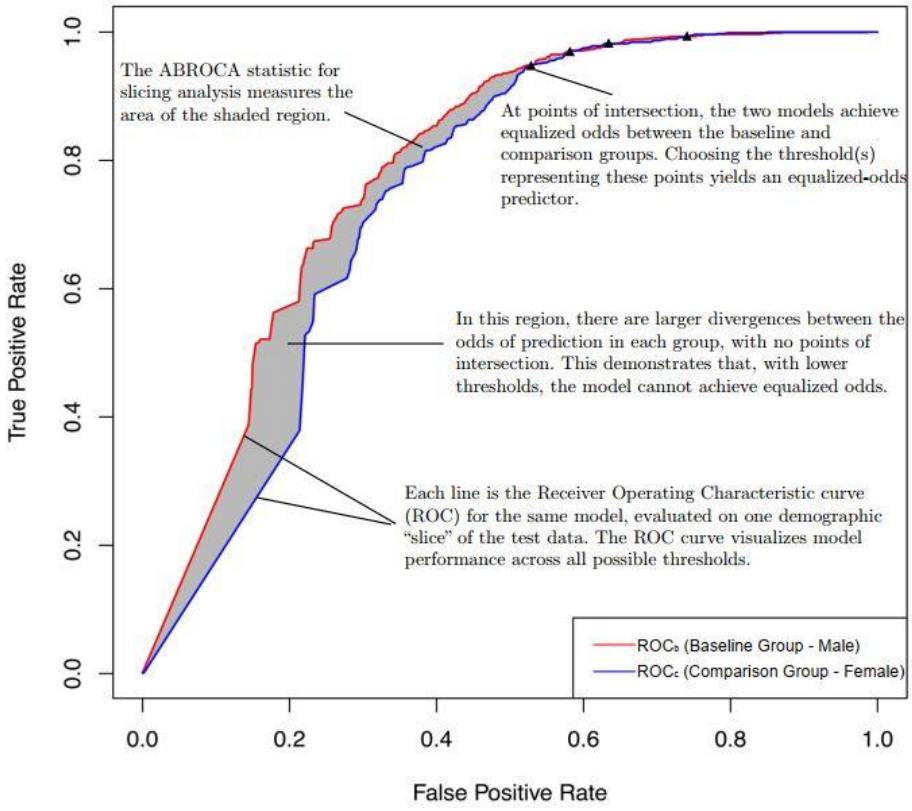



**Figure 37: Example of an annotated slice plot of ABROCA statistic** [8]

In contrast to the *fairlearn* toolkit above, the authors note that using this technique shows no evidence of a strict trade-off between fairness and model performance (Gardner, Brooks & Baker, 2019). By utilizing this method for fairness evaluation, researchers can better identify

[8] Source: Reprinted from Evaluating the fairness of predictive student models through slicing analysis. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge (pp. 225-234).

which statistical algorithms produce the best fairness when designing the EWS. They can also assess the severity of discrimination across a range of possible thresholds to better tune the model's classification prediction. Lastly, researchers can understand how the data used to train the models can be manipulated (ex: through sampling) to improve model fairness across subgroups. The last method for mitigating bias created by predictive modeling is to provide educators with an EWS *factsheet* to increase transparency into the design and methodology. This documentation should detail the model's intended purpose, performance, safety, security, and provenance information. While this strategy does not strictly change the underlying model or manipulate the produced predictions using post-hoc transformations, it does improve user (educator) knowledge on both the design and existing limitations of the EWS. Having a deep understanding of the EWS would enable educators to make more informed equitable decisions on dropout risk interventions (Arnold, Bellamy, Hind, Houde, Mehta, Mojsilović, ... & Reimer, 2019). Suppose an EWS is well documented and known to underperform for a specific sub-group. In that case, the educator could leverage additional data (ex. classroom observations, qualitative data, etc.) in addition to the quantitative output of the model to better mitigate the dropout likelihood more equitably.

## 5.5 Limitations

The difficulty involved of replicating the Bowers GMM was a limiting factor in this research, requiring me to conduct a direct comparison against the reported AUC of the GMM, published in 2010 and 2012, rather than reporting the results of his duplicated method on my research data (Bowers, 2010; Bowers & Sprott, 2012). As mentioned earlier, the decision to compare against the published results is primarily due to the GMM's structural equation modeling approach, as these types of models are traditionally built using proprietary software

and used for theory testing instead of generating on-demand risk predictions (Evermann & Tate, 2016). While identifying and measuring the differences in change among potential un-observed dropout sub-populations would have been a fascinating endeavor, this type of modeling often takes several days to converge (Ram & Grimm, 2009). As the computational resources available for this research were limited to my local computer, and the cost of licensing the required software was prohibitive, I could not apply this method to my data.

Despite my inability to replicate the Bowers GMM, my results highlight the value that this single non-cumulative GPA feature has on dropout risk detection performance across all the machine learning driven EWSs created in this research, matching prior evidence of this feature's value (Bowers, 2012a, 2012b). While this result holds true for most districts, there are a few districts where it doesn't. Generally, where non-cumulative GPA importance was low, absences is high, suggesting these districts may suffer from student attendance problems. Given that GPA and attendance interplay (i.e. if a student is not present, then their GPA goes down), attendance related features become more important for these districts than GPA.

Given these findings, there is an opportunity for future work to conduct deeper examinations of the correlations between these features as well as the cutoff values, in order to identify the optimal hand off from absence to non-cumulative GPA. This can lead to the creation of a threshold effect that is non-cumulative inside the algorithm, where educators can better focus on the indicator that is truly dominating the risk prediction and provide better intervention.

Additionally, the results of this research do not account for difference in survival versus hazard rates within grade (i.e. the risk set is conditionally dependent on time, yet it's considered time invariant in all models tested). My reported findings assume the at-risk population to be stable and unchanging through each grade, an incorrect assumption (Bowers, 2010; Singer &

Willett, 2003). According to Bowers (2010), "aggregated overall rates of dropping out do not acknowledge the time-sensitive nature of schooling and dropout processes" (p. 7). This presents several challenges for educators attempting to interpret and utilize EWSs within their school district. Given this existing limitation, future research should attempt to account for differences in risk populations within grade to better improve the internal validity of this research.

This can be accomplished by taking a similar approach to a discrete-time hazard model and restructuring the data and method used for analysis to evaluate the risk of dropout within each grade, rather than aggregating all the years together (Singer & Willett, 1993, 2003; Willett & Singer, 1991). Implementing this method would require removing students at each grade level that have either 1) dropped out before that grade or 2) transferred or left the school district for another (valid) reason. Removing these students from the data would then make the dropout risk conditionally relative to the grade population in which the student represents.

Enabling school stakeholders with the ability to review dropout risk conditional on time within each grade level would be impactful in two ways. The first is that they would be able to measure the number of students currently enrolled versus the number of students who began in that cohort in an earlier grade (i.e., grade one or perhaps even the beginning grade in that school building such as grade nine for a high school) to better understand how many students they have lost over time. This will show how significant the threat of dropout is to their student population to better focus and apply resources that improve student outcomes. The second way this would be impactful for educators is that having the ability to view students currently enrolled during X month of a school year versus the number of students who started that grade in the beginning of the year would show schools their risk relative to their actual population (i.e. who is attending, where should interventions be focused, etc.). If the data is not narrowed year by year, then the

risk set could be artificially inflated and would not present findings that are actionable for the school district quickly. In addition, it would result in a model showing incorrect lower dropout rates, suggesting a higher graduation rate then what is occurring in that school.

The binary categorization of graduates and dropouts can be considered a limitation encountered in this research. Mentioned earlier, research by Bowers and Sprott (2012a) found evidence to suggest there are several types of potential student dropouts, all with various trajectories. The authors assert that rather than one binary category of either graduation or dropout; there are several latent levels of dropout trajectory. The original four trajectories identified in 2012 were Mid-Decreasing, Low-Increasing, Mid-Achieving, and High-Achieving and account for 91.8% of dropouts. A follow-up study conducted by Bowers and Sprott (2012) identified the remaining 9% of students as either "Involved" or *lost at the last minute*. The results of this research suggest that indicators used to predict dropout (in their case, non-cumulative GPA) impacted the dropout trajectory differently for each typology, leading the authors to conclude that understanding the different types of dropout typologies could better enable schools to provide better, more personalized interventions for students (Bowers & Sprott, 2012a; Bowers & Sprott, 2012b, Bowers & Zhou, 2019).

My research does not account for these different typologies and instead limits the potential student outcome to a binary problem, the student either graduates, or the student drops out. While binary classification is a common method application of machine learning (Kumari & Srivastava, 2017), it presents some challenges to EWS researchers attempting to better intervene on students at-risk of dropping out (Bowers & Sprott, 2012a; Ananga, 2011). One future strategy that can be used to overcome this limitation would be to implement a two-step approach to EWS design to conduct an analysis of the data using a Growth-Mixture Model or Latent Class

Analysis approach (Bowers & Sprott, 2012a; Bowers & Sprott, 2012b, Bowers & Zhou, 2019) to create and classify the historical data into multiple dropout typologies. Once the data has been re-labeled with the various typologies, a multi-classification machine-learning model could be trained to predict not only the student risk of dropout, but the students risk of dropout-type, leading to better insights into potential interventions (Janosz, Le Blanc, Boulerice, & Tremblay, 2000).

Lastly, the data for this research was sourced in partnership with a private entity that school districts pay to use, the sample cannot be considered truly random nor nationally representative. The BrightBights Clarity platform is offered to schools that have the capacity to purchase the licensing, potentially biasing the sample data. Given the way resources and funding are provide in the U.S. education system, the student data represented in this research stems from two types of school districts; 1) high-performing schools in more affluent parts of the country, and 2) underperforming schools that rely on Title I funding. The funding available to purchase an educational technology tool (Title I), and the factors motivating the decision to choose the BrightBytes ed-tech solution used to collect the data within this analysis introduces potential bias impacting the external validity of my research. The data does not fully represent every type of school district interested in utilizing an early warning system to improve student graduation rates. Future work on EWSs should consider using truly nationally representative data, similar to the GMM research completed by Bowers and Sprott (2012a).

## 5.6 Future Work

There are several ways in which the models presented here could be improved. Currently, I only look at the following characteristics: student/school demographics, school size, district-level census data, and graduation rate. Research has shown that contextual factors can help identify

students at risk of dropping out and that the factors associated with dropout can differ between populations (Balfanz & Legters, 2004; Christle, Jolivette, Nelson, 2007). Some additional features to include in future work could be 1) measures of the distance between the school and the nearest city, 2) the percentage of students that continue to postsecondary enrollment, 3) the percentage of students proficient on state exams, 4) the parent or student satisfaction with the school, 5) the proportion of military-connected or otherwise highly mobile students (Baker, Berning, & Gowda, 2020), 6) rates of teen pregnancy within the school district, 7) participation in after-school activities, and 8) crime rate by city or zip code.

Exploring alternative forms of distance calculation could also improve the performance of the DSEE relative to the Mean model. In the current approach, I measure district-to-district similarity with the use of a Euclidean distance measure. Future iterations of the DSEE method could take an empirical approach to select the measure of similarity based on model performance (McCune, Grace, & Urban, 2002), rather than being limited to the simplistic distance calculation method used in this research, where all demographic features are weighted equally. Given the type and quality of the data used in the similarity calculation, there is evidence to suggest that substituting the Euclidian distance measure with an alternative approach, such as the Mahalanobis distance which controls for covariance in the data (De Maesschalck, Jouan-Rimbaud & Massart, 2000) or the Hassanat distance which is invariant to different scales, noise and outliers (Alkasassbeh, Altarawneh & Hassanat, 2015) can better improve the calculation of similarity between two districts as research suggests that that datasets favor a specific distance metric (Prasatha, Alfeilate, Hassanate, Lasassmehe, Tarawnehf, Alhasanatg & Salmane, 2017; Ho & Pepyne, 2002).

Taking a model-based collaborative filtering recommender system approach to determine Pillar model selection at the student level, instead of at the district level, could also potentially improve the DSEE performance. This could be completed by first building a multi-class model on Pillar student features using *District ID* as the label, then scoring the Target students against this model to probabilistically determine which Pillar model the Target students belong to (Jiang, Qian, Shen, Fu & Mei, 2015). This *District ID* prediction would then be used to select the model(s) used to assign the final dropout prediction, based on data properties collected at the student level. Implementing strategies from the better-performing Knowles model EWS into the design of the DSEE could also increase both the performance and prediction equity. This research shows that the Mean model and DSEE failed to outperform the computationally-intensive stacked model approach used in the Knowles model. By combining the methods and building the best possible district-level Pillar Knowles models, and then generalizing these models to new districts using the weighting algorithm, there is an opportunity for improved DSEE model performance.

Creating Target District personalization of Pillars to only include the Pillar Models that provide the best performance in the pool when generating risk predictions could also increase DSEE AUC scores. Using the historical records available in a Target District as a test set, I could select Pillar Models to include in the pool used for scoring based on performance, rather than electing to use all the Pillar Models. This process is similar to the Knowles approach and could improve overall DSEE performance as the Pillar Model pool used for scoring would only include the optimal, best-performing detectors. Additionally, with the reduced number of models used, the degree of difference in data for the distance calculation could shift, potentially improving the similarity weighting function.

Additionally, future DSEE work that builds on Bowers' research and includes more longitudinal-based non-cumulative features to align with the Bowers GMM approach more closely could improve the performance of the DSEE. Currently, the data used in the DSEE relies on student-level data recorded at the grade level, with each student containing a maximum of 12 records, one for each grade. Increasing the granularity of this feature set to include data collected at the semester-grade level (ex: $9^{th}$ Grade Semester 1 non-cumulative GPA, $9^{th}$ Grade Semester 2 non-cumulative GPA, etc.) could provide additional information for predicting student dropout risk using the DSEE.

Lastly, future work should explore the 'recursive' impact of an intervention on both model performance and design. An effective implementation of an EWS in a school district provides two key outputs; 1) the students at-risk of dropping out and 2) *why* the student is at risk. This output is then used to inform the appropriate intervention needed to put the student back on path to graduation. The successful application of this intervention essentially changes the underlying student data used to both create the EWS and generate the predictions, making EWS design recursive (i.e. there is a half-life on the performance of the EWS before it must be retrained/refreshed to reflect the change of data). While this research explores both EWS design and potential data-driven intervention strategies to be used to mitigate dropout for at-risk students, future research should expand on this recursive issue and explore the impact that these interventions have on EWS design.

Exploring this issue would hopefully lead to new strategies on the frequency at which EWS models should be retrained based on changes in the underlying data. Considering the effort involved with implementing a machine-learning driven EWS, having a better understanding of the cadence at which the model needs to be trained to produce an accurate at-risk prediction

would help educators interested in EWS applications with balancing the resources required with not only implementing the EWS, but also maintaining the successful use of these systems moving forward.

## 5.7 Concluding Remarks

Given the results of this research, there are several conclusions that can be drawn for educators and researchers interested in implementing an early warning system in the school district.

First, creating an early warning system is difficult and costly for educators. The results of this research are the culmination of several years of work, completed with the support of a company that specializes in K-12 student data storage. I was in the fortunate position to work with data that had already been collected and prepared for data modeling. Many schools and districts within the U.S. are not offered this opportunity, and further lack the financial resources to hire a researcher internally gather, clean, prepare, model, and deploy an early warning system into the educational environment. While recent calls by researchers have advocated for EWS code to be published publicly (Agasisti & Bowers, 2017; Bowers et. Al., 2019; Bowers, 2021) improving their accessibility and alleviating some of this burden, there still exist significant barriers to implementing an early warning system in the short-term.

Second, there is a common data theme across the EWSs replicated in this research, with non-cumulative GPA and student absence records often providing the most information for detecting student dropout risk. While I do not advocate for the removal of the other features in EWS design, as the information provided by the other student records (assessments and behavioral data) most likely improves EWS performance; the level of importance of these two features provide an interesting opportunity for educational researchers and school districts

interested in deploying an early warning system. Given the level of effort required to collect and clean student records, there exists the possibility of creating an EWS that simply relies on non-cumulative GPA, absences and student age (as a proxy for retention) to detect risk, significantly reducing the effort required to design and deploy the EWS.

Third, there seems to be an observable performance *ceiling* for early warning systems. While the performance of the EWSs tested in this research shift significantly, depending on design, a perfect detector is never achieved. This *ceiling* is likely due to the data captured on students by educational systems. Research has shown that students dropout for many different reasons, and not accounting for the differences in dropout typology can limit EWS performance (Bowers, 2012a, 2012b). While future work can attempt to address this issue, EWS performance will still be limited to the simple fact that schools will never be able to collect *all* the meaningful data on a student. Additionally, there exists a set of trade-offs between model interpretability and model performance, with the best-performing models using complex opaque methods of analysis and the lowest-performing models using easily interpretable thresh-hold based methods.

Lastly, predicting student performance outcomes is difficult work. A recent systematic literature review of published material between 2010 and 2020 completed by Namoun and Alshanqiti (2021) highlight the major challenges faced by researchers focused on predicting student performance. The results of this review according with my findings: machine-learned driven methods (Random Forest, Hybrid/Stacked models, etc.) outperform traditional methods (linear regression, discriminant analysis, etc.) of EWS design. In addition, Namoun and Alshanqiti (2021)suggest cthat urrent studies implementing machine learning models to predict student outcomes have difficulty with; 1) exploring how student outcomes predictions can assist with automated course and program-level assessments, 2) using multiple datasets from various

disciplines to strengthen predictive model validity, 3) shifting from predictive analytics to explanatory analytics in order to understand the effects of different features on student outcomes to enable better applications of focused interventions, 4) using multiple different metrics for model performance to better evaluate the quality of the predictive solution, 5) exploring unsupervised learning techniques, and 6) applying new technologies such as automated machine-learning to improve efficiency and accessibility to non-technical audiences (Namoun & Alshanqiti, 2021).

To address these challenges, Namoun and Alshanqiti recommend that future studies should focus on 1) formalizing a clear definition of the outcome variable of prediction, 2) build predictive models for non-technical audiences, 3) produce and share datasets for other researchers to explore, 4) build models that predict at the program or cohort level, and 5) implement methods that explain and justify the prediction in way that is actionable to educators (Namoun & Alshanqiti, 2021). While my research provides a clear definition of dropout and incorporates components that explain the model for non-technical stakeholders, I only address three of the five recommendations provided by the authors. As my research was completed in corporation with a private entity (BrightBytes), the choice to share the data with other researchers is this organization's decision rather than mine. Additionally, the data provided by BrightBytes did not include student cohort and program level data which made it infeasible to model at this level when generating at-risk predictions.

In conclusion, this dissertation presents new opportunities in identifying students at risk of dropping out for districts with minimal or no data. Students educated by districts where data is insufficient can now be presented with greater opportunities using proactive interventions driven by predictive modeling rather than being limited to receiving reactive interventions that are often

applied too late, if ever. Research into the design, application, and performance of early warning systems in K-12 education needs to receive continued community support, given the potential benefit of improving student outcomes through proactive data-driven interventions for at-risk students. Ultimately, there needs to be a balance between the effectiveness of the early system and the lift to execute and deploy such a system into practice.

# References


Adelman, C. (2002). The relationship between urbanicity and educational outcomes. Increasing access to college: Extending possibilities for all students, 35-64.

Agasisti, T., & Bowers, A. J. (2017). Data Analytics and Decision-Making in Education: Towards the Educational Data Scientist as a Key Actor in Schools and Higher Education Institutions. In G. Johnes, J. Johnes, T. Agasisti, & L. López-Torres (Eds.), Handbook on the Economics of Education (pp. 184-210). Cheltenham, UK: Edward Elgar Publishing. https://doi.org/10.7916/D8PR95T2

Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B., & Addison, K. L. (2015, March). Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (pp. 93-102).

Aguilar, S., Lonn, S., & Teasley, S. D. (2014). Perceptions and use of an early warning system during a higher education transition program. In Proceedings of the fourth international conference on learning analytics and knowledge (pp. 113-117).

Alkasassbeh, M., Altarawneh, G. A., & Hassanat, A. (2015). On enhancing the performance of nearest neighbour classifiers using hassanat distance metric. arXiv preprint arXiv:1501.00687.

Alliance, E. F. (2015). How the Technology Works. Retrieved, 10(12), 2015.

Allensworth, E. (2013). The Use of Ninth-Grade Early Warning Indicators to Improve Chicago Schools. Journal of Education for Students Placed at Risk (JESPAR), 18(1), 68–83. https://doi.org/10.1080/10824669.2013.745181

Allensworth, E. M., & Easton, J. Q. (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. Research Report. Consortium on Chicago School Research.

Allensworth, E. M., & Easton, J. Q. (2005). The on-track indicator as a predictor of high school graduation. Research Report. Consortium in Chicago Schools Research.

Allensworth, E. M., Nagaoka, J., & Johnson, D. W. (2018). High School Graduation and College Readiness Indicator Systems: What We Know, What We Need to Know. Retrieved from Chicago, IL:

https://consortium.uchicago.edu/sites/default/files/publications/High%20School%20Graduation%20and%20College-April2018-Consortium.pdf

Amatriain, X., Jaimes, A., Oliver, N., & Pujol, J. (2011). Data Mining Methods for Recommender Systems. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), Recommender Systems Handbook (pp. 39-71): Springer US.

Amos, J. (2008). Dropouts, Diplomas, and Dollars: US High Schools and the Nation's Economy. Alliance for Excellent Education. Retrieved from https://all4ed.org/wp-content/uploads/2008/08/Econ2008.pdf

Ananga, E. D. (2011). Typology of school dropout: The dimensions and dynamics of dropout in Ghana. International Journal of Educational Development, 31(4), 374-381.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. Psychological Bulletin, 103(3), 411.

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Reimer, D. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), 6-1.

Baker, R. S. J. D., Gowda, S. M., Corbett, A. T., & Ocumpaugh, J. (2012). Towards Automatically Detecting Whether Student Learning Is Shallow. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), Intelligent Tutoring Systems (pp. 444–453). Springer Berlin Heidelberg.

Baker, R. S., & Koedinger, K. R. (2018). Towards demonstrating the value of learning analytics for K-12 education. Learning analytics in education, 49-62.

Balfanz, R., & Byrnes, V. (2019). Early Warning Indicators and Intervention Systems: State of the Field. In Handbook of Student Engagement Interventions (pp. 45–55). Elsevier.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology, 12, 387-415. DOI: 10.1016/0022-2496(75)90001-2

Barfield, K. A., Hartman, J., & Knight, D. (2012). Early Warning Systems: It's Never Too Early: Researchers from Edvance research understand that longitudinal data

must be available at the school and district level in order to be useful and effective. THE Journal (Technological Horizons In Education), 39(2), 18.

Barnet, B., Arroyo, C., Devoe, M., & Duggan, A. K. (2004). Reduced school dropout rates among adolescent mothers receiving school-based prenatal care. Archives of Pediatrics & Adolescent Medicine, 158(3), 262-268.

Barrington, B. L., & Hendricks, B. (1989). Differentiating characteristics of high school graduates, dropouts, and nongraduates. The journal of educational research, 82(6), 309-319.

Basilico, J., & Hofmann, T. (2004, July). Unifying collaborative and content-based filtering. In Proceedings of the twenty-first international conference on Machine learning (p. 9).

Batista, G. E., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. Applied artificial intelligence, 17(5-6), 519-533.

Bearden, L. J., Spencer, W. A., & Moracco, J. C. (1989). A study of high school dropouts. The School Counselor, 37(2), 113-120.

Bei, Z., Yu, Z., Zhang, H., Xiong, W., Xu, C., Eeckhout, L., & Feng, S. (2015). RFHOC: a random-Forest approach to auto-tuning Hadoop's configuration. IEEE Transactions on Parallel and Distributed Systems, 27(5), 1470-1483.

Bernhardt, V. L. (2004). Data Analysis for Continuous School Improvement (2 ed.). Eye On Education.

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. Journal of machine learning research, 5(Sep), 1089-1105.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V. & Walker, K. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020. URL https://www. microsoft. com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai.

Betts, J. R., & Shkolnik, J. L. (2000). The effects of ability grouping on student achievement and resource allocation in secondary schools. Economics of Education Review, 19(1), 1-15.

Blum, A., Kalai, A., & Langford, J. (1999, July). Beating the hold-out: Bounds for k-fold and progressive cross-validation. In Proceedings of the twelfth annual conference on Computational learning theory (pp. 203-208).

Bowers, A. J. (2008). 'Promoting Excellence: Good to great, NYC's District 2, and the case of a high performing school district'. Leadership and Policy in Schools, 7(2), 154-177.

Bowers, A. J. (2009). Reconsidering grades as data for decision making: More than just academic knowledge. Journal of Educational Administration.

Bowers, A. J. (2010). Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis. Practical Assessment Research and Evaluation, 15(7), 1-18.

Bowers, A. J. (2010). Grades and graduation: A longitudinal risk perspective to identify student dropouts. The Journal of Educational Research, 103(3), 191–207.

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. Educational Research and Evaluation, 17(3), 141-159.

Bowers, A. J. (2017). Quantitative Research Methods Training in Education Leadership and Administration Preparation Programs as Disciplined Inquiry for Building School Improvement Capacity. Journal of Research on Leadership Education, 12(1), 72 - 96. doi:10.1177/1942775116659462

Bowers, A.J. (2021) Early Warning Systems and Indicators of Dropping Out of Upper Secondary School: The Emerging Role of Digital Technologies. In Smart Data and Digital Technology in Education: Learning Analytics, AI and Beyond, Organisation for Economic Co-Operation and Development (OECD) Publishing: Paris, France

Bowers, A. J., Bang, A., Pan, Y., & Graves, K. E. (2019). Education Leadership Data Analytics (ELDA): A White Paper Report on the 2018 ELDA Summit. Retrieved from https://doi.org/10.7916/d8-31a0-pt97

Bowers, A. J., & Sprott, R. (2012). Examining the multiple trajectories associated with dropping out of high school: A growth mixture model analysis. The Journal of Educational Research, 105(3), 176–195.

Bowers, A. J., & Sprott, R. (2012). Why tenth graders fail to finish high school: A dropout typology latent class analysis. Journal of Education for Students Placed at Risk, 17(3), 129-148. doi:10.1080/10824669.2012.692071

Bowers, A. J., Sprott, R., & Taff, S. A. (2012). 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, 77–100.

Bowers, A. J., & Zhou, X. (2019). Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes. Journal of Education for Students Placed at Risk (JESPAR), 24(1), 20–46. https://doi.org/10.1080/10824669.2018.1523734

Bramer, M. (2007). Avoiding overfitting of decision trees. Principles of data mining, 119-134.

Breiman, L., & Cutler, A. (2007). Random forests-classification description. Department of Statistics, Berkeley, 2.

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. Journal of the American society for information science, 45(1), 12-19.

Byrd, R. H., Chin, G. M., Nocedal, J., & Wu, Y. (2012). Sample size selection in optimization methods for machine learning. Mathematical Programming, 134(1), 127-155.

Carlson, S. E. (2018). Identifying students at risk of dropping out: indicators and thresholds using ROC analysis.

Catterall, J. S. (1998). Risk and resilience in student transitions to high school. American journal of Education, 106(2), 302-333.

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11(Jul), 2079-2107.

Centering EquityActionable Intelligence for Social Policy (2020). Centering Racial Equity Throughout Data Integration. Retrieved from https://www.aisp.upenn.edu/centering-equity/

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. City, 1(2), 1.

Chaifetz, J., & Kravitz, R. (2004). Holding Back Students Damages Their Educational Progress: An Advocacy Report. Clearinghouse Rev., 38, 690.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

Christenson, S. L., & Thurlow, M. L. (2004). School dropouts: Prevention considerations, interventions, and challenges. Current Directions in Psychological Science, 13(1), 36-39.

Christie, S. T., Jarratt, D. C., Olson, L. A., & Taijala, T. T. (n.d.). Machine-Learned School Dropout Early Warning at Scale.

Christofferson, R. D., & Callahan, K. (2015). Positive Behavior Support in Schools (PBSIS): An Administrative Perspective on the Implementation of a Comprehensive School-Wide Intervention in an Urban Charter School. Education Leadership Review of Doctoral Research, 2(2), 35-49.

Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. Children and Youth Services Review, 96, 346-353.

Chute, E. (2019). Computer science is elementary: Comprehensive plan for computer science implementation at the elementary level.

Clarke, J. H. (2013). Personalized learning: Student-designed pathways to high school graduation. Corwin Press.

Clark, D., & Martorell, P. (2014). The signaling value of a high school diploma. Journal of Political Economy, 122(2), 282-318.

Clune, B., & Knowles, J. (2016). Delivery of State-Provided Predictive Analytics to Schools: Wisconsin's DEWS and the Proposed EWIMS Dashboard. WCER Working Paper No. 2016-3. Wisconsin Center for Education Research.

Coleman, C., Baker, R. S., & Stephenson, S. (2020). A Better Cold-Start for Early Prediction of Student At-Risk Status in New School Districts.

Cortes, C., Jackel, L. D., & Chiang, W. P. (1995). Limits on learning machine accuracy imposed by data quality. In Advances in Neural Information Processing Systems (pp. 239-246).

D'Agostino, M., & Dardanoni, V. (2009). What's so special about Euclidean distance?. Social Choice and Welfare, 33(2), 211-233.

Darling-Hammond, L., Zielezinski, M. B., & Goldman, S. (2014). Using technology to support at-risk students' learning. Washington, DC: Alliance for Excellent Education.

DasGupta, A. (2011). Probability for statistics and machine learning: fundamentals and advanced topics. Springer Science & Business Media.

Davis, M., Herzog, L., & Legters, N. (2013). Organizing schools to address early warning indicators (EWIs): Common practices and challenges. Journal of Education for Students Placed at Risk (JESPAR), 18(1), 84-100.

De Hoon, M. J., Imoto, S., Nolan, J., & Miyano, S. (2004). Open source clustering software. Bioinformatics, 20(9), 1453-1454.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. Chemometrics and intelligent laboratory systems, 50(1), 1-18.

Delgado R, Tibau X-A (2019) Why Cohen's Kappa should be avoided as performance measure in classification. PLoS ONE 14(9): e0222916. https://doi.org/10.1371/journal.pone.0222916

DeLong, E., DeLong, D., & Clarke-Pearson, D. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics, 44(3), 837-845. DOI: 10.2307/2531595. http://www.jstor.org/stable/2531595

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. ACM computing surveys (CSUR), 27(3), 326-327.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation, 10(7), 1895-1923.

DePaoli, J. L., Fox, J. H., Ingram, E. S., Maushard, M., Bridgeland, J. M., & Balfanz, R. (2015). Building a Grad Nation: Progress and Challenge in Ending the High School Dropout Epidemic. Annual Update 2015. Civic Enterprises.

Dwyer, K., Osher, D., & Warger, C. (1998). Early warning, timely response: A guide to safe schools.

Dwyer, K. P., Osher, D., & Hoffman, C. C. (2000). Creating responsive schools: Contextualizing early warning, timely response. Exceptional Children, 66(3), 347-365. Retrieved from http://ezproxy.cul.columbia.edu/login?url=https://search-proquest-com.ezproxy.cul.columbia.edu/docview/201092749?accountid=10226

Driscoll, A. K. (1999). Risk of high school dropout among immigrant and native Hispanic youth. International Migration Review, 33(4), 857–875.

Duarte, E., & Wainer, J. (2017). Empirical comparison of cross-validation and internal metrics for tuning SVM hyperparameters. Pattern Recognition Letters, 88, 6-11.

Dunn, C., Chambers, D., & Rabren, K. (2004). Variables affecting students' decisions to drop out of school. Remedial and Special Education, 25(5), 314-323.

Dunn, M. C., Kadane, J. B., & Garrow, J. R. (2003). Comparing harm done by mobility and class absence: Missing students and missing data. Journal of Educational and Behavioral Statistics, 28(3), 269–288.

Dupéré, V., Leventhal, T., Dion, E., Crosnoe, R., Archambault, I., & Janosz, M. (2015). Stressors and Turning Points in High School and Dropout: A Stress Process, Life Course Framework. Review of Educational Research, 85(4), 591-629. doi:10.3102/0034654314559845

Easton, J. Q., Johnson, E., & Sartain, L. (2017). The predictive power of ninth-grade GPA. Chicago, IL: University of Chicago Consortium on School Research.

Eide, E. R., & Showalter, M. H. (2001). The effect of grade retention on educational and labor market outcomes. Economics of Education Review, 20(6), 563–576.

Ekstrom, R. B. (1986). Who drops out of high school and why? Findings from a national study. Teachers College Record, 87(3), 356–73.

Ensminger, M. E., & Slusarcick, A. L. (1992). Paths to high school graduation or dropout: A longitudinal study of a first-grade cohort. Sociology of Education, 95–113.

Epstein, J. L., & Sheldon, S. B. (2002). Present and accounted for: Improving student attendance through family and community involvement. The Journal of Educational Research, 95(5), 308–318.

Evermann, J., & Tate, M. (2016). Assessing the predictive performance of structural equation model estimators. Journal of Business Research, 69(10), 4565-4582.

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In Automated Machine Learning (pp. 3-33). Springer, Cham.

Finn, C. E. (1987). The high school dropout puzzle. The Public Interest, 87, 3.

Flach, P. A., Hernández-Orallo, J., & Ramirez, C. F. (2011, January). A coherent interpretation of AUC as a measure of aggregated classification performance. In ICML.

Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. Acm Sigkdd Explorations Newsletter, 12(1), 49-57.

Frazelle, S., & Nagel, A. (2015). A Practitioner's Guide to Implementing Early Warning Systems. REL 2015-056. Regional Educational Laboratory Northwest.

Frazer, L. (1991). At-risk students three years later: We know which ones will drop out. Austin Independent School District.

Freeman, J., & Simonsen, B. (2015). Examining the Impact of Policy and Practice Interventions on High School Dropout and School Completion Rates: A Systematic Review of the Literature. Review of Educational Research, 85(2), 205-248. doi:10.3102/0034654314554431

French, D. C., & Conrad, J. (2001). School dropout as predicted by peer rejection and antisocial behavior. Journal of Research on Adolescence, 11(3), 225–244.

García, S., Fernández, C., & Weiss, C. (2013). Does lengthening the school day reduce the likelihood of early school dropout and grade repetition: Evidence from Colombia. Available at SSRN 2356438.

Gardner, J., & Brooks, C. (2017). Toward Replicable Predictive Model Evaluation in MOOCs. EDM.

Gardner, J., Brooks, C., & Baker, R. (2019, March). Evaluating the fairness of predictive student models through slicing analysis. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge (pp. 225-234).

Gedikli, F., Jannach, D., & Ge, M. (2014). How should I explain? A comparison of different explanation types for recommender systems. International Journal of Human-Computer Studies, 72(4), 367-382.

Gideon, R. A., & Hollister, R. A. (1987). A rank correlation coefficient resistant to outliers. Journal of the American Statistical Association, 82(398), 656-666.

Gilbert, S. N. (1993). Leaving school: Results from a national survey comparing school leavers and high school graduates 18 to 20 years of age. Human Resources and Labour Canada.

Ghojogh, B., & Crowley, M. (2019). The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. arXiv preprint arXiv:1905.12787.

Gleason, P., & Dynarski, M. (2002). Do we know whom to serve? Issues in using risk factors to identify dropouts. Journal of Education for Students Students Placed At Risk, 7(1), 25-41.

Goldschmidt, P., & Wang, J. (1999). When can schools affect dropout behavior? A longitudinal multilevel analysis. American Educational Research Journal, 36(4), 715-738.

Gordon, R., Piana, L. D., & Keleher, T. (2000). Facing the Consequences: An Examination of Racial Discrimination in US Public Schools.

Gregg, N. (2009). Adolescents and adults with learning disabilities and ADHD: Assessment and accommodation. Guilford Press.

Halverson, R., & Smith, A. (2009). How new technologies have (and have not) changed teaching and learning in schools. Journal of Computing in Teacher Education, 26(2), 49-54.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29–36.

Hartnett, S. (2007). Does peer group identity influence absenteeism in high school students? The High School Journal, 91(2), 35–44.

Hassanat, A. B. (2014). Dimensionality invariant similarity measure. arXiv preprint arXiv:1409.0923.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). Random Forests. Springer Series in Statistics, 587–604. doi:10.1007/978-0-387-84858-7_15

Hawkins, D. M. (2004). The Problem of Overfitting. Journal of Chemical Information and Computer Sciences, 44(1), 1–12. doi:10.1021/ci0342472

Heckman, J. J. (2011). The economics of inequality: The value of early childhood education. American Educator, 35(1), 31.

Heppen, J. B., & Therriault, S. B. (2008). Developing Early Warning Systems to Identify Potential High School Dropouts. Issue Brief. National High School Center.

Hesterberg, T. (2011). Bootstrap. Wiley Interdisciplinary Reviews: Computational Statistics, 3(6), 497-526.

Ho, Y. C., & Pepyne, D. L. (2002). Simple explanation of the no-free-lunch theorem and its implications. Journal of optimization theory and applications, 115(3), 549-570.

Hocking, C. (2008). The contributing factors to student absenteeism/truancy and the effectiveness of social services and interventions. Social Work Student Papers, 18.

Hoff, J. (2019). The Impact of Freshmen On-Track Status, Absenteeism, and Associated Demographic Variables on Four-Year Graduation Attainment within a Rural Community: A Predictive Validity Study.

Hoffman, N., Vargas, J., Venezia, A., & Miller, M. S. (2007). Minding the Gap: Why Integrating High School with College Makes Sense and How to Do It. ERIC.

Hoyle, R. H. (1995). The structural equation modeling approach: Basic concepts and fundamental issues.

Hsu, Henry, and Peter A. Lachenbruch. "Paired t test." Encyclopedia of Biostatistics 6 (2005).

Hu, L. Y., Huang, M. W., Ke, S. W., & Tsai, C. F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus, 5(1), 1-9.

Huysamen, J. E. (1999). Demographic-group differences in the prediction of tertiary-academic performance. South African journal of higher education, 13(1), 171-177.

Ingels, S. J., Pratt, D. J., Rogers, J. E., Siegel, P. H., & Stutts, E. S. (2004). Education Longitudinal Study of 2002: Base Year Data File User's Manual. NCES 2004-405. National Center for Education Statistics.

Jabbar, H., & Khan, R. Z. (2015). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Computer Science, Communication and Instrumentation Devices.

Jamalabadi, H., Alizadeh, S., Schönauer, M., Leibold, C., & Gais, S. (2016). Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. Human brain mapping, 37(5), 1842-1855.

Janosz, M., Le Blanc, M., Boulerice, B., & Tremblay, R. E. (2000). Predicting different types of school dropouts: A typological approach with two longitudinal samples. Journal of educational psychology, 92(1), 171.

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data–recommendations for the use of performance metrics. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, 245–251. IEEE.

Jiang, S., Qian, X., Shen, J., Fu, Y., & Mei, T. (2015). Author topic model-based collaborative filtering for personalized POI recommendations. IEEE transactions on multimedia, 17(6), 907-918.

Johnson, E., & Semmelroth, C. (2010). The predictive validity of the early warning system tool. Nassp Bulletin, 94(2), 120–134.

Joksimović, S., Poquet, O., Kovanović, V., Dowell, N., Mills, C., Gašević, D., … Brooks, C. (2018). How Do We Model Learning at Scale? A Systematic Review of Research on MOOCs. Review of Educational Research, 88(1), 43–86. https://doi.org/10.3102/0034654317740335

Jordan, L., Kostandini, G., & Mykerezi, E. (2012). Rural and urban high school dropout rates: Are they different?. Journal of Research in Rural Education (Online), 27(12), 1.

Kassambara, A. "rstatix: pipe-friendly framework for basic statistical tests. R package version 0.4. 0." (2020).

Katuwal, G. J., & Chen, R. (2016). Machine learning model interpretability for precision medicine. arXiv preprint arXiv:1610.09045.

Kaznowski, K. (2004). Slow learners: Are educators leaving them behind?. NASSP Bulletin, 88(641), 31-45.

Kelley, C. T. (1999). Iterative methods for optimization. Society for Industrial and Applied Mathematics.

Kennelly, L., & Monrad, M. (2007). Approaches to Dropout Prevention: Heeding Early Warning Signs with Appropriate Interventions. American Institutes for Research.

Kim, Y., Kim, T. H., & Ergün, T. (2015). The instability of the Pearson correlation coefficient in the presence of coincidental outliers. Finance Research Letters, 13, 243-257.

Koren, Y. (2011). U.S. Patent No. 8,037,080. Washington, DC: U.S. Patent and Trademark Office.

Kostrikov, I., & Gall, J. (2014, September). Depth Sweep Regression Forests for Estimating 3D Human Pose from Images. In BMVC (Vol. 1, No. 2, p. 5).

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003, September). Preventing student dropout in distance learning using machine learning techniques. In International conference on knowledge-based and intelligent information and engineering systems (pp. 267-274). Springer, Berlin, Heidelberg.

Knowles, J. E. (2015). Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin. JEDM | Journal of Educational Data Mining, 7(3), 18–67.

Koon, S., & Petscher, Y. (2015). Comparing Methodologies for Developing an Early Warning System: Classification and Regression Tree Model versus Logistic Regression. REL 2015-077. Regional Educational Laboratory Southeast.

Krislock, N., & Wolkowicz, H. (2012). Euclidean distance matrices and applications. In Handbook on semidefinite, conic and polynomial optimization (pp. 879-914). Springer, Boston, MA.

Krumm, A. E., Waddington, R. J., Teasley, S. D., & Lonn, S. (2014). A learning management system-based early warning system for academic advising in undergraduate engineering. In Learning analytics (pp. 103-119). Springer, New York, NY.

Kuhn, M., & Johnson, K. (2013). Over-fitting and model tuning. In Applied predictive modeling (pp. 61-92). Springer, New York, NY.

Kumari, R., & Srivastava, S. K. (2017). Machine learning: A review on binary classification. International Journal of Computer Applications, 160(7).

Kupersmidt, J. B., & Coie, J. D. (1990). Preadolescent peer status, aggression, and school adjustment as predictors of externalizing problems in adolescence. Child Development, 61(5), 1350–1362.

Kvålseth, T. O. (1989). Note on Cohen's kappa. Psychological reports, 65(1), 223-226.

Lakshminarayan, K., Harp, S. A., Goldman, R. P., & Samad, T. (1996, August). Imputation of Missing Data Using Machine Learning Techniques. In KDD (pp. 140-145).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. biometrics, 159-174.

Lane, K. L., Fletcher, T., Carter, E. W., Dejud, C., & Delorenzo, J. (2007). Paraprofessional-led phonological awareness training with youngsters at risk for reading and behavioral concerns. Remedial and Special Education, 28(5), 266-276.

Lee, J. C., & Staff, J. (2007). When work matters: The varying impact of work intensity on high school dropout. Sociology of Education, 80(2), 158–178.

Lee, S., & Chung, J. Y. (2019). The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. Applied Sciences, 9(15), 3093. https://doi.org/10.3390/app9153093

Legault, L., Green-Demers, I., & Pelletier, L. (2006). Why do high school students lack motivation in the classroom? Toward an understanding of academic amotivation and the role of social support. Journal of Educational Psychology, 98(3), 567.

Lever, J., Krzywinski, M., & Altman, N. (2016). Points of significance: model selection and overfitting.

Li, Q., & Kim, B. M. (2003, October). Clustering approach for hybrid recommender system. In Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003) (pp. 33-38). IEEE.

Lonn, S., Aguilar, S. J., & Teasley, S. D. (2015). Investigating student motivation in the context of a learning analytics intervention during a summer bridge program. Computers in Human Behavior, 47, 90-97.

Mac Iver, M. A., & Mac Iver, D. J. (2009). Beyond the Indicators: An Integrated School-Level Approach to Dropout Prevention. George Washington University Center for Equity and Excellence in Education.

Mac Iver, M. A., & Mac Iver, D. J. (2010). How do we ensure that everyone graduates? An integrated prevention and tiered intervention model for schools and districts. New Directions for Youth Development, 2010(127), 25-35.

Mac Iver, M. A., Stein, M. L., Davis, M. H., Balfanz, R. W., & Fox, J. H. (2019). An Efficacy Study of a Ninth-Grade Early Warning Indicator Intervention. Journal of Research on Educational Effectiveness, 12(3), 363–390.

Managing performance vs. accuracy trade-offs with loop perforation. In Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering (pp. 124-134).

Mandinach, E. B., Honey, M., Light, D., & Brunner, C. (2008). 'A Conceptual Framework for Data-Driven Decision Making'. In E. B. Mandinach & M. Honey (Eds.), Data-Driven School Improvement: Linking Data and Learning (pp. 13-31). New York: Teachers College Press

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. Journal of Thoracic Oncology, 5(9), 1315-1316.

Marlin, B. M., Zemel, R. S., Roweis, S. T., & Slaney, M. (2011, June). Recommender systems: missing data and statistical model estimation. In Twenty-Second International Joint Conference on Artificial Intelligence.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. Expert Systems, 33(1), 107-124.

Marlin, B. (2008). Missing data problems in machine learning (Doctoral dissertation).

Marsh, H. W., & Yeung, A. S. (1997). Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. Journal of educational psychology, 89(1), 41.

Martin, A. J. (2011). Holding back and holding behind: Grade retention and students' non-academic and academic outcomes. British Educational Research Journal, 37(5), 739-763.

Mawhinney-Rhoads, L., & Stahler, G. (2006). Educational policy and reform for homeless students: An overview. Education and Urban Society, 38(3), 288–306.

Mayer, G. R., Sulzer, B., & Cody, J. J. (1968). The use of punishment in modifying student behavior. The journal of special education, 2(3), 323-328.

McCallumore, K. M., & Sparapani, E. F. (2010). The Importance of the Ninth Grade on High School Graduation Rates and Student Success in High School. Education, 130(3).

McCaul, E. (1989). Rural Public School Dropouts: Findings from High School and Beyond. Research in Rural Education, 6(1), 19–24.

McFarland, J., Cui, J., Rathbun, A., & Holmes, J. (2018). Trends in High School Dropout and Completion Rates in the United States: 2018. Compendium Report. NCES 2019-117. National Center for Education Statistics.

McIntire, T. (2004). Student Information Systems Demystified: The Increasing Demand for Accurate, Timely Data Means Schools and Districts Are Relying Heavily on SIS Technologies. Technology & Learning, 24(10), 9.

McKee, M. T., & Caldarella, P. (2016). Middle school predictors of high school performance: A case study of dropout risk indicators. Education, 136(4), 515–529.

McMahon, B. M., & Sembiante, S. F. (2020). Re-envisioning the purpose of early warning systems: Shifting the mindset from student identification to meaningful prediction and intervention. Review of Education, 8(1), 266-301. doi:10.1002/rev3.3183

McNeal Jr, R. B. (1997). Are students being pulled out of high school? The effect of adolescent employment on dropping out. Sociology of Education, 206–220.

McPartland, J., & Jordan, W. (2001). Essential components of high school dropout prevention reforms.

Mensch, B. S., & Kandel, D. B. (1988). Dropping out of high school and drug involvement. Sociology of Education, 95–113.

Miller, S. R., Allensworth, E. M., & Kochanek, J. R. (2002). Student performance: course taking, test scores, and outcomes: the state of chicago public high schools, 1993 to 2000. Consortium on Chicago School Research.

Miller, Shazia, Stuart Luppescu, R. Matt Gladden, and John Q. Easton. 1999. How do [elementary school] graduates perform in CPS high schools? Chicago: Consortium on Chicago School Research.

Moore, A. W. (2001). Cross-validation for detecting and preventing overfitting. School of Computer Science Carneigie Mellon University.

Moreno-Torres, J. G., Sáez, J. A., & Herrera, F. (2012). Study on the impact of partition-induced dataset shift on k-fold cross-validation. IEEE Transactions on Neural Networks and Learning Systems, 23(8), 1304-1312.

Mulligan, S. (2001). Classroom strategies used by teachers of students with attention deficit hyperactivity disorder. Physical & Occupational Therapy in Pediatrics, 20(4), 25-44.

Muthén, B. O. (2001). Latent variable mixture modeling. In New developments and techniques in structural equation modeling (pp. 21–54). Psychology Press.

Nagrecha, S., Dillon, J. Z., & Chawla, N. V. (2017, April). MOOC dropout prediction: lessons learned from making pipelines interpretable. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 351-359).

Namoun, A.; Alshanqiti, A. (2021) Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. Appl. Sci. 2021, 11, 237. https://doi.org/10.3390/ app11010237

Neild, R. C., Balfanz, R., & Herzog, L. (2007). An early warning system. *Educational Leadership*, 65(2), 28–33.

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance?. Bioinformatics, 34(21), 3711-3718.

Niemi, R.E. Clark, B. Saxberg, R. Pea (Eds.) Learning Analytics in Education. Charlotte, NC: Information Age Publishing.

Nowicki, S., Duke, M. P., Sisney, S., Stricker, B., & Tyler, M. A. (2004). Reducing the drop-out rates of at-risk high school students: The effective learning program (ELP). Genetic, social, and general psychology monographs, 130(3), 225-240.

O'Cummings, M., & Therriault, S. B. (2015). From Accountability to Prevention: Early Warning Systems Put Data to Work for Struggling Students. American Institutes for Research.

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. British Journal of Educational Technology, 45(3), 487–501. https://doi.org/10.1111/bjet.12156

Pallas, A. M. (1985). The determinants of high school dropout (educational attainment, transitions).

Pagani, L. S., Vitaro, F., Tremblay, R. E., McDuff, P., Japel, C., & Larose, S. (2008). When predictions fail: The case of unexpected pathways toward high school dropout. Journal of Social Issues, 64(1), 175–194.

Partin, T. C. M., Robertson, R. E., Maggin, D. M., Oliver, R. M., & Wehby, J. H. (2009). Using teacher praise and opportunities to respond to promote appropriate student behavior. Preventing School Failure: Alternative Education for Children and Youth, 54(3), 172-178.

Patel, H. H., Messiah, S. E., Hansen, E., & D'Agostino, E. M. (2020). The relationship between transportation vulnerability, school attendance, and free transportation to an afterschool program for youth. Transportation, 1-19.

Patrick, L., Care, E., & Ainley, M. (2011). The relationship between vocational interests, self-efficacy, and achievement in the prediction of educational pathways. Journal of Career Assessment, 19(1), 61-74.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Pinkus, L. (2008). Using early-warning data to improve graduation rates: Closing cracks in the education system. Washington, DC: Alliance for Excellent Education.

Prasatha, V. S., Alfeilate, H. A. A., Hassanate, A. B., Lasassmehe, O., Tarawnehf, A. S., Alhasanatg, M. B., & Salmane, H. S. E. (2017). Effects of distance measure choice on knn classifier performance-a review. arXiv preprint arXiv:1708.04321.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301.

Quinlan, J. R. (1996, August). Bagging, boosting, and C4. 5. In AAAI/IAAI, Vol. 1 (pp. 725-730).

Ram, N., & Grimm, K. J. (2009). Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. International journal of behavioral development, 33(6), 565-576.

Resnick, P., & Varian, H. R. (1997). Recommender systems. Communications of the ACM, 40(3), 56-58.

Reardon, S. F. (2019). Educational opportunity in early and middle childhood: Using full population administrative data to study variation by place and age. RSF: The Russell Sage Foundation Journal of the Social Sciences, 5(2), 40-68.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Rickles, J., Heppen, J. B., Allensworth, E., Sorensen, N., & Walters, K. (2018). Online credit recovery and the path to on-time high school graduation. educational researcher, 47(8), 481-491.

Roderick, M. (1994). Grade retention and school dropout: Investigating the association. American Educational Research Journal, 31(4), 729–759.

Roderick, M., & Camburn, E. (1996). Academic difficulty during the high school transition. Section III. In PB Sebring, AS Bryk, M. Roderick, and E. Camburn, Charting Reform in Chicago: The Students Speak. Chicago: Consortium on Chicago School

Roderick, M., & Nagaoka, J. (2005). Retention under Chicago's high-stakes testing program: Helpful, harmful, or harmless?. Educational evaluation and policy analysis, 27(4), 309-340. Research.

Romero, C., & Ventura, S. (2010). 'Educational data mining: a review of the state of the art'. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618.

Ross, T. (2016). The differential effects of parental involvement on high school completion and postsecondary attendance. Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, 24, 1-38.

Rumberger, R. W. (1987). High school dropouts: A review of issues and evidence. Review of Educational Research, 57(2), 101–121.

Rumberger, R. W. (2011). Dropping out. Harvard University Press.

Rumberger, R. W. (2020). The economics of high school dropouts. In The Economics of Education (pp. 149-158). Academic Press.

Rumberger, R., Addis, H., Allensworth, E., Balfanz, R., Bruch, J., Dillon, E., & Tuttle, C. (2017). Preventing dropout in secondary schools. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education. Retrieved from https://ies. ed. gov/ncee/wwc/Docs/PracticeGuide/wwc_dropout_092617. pdf.

Russell, D. W., Kahn, J. H., Spoth, R., & Altmaier, E. M. (1998). Analyzing data from experimental studies: A latent variable structural equation modeling approach. Journal of Counseling Psychology, 45(1), 18.

Sansone, D. (2019). Beyond Early Warning Indicators: High School Dropout and Machine Learning. Oxford Bulletin of Economics and Statistics, 81(2), 456–485.

Sao Pedro, M. A., Baker, R. S. J. D., & Gobert, J. D. (2013). What Different Kinds of Stratification Can Reveal About the Generalizability of Data-mined Skill Assessment Models. Proceedings of the Third International Conference on Learning Analytics and Knowledge, 190–194. https://doi.org/10.1145/2460296.2460334

Sara, N. B., Halland, R., Igel, C., & Alstrup, S. (2015). High-school dropout prediction using machine learning: A Danish large-scale study. In ESANN 2015 Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence (pp. 319-24).

Sarle, W. S. (1996). Stopped training and other remedies for overfitting. Computing science and statistics, 352-360.

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In The adaptive web (pp. 291-324). Springer, Berlin, Heidelberg

Schaffer, C. (1993). Selecting a classification method by cross-validation. Machine Learning, 13(1), 135-143.

Schoeneberger, J. A. (2012). Longitudinal attendance patterns: Developing high school dropouts. The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 85(1), 7–14.

Selena, S., & Kenney, M. (2019). Algorithms, Platforms, and Ethnic Bias: A Diagnostic Model. Communications of the Association of Computing Machinery, Forthcoming November.

Sheldon, S. B., & Epstein, J. L. (2002). Improving student behavior and school discipline with family and community involvement. Education and urban society, 35(1), 4-26.

Shernoff, D. J., Csikszentmihalyi, M., Schneider, B., & Shernoff, E. S. (2014). Student engagement in high school classrooms from the perspective of flow theory. In Applications of flow in human development and education (pp. 475-494). Springer, Dordrecht.

Sidiroglou-Douskos, S., Misailovic, S., Hoffmann, H., & Rinard, M. (2011, September).

Siemens, G. (2013). Learning analytics: The emergence of a discipline. American Behavioral Scientist, 57(10), 1380-1400.

Siemens, G., & Long, P. (2011). 'Penetrating the Fog: Analytics in Learning and Education'. EDUCAUSE Review, 46(5), 30.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. Journal of educational statistics, 18(2), 155-195.

Singer, J. D. & Willett, J. B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence. Oxford university press.

Snoek, L., Miletić, S., & Scholte, H. S. (2019). How to control for confounds in decoding analyses of neuroimaging data. NeuroImage, 184, 741-760.

Snyder, T. D., De Brey, C., & Dillow, S. A. (2018). Digest of Education Statistics 2016, NCES 2017-094. National Center for Education Statistics.

Soland, J. (2013). Predicting high school graduation and college enrollment: Comparing early warning indicator data and teacher intuition. Journal of Education for Students Placed at Risk (JESPAR), 18(3-4), 233-262.

Somers, C. L., Owens, D., & Piliawsky, M. (2009). A STUDY OF HIGH SCHOOL DROPOUT PREVENTION AND AT-RISK NINTH GRADERS'ROLE MODELS AND MOTIVATIONS FOR SCHOOL COMPLETION. Education, 130(2).

Sonobe, R., Tani, H., Wang, X., Kobayashi, N., & Shimamura, H. (2014). Parameter tuning in the support vector machine and random forest and their performances in cross-and same-year crop classification using TerraSAR-X. International Journal of Remote Sensing, 35(23), 7898-7909.

Stockwell, D. R., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. Ecological Modelling, 148(1), 1-13.

Strudler, N., & Schrader, P. G. (2016). Engage, empower, explore: An evaluation study of one-toone implementation at twelve urban middle schools. In Presentation at SITE 2016 Conference: Savannah, GA.

Suh, S., & Suh, J. (2007). Risk factors and levels of risk for high school dropouts. Professional School Counseling, 10(3), 2156759X0701000312.

Suh, S., Suh, J., & Houston, I. (2007). Predictors of categorical at-risk high school dropouts. Journal of Counseling & Development, 85(2), 196–203.

Sullivan, R. (2017). Early Warning Signs. A Solution-Finding Report. Retrieved from https://eric.ed.gov/?id=ED583010

Supik, J. D., & Johnson, R. L. (1999). Missing: Texas Youth. Dropout and Attrition Rates in Texas Public High Schools. A Policy Brief.

Sutphen, R. D., Ford, J. P., & Flaherty, C. (2010). Truancy interventions: A review of the research literature. Research on social work practice, 20(2), 161-171.

Swanson, C. B. (2004). Who graduates? Who doesn't?: A statistical portrait of public high school graduation, class of 2001.

Tan, P. N., Kumar, V., & Srivastava, J. (2002, July). Selecting the right interestingness measure for association patterns. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 32-41).

Tan, P. N., Kumar, V., & Srivastava, J. (2004). Selecting the right objective measure for association analysis. Information Systems, 29(4), 293-313.

Tyler, J. H., & Lofstrom, M. (2009). Finishing high school: Alternative pathways and dropout recovery. The future of children, 77-103.

U.S. Bureau of Labor Statistics. (2002) National longitudinalsurvey of youth 1997. Retrieved from http://www.bls.gov/nls/nlsy97.htm

U.S. Census Bureau (2019). Methodology: input data, methodology, and processes for the creation of population and housing unit estimates for the listed geographies. Retrieved from https://www.census.gov/programs-surveys/popest/technical-documentation/methodology.html

U.S. Census Bureau (2019). QuickFacts provides statistics for all states and counties, and for cities and towns with a population of 5,000 or more. Retrieved from https://www.census.gov/quickfacts/fact/table/US/PST045219

U.S. Census Bureau (2020). Urban and Rural. Retrieved from https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html

The U.S. Department of Education. Common Data Education Standards. (2019). WHY CEDS?. Common Data Education Standards. Retrieved from https://ceds.ed.gov/pdf/why-ceds.pdf

The U.S. Department of Education. Common Data Education Standards. (2019). CEDS 101. Common Data Education Standards. Retrieved from https://ceds.ed.gov/pdf/ceds-101.pdf

U.S. Department of Education, National Center for Education Statistics. (2020). The Condition of Education 2020 (NCES 2020-144), Public High School Graduation Rates.

Therriault, S. B., O'Cummings, M., Heppen, J., Yerhot, L., & Scala, J. (2013). High school early warning intervention monitoring system implementation guide. Retrieved from National High School Center at the American Institutes for Research website: http://www.earlywarningsystems.org/wpcontent/uploads/documents/EWSHSImplementationguide2013. pdf.

Upchurch, D. M., & McCarthy, J. (1990). The timing of a first birth and high school completion. American Sociological Review, 224-234.

Venables, W. N., & Ripley, B. D. (2002). Tree-based methods. In Modern Ap-plied Statistics with S (pp. 251-269). Springer, New York, NY.

Vezhnevets, A., & Barinova, O. (2007). Avoiding boosting overfitting by removing confusing samples. In European Conference on Machine Learning (pp. 430-441). Springer, Berlin, Heidelberg.

Wang, B., Liao, Q., & Zhang, C. (2013, August). Weight based KNN recommender system. In 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics (Vol. 2, pp. 449-452). IEEE.

Wang, M., & Bodner, T. E. (2007). Growth mixture modeling: Identifying and predicting unobserved subpopulations with longitudinal data. Organizational Research Methods, 10(4), 635–656.

Wardenaar, K. J. (2020, April 7). Latent Class Growth Analysis and Growth Mixture Modeling using R: A tutorial for two R-packages and a comparison with Mplus. https://doi.org/10.31234/osf.io/m58wx

West, M. R. (2012). Is retaining students in the early grades self-defeating. CCF Brief, 49.

Wiens, T. S., Dale, B. C., Boyce, M. S., & Kershaw, G. P. (2008). Three way k-fold cross-validation of resource selection functions. Ecological Modelling, 212(3-4), 244-255.

Willett, J. B., & Singer, J. D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. Review of educational research, 61(4), 407-450.

Wilson, S. E. (2015). Methods for clustering data with missing values. url: https://www. math. leidenuniv. nl/scripties/MasterWilson. pdf (visited on 11/02/2016).

Wood, L., Kiperman, S., Esch, R. C., Leroux, A. J., & Truscott, S. D. (2017). Predicting dropout using student-and school-level factors: An ecological perspective. School Psychology Quarterly, 32(1), 35.

Wong, T. T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition, 48(9), 2839-2846.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). “pROC: an open-source package for R and S+ to analyze and compare ROC curves”. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77

Yadav, S., & Shukla, S. (2016, February). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In 2016 IEEE 6th International conference on advanced computing (IACC) (pp. 78-83). IEEE.

Yordanova, Katerina, and Ivo Emanuilov. "Do You Believe in FAIR-y-tales? An Overview of Microsoft’s New Toolkit for Assessing and Improving Fairness of Algorithms." (2020).

Zhang, Z., Kwok, J. T., & Yeung, D. Y. (2003, August). Parametric distance metric learning with label information. In IJCAI (Vol. 1450).

Zhang, S., Zhang, J., Zhu, X., Qin, Y., & Zhang, C. (2008). Missing value imputation based on data clustering. In Transactions on computational science I (pp. 128-138). Springer, Berlin, Heidelberg.

Žliobaitė, I., Bifet, A., Read, J., Pfahringer, B., & Holmes, G. (2015). Evaluation methods and decision theory for classification of streaming data with temporal dependence. Machine Learning, 98(3), 455-482.

# Appendix A: Study Participant Descriptive Tables & Figures

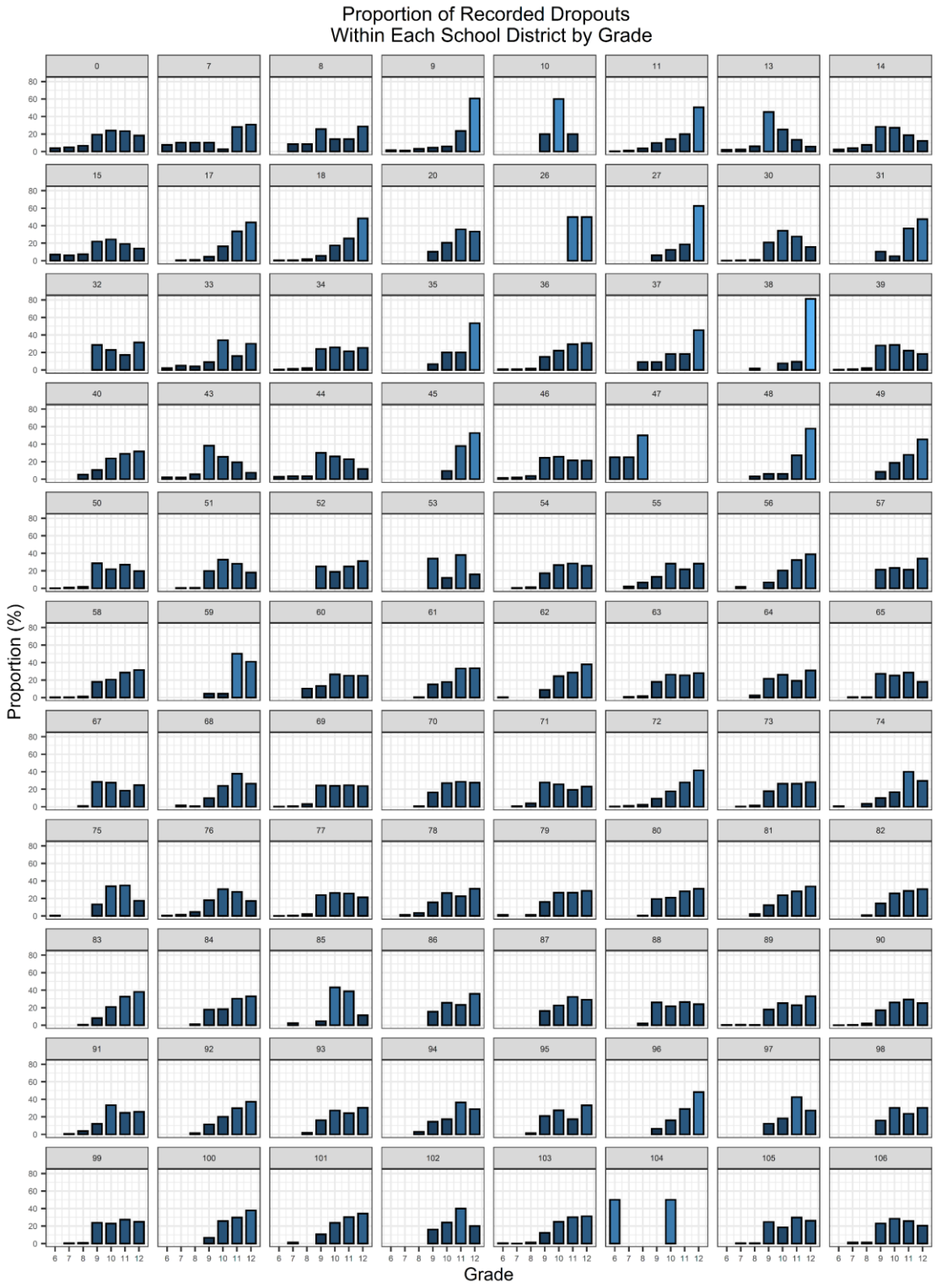



**Figure 38: Proportion of Dropout Records by Grade and School District**

**Table 12: Student Gender Distribution Within Each District[9]**

| Org ID | Female | Male | Missing | (%) Female | (%) Male | (%) Missing | Total Students |
|---|---|---|---|---|---|---|---|
| 0 | 20,000 | 16,000 | 1000 | 49% | 48% | 3% | 33,000 |
| 7 | 2,000 | 2,000 | 0 | 49% | 51% | 1% | 3,000 |
| 8 | 1,000 | 1,000 | 0 | 51% | 48% | 1% | 2,000 |
| 9 | 2,000 | 2,000 | 0 | 49% | 50% | 1% | 5,000 |
| 10 | 0 | 0 | 0 | 67% | 33% | 0% | 0 |
| 11 | 5,000 | 5,000 | 0 | 50% | 46% | 4% | 11,000 |
| 13 | 4,000 | 4,000 | 0 | 49% | 47% | 4% | 9,000 |
| 14 | 7,000 | 6,000 | 0 | 50% | 47% | 3% | 13,000 |
| 15 | 7,000 | 7,000 | 1000 | 49% | 48% | 3% | 15,000 |
| 17 | 1000 | 1000 | 0 | 46% | 53% | 1% | 2,000 |
| 18 | 9,000 | 9,000 | 1,000 | 47% | 47% | 7% | 20,000 |
| 20 | 0 | 0 | 0 | 37% | 63% | 0% | 0 |
| 26 | 0 | 0 | 0 | 51% | 49% | 0% | 0 |
| 27 | 1000 | 1000 | 0 | 50% | 50% | 0% | 1,000 |
| 30 | 16,000 | 16,000 | 2,000 | 48% | 47% | 5% | 34,000 |
| 31 | 0 | 0 | 0 | 22% | 33% | 45% | 0 |
| 32 | 1000 | 1000 | 0 | 51% | 49% | 0% | 1,000 |
| 33 | 0 | 0 | 0 | 46% | 53% | 1% | 1000 |
| 34 | 1000 | 1000 | 0 | 50% | 49% | 1% | 1,000 |
| 35 | 0 | 0 | 0 | 43% | 52% | 5% | 0 |
| 36 | 1000 | 1000 | 0 | 47% | 48% | 5% | 2,000 |
| 37 | 0 | 0 | 0 | 51% | 49% | 0% | 0 |
| 38 | 0 | 0 | 0 | 42% | 56% | 2% | 0 |
| 39 | 1000 | 1,000 | 0 | 39% | 56% | 5% | 3,000 |
| 40 | 0 | 0 | 0 | 43% | 55% | 2% | 0 |
| 43 | 2,000 | 2,000 | 0 | 50% | 49% | 2% | 5,000 |
| 44 | 3,000 | 3,000 | 0 | 50% | 49% | 1% | 5,000 |
| 45 | 0 | 0 | 1000 | 13% | 10% | 77% | 1000 |
| 46 | 8,000 | 8,000 | 1000 | 50% | 45% | 5% | 17,000 |
| 47 | 0 | 0 | 0 | 44% | 22% | 33% | 0 |
| 48 | 0 | 0 | 0 | 32% | 68% | 0% | 0 |
| 49 | 1000 | 1000 | 0 | 44% | 55% | 0% | 1,000 |
| 50 | 4,000 | 4,000 | 0 | 50% | 49% | 1% | 8,000 |
| 51 | 1,000 | 1,000 | 0 | 48% | 52% | 1% | 2,000 |
| 52 | 1000 | 1000 | 0 | 50% | 50% | 0% | 1,000 |
| 53 | 1000 | 1000 | 0 | 50% | 50% | 0% | 2,000 |
| 54 | 3,000 | 3,000 | 0 | 50% | 50% | 1% | 6,000 |
| 55 | 0 | 0 | 0 | 50% | 50% | 0% | 1000 |
| 56 | 0 | 0 | 0 | 52% | 48% | 0% | 1000 |
| 57 | 0 | 0 | 0 | 48% | 52% | 0% | 1000 |
| 58 | 2,000 | 2,000 | 0 | 46% | 53% | 1% | 4,000 |

[9] Values rounded to obfuscate school individual districts

| Org ID | Female | Male | Missing | (%) Female | (%) Male | (%) Missing | Total Students |
|---|---|---|---|---|---|---|---|
| 59 | 0 | 0 | 0 | 47% | 53% | 0% | 1000 |
| 60 | 0 | 1000 | 0 | 46% | 54% | 0% | 1000 |
| 61 | 1,000 | 1,000 | 0 | 48% | 52% | 0% | 3,000 |
| 62 | 1000 | 1000 | 0 | 47% | 53% | 0% | 2,000 |
| 63 | 1,000 | 1,000 | 0 | 49% | 51% | 0% | 2,000 |
| 64 | 1000 | 1000 | 0 | 50% | 50% | 0% | 1,000 |
| 65 | 3,000 | 3,000 | 0 | 50% | 50% | 1% | 6,000 |
| 67 | 1,000 | 1,000 | 0 | 47% | 53% | 0% | 2,000 |
| 68 | 2,000 | 2,000 | 0 | 49% | 50% | 1% | 4,000 |
| 69 | 7,000 | 7,000 | 0 | 48% | 51% | 1% | 14,000 |
| 70 | 1000 | 1000 | 0 | 49% | 51% | 0% | 1,000 |
| 71 | 1000 | 1000 | 0 | 49% | 50% | 1% | 1,000 |
| 72 | 2,000 | 2,000 | 0 | 49% | 51% | 1% | 3,000 |
| 73 | 2,000 | 2,000 | 0 | 49% | 50% | 1% | 4,000 |
| 74 | 1,000 | 1,000 | 0 | 49% | 51% | 0% | 3,000 |
| 75 | 1,000 | 1,000 | 0 | 47% | 52% | 0% | 2,000 |
| 76 | 1000 | 1000 | 0 | 46% | 53% | 1% | 1,000 |
| 77 | 2,000 | 2,000 | 0 | 49% | 51% | 1% | 5,000 |
| 78 | 1,000 | 1,000 | 0 | 49% | 51% | 0% | 2,000 |
| 79 | 1000 | 1000 | 0 | 48% | 52% | 1% | 1,000 |
| 80 | 3,000 | 3,000 | 0 | 48% | 51% | 1% | 6,000 |
| 81 | 0 | 0 | 0 | 49% | 52% | 0% | 1000 |
| 82 | 1000 | 1000 | 0 | 49% | 51% | 0% | 1,000 |
| 83 | 1000 | 1,000 | 0 | 47% | 52% | 1% | 2,000 |
| 84 | 1,000 | 1,000 | 0 | 49% | 50% | 0% | 3,000 |
| 85 | 0 | 0 | 0 | 48% | 52% | 0% | 1000 |
| 86 | 0 | 0 | 0 | 47% | 53% | 0% | 1000 |
| 87 | 0 | 0 | 0 | 46% | 54% | 0% | 1000 |
| 88 | 1,000 | 1,000 | 0 | 49% | 51% | 0% | 2,000 |
| 89 | 2,000 | 3,000 | 0 | 48% | 52% | 1% | 5,000 |
| 90 | 3,000 | 3,000 | 0 | 50% | 50% | 1% | 6,000 |
| 91 | 1,000 | 1,000 | 0 | 48% | 51% | 1% | 2,000 |
| 92 | 0 | 0 | 0 | 52% | 48% | 0% | 1000 |
| 93 | 1000 | 1000 | 0 | 46% | 54% | 0% | 1,000 |
| 94 | 0 | 0 | 0 | 49% | 51% | 0% | 1000 |
| 95 | 1000 | 1000 | 0 | 48% | 52% | 0% | 1,000 |
| 96 | 0 | 0 | 0 | 51% | 50% | 0% | 1000 |
| 97 | 0 | 0 | 0 | 49% | 50% | 0% | 1000 |
| 98 | 1000 | 1000 | 0 | 49% | 51% | 0% | 2,000 |
| 99 | 2,000 | 2,000 | 0 | 48% | 52% | 0% | 4,000 |
| 100 | 0 | 0 | 0 | 47% | 53% | 0% | 1000 |
| 101 | 1000 | 1000 | 0 | 48% | 51% | 1% | 2,000 |
| 102 | 0 | 0 | 0 | 48% | 52% | 0% | 1000 |
| 103 | 3,000 | 3,000 | 0 | 48% | 51% | 1% | 6,000 |

| Org ID | Female | Male | Missing | (%) Female | (%) Male | (%) Missing | Total Students |
|---|---|---|---|---|---|---|---|
| 104 | 0 | 0 | 0 | 10% | 30% | 60% | 0 |
| 105 | 1,000 | 1,000 | 0 | 49% | 51% | 0% | 2,000 |
| 106 | 0 | 0 | 0 | 48% | 52% | 0% | 1000 |
| **$\bar{X}$** | **2,000** | **2,000** | **0** | **47%** | **50%** | **3%** | **4,000** |

**Table 13: Student Ethnicity Distribution Within Each District[10][11]**

| Org ID | AS | AA | HIS | IND | MU | PI | UN | WH | NA | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1,100 | 2,300 | 15,100 | 900 | 200 | - | - | 12,900 | 900 | 33,400 |
| 7 | 300 | 300 | 1,700 | 0 | 100 | 0 | 0 | 1,000 | 0 | 3,400 |
| 8 | 100 | 400 | 1,400 | 0 | 0 | 0 | 0 | 200 | 0 | 2,200 |
| 9 | 400 | 200 | 2,700 | 0 | 0 | 200 | 900 | 300 | 100 | 4,700 |
| 10 | - | - | 0 | - | - | - | - | 0 | - | 0 |
| 11 | 200 | 1,400 | 2,500 | 0 | 200 | 0 | - | 6,200 | 400 | 11,000 |
| 13 | 100 | 6,400 | 200 | 0 | 100 | - | 0 | 1,800 | 300 | 9,000 |
| 14 | 800 | 9,000 | 1,500 | 0 | 200 | 0 | - | 1,500 | 400 | 13,500 |
| 15 | 400 | 7,200 | 900 | 0 | 500 | 0 | - | 5,600 | 500 | 15,200 |
| 17 | 0 | 100 | 0 | 0 | 0 | - | 600 | 800 | 0 | 1,600 |
| 18 | 900 | 2,200 | 2,300 | 100 | 700 | 0 | 6,100 | 6,200 | 1,300 | 19,800 |
| 20 | - | 0 | - | - | - | - | 0 | 0 | - | 0 |
| 26 | 0 | 0 | 0 | - | 0 | - | 100 | - | - | 100 |
| 27 | 0 | 0 | 100 | 0 | 0 | 0 | - | 1,000 | - | 1,200 |
| 30 | 1,100 | 11,500 | 2,000 | 0 | 500 | 0 | 0 | 17,200 | 1,600 | 33,900 |
| 31 | - | 0 | 0 | - | - | - | - | 100 | 100 | 200 |
| 32 | 0 | 0 | 0 | 0 | 0 | - | - | 1,100 | 0 | 1,200 |
| 33 | 0 | 100 | 100 | 0 | 0 | - | 0 | 400 | 0 | 700 |
| 34 | 0 | 100 | 0 | - | 0 | - | 0 | 1,300 | 0 | 1,400 |
| 35 | 0 | 0 | 0 | - | 0 | - | - | 300 | 0 | 400 |
| 36 | 0 | 500 | 0 | 0 | 200 | 0 | - | 1,000 | 100 | 1,800 |
| 37 | 0 | 0 | 200 | 0 | 0 | - | 0 | 200 | 0 | 500 |
| 38 | - | - | 100 | 0 | - | - | - | 0 | 0 | 100 |
| 39 | 0 | 400 | 1,500 | 0 | 0 | - | - | 500 | 100 | 2,500 |
| 40 | 0 | 0 | 200 | - | 0 | - | - | 0 | 0 | 200 |
| 43 | 100 | 3,600 | 100 | 0 | 100 | 0 | - | 1,000 | 100 | 4,900 |
| 44 | 100 | 3,200 | 300 | 0 | 100 | 0 | - | 1,500 | 100 | 5,200 |
| 45 | 0 | 0 | 0 | 0 | - | - | - | 200 | 700 | 900 |
| 46 | 300 | 12,500 | 200 | 0 | 1,100 | 0 | 0 | 1,900 | 800 | 17,000 |
| 47 | - | - | 0 | - | - | - | 0 | 0 | 0 | 0 |
| 48 | - | - | 0 | 0 | - | - | 0 | 0 | - | 100 |
| 49 | 0 | 0 | 0 | 0 | 0 | - | - | 1,200 | 0 | 1,200 |
| 50 | 100 | 900 | 400 | 0 | 100 | - | - | 6,800 | 100 | 8,300 |

[10] AS = Asian, AA = African American, HIS = Hispanic, IND = Indigenous, PI = Pacific Islander, UN = Undefined, WH = White, NA = Missing Record

[11] Values rounded to nearest hundred obfuscate school individual districts

| Org ID | AS | AA | HIS | IND | MU | PI | UN | WH | NA | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| 51 | 0 | 0 | 0 | 0 | 0 | - | - | 2,100 | 0 | 2,100 |
| 52 | 0 | 0 | 0 | - | 0 | - | - | 1,100 | 0 | 1,200 |
| 53 | 0 | 0 | 0 | 0 | 0 | - | - | 1,700 | 0 | 1,700 |
| 54 | 100 | 400 | 100 | 0 | 100 | - | - | 5,400 | 0 | 6,000 |
| 55 | 0 | 0 | 0 | 0 | 0 | - | - | 600 | - | 600 |
| 56 | 0 | 0 | 0 | - | - | - | - | 900 | - | 900 |
| 57 | 0 | 0 | 0 | 0 | 0 | 0 | - | 700 | - | 700 |
| 58 | 0 | 200 | 0 | 0 | 0 | - | - | 3,300 | 0 | 3,600 |
| 59 | - | 0 | 0 | - | 0 | - | - | 500 | 0 | 500 |
| 60 | 0 | 0 | 0 | - | - | - | - | 1,000 | 0 | 1,000 |
| 61 | 0 | 100 | 0 | 0 | 0 | - | - | 2,500 | 0 | 2,700 |
| 62 | 0 | 0 | 0 | 0 | 0 | - | - | 1,800 | 0 | 1,900 |
| 63 | 0 | 100 | 0 | 0 | 0 | - | - | 2,000 | 0 | 2,100 |
| 64 | 0 | 0 | 0 | - | 0 | - | - | 1,100 | - | 1,200 |
| 65 | 0 | 100 | 0 | 0 | 0 | - | - | 5,300 | 0 | 5,500 |
| 67 | 0 | 0 | 0 | 0 | 0 | - | - | 2,400 | 0 | 2,500 |
| 68 | 100 | 400 | 200 | 0 | 100 | 0 | - | 3,400 | 0 | 4,200 |
| 69 | 200 | 1,600 | 100 | 0 | 0 | 0 | - | 12,300 | 200 | 14,400 |
| 70 | 0 | 0 | 0 | 0 | 0 | - | - | 1,400 | 0 | 1,400 |
| 71 | - | 0 | 0 | - | 0 | - | - | 1,400 | 0 | 1,500 |
| 72 | 0 | 100 | 0 | - | 0 | - | - | 3,100 | 0 | 3,200 |
| 73 | 0 | 200 | 0 | 0 | 0 | 0 | - | 3,800 | 0 | 4,200 |
| 74 | 0 | 0 | 0 | 0 | 0 | 0 | - | 2,600 | 0 | 2,600 |
| 75 | 0 | 0 | 0 | 0 | 0 | - | - | 2,100 | 0 | 2,200 |
| 76 | 0 | 200 | 0 | - | 0 | - | - | 1,200 | 0 | 1,400 |
| 77 | 0 | 400 | 0 | 0 | 0 | 0 | - | 4,000 | 0 | 4,500 |
| 78 | 0 | 100 | 0 | 0 | 0 | - | - | 2,200 | 0 | 2,300 |
| 79 | 0 | 0 | 0 | - | 0 | - | - | 1,300 | 0 | 1,300 |
| 80 | 200 | 200 | 100 | 0 | 100 | 0 | - | 5,000 | 0 | 5,600 |
| 81 | - | 0 | 0 | - | 0 | - | - | 1,000 | - | 1,000 |
| 82 | 0 | 0 | 0 | 0 | 0 | - | - | 1,300 | 0 | 1,400 |
| 83 | 0 | 0 | 0 | - | 0 | - | - | 2,100 | 0 | 2,100 |
| 84 | 0 | 200 | 0 | 0 | 0 | 0 | - | 2,600 | 0 | 2,800 |
| 85 | 0 | 0 | - | 0 | 0 | - | - | 500 | - | 600 |
| 86 | 0 | 0 | 0 | - | 0 | - | - | 700 | 0 | 700 |
| 87 | - | 0 | 0 | - | 0 | - | - | 600 | 0 | 600 |
| 88 | 0 | 0 | 0 | - | 0 | - | - | 2,200 | 0 | 2,200 |
| 89 | 0 | 100 | 0 | 0 | 0 | - | - | 4,800 | 0 | 5,000 |
| 90 | 100 | 600 | 0 | 0 | 0 | - | - | 5,200 | 0 | 6,000 |
| 91 | 0 | 0 | 0 | 0 | 0 | - | - | 2,000 | 0 | 2,100 |
| 92 | 0 | 0 | 0 | - | 0 | - | - | 800 | 0 | 800 |

| Org ID | AS | AA | HIS | IND | MU | PI | UN | WH | NA | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| 93 | 0 | 0 | 0 | 0 | 0 | - | - | 1,200 | 0 | 1,200 |
| 94 | 0 | 0 | 0 | - | 0 | 0 | - | 700 | 0 | 800 |
| 95 | 0 | 0 | 0 | 0 | 0 | - | - | 1,200 | 0 | 1,200 |
| 96 | 0 | 0 | 0 | 0 | - | - | - | 600 | - | 600 |
| 97 | 0 | 0 | - | 0 | 0 | - | - | 800 | 0 | 800 |
| 98 | 0 | 0 | 0 | 0 | 0 | - | - | 1,900 | 0 | 1,900 |
| 99 | 0 | 0 | 0 | 0 | 0 | - | - | 3,700 | 0 | 3,800 |
| 100 | - | 0 | 0 | - | 0 | - | - | 800 | 0 | 800 |
| 101 | 0 | 0 | 0 | - | 0 | - | - | 1,500 | 0 | 1,600 |
| 102 | 0 | 0 | 0 | 0 | - | - | - | 500 | - | 500 |
| 103 | 100 | 100 | 0 | 0 | 0 | - | - | 6,100 | 0 | 6,400 |
| 104 | - | - | - | - | - | - | - | 0 | 0 | 0 |
| 105 | 0 | 0 | 0 | 0 | 0 | - | - | 2,100 | 0 | 2,100 |
| 106 | - | - | - | - | - | 800 | - | - | 0 | 800 |
| **$\bar{X}$** | **100** | **800** | **400** | **0** | **100** | **0** | **100** | **2,200** | **100** | **3,700** |
| **Σ** | **7,100** | **68,000** | **34,600** | **1,300** | **4,900** | **1,200** | **7,900** | **193,300** | **8,300** | **326,500** |

**Table 14: Count of Recorded Student Outcomes Within Each District[12]**

| Org ID | Dropout | Graduated | Total Students | Dropout Rate |
|---|---|---|---|---|
| 0 | 4,700 | 28,700 | 33,400 | 14% |
| 7 | 0 | 3,400 | 3,400 | 1% |
| 8 | 0 | 2,100 | 2,200 | 2% |
| 9 | 400 | 4,300 | 4,700 | 10% |
| 10 | 0 | 0 | 0 | 83% |
| 11 | 1,200 | 9,700 | 11,000 | 11% |
| 13 | 1,600 | 7,300 | 9,000 | 18% |
| 14 | 1,100 | 12,400 | 13,500 | 8% |
| 15 | 1,700 | 13,600 | 15,200 | 11% |
| 17 | 200 | 1,500 | 1,600 | 11% |
| 18 | 3,100 | 16,700 | 19,800 | 16% |
| 20 | 0 | 0 | 0 | 85% |
| 26 | 0 | 100 | 100 | 2% |
| 27 | 0 | 1,200 | 1,200 | 3% |
| 30 | 3,400 | 30,500 | 33,900 | 10% |
| 31 | 0 | 200 | 200 | 11% |
| 32 | 0 | 1,200 | 1,200 | 3% |
| 33 | 100 | 600 | 700 | 17% |
| 34 | 300 | 1,100 | 1,400 | 22% |
| 35 | 0 | 400 | 400 | 5% |
| 36 | 400 | 1,400 | 1,800 | 23% |
| 37 | 0 | 500 | 500 | 2% |
| 38 | 100 | 0 | 100 | 52% |
| 39 | 1,900 | 600 | 2,500 | 76% |
| 40 | 100 | 200 | 200 | 32% |
| 43 | 600 | 4,300 | 4,900 | 12% |
| 44 | 400 | 4,800 | 5,200 | 7% |
| 45 | 100 | 900 | 900 | 8% |
| 46 | 1,800 | 15,200 | 17,000 | 10% |
| 47 | 0 | 0 | 0 | 89% |
| 48 | 0 | 0 | 100 | 61% |
| 49 | 100 | 1,100 | 1,200 | 8% |
| 50 | 700 | 7,600 | 8,300 | 9% |
| 51 | 300 | 1,900 | 2,100 | 12% |
| 52 | 100 | 1,000 | 1,200 | 11% |
| 53 | 100 | 1,700 | 1,700 | 3% |

[12] Values rounded nearest hundred to obfuscate school individual districts

| Org ID | Dropout | Graduated | Total Students | Dropout Rate |
|---|---|---|---|---|
| 54 | 700 | 5,400 | 6,000 | 12% |
| 55 | 0 | 500 | 600 | 8% |
| 56 | 100 | 900 | 900 | 6% |
| 57 | 0 | 600 | 700 | 7% |
| 58 | 500 | 3,100 | 3,600 | 14% |
| 59 | 0 | 500 | 500 | 4% |
| 60 | 100 | 900 | 1,000 | 7% |
| 61 | 300 | 2,400 | 2,700 | 10% |
| 62 | 200 | 1,700 | 1,900 | 12% |
| 63 | 100 | 1,900 | 2,100 | 6% |
| 64 | 100 | 1,100 | 1,200 | 7% |
| 65 | 500 | 5,000 | 5,500 | 9% |
| 67 | 200 | 2,300 | 2,500 | 9% |
| 68 | 300 | 3,900 | 4,200 | 7% |
| 69 | 2,100 | 12,300 | 14,400 | 14% |
| 70 | 200 | 1,200 | 1,400 | 14% |
| 71 | 200 | 1,300 | 1,500 | 13% |
| 72 | 300 | 2,900 | 3,200 | 8% |
| 73 | 400 | 3,800 | 4,200 | 9% |
| 74 | 200 | 2,400 | 2,600 | 8% |
| 75 | 200 | 2,000 | 2,200 | 9% |
| 76 | 200 | 1,200 | 1,400 | 14% |
| 77 | 600 | 4,000 | 4,500 | 12% |
| 78 | 100 | 2,200 | 2,300 | 4% |
| 79 | 100 | 1,200 | 1,300 | 7% |
| 80 | 600 | 5,000 | 5,600 | 10% |
| 81 | 100 | 900 | 1,000 | 9% |
| 82 | 100 | 1,300 | 1,400 | 8% |
| 83 | 200 | 1,900 | 2,100 | 9% |
| 84 | 200 | 2,600 | 2,800 | 7% |
| 85 | 0 | 500 | 600 | 8% |
| 86 | 0 | 700 | 700 | 5% |
| 87 | 100 | 600 | 600 | 10% |
| 88 | 200 | 2,000 | 2,200 | 11% |
| 89 | 300 | 4,600 | 5,000 | 7% |
| 90 | 700 | 5,400 | 6,000 | 11% |
| 91 | 200 | 1,900 | 2,100 | 9% |
| 92 | 100 | 800 | 800 | 9% |
| 93 | 100 | 1,100 | 1,200 | 9% |
| 94 | 100 | 700 | 800 | 14% |
| 95 | 200 | 1,100 | 1,200 | 13% |
| 96 | 0 | 600 | 600 | 5% |

| Org ID | Dropout | Graduated | Total Students | Dropout Rate |
|---|---|---|---|---|
| 97 | 0 | 800 | 800 | 4% |
| 98 | 300 | 1,700 | 1,900 | 13% |
| 99 | 500 | 3,300 | 3,800 | 12% |
| 100 | 100 | 800 | 800 | 9% |
| 101 | 100 | 1,500 | 1,600 | 5% |
| 102 | 0 | 500 | 500 | 5% |
| 103 | 600 | 5,900 | 6,400 | 9% |
| 104 | 0 | 0 | 0 | 20% |
| 105 | 300 | 1,900 | 2,100 | 13% |
| **$\bar{X}$** | **400** | **3,300** | **3,700** | **14%** |

# Appendix B: Similarity Data Descriptive Tables

**Table 15: Reported Demographic Descritpives of Current Student Populations[13][14]**

| Org ID | Graduation Rate[15] | Total Students | Expressed as a % | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | PI | IND | AS | AA | WH | HIS | UN |
| 0 | 83% | 45000 | 0% | 4% | 2% | 9% | 20% | 64% | 0% |
| 7 | 94% | 11000 | 0% | 0% | 7% | 8% | 25% | 57% | 1% |
| 8 | 80% | 10000 | 2% | 0% | 2% | 18% | 8% | 67% | 1% |
| 9 | 94% | 9000 | 4% | 0% | 11% | 3% | 5% | 76% | 0% |
| 10 | 88% | 13000 | 0% | 1% | 4% | 19% | 61% | 15% | 0% |
| 11 | 85% | 48000 | 0% | 0% | 2% | 13% | 45% | 36% | 0% |
| 13 | 79% | 21000 | 0% | 0% | 1% | 77% | 13% | 5% | 0% |
| 14 | 78% | 99000 | 0% | 0% | 7% | 61% | 12% | 18% | 0% |
| 15 | 87% | 42000 | 0% | 0% | 3% | 54% | 28% | 10% | 0% |
| 17 | 92% | 5000 | 0% | 0% | 2% | 10% | 76% | 6% | 0% |
| 18 | 88% | 31000 | 0% | 1% | 8% | 20% | 37% | 28% | 0% |
| 20 | 94% | 2000 | 0% | 0% | 1% | 2% | 95% | 3% | 0% |
| 26 | 93% | 0 | 0% | 0% | 0% | 0% | 0% | 100% | 0% |
| 27 | 92% | 1000 | 0% | 0% | 8% | 0% | 65% | 25% | 0% |
| 30 | 89% | 9000 | 0% | 0% | 5% | 37% | 41% | 12% | 0% |
| 31 | 96% | 2000 | 0% | 1% | 1% | 5% | 82% | 11% | 0% |
| 32 | 90% | 4000 | 0% | 0% | 2% | 2% | 92% | 3% | 0% |
| 33 | 67% | 2000 | 0% | 0% | 2% | 10% | 46% | 37% | 0% |
| 34 | 67% | 3000 | 0% | 0% | 1% | 5% | 79% | 9% | 0% |
| 35 | 92% | 1000 | 0% | 0% | 1% | 12% | 72% | 8% | 0% |
| 36 | 84% | 9000 | 0% | 0% | 1% | 22% | 53% | 8% | 0% |
| 37 | 85% | 2000 | 0% | 0% | 0% | 0% | 41% | 57% | 0% |
| 38 | 82% | 1000 | 0% | 1% | 1% | 0% | 25% | 73% | 0% |
| 39 | 0% | 0 | 0% | 0% | 0% | 33% | 0% | 67% | 0% |
| 40 | 91% | 3000 | 0% | 0% | 15% | 38% | 22% | 16% | 0% |
| 43 | 81% | 23000 | 0% | 0% | 1% | 70% | 19% | 6% | 0% |
| 44 | 89% | 28000 | 0% | 0% | 3% | 57% | 20% | 12% | 0% |
| 45 | 92% | 4000 | 0% | 0% | 1% | 5% | 90% | 4% | 0% |
| 46 | 0% | 0 | 0% | 1% | 0% | 84% | 1% | 11% | 0% |
| 47 | 98% | 4000 | 0% | 1% | 0% | 3% | 50% | 44% | 0% |
| 48 | 87% | 3000 | 0% | 55% | 0% | 0% | 37% | 0% | 6% |

[13] AS = Asian, AA = African American, HIS = Hispanic, IND = Indigenous, PI = Pacific Islander, UN = Undefined, WH = White, NA = Missing Record

[14] Values rounded to nearest thousand obfuscate school individual districts

[15] As Reported to NCES for all public institutions.

| Org ID | Graduation Rate[15] | Total Students | Expressed as a % | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | PI | IND | AS | AA | WH | HIS | UN |
| 49 | 84% | 2000 | 0% | 2% | 0% | 1% | 94% | 1% | 0% |
| 50 | 85% | 20000 | 0% | 0% | 1% | 9% | 74% | 8% | 0% |
| 51 | 82% | 4000 | 0% | 0% | 0% | 1% | 99% | 0% | 0% |
| 52 | 83% | 2000 | 0% | 0% | 0% | 0% | 97% | 1% | 0% |
| 53 | 91% | 3000 | 0% | 0% | 0% | 2% | 95% | 1% | 0% |
| 54 | 78% | 12000 | 0% | 0% | 1% | 6% | 84% | 2% | 0% |
| 55 | 87% | 1000 | 0% | 0% | 0% | 0% | 99% | 1% | 0% |
| 56 | 83% | 2000 | 0% | 0% | 0% | 0% | 99% | 0% | 0% |
| 57 | 88% | 1000 | 0% | 0% | 0% | 0% | 97% | 1% | 0% |
| 58 | 80% | 3000 | 0% | 0% | 0% | 4% | 94% | 1% | 0% |
| 59 | 90% | 1000 | 0% | 0% | 1% | 1% | 97% | 1% | 0% |
| 60 | 89% | 2000 | 0% | 0% | 0% | 1% | 97% | 1% | 0% |
| 61 | 84% | 5000 | 0% | 0% | 1% | 3% | 92% | 2% | 0% |
| 62 | 81% | 3000 | 0% | 0% | 0% | 1% | 95% | 2% | 0% |
| 63 | 87% | 4000 | 0% | 0% | 0% | 3% | 93% | 1% | 0% |
| 64 | 85% | 2000 | 0% | 0% | 0% | 3% | 85% | 9% | 0% |
| 65 | 83% | 11000 | 0% | 0% | 1% | 2% | 93% | 2% | 0% |
| 67 | 86% | 4000 | 0% | 0% | 0% | 1% | 97% | 1% | 0% |
| 68 | 87% | 9000 | 0% | 0% | 1% | 6% | 76% | 10% | 0% |
| 69 | 78% | 25000 | 0% | 0% | 1% | 10% | 82% | 1% | 0% |
| 70 | 81% | 3000 | 0% | 0% | 0% | 1% | 98% | 1% | 0% |
| 71 | 82% | 3000 | 0% | 0% | 0% | 1% | 99% | 0% | 0% |
| 72 | 84% | 6000 | 0% | 0% | 0% | 2% | 97% | 0% | 0% |
| 73 | 87% | 8000 | 0% | 0% | 1% | 5% | 90% | 1% | 0% |
| 74 | 88% | 4000 | 0% | 0% | 0% | 1% | 96% | 1% | 0% |
| 75 | 84% | 4000 | 0% | 0% | 0% | 1% | 96% | 1% | 0% |
| 76 | 79% | 3000 | 0% | 0% | 0% | 8% | 88% | 0% | 0% |
| 77 | 82% | 9000 | 0% | 0% | 0% | 9% | 84% | 1% | 0% |
| 78 | 92% | 4000 | 0% | 0% | 0% | 4% | 93% | 1% | 0% |
| 79 | 88% | 4000 | 0% | 0% | 0% | 2% | 96% | 1% | 0% |
| 80 | 84% | 12000 | 0% | 0% | 3% | 4% | 86% | 2% | 0% |
| 81 | 86% | 2000 | 0% | 0% | 0% | 1% | 96% | 1% | 0% |
| 82 | 89% | 2000 | 0% | 0% | 0% | 1% | 95% | 2% | 0% |
| 83 | 85% | 4000 | 0% | 0% | 0% | 1% | 98% | 0% | 0% |
| 84 | 88% | 5000 | 0% | 0% | 1% | 7% | 85% | 1% | 0% |
| 85 | 88% | 1000 | 0% | 0% | 0% | 2% | 94% | 2% | 0% |
| 86 | 91% | 1000 | 0% | 0% | 0% | 0% | 97% | 1% | 0% |
| 87 | 84% | 1000 | 0% | 0% | 0% | 0% | 99% | 0% | 0% |

| Org ID | Graduation Rate[15] | Total Students | Expressed as a % | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | PI | IND | AS | AA | WH | HIS | UN |
| 88 | 82% | 4000 | 0% | 0% | 0% | 1% | 98% | 1% | 0% |
| 89 | 88% | 10000 | 0% | 0% | 1% | 2% | 94% | 1% | 0% |
| 90 | 84% | 12000 | 0% | 0% | 1% | 8% | 85% | 1% | 0% |
| 91 | 85% | 4000 | 0% | 0% | 0% | 2% | 96% | 1% | 0% |
| 92 | 87% | 1000 | 0% | 0% | 0% | 0% | 99% | 1% | 0% |
| 93 | 83% | 2000 | 0% | 0% | 0% | 1% | 97% | 1% | 0% |
| 94 | 78% | 1000 | 0% | 0% | 0% | 2% | 93% | 1% | 0% |
| 95 | 82% | 2000 | 0% | 0% | 0% | 1% | 99% | 0% | 0% |
| 96 | 90% | 1000 | 0% | 0% | 0% | 1% | 97% | 1% | 0% |
| 97 | 91% | 1000 | 0% | 0% | 0% | 0% | 98% | 0% | 0% |
| 98 | 80% | 4000 | 0% | 0% | 0% | 1% | 97% | 2% | 0% |
| 99 | 82% | 7000 | 0% | 0% | 0% | 1% | 98% | 0% | 0% |
| 100 | 86% | 1000 | 0% | 0% | 0% | 1% | 99% | 1% | 0% |
| 101 | 91% | 2000 | 0% | 0% | 1% | 1% | 97% | 1% | 0% |
| 102 | 89% | 1000 | 0% | 0% | 0% | 1% | 97% | 1% | 0% |
| 103 | 85% | 12000 | 0% | 0% | 1% | 2% | 93% | 1% | 0% |
| 104 | 35% | 0 | 0% | 0% | 0% | 2% | 97% | 0% | 0% |
| 105 | 83% | 4000 | 0% | 0% | 0% | 1% | 98% | 0% | 0% |
| 106 | 95% | 3000 | 100% | 0% | 0% | 0% | 0% | 0% | 0% |
| **$\bar{X}$** | **83%** | **8359** | **1%** | **1%** | **1%** | **9%** | **73%** | **12%** | **0%** |

**Table 16: Summary Statistics of Data Used to Determine EWS Similarity Scores[16]**

| Org ID | Population | Median Income | (%) Of Population | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Urban | UA | UC | Rural | Employed | Highschool Educated |
| 0 | 660,000 | 50,000 | 1 | 1 | 0 | 0 | 0.8 | 1 |
| 7 | 110,000 | 70,000 | 1 | 1 | 0 | 0 | 0.8 | 1 |
| 8 | 140,000 | 50,000 | 1 | 1 | 0 | 0 | 0.7 | 0.9 |
| 9 | 190,000 | 80,000 | 1 | 1 | 0 | 0 | 0.8 | 0.9 |
| 10 | 100,000 | 50,000 | 0.7 | 0.7 | 0 | 0.3 | 0.8 | 0.9 |
| 11 | 450,000 | 60,000 | 0.9 | 0.9 | 0 | 0.1 | 0.8 | 1 |
| 13 | 170,000 | 50,000 | 0.8 | 0.8 | 0 | 0.3 | 0.8 | 0.9 |
| 14 | 640,000 | 60,000 | 0.9 | 0.9 | 0 | 0.1 | 0.8 | 1 |
| 15 | 190,000 | 50,000 | 0.5 | 0.5 | 0.1 | 0.5 | 0.8 | 0.9 |
| 17 | 40,000 | 50,000 | 1 | 1 | 0 | 0 | 0.8 | 1 |
| 18 | 270,000 | 60,000 | 0.9 | 0.9 | 0 | 0.1 | 0.8 | 1 |
| 20 | 10,000 | 50,000 | 0 | 0 | 0 | 1 | 1 | 0.9 |
| 26 | 50,000 | 90,000 | 1 | 0 | 1 | 0 | 1 | 0.9 |
| 27 | 20,000 | 60,000 | 1 | 1 | 0 | 0 | 0.8 | 1 |
| 30 | 750,000 | 50,000 | 1 | 1 | 0 | 0 | 0.8 | 0.9 |
| 31 | 10,000 | 50,000 | 0.7 | 0 | 0.7 | 0.3 | 0.8 | 1 |
| 32 | 30,000 | 80,000 | 0.7 | 0.7 | 0 | 0.3 | 0.8 | 1 |
| 33 | 10,000 | 50,000 | 0.7 | 0 | 0.7 | 0.3 | 0.8 | 0.9 |
| 34 | 20,000 | 40,000 | 0.7 | 0 | 0.7 | 0.3 | 0.8 | 0.9 |
| 35 | 0 | 60,000 | 0 | 0 | 0 | 1 | 1 | 0.9 |
| 36 | 120,000 | 40,000 | 1 | 1 | 0 | 0.1 | 0.8 | 0.9 |
| 37 | 30,000 | 50,000 | 0.8 | 0.3 | 0.5 | 0.2 | 0.8 | 0.9 |
| 38 | 10,000 | 40,000 | 0.7 | 0 | 0.7 | 0.3 | 0.8 | 0.9 |
| 39 | 160,000 | 50,000 | 1 | 1 | 0 | 0 | 0.8 | 0.9 |
| 40 | 30,000 | 50,000 | 1 | 1 | 0 | 0 | 0.8 | 0.9 |
| 43 | 290,000 | 50,000 | 0.8 | 0.8 | 0 | 0.3 | 0.8 | 1 |
| 44 | 130,000 | 60,000 | 0.4 | 0.4 | 0 | 0.6 | 0.8 | 1 |
| 45 | 40,000 | 40,000 | 0.8 | 0.8 | 0 | 0.2 | 0.8 | 0.9 |
| 46 | 950,000 | 60,000 | 1 | 0.9 | 0 | 0.1 | 0.8 | 1 |
| 47 | 20,000 | 60,000 | 0.5 | 0.1 | 0.5 | 0.5 | 0.8 | 0.9 |
| 48 | 10,000 | 60,000 | 0.4 | 0 | 0.4 | 0.7 | 0.7 | 1 |
| 49 | 10,000 | 30,000 | 0.1 | 0 | 0.1 | 0.9 | 0.8 | 0.9 |
| 50 | 90,000 | 60,000 | 0.5 | 0.5 | 0 | 0.5 | 0.8 | 0.9 |
| 51 | 110,000 | 40,000 | 0.3 | 0.1 | 0.2 | 0.7 | 0.8 | 0.9 |
| 52 | 0 | 40,000 | 0 | 0 | 0 | 1 | 0.8 | 0.9 |
| 53 | 20,000 | 50,000 | 0.5 | 0.5 | 0 | 0.5 | 0.8 | 0.9 |
| 54 | 100,000 | 50,000 | 0.7 | 0.7 | 0 | 0.3 | 0.8 | 1 |
| 55 | 0 | 50,000 | 0 | 0 | 0 | 1 | 0.8 | 0.9 |
| 56 | 60,000 | 80,000 | 1 | 1 | 0 | 0.1 | 0.8 | 1 |
| 57 | 20,000 | 40,000 | 0.6 | 0 | 0.6 | 0.4 | 0.8 | 0.9 |

[16] Values rounded to nearest thousand to obfuscate school individual districts.

| Org ID | Population | Median Income | (%) Of Population | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Urban | UA | UC | Rural | Employed | Highschool Educated |
| 58 | 30,000 | 40,000 | 0.4 | 0.4 | 0 | 0.6 | 0.8 | 0.9 |
| 59 | 30,000 | 70,000 | 0.9 | 0.9 | 0 | 0.1 | 0.8 | 1 |
| 60 | 10,000 | 30,000 | 0.4 | 0 | 0.4 | 0.6 | 0.9 | 0.9 |
| 61 | 20,000 | 40,000 | 0.3 | 0 | 0.3 | 0.7 | 0.8 | 0.9 |
| 62 | 20,000 | 40,000 | 0 | 0 | 0 | 1 | 0.8 | 0.9 |
| 63 | 80,000 | 70,000 | 0.6 | 0.6 | 0 | 0.4 | 0.8 | 1 |
| 64 | 10,000 | 40,000 | 0.4 | 0 | 0.4 | 0.6 | 0.8 | 0.9 |
| 65 | 100,000 | 50,000 | 0.6 | 0.1 | 0.5 | 0.4 | 0.8 | 0.9 |
| 67 | 20,000 | 40,000 | 0.3 | 0 | 0.3 | 0.7 | 0.8 | 0.9 |
| 68 | 40,000 | 70,000 | 0.5 | 0 | 0.5 | 0.5 | 0.8 | 0.9 |
| 69 | 280,000 | 50,000 | 0.7 | 0.6 | 0.1 | 0.3 | 0.8 | 0.9 |
| 70 | 10,000 | 40,000 | 0.3 | 0 | 0.3 | 0.7 | 0.8 | 1 |
| 71 | 10,000 | 40,000 | 0 | 0 | 0 | 1 | 0.8 | 1 |
| 72 | 20,000 | 50,000 | 0.3 | 0 | 0.3 | 0.7 | 0.9 | 0.9 |
| 73 | 110,000 | 50,000 | 0.5 | 0.3 | 0.3 | 0.5 | 0.8 | 0.9 |
| 74 | 20,000 | 40,000 | 0.5 | 0.5 | 0 | 0.5 | 0.9 | 0.9 |
| 75 | 10,000 | 60,000 | 0.2 | 0 | 0.2 | 0.8 | 0.8 | 0.9 |
| 76 | 30,000 | 70,000 | 0.8 | 0.5 | 0.3 | 0.2 | 0.8 | 1 |
| 77 | 80,000 | 40,000 | 0.5 | 0 | 0.5 | 0.6 | 0.8 | 1 |
| 78 | 20,000 | 50,000 | 0.1 | 0 | 0.1 | 0.9 | 0.8 | 1 |
| 79 | 100,000 | 60,000 | 0.5 | 0.3 | 0.3 | 0.5 | 0.8 | 0.9 |
| 80 | 100,000 | 50,000 | 0.6 | 0.6 | 0 | 0.4 | 0.8 | 1 |
| 81 | - | - | - | - | - | - | - | - |
| 82 | 20,000 | 50,000 | 0 | 0 | 0 | 1 | 0.9 | 0.9 |
| 83 | 40,000 | 30,000 | 0.4 | 0.3 | 0.1 | 0.6 | 0.8 | 0.9 |
| 84 | 40,000 | 50,000 | 0.8 | 0.8 | 0 | 0.2 | 0.8 | 1 |
| 85 | 40,000 | 50,000 | 0.3 | 0.1 | 0.2 | 0.7 | 0.8 | 0.9 |
| 86 | 10,000 | 50,000 | 0.9 | 0 | 0.9 | 0.2 | 0.8 | 1 |
| 87 | 0 | 40,000 | 0 | 0 | 0 | 1 | 0.8 | 0.9 |
| 88 | 40,000 | 40,000 | 0.3 | 0 | 0.3 | 0.7 | 0.8 | 0.9 |
| 89 | 40,000 | 60,000 | 0.6 | 0.6 | 0 | 0.4 | 0.8 | 0.9 |
| 90 | 10,000 | 50,000 | 0.4 | 0.4 | 0 | 0.6 | 0.8 | 1 |
| 91 | 10,000 | 40,000 | 0.2 | 0 | 0.2 | 0.8 | 0.8 | 0.9 |
| 92 | 10,000 | 40,000 | 0 | 0 | 0 | 1 | 0.8 | 0.9 |
| 93 | 20,000 | 40,000 | 0.4 | 0 | 0.4 | 0.6 | 0.8 | 0.9 |
| 94 | 10,000 | 40,000 | 0.4 | 0 | 0.4 | 0.6 | 0.9 | 0.9 |
| 95 | 10,000 | 40,000 | 0.5 | 0 | 0.5 | 0.5 | 0.9 | 0.9 |
| 96 | - | - | - | - | - | - | - | - |
| 97 | - | - | - | - | - | - | - | - |
| 98 | 20,000 | 40,000 | 0.2 | 0 | 0.2 | 0.8 | 0.8 | 0.9 |
| 99 | 50,000 | 40,000 | 0.5 | 0.3 | 0.2 | 0.5 | 0.8 | 0.9 |
| 100 | 60,000 | 40,000 | 0.5 | 0.5 | 0 | 0.5 | 0.8 | 1 |
| 101 | - | 40,000 | 0.7 | 0 | 0.7 | 0.4 | 0.8 | 0.9 |

| Org ID | Population | Median Income | (%) Of Population | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Urban | UA | UC | Rural | Employed | Highschool Educated |
| 102 | - | - | - | - | - | - | - | - |
| 103 | 40,000 | 50,000 | 0.6 | 0.5 | 0.2 | 0.4 | 0.8 | 1 |
| 104 | - | - | - | - | - | - | - | - |
| 105 | 0 | 40,000 | 0 | 0 | 0 | 1 | 0.8 | 1 |
| 106 | 80,000 | 60,000 | 1 | 1 | 0 | 0 | 0.9 | 1 |
| **$\bar{X}$** | **90,000** | **50,000** | **0.6** | **0.4** | **0.2** | **0.5** | **0.8** | **0.9** |

# Appendix C: Data for Modeling

**Table 17: Percent of Missing Data Across Model Features**

| Feature Name | (%) Missing |
|---|---|
| avg_social_science_norm_interim_score | 99.96% |
| avg_science_norm_interim_score | 99.27% |
| avg_math_norm_interim_score | 97.40% |
| avg_reading_norm_interim_score | 97.29% |
| avg_social_science_norm_summative_score | 93.56% |
| avg_science_norm_summative_score | 83.79% |
| norm_grad_credit_ratio | 72.54% |
| stddev_social_science_grade | 70.29% |
| stddev_science_grade | 67.87% |
| grad_credit_ratio | 64.79% |
| stddev_math_grade | 64.14% |
| stddev_reading_grade | 54.37% |
| norm_avg_social_science_grade | 51.42% |
| norm_avg_science_grade | 48.52% |
| norm_avg_elective_grade | 47.04% |
| norm_avg_math_grade | 46.61% |
| norm_avg_reading_grade | 45.86% |
| avg_social_science_grade | 41.47% |
| stddev_elective_grade | 41.19% |
| norm_avg_all_course_grade | 40.93% |
| avg_science_grade | 38.68% |
| avg_math_grade | 36.30% |
| avg_reading_grade | 35.51% |
| sum_reading_grade | 35.51% |
| stddev_credits_earned | 35.26% |
| avg_credits_earned | 34.65% |
| stddev_core_course_grade | 34.51% |
| avg_core_course_grade | 33.90% |
| avg_math_norm_summative_score | 33.72% |
| avg_reading_norm_summative_score | 33.40% |
| avg_all_course_grade | 30.38% |
| pass_rate | 29.64% |
| gpa_for_grade_band | 27.32% |
| norm_age_for_class_number | 26.67% |
| attend_ratio | 2.85% |
| absent_in_first_30 | 2.82% |

| Feature Name | (%) Missing |
|---|---|
| absent_in_first_45 | 2.82% |
| absent_in_first_60 | 2.82% |
| absent_in_first_90 | 2.82% |
| attnd_100 | 2.82% |
| chronic_absent | 2.82% |
| sum_absent_ratio | 2.82% |
| sum_attendance_ratio | 2.82% |
| sum_tardy_ratio | 2.82% |
| total_absent_in_first_30 | 2.82% |
| total_absent_in_first_45 | 2.82% |
| total_absent_in_first_60 | 2.82% |
| total_absent_in_first_90 | 2.82% |
| algebra_passed | 0.00% |
| count_major | 0.00% |
| count_math | 0.00% |
| count_minor | 0.00% |
| count_reading | 0.00% |
| count_science | 0.00% |
| count_social_science | 0.00% |
| $\bar{X}$ | 33.33% |

# Appendix D: Pillar & Target Model Designation

**Table 18: Results of Data Based Pillar and Target Organization Assignment[17]**

| Org ID | Contains Missing Records | Total Records | Model Designation |
|---|---|---|---|
| 0 | No | 292,000 | Pillar |
| 7 | Yes | 13,000 | Target |
| 8 | Yes | 4,000 | Target |
| 9 | No | 34,000 | Pillar |
| 10 | Yes | 0 | Target |
| 11 | No | 93,000 | Pillar |
| 13 | No | 71,000 | Pillar |
| 14 | Yes | 37,000 | Target |
| 15 | Yes | 68,000 | Target |
| 17 | No | 17,000 | Target |
| 18 | No | 87,000 | Pillar |
| 20 | Yes | 0 | Target |
| 26 | Yes | 0 | Target |
| 27 | Yes | 5,000 | Target |
| 30 | No | 217,000 | Pillar |
| 31 | Yes | 1,000 | Target |
| 32 | No | 5,000 | Target |
| 33 | No | 5,000 | Target |
| 34 | No | 8,000 | Target |
| 35 | No | 2,000 | Target |
| 36 | No | 7,000 | Target |
| 37 | No | 4,000 | Target |
| 38 | No | 1,000 | Target |
| 39 | No | 10,000 | Target |
| 40 | No | 1,000 | Target |
| 43 | Yes | 32,000 | Target |
| 44 | No | 39,000 | Pillar |
| 45 | No | 6,000 | Target |
| 46 | Yes | 39,000 | Target |
| 47 | No | 0 | Target |
| 48 | No | 0 | Target |
| 49 | No | 10,000 | Target |
| 50 | No | 72,000 | Pillar |
| 51 | No | 20,000 | Pillar |
| 52 | No | 10,000 | Target |
| 53 | No | 16,000 | Target |

[17] Values rounded to nearest thousand to obfuscate school individual districts.

| Org ID | Contains Missing Records | Total Records | Model Designation |
|---|---|---|---|
| 54 | No | 51,000 | Pillar |
| 55 | No | 5,000 | Target |
| 56 | No | 8,000 | Target |
| 57 | No | 6,000 | Target |
| 58 | No | 29,000 | Pillar |
| 59 | No | 4,000 | Target |
| 60 | No | 9,000 | Target |
| 61 | No | 23,000 | Pillar |
| 62 | No | 16,000 | Target |
| 63 | No | 18,000 | Target |
| 64 | No | 10,000 | Target |
| 65 | No | 48,000 | Pillar |
| 67 | No | 22,000 | Pillar |
| 68 | No | 36,000 | Pillar |
| 69 | No | 131,000 | Pillar |
| 70 | No | 12,000 | Target |
| 71 | No | 12,000 | Target |
| 72 | No | 28,000 | Pillar |
| 73 | No | 36,000 | Pillar |
| 74 | No | 22,000 | Pillar |
| 75 | No | 18,000 | Target |
| 76 | No | 13,000 | Target |
| 77 | No | 42,000 | Pillar |
| 78 | No | 19,000 | Target |
| 79 | No | 12,000 | Target |
| 80 | No | 47,000 | Pillar |
| 81 | No | 8,000 | Target |
| 82 | No | 12,000 | Target |
| 83 | No | 19,000 | Target |
| 84 | Yes | 23,000 | Target |
| 85 | No | 5,000 | Target |
| 86 | No | 6,000 | Target |
| 87 | No | 6,000 | Target |
| 88 | No | 20,000 | Pillar |
| 89 | No | 44,000 | Pillar |
| 90 | No | 53,000 | Pillar |
| 91 | No | 19,000 | Target |
| 92 | No | 12,000 | Target |
| 93 | No | 10,000 | Target |
| 94 | No | 6,000 | Target |
| 95 | No | 10,000 | Target |
| 96 | No | 5,000 | Target |

| Org ID | Contains Missing Records | Total Records | Model Designation |
|---|---|---|---|
| 97 | No | 7,000 | Target |
| 98 | No | 16,000 | Target |
| 99 | No | 32,000 | Pillar |
| 100 | No | 7,000 | Target |
| 101 | No | 14,000 | Target |
| 102 | No | 4,000 | Target |
| 103 | No | 57,000 | Pillar |
| 104 | Yes | 0 | Target |
| 105 | No | 20,000 | Target |
| 106 | Yes | 3,000 | Target |

**Table 19: Results of All Pillar Models During Training**

| Org ID | Missing (%) | AUC | Model Designation |
|---|---|---|---|
| 0 | 0.272 | 0.878 | Pillar |
| 9 | 0.625 | 0.762 | Target |
| 11 | 0.394 | 0.889 | Pillar |
| 13 | 0.291 | 0.904 | Pillar |
| 18 | 0.343 | 0.851 | Pillar |
| 30 | 0.208 | 0.874 | Pillar |
| 44 | 0.499 | 0.821 | Target |
| 50 | 0.264 | 0.874 | Pillar |
| 51 | 0.234 | 0.859 | Pillar |
| 54 | 0.289 | 0.878 | Pillar |
| 58 | 0.375 | 0.865 | Pillar |
| 61 | 0.360 | 0.807 | Pillar |
| 65 | 0.224 | 0.895 | Pillar |
| 67 | 0.242 | 0.89 | Pillar |
| 68 | 0.271 | 0.864 | Pillar |
| 69 | 0.263 | 0.852 | Pillar |
| 72 | 0.240 | 0.856 | Pillar |
| 73 | 0.225 | 0.873 | Pillar |
| 74 | 0.227 | 0.878 | Pillar |
| 77 | 0.249 | 0.871 | Pillar |
| 80 | 0.281 | 0.891 | Pillar |
| 88 | 0.302 | 0.878 | Pillar |
| 89 | 0.227 | 0.885 | Pillar |
| 90 | 0.233 | 0.848 | Pillar |
| 99 | 0.339 | 0.848 | Pillar |
| 103 | 0.246 | 0.858 | Pillar |
| 108 | 0.396 | 0.847 | Pillar |
| **$\bar{X}$** | **0.301** | **0.862** | **-** |

# Appendix E: Aggregate Data Model Performance

**Table 20: Performance of Aggregate Data model Across All Districts in Data Test Set[18]**

| Org ID | Graduated | Dropout | Total Records | Aggregate Data model AUC |
|---|---|---|---|---|
| 0 | 80000 | 7000 | 87000 | 0.768 |
| 7 | 4000 | 0 | 4000 | 0.707 |
| 8 | 1000 | 0 | 1000 | 0.851 |
| 9 | 10000 | 1000 | 10000 | 0.622 |
| 10 | 0 | 0 | 0 | 0.5 |
| 11 | 25000 | 3000 | 28000 | 0.733 |
| 13 | 19000 | 3000 | 21000 | 0.768 |
| 14 | 10000 | 1000 | 11000 | 0.791 |
| 15 | 19000 | 2000 | 21000 | 0.696 |
| 17 | 5000 | 0 | 5000 | 0.747 |
| 18 | 22000 | 4000 | 26000 | 0.715 |
| 20 | 0 | 0 | 0 | 0.864 |
| 26 | 0 | 0 | 0 | 0 |
| 27 | 1000 | 0 | 1000 | 0.621 |
| 30 | 60000 | 5000 | 65000 | 0.749 |
| 31 | 0 | 0 | 0 | 0.685 |
| 32 | 1000 | 0 | 1000 | 0.846 |
| 33 | 1000 | 0 | 1000 | 0.706 |
| 34 | 2000 | 0 | 2000 | 0.705 |
| 35 | 1000 | 0 | 1000 | 0.771 |
| 36 | 2000 | 0 | 2000 | 0.691 |
| 37 | 1000 | 0 | 1000 | 0.566 |
| 38 | 0 | 0 | 0 | 0.501 |
| 39 | 1000 | 2000 | 3000 | 0.533 |
| 40 | 0 | 0 | 0 | 0.69 |
| 43 | 9000 | 1000 | 10000 | 0.752 |
| 44 | 11000 | 1000 | 12000 | 0.682 |
| 45 | 2000 | 0 | 2000 | 0.732 |
| 46 | 11000 | 1000 | 12000 | 0.804 |
| 47 | 0 | 0 | 0 | 0.875 |
| 48 | 0 | 0 | 0 | 0.74 |
| 49 | 3000 | 0 | 3000 | 0.811 |
| 50 | 20000 | 1000 | 22000 | 0.779 |

[18] Values rounded to nearest thousand to obfuscate school individual districts.

| Org ID | Graduated | Dropout | Total Records | Aggregate Data model AUC |
|---|---|---|---|---|
| 51 | 5000 | 1000 | 6000 | 0.75 |
| 52 | 3000 | 0 | 3000 | 0.76 |
| 53 | 4000 | 0 | 5000 | 0.771 |
| 54 | 14000 | 1000 | 15000 | 0.776 |
| 55 | 1000 | 0 | 1000 | 0.798 |
| 56 | 2000 | 0 | 2000 | 0.771 |
| 57 | 2000 | 0 | 2000 | 0.726 |
| 58 | 8000 | 1000 | 9000 | 0.751 |
| 59 | 1000 | 0 | 1000 | 0.785 |
| 60 | 3000 | 0 | 3000 | 0.759 |
| 61 | 7000 | 0 | 7000 | 0.728 |
| 62 | 4000 | 0 | 5000 | 0.766 |
| 63 | 5000 | 0 | 5000 | 0.767 |
| 64 | 3000 | 0 | 3000 | 0.785 |
| 65 | 13000 | 1000 | 14000 | 0.804 |
| 67 | 6000 | 0 | 7000 | 0.8 |
| 68 | 10000 | 1000 | 11000 | 0.781 |
| 69 | 35000 | 4000 | 39000 | 0.743 |
| 70 | 3000 | 0 | 4000 | 0.776 |
| 71 | 3000 | 0 | 4000 | 0.738 |
| 72 | 8000 | 0 | 9000 | 0.769 |
| 73 | 10000 | 1000 | 11000 | 0.781 |
| 74 | 6000 | 0 | 6000 | 0.776 |
| 75 | 5000 | 0 | 6000 | 0.773 |
| 76 | 3000 | 0 | 4000 | 0.756 |
| 77 | 11000 | 1000 | 13000 | 0.759 |
| 78 | 5000 | 0 | 6000 | 0.787 |
| 79 | 3000 | 0 | 4000 | 0.712 |
| 80 | 13000 | 1000 | 14000 | 0.786 |
| 81 | 2000 | 0 | 2000 | 0.736 |
| 82 | 3000 | 0 | 4000 | 0.772 |
| 83 | 5000 | 0 | 6000 | 0.796 |
| 84 | 7000 | 0 | 7000 | 0.79 |
| 85 | 1000 | 0 | 2000 | 0.825 |
| 86 | 2000 | 0 | 2000 | 0.856 |
| 87 | 2000 | 0 | 2000 | 0.743 |
| 88 | 6000 | 0 | 6000 | 0.772 |
| 89 | 13000 | 1000 | 13000 | 0.793 |
| 90 | 15000 | 1000 | 16000 | 0.748 |

| Org ID | Graduated | Dropout | Total Records | Aggregate Data model AUC |
|---|---|---|---|---|
| 91 | 5000 | 0 | 6000 | 0.768 |
| 92 | 3000 | 0 | 4000 | 0.723 |
| 93 | 3000 | 0 | 3000 | 0.718 |
| 94 | 2000 | 0 | 2000 | 0.776 |
| 95 | 3000 | 0 | 3000 | 0.778 |
| 96 | 1000 | 0 | 2000 | 0.731 |
| 97 | 2000 | 0 | 2000 | 0.683 |
| 98 | 4000 | 0 | 5000 | 0.746 |
| 99 | 9000 | 1000 | 10000 | 0.738 |
| 100 | 2000 | 0 | 2000 | 0.783 |
| 101 | 4000 | 0 | 4000 | 0.712 |
| 102 | 1000 | 0 | 1000 | 0.791 |
| 103 | 16000 | 1000 | 17000 | 0.772 |
| 104 | 0 | 0 | 0 | 0.75 |
| 105 | 5000 | 1000 | 6000 | 0.749 |
| 106 | 1000 | 0 | 1000 | 0.623 |
| $\bar{X}$ | 7000 | 1000 | 8000 | 0.746 |

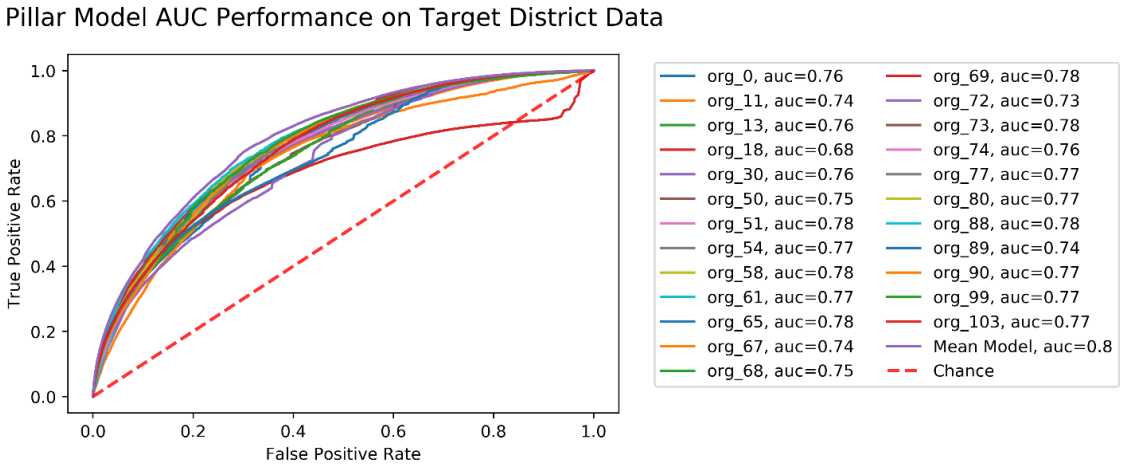



**Figure 39: AUC Curve Performance of Pillar Models and Mean model on Target District Data**

# Appendix F: District Similarity Ensemble Extrapolation

**Table 21: AUC Performance of DSEE on Target Districts[19]**

| | Count of Records | | |
|---|---|---|---|
| **Org ID** | **Graduates** | **Dropout** | **AUC** |
| 7 | 12,900 | 100 | 0.774 |
| 8 | 4,000 | 0 | 0.791 |
| 9 | 32,200 | 1,900 | 0.672 |
| 10 | 0 | 100 | 0.307 |
| 14 | 34,700 | 2,600 | 0.84 |
| 15 | 63,200 | 5,000 | 0.823 |
| 17 | 15,200 | 1,500 | 0.817 |
| 20 | 0 | 100 | 0.742 |
| 26 | 400 | 0 | 0.902 |
| 27 | 4,400 | 100 | 0.779 |
| 31 | 500 | 100 | 0.821 |
| 32 | 4,900 | 100 | 0.878 |
| 33 | 4,400 | 600 | 0.806 |
| 34 | 6,400 | 1,400 | 0.802 |
| 35 | 2,100 | 100 | 0.845 |
| 36 | 5,400 | 1,500 | 0.772 |
| 37 | 4,300 | 100 | 0.706 |
| 38 | 500 | 400 | 0.497 |
| 39 | 2,900 | 7,500 | 0.673 |
| 40 | 800 | 300 | 0.725 |
| 43 | 29,100 | 3,300 | 0.82 |
| 44 | 37,300 | 1,600 | 0.762 |
| 45 | 5,300 | 400 | 0.868 |
| 46 | 35,200 | 4,100 | 0.747 |
| 47 | 0 | 0 | 0.287 |
| 48 | 100 | 100 | 0.852 |
| 49 | 9,700 | 500 | 0.87 |
| 52 | 9,200 | 800 | 0.829 |
| 53 | 15,300 | 400 | 0.879 |
| 55 | 4,600 | 200 | 0.853 |
| 56 | 7,600 | 300 | 0.831 |
| 57 | 5,500 | 300 | 0.832 |
| 59 | 4,200 | 100 | 0.867 |
| 60 | 8,700 | 300 | 0.828 |
| 62 | 14,900 | 1,500 | 0.848 |
| 63 | 16,900 | 700 | 0.873 |
| 64 | 9,700 | 500 | 0.865 |
| 70 | 11,000 | 1,100 | 0.829 |

[19] Values rounded to nearest thousand to obfuscate school individual districts.

| | Count of Records | | |
|---|---|---|---|
| **Org ID** | **Graduates** | **Dropout** | **AUC** |
| 71 | 11,300 | 1,100 | 0.839 |
| 75 | 17,200 | 900 | 0.853 |
| 76 | 11,300 | 1,300 | 0.85 |
| 78 | 18,900 | 300 | 0.9 |
| 79 | 11,800 | 500 | 0.809 |
| 81 | 7,400 | 500 | 0.824 |
| 82 | 11,300 | 600 | 0.879 |
| 83 | 17,500 | 1,000 | 0.876 |
| 84 | 21,800 | 800 | 0.895 |
| 85 | 4,800 | 200 | 0.907 |
| 86 | 6,300 | 200 | 0.872 |
| 87 | 5,300 | 500 | 0.861 |
| 91 | 17,500 | 1,200 | 0.854 |
| 92 | 11,100 | 700 | 0.826 |
| 93 | 9,700 | 600 | 0.824 |
| 94 | 5,600 | 700 | 0.839 |
| 95 | 9,400 | 1,000 | 0.847 |
| 96 | 4,900 | 200 | 0.878 |
| 97 | 6,800 | 200 | 0.842 |
| 98 | 14,900 | 1,600 | 0.806 |
| 100 | 6,800 | 400 | 0.863 |
| 101 | 13,400 | 500 | 0.842 |
| 102 | 4,300 | 200 | 0.902 |
| 104 | 0 | 0 | 0.78 |
| 105 | 17,800 | 2,000 | 0.818 |
| 106 | 2,900 | 200 | 0.72 |
| **$\bar{X}$** | **10,800** | **900** | **0.805** |

# Appendix G: Chicago Model Performance

**Table 22: Chicago model Performance on Target Districts (9th Grade Records)[20]**

| Org ID | Graduated | Dropout | Total Records | Chicago model AUC |
|---|---|---|---|---|
| 7 | 1700 | 0 | 1700 | 0.537 |
| 8 | 0 | 0 | 0 | 0.611 |
| 9 | 3400 | 200 | 3600 | 0.498 |
| 10 | 0 | 0 | 0 | 0.5 |
| 14 | 4400 | 800 | 5100 | 0.652 |
| 15 | 8300 | 900 | 9200 | 0.742 |
| 17 | 1800 | 200 | 2000 | 0.523 |
| 20 | 0 | 0 | 0 | 0 |
| 26 | 100 | 0 | 100 | 0 |
| 27 | 1000 | 0 | 1100 | 0.517 |
| 31 | 100 | 0 | 100 | 0.748 |
| 32 | 700 | 0 | 700 | 0.9 |
| 33 | 500 | 100 | 600 | 0.746 |
| 34 | 800 | 300 | 1100 | 0.725 |
| 35 | 200 | 0 | 300 | 0.529 |
| 36 | 800 | 300 | 1100 | 0.704 |
| 37 | 400 | 0 | 400 | 0.862 |
| 38 | 0 | 0 | 100 | 0.499 |
| 39 | 500 | 1900 | 2400 | 0.579 |
| 40 | 100 | 100 | 200 | 0.5 |
| 43 | 4300 | 900 | 5200 | 0.674 |
| 44 | 4200 | 300 | 4500 | 0.675 |
| 45 | 800 | 100 | 800 | 0.7 |
| 46 | 3400 | 1000 | 4400 | 0.395 |
| 47 | 0 | 0 | 0 | 0 |
| 48 | 0 | 0 | 0 | 0.875 |
| 49 | 1100 | 100 | 1200 | 0.808 |
| 52 | 1100 | 200 | 1200 | 0.707 |
| 53 | 1800 | 100 | 1900 | 0.832 |
| 55 | 500 | 0 | 600 | 0.704 |
| 56 | 900 | 100 | 1000 | 0.655 |
| 57 | 600 | 100 | 700 | 0.764 |
| 59 | 500 | 0 | 500 | 0.835 |
| 60 | 1000 | 100 | 1000 | 0.732 |
| 62 | 1700 | 200 | 1900 | 0.685 |
| 63 | 1900 | 100 | 2100 | 0.823 |
| 64 | 1100 | 100 | 1200 | 0.819 |

[20] Values rounded to nearest hundred to obfuscate school individual districts.

| Org ID | Graduated | Dropout | Total Records | Chicago model AUC |
|---|---|---|---|---|
| 70 | 1300 | 200 | 1500 | 0.628 |
| 71 | 1400 | 300 | 1700 | 0.787 |
| 75 | 2100 | 200 | 2300 | 0.762 |
| 76 | 1100 | 200 | 1300 | 0.787 |
| 78 | 2200 | 100 | 2200 | 0.78 |
| 79 | 1100 | 100 | 1200 | 0.648 |
| 81 | 900 | 100 | 1000 | 0.744 |
| 82 | 1400 | 100 | 1500 | 0.758 |
| 83 | 2000 | 200 | 2200 | 0.591 |
| 84 | 2600 | 200 | 2800 | 0.745 |
| 85 | 500 | 0 | 600 | 0.81 |
| 86 | 800 | 0 | 800 | 0.746 |
| 87 | 600 | 100 | 700 | 0.7 |
| 91 | 2000 | 200 | 2200 | 0.768 |
| 92 | 1300 | 100 | 1500 | 0.72 |
| 93 | 1100 | 100 | 1200 | 0.789 |
| 94 | 700 | 100 | 800 | 0.728 |
| 95 | 1100 | 200 | 1300 | 0.78 |
| 96 | 600 | 0 | 600 | 0.673 |
| 97 | 800 | 0 | 800 | 0.512 |
| 98 | 1900 | 400 | 2200 | 0.724 |
| 100 | 700 | 100 | 800 | 0.621 |
| 101 | 1500 | 100 | 1600 | 0.706 |
| 102 | 500 | 0 | 500 | 0.846 |
| 104 | 0 | 0 | 0 | 0 |
| 105 | 2000 | 300 | 2400 | 0.655 |
| 106 | 700 | 0 | 700 | 0.631 |
| **$\bar{X}$** | **1200** | **100** | **1300** | **0.682** |

# Appendix H: Balfanz Model Performance

**Table 23: Balfanz model Performance on Target Districts (Grades 6th – 12th)[21]**

| Org ID | Graduated | Dropout | Total Records | Balfanz model AUC |
|---|---|---|---|---|
| 7 | 500 | 0 | 500 | 0.644 |
| 9 | 2600 | 200 | 2800 | 0.614 |
| 10 | 0 | 0 | 0 | 0.5 |
| 14 | 0 | 100 | 100 | 0 |
| 15 | 3900 | 400 | 4400 | 0.478 |
| 17 | 1200 | 100 | 1300 | 0.557 |
| 20 | 0 | 0 | 0 | 0 |
| 32 | 200 | 0 | 200 | 0.884 |
| 33 | 300 | 0 | 300 | 0.641 |
| 34 | 500 | 100 | 600 | 0.583 |
| 35 | 100 | 0 | 100 | 0.664 |
| 36 | 300 | 100 | 400 | 0.551 |
| 37 | 400 | 0 | 400 | 0.622 |
| 38 | 0 | 0 | 100 | 0.494 |
| 39 | 100 | 600 | 700 | 0.547 |
| 40 | 0 | 0 | 0 | 0.567 |
| 43 | 1500 | 300 | 1700 | 0.547 |
| 44 | 3200 | 100 | 3300 | 0.463 |
| 45 | 400 | 0 | 400 | 0.572 |
| 46 | 0 | 200 | 200 | 0.533 |
| 47 | 0 | 0 | 0 | 0.5 |
| 48 | 0 | 0 | 0 | 0 |
| 49 | 700 | 0 | 800 | 0.622 |
| 52 | 700 | 100 | 700 | 0.724 |
| 53 | 1100 | 0 | 1100 | 0.617 |
| 55 | 300 | 0 | 400 | 0.727 |
| 56 | 600 | 0 | 600 | 0.742 |
| 57 | 400 | 0 | 400 | 0.794 |
| 59 | 300 | 0 | 300 | 0.771 |
| 60 | 700 | 0 | 700 | 0.67 |
| 62 | 1100 | 100 | 1200 | 0.642 |
| 63 | 1200 | 100 | 1200 | 0.762 |
| 64 | 700 | 0 | 800 | 0.764 |
| 70 | 800 | 100 | 900 | 0.665 |

[21] Values rounded to nearest hundred to obfuscate school individual districts.

| Org ID | Graduated | Dropout | Total Records | Balfanz model AUC |
|---|---|---|---|---|
| 71 | 800 | 100 | 1000 | 0.638 |
| 75 | 1400 | 100 | 1400 | 0.546 |
| 76 | 900 | 100 | 1000 | 0.682 |
| 78 | 1400 | 0 | 1400 | 0.672 |
| 79 | 1100 | 0 | 1200 | 0.627 |
| 81 | 500 | 0 | 600 | 0.69 |
| 82 | 800 | 0 | 900 | 0.701 |
| 83 | 1300 | 100 | 1400 | 0.63 |
| 84 | 1500 | 0 | 1500 | 0.803 |
| 85 | 400 | 0 | 400 | 0.647 |
| 86 | 500 | 0 | 500 | 0.698 |
| 87 | 400 | 0 | 400 | 0.58 |
| 91 | 1300 | 100 | 1400 | 0.677 |
| 92 | 800 | 0 | 800 | 0.66 |
| 93 | 700 | 100 | 800 | 0.759 |
| 94 | 400 | 100 | 500 | 0.721 |
| 95 | 700 | 100 | 800 | 0.754 |
| 96 | 300 | 0 | 400 | 0.47 |
| 97 | 500 | 0 | 500 | 0.552 |
| 98 | 1200 | 100 | 1300 | 0.654 |
| 100 | 500 | 0 | 500 | 0.762 |
| 101 | 1000 | 0 | 1000 | 0.658 |
| 102 | 300 | 0 | 300 | 0.674 |
| 104 | 0 | 0 | 0 | 0 |
| 105 | 1400 | 200 | 1600 | 0.627 |
| 106 | 100 | 0 | 100 | 0.5 |
| **$\bar{X}$** | **700** | **100** | **700** | **0.657** |

# Appendix I: Knowles Model Performance

**Table 24: Knowles model Algorithm Search Results**

| Method | AUC | AUC SD |
|---|---|---|
| gbm | 88.213 | 0.0043 |
| rf | 87.960 | 0.0054 |
| earth | 87.132 | 0.0036 |
| glmnet | 86.912 | 0.0064 |
| multinom | 86.907 | 0.0040 |
| glm | 86.907 | 0.0050 |
| treebag | 86.856 | 0.0042 |
| glmboost | 86.645 | 0.0029 |
| lda | 86.628 | 0.0014 |
| lda2 | 86.628 | 0.0044 |
| sda | 86.624 | 0.0012 |
| nnet | 86.548 | 0.0044 |
| ctree | 85.730 | 0.0009 |
| ctree2 | 85.152 | 0.0033 |
| pda2 | 84.203 | 0.0062 |
| knn | 83.429 | 0.0037 |
| LogitBoost | 83.022 | 0.0006 |
| rpart | 79.301 | 0.0106 |

**Table 25: Knowles model Performance on Target Districts (Grades 6th – 12th)[22]**

| Org ID | Graduated | Dropout | Total Records | Knowles model AUC |
|---|---|---|---|---|
| 7 | 11400 | 100 | 11500 | 0.823 |
| 8 | 3600 | 0 | 3600 | 0.943 |
| 9 | 13200 | 500 | 13700 | 0.79 |
| 10 | 0 | 0 | 0 | 0.57 |
| 14 | 31100 | 2200 | 33400 | 0.927 |
| 15 | 51100 | 3200 | 54300 | 0.866 |
| 17 | 10800 | 1000 | 11800 | 0.849 |
| 20 | 0 | 100 | 100 | 0.857 |
| 26 | 400 | 0 | 400 | 0.805 |
| 27 | 300 | 0 | 300 | 0.996 |
| 31 | 500 | 0 | 500 | 0.839 |
| 32 | 4300 | 100 | 4400 | 0.938 |
| 33 | 2200 | 300 | 2400 | 0.921 |
| 34 | 4900 | 1100 | 6000 | 0.835 |
| 35 | 1400 | 0 | 1500 | 0.856 |
| 36 | 4700 | 1200 | 5900 | 0.785 |
| 38 | 300 | 300 | 500 | 0.563 |
| 39 | 600 | 1500 | 2100 | 0.79 |
| 40 | 700 | 200 | 1000 | 0.782 |
| 43 | 25600 | 2400 | 28000 | 0.884 |
| 44 | 16300 | 700 | 17000 | 0.905 |
| 45 | 4400 | 300 | 4700 | 0.88 |
| 46 | 31800 | 3400 | 35200 | 0.938 |
| 47 | 0 | 0 | 0 | 0.771 |
| 48 | 100 | 100 | 100 | 0.886 |
| 49 | 7200 | 400 | 7700 | 0.891 |
| 52 | 6500 | 600 | 7100 | 0.867 |
| 53 | 10800 | 300 | 11100 | 0.914 |
| 55 | 3300 | 200 | 3500 | 0.877 |
| 56 | 5500 | 300 | 5800 | 0.851 |
| 57 | 4000 | 200 | 4200 | 0.871 |
| 59 | 3100 | 100 | 3100 | 0.906 |
| 60 | 6100 | 300 | 6400 | 0.853 |
| 62 | 10700 | 1100 | 11800 | 0.875 |
| 63 | 12200 | 500 | 12700 | 0.91 |
| 64 | 6900 | 400 | 7300 | 0.891 |
| 70 | 8000 | 900 | 8900 | 0.875 |
| 71 | 8100 | 900 | 9000 | 0.857 |
| 75 | 12100 | 700 | 12800 | 0.894 |
| 76 | 7800 | 900 | 8700 | 0.882 |
| 78 | 13600 | 300 | 13900 | 0.92 |
| 79 | 7600 | 300 | 8000 | 0.847 |

[22] Values rounded to nearest hundred to obfuscate school individual districts.

| Org ID | Graduated | Dropout | Total Records | Knowles model AUC |
|---|---|---|---|---|
| 81 | 5300 | 400 | 5800 | 0.85 |
| 82 | 8200 | 500 | 8600 | 0.881 |
| 83 | 12400 | 800 | 13300 | 0.898 |
| 84 | 15900 | 700 | 16600 | 0.91 |
| 85 | 3300 | 200 | 3500 | 0.919 |
| 86 | 4500 | 100 | 4700 | 0.919 |
| 87 | 3700 | 300 | 4000 | 0.886 |
| 91 | 12300 | 900 | 13200 | 0.886 |
| 92 | 8100 | 500 | 8700 | 0.861 |
| 93 | 6900 | 500 | 7300 | 0.856 |
| 94 | 4100 | 500 | 4700 | 0.858 |
| 95 | 6600 | 800 | 7400 | 0.876 |
| 96 | 3600 | 200 | 3800 | 0.905 |
| 97 | 4900 | 200 | 5000 | 0.873 |
| 98 | 10900 | 1300 | 12200 | 0.847 |
| 100 | 4800 | 300 | 5100 | 0.89 |
| 101 | 9400 | 300 | 9800 | 0.876 |
| 102 | 3100 | 100 | 3300 | 0.912 |
| 104 | 0 | 0 | 0 | 0.905 |
| 105 | 12400 | 1400 | 13800 | 0.852 |
| 106 | 2600 | 200 | 2800 | 0.716 |
| **$\bar{X}$** | **6900** | **500** | **7300** | **0.874** |

# Appendix J: AUC Performance at 95% CI by Populations

**Table 26: Calculated Mean, Standard Deviation, Standard Error, and 95% Confidence Interval of EWS AUC Performance**

| Grade Prediction | EWS Model | Mean | SD | SE | 95% CI |
|---|---|---|---|---|---|
| 1st - 12th | Aggregate Data model | 0.739 | 0.078 | 0.010 | 0.020 |
| | DSEE model | 0.805 | 0.114 | 0.014 | 0.028 |
| | Mean model | 0.805 | 0.111 | 0.014 | 0.028 |
| 6th | Aggregate Data model | 0.708 | 0.102 | 0.014 | 0.029 |
| | Balfanz model | 0.640 | 0.096 | 0.013 | 0.026 |
| | DSEE model | 0.747 | 0.144 | 0.019 | 0.039 |
| | Knowles model | 0.757 | 0.133 | 0.018 | 0.036 |
| | Mean model | 0.747 | 0.144 | 0.019 | 0.039 |
| 9th | Aggregate Data model | 0.774 | 0.083 | 0.011 | 0.022 |
| | Chicago model | 0.683 | 0.141 | 0.018 | 0.037 |
| | DSEE model | 0.846 | 0.089 | 0.012 | 0.023 |
| | Knowles model | 0.863 | 0.073 | 0.010 | 0.019 |
| | Mean model | 0.846 | 0.089 | 0.012 | 0.023 |