Journal of Liberal Arts and Humanities (JLAH)
Issue: Vol. 2; No. 9; September 2021 pp. 1-10
ISSN 2690-070X (Print) 2690-0718 (Online)
Website: www.jlahnet.com
E-mail: editor@jlahnet.com
Doi: 10.48150/jlah.v2no9.2021.a1

PREDICTION MODELS FOR IDENTIFYING STUDENTS AT RISK FOR GRADUATING LATE OR DROPPING OUT

Dana H. Morris, Ed.D.
Director of Federal Programs
Houston County School District
1100 Main St. Perry, GA 31069
E-mail: dana.h.morris@hcbe.net
Tel: 478-988-6200

Lantry L. Brockmeier*, Ph.D.
Leadership, Technology, and Workforce Development
Valdosta State University
1500 N. Patterson St.Valdosta, GA 31698
E-mail: llbrockmeier@valdosta.edu
Tel: 229-333-5633

James L. Pate, Ph.D.
Leadership, Technology, and Workforce Development
Valdosta State University
1500 N. Patterson St.Valdosta, GA 31698
E-mail: jlpate@valdosta.edu
Tel: 229-333-5633

Gerald R. Siegrist, Ed.D.
Leadership, Technology, and Workforce Development
Valdosta State University
1500 N. Patterson St.Valdosta, GA 31698
E-mail: rgreen@valdosta.edu
Tel: 229-333-5633

Abstract

The problem of high school dropouts has been identified for decades but utilizing readily obtainable student data and data mining can aid school leaders to more accurately detect students likely to drop out. This early warning information can be used by educators to help identify potential high school dropouts and students who will not graduate on time. The purpose of this nonexperimental correlational study was to use longitudinal data from a mid-sized school district to create a dropout early warning system to predict students who are at risk for not graduating on time. Three statistical models (logistic regression, linear discriminant analysis, and quadratic discriminant analysis) were utilized to identify the most accurate indicators to predict both sixth-grade and ninth-grade students who were at risk for not completing high school on time. Statistical models with the lowest false-positive rate and high accuracy were identified for both sixth grade and ninth grade.

Keywords: student retention, high school dropouts, on-time graduation, classification models

Introduction

With improvements in student information systems over the past decade, there has been a growth of early warning systems that address the dropout crisis more effectively (Jobs for the Future, 2014). An early warning system involves analysis of school data to monitor students at risk for falling offthe path to graduation and implementing interventions to help students graduate (Davis, Herzog, & Legters, 2013). Statistical evidence is available to show that dropouts can be identified long before they fail to graduate(Allensworth & Easton, 2007; Balfanz, Herzog, & Mac Iver, 2007; Neild, 2009). A few key data points allow schools and districts to identify those students who are most likely to drop out of high school. As of 2014, the National Governors Association identified 16 states and a growing number of school districts that use some form of an early warning system to predict students, as early as elementary school, who are likely to struggle to graduate from high school within four years (Jobs for the Future, 2014). Most early warning systems include measurements of attendance, behavior, and course performance, also known as the "ABCs," because together they are strong predictors of high school graduation at all levels (U.S. Department of Education, 2016). Identifying the best combination of early warning system indicators for each user's state, district, school, or grade is key to accurately detecting potential dropouts.

Some benefits of an early warning system include its use of readily available student data, its ability to cull through large amounts of data and focus only on the most important indicators, its use in real-time, and its ability to monitor throughout the school year(Davis et al., 2013). Another advantage of an early warning system is it accurately predicts a high percentage of students who will not graduate from high school within four years based only on academic data instead of background characteristics (Brundage, 2014). School personnel cannot impact all the students' background characteristics, but they can intervene with students' school academic performance. Interventions focused simply on attendance and behavior have been effective(Balfanz et al., 2007).

When developing and using an early warning system, stakeholder input is essential to improving and refining it. The system should be regularly revisited and risk models revaluated. Before the development of an early warning system, quality research should be performed to identify the best indicators or combinations of indicators. Proper training of the users and clear visuals help ensure effective use of the system. Although there is valuable information about risk factors for dropping out obtained from studies of large districts,such as Chicago and Philadelphia, school systems should look at their longitudinal data to identify factors most strongly associated with their student dropouts (Heppen & Therriault, 2008). To identify who is at risk of dropping out, districts can save time and money by investigating longitudinal data of past cohorts to predict what will happen to students in future cohorts (Jerald, 2006). This personal data can help a system more accurately predict students who are most at risk for dropping out.

The problem of high school dropouts has been studied for decades but, now more than ever, better research and data are available for school leaders to learn both who will likely drop out and what interventions to implement. There are several factors for why a student chooses to drop out, and there are variables along the way that can identify who is likely to dropout(Rumberger & Lim, 2008). Early warning systems could aid dropout prevention by testing local indicators to identify accurately students who are likely to drop out and to aid schools in identifying those who need interventions (Pinkus, 2008). In the past, dropout prevention efforts generated poor results (Jerald, 2006). Accurate and early identification of students at risk for not graduating can lead to appropriate and timely interventions that increase the likelihood of students completing high school (McKee & Caldarella, 2016). An early warning and multi-tiered response system are essential to ending the dropout crisis in schools, districts, and the nation.

High school dropouts not only negatively impact themselves; theyalso negatively impact their families and society. Society needs an educated and trained workforce to compete in the world marketplace (Neild, 2009). Members of society without even a high school diploma can become a burden because higher rates of unemployment and higher crime rates are associatedwith high school dropouts(Alliance for Excellent Education, 2011). During the third quarter of 2019, all full-time workers aged 25 and older had a median weekly income of $975 while full-time workers without a high school diploma earned 62% of that amount. A high school graduate earned 77% of that amount and employees with a bachelor's degree earned 131% of the $975 median weekly income (Bureau of Labor Statistics, 2019).Dropouts also experienced more unemployment, utilized more government aid, or spent more time incarcerated than their peers who graduated high school (Zvoch, 2006).

Dropouts also report more health problems, and on average, die at a younger age than those who graduate (Laird, Kienzl, DeBell, & Chapman, 2007). Students who drop out of high school simply limit their life opportunities and personal wellbeing(McKee & Caldarella, 2016).

The vast evidence of how dropping out of high school significantly impedes a person's quality of life cannot be ignored. Because of the detrimental consequences caused by dropping out of high school, schools must take steps to ensure all students graduate. Not only do students' futures depend on it, but society's future as well. The dramatic economic benefit of improving the outcomes of academically at-risk students should be a wake-up call to the nation. With global competition and the grim outlook for dropouts, high schools must keep students in school and prepare them for life after high school, including college, the workforce, or possible military service(Amos, 2008). The social and economic contributions of these young people cannot be underestimated.

Purpose of the Study

The purpose of this study is multifaceted.The primary purpose of this nonexperimental, correlational study was to generate a dropout early warning system to predict both sixth-grade middleschoolandninth-grade highschool students who are at risk for not graduating on timeand to identify the most accurate indicators at each grade level. A secondary purpose was to identify the most accurate statistical model from the selected statistical models to generate high levels of true classification and low levels of false classification.

Methodology

The methodology section is divided into three subsections. First, we will discuss the research design. This will be followed by a discussion of the participants. Finally, we will discuss the data analysis.

Research Design

A nonexperimental, ex post facto, multivariate correlational research design was employed. The school district had previously collected the data used in this study. Three statistical models (i.e., logistic regression, linear discriminate analysis, and quadratic discriminate analysis) were used to predict individual students' risk of not graduating on time. The dependent variable was whether the student graduated high school on time or the student did not graduate high school on time.

The ninth-grade independent variables were: attending school less than 90% of the time, earning sufficient credits to move to the tenth grade (5 Carnegie units), number of days suspended out of school, number of school moves, End of Course Test (EOCT) standardized reading and math scores (ranges from 200 to 600), failing no more than one semester of a core course, school minority percentages (from 1 to 100), school poverty percentages (from 1 to 100), ELL status (yes or no), SWD status (yes or no), free/reduced meal status (yes or no), race, and gender (Allensworth & Easton, 2005; DePaoli et al., 2015; Kemple et al., 2013; Lee et al., 2011; Mac Iver & Mac Iver, 2010; Mac Iver & Messel, 2012; 2013; Neild, 2009; Zvoch, 2006).

The sixth-grade independent variables were: failing English with an average below a 70, failing math with an average below a 70, attending school less than 80% of the time, receiving out-of-school suspension (yes or no), number of school moves, Criterion-Reference Competency Test (CRCT) standardized reading and math scores (ranges from 650 to 900), school minority percentages (from 1 and 100), school poverty percentages (from 1 and 100), ELL status (yes or no), SWD status (yes or no), free or reduced meal status (yes or no), race, and gender (Balfanz, 2009; Balfanz et al., 2007; Jerald, 2006; Mac Iver, 2010; Rumberger, 2004; Silver et al., 2008).

Participants

Data came from six middle schools and six high schools in a mid-sized Georgia school district. The race or ethnicity of the student population consisted of 73% Black students, 18% White students, 5% Hispanic students, 2% Asian students, and 2% multiracial students. Each school had a free and reduced lunch percentage close to 99%. Of the student population, 2% are English language learners (ELL) and 10% are students with disabilities (SWD). The total number of students in each cohort was approximately 1,000 students per year. This study included all students who entered sixth grade in the 2010 and 2011 school years and ninth grade in the 2013 and 2014 school years. Their on-time graduation years were 2017 and 2018. The data set was split so that the 2017 cohort data were used for training and the 2018 cohort data were used to evaluate the models.

Data Collection

Once the IRB approval was received, the school district's data analyst provided the longitudinal data for the study. Before providing the data, the data analyst assigned random identification numbers to each student, so all students remained anonymous. Except for school minority, school poverty, and free or reduced meal status, the district's database contained all the data. The Georgia Department of Education College and Career Performance Index contained the school minority and school poverty data, while the Georgia School Nutrition database contained the free or reduced meal status based on family income.

Data Analysis

Initially, descriptive statistics were generated to examine measures of central tendency and variability of the variables. A few variables were found to have a small amount of missing data. These missing data were imputed using a bagged tree model found in R's caret package. The continuous level variables were checked for outliers. Correlations among the independent variables were generated examined along with the correlations between the independent variables and the dependent variable.

To help with the class imbalance of the dependent variable (students graduating ontime and students not graduating ontime), upsampling of the minority class and downsampling of the majority class was performed before training the model. The result for both upsampling and downsampling is the same number of observations from the minority and majority classes in the training data set analysis. All three statistical analyses were trained with upsampled and downsampled data.

The first statistical analysis used was logisticregression. Assumptions for logistic regression include observations must be independent, and independent variables must be linearly related to the logit of the dependent variable (Leech, Barrett, & Morgan, 2008). The dependent variable must be dichotomous,and the independent variables must be continuous or categorical. The second and third statistical procedures used were the linear discriminate analysis (LDA) and quadratic discriminate analysis (QDA), respectively. Linear discriminant analysis and quadratic discriminant analysis help to find the boundaries around the classification choices(James, Witen, Hastie, & Tibshirani, 2013). For the multiple variables, the models estimate the mean and variance from the data for each class. Both the LDA and QDA algorithms make predictions by estimating the probability that a new set of inputs belongs to a particular class or group. The class with the highest probability is the output class, and therefore, the prediction(James et al., 2013). LDA and QDA have assumptions that are often more restrictive than logistic regression. Both LDA and QDA assume that the predictor variables are drawn from a normal distribution. LDA assumes equality of covariances among the predictor variables X across all levels of Y,butQDA does not. Both LDA and QDA require the number of predictor variables to be less than the sample size("Linear & quadratic discriminant analysis," n.d.).

Results

The results section consists of two subsections. First, the: ninth-grade results by statistical procedureand accuracy of the models are presented. Second, thesixth-grade results by statistical procedure and accuracy of the models are presented.

Ninth Grade Results

In this study, logistic regression was one of three statistical models used to predict whether or not a student would graduate within four years based on known student ninth-grade data. For the upsampled logistic regression prediction, the accuracy was 0.89, true-positive rate of 0.97, and false-positive rate of 0.66. The downsampled version had an accuracy of 0.88, true-positive rate of 0.94, and false-positive rate of 0.54. Another measure of the quality of the model is the Akaike Information Criterion (AIC) value. If two similar models are compared, then the model with the lower Akaike Information Criterion (AIC) value is superior. The downsampled model was identified as the preferred model with an AIC level of 272.36 compared to the upsampled model with an AIC level of 1545.20.

The Nagelkerkepseudo R-squared value is helpful when it is compared to another pseudo R-squared of the same type and predicting the same outcome. A higher pseudo R-squared indicates which model does a better job of predicting the outcome (UCLA: Statistical Consulting Group., n.d.). The upsampled logistic regression model had an Nagelkerkepseudo R-squared value of 0.492 and can be compared to the downsampled logistic regression model's pseudo R-squared value of 0.518. Thedownsampledlogistic regression model is again identified as the preferred model.

The upsampled logistic regression model showed that by earning sufficient credits to advance to the tenth grade, a student increases their odds of graduating by a factor of 5.2, given all other variables are unchanged. If a student received an out-of-school suspension, the odds of a student graduating decreased by 57.9% (0.421 – 1 = -0.579), keeping other variables constant.The downsampled logistic regression model showed that by earning sufficient credits to advance to the tenth grade, a student increases their odds of graduating by a factor of 4.3, given that all other variables are unchanged. If a student had multiple moves in the ninth grade, the odds of a student graduating decreased by 28.9% (0.711 – 1 = -.289), keeping all other variables constant.

The upsampled LDA model with an accuracy of 89.0%, true-positive rate of 0.97, and false-positive rate of 0.66 had results that closely aligned with those of the upsampled logistic regression models.Earning sufficient credits to advance to the tenth grade was still the most influential predictor with a coefficient of 0.985, but now gender had the second-highest weighted coefficient (0.568). Another strong predictor in the upsampled LDA model was fail9 (coefficient of 0.499), suggesting that the students who fail less than two classes are much more likely to graduate. The downsampled LDA model had an accuracy score of 88.7%, true-positive rate of 0.98, and false-positive rate of 0.75, which was slightly less than the upsampled LDA model. A strong predictor in the downsampled LDA model was earning sufficient credits to advance to tenth grade, which had a coefficients of linear discriminants value of 1.064. The group means showed that 90.6% of the graduates earned sufficient credits to advance to the 10th grade, while only 46.4% of the nongraduates did.

The upsampled QDA group means showed that 93.9% of the graduates earned sufficient credits to advance to the 10th grade, while only 48.3% of the nongraduates did. Similarly, only 16.2% of graduates received out-of-school suspension while 53.4% of nongraduates received out-of-school suspension, and 88.1% of graduates attended at least 90% of school days while only 49.7% nongraduates attended at least 90% of school days. The downsampled QDA group means show the drastic difference between the means of graduates and nongraduates for the variables of earning sufficient credits to advance to tenth grade, attending at least 90% of school days, and if suspended from school just as it did for the upsampled QDA model.

Overall, variables consistently identified in a majority of the ninth-grade models as able to predict students who would not complete high school within four years were: (a) if a student did not receive enough credits to advance to the tenth grade, (b) if a student did not attend school at least 90% of the time, (c) if a student was suspended from school, (d) if a student had multiple school moves in the ninth grade, and (e) male gender.

To determine the answer of which statistical model was most accurate at predicting future dropouts or late graduates utilizing ninth-grade variables, both a ROC curve and confusion matrix were used to identify the most accurate statistical model. Of the three statistical models, the upsampled and downsampled logistic regression analyses had the highest area under the curve values at 0.842 each (see Table 1). Linear discriminant analysis for both the upsampled and downsampleddata sets hadan area under the curve values of 0.841. For the quadratic discriminant analysis,the area under the curve was 0.812 for both the upsampled data set and the downsampleddata set (see Figure 1).

Table 1Analysis Results Based on Ninth Grade Variables for all Data Types and Statistical Models Using Test Data

Statistical Model Used for Ninth Grade Data	AUROC	Accuracy	Kappa	Sensi-tivity	Speci-ficity	F1	Balanced Accuracy
Upsampled LR	0.842	.89	.39	.97	.34	.94	.66
Downsampled LR	0.842	.88	.43	.94	.46	.93	.70
Upsampled LDA	0.841	.89	.39	.97	.34	.94	.66
Downsampled LDA	0.841	.89	.31	.98	.25	.94	.62
Upsampled QDA	0.812	.87	.34	.94	.37	.92	.66
Downsampled QDA	0.812	.85	.41	.90	.57	.91	.73

Note. AUROC is Area Under the ROC Curve.

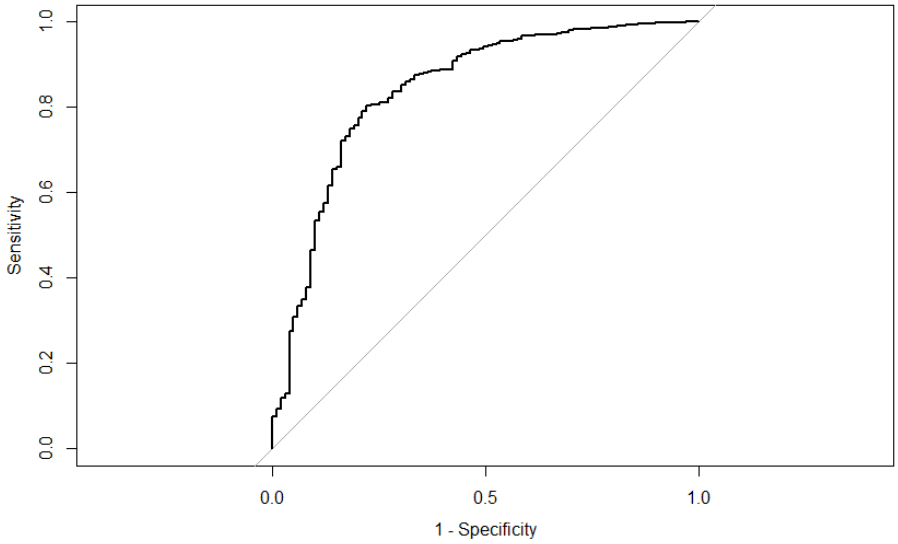

Figure 1 ROC curve results based on ninth-grade variables used to predict graduation utilizing downsampled quadratic discriminant analysis.

Because a main goal is to identify nongraduates early in their schooling, then it is wiser to choose a model with the lowest false-positive rate even if the true-positive rate suffers slightly from that choice. It is a better option to offer help to more students than need it than to fail to identify students who need help. This judgment leads to the conclusion that the downsampled QDA model is the best model for minimizing false-positive observations, as its false-positive rate was the lowest at 0.43, and its true-negative was the highest at 0.57. The downsampled QDA model also had the second-highest kappa value and the highest balanced accuracy.

Sixth Grade Results

In this study, logistic regression was used to predict whether or not a student would graduate high school on-time based on student sixth-grade data. For the upsampled logistic regression prediction, the accuracy was 0.679, true-positive rate was 0.71, and false-positive rate of 0.40. The downsampled version had an accuracy of 0.728, true-positive rate was 0.68, and false-positive rate of 0.36. The upsampled logistic regression model had an AIC level of 1944.0 and could be compared to the downsampled logistic regression model's AIC level of 346.4. Thus, the downsampled model is identified as the preferred model.

The Nagelkerkepseudo R-squared value is helpful when it is compared to another pseudo R-squared of the same type and predicting the same outcome. The upsampled logistic regression model had a Nagelkerkepseudo R-squared value of 0.312 and can be compared to the downsampled logistic regression model's pseudo R-squared value of 0.366. When compared to the upsampled logistic regression model's pseudo R-squared level, the downsampled model is again identified as the preferred model.

For the upsampled logistic regression modelshowed that bypassing English in the sixth grade, a student increases their odds of graduating by a factor of 4.7, given all other variables are unchanged. If a student received free or reduced lunch, the odds of a student graduating decreased by 70.2% (0.298 – 1 = -0.702), keeping other variables constant. For the downsampled logistic regression model showed that if a student had multiple moves in the sixth grade, the odds of a student graduating decreased by 48.9% (0.511 – 1 = -0.489), keeping all other variables constant.

Linear discriminant analysis makes predictions by estimating the probability that new inputs belong to a particular class. The upsampled LDA model had an accuracy of 0.693, true-positive rate of 0.69, and false-positive rate of 0.38 and had results that closely aligned with those of the upsampled logistic regression models. That is, the coefficients of linear discriminants—gender (0.768) and passing English in the sixth grade (0.733)—show that those predictors were the most influential in determining a student's likelihood of graduating, as they were the ones with the highest magnitude. Having a lower chance of graduating from high school was associated with the student receiving free or reduced lunch (-0.772) and being suspended out of school (-0.847).

The downsampled LDA model,which had an accuracy level of 0.670, true-positive rate of 0.67, and false-positive rate of 0.34, performed worse than the upsampled model. In this model, the most important features were passing English (1.416), race (1.22), and gender(1.02).

Using quadratic discriminant analysis (QDA), each observation was classified in the group that had the least squared distance. For the upsampled QDA prediction, the accuracy was 0.830, the true-positive rate was 0.88, and the false-positive rate was 0.61. The downsampled version had an accuracy of 0.746, true-positive rate of 0.78, and false-positive rate of 0.49. For the upsampled data set, there was a big difference in group means for being suspended out of school. The group means showed that 13.7% of the graduates had been suspended in the sixth grade, while 35.7% of the nongraduates were suspended.The downsampled QDA group means show the drastic difference between graduates and nongraduates means for the attendance and behavior variables. This difference means that sixth-grade students who missed more than 20% of school days and were suspended are less likely to graduate or graduate on time than students who missed 20% or fewer school days and were not suspended in the sixth grade.

The QDA models outperformed the upsampled and downsampled LDA models, suggesting that the decision boundary for that data might be better fit by a quadratic curve rather than a linear one. Of the three statistical models, the upsampled and downsampled QDA models had the highest accuracy at 82.0% and 74.6%, respectively. The downsampled LDA model had the lowest accuracy at 67.0%.

Overall, variables consistently identified in most of the sixth-grade models as able to predict students who would not complete high school within four years were: (a) the student's gender (male), (b) if the student were suspended from school, (c) if the student had multiple school moves, and(d) if the student did not pass English.

To determine the answer of which statistical model was most accurate at predicting future dropouts or late graduates utilizing sixth-grade variables, the area under the curve shows that sixth-grade variables predict with relatively high accuracy for all the statistical model (see Table 2). The area under the curve was 0.700or higher for every test. The results for the area under the curve values for the upsampled and downsampled logistic regression analyses were 0.752 and 0.736, respectively. Linear discriminant analysis had an area under the curve value of 0.754 for the upsampleddata setand an area under the curve value of 0.736 for the downsampled data set.The results for the area under the curve values for theupsampled and downsampledquadratic discriminant analyses were 0.700 and 0.703, respectively.

The results of the confusion matrix show that sixth-grade variables predict if a student will graduate within four years of high school with reasonable accuracy for all the statistical models because the accuracy was 0.67 and higher for all analyses. The accuracyofthe quadratic discriminant analysis showed good results of 0.83 for the upsampled data set and 0.75 for the downsampled data set. Logistic regression was close behind with an accuracy of 0.68 for both the upsampled and downsampled logistic regression. The results of the upsampled and downsampled linear discriminant analysis had an accuracy of 0.69and 0.67, respectively.

Table 2Analysis Results Based on Sixth Grade Variables for all Data Types and Statistical Models Using Test Data

Statistical Model Used for Sixth Grade Data	AUROC	Accuracy	Kappa	Sensi-tivity	Speci-ficity	F1	Balanced Accuracy
Upsampled LR	0.752	.68	.17	.71	.60	.79	.65
Downsampled LR	0.736	.68	.17	.68	.64	.79	.65
Upsampled LDA	0.754	.69	.19	.69	.62	.80	.66
Downsampled LDA	0.736	.67	.18	.67	.66	.78	.66
Upsampled QDA	0.700	.83	.26	.88	.39	.90	.64
Downsampled QDA	0.703	.75	.20	.78	.51	.84	.64

Note. AUROC is Area Under the ROC Curve.

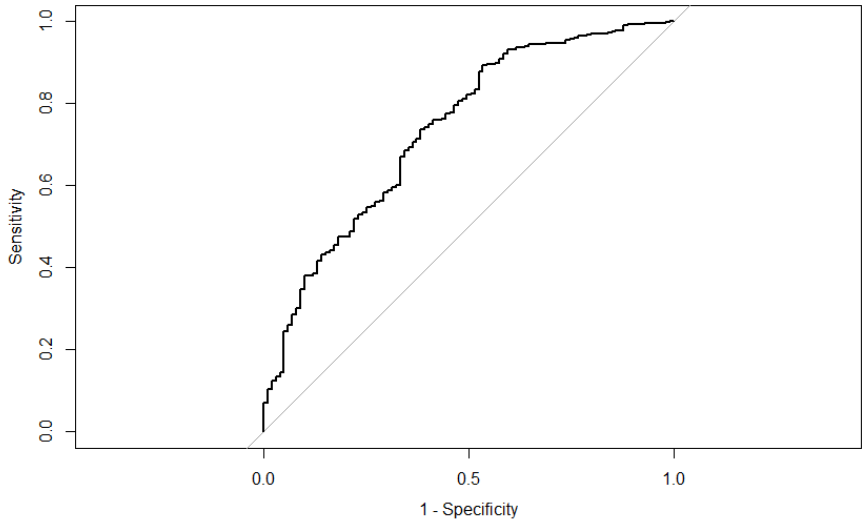

Figure 2. ROC curve results based on sixth-grade variables used to predict graduation utilizing downsampled linear discriminant analysis.

The sixth-grade models displayed the same tradeoff between thetrue-positive rate and the false-positive rate that the ninth-grade models did, albeit with generally lower accuracy than the ninth-grade models. In terms of accuracy and true-positive rate, the upsampled QDA model is best, with an accuracy level of 0.83, a true-positive rate of 0.88, and a kappa value of 0.26. However, the high levels come at a price because the upsampled QDA also had the lowest balanced accuracy and highest false-positive rate. The model with the lowest false-positive rate was the downsampled LDA model (false-positive rate = 0.34), which showed that the downsampled LDA model performed much better in reducing false-positive predictions when it had small, equally-sized classes for its training. The downsampled LDA model also had the third-highest kappa value and the highest balanced accuracy.

Discussion and Conclusion

Due to the negative consequences caused by dropping out of high school, schools must take steps to ensure all students graduate(Zvoch, 2006). An early warning system could be used at the school-level as well as the district-level to identify the students who need guidance and interventions to graduate within four years of starting high school. Creating an accurate dropout early warning system involves finding the best combination of indicators considering both true-positive and false-positive results and identifying when it is best to begin identifying the students(Jerald, 2007).By accurately identifying the students most at risk for not graduating on time, schools could use their limited resources effectively. The results used in this study support previous research that identified ninth-grade dropout predictors and sixth-grade dropout indicators.

District staff or school staff can identify those students who are at risk of not graduating high school within four years. In the ninth grade, variables such as checking credits at the end of the school year, school attendance, school suspensions, gender, and the number of schools attended in the ninth grade can all easily be pulled from the school's student database. Sixth-grade variables can also be easily pulled from a school's database and used to predict students who are at risk for not graduating high school on-time. Those sixth-grade variables include school suspensions, gender, number of schools attended in the sixth grade, and if the student did not pass English.

Although school staff can more accurately identify students who will not graduate on time when students are in the ninth grade, it is important to utilize sixth grade to intervene and help students get back earlier on the path to graduation(Balfanz, 2009). The reduction in accuracy is negligible when compared to the benefits of helping identified students get back on the path to graduation three years sooner.

The identified variables by grade in statistical models can predict with a relatively high true-positive rate and a low false-positive rate. It is wiser to choose a model with the lowest false-positive rate even if the true-positive rate suffers slightly from that choice. It is a better option to offer help to more students than need it than to fail to identify students who need help.

It is this option that leads to the conclusion that for ninth grade, the downsampled QDA model is the best model for minimizing false-positive ratewhich was the lowest at 0.43, and its true-negative value which was the highest at 0.56. It also had an accuracy level of 0.85 and a true-positive rate of 0.90. For the sixth grade, the downsampled LDA model had the lowest false-positive rate at 0.34 and accuracy and true-positive rate of 0.67. Note the sixth-grade accuracy in these models was higher than inprevious sixth-grade studies reviewed.

This study identified both middle school and high school students as likely to drop out or not graduate within four years of entering high school. Identification of those students allows schools to provide interventions to get them on track for on-time graduation. The impact of putting students back on track to graduation could be life-changing for the students and result in a better quality of life for both the students, their families, and society.

References

Allensworth, E., & Easton, J. (2005). The on-track indicator as a predictor of high school graduation. Consortium on Chicago School Research at the University of Chicago. Retrieved from http://www.aaronjmeyer.com/storage/OnTrackIndicator.pdf

Allensworth, E., & Easton, J. (2007). What matters for staying on-track and graduating in Chicago Public High Schools. Consortium on Chicago School Research. Retrieved from http://files.eric.ed.gov/fulltext/ED498350.pdf

Alliance for Excellent Education. (2011). The high cost of high school dropouts: What the nation pays for inadequate high schools. Retrieved from http://www.all4ed.org/files/HighCost.pdf

Amos, J. (2008). Dropouts, diplomas, and dollars: U.S. high schools and the nation's economy. Alliance for Excellent Education. Retrieved from http://all4ed.org/wp-content/uploads/2008/08/Econ2008.pdf

Balfanz, R. (2009). Putting middle grades students on the graduation path: A policy and practice brief. National Middle School Association. Retrieved from file:///Users/mbienkowski/Dropbox/zPapers/Library.papers3/Articles/2009/Balfanz/2009 Balfanz.pdf%5Cnpapers3://publication/uuid/E7AC62A6-F5F0-4EBD-81C3-565AE9C35E8A

Balfanz, R., Herzog, L., & Mac Iver, D. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. Educational Psychologist, 42(4), 223–235. https://doi.org/10.1080/00461520701621079

Brundage, A. (2014). The use of early warning systems to promote success for all students. Retrieved from http://www.fldoe.org/core/fileparse.php/5423/urlt/ews.pdf

Bureau of Labor Statistics, U. S. D. of L. (2019). Median weekly earnings $606 for high school dropouts, $1,559 for advanced degree holders. Retrieved August 25, 2020, from https://www.bls.gov/opub/ted/2019/median-weekly-earnings-606-for-high-school-dropouts-1559-for-advanced-degree-holders.htm

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, 18(1), 84–100.https://doi.org/10.1080/10824669.2013.745210

DePaoli, J., Fox, J., Ingram, E., Maushard, M., Bridgeland, J., &Balfanz, R. (2015). Building a grad nation: Progress and challenge in ending the high school dropout epidemic. Retrieved from https://files.eric.ed.gov/fulltext/ED530320.pdf

Heppen, J., &Therriault, S. (2008). Developing early warning systems to identify potential high school dropouts. Retrieved from http://files.eric.ed.gov/fulltext/ED521558.pdf

James, G., Witen, D., Hastie, T., &Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. ScienceDirect, 64(9-12),856-875.https://doi.org/10.1016/j.peva.2007.06.006

Jerald, C. (2006). Identifying potential dropouts: Key lessons for building an early warning data system. Achieve, Inc. Retrieved from http://www.achieve.org/files/IdentifyingPotentialDropouts.pdf

Jerald, C. (2007). Keeping kids in school: What research tells us about preventing dropouts. Center for Public Education, 1–18. Retrieved from http://www. centerforpubliceducation.org/Main-Menu/Staffingstudents/Keeping-kids-in-school-At-a-glance/Keeping-kids-in-school-Preventing-dropouts.html

Jobs for the Future. (2014). Early warning indicators and segmentation analysis: A technical guide on data studies that inform dropout prevention and recovery. Retrieved from http://www.jff.org/sites/default/files/publications/materials/earlywarningindicators.pdf

Kemple, J. J., Segeritz, M. D., & Stephenson, N. (2013). Building On-Track Indicators for High School Graduation and College Readiness: Evidence from New York City. Journal of Education for Students Placed at Risk, 18(1), 7–28. https://doi.org/10.1080/10824669.2013.747945

Laird, J., Kienzl, G., DeBell, M., & Chapman, C. (2007). Dropout rates in the United States: 2005. U.S. Department of Education. Retrieved from http://files.eric.ed.gov/fulltext/ED497226.pdf

Lee, T., Cornell, D., Gregory, A., & Fan, X. (2011). High suspension schools and dropout rates for black and white students. Education and Treatment of Children, 34(2), 167–192. https://doi.org/10.1353/etc.2011.0014

Leech, N., Barrett, K., & Morgan, G. (2008). SPSS for intermediate statistics: Use and interpretation (3rd ed.). New York: Taylor & Francis Group, LLC.

Linear & quadratic discriminant analysis. (n.d.). Retrieved from http://uc-r.github.io/discriminant_analysis

Mac Iver, M. (2010). Gradual disengagement: A portrait of the 2008-09 dropouts in the Baltimore City Schools. Baltimore Education Research Consortium, 1–16. Retrieved from http://acy.hs-cluster-1.net/upimages/Gradual_Disengagement.pdf

Mac Iver, M., & Mac Iver, D. (2010). Keeping on track in ninth grade and beyond: Baltimore's ninth graders in 2007-08. Retrieved from http://new.every1graduates.org/wp-content/uploads/2012/03/Keeping_On_Track_in_Ninth_Grade_and_Beyond.pdf

Mac Iver, M., & Messel, M. (2012). Predicting high school outcomes in the Baltimore City Public Schools. The Senior Urban Education Research Fellowship Series, VII(Summer 2012). Retrieved from http://baltimore-berc.org/pdfs/PredictingHighSchoolOutcomes.pdf

Mac Iver, M., & Messel, M. (2013). The ABCs of keeping on track to graduation: Research findings from Baltimore. Journal of Education for Students Placed at Risk, 18(1), 50–67. https://doi.org/10.1080/10824669.2013.745207

McKee, M., & Caldarella, P. (2016). Middle school predictors of high school performance: A case study of dropout risk indicators. Education, p. 515. Retrieved from http://eds.a.ebscohost.com/eds/pdfviewer/pdfviewer?vid=1&sid=44c5b4e5-47a6-4f39-aca7-b38c2d219f86%40sessionmgr4006

Neild, R. (2009). Falling off track during the transition to high school: What we know and what can be done. Future of Children, 19(1), 53–76. https://doi.org/10.1353foc.0.0020

Pinkus, L. (2008). Using early-warning data to improve graduation rates: Closing cracks in the education system. Washington, DC: Alliance for Excellent Education, (August). Retrieved from http://beta.fresnounified.org/gradtf/Shared Documents/Using Early Warning Data to Improve Graduation Rates, Closing Cracks in the Education System.pdf

Rumberger, R. (2004). Why students drop out of school. In Dropouts in America: Confronting the graduation rate crisis (pp. 131–155). Cambridge: Harvard Education Press.

Rumberger, R., & Lim, S. (2008). Why students drop out of school: A review of 25 years of research. Retrieved from http://www.cdrp.ucsb.edu/pubs_reports.htm

Silver, D., Saunders, M., & Zarate, E. (2008). What factors predict high school graduation in the Los Angeles Unified School District. Retrieved from www.lmri.ucsb.edu/dropouts

U.S. Department of Education. (2016). Issue brief: Early warning systems. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/ews.asp.

UCLA: Statistical Consulting Group. (n.d.). Introduction to SAS. Retrieved from https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/

Zvoch, K. (2006). Freshman year dropouts: Interactions between student and school characteristics and student dropout status. Journal of Education for Students Placed at Risk, 11(1), 97–117. https://doi.org/10.1207/s15327671espr1101