Enhancing Student Success through Predictive Modeling in Data-Driven Learning Analytics

Ethan J. Blackwood and Ava L. Thompson

Ethan J. Blackwood, Department of Education and Learning Technologies, University of California, Irvine; Ava L. Thompson, Center for Data-Driven Education, University of Michigan, Ann Arbor

ABSTRACT

Analytic tools are useful for detecting patterns in education data and providing insights about student performance and learning. This study compared six supervised learning algorithms (e.g., linear regression, ridge regression, lasso, regression trees, random forests regression, gradient boosted regression) and identified features important for predicting student performance. The dataset consisted of N=1044 observations from two secondary schools in Portugal [1]. Performance was assessed by final grades (*range*: 0-20) in two courses, mathematics and Portugese. The models were fit to training data with 27 independent variables and evaluated on a testing subset. Overall, performance was lower for students in mathematics than Portugese. The models selected a similar set of variables as important for predicting performance: Mother's education level, student plans for higher education, and weekly study time were positively related to predicted performance, whereas course subject, school educational support, and romantic relationships were associated with decreased student performance. The models differed in the number, weighting, order and importance given to the predictor variables. Linear regression provided a model with 13 predictors. Ridge regression shrank the coefficient estimates toward zero; the lasso performed variables selection for a model with 20 predictors. There was a tradeoff between model complexity and interpretability. The single pruned regression tree provided a simple, interpretable non-linear model that branched on four features. Random forests regression and gradient boosting reduced overfitting, but were more difficult to interpret. Advantages and limitations of the different models are considered. Applications for educational data mining (EDM) and learning analytics (LA) are discussed. [1]

KEYWORDS

Predictive Modeling, Variable Importance, Learning Analytics

1 INTRODUCTION

Education institutions have generated very large amounts of student data in recent decades due to dramatic increases in computing speed and processing power [2, 3]. The development and use of analytic approaches for predictive modeling allows researchers and educators to discover patterns in data and provide insights about learning for effective decision making. Educational data mining (EDM) and learning analytics (LA) are multidisciplinary fields at the intersection of learning science, social science, statistics, and computer science that leverage big data to understand learning and the environments in which it occurs [4, 5]. Predictive modeling provides useful methods for analyzing the factors that contribute to student success and identifying individuals at risk for dropping out. Evaluating different predictive models and approaches to feature selection is useful for determining which approach is best for predicting student performance.

Historically, education institutions have tracked student performance, dropout, retention, and used analytic tools to identify factors central to learning such as persistence and social integration [6]. As extensive education datasets became available for analysis, EDM/LA researchers have applied a diverse range of descriptive, correlational, and predictive methodologies to discover potentially useful patterns in the data for understanding learners and learning in different contexts [7]. The fields of EDM and LA both share the goal of using research methods and predictive analysis to improve student performance and instructional design [6, 8, 9]. EDM research has focused more on the technical challenges of extracting value from big data in education [10, 11], whereas LA takes a more holistic, education-focused approach to learning that seeks to inform and empower instructors and learners [12, 13]. Despite differences in their respective origins and emphases, LA and EDM are complementary approaches that use similar methodologies.

Together, EDM and LA represent an ecosystem of techniques for gathering, processing, and acting on data to promote learning. The main analytic approaches used in these areas include discovery with models (i.e. modeling), similarity grouping, relationship mining, content analysis, and social network analysis (SNA) [5, 8]. These procedures facilitate the preparation, measurement, and collection of data about learning activities for subsequent analysis, interpretation, and reporting. Student characteristics are often modeled in terms of domain knowledge, motivation, metacognitive abilities (i.e. thinking about thinking), learning strategies, attitudes, and affect [13]. Analyses of learner data and patterns identified within these data have been directed at predicting learning outcomes, recommending resources, and detecting error patterns [14]. The output from EDM/LA research has provided insights for various stakeholders, including learners, educators, and administrators. This paper focuses on predictive modeling of student performance as a form of data-driven learning analytics.

[1]

1.1 Predictive Modeling

Predictive modeling involves a set of statistical procedures and automated processes for extracting knowledge from data [15, 16]. Two main branches of predictive modeling are supervised learning and unsupervised learning. Supervised learning problems involve prediction about a specific outcome or target variable (i.e. course grade) when examples of input/output pairs are available in the data. If a dataset has no target outcome, unsupervised learning methods

(e.g. clustering) can reveal structure in unlabeled data. Clustering can be used to group individuals based on similar learning profiles.

In this study, student performance is analyzed as a supervised learning problem based on final course grades.

Two main approaches for supervised learning problems are classification and regression. For a binary or categorical outcome that is represented as a class label (e.g., pass, fail), a classification model will predict which class or category that new instances are assigned to. When the target variable to be predicted is measured on a continuous scale (e.g. GPA), a regression model tests how a set of attributes or features predicts the target outcome. Classification is the most commonly used data analytic method for modeling students and their behavior and can include methods such as logistic regression, support vector machines, naive Bayes, decision trees, and neural networks [17].

The present study compares several regression models of student performance to identify the set of features that best predict student performance. Each model was first trained on a set of input-output pairs and then used to make predictions about new observations that were previously set aside. Comparing different predictive models can help determine which model is best for a given problem with the data available [18]. Past empirical findings indicate that, in addition to course assessments (i.e. number of quizzes passed), student engagement and participation in course activities are the most influential predictors of final grades [11, 13]. A student's sense of belonging is also essential for engagement and improved course satisfaction, which can in turn lead to reduced student dropout.

1.2 Linear Models

1.2.1 Linear Regression. A general assumption of linear regression is that the target outcome can be represented as a linear function of the input features. The standard linear model describes the relationship between predicted target variable (Y) from a set of features ($X_1 \ldots X_p$), including some measure of error (equation 1). The predicted value of the target outcome can be thought of as the weighted sum of the input features with the weights or coefficients (i.e., beta values) indicating the influence of a given feature on the outcome. Ordinary least squares (OLS) regression miminizes the distance (i.e., error) between the predicted values of *Y* and the observed values in the dataset. If the number of observations (n) is much larger than the number of features (p), OLS coefficient estimates will have low variance and perform well on test observations; however, if the number of observations n is not much larger than the number of features p, high variability in the OLS fit can result in overfitting and poor prediction on the test observations. For high-dimensional datasets ($p >> n$), the least squares coefficient estimate breaks down. The simple linear model can be improved by using alternative fitting approaches that produce better prediction accuracy and model interpretability [15].

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \quad (1)$$

In many regression analyses, it is often the case that multiple independent variables or features will not be correlated with the target outcome. Three methods for improving the fit of linear models are: (a) subset selection, (b) dimension reduction, and (c) regularization (i.e.., shrinkage). Determining which set of features is best for representing the predicted outcome is essential for model interpretation. A straightforward approach to feature selection is to first conduct a regression including all the independent variables and then rerun the regression while excluding non-significant variables from the model. Another approach, termed *regularization*, includes all p predictor variables, but constrains (i.e., regularizes) the coefficient estimates of the independent variables by shrinking them towards zero. Regularization reduces variablity, which improves accuracy on the testing set with a slight increase in bias. Shrinking the coefficient estimates of irrelevant features toward zero reduces overfitting and can aid model interpretation.

1.2.2 Ridge Regression: L2 Penalty. As with OLS, ridge regression seeks coefficient estimates that fit the data well by reducing error, but ridge regression introduces a shrinkage penalty (L2) that has the effect of shrinking the coefficient estimates towards zero. When the tuning parameter (lambda) is set to zero, the shrinkage penalty has no effect and ridge regression produces the least squares estimates. As the value of lambda increases, the estimated regression coefficients approach zero [15]. The advantage of ridge regressions over least squares is based on the bias-variance tradeoff. As the tuning parameter lambda increases, the flexibility of the ridge regression decreases, leading to decreased variance but increased bias. Lower variance is associated with reduced overfitting, whereas higher bias can lead the model to miss relevant relations between features and target outputs (underfitting). Ridge regression is often applied after standardizing the predictor variables so that they are all on the same scale (e.g., M=0, SD=1). Ridge regression performs well with high-dimensional datasets (p>>n) by trading off a small increase in bias for a large decrease in variance. A disadvantage of ridge regression is that, because it includes all predictors in the model, the penalty shrinks the coefficients toward zero, but does not set any of them exactly to zero. This can create a problem for model interpretation when working with a very large number of features.

1.2.3 The Lasso: L1 Penalty. The lasso and ridge regression have similar formulations, but the lasso has a major advantage over

ridge regression as it produces simpler, more interpretable models based on a subset of features. The lasso uses the L1 penalty which has the effect of forcing some of the coefficient estimates to be equal exactly to zero when the tuning parameter lambda is sufficiently large [16]. The lasso performs variable selection and produces sparse models based on a subset of features, which are generally easier to interpret than ridge regression. The lasso implicitly assumes that a number of the feature coefficients or weight truly equal to zero. In general, the lasso performs better than ridge regression in situations where a small number of features account for most of the variability in the target outcome, and the remaining features have coefficients that are very small or equal to zero. By contrast, ridge regression performs better when the target is a function of a large number of predictors that contribute approximately equally to the coefficients. Cross-validation is used to determine which value of the lambda parameter is optimal. In general, least squares regression (OLS) performs well when the number of observations is larger than the number of features ($n >> p$); however, ridge regression and the lasso are preferred when working with a very large number of predictors ($p >> n$).

1.3 Non-Linear Models

1.3.1 Regression Trees. Decision tree models are widely used for classification and regression. Tree models are built on a hierarchy of *if-else* questions that proceeds from a root node as the starting point and continues through a series of decisions. Each node in the tree represents either a question or a terminal node (i.e.,leaf) that contains the outcome. In constructing the tree, the algorithm searches through all possible decisions, or tests, and finds a solution that is most informative about the target outcome. The recursive branching process of tree based models yields a binary tree of decisions, with each node representing a test that considers a single feature. This process of recursive partitioning is repeated until each leaf in the decision tree contains only a single target. Prediction for a new data point proceeds by checking which region of the partition the new point falls into and predicting the majority in that feature space. Tree based models require little adjustment and are easy to interpret. A drawback is that they can lead to very complex models that highly overfit data used to train the model. A good strategy for building a regression tree is to grow a very large tree and then prune it back to obtain a subtree that provides the lowest test error rate. A good way to prevent overfitting is to use pre-pruning to limit the maximum depth of the tree.

1.3.2 Random Forest Regression. A random forest is a collection of decision trees that are each slightly different, with each tree overfitting the data in a different way. This approach reduces overfitting by building many trees and averaging their results. Randomness is introduced into the tree building process in two ways: first, by drawing a random subset (i.e. bootstrap sample) of the data, and second by selecting a random subset of features at each node branch [19]. In building the random forest, the user must first decide how many trees to build and the algorithm makes different random choices so that each tree is distinct [18, 20]. The bootstrapping method repeatedly draws random samples of size n from the dataset with replacement. The decision trees are built on these random samples that are the same size as the original data, with some points missing and some data points repeated. The algorithm also selects a random subset of p features, that are repeated separately at each node, so that each decision at the node branch is based on a different subset of features. These two processes help ensure that all of the decision trees in the random forest are different.

1.3.3 Gradient Boosting. Similar to random forests, gradient boosting is an ensemble approach that builds many smaller trees; however, with each new tree the gradient boosting algorithm attempts to correct for deficiencies of the current ensemble. In contrast to random forests, gradient boosting grows smaller, stubbier trees, and goes after bias [15, 16]. Gradient boosted regression trees use strong prepruning, with shallow trees of a depth of one to five. Thus, each tree provides an estimate of part of the data. Combining many shallow trees iteratively improves model performance. Gradient boosting and random forests perform well on similar tasks and data. A common practice is to first construct random forests and then use gradient boosting to improve model accuracy [20].

1.4 Study Goals

This project examines the relationships between student characteristics, behavior, and performance. Data on student performance from two secondary schools in Portugal was obtained from the UC-Irvine machine learning repository (UCI-MLR) [1]. The dataset included information from a student survey and school grade records. The data were fit using several supervised learning regression models (described above). The predictor variables of interest were demographic features, family characteristics, and student behaviors (e.g., weekly study hours, romantic relationships). The different models explored various dimensions of student performance by: (i) Analyzing the combination of factors that best predict student performance, (ii) Selecting the variables most important for predicting performance, and (iii) Identifying the most accurate and interpretable model of predicted student performance.

2 METHOD

2.1 Data

The student performance dataset downloaded from the UCI-MLR was saved as a data frame object in a python interactive notebook. The data was collected from two secondary schools in the Alentejo region of Portugal during the 2005-2006 school year and contained information from a questionnaire and school reports of student grades [1]. The sample consisted of 1044 students (56.6% female, *M*age=16.71 years, *SD*=1.19, *Median*=17 years, *range*=15-22). Age was measured as a categorical variable (n=10 individuals between the ages of 20 to 22 were included in the category: 19+ years). The

dataset consisted of 30 independent variables, including demographic information, social/ emotional attributes, school-related variables, and student behaviors (see Table 1). The target variable, student performance was evaluated on a 20 point scale as in other European countries (e.g. France) at three points during the school year (i.e., Grade1, Grade2, Grade3) for two courses: Mathematics (*n=395*) and Portugese (*n=649*). The target variable of interest was the final course grade (G3). A binary dummy variable of student performance was calculated based on the measure of final exam grades (Pass: *G3*>10, Fail: *G3*<=10) for descriptive purposes.

2.2 Model Construction

2.2.1 Linear Regression (OLS). All models were constructed in *R* (using *Rstudio*) [15]. After preliminary exploration of the data, the sample was divided into the training set (n_1=731) and testing set (n_2=313) using a 70 to 30 percent split. Each model was first fit to training data and evaluated on the testing set. Student performance was regressed on 27 independent variables shown in Table 1 using the general linear model (OLS). The regression model was run on the training set with the full set of predictor variables; the model was then rerun excluding all non-significant predictors variables from model. The final model was then evaluated on a subset of hold-out data in the testing set.

2.2.2 Ridge regression (L2 penalty). The glmnet package was used to fit the ridge regression and lasso models. The glmnet() function does not use model formula language, so the *X* matrix of predictors and target vector *Y* were passed to the model. The model.matrix() function produced a matrix corresponding to the 27 predictors and automatically transformed any qualitative variables into dummy variables. The *alpha* parameter in the glmnet() function determines what kind of model is fit: *alpha* = 0 is used to fit ridge regression. It is important to select an appropriate value of the parameter *lambda*, as the algorithm generates a different set of coefficients for each value of lambda. By default, the glmnet() function performs ridge regression for an automatically selected range of lambda values (e.g. 100). The glmnet function also standardizes the variables so they are all on the same scale. The shrinkage penalty is applied to every feature, but not the intercept.

2.2.3 The Lasso (L1 Penalty). The Lasso model was fit using the glmnet() function with alpha=1. The model automatically calculates correlation estimates for a wide range of lambda values. Cross-validation was used to select an optimal value of the tuning parameter lambda. The lasso is similar to best subset selection as it tries to find the set of coefficient estimates that leads to the smallest error (RSS). In terms of the bias-variance tradeoff, the lasso is qualitatively similar to ridge regression. As the value of *lambda* increases, the variance decreases and bias increases somewhat.

2.2.4 Regression Trees. The regression tree model of student performance was fit to the training data using the rpart() function in R, with all 27 independent variables. The decision tree uses recursive binary splitting to construct a large tree on the training data. Cross-validation was used to determine the optimal tree complexity. The model was prepruned to a maximum depth of 3, which means the algorithm split on three consecutive features.

2.2.5 Random Forests Regression. The random forest model was fit using 1000 trees, with all of the features considered at each node to determine the randomness of each tree. In general, random forests work well without very much parameter tuning or scaling of data. The important parameters for the random forests algorithm are the number of sampled data points and the maximum number of features; the algorithm can look at all of the features in the dataset or a limited number. A high value for *maximum-features* will produce trees in the random forest that are very similar and will fit the data easily based on the most distinctive features, whereas a low value will produce trees that are very different from each other, which reduces overfitting.

Table 2: Correlation Matrix of Previous Course Failures and Course Grade Variables

Variable	Failures	Grade 1	Grade 2	Grade 3
Failures	1.00	0.37***	0.38***	0.38***
Grade 1		1.00	0.86***	0.81***
Grade 2			1.00	0.91***
Grade 3				1.00

Note. *** p<0.001

2.2.6 Gradient Boosted Regression Trees. In addition to prepruning and the number of trees, an important parameter for gradient boosting is the *learning rate* which determines how strongly each tree tries to correct for mistakes of previous trees. A high learning rate produces stronger corrections, allowing for more complex models. The gbm package was loaded, and the gbm() function was called on student performance (final grade) using the Gaussian distribution, with 1000 shallow trees, a shrinkage parameter = 0.01, and interaction depth of 4 splits.

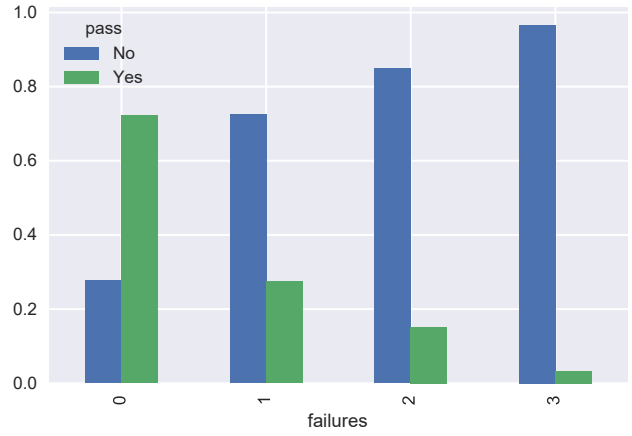

Figure 1: Proportion of Passing or Failing Final Grades as a Function of Previous Course Failures

3 RESULTS

3.1 Exploratory Data Analysis

Preliminary examination of the data revealed significant intercorrelations between the course evaluation variables Grades 1, Grade 2, and Grade 3 which accounted for significant portions of variance in the target outcome (see Table 2). In addition, previous course failures was significantly correlated with all three course grade measures. To address the issue of multicolinearity, Grade 1, Grade 2, and past course failures were not included in the regression analyses reported below.

Table 3 provides descriptive statistics for selected attributes. Chisquared tests of independence were used to compare the proportion of students who received passing and failing grades by attribute. There was no relationship between performance and sex; males Table 3: Summary Table of Student Performance by Final Course Grade (Pass>=10, Fail<10) for Selected Variables

	Pass		Fail	
Attribute	*N*	%	N	%
Total	661	63.3%	383	36.7%
Male	277	61.1%	176	38.9%
Female	384	65.0%	207	35.0%
Course				
Portugese	452	69.6%	197	30.4%
Math	209	52.9%	186	47.1%
Mother's Education				
Higher Ed	235	76.8%	71	23.2%
Secondary	143	49.5%	95	32.9%
Grades 5 to 9	180	75.6%	109	45.8%
Primary	98	47.5%	106	52.5%
None	7	77.8%	2	22.2%
Higher Education Plans				
Planned	640	67.0%	315	33.0%
No Plans	21	23.6%	68	76.4%
School Support				
Received	63	52.9%	56	47.1%
None	598	64.6%	327	33.4%
Study Time				
More than 10 hrs.	45	72.6%	17	27.4%
5 to 10 hrs.	123	75.9%	39	24.1%
2 to 5 hrs.	321	63.8%	182	36.2%
Less than 2 hrs.	172	54.3%	145	45.7%
Romantic Relationship				
Yes	221	59.6%	150	40.4%
None	440	65.4%	233	34.6%
Internet Access				
Yes	543	65.7%	284	34.3%
None	118	54.4%	99	45.6%

and females did not differ significantly in performance (p=0.20). Student performance did vary according to mother's level of education (p<0.05), but as seen in Table 3, the relationship between performance and mother's education was non-linear. Performance also varied significantly by course subject (p<0.001); more than two-thirds of students in the Portugese course successfully passed, whereas just over half of students in the mathematics course received a passing grade. The relationship between student performance and plans for higher education was significant (p<0.001). Two-thirds of students with plans for higher education received a passing grade, whereas less than one-quarter of students with no plans for higher education passed their course. Extra educational school support was significantly related to performance (p<0.001). Just over half of students who received extra educational support at school received a passing grade compared to nearly two-thirds of students who did not receive extra support.

Table 4: Coefficient Estimates for Regression Models of Student Performance on Training Set and Testing Set

	Training Set			Testing Set		
Variables	Coefficient	S.E.	t-value	Coefficient	S.E.	t-Value
Intercept	13.289***	2.166	6.14	6.955*	3.629	1.92
Course	-2.225***	0.261	-8.53	-1.162***	0.445	-2.61
Mother's Education	0.485***	0.122	3.97	0.473**	0.213	2.25
Go Out with Friends	-0.442***	0.114	-3.89	-0.097	0.178	-0.55
Higher Ed	1.695***	0.487	3.48	4.101***	0.772	5.31
School Support	-1.435***	0.416	-3.45	-1.481**	0.647	-2.29
Health	-0.262***	0.088	-2.97	-0.008	0.152	-0.05
Study Time	0.454***	0.156	2.91	0.893***	0.260	3.43
Internet Access	0.819**	0.321	2.55	0.420	0.542	0.77
Family Relations	0.340**	0.140	2.43	-0.096	0.215	-0.45
Romantic Relation	-0.600**	0.266	-2.26	-1.127**	0.461	-2.45
Age	-0.250**	0.114	-2.20	-0.034	0.191	-0.18
Family Size	-0.500*	0.278	-1.80	-0.594	0.458	-1.30
Father's Job	0.278*	0.145	1.92	-0.050	0.245	-0.20
n	730			314		
F-Value	15.41***			5.96***		
df	(13, 716)			(13, 300)		
$R2$	0.219			0.205		
Adj. R^2	0.204			0.171		
Resid. S.E	3.396			3.643		

Note. Significance levels *<0.10 **<0.05 ***<0.01

As expected, weekly study time was significantly associated with student performance (p<0.001). The proportion of passing and failing grades significantly different for students who studied 5 hours or more per week compared to students who studied less than 5 hours per week. The association between romantic relationships and student performance was marginally significant (p<0.06). The proportion of passing and failing grades was significantly different for students in a romantic relationship than students not in a romantic relationship. The relation between internet access and student performance was also marginally significant (p<0.06). The proportion of passing and failing grades was significantly different for students with access to the internet at home compared to students without home internet access (p<0.05).

3.2 Linear Regression and Regularization

3.2.1 General Linear Model. Student performance was first regressed on the 27 independent variables (Table 1) with the training set; this regression was statistically significant and accounted for 19.7% of the variance in the predicted value of student performance, taking into account the number of independent variables, $F(27, 702) = 7.62$, $p < 0.001$ (R^2=0.227, adjusted R^2=0.197). The regression model was rerun, excluding the non-significant predictors, and this regression also yielded a significant relationship between student performance and the independent variables, accounting for 20.4% of the variability in predicted performance, $F(12, 717) = 21.79$, $p < 0.001$ (R^2=0.219, adjusted R^2=0.204). An ANOVA test showed no significant difference between the two models (F< 1.0, p = 0.91) and the simpler model with thirteen predictor variables was retained as the final model. The estimated coefficients, standard error, and t-value on the testing set (ranked by t-Value) are presented in the left side of Table 3.

The final model was evaluated on the testing set and the regression yielded a significant relationship between student performance and the independent variables, accounting for 17.1% of the variability of the predicted value of student performance, taking into account the number of independent variables, $F(13, 300) = 5.96$, p <

0.001 (R^2=0.205, adjusted R^2=0.171). As shown in Table 3, 6 of the 13 independent variables in the testing set were significant, which suggests that the model was overfit to data in the training set, The predicted values of student performance are be explained by the combined effect, or weighted average, of the coefficient estimates and the observed values for each significant independent variable in the model (Equation 2).

$$Y = 6.955 - 1.162(Course) + 0.473(MotherEd) + 4.101(HiðherEd) + 0.893(StudyTime) - 1.481(SchoolSupport) - 1.127(RomanticRel)$$

On the testing set, there was a 1.16 decrease in predicted performance for students in the math course compared to students in the Portugese course, controlling for all other independent variables. A unit change in mother's level of education was associated with a 0.47 increase in predicted student performance, controlling for all other variables. Students with plans to pursue higher education had a 4.10 higher predicted final grade than students with no plans for higher education, controlling for other variables. A one-unit change in weekly study time resulted in a 0.89 increase in predicted student performance, holding constant the effect of other variables. Students receiving school support had a 1.48 lower predicted final course grade than students who did not receive school support, holding all other variables constant. Finally, there was a -1.13 decrease in predicted performance for students in a romantic relationship compared to students not in a romantic relationship, controlling for all other independent variables.

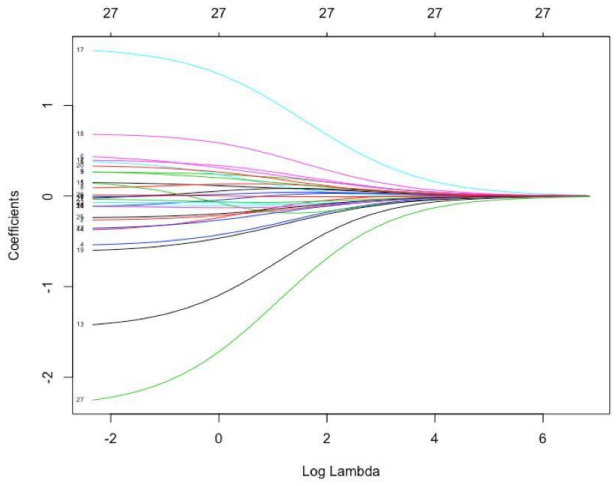

Figure 2: Coefficient Estimates for the Ridge Regression model (L2 Penalty) as a function of the log Values of Lambda

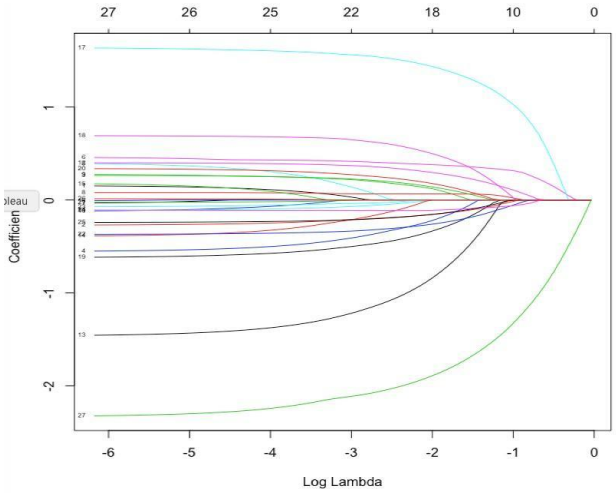

Figure 3: Coefficient Estimates for the Lasso Regression model (L1 Penalty) as a function of the log Values of Lambda

3.2.2 Ridge Regression (L2 Penalty). Figure 2 plots the coefficient estimated from the ridge regression model (L2) as a function of the log values of lambda (x-axis). As the values of lambda become very large, the model shrinks the coefficient values of non-relevant predictor variables towards zero, but the values are never exactly Table 5: Coefficient Estimates for Ridge Regression and the Lasso Model of Student Performance using Best Value of Lambda from Cross-Validation

	Ridge (L2 Penalty)	Lasso (L1 Penalty)
Best lambda (CV)	0.819	0.071
MSE	13.779	13.895
Predictor Variables	*Coefficients*	*Coefficients*
Intercept	13.498	12.899
Course	-1.801	-2.058
Higher Ed	1.390	1.534
School Support	-1.148	-1.121
Internet Access	0.602	0.614
Romantic Relation	-0.487	-0.461
Family Size	-0.447	-0.353
Study Time	0.345	0.353
Mother's Education	0.326	0.403
Going Out	-0.280	-0.316
Family Relations	0.274	0.247
Family Support	-0.258	-0.152
Area (Urban/Rural)	0.242	0.192
Age	-0.223	-0.196
Father's Job	0.209	0.205
Health	-0.203	-0.199

Weekly Alcohol Cons.	-0.130	-0.113
Mother's Job	0.127	0.066
Travel Time to School	-0.109	-0.053
Parents Rel. Status	0.258	0.045
Daily Alcohol Cons.	-0.072	-0.029
Sex	0.118	0.
Paid Extra Classes	-0.050	0.
Extra Activities	0.015	0.
Free Time	-0.062	0.
Absences	0.002	0.
Father's Education	0.046	0.
Student Guardian	-0.052	0.

equal to zero. Cross-validation was used to obtain the best value of lambda, which was, $\lambda = 0.819$ ($\ln\lambda = -0.200$). As shown in Figure 2, the predictors with the highest coefficient values were course subject (27), students' plans for higher education (17), extra school support (13), and internet access at home (18). The ridge regression (L2) was rerun using the best value of lambda from cross-validation with all 27 predictor variables with coefficient estimates shown in Table 5. The ridge regression model had a mean squared error (MSE) of 13.78.

3.2.3 The Lasso (L1 Penalty). Figure 3 plots the estimated coefficients from the lasso (L1) regression as a function of the log value of lambda, with the number of associated features listed across the top of the plot. The plot shows that as the values of lambda increase, the L2 penalty shrinks many of the coefficient values to be equal exactly to zero. Cross-validation was used to obtain the best value of lambda, $\lambda = 0.071$ ($\ln\lambda = -2.65$). The lasso model was rerun using the optimal value of lambda selected by cross validation with 20 predictor variables; the model had a mean squared error (MSE) of 13.895 and accounted for approximately 20 percent of the variability in student performance. Similar to the ridge regression, the predictors with the highest coefficients were course subject (27), plans for higher education (17), extra school support (13), and internet access at home (18). The error from the lasso model is very similar to the ridge regression, but the lasso has an advantage over ridge regression in that the resulting coefficient estimates are sparse and the model selected a subset of the predictor variables.

3.3 Decision Tree Models

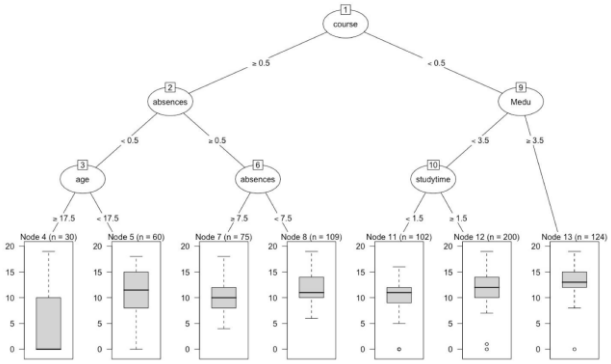

Figure 4: Regression Tree Model of Student Performance on the Traing Set (*n*=700)

3.3.1 Regression Trees. The decision tree model was fit to the training set, with a maximum depth of 3; Figure 4 shows the resulting regression tree with course subject as the root node and 7 terminal nodes. Course subject was a dummy variable; the branch to the left represents students in the mathematics course (39%) and the branch to the right represents students in the Portugese course (61%). In addition to course subject, the algorithm split on mother's education level, weekly study time. student absences, and age in constructing the tree. The values of student performance ranged from 5, for students in the mathematics course with no absences who were 18 years or older, to 13 for students in the Portugese course whose mother's had some higher education. The regression tree model was evaluated on the test set (maximum depth=3) which yielded a tree with plans for higher education (rather than course subject) as the root node and 6 terminal nodes (see Figure 5). The MSE for the regression tree on the testing set was 16.385.

As seen in Figure 5, from the root node of plans for higher education, the algorithm split at nodes for mother's education level, area (urban / rural), course subject, and student absences in constructing the tree. Following the right branch from the root node, students with plans for higher education (91%) had a mean predicted performance of 12, whereas on the left branch, students with no plans for higher education (9%) had a mean predicted performance of only 7. For students with plans for higher education, the next split on mother's education level: Following the branch to the right, students whose mothers had 5th grade level of education or

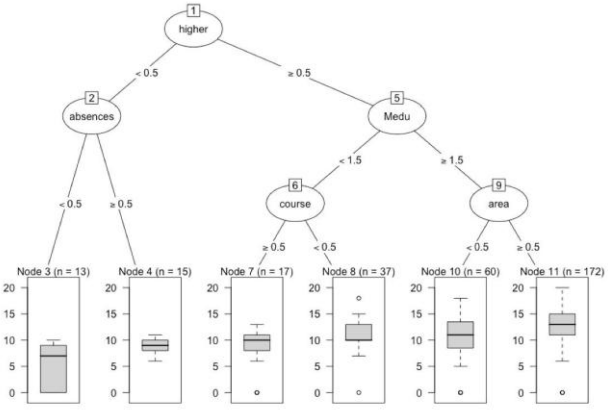

Figure 5: Regression Tree Model of Student Performance on the Testing Set (*n*=314)

higher (74%) had a mean predicted performance of 12. Following this branch to the next node of area, on the right branch students in urban areas (55%) had a mean predicted performance of 13, whereas students in rural areas (19%) had a mean performance of 11. On the left branch, students whose mother's had attended secondary school or lower (17%), the mean predicted performance was 10. This branch split next on course topic, and students in the Portugese course (12%) had a mean predicted performance of 11, whereas the mean performance for students in the mathematics course (5%) was 8.2. For students with no plans for higher education (9%), the next node split on absences, where the mean predicted performance for students with no absences (5%) was 9.1, and students with one or more absences, had a mean predicted performance of 4.6.

3.3.2 Random Forests Regression. The mean squared error (MSE) for the random forests regression was 6.278, which indicates better performance for RF than a single decision tree. The random forests (RF) algorithm provides feature importance as a model summary; for regression, this is measured in terms of percent increase in MSE. The left side of Table 6 provides the feature importance for the RF regression sorted by percent increase in MSE. The algorithm selected mother's education as the most informative feature for predicting student performance (final grade). In contrast to the single decision tree, number of absences and area (urban/rural) were selected as the second and third most important features in the model. Plans for higher education and course subject were also among the most influential predictor variables in the random forest model, but these variables were not given as prominent a position as in the single tree.

3.3.3 Gradient Boosted Regression. The mean squared error (MSE) for the gradient boosted regression tree model was 18.124. Feature importance for the gradient boosted regression trees is presented on the right side of Table 6. Absences and course subject were selected as the two most important features for predicting student performance. The algorithm selected mother's education level, student age, going out with friends, and weekly study time as the next most informative variables, in descending order of importance. Plans for higher education was not selected among the most important variables for predicting student performance in the gradient boosted model.

4 DISCUSSION

The different models identified many of the same variables as important for predicting student performance: course subject, mother's education, student plans for higher education, weekly study time, school support, absences, going out with friends, and romantic relationships. Overall, the performance of students in mathematics was lower than students in the Portugese course, likely owing to the a difference in the difficulty of the subject material. The models differed in the number, weighting, order, and importance given to selected predictors. With linear regression, the predicted outcome is based on the weighted average or combination of all the predictor variables. The OLS linear regression identified thirteen regressor variables in the training set, of which, only six were significant predictors of student performance in the test set, which indicates overfitting. Mother's education level, student plans for higher education, and weekly study time were associated with improved performance, whereas course subject, school support, and romantic relationships were related to decreased performance. The ridge regression reduced model error by shrinking the coefficient estimates towards zero, but with slightly higher bias and risk of underfitting. The lasso model performed variable selection by shrinking the coefficient values of seven non-relevant predictors to exactly zero, which yielded a model with a subset of twenty features. Although the lasso model was simpler than ridge regression, the final OLS model was more parsimonious than the lasso.

The single regression tree provided a simple model that is easy to interpret and gives the mean predicted value of student performance at each node. The algorithm selected course subject as the root node in the training set and branched on mother's education, weekly study time, absences, mother's job, and student age as key nodes for predicting performance. In the testing set, plans for higher education was selected as the root node (which may indicate overfitting) and the algorithm branched on mother's education, residence area (i.e., rural, urban), course subject, and student absences as key nodes. Predicted performance was highest among students from urban areas, whose mothers had attained at least a 5th grade education, with plans for higher education. The lowest predicted performance was found among students who had one or more absences, with no plans for higher education. Random forests regression corrected for overfitting by constructing many trees and averaging across the predicted values. Feature importance in the random forests model identified mother's education as the most informative variable, followed in descending order by student absences, residence area, plans for higher education, course subject, weekly alcohol consumption,

and mother's job. A surprising finding is that the gradient boosted model had a higher error than the random forests regression, given that the gradient boosted algorithm tries to correct for mistakes in previous trees in an iterative process. In terms of feature importance, the gradient boosted model selected student absences as the most informative feature for predicting performance, followed by course subject, mother's education, student age, going out with friends, and study time.

4.0.1 Limitations. A limitation of the present study is that variable importance is not a well-defined concept and lacks a theoretically based quantitative metric [21]. The linear regression model uses significance tests to select the regressors that best predict the target outcome, with non-significant variables excluded from the final model. In this sense, the t-value provides a measure of the importance of a given predictor, and model construction involves a form of variables selection. With random forests, node purity measures branch homogeneity for classification tasks, whereas MSE reduction is used for variable selection on regression tasks. Variable importance with random forests is affected by the number of categories and scale of measurement of the predictor variables [22], but it does not provide a direct indication of the true importance of the variable. In addition, variable importance with random forests can be biased when predictors are measured on different scales or vary in number of categories. Given that much of student data collected by educational institutions is measured on different scales, researchers typically transform variables of interest to the same scale. Transformed variables must be converted back to their original scales for meaningful interpretation of the relationship between the predictors and the target outcome. Furthermore, it may not be feasible to assess the broad construct of learning using a single measure of performance such as a final course grade. A comprehensive analysis of predicted student performance could include multiple dependent measures.

4.0.2 EDM and Learning Analytics. This study revealed several demographic characteristics that play an important role in predicting student performance. It can be difficult to obtain student information from educational institutions owing to privacy protections and confidentiality of student data. This study used archived data from a public repository (UCI-MLR) that included information from both school records and a student survey. Where potential bias in self-report data is a concern, assessing student behaviour on online learning platforms can provide a more direct measures of student performance. Education data mining (EDM) and learning analytics (LA) developed out of the increase in big data in education and shift toward online learning. Much LA/EDM research data is collected within a learning management system (LMS), virtual learning environment (VLE), or massive open online course (MOOC). One of the most widely known platforms for tracking student performance by analyzing LMS data is the Course Signals program [23]. In addition to grades, course signals combined student demographic information, academic history, and student interaction on the Blackboard LMS to track performance. A predictive algorithm was used to calculated the likelihood of student success based on performance, effort, history, and student characteristics. Course signals provided students with real-time feedback about their status in the LMS as traffic indicators (i.e., red, yellow, green). The assessment allowed instructors to enact interventions for high risk students via emails, texts, or face to face meetings with referrals to academic advising and student resource center. Courses that implemented the signals program and provided feedback showed an increase in satisfactory grades, decrease in withdrawals, and improved retention.

4.0.3 Student Performance, Affect, and Motivation. Learning is a complex phenomenon that is not always directly observable and often inferred from behavior. In addition to quantitative measures of online activities, ideally, a meaningful analytics system could include qualitative measures of students' affective states (i.e., boredom, frustration, confusion) or motivation to model engagement in learning activities [24, 25]. LA/EDM researchers has also investigated interactions among learners online [26]. Theories of Social learning demonstrate the importance of collaboration in learning [27]. From a social constructivist perspective, knowledge is constructed through interaction with more knowledgeable partners, including parents, siblings, teachers, or peers. Sociologists have investigated the structure of connections in social networks [28]. Social network analysis (SNA) provides a useful methodology for exploring the role of collaboration in learning and visualizing connections among learners [4]. Past research has revealed that individual differences in students metacognitive abilities (i.e., selfawareness, self-reflection), disposition, experience, and motivation are influential for developing learning relationships [9, 13?]. Future research could help our understanding of the relations among students that contribute to performance.

5 CONCLUSION

Predictive modeling offers a set of analytic tools for detecting patterns in education data and understanding the factors that contribute to successful learning. This study compared six supervised learning models of student performance that varied in complexity and interpretability. Simpler models provided solutions that are easier to interpret but prone to overfitting, whereas more complex models reduce overfitting and decreased error, but are more difficult to interpret. A general conclusion is that comparing the results of several models provides a more complete picture of factors that contribute to student performance than examining any single model. LA/EDM research is increasingly focused on learning outcomes in online platforms. Merging data from LMS or online learning platforms with institutional data about student demographic characteristics can provide a better understanding of student learning and the conditions in which it occurs [29, 30]. Although student absences or past failures are powerful indicators of future performance, motivational factors, such as a student's future orientation toward higher education can reveal a great deal about his or her motivation to succeed. Using a data-driven approach to learning analytics based on student profiles and LMS

activity can also facilitate early detection of at-risk students. These efforts may help inform decision making and policy efforts to address student retention and allocation of resources to improve successful learning outcomes.

REFERENCES

[1] P. Cortez and A. Silva. Using data mining to predict secondary school student performance. In A. Brito and J. Teixeira, editors, *Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008)*, pages 5–12, April 2008.

[2] B. Daniel. Big data and analytics in higher education: Opportunities and challenges. *British journal of educational technology*, 46(5):903–920, 2015.

[3] B. K. Daniel. *Big data and learning analytics in higher education: Current theory and practice*. Springer, 2016.

[4] G. Siemens. Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10):1380–1400, 2013.

[5] G. Siemens and R. S. J.d. Baker. Learning analytics and educational data mining: Communication and collaboration. In *Proceedings of 2nd International Conference on Learning, Analytics, and Knowledge (LAK12)*, May 2012.

[6] R. Ferguson. Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5fi??6):304fi??317, 2012.

[7] G. Siemens and P. Long. Penetrating the fog: analytics in learning and education, 2011.

[8] R. S. J.d. Baker and K. Yacef. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17, 2009.

[9] J. Lester, C. Klein, A. Johri, and H. Rangwala. *Learning Analytics in Higher Education*. Routledge, 2019.

[10] A. Pea-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 41:1432fi??1462, 2014.

[11] C. Romero, S. Ventura, M. Pechenizkiy, and R. Baker. *Handbook of educational data mining*. Chapman and Hall / CRC, 2010.

[12] C. Lang, G. Siemens, A. Wise, and D. Gaevifj. *The Handbook of Learning Analytics, First Edition*. Society for Learning Analytics Research (SoLAR), 2017.

[13] Z. Papamitsiou and A. Economides. Learning analytics and educational data mining: A systematic literature review of empirical evidence. *Educational Technology & Society*, 17(4):49–64, 2014.

[14] K. Verbert, N. Manouselis, W. Drachsler, and E. Duval. Dataset-driven research to support learning and knowledge analytics. *Educational Technology and Society*, 15(3):133fi??148, 2012.

[15] Gareth James, Daniela Witten, Tervor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer Media, 2013.

[16] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, New York, NY, 2013.

[17] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers and Education*, 53(3):950fi??965, 2009.

[18] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning, Second Edition*. Packt, Birmingham, UK, 2017.

[19] R. S. J.d. Baker and K. Yacef. Random forests. *Machine Learning*, 45:5–32, 2001.

[20] Andreas C. Muller and Sarah Guido. *Introduction to Machine Learning*. O'Reilly, Sebastopol, CA, 2017.

[21] U. Gromping. Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, 63(4):308fi??319, 2009.

[22] C. Strobl, A.L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(25), 2007.

[23] K. E. Arnold and M. D. Pistilli. Course signals at purdue: Using learning analytics to increase student success. In *Proceedings from the 2nd International Learning Analytics and Knowledge Conference*, page 267fi??270. ACM, April 2012.

[24] R. S. Baker and A. T. Corbett. Assessment of robust learning in educational data mining. *Research and Practice in Assessment*, 9:38fi??50, 2014.

[25] Z.A. Pardos, R.S. Baker, M.O.C.Z. San Pedro, S.M. Gowda, and S.M. Gowda.) affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes. *Journal of Learning Analytics*, 1(1):107–128, 2014.

[26] S. Dawson, D. Gaevifj, G. Siemens, and S. Joksimovic. Current state and future trends: A citation network analysis of the learning analytics field. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge.*, page 231fi??240. ACM, March 2014.

[27] L. S. Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard University Press, 1978.

[28] M. S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[29] Matthew T. Hora with Ross J. Benbow and Amanda K. Oleson. *Beyond the skills gap: Preparing college students for life and work*. Harvard University Press, 2016.

[30] G. Siemens. Learning analytics: Envisioning a research discipline and domain of practice. In *Proceedings of 2nd International Conference on Learning, Analytics, and Knowledge (LAK12)*, 2012.