Predicting Student Success in Math Class Using Random Forests and Support Vector Machines

Kristen Keller

Introduction

Motivation and Objective

Education level is known to be a strong predictor of economic success, yet every year students all across the world drop out of school prematurely. Many of these dropouts can be attributed to students failing classes that they feel unable or unwilling to repeat. It would be helpful if schools could use available information to predict which students would struggle in a class before the class even started. Then they might be able to extend additional resources to those students to help them succeed. The main objective of this project is to develop a model to predict whether a student is going to pass of fail math class based on their covariate values. A secondary objective is to determine which covariates are most useful for predicting whether a student will pass. Machine learning methods such as random forest and support vector machines were used to achieve these goals.

Data source

The data we use in this study were collected on 395 students attending 2 secondary schools in the Alentejo region of Portugal during the 2005-2006 school year.[1] The outcome variable represents a student's grade in math class for the final quarter of the year. In the original data, the outcome variable took on integer values between 1 and 20 (inclusive). For our study, we recoded the outcome variable as an indicator that represents whether the student passed or failed. A score of 10 or more was considered a passing grade and a score of less than 10 was considered failing. There are a total of 32 covariates available in this dataset. Two of these covariates represent previous grades the students received in math class during the first two quarters of the year. These covariates will not be included in the model, but all other covariates will be included. See the appendix for more information on the available covariates. There are no missing values in this dataset.

Results

Summary of Data

Before the data were split into a test and training set, the correlation matrix for continuous variables in the dataset was examined. The variables representing mother's education level and father's education level were strongly correlated, as were the variables representing weekly alcohol consumption and daily alcohol consumption. The variable representing failures was positively correlated with age and negatively correlated with mother's education and father's education. The variable representing study time was negatively correlated with weekly alcohol consumption and daily alcohol consumption. The continuous categorization of the outcome variable was included in the correlation matrix to help visualize the relationship between the continuous covariates and the outcome variable on this scale. It appeared that the outcome variable had the highest correlation with failures, followed by age, mother's education, and father's education (Figure 1).

Figure 1. Correlation (Spearman's) between continuous variables in model.

After the correlation matrix was examined, the data were split into a test data set and a training data set. 80% of the data was allocated to the training set (n=318) and 20% was allocated to the test dataset (n=77). Summary statistics were computed for the continuous variables in the model, as well as the continuous version of the outcome variable. The students in the test set appeared to have more absences on average than the students allocated to the training set. Other than that, no major differences in continuous covariate distribution were observed (Table 1).

Table 1. Summary statistics for continuous variables in test and train dataset

	Test			**Train**		
	Mean	**Median**	**SD**	**Mean**	**Median**	**SD**
age	16.69	17	1.28	16.74	17	1.27
Medu	2.75	3	1.1	2.73	3	1.08
Fedu	2.49	2	1.09	2.66	3	1.07
traveltime	1.48	1	0.73	1.31	1	0.54
studytime	2.04	2	0.85	2.01	2	0.8
failures	0.33	0	0.74	0.35	0	0.77
famrel	3.96	4	0.9	3.9	4	0.87
freetime	3.24	3	0.99	3.22	3	1.05
goout	3.13	3	1.11	3.01	3	1.13
Dalc	1.47	1	0.9	1.51	1	0.84
Walc	2.3	2	1.29	2.26	2	1.29
health	3.52	4	1.4	3.71	4	1.35
absences	5.89	4	8.1	4.97	2	7.59
G1	10.89	11	3.37	10.97	11	3.11
G2	10.73	11	3.79	10.66	10	3.66
G3	10.44	11	4.58	10.3	11	4.62

Summary statistics were also computed for the categorical variables in the test and training set. Specifically, the proportion of students falling in each category was computed for each variable. There was a fairly large difference in the distribution of the covariates family size, parents cohabitation status, school support, and internet across the two datasets. The categorized response variable was also distributed differently across datasets, with there being many more failures in the training dataset then the test dataset (Table 2).

Table 2. Summary statistics for continuous variables in test and train dataset. If there was not an i^{th} response category for a given variable, a value of 0 was displayed.

	Test					Train				
	Cat. 1	**Cat. 2**	**Cat. 3**	**Cat. 4**	**Cat. 5**	**Cat. 1**	**Cat. 2**	**Cat. 3**	**Cat. 4**	**Cat. 5**
school	0.89	0.11	0.00	0.00	0.00	0.87	0.13	0.00	0.00	0.00
sex	0.54	0.46	0.00	0.00	0.00	0.52	0.48	0.00	0.00	0.00
address	0.77	0.23	0.00	0.00	0.00	0.82	0.18	0.00	0.00	0.00
famsize	0.72	0.28	0.00	0.00	0.00	0.31	0.69	0.00	0.00	0.00
Pstatus	0.11	0.89	0.00	0.00	0.00	0.91	0.09	0.00	0.00	0.00
Mjob	0.15	0.08	0.37	0.26	0.14	0.31	0.27	0.1	0.16	0.16
Fjob	0.07	0.55	0.30	0.04	0.05	0.56	0.08	0.08	0.22	0.06
reason	0.37	0.1	0.26	0.27	0.00	0.32	0.36	0.26	0.05	0.00
guardian	0.69	0.23	0.08	0.00	0.00	0.7	0.23	0.06	0.00	0.00
schoolsup	0.12	0.88	0.00	0.00	0.00	0.82	0.18	0.00	0.00	0.00
famsup	0.38	0.62	0.00	0.00	0.00	0.43	0.57	0.00	0.00	0.00
paid	0.55	0.45	0.00	0.00	0.00	0.51	0.49	0.00	0.00	0.00
activities	0.49	0.51	0.00	0.00	0.00	0.51	0.49	0.00	0.00	0.00
nursery	0.81	0.19	0.00	0.00	0.00	0.75	0.25	0.00	0.00	0.00
higher	0.94	0.06	0.00	0.00	0.00	0.97	0.03	0.00	0.00	0.00
internet	0.17	0.83	0.00	0.00	0.00	0.83	0.17	0.00	0.00	0.00
romantic	0.65	0.35	0.00	0.00	0.00	0.73	0.27	0.00	0.00	0.00
G3	0.66	0.34	0.00	0.00	0.00	0.33	0.67	0.00	0.00	0.00

Prediction Using Random Forests

After the data were examined, random forests were used to classify the data. For each tree, 70% of the data was used to build the tree and 30% of the data was set aside for validation. In total 1000 trees were build. In the initial model, mtry was set to 15 and the minimum number of observations in the terminal nodes was set to 1. If the model said a student had a 50% chance of passing or better, that student was classified as a pass. Otherwise, the student was classified as a fail. The out of bag prediction error for this tree was 29.56%. In an attempt to improve upon this value, models were created using a range of different values of mtry and minimum node size. Specifically, 5, 10, 15, 20, and 25 were used for mtry and 1, 3, 5, and 8 were used for node size (Table 3). The models with an mtry of 20 tended to have lower prediction errors than other models. The models with an mtry of 20 and a terminal node size of 1 and 5 preformed best. 1 node was chosen as the optimal parameter because models with this value preformed better than other models across different mtry values.

Table 3. Out of bag prediction error for different options for mtry and maximum terminal node size (** denotes the minimum value; * denote the next 3 lowest values).

	Mtry: 5	**Mtry: 10**	**Mtry: 15**	**Mtry: 20**	**Mtry: 25**
Node size: 1	0.3029	0.2884	0.2925	0.2769*	0.2812
Node size: 3	0.3071	0.2969	0.2778*	0.2829	0.2800
Node size: 5	0.3030	0.2866	0.2855	0.2761**	0.2863
Node size: 8	0.3026	0.2994	0.2786*	0.2806	0.282

The model with an mtry of 20 and a minimum terminal node size of 1 was used to generate predictions for the test data. The sensitivity and specificity values were 0.9020 and 0.4231, respectively (Table 4). This suggests that the model classifies a large percentage of the students as passing. The model classifies 20.78% of students as failing and 79.22% of the students as passing, whereas in reality 33.77% of the students failed and 66.23% of the students passed.

Table 4. Measures of predictive accuracy for random forest with selected mtry and maximum terminal node size. Measures calculated using test data.

	Original Model	**Cutoff: 0.55**	**Cutoff: 0.6**
Prediction Error	25.97	31.17	35.06
Sensitivity	0.9020	0.8039	0.7059
Specificity	0.4231	0.4615	0.5385
PPV	0.7541	0.7455	0.75
NPV	0.6875	0.5455	0.4828

In this setting, it seems possible that we might be more interested in classifying true negatives than true positives correctly. In the previous model, we said that students with a 50% chance of passing would be classified as a pass. In order to increase the specificity of our model, we tested different value for the probability cutoff at which a student went from being classified as a pass to a fail. Specifically, we tested cutoffs ranging from 0.35 to 0.75 in steps of 0.025 (Figure 2). We found that if we increased the cutoff to 55%, the average OOB sensitivity using the training data was 0.8458 and the specificity was 0.5288. Here the specificity increased by 0.144 and sensitivity only decreased by 0.047. If we increased the cutoff to 60%, the average OOB sensitivity was 0.7757 and the specificity was 0.0657. Test data was used to calculate out of sample prediction error and other quantities (Table 4). Our choice of model here would depend on how invested we were in classifying true failures correctly at the expense of misclassifying more true passes.

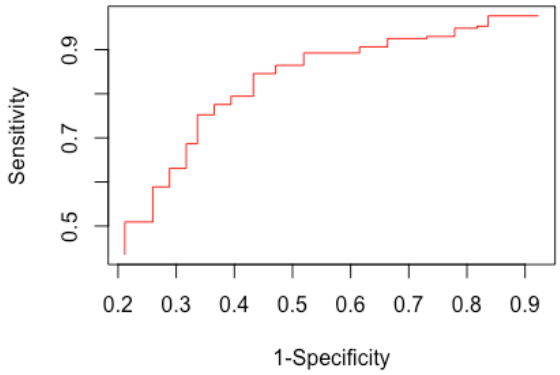

Figure 2. ROC curve generated using different probability cutoffs for random forests classification. Curve generated based on OOB values using training data.

Variable Importance Using Random Forests

Variance importance scores were extracted from the original random forests model that was run. Specifically, mean decrease in the Gini Index was used to distinguish between variables. The variables with the highest scores were absences, failures, goout, Mjob, age, and free time. The variables with the lowest variable importance scores were school, Pstatus, nursery, address, and activities (Table 5).

Table 5. Variable importance scores from original random forests model.

Variable	Variable Importance Score	Variable	Variable Importance Score
absences	11.25	guardian	2.9
failures	10.15	traveltime	2.53
goout	7.94	Dalc	2.25
Mjob	5.57	romantic	1.33
age	5.39	paid	1.11
freetime	4.79	famsup	1.07
health	4.19	higher	1.04
famrel	4.15	sex	0.99
Fedu	4.14	internet	0.99
Walc	4.01	famsize	0.98
studytime	3.67	activities	0.93
reason	3.39	address	0.84
Fjob	3.32	nursery	0.78
Medu	3.24	Pstatus	0.77
schoolsup	2.98	school	0.56

Support Vector Machines

Prediction models were also generated using support vector machines. The following 4 kernels were use:

2^{nd} Degree Polynomial – $(\gamma uv+\beta_0)^2$
3^{rd} Degree polynomial – $(\gamma uv+\beta_0)^3$
Radial Basis – $\exp(-\gamma |u-v|^2)$
Sigmoid – $\tanh(\gamma\, u'v+\beta_0)$

The β_0 terms are coefficient terms that were set to 0. An additional parameter is that was examined was the cost function of misclassifying points. When the cost function is large, a smaller margin hyperplane will be chosen if it does a better job classifying points. When the cost fun is small, a wider margin hyperplane is preferred, even if using a wider margin leads to the misclassification of additional points.

For each kernel choice, a grid search using 5-fold cross validation was used to select the optimal cost and gamma values from the following list: 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{-4}, 2^{1}, 2^{2}. Models were run using the selected parameter values for each kernel. All of the models did a good job classifying students that passed but did not do as well classifying students that failed. The quadratic polynomial and radial model had the lowest out of sample prediction error. The Radial model had the highest specificity out of all of the models, with a value of 0.2692 (Table 6). We would choose the radial kernel model as the best model among the support vector machine models because of the specificity.

Table 6. Prediction error and other measures of prediction accuracy for SVM models constructed using different kernels.

Kernel	Prediction Error	Sensitivity	Specificity	PPV	NPV
Quadratic Polynomial	27.27%	1	0.1923	0.7083	1.0000
Cubic Polynomial	31.17%	0.9216	0.2308	0.7015	0.6000
Radial	27.27%	0.9608	0.2692	0.7206	0.7778
Sigmoid	28.57%	0.9804	0.1923	0.7042	0.8333

Conclusion

In this study we built prediction models using support random forests and support vector machines to predict whether students would pass or fail math class. For all of the models we created, the sensitivity was much higher than the specificity. This is of concern, because low specificity means that our models are not successfully identifying students that are failing

classes. The support vector machine models had lower specificity than the random forest model, so if we had to choose one model we would go with a random forests model. In order to determine the optimal cutoff for classifying passes and fails using the random forests, we would have to speak to someone who is looking to use the model and ask them how many true positives we are willing to misclassify in order to increase our specificity.

The most important variables in our prediction models appeared to be absences, failures, amount of time spent going out with friends, mother's job, and age, followed by amount of free time, health status, quality of family relationships, father's education, and weekly alcohol use.

For future work, our primary interest would be to look into new methods and modifications of our current methods that would help to increase the specificity without majorly decreasing the sensitivity.

Appendix: Variable Descriptions

school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
sex - student's sex (binary: "F" - female or "M" - male)
age - student's age (numeric: from 15 to 22)
address - student's home address type (binary: "U" - urban or "R" - rural)
famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
guardian - student's guardian (nominal: "mother", "father" or "other")
traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures - number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup - extra educational support (binary: yes or no)
famsup - family educational support (binary: yes or no)
paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities - extra-curricular activities (binary: yes or no)
nursery - attended nursery school (binary: yes or no)
higher - wants to take higher education (binary: yes or no)
internet - Internet access at home (binary: yes or no)
romantic - with a romantic relationship (binary: yes or no)
famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime - free time after school (numeric: from 1 - very low to 5 - very high)
goout - going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health - current health status (numeric: from 1 - very bad to 5 - very good)
absences - number of school absences (numeric: from 0 to 93)

Adapted from Cortex and Silva 2008.

References

(1) P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.