Random forest for feature selection in a binary logistic regression of students performance

C Hao, M Ilagan

Department of Mathematics & Statistics
University of Calgary
April 15, 2019

Contents

List of Tables

1 Introduction

School education is a vital channel that allows individuals to acquire basic knowledge and skills, and then apply them consistently to their works and in their lives to fulfill their own goals and ambitions. Therefore, how much an individual would be able to benefit from school education plays an enormous role in his or her lifelong growth. In the field of social science, it is usually of researchers' interest to study what factors would affect the effectiveness of school education.

Machine learning, an alternative to traditional statistics, with its emphasis on prediction rather than explanation, has recently grown in popularity in psychological research (Yarkoni & Westfall, 2017). It is then of interest to compare machine learning techniques to the more traditional approach on performance in modeling educational outcomes. On one hand, a researcher can choose predictors naively—as it has been traditionally done, by personal judgment and understanding of the problem from the literature. On the other hand, the same researcher can choose predictors algorithmically, as done in machine learning studies. Which approach yields a better model? To answer this question, we employ both approaches and compare their results.

In the present study, we use logistic regression to model a binary outcome Y in terms of predictors $X_1, \ldots, X_p$. We build this logistic regression by choosing its predictors both naively and algorithmically. For the algorithmic model, we use the machine learning techniques of random forest and cross-validation. A brief description of these techniques is in order.

Logistic regression. In logistic regression, we fit the data to the best-fitting model of the form

$$Y \sim \text{Bernouilli}\left(\frac{1}{1+\exp(-(\beta_0+\sum_{j=1}^{p}\beta_j x_j))}\right),$$

i.e. the response is a coin flip with the event probability being a nondecreasing function of the linear combination of the features. Iterative numerical methods are used to find the best-fitting values for the coefficients $\beta_0, \beta_1, \ldots, \beta_p$.

Random forest. In constructing a decision tree, recursive binary splitting is used to build a flowchart-like series of decisions that yields interpretable predictions. However, decision trees are known to have high variance. Random forest is a tree-based data mining method proposed by Breiman (2001). In a random forest, bootstrap resampling generates a multitude of trees, thereby reducing variance. This reduction, however, comes at the cost of an interpretable flowchart. In lieu of interpretable predictions, a random forest yields for each predictor a measure of its importance.

Cross-validation. In general, more complicated models yield more predictive power on the dataset they were trained on. However, since we aim for a model with good predictive performance on observations it has not seen, there is a limit to how well one can optimize to the data at hand—beyond this limit, the model loses generalizability. In cross-validation, we find this limit, by partitioning the data into folds. We randomly assign observations to k folds of roughly equal size; and each observation is predicted upon using only the observations not in the same fold. Doing so is a compromise between ensuring that all of the data is used for fitting a model and ensuring that some observations are left out to test a model's robustness.

2 Data and method

We used the Student Performance Data Set (SPDS), a dataset from the UCI Machine Learning Repository. SPDS is regarding student performance in the secondary education of two Portugese schools. Observations are students. Variables are the student's mathematics grade for three grading periods, as well as possibly relevant predictors, such as parents' occupation, time spent studying at home, and interent access at home. To create a binary response variable for logistic regression, we take the average grade over three grading periods, then assign to 1 and 0 by a median split. In other words, we endeavor to predict whether or not a student's aggregate score is above the median of his peers. There are a total of 395 observations and 30 features.

The dataset was partitioned into a training set and test set. Assignment was stratified by response class, such that roughly three-fourths of observations in each class went to the training set; the remainder went to the test set. The resulting training set had 296 observations; the test set had 99 observations.

Using the same training data, two models were constructed independently: one choosing predictors naively, which we call the naive model; and the other, choosing algorithmically, which we call the algorithmic model. As a measure against overfitting, we limit our selection to five predictors.

Towards a naive model, we selected predictors based on our own personal judgment and complemented by social science literature on education outcomes.

Towards an algorithmic model, we did the following procedure on the training set.

1. Grow a random forest, as implemented in R in function `randomForest` in package `randomForest`. Set the number of trees to be 10000; otherwise, default parameters are used.

2. Using the variable importances obtained from the random forest, sort the 30 features by importance.

3. For $p^* = 1, 2, \dots 30$, consider the logistic regression model with the p^* most important features. For each of these 30 models, obtain the 5-fold cross-validated misclassification rate.

4. Out of the 30 logistic regression models considered, choose the one that minimizes the cross-validated misclassification rate.

Having both a naive model and an algorithmic model, we assessed both. On the same test data, we obtained each model's misclassification rate. Finally, we judged the model with the smaller misclassification rate as predictively superior.

3 Results

3.1 The naive model

Based on research conducted by the Haramaya University (Meltem Dayiogule, Aden Kadir Geleto, 2004), we identified 5 variables in SPDS to be the predictors in the logistic regression model.

School. The school a student attend determines the learning environment, resources, qualities of teachers etc., all of which have an impact on the student's academic performance, holding other factors at the individual level, such as intelligence and interest in learning, constant.

Sex. Gender has long been regarded as a significant factor affecting academic performance among students. Social studies have claimed that variations in learning attitudes, personality, teacher's expectation and differential course taking as a result of gender difference have an impact on academic performance. On the other hand, from a biological perspective some researchers have asserted that males have larger average brain sizes than females and therefore, they would be expected to have higher average IQs.

Father's education. There has been increasing studies attributing students' academic successes to their family background. Some studies have indicated that students with more educated father or parent in general would tend to perform better than students having less educated or illiterate parents, due to the fact that families are the one of the most significant places where students learn their knowledge and skills from.

Family support. Same as father's education, this variable is selected based on the belief that the more resources a student's family has, the more likely he or she could be able to learn outside the classroom as a supplement to school education, which has a positive effect on his or her academic performance.

Study time. This one last variable is obviously significant in our subjective thinking since acquiring knowledge is often considered a repetitious process: the more time a person spend on memorizing and using a piece of knowledge, the better he or she would be able to apply it in the exams.

Having such predictors, we formulate the model as

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1\texttt{school} + \beta_2\texttt{sex} + \beta_3\texttt{Fedu} + \beta_4\texttt{studytime} + \beta_5\texttt{famsup}$$

where

`school` is either Gabriel Pereira (GP) or Mousinho da Silveira (MS).

`sex` is student's sex, female (F) or male (M).

`Fedu` is father's education: none (0), primary education (4th grade) (1), 5th to 9th grade (2), secondary education (3) or higher education (4).

`studytime` is weekly study time: less than 2 hours (1), 2 to 5 hours (2), 5 to 10 hours (3), or larger than 10 hours (4).

`famsup` is family educational support: yes or no.

The result is reported below. Thus, the fitted model is

Table. 1. The naive model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.2914	0.4882	-2.65	0.0082
`schoolMS`	-0.1330	0.3813	-0.35	0.7271
`sexM`	0.2300	0.2534	0.91	0.3641
`Fedu`	0.3343	0.1100	3.04	0.0024
`studytime`	0.2577	0.1509	1.71	0.0876
`famsupyes`	-0.3489	0.2532	-1.38	0.1683

$$\log\left(\frac{p(x)}{1-p(x)}\right) = -1.2914 - 0.1330\texttt{school} + 0.2300\texttt{sex} + 0.3343\texttt{Fedu} + 0.2577\texttt{studytime} - 0.3489\texttt{fam}$$

which we interpret as follows.

1. The estimated parameter $\hat{\beta}_1$ predicts an approximate 13.30% decrease in the expected log-odds ratio for a unit increase in the explanatory variable `school`, holding other variables fixed. However, the p-value is $0.7271 > 0.05$ is not significant which means that school is not an influential factor for students' performance.

2. The estimated parameter $\hat{\beta}_2$ predicts an approximate 23.00% increase in the expected log-odds ratio for a unit increase in the explanatory variable `sex`, holding other variables fixed. However, the p-value is $0.3641 > 0.05$ is not significant which means that there is not any difference in students performance at school between boys and girls.

3. The estimated parameter $\hat{\beta}_3$ predicts an approximate 33.43% increase in the expected log-odds ratio for a unit increase in the explanatory variable `Fedu`, holding other variables fixed. Meanwhile, the p-value is $0.0024 < 0.05$ which means that the higher father's educational level, the better grades a children can get.

4. The estimated parameter $\hat{\beta}_4$ predicts an approximate 25.77% increase in the expected log-odds ratio for a unit increase in the explanatory variable `studytime`,

holding other variables fixed. Meanwhile, the p-value is 0.0876 which is marginally insignificant and is close to 5% significant level. Hence, we can get that students are more likely to get higher grades when they spend more time on study every week.

5. The estimated parameter $\hat{\beta}_4$ predicts an approximate 34.89% decrease in the expected log-odds ratio for a unit increase in the explanatory variable `famsup`, holding other variables fixed. However, the p-value is $0.1683 > 0.05$ which not significant. Therefore, we can get that whether students can get family financial support do not have impact on their performance at school.

The naive model has a training misclassification rate of 43.24%.

3.2 The algorithmic model

Random forest and cross-validation yielded the following set of predictors—18 of them—in decreasing order of importance, with their corresponding descriptions.

`absences` number of school absences (numeric: from 0 to 93)

`Mjob` mother'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')

`freetime` free time after school (numeric: from 1 - very low to 5 - very high)

`failures` number of past class failures (numeric: n if 1<=n<3, else 4)

`goout` going out with friends (numeric: from 1 - very low to 5 - very high)

`Fedu` father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

`Fjob` father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

`health` current health status (numeric: from 1 - very bad to 5 - very good)

`Medu` mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

`age` student's age (numeric: from 15 to 22)

`Walc` weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

`studytime` weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

`famrel` quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

`Dalc` workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

`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)

`schoolsup` extra educational support (binary: yes or no)

`famsize` family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

The algorithmic model has a training misclassification rate of 27.70%. The resulting logistic regression model is reported below.

4 Comparison

Having both models, we are in a position to compare the two approaches. Test misclassification rate is 44.44% for the naive model and 32.32% for the algorithmic model. Confusion tables are reported below.

Overall, the algorithmic model is superior. Four notes are in order. First, the two models intersect on only two predictors, Father's education (`Fedu`) and study time (`studytime`)—the remaining naive predictors are not even in the top 18 predictors. Second, very few of the regression coefficients of the algorithmic model are significantly different from zero. Thus, filtering based on p-values would not have been effective in finding a good model. Third, while machine learning is typically associated with less interpretability, because both approaches use the same family of regression models in the present study,

Table. 2. The naive model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.2900	2.4952	0.12	0.9075
absences	-0.0067	0.0164	-0.41	0.6805
Mjobhealth	0.6996	0.7883	0.89	0.3748
Mjobother	0.4202	0.4735	0.89	0.3748
Mjobservices	1.0617	0.5290	2.01	0.0448
Mjobteacher	-0.2462	0.6630	-0.37	0.7104
reasonhome	0.3502	0.3491	1.00	0.3159
reasonother	0.3662	0.4992	0.73	0.4632
reasonreputation	0.7308	0.3794	1.93	0.0541
freetime	0.1733	0.1551	1.12	0.2637
failures	-1.3212	0.3305	-4.00	0.0001
goout	-0.1814	0.1483	-1.22	0.2214
Fedu	0.1956	0.1698	1.15	0.2495
Fjobhealth	1.0556	1.2309	0.86	0.3911
Fjobother	-0.6956	0.6963	-1.00	0.3178
Fjobservices	-0.7762	0.7160	-1.08	0.2783
Fjobteacher	-0.1139	0.8760	-0.13	0.8965
health	-0.0593	0.1035	-0.57	0.5667
Medu	0.1548	0.2152	0.72	0.4720
age	0.0784	0.1288	0.61	0.5431
Walc	-0.1915	0.1610	-1.19	0.2345
studytime	-0.0215	0.1782	-0.12	0.9042
famrel	-0.4014	0.1752	-2.29	0.0220
Dalc	0.1202	0.2316	0.52	0.6036
traveltime	-0.1319	0.2169	-0.61	0.5431
schoolsupyes	-1.7094	0.4760	-3.59	0.0003
famsizeLE3	0.5335	0.3183	1.68	0.0937

the two resulting models in fact have similar interpretation. Thus, with a machine learning approach, we were able to reap the benefit of added predictive power without compromising on interpretability. Finally, to the credit of the naive model, it had a smaller shrinkage of accuracy going from training set to test set.

5 Conclusion

In conclusion, we found that a machine learning approach can greatly benefit research on educational outcomes. In the present study, we saw that a machine learning approach can enhance a model's predictive power even with little compromise on interpretability. We encourage education researchers to continue to use machine learning

Table. 3. Confusion table for naive model

Actual value Predicted value	0	1
0	0.2828283	0.2323232
1	0.2121212	0.2727273

Table. 4. Confusion table for algorithmic model

Actual value Predicted value	0	1
0	0.3030303	0.2121212
1	0.1111111	0.3737374

approaches, so as to be able to leverage big data.

6 References

1.Breiman, L. (2001) Random forests. *Machine Learning, 45,* 05–32.
`https://doi.org/10.1023/A:1010933404324`

2.Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122.
`https://doi.org/10.1177/1745691617693393`

3.Cortez, P. & Silva, A. (2018). 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), 5–12.
`http://archive.ics.uci.edu/ml/datasets/Student+Performance`

4.Essays, UK. (November 2018). Factors Which Influence The Students Academic Performance.
`https://www.ukessays.com/essays/education/factors-which-influence-the-students-academic-performance-education-essay.php?vref=1`