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

C.Hao, M.Ilagan

April 5, 2019

Outline

- Introduction
- Analysis
- Comparison
- Conclusion
- References

Introduction

- Under what circumstances are school education most effective
- Indicator: academic performance measured by grades
- Focus in social science and psychology research
- Traditional statistics: personal judgement and literature review in predictor selection
- Machine learning: through algorithms
- Question: which way is better? Let's compare!

Data and methods

- Student Performance Data Set (SPDS) from the UCI Machine Learning Repository
- On secondary education of two Portugese schools.
- Variables are the student's mathematics grade for three grading periods
- Relevant predictors e.g. parents' occupation, time spent studying at home, and internet access at home

Model with traditional way v.s. machine learning techniques

- Binary response: first take the average grade over 3 periods, then assign 1 and 0 respectively to the upper and lower half split by median average score
- 395 observations and 30 features
- Dataset was partitioned into a training set and test set
- Assignment was stratified by response class (about 3/4 observations in each class went to the training set; 1/4 to the test set)
- Training set had 296 observations; test set had 99
- Same training data for 2 models, constructed independently
- Aim at 5 predictors to avoid overfitting

Logistic model with naive variable selection

Based on our own personal judgment and literature review

- School
- Sex
- Father's education
- Family support
- Study time

Model formulation

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

school - student's school: Gabriel Pereira(GP) or Mousinho da Silveira(MS).
sex - student's sex: female(F) or male(M).
Fedu - father's education: none(0), primary education (4th grade)(1), 5th to 9th grade(2), secondary education(3) or higher education(4).
studytime - 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 - family educational support: yes or no.

	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

$$\begin{aligned}\ln\left(\frac{p(x)}{1-p(x)}\right) = &-1.2914 - 0.1330 school + 0.2300 sex \\ &+ 0.3343 Fedu + 0.2577 studytime - 0.3489 famsup\end{aligned}$$

Analysis

Table: Predicted result

Predicted value \ Actual value	False	True
False	0.283	0.232
True	0.212	0.273

The accuracy is 55.56%.

Choosing algorithmically

Procedure for algorithmic model:

1. Grow random forest of 10000 trees
2. Using variable importances from random forest, sort features by importance
3. For $p^* = 1, 2, \ldots 30$, consider logistic regression model with the p^* most important features
4. Select model (out of the 30) with lowest 5-fold cross-validated residual

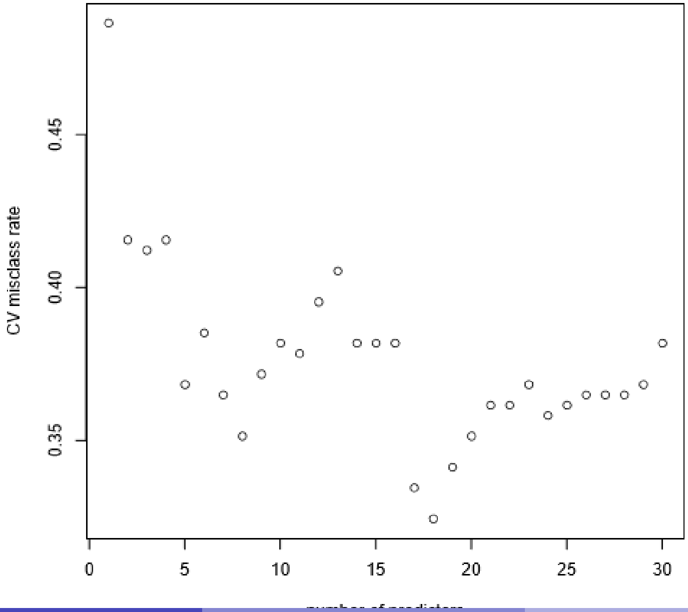
CV misclass rate
0.35
0.40
0.45
0
5
10
15
20
25
30
number of predictors

Table: Variable importance ranks and cross-validated misclassification rate

Rank	Feature	CV misclass	Rank	Feature	CV misclass
1	absences	0.4864865	10	Medu	0.3817568
2	Mjob	0.4155405	11	age	0.3783784
3	reason	0.4121622	12	Walc	0.3952703
4	freetime	0.4155405	13	studytime	0.4054054
5	failures	0.3682432	14	famrel	0.3817568
6	goout	0.3851351	15	Dalc	0.3817568
7	Fedu	0.3648649	16	traveltime	0.3817568
8	Fjob	0.3513514	17	schoolsup	0.3344595
9	health	0.3716216	18	famsize	0.3243243

Table: Predicted result

Predicted value \ Actual value	False	True
False	0.3030	0.2121
True	0.1111	0.3737

Test accuracy is 67.68%.

Comparison

- We compare the Logistic regression method and Random forest method. Test misclassification rate is 44.44% and 32.32% respectively. Confusion tables are reported below.

Table: Confusion table for naive model

Predicted value \ Actual value	0	1
0	0.283	0.232
1	0.212	0.273

Table: Confusion table for algorithmic model

Predicted value \ Actual value	0	1
0	0.303	0.212
1	0.111	0.374

Comparison

- Overall, the algorithmic model is superior.
- Two models intersect on two predictors, (`Fedu`)(`studytime`), the remaining naive predictors are not in the 18 predictors.
- 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.
- Machine learning is typically associated with less interpretability, because both approaches use the same family of regression models in our study, the two resulting models in fact have similar interpretation.

Conclusion

- A machine learning approach can greatly benefit research on educational outcomes.
- 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 approaches, so as to be able to leverage big data.

References

- Breiman, L. (2001). Random forests. *Machine Learning*, *45*,0 5–32.
- Yarkoni, T., & Westfall, J. (2017). Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.
- 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.