Student Performance Evaluation using Naïve Bayes and Random Forests

INM431 Machine Learning coursework – Daniel Sikar - daniel.sikar@city.ac.uk

Description and motivation

Studies in Europe and the USA have shown the fiscal and social costs of early school leaving, which typically reduces lifetime earnings and leads to higher unemployment incidence and duration, while the benefits of education include gain in net earnings and wealth, improved health status / life expectancy, lower reliance on government health and welfare programmes and reduced expenditures on criminal justice [1][2][3].

One important problem is to automatically detect students that are going to do poorly in a course early enough to be able to take remedial actions [4]. A number of Machine Learning algorithms have been used in student grade prediction. We base our study on prior work [5], using Naïve Bayes and Random Forests applied to secondary school student performance data.

Initial analysis of the data set including basic statistics

The data sets being investigated, made publicly available at the UCI Machine Learning Repository[6], were collected by using school reports and questionnaires, at two Portuguese secondary schools and refers to subjects of Mathematics (student-mat.csv) and Portuguese (student-por.csv). Both data sets contain the same column headers and were combined for the purpose of our analysis (student-labelled.csv).

The combined data set has 1044 observations and 33 attributes, consisting of 4 nominal, 13 binary and 16 numeric attributes, three numeric attributes being grades (G1, G2 and G3) following a 20-point grading scale. Attribute G3 contains the final grade. A "Result" binary attribute column was added with a value of 0 for fails (G3 less than 10) and a value of 1 for passes (G3 greater or equal to 10). Based on the "Result" attribute, the dataset was found to be unbalanced, with 78% passes and 22% fails, which shall be considered when evaluating predictive model performance.

In Fig. 1, the scatter plots show that there is strong correlation between the prior (G1, G2) and final (G3) grades and also between prior grades. The plots also show that there are outliers: data points where the final grade is zero. The data source does not explain why these grades are zero - if we assume the most likely explanation is that the student did not take the exam we could consider eliminating these from the analysis or perform analyses with and without these points.

The histograms imply that Final Grades are higher for students with internet access, no romantic relationship and extra-curricular activities. Statistical means are higher for these subsets.

The box plots suggest that students with some free time (freetime) and moderate amounts of alcohol intake during weekdays (Dalc) and weekends (Walc) have higher mean grades. According to the data source[5], attribute "Health" is described as "current health status (numeric: from 1 – very bad to 5 – very good)". Given the "health" by final grade "G3" box plot shows that students in poorest health states have achieved highest mean grades, the described gradient is assumed to be inverted.

Additional data analysis box plots (not shown in Fig. 1), plotting "quality of family relationships (numeric: from 1 – very bad to 5 – excellent) versus final grade, show students with better family relationships have higher mean grades, the same being the case for students with parents with higher levels of education (numeric attributes Medu and Fedu).

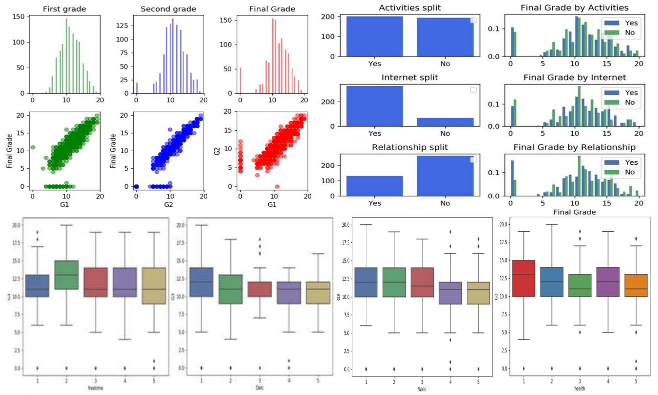

Fig. 1

Two models with their pros and cons

Naïve Bayes

Random Forests

Hypothesis statement

Description of choice of training and evaluation methodology

Choice of parameters and experimental results

Naïve Bayes

Random Forests

IMAGE HOLDER

Analysis and critical evaluation of results

IMAGE HOLDER

Lessons learned and future work

References
[1] Belfield, C. (2008) Cost of Early School-Leaving and School Failure (p. 48). New York: Economics department, City University of New York
[2] Brunello, G. & Paola, M.D. 2014, "The costs of early school leaving in Europe", IZA Journal of Labor Policy, vol. 3, no. 1, pp. 1-31.
[3] Fernandez-Gutierrez, M. & Martinez, J. 2014, "THE NON-MONETARY COSTS OF EARLY SCHOOL LEAVING: ESTIMATION IN TERMS OF YEARS OF GOOD HEALTH", EDUCACION XX1, vol. 17, no. 2, pp. 241-263.
[4] Meier, Y., Xu, J., Atan, O. & van der Schaar, M. 2016;2015;, "Predicting Grades", IEEE Transactions on Signal Processing, vol. 64, no. 4, pp. 959-972.
[5] 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.
[6] https://archive.ics.uci.edu/ml/datasets/Student+Performances Retrieved 23.11.2019