DS 740 Final

Paul Isaacson

December 13, 2017

What factors make a secondary school student successful? Knowing the factors, can we identify students who are at risk? After identifying students who are at risk, can we help them? How universal are the risk factors? Can they be applied internationally?

To explore these questions, the author looked at a dataset of 649 secondary students from Portugal. This dataset showed grade achievement against several demographic, social/emotional and school-related categories. The data was collected between 2005 and 2006 from two sources including reports from the school as well as in-class questionnaires.

This data set was geographically isolated to Portugal. As such, the author took care not to extrapolate the results beyond Portugal. Even so, entities such as the Portugal government would be very interested in this information.

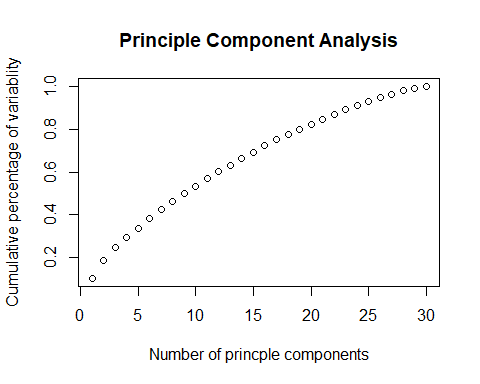
There is evidence that the Portugal government is placing greater importance on academic performance. From 1973 to 2014, the Portugal government increased expenditure on education as a percent of GDP from 1.6% to 5.1% (1). This would strongly suggest that the government is taking education improvement seriously. As part of this, the Portugal government needs strong analysis on what factors affect academic performance. The author believes the analysis found from this study can help.

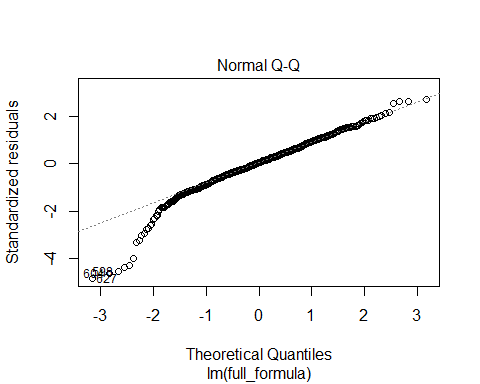
The data set contained a rich set of predictor variables to explore. The author conducted a rather robust exploration of which variables are important. First, principle component analysis was used. However, to explain over 90% of the variability of the data, 24 variables were needed (Figure 1). Next, backward step selection comparing models AIC was conducted. The model with the best results had 15 predictor variables. The subset of variables included sex, age, address, father’s job, reason to choose school, weekly study time, past class failures, extra school support, desire to take higher education, having access to internet at home, being in a romantic relationship, quality of family relationships, how often they go out with friends, workday alcohol consumption, and current health. The reader should note that the particular variables selected are correlated with success but does not constitute the entire list of variables of importance.

Both the full set of variables as well as the reduced set were used in modeling the prediction of final grades. Multiple regression (lm), support vector machine (svm), and artificial neural network (ann) were all explored. The lm was used as a baseline to compare the svm and ann. The qqplot for the lm model shows that lm is a reasonable model (Figure 2). Both svm and ann are appropriate for regression problems. Results from CV show all results produce similar scores with a tuned svm having the best CV score

Perhaps most actionable in this list is having access to internet at home. To the author, this is low-hanging fruit. As such, the Portugal government may do well to ensure the student body has access to internet at home through assistance programs. Additionally, several factors such as romantic relationships and quality of family relationships, suggest social/emotional factors are important. With this in mind, guidance counselors should take care to check in with students to find how the student’s relationships are going.

The author would like to see these same in-class questionnaires given to students from other countries. The purpose would be see how universal the factors are. Additionally, using many modern modeling techniques, the author was able to produce a model that fairly accurately predicts a students grade based on the factors described above. The author suggests schools would do well to have students take the in-class questionnaire in the beginning of the year. Immediately after, the model could help identify students who may be at risk. Efforts from school staff, most likely already stretched thin, could be applied to the greatest at-risk population in a pro-active manner.

*Figure 1. PCA suggests that the variability of the data cannot be reduced to a few variables.* 

*Figure 2. Qqplot of linear regression using all the variables shows resonable but left-skewed results* 

References

1. World Bank. (n.d.). Government expenditure on education, total (% of GDP). Retrieved from <https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS?locations=PT>.