Performance Prediction of High School Student using Machine Learning Tools

John AOUSSOU
Originally submitted as final project for the 15.077 course
(Statistical Learning and Data Mining)
MIT, Spring 2015

Introduction

Amongst countries of the European Union, Portugal is affected by a relatively higher number of students who fail to pursue higher education, or altogether drop out of school. Data was collected during the 2005 – 2006 school year in two high schools of the Alentejo region of Portugal by the Government so as to identify factors that are key to academic success. The data collected included the grades obtained for the three terms of the school year in two subjects, namely mathematics and Portuguese. Those are two fundamental subjects that allow success in other subjects. Information regarding their socio-economic background and personal habits was also gathered.

The objective of this project is primarily to identify what factors are crucial to success in high school. Models that help predict the average performance over the three terms are also developed. Performance is evaluated using numerical grades, grade classification as well as a pass or fail criterion. The data set is relatively small, and variation between results using different training and test sets can be quite important. Similar models to predict the median were also built, however they did not show any significant difference in the trends and therefore are not presented in the report for concision.

A paper addressing the issue using the same data set was produced in 2008 by Cortez and Silva and was attached to this report. Major differences between their work and this work is that they evaluate performance based on the grade obtained for the third period only, whereas here the predicted variable is the average. Moreover, in this work, in addition to models that try to predict the numerical grade, the grade class and pass or fail, models are also built that can identify students who are particularly at risk. Also, this work only takes into consideration grades obtained in mathematics.

The software used to develop those model is R. The training sample constitutes 80% of the data and the test sample 20% of it. Each simulation is run using

seed(1) so as to ensure that the exact same sets are compared. When trying to identify important factors, runs with seeds 2, 3, 4 and 5 are also performed to check for consistency. Each model is tuned for the best parameters using cross validation, such as coefficients for regression methods, number of neighbors for the nearest neighbor method, decay and number of nodes per layer in Neural Networks, and best pruning parameters for classification trees.

When making predictions, higher importance is given to identify students who are likely to fail. Therefore, for all models built to determine pass or fail, the overall misclassification is less relevant than the confusion matrix.

I Description of the data

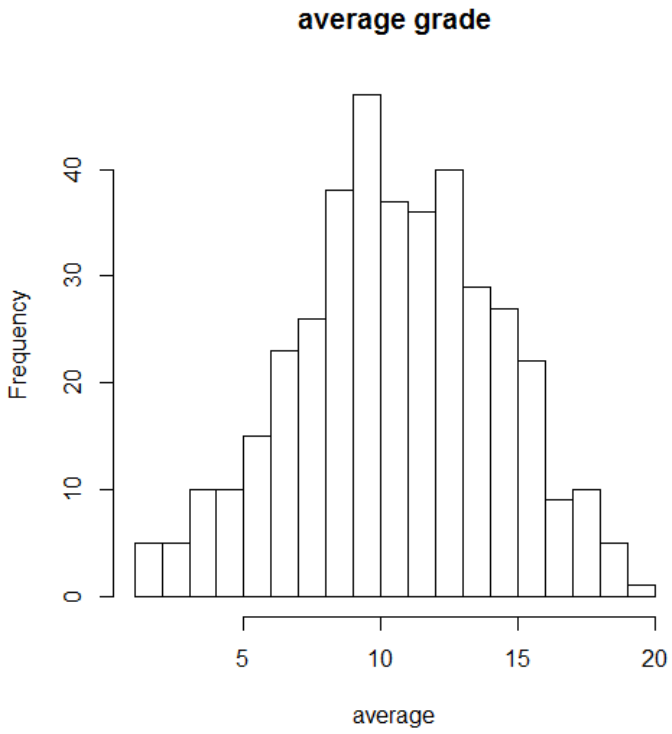
The data consists in thirty-three variables. Data regarding the socio-economic background of students include their parent's education and occupation, and whether they live in a rural or urban environment. Data regarding their personal habits include their daily study time and alcohol consumption during the week and during the weekend. Data regarding their motivation was also collected, such as the reason why they chose their school and whether they hope to pursue higher education. A list of those variables along with their labels are described in more detail in the appendix. Over half of the predictive variables are categorical, and as such have to be converted to dummy variables.

The grading scale is out of 20 marks. The data looks roughly normally distributed, and passes the Anderson – Darling and Shapiro – Wilk tests, while it fails the Kolmogoroff – Smirnoff test. Since the sample is small, it is not possible to draw any clear conclusion.

The correlation matrix reveal interesting facts about the relationship between some variables, for example that girls study longer hours and that people are more likely to marry someone with the same educational level. However, no variables are strongly correlated. Similar trends can be observed for the mean. The highest correlation between average grades and the predictive variables is that with the number of past failures and the parents' educational level. Correlation between the predictive variables and the average grade is shown in the appendix.

A Principal Component Analysis shows that the two first eigenvectors account for only 8.8% and 6.8% percent of the variability. Also, when projecting on to the space form by those two vectors, no clear pattern can be distinguished.

The data has already been preprocessed, and no further work needs to be done in that regard.



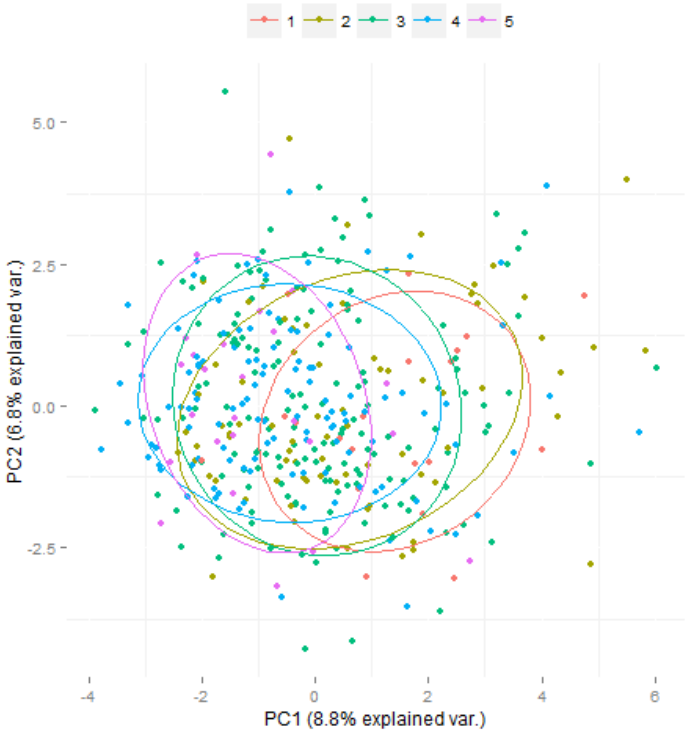
(a) mean: 10.68 - median: 10.67 - standard deviation: 3.70

II Linear Regression models and nearest neighbors

II.1 Exact grade prediction

The Ordinary Least Squares, Ridge and LASSO regression models as well as the k-nearest neighbor methods are applied to try and predict the exact numerical grades. For all models, the RMSE of the train test is around 3 while that of the test case is above 3.5. The list of significance and the standard error values of the Ordinary Least Square coefficients is attached in the appendix.

The coefficients obtained with seed(1) are compared to those obtained with the other seeds. The number of previous failures was universally classed as highly significant, which is coherent with the observations from the correlation matrix. Interestingly, the father being a teacher, school support, the time spent outside, the time spent studying and the gender are systematically classified as significant which was not particularly obvious from the correlation matrix. The corresponding LASSO coefficients are non-zero, which is consistent.



(a) principal component analysis

According to the OLS model, a girl whose father is a teacher, who has never failed and benefits from school support starts with an advantage of five marks out of twenty or more.

II.2 Classification

While the models built so far help identify interesting factors, they are not accurate enough to predict an exact grade. Moreover, since the mean of the predicted value is close to that of the actual values, they do not provide much information about whether the model is more likely to overestimate or underestimate a student's grade. The problem can thus be converted into a classification problem.

Classes are defined by bands of 4 marks out of the 20-mark scale (see appendix for detailed description). As determined above, the RMSE is of 3.5 or more, which is almost the difference between two classes, so it is not obvious whether

classification will help perform better predictions.

The confusion matrices are attached in the appendix. It seems as if RIDGE and LASSO are better at predicting whether someone is around the average, whereas OLS is arguably slightly better at predicting classes away from the average. This could be related to the fact that OLS is more sensitive to outliers.

The k-nearest neighbor methods performs particularly poorly. The projection plot of the two largest PCA component in section I. shows that there is no strong clustering pattern, so this could explain why.

II.3 Pass and fail

All four algorithms seem to be passing more people than expected, with k-NN being the most generous and OLS being less so.

III Classification algorithms

III.1 Classification

Results are shown in part IV of the appendix

a. Classification Tree

The minimum number of observations in a leaf node was varied. The tree below was produced with a minimum of 10 observations. While there is a high variability regarding the exact factors and their importance for each simulation, as before, it appears that the number of past failures is the primary factor, and that school support is also important. A recurrent variable that did not manifest itself in the linear regression analysis but that seems to be clearly of importance is the number of absences.

Interestingly, it now also appears that the job of the mother is influential: a student whose mother is a teacher or stays at home, as opposed to being a health or civil services worker has more hope to be in the 4thclass than in lower classes.

Overall, the method performs reasonably well, especially when it comes to classify students with low grades.

b. Random forest

The plots of the Mean Decrease in Accuracy and the Mean Decrease in Gini were obtained using all the data and not only the train set.

The results confirm what was inferred from observing the classification tree, i.e. that the number of failures and school support are important factors. In particular they strongly confirm that the number of absences is crucial.

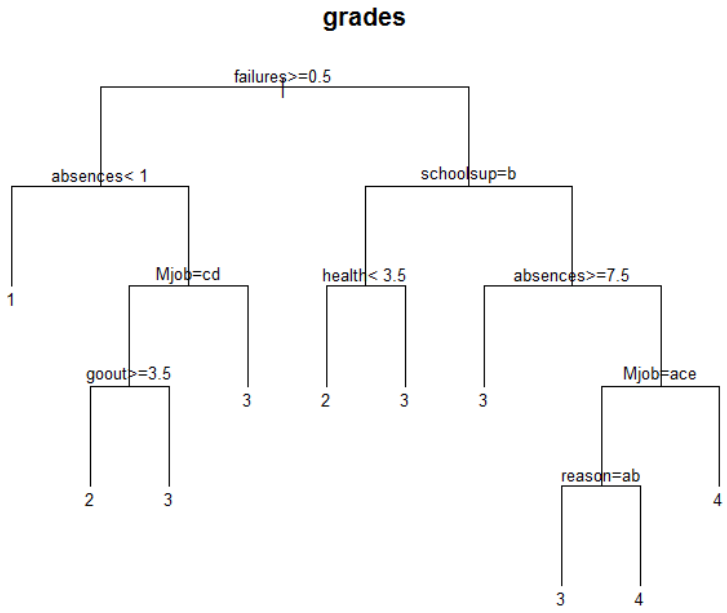


Figure 3: classification tree - seed(1)

Interestingly, the presence of some variable negatively impact accuracy, such as nursery to decrease accuracy, as if causing overfitting.

c. Other methods

Naïve Bayes performs overall poorly, possibly because of interdependence of many of the predictive variable. However, it performs surprisingly well to predict that a student will fall in the lowest range, and that systematically. The SVM method performs poorly, perhaps because of the presence of irrelevant inputs (Hastie et al., 2009) The multinomial logit method also performs poorly.

III.2 Pass and fail

All method perform quite poorly, especially the SVM, Random Forest and Neural Network methods. The logistic binomial and Naïve Bayes methods perform slightly better, and can be compared to the results obtained using regression methods. They tend to be more accurate at predicting that a student will fail and less accurate at predicting that they will pass.

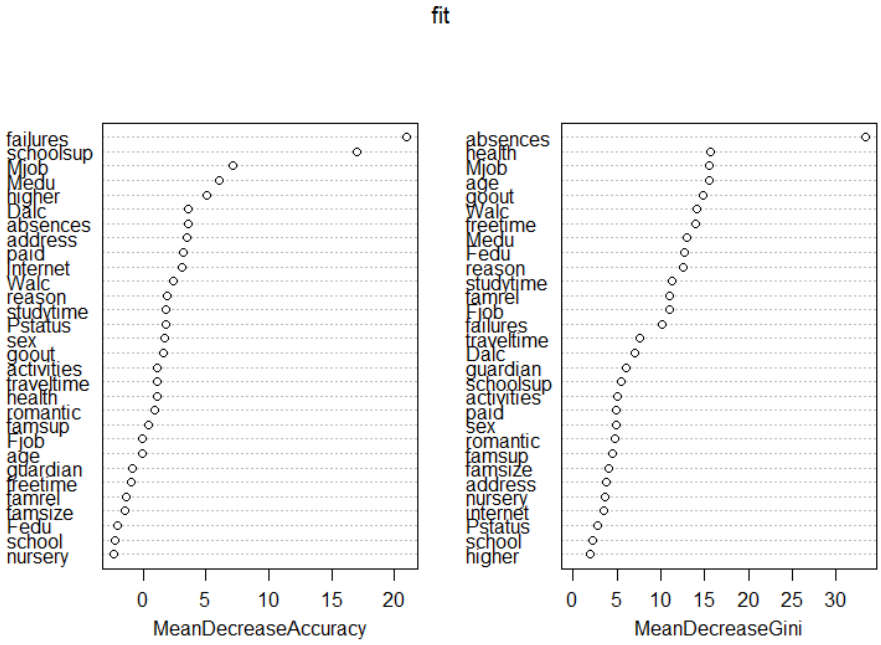


Figure 4: mean decrease in accuracy and in Gini

IV Predicting that a student is likely to drop out of school

IV.1 Description of the problem and model design

The lack of accuracy of the models tried so far could be due to the high concentration of grades around the mean. The effect of that concentration might be particularly acute when classifying a grade, and this effect is possibly even more pronounced when predicting pass and fail, since the cutoff grade is 10 which very close to the mean and the median.

When observing the data closely, it appears that an unexpected number of students obtain a grade of zero for the final period. In fact, while no student obtained a grade of zero for the first term, it was the case of 13 out 395 (3.29%) and 38 out of 395 students (9.62%) for the second and third term respectively. Also worth noting is that for both terms, the next higher grade is 4. In other words, no student got a grade of 1, 2 or 3. A closer observation also reveals that all students who got zero for the second term also got zero for the third term.

This gap and the high number of students getting a zero suggest, that, possibly, those students simply did not sit for the exam or perhaps just decided not to study for it. One can only speculate about potential reasons for this trend, nev-

ertheless, it seems reasonable to assume to those students represent a minority within the group, and that this minority is particularly at risk. They will from then on be referred to as "students at risk".

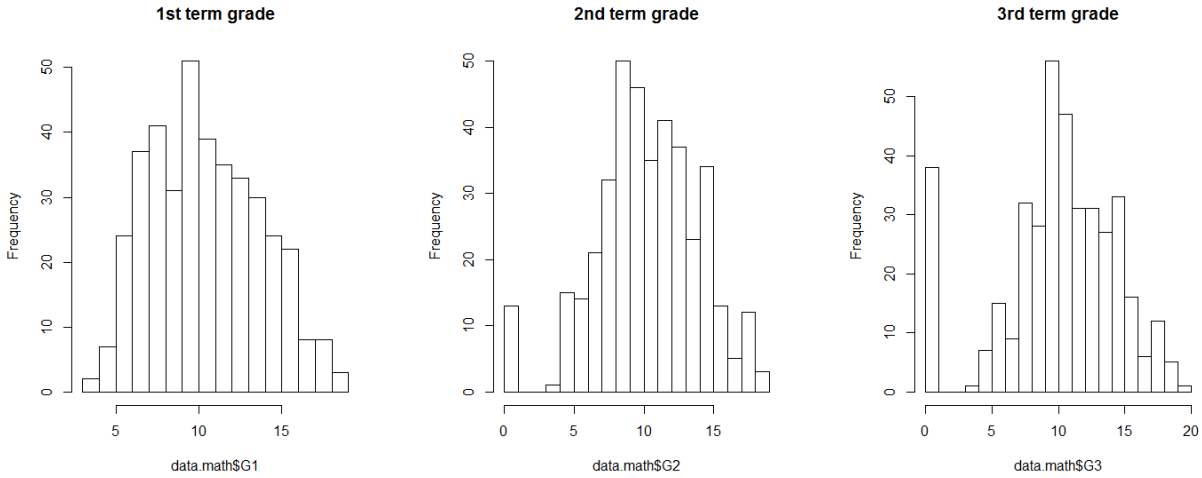


Figure 5: grades by term

The same models as those used the previous sections are used to try to identify students at risk. Random Forest and SVM completely fail to identify any student at risk. The Classification Tree methods performs decently, while Naïve Bayes almost systematically identifies all students at risk, but also predicts that mainly students are at risk when they are not. k-NN performs poorly, regression methods perform reasonably well, with RIDGE identifying all students at risk but also falsely classifying as being at risk students who are not. OLS is more conservative (higher precision, smaller recall) and LASSO is in the middle. The multinomial logit and Neural Network methods are very effective and the next section will attempt to decide between them.

IV.2 Comparison between the Multinomial and Neural Network models

The Multinomial and Neural Network methods appear to be the models that perform best and to further determine which of the two models is the best is an interesting exercise. One difficulty that arises when using such models is to find the optimal threshold. One way of doing this is to find the best trade-off between True Positive and False Positives (in this case misclassifying a student that is not at risk as being so) (Chawla, 2005). This can be achieved by finding the threshold where the ROC curve is closest to the upper left corner. The area under the ROC curve is also a common measure of the performance of a model.

The closer the value is to one, the stronger the discriminative power of the test.

Nevertheless, the problem is class unbalanced. Since the objective is to identify students at risk, the error cost of failing to identify one of those students (False Negative) is higher than misclassifying at student that is not at risk as being so (False Positive). In that case, a better measure would be the area under the Precision-Recall curve (Davis and Goadrich, 2006). For a PR curve, the closer the curve is to the upper right corner, the better. Also the closer the value of the area under the curve is to one, the better. Other measures can be used, such as the F1 – score, which is one possible indicator of the trade- off between True Positive, False Positive and False Negative.

The table below shows the results obtained for seed(1) and those for four different seeds are attached in the appendix. They reveal that the Neural Network model systematically performs better in terms of ROC and PR area under the curve.

Table 1: seed(1)

method	**binomial**			**neural network**		
Confusion Matrix		0	1		0	1
	0	**68**	**2**	0	**64**	**1**
	1	**2**	**7**	1	**6**	**8**
ROC AUC			.96			.97
PR AUC			.66			.75
precision			.78			.57
recall			.78			.89
F-value			.78			.70

A Student at risk identification - binomial vs Neural Network

Table 2: seed(2)

method	**binomial**			**neural network**		
Confusion Matrix		0	1		0	1
	0	**65**	**4**	0	**58**	**1**
	1	**4**	**6**	1	**11**	**9**
ROC AUC			.89			.91
PR AUC			.25			.47
precision			.6			.55
recall			.6			.9
F-value			.6			.6

Table 3: seed(3)

method	**binomial**			**neural network**		
Confusion Matrix		0	1		0	1
	0	**66**	**3**	0	**64**	**2**
	1	**4**	**6**	1	**6**	**7**
ROC AUC			.91			.93
PR AUC			.48			.63
precision			.6			.54
recall			.67			.78
F-value			.63			.64

Table 4: seed(4)

method	**binomial**			**neural network**		
Confusion Matrix		0	1		0	1
	0	**64**	**3**	0	**59**	**1**
	1	**6**	**6**	1	**11**	**8**
ROC AUC			.91			.93
PR AUC			.5			.42
precision			.30			.64
recall			.67			.89
F-value			.57			.57

Table 5: seed(5)

method	**binomial**			**neural network**		
Confusion Matrix		0	1		0	1
	0	**64**	**3**	0	**59**	**1**
	1	**6**	**6**	1	**11**	**8**
ROC AUC			.76			.87
PR AUC			.15			.31
precision			.25			.2
recall			.67			1
F-value			.36			.36

variable	correlation with grade
school	0.0432849523
sex	-0.1011220047
age	-0.1345893739
address	-0.107297478
famsize	0.0825602049
Pstatus	0.0430475813
Medu	0.2242598684
Fedu	0.1758521351
traveltime	-0.1281971628
studytime	0.1345647189
failures	-0.3757588959
schoolsup	-0.1376435544
famsup	-0.0615533691
paid	0.0895107463
activities	0.040858614
nursery	0.0651451589
higher	0.1894834951
internet	0.1026280835
romantic	-0.1027310828
famrel	0.0216525209
freetime	0.0037731403
goout	-0.1545113365
Dalc	-0.0725081775
Walc	-0.0880246711
health	-0.080380376
absences	-0.0059088061
Mteacher	0.066383231
Mservices	0.0849021135
Mhealth	0.1293340952
Mhome	-0.1040323869
Fteacher	0.1225968066
Fservices	-0.0016348988
Fhealth	0.0474734617
Fhome	0.0085906926
reason_course	-0.0993454672
reason_home	-0.0123433275
reason_rep	0.0988264999
g_mother	0.0008078045
g_father	0.045000838

variable	**estimate**	**std. error**	t **value**	$Pr(>\|t\|)$	**relevance**
(Intercept)	13.664992	4.385539	3.116	0.002027	**
school	-0.123402	0.702289	-0.176	0.860647	
sex	-0.927797	0.443516	-2.092	0.03736	*
age	-0.238607	0.201814	-1.182	0.238099	
address	-0.045632	0.533277	-0.086	0.93187	
famsize	0.946456	0.428521	2.209	0.02802	*
Pstatus	0.260897	0.6452	0.404	0.686257	
Medu	0.083987	0.294964	0.285	0.776059	
Fedu	0.071776	0.25209	0.285	0.77607	
traveltime	-0.270258	0.312125	-0.866	0.387316	
studytime	0.687527	0.256459	2.681	0.007786	**
failures	-1.2711	0.334952	-3.795	0.000182	***
schoolsup	-2.155644	0.618672	-3.484	0.000574	***
famsup	-0.714101	0.434309	-1.644	0.101268	
paid	0.090716	0.433401	0.209	0.834359	
activities	-0.268706	0.396752	-0.677	0.498805	
nursery	0.067096	0.502822	0.133	0.893943	
higher	1.12656	0.970685	1.161	0.246815	
internet	0.797706	0.560192	1.424	0.15558	
romantic	-0.736477	0.42087	-1.75	0.081249	0
famrel	0.008743	0.213899	0.041	0.967424	
freetime	0.144771	0.211756	0.684	0.494758	
goout	-0.353205	0.204113	-1.73	0.08467	0
Dalc	0.033438	0.29886	0.112	0.910997	
Walc	-0.027891	0.225071	-0.124	0.901468	
health	-0.236491	0.146057	-1.619	0.106552	
absences	0.023604	0.026911	0.877	0.381185	
Mteacher	-0.241687	0.753156	-0.321	0.74853	
Mservices	1.609732	0.533106	3.02	0.002769	**
Mhealth	1.43355	0.834101	1.719	0.086794	0
Mhome	1.137286	0.665775	1.708	0.088721	0
Fteacher	2.554004	0.819175	3.118	0.002015	**
Fservices	0.150162	0.487953	0.308	0.758513	
Fhealth	1.295869	1.001872	1.293	0.196938	
Fhome	0.400254	0.978835	0.409	0.682924	
reason_course	-0.574665	0.727412	-0.79	0.430198	
reason_home	-0.141508	0.754481	-0.188	0.851362	
reason_rep	0.249587	0.767304	0.325	0.745217	
g_mother	-0.038056	0.793055	-0.048	0.961761	
g_father	0.465149	0.876628	0.531	0.596115	