Analysing Student Performance - Portuguese Language Class

smck0583, mcza2974, ttie8223, dpha3771, kmao8943

Git Repository - https://github.sydney.edu.au/smck0583/CC05E4

List of packages used: ggplot2Wickham (2016) dplyr olsrr pinp tinytex(Xie, 2022) tidyverse(Wickham *et al.*, 2019) stargazer(Hlavac, 2022) Data source: Ünal (2020)

Abstract. This report investigates the effect of various attributes on final grade in Portuguese language class from data of 649 secondary students in Portugal, and aims to use multiple linear regression to produce a model that closely predicts final grade in Portuguese for students in Portuguese public secondary schools. The dataset contains 32 attributes school-related, demographic, lifestyle and socioeconomic factors. The final model was chosen by AIC in a stepwise algorithm and has an RMSE of 0.84 and an R-squared value of 0.90. The predictor variables are first period grade, second period grade, student age, number of previous class failures, degree of workday alcohol consumption, desire to pursue higher education, presence of extra educational support, health status, and degree of going out with friends.

Introduction. Portuguese ability level plays a significant role in the lives of people living in Portugal, thus it is important to provide early support to students who perform poorly in Portuguese relative to their peers before the variance of ability levels between students compound. However, most Portuguese public schools store student data such as school reports on physical paper rather than digitally, which makes it difficult to track student scores. Thus, an alternative strategy to provide early support to students would be to direct additional resources to students that are predicted to perform poorly in Portuguese. This report explores the questions: What factors have significant impacts on final Portuguese grades for students in Portuguese public secondary schools? What category of the variables was most prominent? The results of the resulting investigations can help predict which students will require additional Portuguese support and help improve education outcomes.

Data Set Description. This report uses student data collected by Paulo Cortez and Alice Silva (University of Minho, 2005-2006) from 649 students across two Portuguese public secondary schools. The dataset consists of 33 attributes and was formed by combining data from school reports, which contained student grades and number of absences, and from questionnaire responses, which consisted of demographic, lifestyle and socioeconomic variables that were predicted to impact student performance. More specifically, the dataset contains 3 attributes which are related to student grade, namely G1, G2, and G3, which are values between 0 and 20 representing the first period grade, second period grade, and final grade respectively.

Analysis.

Full Model Assumption Checking. The initial model applied to the data set was a full linear model constructed from all of the variables on the untouched data set. Importantly, the assumptions of the underlying model seemingly weren't sufficiently met when initially inspected. Both homoscedasticity and normality of residuals assumptions were not upheld, the best example of this is shown in figure-1 which features heteroscedastic placement of residuals in the lower left hand quadrant of the figure.

After inspection these heteroscedastic data points correspond to zero scoring students in the final exams (G3 = 0), this is indicative of students who were absent or got caught cheating on their final exam, as this result differs from the aim of this investigation they shall be removed from further analysis. Additionally two outliers were detected using Cook's Distance in figure-5 and so were removed to reduce their effect on the final model.

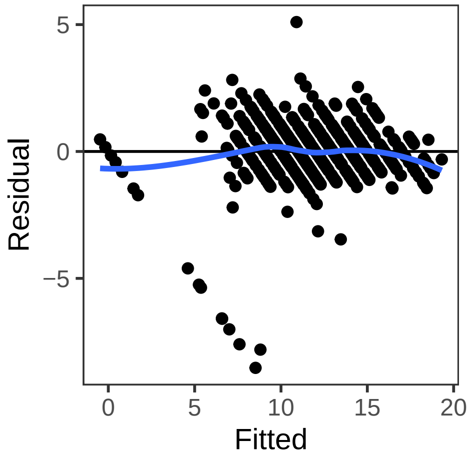

Fig. 1. Residual Plot of the Full Model

Derivation of Final Model. The data set has now been transformed to fit the assumptions of the linear model, however, in order to derive an optimal model, forward and backward AIC and BIC model derivations may be used. The optimal model was found to be the forward AIC model as this model outperformed both the full baseline model but also forward and backward BIC models in terms of coefficient of determination (noting that both AIC and BIC forward and backward models had identical results) (figure-2) (equation-1). With AIC deriving an adjusted coefficient of determination of 0.909 and AIC score of 1524.877, hence the forward AIC model will be selected as the final model (table-1).

$$\begin{aligned} G3 = &-0.586 + 0.741\text{G2} + 0.199\text{G1} + 0.119\text{age} \\ &-0.173\text{failures} - 0.091\text{Dalc} + 0.310\text{higher[yes]} \\ &-0.236\text{schoolsup[yes]} - 0.047\text{health} \ - 0.044\text{goout} \end{aligned} \quad [1]$$

Final Model Assumptions. Unlike the first full model, the final model seems to satisfy the assumptions of the linear model. The assumption of independence is upheld given the fact that G3 is a test score and each student performed the final test strictly independently (otherwise receiving a score of 0). Homoscedasticity was assessed with the use of a residual plot shown in figure-3, wherein the points seem to be evenly scattered above and below the identity line indicating homoscedasticity. Linearity was also assessed with the residual plot, wherein the points are relatively flat across the fitted values this is further supported by a relatively flat Loess fitted line indicating linearity. Normally distributed residuals were assessed through a QQ-plot in figure-4, featuring data points closely straddling the line across the quantile range supporting the normality assumption.

Results. The final linear model is shown above (1). Here it is interesting to note that the AIC fell from 1579.18 in the backward model and 1578.6 in the forward model to 1524.877 and 1524.877 respectively, in the final model. Indicating the final model is a better model as a result of the lower AIC. This also the case with the BIC falling from 1581.603 in both forward and backward, and in the final BIC being 1533.832. Furthermore, in the original AIC model the adjusted r^2 for the forward model was 0.901 and 0.902 in the backward model, whereas in the final model, the adjusted R^2 was 0.909 and 0.909 respectively, indicating the final better fit the data, this was mostly likely due to the removal of outliers. This is also the case with the BIC model with the r^2 increasing from 0.900 (in both forward and back) to 0.907 (again in both), indicating the BIC model improved as well in the final model.

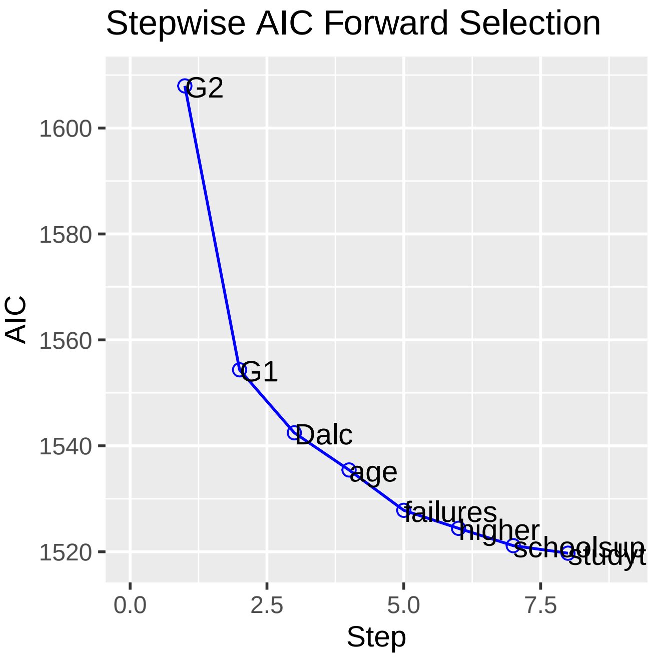

Fig. 2. AIC Optimzed Model Derivation

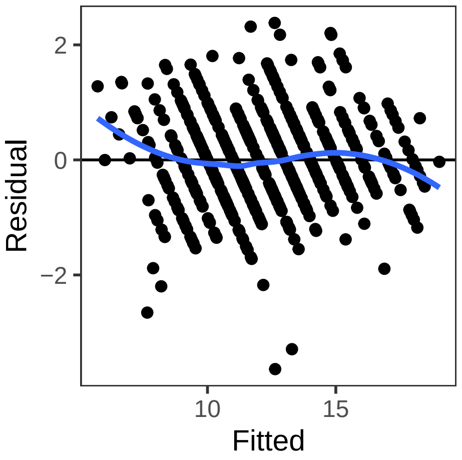

Fig. 3. Residual Plot of Final Model

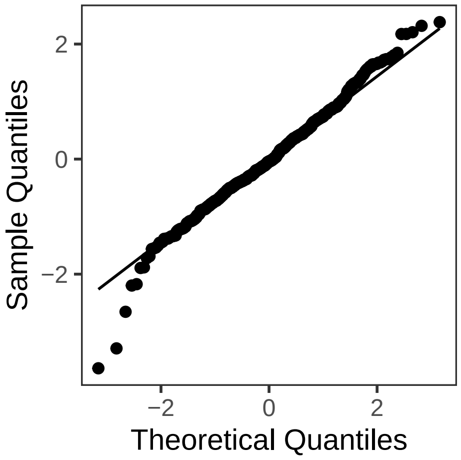

Fig. 4. QQ-Plot of Final Model

Discussion and Limitations. In answering our key questions, the variables found to predict student performance from the forward AIC model were first period grade, second period grade, age, number of past class failures, average workday alcohol consumption, whether students want to take higher education, if they have extra educational support, current health status and how often they go out with friends. Categorising these under demographic, lifestyle and school related variables as outlined in Cortex and Silva (2008), we infer lifestyle and school related variables to equally contribute to predicting student achievement. Considering first and second period grades are well correlated with third period grades, to explore other relevant factors that influence student performance, these were not included in the three categories.

A limitation of this report is that there are a large number of parameters (32) relative to the sample size (632 after pre-processing), which makes it harder to determine whether factors have a significant impact on the response variable. A potential improvement would be to reduce the number of parameters in the pre-processing step by eliminating one of the factors for any pairs of factors that are correlated.

A further limitation is the dataset wasn't from a random sample of Portuguese secondary school students. We already observed that there may be a significant difference between average final scores of students in different schools, such as between MS and GP. Since the dataset is limited to two schools, the model may be very inaccurate for other public secondary schools, particularly for those with significantly different score distributions. An improvement would be to include a larger variety of schools so that the sample is closer to a random population. Further, only 227 of the observations were from Mousinho da Silveira (MS) secondary school. An analysis by Molin (2020) of the dataset had findings that Gabriel Pereira (GP) secondary school excelled academically compared to GP. GP had mainly urban students, the lowest proportion of failures, highest study time and offered school support. The discrepancy between these two schools may potentially skew the data. In addition, cross-cultural limitations may inhibit the generalisability of findings to other populations as this study was conducted in Portugal; a western, educated, industrialised, rich and democratic (WEIRD) society . Future prospects for this could include acquiring data of schools from a multitude of countries as well as ensuring an equivalent number of observations from different schools within the same region. Ensuring a large sample size to account for the number of attributes will counter the issue of overfitting. On top of this, the inclusion of both G1 and G2 in the final model results in multicollinearity since they are correlated (see Appendix-6), which may reduce the accuracy of the coefficients and increase the standard error. To avoid this, we could remove one of the variables before performing the model selection.

Conclusion. In conclusion, the report aims to identify attributes that influence student performance in secondary school which can then be used to improve educational outcomes. These were found to be first period grade, second period grade, age, number of previous class failures, workday alcohol consumption, desire to pursue higher education, extra educational support, current health status, and degree of going out with friends. Lifestyle and school-related factors were also found to equally contribute to the success of students in school.

References

Hlavac M (2022). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Social Policy Institute, Bratislava, Slovakia. R package version 5.2.3, URL https://CRAN.R-project.org/package=stargazer.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, **4**(43), 1686. doi:10.21105/joss.01686.

Xie Y (2022). *tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents*. R package version 0.42, URL https://github.com/rstudio/tinytex.

Ünal F (2020). "Data Mining for Student Performance Prediction in Education." In D Birant (ed.), *Data Mining*, chapter 9. IntechOpen, Rijeka. doi:10.5772/intechopen.91449. URL https://doi.org/10.5772/intechopen.91449.

Table 1. Appendix: AIC vs Baseline Model Comparison

	Dependent variable:	
	G3	
	AIC Models (Forward and Back)	Full Baseline Model
	(1)	(2)
schoolMS		−0.198 (0.128)
sexM		−0.123 (0.118)
G2	0.741*** (0.026)	0.870*** (0.035)
G1	0.199*** (0.026)	0.129*** (0.038)
age	0.119*** (0.029)	0.029 (0.048)
addressU		0.114 (0.123)
famsizeLE3		0.016 (0.115)
PstatusT		−0.097 (0.163)
Medu		−0.092 (0.071)
Fedu		0.050 (0.065)
Mjobhealth		0.266 (0.252)
Mjobother		−0.094 (0.142)
Mjobservices		0.173 (0.175)
Mjobteacher		0.221 (0.236)
Fjobhealth		−0.444 (0.353)
Fjobother		−0.338 (0.214)
Fjobservices		−0.471** (0.225)
Fjobteacher		−0.544* (0.316)
reasonhome		−0.079 (0.134)
reasonother		−0.362** (0.172)
reasonreputation		−0.169 (0.140)
guardianmother		−0.025 (0.125)
guardianother		0.217 (0.249)
traveltime		0.139* (0.075)
studytime		0.050 (0.066)
failures	−0.173*** (0.063)	−0.255** (0.099)
Dalc	−0.091** (0.037)	−0.052 (0.072)
Walc		−0.017 (0.056)
higheryes	0.310*** (0.117)	0.207 (0.183)
internetyes		0.085 (0.130)
romanticyes		−0.042 (0.108)
famrel		−0.016 (0.055)
freetime		−0.050 (0.053)
schoolsupyes	−0.236** (0.107)	−0.184 (0.173)
famsupyes		0.095 (0.107)
paidyes		−0.192 (0.217)
activitiesyes		0.012 (0.105)
nurseryyes		−0.096 (0.127)
health	−0.047** (0.022)	−0.055 (0.036)
absences		0.014 (0.012)
goout	−0.044 (0.029)	−0.019 (0.050)
Constant	−0.586 (0.531)	0.638 (0.964)
Observations	632	649
R^2	0.910	0.860
Adjusted R^2	0.909	0.851
Residual Std. Error	0.801 (df = 622)	1.249 (df = 607)
F Statistic	701.120*** (df = 9; 622)	90.951*** (df = 41; 607)
Note:		*p<0.1; **p<0.05; ***p<0.01

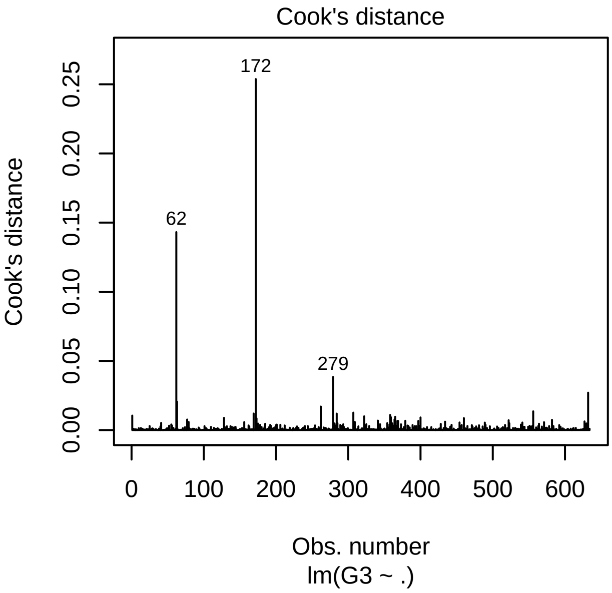

Fig. 5. Appendix: Cook's Distance Plot of Untouched Data Set

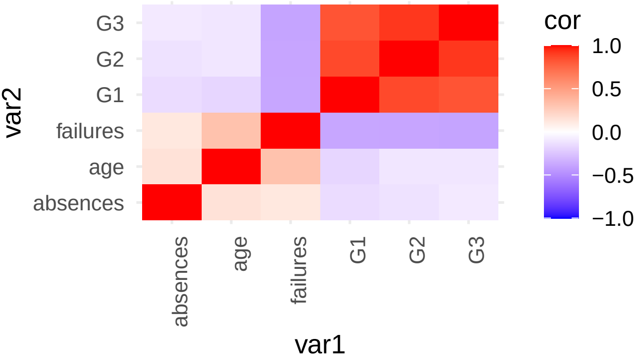

Fig. 6. Appendix: Correlation Matrix of Numberic Variables in the Data Set