Modeling Student Performance

Regression Analysis of the Portuguese Student Dataset

Abstract:
In this study, we investigate which factors best predict final course grades for Portuguese secondary school students and whether including interaction terms improves model performance. We use the Portuguese language subset of the Student Performance dataset from the UCI Machine Learning Repository and fit linear regression models using stepwise AIC and BIC selection. Models are compared using leave-one-out cross-validation. While prior grades are the strongest predictors of final performance, the selected AIC interaction model identifies additional factors. Specifically, the effect of age on final grade differs by failure history, and earlier grades are more predictive for students with higher education plans. Overall, the final model explains over 91% of the variation in final grades.

1. Introduction

Academic performance in secondary schools can be a reflection of both individual effort and unequal access to resources and opportunities [2]. Detection of these factors, especially those that are outside of a students' control, is key to highlighting areas where intervention or additional support may be needed to reduce drop out at critical stages in a child's educational development [1]. This phenomenon is especially relevant in Portugal. At the time the Portuguese student data was collected, Portugal recorded an early school leaving rate of 38.3% for 18 to 24 year olds, while the European Union average value was approximately 15% [4].

This study investigates the following research question: What factors best predict final course performance in the Portuguese student dataset, and do interaction effects among these predictors improve model fit? Previous work on this data indicates that a student's first and second period grades are the most important predictors of final grade, but we are interested in exploring other predictors as well [3].

2. Data

2.1 Data Description

For this study, we used the Student Performance dataset from the UC Irvine Machine Learning Repository [5]. The dataset is originally from a paper titled *Using Data Mining to Predict Secondary School Student Performance* [3]. This data was collected through a combination of administrative school reports and a questionnaire distributed to students during the 2005-2006 school year from two public secondary schools in Portugal. Observations were recorded for 395 students in mathematics classes and 649 students in Portuguese language classes. The dataset contains 32 predictor variables related to demographic, social/emotional, and other school-related factors. The response variable of interest is final grade, denoted *G3*.

2.2 Data Cleaning

While no values are explicitly missing from the dataset, we removed 15 cases where *G3* is observed to be 0. In all 15 cases, students' first and/or second period grades (*G1* and *G2*) are non-zero, thus implying that *G3* is no longer an accurate depiction of final grade. Instead, it is likely that cases with *G3* = 0 imply that a student has dropped out of class in the middle of the school year, which means this *G3* score is no longer relevant to our analysis of factors that influence final academic performance.

Because the Student Performance dataset is split by mathematics and Portuguese language courses, our next step before beginning the model selection process is to determine whether or not we can merge the two datasets together. We compared the distribution of G3 across the two course datasets using side-by-side boxplots (Figure 1). Here, we see that the equal variance assumption is violated, and thus we cannot merge the two datasets. Instead, we choose to focus our analysis on the Portuguese language dataset because it contains twice the amount of observations. In the end, our trimmed and cleaned dataset has 32 predictor variables and a sample size of 634.

3. Methods

3.1 Multicollinearity Assesment

Before model selection, we assessed multicollinearity among the 32 candidate predictors using variance inflation factors (VIFs). All the VIF values are below 4.3, so multicollinearity is not a concern at this stage.

3.2 First-Order Model Selection

We started with a full first-order linear regression model that included all 32 predictors and then used stepwise selection with both the Akaike Information Criterion (AIC) and the Bayesian

Information Criterion (BIC). The AIC procedure retained a broader set of predictors, including sex, age, prior grades (*G1* and *G2*), and failure history, whereas the BIC procedure selected a much smaller model that included only age and prior grades. Because our research question is not limited to the predictive role of prior grades, we carried the AIC-selected model forward when examining interaction effects.

3.3 Interaction Models

Starting from the AIC-selected first-order model, we considered all possible two-way interaction terms. Stepwise selection was again carried out using both AIC and BIC.

The AIC-based procedure produced a larger model with several interaction terms. The BIC-based procedure, however, did not select any interactions and remained identical to the main-effects-only specification.

3.4 Model Comparison

We compared predictive performance using leave-one-out cross-validation (LOOCV) and information criteria; results are summarized in Table 1. The AIC interaction model has a slightly lower cross-validated mean squared error than the BIC model (0.826 versus 0.867). On the original grading scale, this corresponds to prediction errors of about $\sqrt{0.826} \approx 0.91$ and $\sqrt{0.867} \approx 0.93$ grade points. Compared to the BIC-selected model, the AIC interaction model also performs better based on AIC and adjusted R-squared. Therefore, we selected this model for further evaluation.

Table 1: Comparison of candidate models using LOOCV MSE, AIC, BIC, and adj. R-squared

Model	LOOCV MSE	AIC	BIC	Adj. R-squared
AIC Interaction Model	**0.8259691**	**1674.184**	1745.416	**0.8897**
BIC-selected Model	0.8672242	1710.138	**1732.398**	0.8812

3.5 Diagnostic Checks

We examined standard regression diagnostics for the selected interaction model, including residuals versus fitted values, Q-Q plots, leverage, and Cook's distance. No observations appeared to be highly influential, though a small number showed unusually large studentized residuals across the plots (Figure 2). We removed five such observations, which improved residual symmetry and reduced heteroscedasticity.

For the AIC interaction model, we checked a time-sequence plot of studentized residuals to assess independence. The residuals are centered around zero and showed no clear pattern across observations, so the independence assumption appeared reasonable (Figure 3).

3.6 Final Model

After removing outliers, we refit the model on the reduced dataset. The final model includes main effects for sex, age, prior grades (*G1* and *G2*), failure history, and plans for higher education, along with interaction terms between age and failures and between higher education plans and both *G1* and *G2*.

Diagnostic plots for the final model show roughly constant variance and improved residual normality, with no evidence of problematic leverage (Figure 4). We used this model for subsequent inference and interpretation.

Our final model includes the predictors listed on the right (variable descriptions can be found in the Appendix). Out of the 14 predictors, 7 predictors are statistically significant at $\alpha = 0.05$ and 4 predictors are marginally significant at $\alpha = 0.1$. The general *F*-test indicates that the model is overall highly significant ($F(14, 614)=474.6$, $p<.001$). The adjusted R^2 is 91.4%, meaning that after accounting for model complexity, over 91% of variation in the response can be explained by the fitted model.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.47697	0.82077	-0.581	0.56137	
as.factor(sex)M	-0.55153	0.28755	-1.918	0.05557	.
age	0.14054	0.03069	4.579	5.66e-06	***
as.factor(failures)1	-1.38522	1.44650	-0.958	0.33863	
as.factor(failures)2	5.03349	2.41225	2.087	0.03733	*
as.factor(failures)3	7.82334	4.57425	1.710	0.08772	.
as.factor(higher)yes	-0.59192	0.63145	-0.937	0.34893	
G1	0.33693	0.07662	4.397	1.29e-05	***
G2	0.52646	0.08245	6.385	3.37e-10	***
age:as.factor(failures)1	0.06293	0.08334	0.755	0.45050	
age:as.factor(failures)2	-0.27218	0.13187	-2.064	0.03944	*
age:as.factor(failures)3	-0.47236	0.26127	-1.808	0.07111	.
as.factor(sex)M:G1	0.04218	0.02448	1.723	0.08541	.
as.factor(higher)yes:G2	0.25417	0.08639	2.942	0.00338	**
as.factor(higher)yes:G1	-0.17676	0.07981	-2.215	0.02714	*

4. Conclusion

4.1 Discussion

In response to our research question, the model shows that significant predictors of final grade in Portuguese language classes are age, *G1*, *G2*, past failures, and several interaction terms including the interaction between age and having two past failures and the interactions between higher education aspirations and both first and second period grades.

In addition to validating previous work which found that *G1* and G2 are positive predictors of final grade, we also found that older students tend to perform better in Portuguese language classes. Specifically, for every 1-year increase in age, the predicted increase in *G3* is 0.14 points, after adjusting for simultaneous linear change in all other predictors ($p<.001$). However, the interaction between age and failures implies that age is not associated with a positive increase in *G3* for students with two prior failures. Instead, for students with 2 prior failures, every 1-year increase in age is associated with a 0.131 point decrease in final grade, after adjusting for simultaneous linear change in all other predictors ($p=.04$).

Additionally, we also see significant interaction between aspirations to pursue higher education, and first and second period grades. In particular, for students who intend to pursue higher education, a 1-point increase in *G1* score is associated with a 0.160 point increase in *G3* ($p=.003$), whereas a 1-point increase in *G2* score is associated with a 0.780 point increase in *G3* ($p=.03$), while adjusting for simultaneous linear change in all other predictors. This result implies that though improvements in both periods benefit final grade, students who wish to pursue higher education gain a larger benefit on *G3* from improvements in *G2* than *G1*. Therefore, our analysis indicates that older students with more failures and students who are not as motivated by aspirations for higher education are more at-risk of early school leaving.

4.2 Further Considerations

The largest limitation of our research is the time of data collection. First, because the data was collected in the 2005-2006 school year, it is very likely that Portuguese educational trends have changed drastically since then. In fact, though Portugal ranked last among the European Union in early school leaving rate in 2005, by 2024, Portugal had moved up to 10th place which demonstrates a significant decrease in student drop out rate in secondary education [4]. Therefore, future research in this area will be sure to improve our results using newer datasets.

Another limitation of our analysis is the lack of transparency in sampling methodology. Though the description of data collection procedures (i.e., administrative school reports and questionnaires) are quite thorough, there are certain gaps regarding how students were chosen to be surveyed for this dataset. This ambiguity raises concerns regarding potential selection bias and whether the sample adequately represents the Portuguese student population.

References

[1] Chamillard, A. T. (2006). Using student performance predictions in a computer science curriculum. *ACM SIGCSE Bulletin, 38(3)*, 260-264.

[2] Cortez, P. (2008). Student Performance [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.

[3] Cortez, Paulo & Silva, Alice. (2008). Using Data Mining to Predict Secondary School Student Performance. *EUROSIS*.

[4] Eurostat (2024). Early leavers from education and training by sex [Dataset]. https://ec.europa.eu/eurostat/databrowser/view/sdg_04_10/default/table

[5] Farooq, M. S., Chaudhry, A. H., Shafiq, M., & Berhanu, G. (2011). Factors affecting students' quality of academic performance: A case of secondary school level. *Journal of quality and technology management, 7(2)*, 1-14

Appendix

Table 2: Description of Predictors

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – $<$ 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – $>$ 1 hour).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

a 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
b teacher, health care related, civil services (e.g. administrative or police), at home or other.

Source: Cortez, Paulo & Silva, Alice. (2008). Using Data Mining to Predict Secondary School Student Performance. EUROSIS.

Figure 1: Variance of Response G3

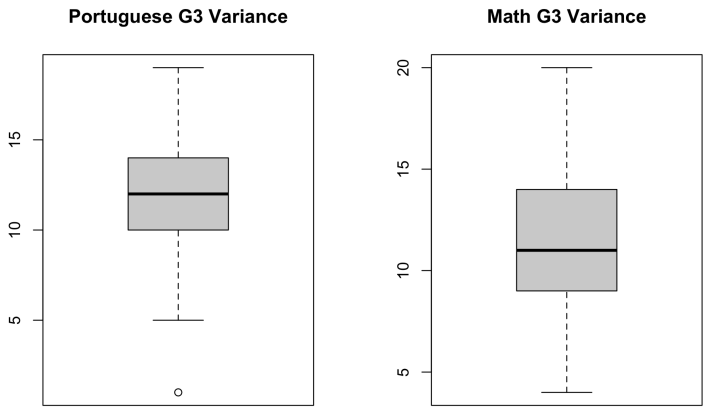

Figure 2: Diagnostic plots for the AIC interaction model

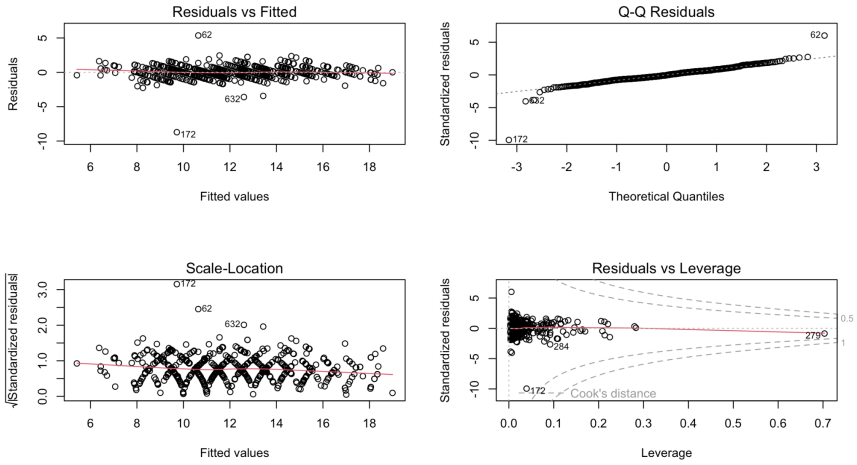

Figure 3: Time sequence plot for the AIC interaction model

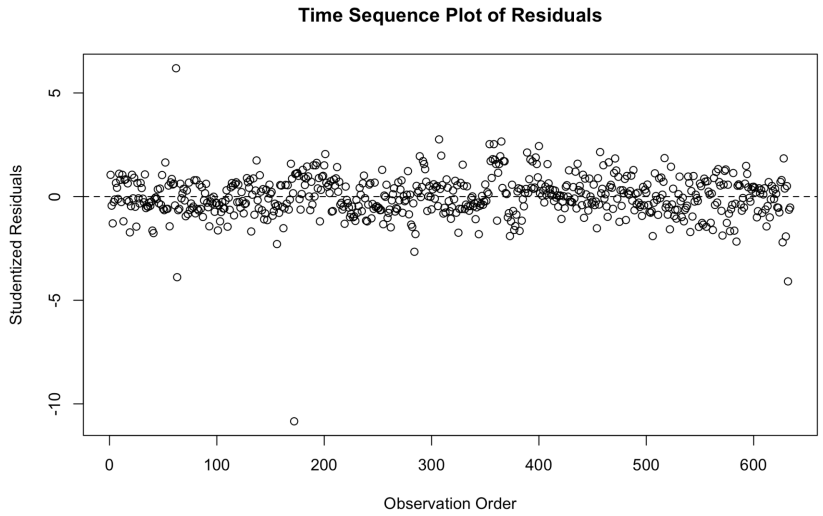

Figure 4: Diagnostic plots after outlier removal

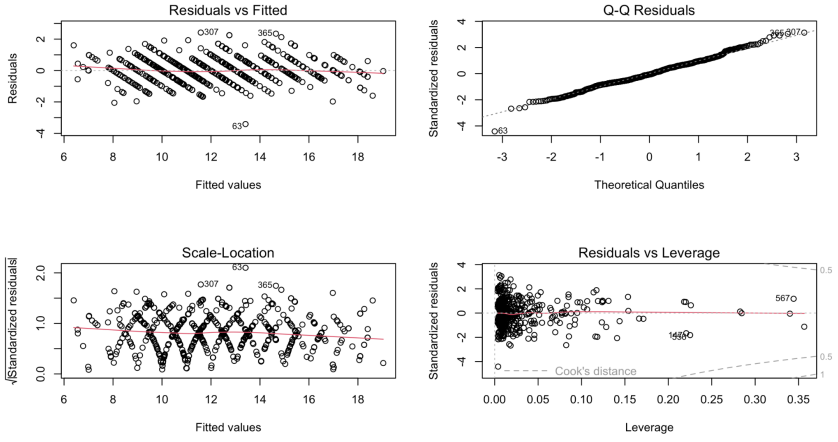

Table 3: Final Model Summary

Term	Estimate	Std. Error	t-value	p-value	95% CI
Intercept	-0.47697	0.82077	-0.581	0.56137	[-2.0888, 1.1349]
Sex = M	-0.55153	0.28755	-1.918	0.05557	[-1.1162, 0.0132]
Age	0.14054	0.03069	4.579	5.66e-06	[0.0803, 0.2008]
Failures = 1	-1.38522	1.44650	-0.958	0.33863	[-4.226, 1.4555]
Failures = 2	5.03349	2.41225	2.087	0.03733	[0.2962, 9.7708]
Failures = 3	7.82334	4.57425	1.710	0.08772	[-1.1597, 16.8064]
Higher = yes	-0.59192	0.63145	-0.937	0.34893	[-1.832, 0.6482]
G1	0.33693	0.07662	4.397	1.29e-05	[0.1865, 0.4874]
G2	0.52646	0.08245	6.385	3.37e-10	[0.3645, 0.6884]
Age*(Failures = 1)	0.06293	0.08334	0.755	0.45050	[-0.1007, 0.2266]
Age*(Failures = 2)	-0.27218	0.13187	-2.064	0.03944	[-0.5312, -0.0132]
Age*(Failures = 3)	-0.47236	0.26127	-1.808	0.07111	[-0.9854, 0.0407]
(Sex = M)*G1	0.04218	0.02448	1.723	0.08541	[-0.0059, 0.0902]
(Higher = yes)*G2	0.25417	0.08639	2.942	0.00338	[0.0845, 0.4238]
(Higher = yes)*G1	-0.17676	0.07981	-2.215	0.02714	[-0.3335, -0.0200]