Brian Ferrell

STAT 534 – 001 – Final Report

December 15, 2021

1 Introduction and Setup

Below is a report on data science techniques and summarized results for predicting a student's final math grade in secondary education for two Portuguese schools. The variables include student grades, demographics, social and school related features and the data was collected by school reports and questionnaires. The goal of this was to develop predictive models for the final year grade (G3) using statistical learning methods. In this report I include a exploratory data analysis, two different multiple linear regression models, 2 classification models with and without past student performances, and communication of results. This analysis is in the following format:

- Prediction of G3 as a continuous response (A1)
- Prediction of G3 as a binary response (A2)
- Prediction of G3 as a binary response without G1 or G2 (A3)

For each of the items A1, A2, and A3, at least two models are applied and compared. Additionally, G3 has a strong correlation with G1 and G2 since student achievement is often affected by previous performances; therefore, we want to see which models are more useful in studying the effect of other relevant features in the dataset. This analysis was carried out using R to which there was a .80 training/testing split (the following code files will be attached as well).

2 Exploratory Data Analysis (EDA)

The dataset comprises of 32 variables (Figure 8). Below shows a breakdown of the different data structures, amount of observations, missing variables (got rid of missing values).

division	metrics	value
size	observations	395
size	variables	32
size	values	12,640
size	memory size (KB)	0
duplicated	duplicate observation	0
missing	complete observation	395
missing	missing observation	0
missing	missing variables	0
missing	missing values	0

division	metrics	value
data type	numerics	4
data type	integers	1
data type	factors/ordered	17
data type	characters	10
data type	Dates	0
data type	POSIXcts	0
data type	others	0

Figure 1: Data Structures

Below shows descriptive statistics for our numeric variables as well as a table showing normality tests of the numerical variables.

variables	missing	mean	sd	min	Q1	median	Q3	max
age	0	16.70	1.28	15	16	17	18	22
failures	0	0.33	0.74	0	0	0	0	3
absences	0	5.71	8.00	0	0	4	8	75
G1	0	10.91	3.32	3	8	11	13	19
G2	0	10.71	3.76	0	9	11	13	19

Figure 2: Descriptive Statistics

variable	min	Q1	median	Q3	max	skewness	kurtosis	balance
age	15	16	17	18	22	0.5	0.0	Balanced
failures	0	0	0	0	3	2.4	5.0	Right-Skewed
absences	0	0	4	8	75	3.7	21.7	Right-Skewed
G1	3	8	11	13	19	0.2	-0.7	Balanced
G2	0	9	11	13	19	-0.4	0.6	Balanced

Figure 3: More Descriptive Statistics

The figure below shows some of the variables in a bivariate analysis comparing numerical variables. You can see that G1 and G2 are highly correlated with each other.

first variable	second variable	correlation coefficient
age	failures	0.24367
age	absences	0.17523
age	G1	-0.06408
age	G2	-0.14347
failures	absences	0.06373
failures	G1	-0.35472
failures	G2	-0.35590
absences	G1	-0.03100
absences	G2	-0.03178
G1	G2	0.85212

Figure 4: Comparing Numerical Variables

The two figures below show a correlation plot as well as scatterplots in order to see how our numerical variables pair with our predictor variable as well as the surrounding ones. You can see that age is negatively correlated with G3; whereas, G2 is highly correlated, not just with G3 but with G1 as well.

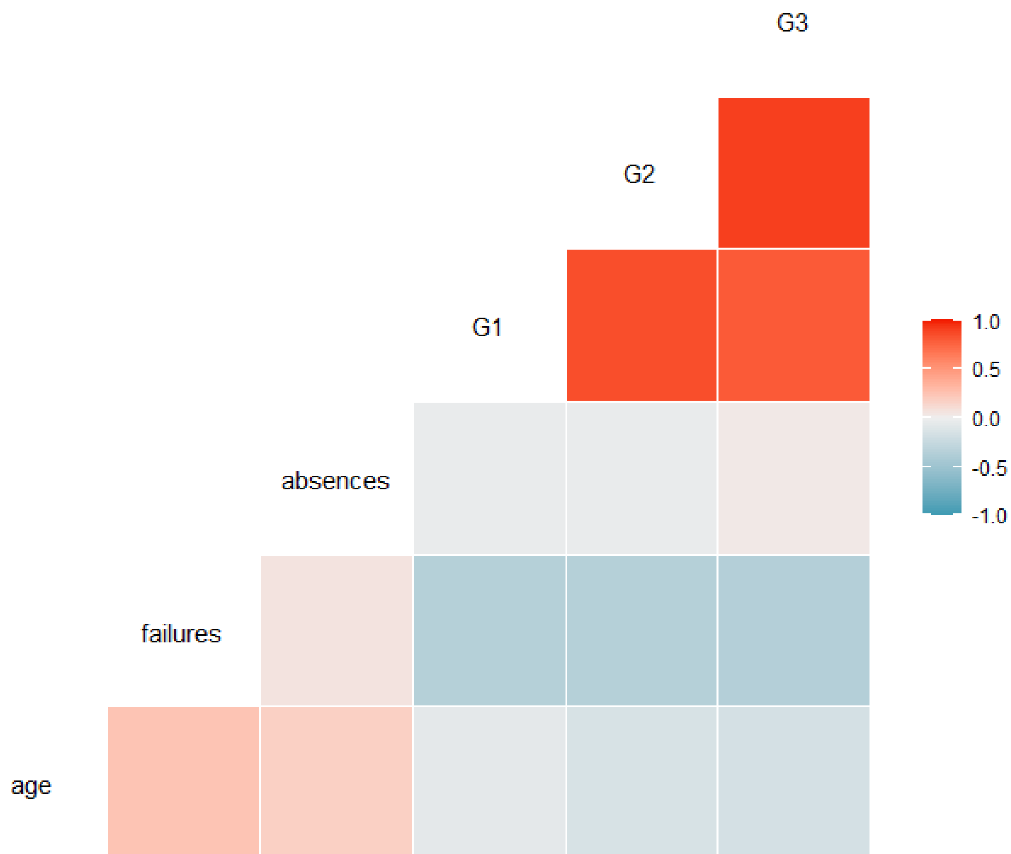

Figure 5: Correlation

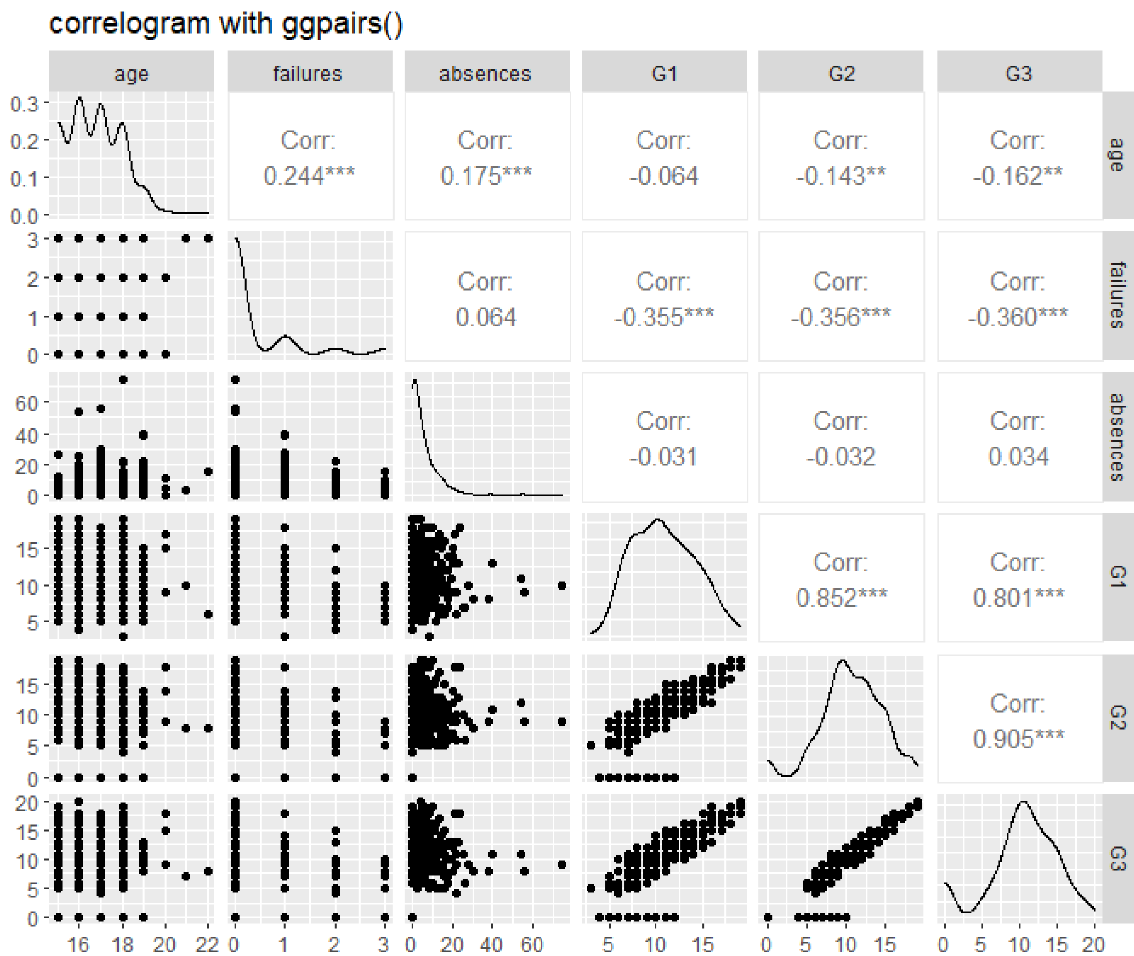

Figure 6: Scatterplots w/ correlation

3 Methods

3.1 Regression (A1)

For feature selection I used the Earth - Mars (Multivariate Adaptive Regression Splines) package to help choose which variables to train the models on. There are obviously other ways of choosing variables, but this was one way of doing it. For this section I performed two different multiple linear regression models, one model contains variables from the Earth package; whereas, the other contains variables that did not appear in the first model. Below shows a plot of which variables were important with respect to our predictor variable as well as a table showing number of subsets, Generalized Cross Validation, and Residual sum-of-squares (RSS).

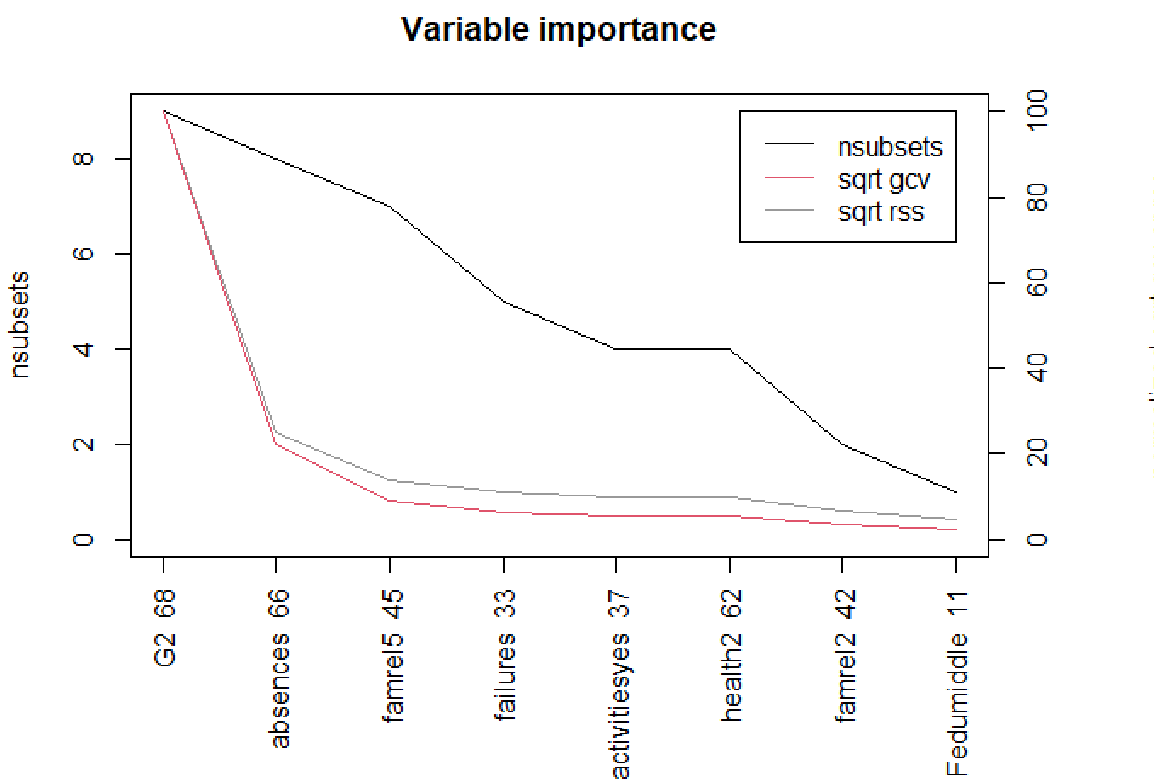

Figure 7: Feature Selection Plot

Table 1: Output of Earth

	nsubsets	gcv	rss
G2	9	100	100
absences	8	22.2	24.8
famrel5	7	8.8	13.8
failures	5	6.4	11
activitiesyes	4	5.6	9.8
health2	4	5.5	9.7
famrel2	2	3.5	6.6
Fedumiddle	1	2.3	4.6

I used all of these to train the majority of our models; however, wherever you see a categorical variable, it gives a value next to it (i.e. famrel5 or Fedumiddle). I replaced those with its original variable name.

3.2 Classification (A2)

The same features from the first model of the previous section were used for our classification models. In this section we used a Naive Bayes classifier as well as a ANN model. In order to make this a a binary classification task - the condition "Pass" attributed to when G3 was greater than or equal to 10, and "Fail" would be for the opposite. Additionally, for our ANN model we implemented 9 hidden units and the Relu activation function during the training (as seen in the figure below).

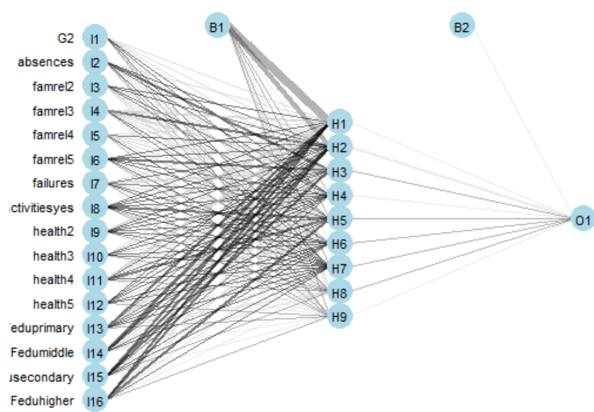

Figure 8: Artificial Neural Network

3.3 Inference (A3)

This section remains a binary classification task, but without the variables G1 and G2 (G1 was never in our original model anyways). Everything else parameter-wise remains the same in this section.

4 Summary of Results

4.1 Regression (A1)

As you can see from the results below in Table 2, model 1 has a higher R-Squared, Ajd. R-Squared, as well as a smaller Mean Sqaured Error (MSE) and AIC. It appears that choosing the variables using the Earth package proved to be significant both statistically as well as performance-wise compared to not using them. The MSE is calculated from the predictions on the testing dataset (79 samples).

Table 2: A1 Results

	r.squared	adj.r.squared	p.value	df	aic	bic	df.residual	mse
model 1	0.832	0.823	1.74e-105	16	1330	1398	299	3.31
model 2	0.737	0.685	1.95e-51	52	1544	1747	263	8.86

4.2 Classification (A2)

The three tables below show the training and testing results, confusion matrices, as well as sensitivity and specificity. The Neural Network outperformed the Naive Bayes model in both the training accuracy and testing; however, both still did fairly well considering this is only a binary classification task. The Naive Bayes model struggled a lot more on correctly labeling a final grade as "Pass" compared to the NN, but did slightly better in not miss-labeling "Fail" grades as passing compared to the NN.

Table 3: A2 Results

	Training Acc.	Testing Acc.	Sensitivity	Specificity
Neural Net	0.987	0.899	0.9	0.897
Naive Bayes	0.915	0.823	0.94	0.621

Table 4: Confusion Matrix for ANN

	fail	pass
fail	26	5
pass	3	45

Table 5: Confusion Matrix for NB

	fail	pass
fail	18	3
pass	11	47

4.3 Inference (A3)

The three tables below show the training and testing results, confusion matrices, as well as sensitivity and specificity again but this time it was trained with out the G2 variable. The Neural Network still outperformed the Naive Bayes model in both the training accuracy and testing (but not as much). Both models struggled a lot more on correctly labeling a final grade as "Pass". In addition, both were highly sensitive to labelling a "Fail" grade as passing.

Table 6: A3 Results

	Training Acc.	Testing Acc.	Sensitivity	Specificity
Neural Net	0.861	0.684	0.86	0.379
Naive Bayes	0.747	0.671	0.92	0.241

Table 7: Confusion Matrix for NN

	fail	pass
fail	11	7
pass	18	43

Table 8: Confusion Matrix for NB

	fail	pass
fail	7	4
pass	22	46

5 Conclusion

In conclusion, we trained 6 models using regression and classification techniques. The best model overall was the ANN, and in section A3 we proved that without G2 the models performances go way down; however, we still learn that the variables failures and absences play a huge part in the predictive performance of these models.

6 Appendix

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.

Figure 9: List of variables