STAT 534 Statistical Data Science I

Fall 2021

Course Project

The data we provided in this project approaches student achievement in secondary education of two Portuguese schools. The variables include student grades, demographic, social and school related features and the data was collected by using school reports and questionnaires. A dictionary of the variables is given in the last page. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). Our goal is to develop predictive models for the final year grade (G3) using statistical learning methods. You can choose one of the subjects in your project. The completion of the project involves exploratory data analysis, statistical modeling, and communication of results. Your analysis should address the following three aspects:

(A1) prediction of G3 as a continuous response,

(A2) prediction of G3 as:

- a binary response - pass if G3 $\geq$ 10, else fail;
- or a 5-level response - A: 16-20, B: 14-15, C: 12-13, D:10-11, E:0-9,

(A3) and inference about G3 without using past grades G1 and G2. (G3 has a strong correlation with G1 and G2 since student achievement is often affected by previous performances. It is more difficult to predict G3 without G1 and G2, but such models are more useful in studying the effect of other relevant features in the dataset.)

You may use any reference material, but the project should be done individually. Please find the explanations of all the variables on the last page. Please contact the instructors with any questions.

If you are really interested in a data science problem, you can use your own data set and submit a project proposal by **December 3**. Remember, you need to apply **both** regression and classification methods to your dataset.

Stage	**Due Date**	Description
1	**Friday, Dec 3**	Project proposal for your own problem (optional)
2	**Friday, Dec 17**	Final report

The document that you turn in at each stage should be narrative with complete sentences.

1 Project Proposal (Optional)

Your project proposal should be one typewritten document that includes:

- a descriptive title,
- a brief description of the dataset including the number of observations and attributes,
- research questions of interest/project aims,
- and a tentative outline of statistical analysis.

2 Final Report (100pt)

Apply the data science techniques to the dataset you choose and summarize the results in a report. Be sure to use proper practices in the empirical evaluation of predictive methods. Submit your report in .pdf format via Canvas. Attach your R code as an appendix or as a separate file. Your final report should be no more than 10 pages (including figures and references, but excluding code) and includes

- (10 pts) an exploratory study of the variables.
- (20 pts) multiple linear regression with your choices of variable selection/dimension reduction methods to address (A1),
- (20 pts) classification analysis to address (A2),
- (20 pts) repeating the regression or the classification analysis without G1 and G2 to address (A3),
- (20 pts) and a summary of analysis results in context.

For each of the items A1, A2, and A3, at least two models should be applied and compared. The report will be evaluated based on completion of the described tasks, the appropriate use of statistical learning methods, and a convincing demonstration of interpretation/conclusions in context, the clarity and organization of your writings.

The last 10 pts are to evaluate the reproducibility of your codes.

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.

You can load the datasets using:

```
d1=read.table("student-mat.csv",sep=";",header=TRUE)
d2=read.table("student-por.csv",sep=";",header=TRUE)
```