EAS506 - Statistical Data Mining I

Homework 1 – Question 2

Paul M Girdler

09/14/18

Abstract

This report summarizes the steps taken to perform multiple-linear-regression to create various models to predict *First Period Grades* (variables *G1.x*, and *G1.y*) . It examines any significant predictors, and interactions. From these recommendations are made to aid students in improving their grades.

Content

1 Introduction

The *Student Performance Dataset* is based upon two datasets of the academic performance of Portuguese students in two different classes: Math and Portuguese. The dataset is available on the **UCI machine learning repository**.

This report follows on from the previous report detailing the preprocessing, analysis and preparation of the data. This report summarizes the steps taken to perform multiple-linear-regression to create various models to predict *First Period Grades* (variables *G1.x*, and *G1.y*) .

2 Method

2.1 Initialization Steps

- Clear the memory
- Install and load all required libraries.
- Import and merge data.
- Briefly examine the data.

2.2 Create Linear Regression Models

- A number of Multiple-Linear-Regression Models were created to predict First Period Grades (*G1.x*, and *G1.y*) in a number of flavours

Model Type	Feature Space
"Kitchen-Sink" Model	all 27 variables
"Kitchen-Sink" Model (with all interactions)	all 27 variables with interactions
"Trimmed" Model	feature selection with only most significant variables.
"Trimmed" Model	feature selection (with all interactions): feature selection with only most significant variables.

See R code for detailed summary of every model.

3 Discussion

3.1 Part A

Which predictors appear to have a significant relationship to the response?

Upon inspection of the Corrplots of numeric variables the following observations are evident:

- studytime and Medu appear to have the most significant positive correlation.
- Failures, Dalc, and absences appear to have the most significant negative correlation.

Look at the "Kitchen Sink" models created (with no interactions) similar correlations are noted. The following significant correlations can be noted.

G1.x Maths

```
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
Medu        0.41310    0.15096   2.737 0.006509 **
failures.x  -1.42994    0.23438  -6.101 2.67e-09 ***
sexM        1.24597    0.33771   3.689 0.000259 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

NOTE: sexM was not noted in the corrplot as it is not a numeric variable

G1.y Portugese

```
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
studytime    0.38575    0.13981   2.759 0.006084 **
Medu        0.38391    0.10616   3.616 0.000340 ***
failures.y  -0.58943    0.23508  -2.507 0.012592 *
Dalc       -0.40210    0.13108  -3.068 0.002318 **
higheryes    1.98276    0.55304   3.585 0.000382 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

NOTE: higher was not noted in the corrplot as it is not a numeric variable.

3.2 Part B

What suggestions would you make to a first-year student trying to achieve good grades?

Looking at the significance of each variable in the "Kitchen Sink" models (see above), and also looking at what is actionable, I would suggest the following:

- Students finding themselves failing in class should get back on track as quickly as possible.
- Parents especially Mother's may have a positive role to play in helping children achieve good grades.
- Drinking alcohol on school nights may negative impact on your grades.
- Female students may need extra support in achieving good Mathematics grades.
- Students who are unsure of whether they want to pursue higher education may need extra support.

3.3 Part C

*Use the * and : symbols to fit models with interactions. Are there any interactions that are significant?*

The most notable interactions were a minor interaction with Medu (Mother's Education Level), had with address (Urban or Rural), and also absences from class.

Additionally, there was a significant relationship between Dalc, (Weekday Alcohol Consumption), and school.

Before speculating what these relationships are it would be prudent to explore them further at another time.

First Period Grade (Maths) "Kitchen-Sink" Model (with all interactions)

- Medu:addressU *
- Dalc:schoolMS:sexM **
- Medu:absences.y *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

First Period Grade (Port) "Kitchen-Sink" Model (with all interactions)

- Medu:addressU *
- Dalc:schoolMS:sexM **
- Medu:absences.y *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1