UNIVERSITÀ
DEGLI STUDI
DI MILANO

Department of Economics, Management and Quantitative Methods

Analyzing the Impact of Alcohol Consumption on Student Academic Performance

RIJIN BABY
954007

Data Science and Economics

Statistical Learning Project

Contents

1 Introduction

This analysis examines whether alcohol consumption has any predictive power over student average grades. Also, look for other factors that are important predictors of student grades. Just a word of warning, the goal is not to predict student grade evolution over marking periods since all the features in the data-set (other than grades) remain constant over marking periods and are general descriptors of student backgrounds.

One part of the analysis focus on supervised learning techniques like linear regression, regression trees and random forest, where we build prediction models to identify the significant variables that contribute for the grade score. The goal of this section is to compare the different supervised techniques performances and evaluate which model performs better.

Second part is Principal component analysis, that allows to summarize the dataset with a smaller number of significant variables, that are able to explain the largest variability of the original set. The aim of this preliminary exploratory analysis is to obtain principal components and exploit them to visualize data in a low-dimensional representation, that captures the largest amount of information.

2 Data set and data pre-processing

2.1 Context

The data were obtained in a survey of students math and portuguese language courses in secondary school. It contains a lot of interesting social, gender and study information about students. Data source: UCI Machine Learning Repository

2.2 Data Dictionary

There are two data-sets student-mat.csv (Math course) and student-por.csv (Portuguese language course). Attributes are as follows:

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')

13. traveltime - home to school travel time (numeric: 1: less than 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - greater than 1 hour)
14. studytime - weekly study time (numeric: 1 - less than 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - greater than 10 hours)
15. failures - number of past class failures (numeric: n if 1¡=n¡3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)
31. G1 - first period grade (numeric: from 0 to 20)
32. G2 - second period grade (numeric: from 0 to 20)
33. G3 - final grade (numeric: from 0 to 20, output target)

student-mat.csv has 395 records and student-por.csv has 649 records. There is higher number of numerical variables, which is good and a couple of categorical ones, which need to be transformed into numerical ones later on. Good news is that the categorical variables have only few levels, so we do not need to care to much about creating too many flag variables.

2.3 Pre-processing

Performed all basic data cleaning steps like missing data check, unique value check, primary key check, duplicate check, spelling check, value range check. No mistakes found, data set is clean and good to go.

The average of the three grades is set as the target variable.

2.3.1 Variable that need to be removed

Variable "failures" is closely related to my target variable, avggrades. Since past failures and avg-grades represent the same general student aptitude (thus it is rather a target rather than a feature), I am inclined to remove variable "failures" from the dataset.

Also, variable freetime (free time after school) which is highly correlated with variable goout (going out with friends)

2.3.2 Comparing 2 subject data

Before merging the two data sets, we need to make sure if math and Portuguese grades are comparable or not. There is a number of students who are repeated in both data sets, so I would use their records to examine if math and Portuguese grades are comparable. There are 85 students who belong to both tables, and I am going to examine their average math and Portuguese grades a test case.

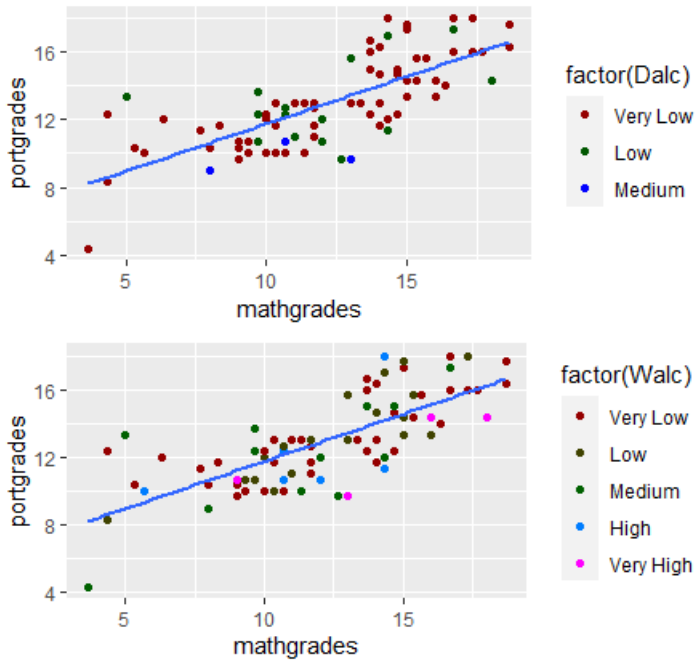


Figure 1: Grade comparison across subjects

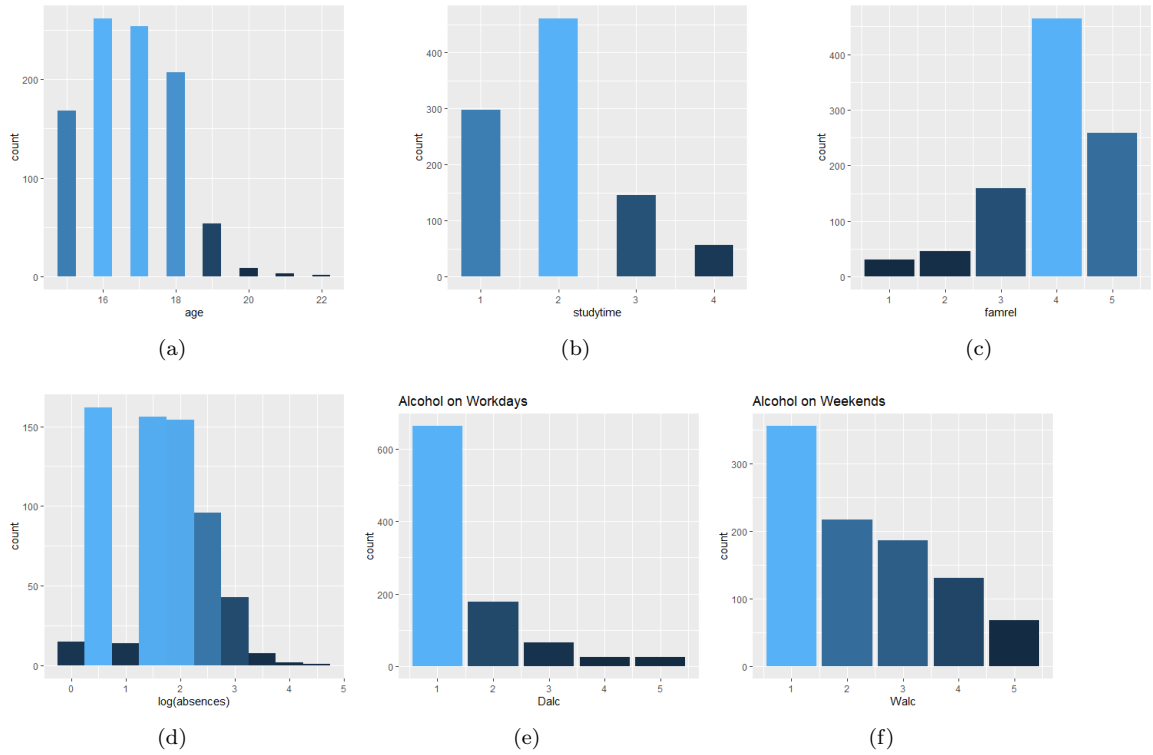
The two scatter plots have few implications. First, among the 85 students, no one consumed high or very high levels of alcohol on daily basis. Second, almost all of those who earned relatively high scores consumed very low levels of alcohol on weekdays. Third, math and Portuguese grades seem to correlate highly with each other. When I regress Portuguese grades on math grades, the adjusted R-squared is 0.55. This means that the correlation coefficient between math and Portuguese grades is about 0.74 and that about 55% of the variation in Portuguese grades can be explained by the variation in math grades. In my view, this is an indication that I can go ahead and combine the two tables together without worrying much about the subject matter, average grades in math or Portuguese reflect general student aptitude. Hence combine both the data sets and take the average of grades for further analysis.

3 Data Exploration through Plots

3.1 Univariate Plots

Basic interpretation of plots:

- we can see that close too 600 students which is more that half the population considered spend 2-5 hours per week studying
- Majority students have excellent relationship with their family which is a nice statistic to find out
- Majority of the number of students are having less absence



(a) (b) (c)

(d) (e) (f)

Figure 2: (a) Age (b) Weekly Study Time (c) Family Relashionship (d) Absence (e) Alcohol on workday (f) Alcohol on weekends

- Workday consumption of alcohol is very minimal among students
- Consumption of alcohol on weekends are higher compared to weekdays

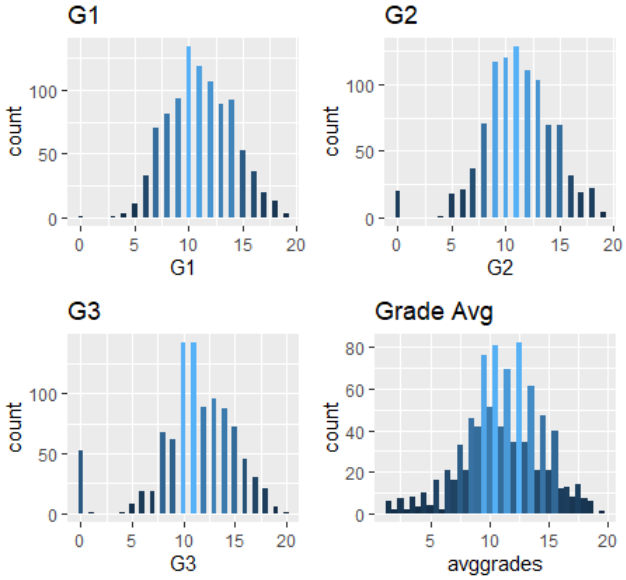
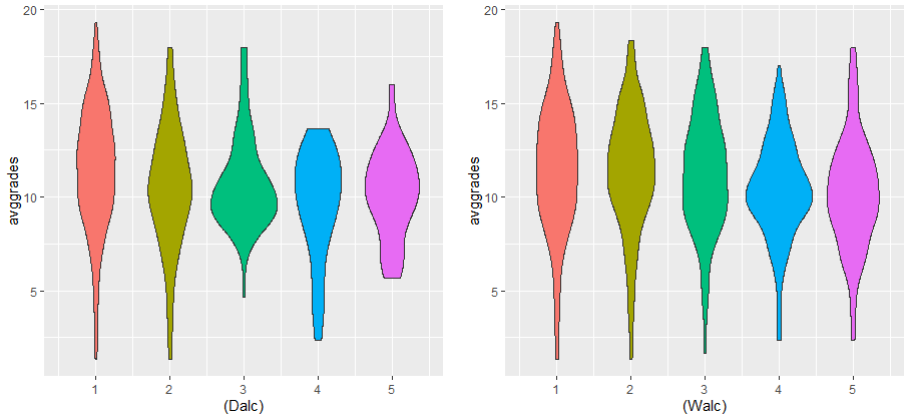


Figure 3: Grade Distribution



(a) (b)

Figure 4: (a) Workday Alcohol Consumption (b) Weekend Alcohol Consumption

3.2 Bivariate Plots

Figure 4(a) is an interesting visualization that tells us that students who dont consume alcohol on daily basis(1 score) have both, least and the highest scores population distributed in both ends, the main difference that shall catch an eye is that students those who consume alcohol more frequently on workday are not the top notch grade achievers as there is no one who has a score(grade) over 13 of 20 and talking about population spread, looking at the width we can say that there is equal distribution at the means for all kinds of students(1 to 5)

Figure 4(b) The main difference that shall catch an eye is that students those who consume alcohol more frequently on weekends are almost equal to top notch grade achievers and none of them who consume alcohol frequently on weekends have a score lesser than 5

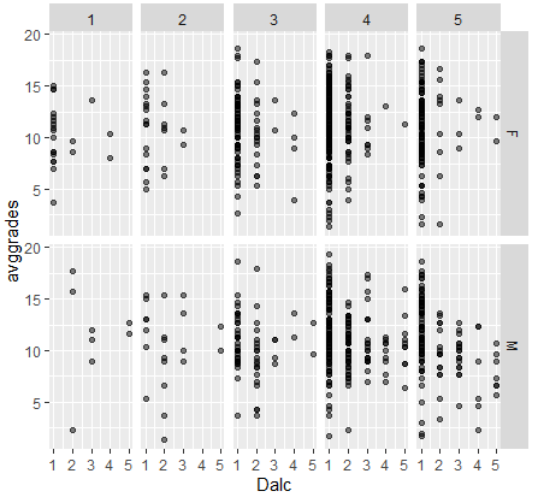


Figure 5: Family relation, student gender and workday alcohol effect on grade

This seems to be a pretty obvious plot because if a student consumes alcohol on workdays, they they are going to perform poorly, as we can see none of the students who belong to a score of 5 at Dalc score top grade.Also, keeping in mind the density of plots at Dalc 1 we can say that there is a spread of population of students of all sorts like the ones who perform best and worse.

4 Supervised Learning Techniques

4.1 Linear Regression

Linear regression is a linear approach to modelling the relationship between a dependent variable and one or more independent variables. Implementing on our data gives the following results:

```
Linear Regression

959 samples
 29 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 863, 863, 863, 864, 862, 864, ...
Resampling results:

  RMSE      Rsquared   MAE
  2.991039  0.1531568  2.293569
```

Figure 6: Cross Validated LR Summary

Adjusted R-squared in the above regression is only 0.17, which is quite low. It implies that only 17% of the variation in the average grades is explained by the variation in everything else. The variables that have statistically significant impact on average grade are studytime, schoolsup (extra educational support), paid, and higher (wants to take higher education) etc.

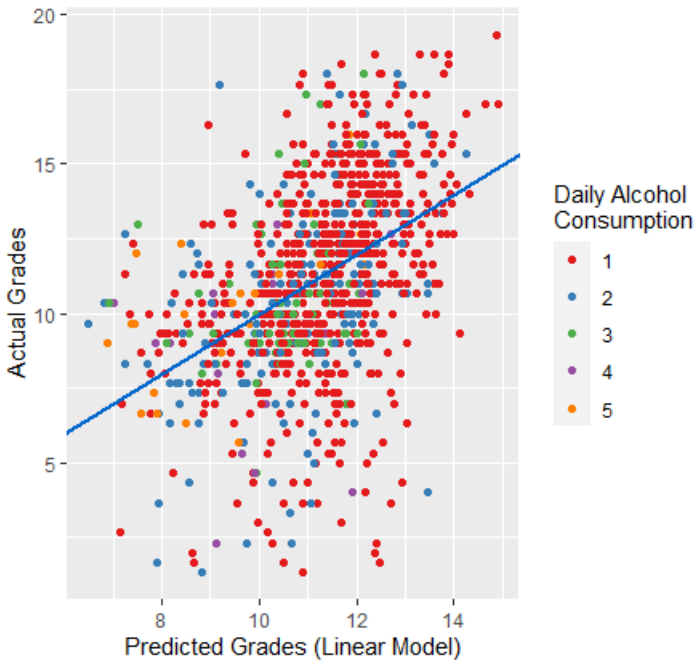


Figure 7: Linear Model Prediction

4.2 Regression Tree

A regression tree is built through a process known as binary recursive partitioning, which is an iterative process that splits the data into partitions or branches, and then continues splitting each partition into smaller groups as the method moves up each branch. Implementing on our data gives the following results:

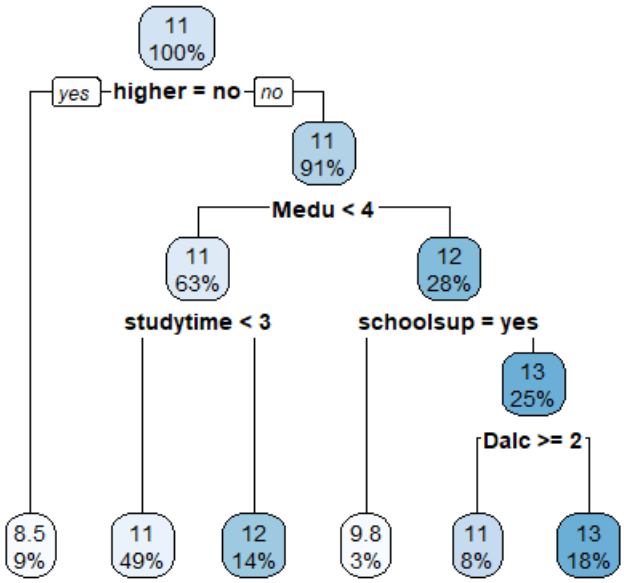


Figure 8: Regression Tree

According to the regression tree analysis, the variable that seems to be important in "higher" that indicates whether the student wants to pursue higher education. The overwhelming majority of surveyed students would like to pursue higher education and their average grade (11/20) is significantly higher than the average grade of those who don't (8.5/20). Regression tree analysis reveals that mother's education is another important feature (interestingly, this feature did not come up as important in the linear regression model). Students whose mothers had higher education has higher grade (12) than the students whose mothers do not (their average grade was 11). I would like to take the first stab at evaluating the relative predictive performance of the two models. It seems that the linear model performs better than the regression tree.

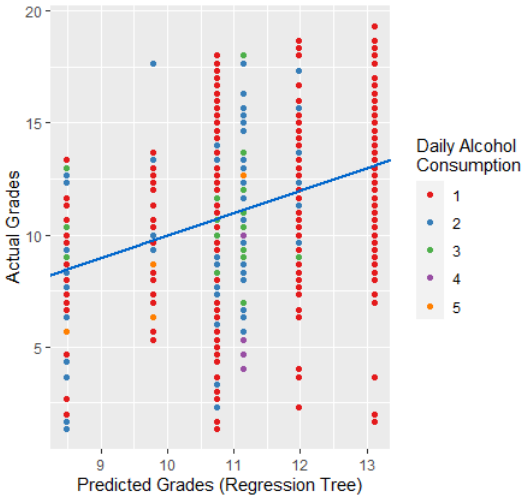


Figure 9: Regression Tree Prediction

In Figure 7 and Figure 9, horizontal axes represent predicted grades while the vertical axes

represent true grades. If the model is accurate in predicting actual grades then predicted grades must be equal to actual grades and thus the scatter points should line up along the 45 degree (blue) line. Unfortunately, as the RMSE value 2.9 and error plots indicate, neither of the two models seems to do a decent job in predicting student average grades. Unsatisfied with how linear regression and regression tree models perform, I am going to take it up a notch and give a random forest a try.

4.3 Random Forest

Random forests are an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees. Implementing on our data gives the following results:

```
> rF_student

Call:
 randomForest(formula = avggrades ~ ., data = student_final, ntree = 500,      importance = T)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 9

          Mean of squared residuals: 8.286178
                    % Var explained: 20.37
>   RMSE(rf.predictions,student_final$avggrades)
[1] 1.459193
```

Figure 10: Random Forest Model

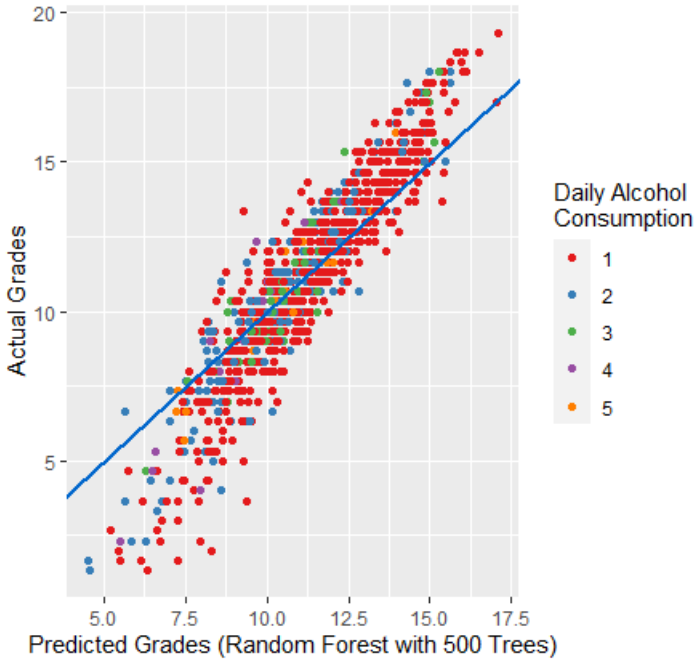


Figure 11: Random Forest Prediction

Even though, random forest seems to systematically under predict the grades of low grade earners and over predict the grades of high grade earners, overall, random forest seems to be a much better predictor of average grades than either the linear regression or regression tree model. 10 fold cross validation confirms that out of the three models that I executed (Linear Regression, Regression

Tree and Random Forest), the Random Forest model with 500 trees is indeed the best predictor of average student grades.

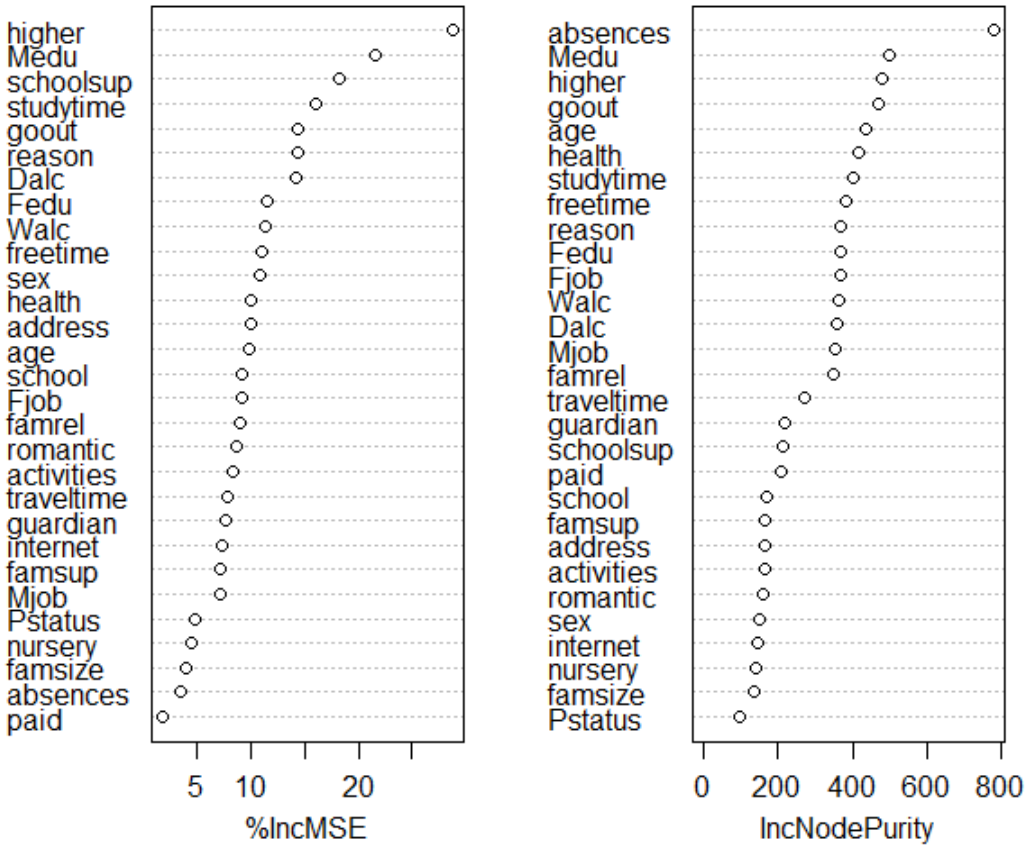


Figure 12: Variable Importance Plot

The top 10 most important variables that impact student average grades are:

- higher- wants to take higher education (binary: yes or no)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 - secondary education or 4 – higher education)
- studytime - weekly study time (numeric: 1 - less than 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - greater than 10 hours)
- schoolsup - extra educational support (binary: yes or no)
- Dalc - workday alcohol consumption (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)
- reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

- sex - student's sex (binary: 'F' - female or 'M' - male)

These results imply that both weekday and weekend alcohol consumption are important predictors of student average grades. Removing either of these two variables will increase the MSE of predictions by between 10-20%.

What's interesting about these results is that some features that would be conventionally thought as important did not end up in the top ten list (I would speculate that the variables such as Pstatus, famsupport, famrel, and absences are among those) while some other variables (such as higher and Medu) turned out to be very important.

As a conclusion, the impact of the most important 2 variables, "higher" and "Medu," on average student grade is as follows:

- higher: willingness to pursue higher education increases average predicted grade from 8.5 to 11.4. Thus, motivate your kids to pursue higher education is the best thing you can do to improve their grades in school!
- Medu: increase in mother's education from none to more than secondary education increases predicted average grade from 10.8 to 11.5. Thus, advice to future parents (especially fathers): If you want your future children do well in school marry someone educated.

5 Unsupervised Techniques: PCA

5.1 Overview

PCA produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated. Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

PCA is used to identify variables in a dataset that represent the most information about the dataset. In this dataset a variable that has a lot of information is e.g. age, because it contains students from 15 to 22 and it is more or less normally distributed. A variable that matters less is e.g. paid (extra paid classes within the course subject). 94% of the students do not have these extra classes. A more extreme case is if I would add the variable country, all students are in the same country, so it does not add any value in describing a student.

PCA ignores all these real world representations and just calculates new variables(principle components) that are optimized for the highest variance in the new variable

5.2 Feature selection based on correlation

Let's see how many variables are actually suited to predict average grades.

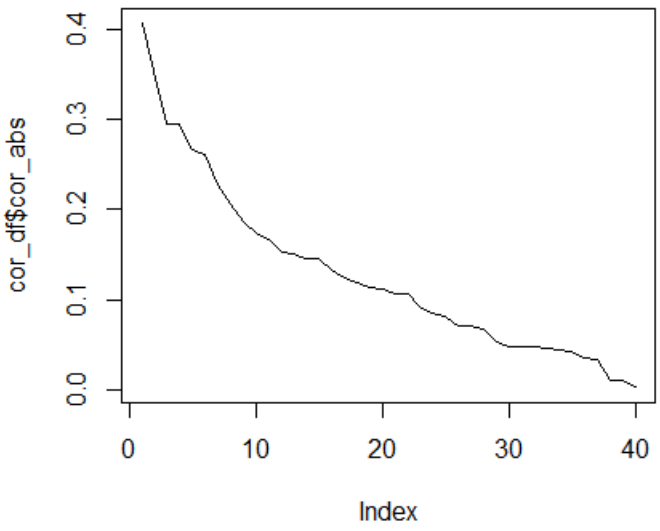
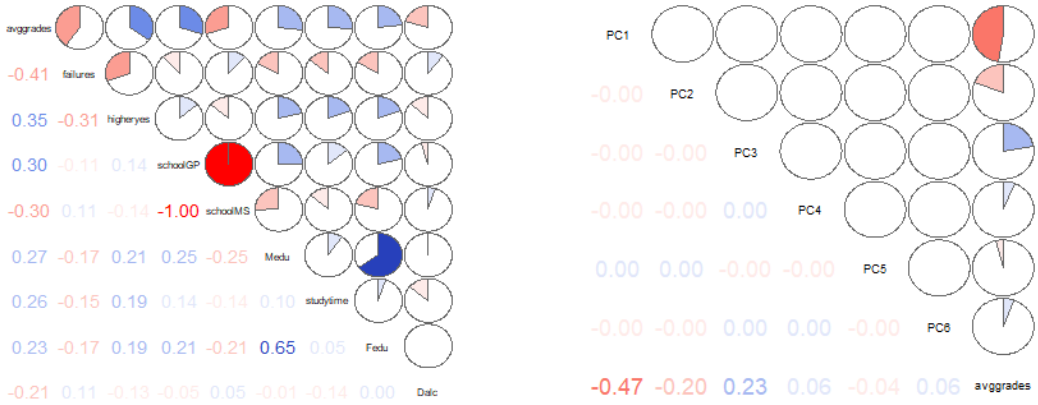


Figure 13: Correlation

Since this is a social study we should not expect really high correlation (predicton power) for each variable, but roughly 8 variables with a correlation between 0.2 and 0.4.

In general we see that all these variables are more or less correlated with each other and we would introduce a lot of multicolinearity if we just throw all of them in a model to predict y (grades). You see that schoolMS and schoolGP or Fedu(father education) and Medu(mother education) are highly correlated and it would be good if we could eliminate them, without losing their prediction power. That is exactly our use case for dimensionality reduction with PCA. Keep this matrix in mind, later on we will compare it with the principle components.



(a) (b)

Figure 14: (a) Correlation Matrix From the most relevant variables (b) Correlation of PC's

5.3 Principle Components

Our data is already transformed into a format that enables us to execute prcomp (principle component analysis). But before we do this, we remove our target variable, since we do not want to mingle this with the principle components. The plot shows us how much variance of the dataset is explained by the 1st, the 2nd ... 8th principle component. We already see that e.g. the 8th PC is really close to zero and therefore unimportant for the dataset.

```
>     summary(pca)
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7       PC8
Standard deviation     1.5942 1.1957 1.1238 0.9422 0.9189 0.82716 0.59139 2.645e-15
Proportion of Variance 0.3177 0.1787 0.1579 0.1110 0.1055 0.08552 0.04372 0.000e+00
Cumulative Proportion  0.3177 0.4964 0.6542 0.7652 0.8708 0.95628 1.00000 1.000e+00
```

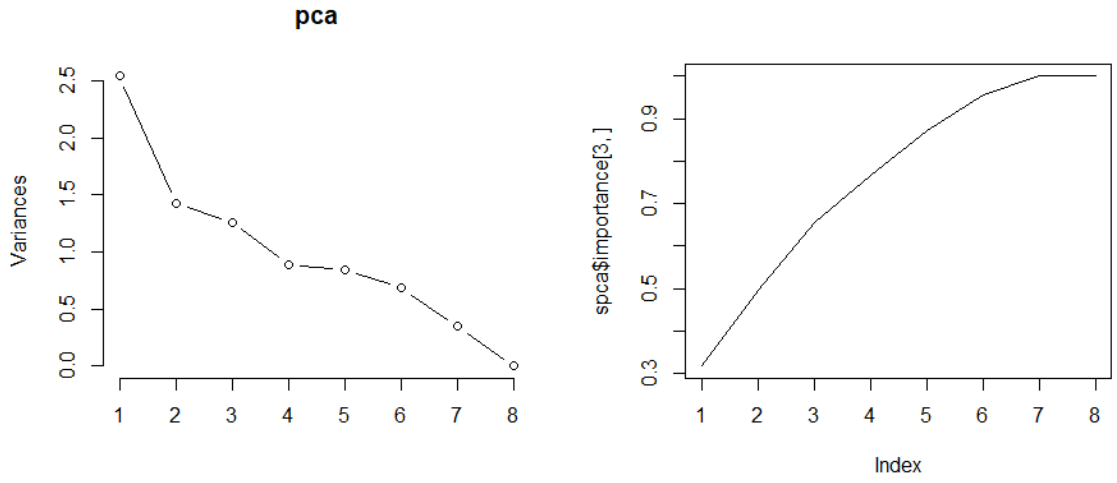
Figure 15: PCA Summary

We can see that the first 6 principle components togehter explain 96% of the variablity in the data. Based on this it, we can forget about PC7 and PC8. We also see that the first three principle components represent already 65% of our data. See a graphical representation of this below.

The outcome of PCA is not only that we might be able to reduce the dimensionality of our data, but also that we get rid of correlation effect between explanatory variables. For the matrix below I removed PC7 and PC8, as well as added our target variable.

Also, look at Figure 14(a) our first correlation matrix with all the correlation between the variables and Figure 14(b) Well the outcome of PCA is not only that we might be able to reduce the dimensionality of our data, but also that we get rid of correlation effect between explanatory variables. For Figure 14(b) PC7 and PC8 are removed, and added our target variable.

However, since a proportion of variance near 95% requires 6 components and given that to perform the PCA we have excluded many attributes, Using PCA for prediction model result in huge lose of information. But out of curiosity I am buiding a linear regreesion model to see the results and make sure that PC's are the best.



(a) (b)

Figure 16: (a) PCA Plot (b) Cummulative Variance by PC's

5.4 Linear Regression Using PCA

Can we use the PCA to predict the average grades better? Yes, we can! Look at the adjusted R-squared, it is 0.31 compared to 0.17 (before PCA).

```
Call:
lm(formula = avggrades ~ ., data = pca_df)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1534  -1.5366  -0.1405   1.4573   6.4247

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.62506    0.09200 126.360  < 2e-16 ***
PC1         -0.84263    0.05775 -14.590  < 2e-16 ***
PC2         -0.46228    0.07700  -6.003 3.23e-09 ***
PC3          0.56803    0.08193   6.933 1.01e-11 ***
PC4          0.17897    0.09772   1.832   0.0675 .
PC5         -0.12537    0.10020  -1.251   0.2113
PC6          0.19677    0.11131   1.768   0.0776 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.344 on 642 degrees of freedom
Multiple R-squared:  0.3221,    Adjusted R-squared:  0.3158
F-statistic: 50.84 on 6 and 642 DF,  p-value: < 2.2e-16
```

Figure 17: Linear Regression using PCA

Using PCA we end up with cryptic variables that nobody understands and that do not really mean something. However, we have some means to still understand what a PC is. Take a look at the following biplot of PC1 and PC2 (explaining 50% of the data):

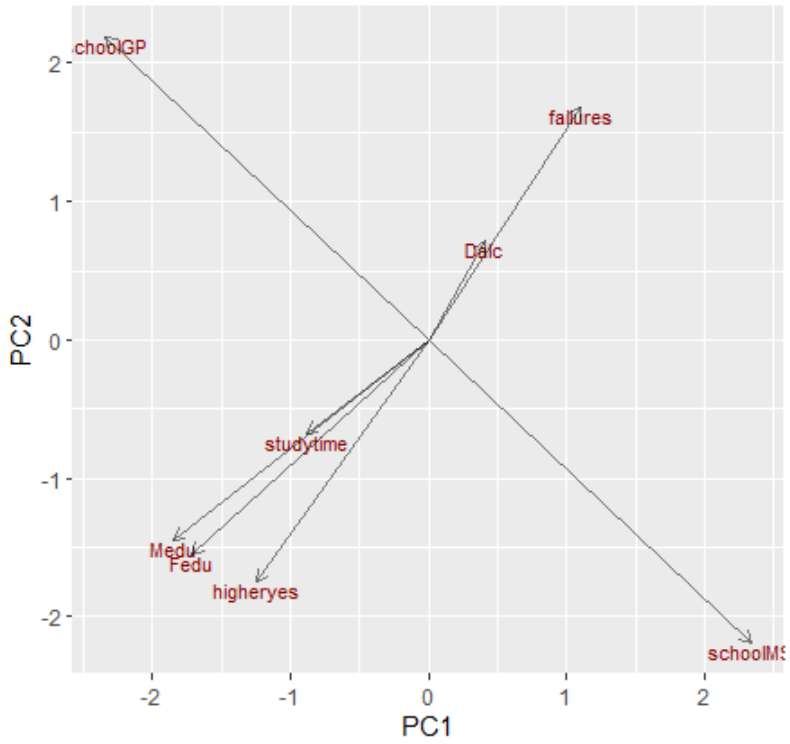


Figure 18: biplot of PC1 and PC2

6 Conclusion

The results using the supervised learning techniques imply that both weekday and weekend alcohol consumption are important predictors of student average grades. Removing either of these two variables will increase the MSE of predictions by between 10-20%.

PCA is a powerful tool to reduce dimensionality or to get a different perspective on your data. At the same time the interpretation of results is more difficult, but possible e.g. with the biplot. With PCA we do not lose prediction power, but we are able to eliminate collinearity.

What's interesting about these results is that some features that would be conventionally thought as important like Pstatus, famsupport, famrel, and absences did not end up in the top list, while some other variables such as higher education and mother's education turned out to be very important. For the grades in school we can say that strive for higher education, longer studytime have a positive relation with your grades, while alcohol consumption during the week and past failures in exams indicate you have worse marks.

Refer Github Repository for the R script.