STAT 311 Project

**INTRODUCTION**

As the accuracy of machine learning has been improving, it is important that there will be symbiotic relationships between human and AI in the future, as MIT media lab puts it, “Extended Intelligence” (Ito, J.2011). Since conducting good education is the foundation for the realization of this future, investing the data of a secondary school seems important.

**DATA**

The data set contains the first, second and third grades of a secondary school in Portugal(Paulo Cortez, University of Minho, GuimarÃ£es, Portugal).The number of observation and the number of predictors are 395, 33 respectively.

|  |  |
| --- | --- |
| Name of variables | Details |
| school | student's school(binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira) |
| sex | student's sex(binary: 'F' - female or 'M' - male) |
| age | student's age (numeric: from 15 to 22) |
| address | student's home address (binary: 'U' - urban or 'R' - rural) |
| famsize | family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3) |
| Pstatus | parent's cohabitation status (binary: 'T' - living together or 'A' - apart) |
| Medu | mother's education(numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education) |
| Fedu | mother's education(numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education) |
| Mjob | mother's education(nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other') |
| Fjob | mother's education(nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other') |
| reason | reason to choose this school(nominal: close to 'home', school 'reputation', 'course' preference or 'other') |
| guardian | student's guardian (nominal: 'mother', 'father' 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<=n<3, else 4) |
| schoolsup | extra educational support(binary: yes or no) |
| famsup | family educational support(binary: yes or no) |
| paid | extra paid classes within the course subject (Math or Portuguese) (binary: yes or no) |
| activities | extra-curricular activities (binary: yes or no) |

|  |  |
| --- | --- |
| nursery | attended nursery school |
| higer | wants to take higher education |
| Internet | Internet access at home (binary: yes or no) |
| romantic | with a romantic relationship (binary: yes or no) |
| famrel | quality of family relationships(numeric: from 1 - very bad to 5 - excellent) |
| 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) |
| Dalc | workday alcohol consumption(numeric: from 1 - very low to 5 - very high) |
| Walc | weekend 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, output target) |

\*Figure: description of data; Source *Student Performance Data Set*

**Methodology and Application**

Order of methods applied to the data

1. Multiple regression analysis

2. Ridge regression and Lasso using cross-validation

3. PCA

4. Artificial neural network

5. Logistic regression, LDA, QDA and KNN

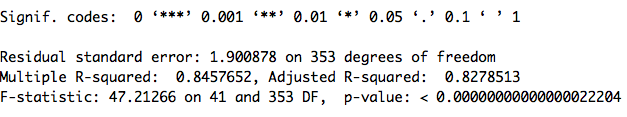
6 Factor analysis

7. Regression tree, bagging and random forest

8. k-means clustering and hierarchical clustering

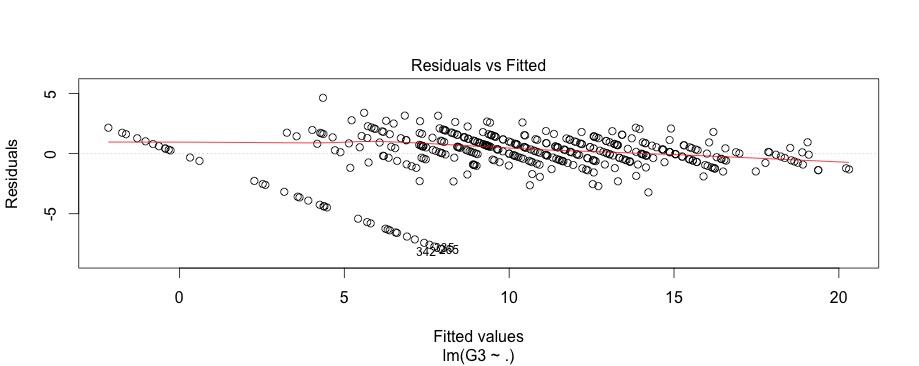
**1 Multiple regression analysis**

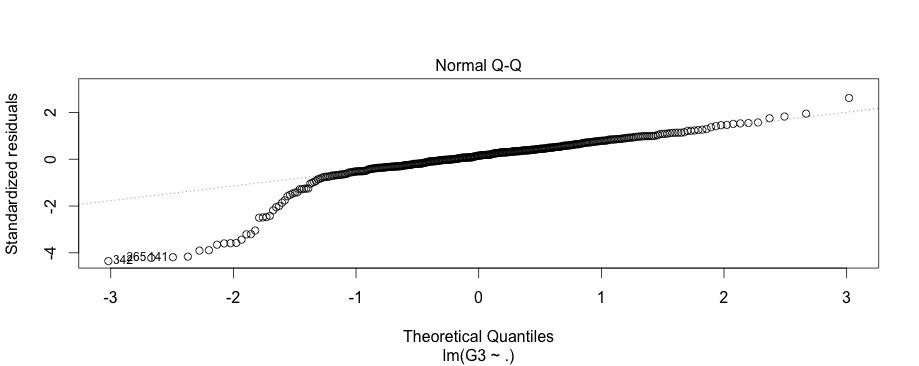
We fit multiple regression to our data set with G3 as a response and other variables as predictors. Below are summary and plots of the fitted model.

******

Test MSE \*\*data is split equally into two; train data set and test data set

[1] 38.59688

******

******

Since regression diagnostics are useful for discovering the problems in the data, which can be decomposed into 6 problems:

1-Nonlinearity,

2-Correlation of error terms

3- Non-constant variances in the errors

4- Outliers

5 High leverage points

6 Collinearity

*Results of the tests for above problems are follows;*

*1 Non-linearity*

*From the first plot, it is clear that data is non-linear.*

*2 Correlation of error terms*

*> durbinWatsonTest(Multreg)*

*lag Autocorrelation D-W Statistic p-value*

*1 0.03681494 1.925051 0.36*

*Alternative hypothesis: rho != 0*

*From the test for correlation, there is little correlation among error terms*

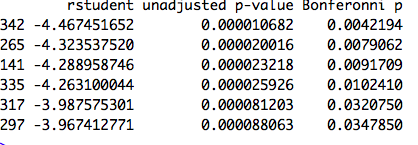
*3 Non-constant variances in the errors*

*Based on the first plot, it is hard to say that this data is heteroscedasticity, the non-presence of funnel shape.*

*4 & 5:Non-normality and possible outliers and high leverage points*

*As shown above qq-plot, this data is skewed, and there is a high possibility that outlier and leverage points exist. Below are further explorations of the data in terms of outliers and high-leverage points.*

*\*\*Bonferroni Outlier Test(p<0.05)*

**

*\*\* suspectful high leverage(greater than (p+1)/n)*

*> highLeverage=hatvalues(Multreg)>(33+1)/395 ## suspectful high leverage(greater than (p+1)/n)*

*> sum(highLeverage==TRUE)*

*[1] 273*

*> sum(highLeverage==TRUE)/395\*100##Percentage of doubtful high leverage points*

*[1] 69.11392405*

*As shown above, there are six outliers according to Bonferroni Outlier Test, and 273 high leverage points, which is about 69 % of data.*

*Below is the computation of variables which are both outliers and suspicious high leverage points*

*> overlap=calculate.overlap(x=list("outlier"=outlier,"levearge"=Leverage))*

*> overlap*

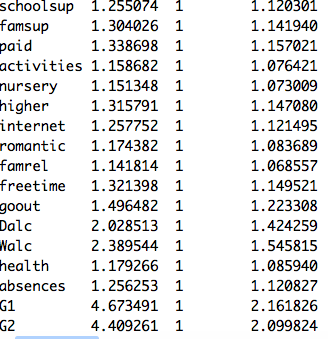
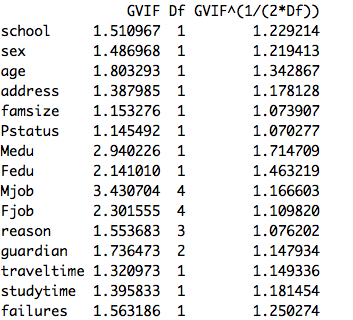
*$a3*

*[1] 141 297*

*It is possible that these two points, 141,297, have large impacts on the least square lines .*

*6. Collinearity*

*This data does not contain collinearity since none of the variables have VIF larger than 5 or 10.*

**

*Given the violation of linear assumptions, standard error of this data set can be predicted better using bootstrap which does not reply on these assumptions.*

*Bootstrap*

*\*\*Standard error estimate using bootstrap (with predictors having significant p-value)*

*Call:*

*boot(data = d1, statistic = boot.fn, R = 1000)*

*Bootstrap Statistics :*

*original bias std. error*

*t1\* -3.40923281 -0.018217232 0.64653115*

*t2\* 0.34252230 0.003824079 0.11231558*

*t3\* 0.03805925 0.002039171 0.01362069*

*t4\* 0.14183194 -0.002858375 0.03989649*

*t5\* 0.99953209 0.002652850 0.03191214*

*Coefficients:Multiple linear regression Fitted model(with predictors having significant p-value)*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) -3.40923 0.53756 -6.342 6.28e-10 \*\*\**

*famrel 0.34252 0.10687 3.205 0.00146 \*\**

*absences 0.03806 0.01195 3.186 0.00156 \*\**

*G1 0.14183 0.05510 2.574 0.01042 \**

*G2 0.99953 0.04862 20.557 < 2e-16 \*\*\**

*---*

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

As shown above, standard errors using bootstrap differs slightly from the ones using a fitted model.

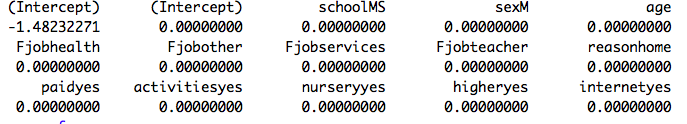
***2. Ridge regression and Lasso using cross-validation***

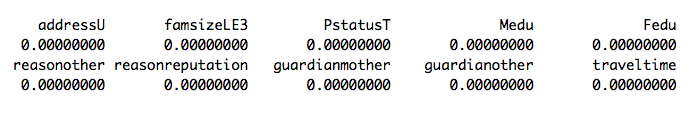
Test MSE :

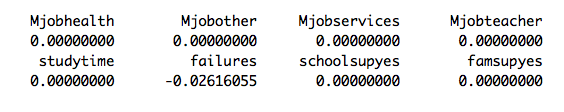
Ridge Lasso

4.375597 4.006323

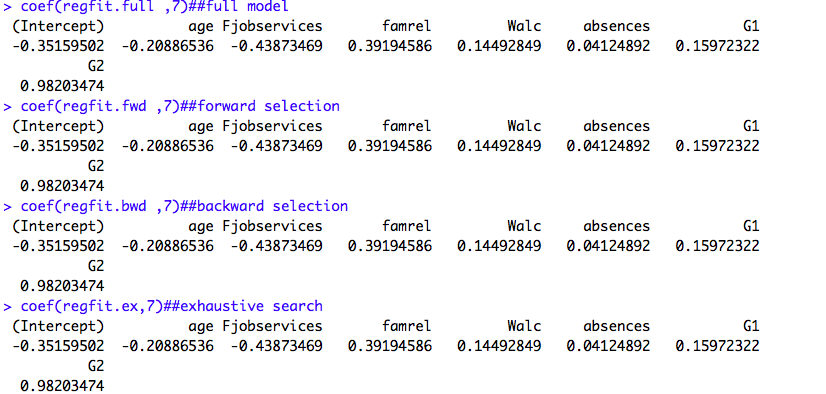
\*\*coefficient estimate of lasso



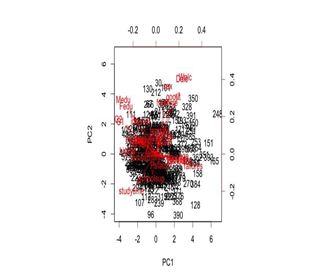




As shown above, lasso with cross-validation uses only one variables, failures. It is clear from the Test MSE of two methods that most of the predictors do not have significant impact on response since the lasso assumes that response is a function of a few variables.

\*\*Feature selection: full model, forward selection, backward selection and exhaustive search 

As shown above, the coefficient estimates using four methods show the same result, which is contrary to usual assumption that each method shows different result.

**3. PCA**

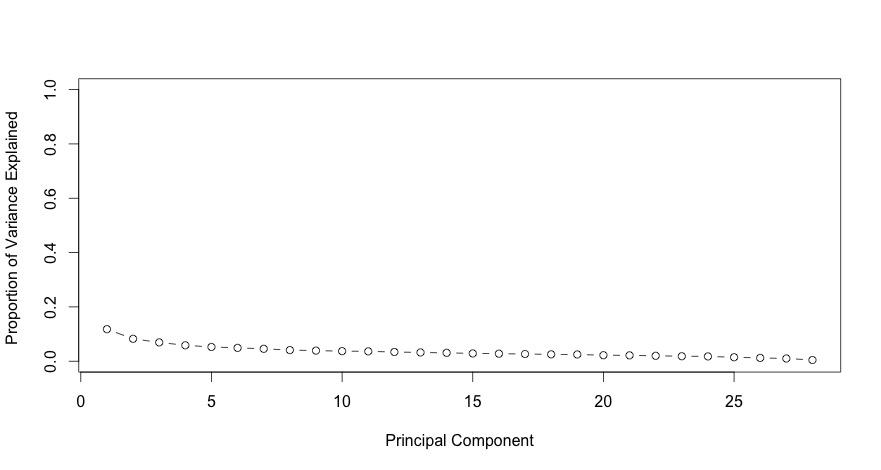
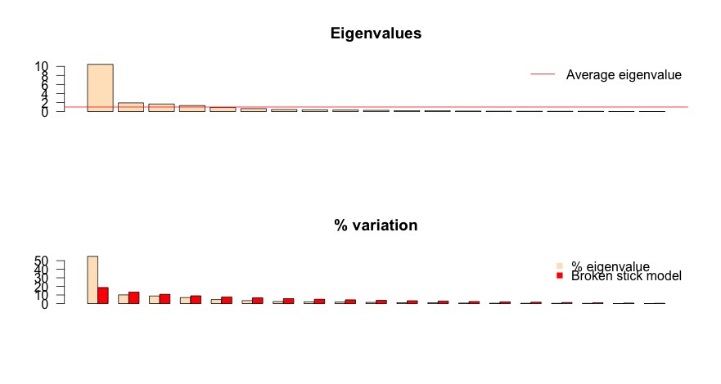


Figure: Proportion of variance explained. A plot is almost flat and each PCAs do not explain this data set so much.



\*\*Plots of eigenvalues \*\*# Usage: evplot(ev) where ev is a vector of eigenvalues

# License: GPL-2 # Author: Francois Gillet, 25 August 2012

As shown above, PCAs above red line(1) are first 4 which should be kept according to Kaiser criterion.

**4. Artificial neural network**

In this method, data is randomly split into two parts, one for training data and the one for test data set.

As a comparison to an artificial neural network, multiple linear regression with numeric variables is fitted first.

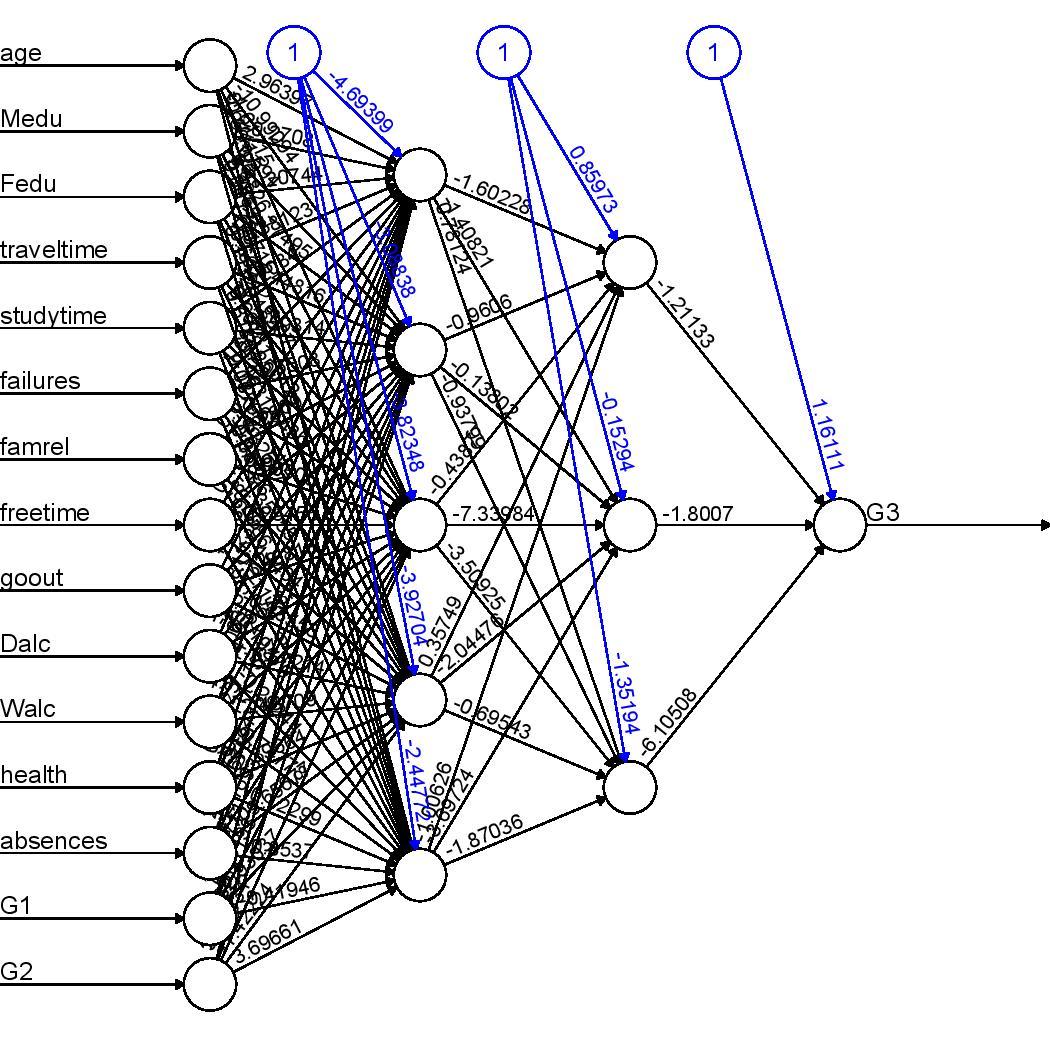


Figure: A plot of neural network with two hidden network, each has five and three neurons.

Test MSE

lm: 4.33279789720453 NN:2.2616425787414

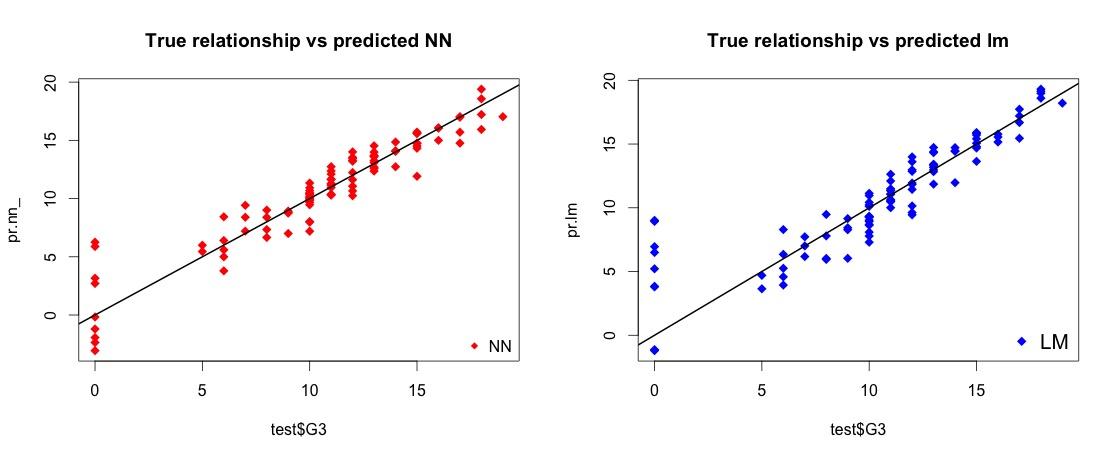


Figure:Plots of predicted neural network and predicted linear regression. Since this is a simulated data, a true line can be shown. (It is hard to measure the goodness of fit of two models since two plots are similar.)

Since test MSE might hugely depend on data split (high variance), use 10 fold cross-validation to estimate the average testMSE.

Screenshot 2016-04-06 21.21.41.png

Screenshot 2016-04-06 21.23.20.png

Screenshot 2016-04-06 21.23.52.png

Even though there are some variability associated with the split, the average of cross validation error, 2.13, is less than that of linear regression model.

In terms of interpretability, Artificial neural network does not perform well. Screenshot 2016-04-06 21.18.18.png

**5. Logistic regression, LDA ,QDA and KNN**

splitting G3(final year’s grades) into pass(G3>10) and fail(G3<=10),LDA and QDA are applied to this data,using cross validation

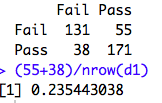
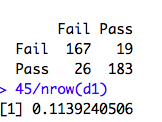
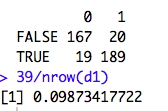
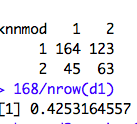


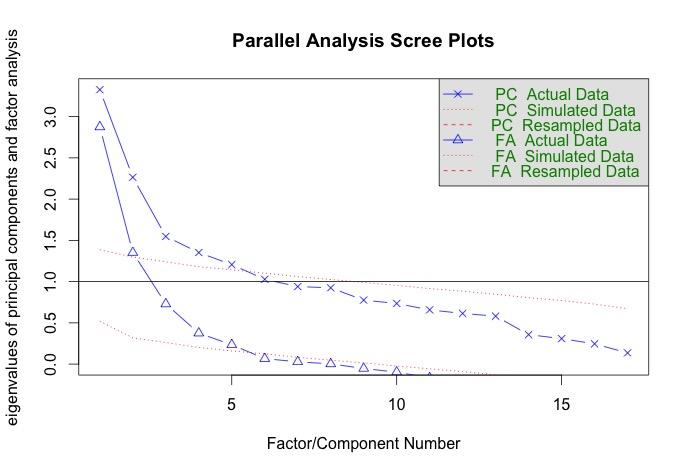
Figure: from left : LDA QDA Logistic regression KNN

Figure: classification tables of LDA, QDA and logistic regression with all of the predictors

Logistic regression leads to the smallest misclassification. This is because Gaussian assumptions is violated. LDA results in a better misclassification rate than QDA. This result suggests that

a true boundary is less flexible and QDA suffers from high variance. Since the training set is small relative to variables, reducing the variance results in a better result. KNN is the worst estimate since it is a non-parametric method and it suffers from high variance when true decision boundary is linear.

**6. Factor analysis**



## Screenshot 2016-04-07 01.10.13.png

## Figure: determinant of the number of factors

Screenshot 2016-04-07 01.12.33.png

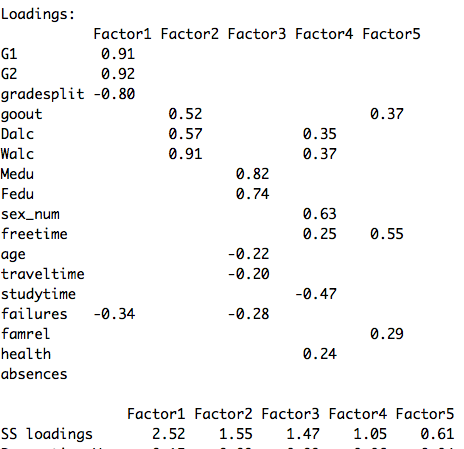


Figure: summary of five factors

Factor 1is highly associated with G1 and G2, which are previous grades. Factor 2 is associated with alcohol consumption. Factors 3 is associated with parents’ education. Factor 4 is associated with gender and alcohol. Factor 5 is associated with free time and friends.

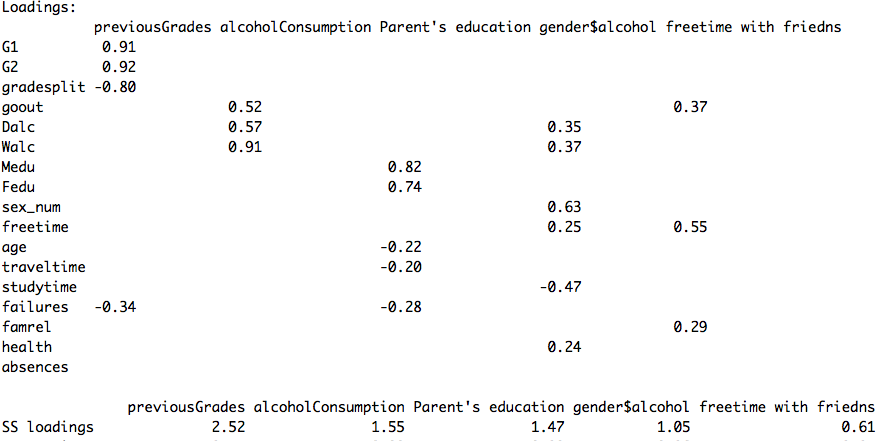
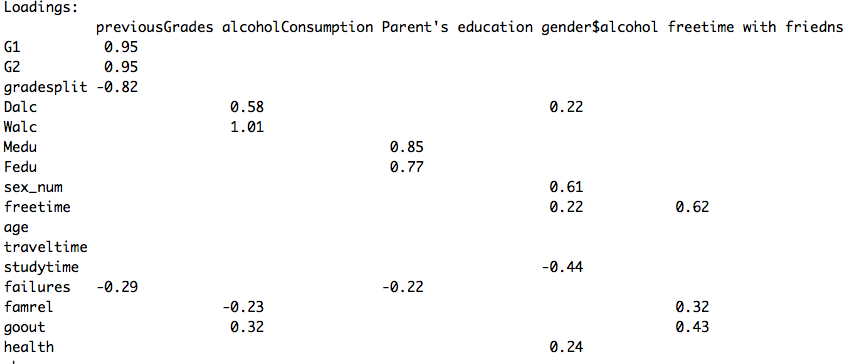


Figure: summary of five factors named after the association with variables

This data can be more interpretable using promax in order to allow variables to be correlate with factors.



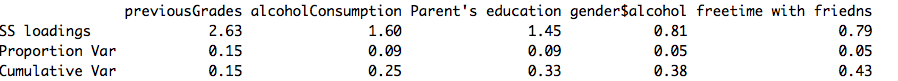


Figure: A summary of an oblique promax solution

**7. Regression tree, bagging and random forest s**

Test MSE

Regression tree: 10.18410744

Bagging: 10.21283965

Random forests: 10.12306077

Random forest results in the smallest Test MSE. This result seems reasonable since random forests decorrelates the tree and reduces the variance.

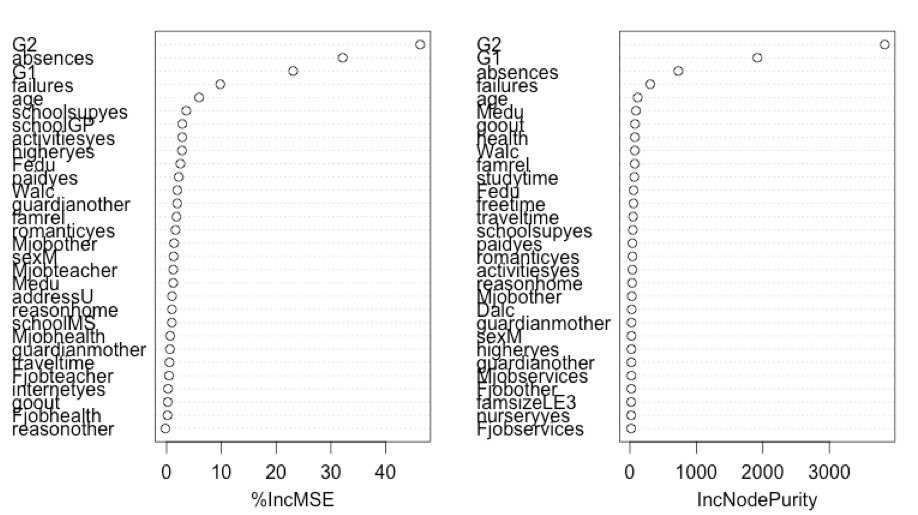
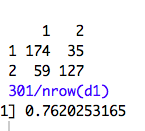
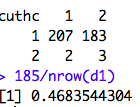


Figure: Plot for importance of variables. G2, G1, absences are quite important variables in both methods.

**8. k-means clustering and hierarchical clustering**



Figure

Left: miss-classification table of k-means clustering on scaled data

Right: miss-classification table of hierarchical clustering on scaled data

References:

Ito.J.(2016).*Extended Intelligence*.Retrieved April 7, 2016, from

<http://pubpub.media.mit.edu/pub/extended-intelligence>

Cortez.P. *Student Performance Data Set* .Retrieved April 7, 2016, from

<http://archive.ics.uci.edu/ml/datasets/Student+Performance>

ALice.M(September,2015). *Fitting a neural network in R; neuralnet package*..Retrieved April 7, 2016, from

<http://www.r-bloggers.com/fitting-a-neural-network-in-r-neuralnet-package/>

Cortez.P. *Student Performance Data Set* .Retrieved April 7, 2016, from

<http://archive.ics.uci.edu/ml/datasets/Student+Performance>

Ito.J.(2016).*Extended Intelligence*.Retrieved April 7, 2016, from

<http://pubpub.media.mit.edu/pub/extended-intelligence>