Kenta Suzuki
73616154

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

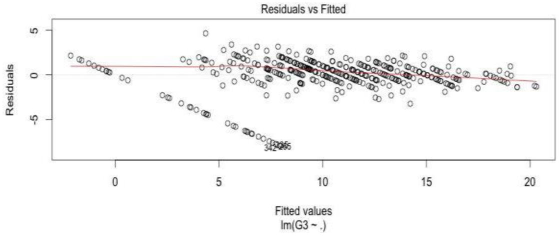
We fit multiple regression to our data set with G3 as a response and other variables as predictors. Below are summary and plots of the fitted model.

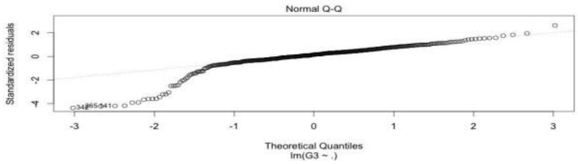
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.900878 on 353 degrees of freedom
Multiple R-squared:  0.8457652, Adjusted R-squared:  0.8278513
F-statistic: 47.21266 on 41 and 353 DF,  p-value: < 0.00000000000000022204
```

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)*

```
      rstudent unadjusted p-value Bonferonni p
342 -4.467451652        0.000010682    0.0042194
265 -4.323537520        0.000020016    0.0079062
141 -4.288958746        0.000023218    0.0091709
335 -4.263100044        0.000025926    0.0102410
317 -3.987575301        0.000081203    0.0320750
297 -3.967412771        0.000088063    0.0347850
```

*** 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

Kenta Suzuki
73616154

[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.

	GVIF	Df	GVIF^(1/(2*Df))
school	1.510967	1	1.229214
sex	1.486968	1	1.219413
age	1.803293	1	1.342867
address	1.387985	1	1.178128
famsize	1.153276	1	1.073907
Pstatus	1.145492	1	1.070277
Medu	2.940226	1	1.714709
Fedu	2.141010	1	1.463219
Mjob	3.430704	4	1.166603
Fjob	2.301555	4	1.109820
reason	1.553683	3	1.076202
guardian	1.736473	2	1.147934
traveltime	1.320973	1	1.149336
studytime	1.395833	1	1.181454
failures	1.563186	1	1.250274
schoolsup	1.255074	1	1.120301
famsup	1.304026	1	1.141940
paid	1.338698	1	1.157021
activities	1.158682	1	1.076421
nursery	1.151348	1	1.073009
higher	1.315791	1	1.147080
internet	1.257752	1	1.121495
romantic	1.174382	1	1.083689
famrel	1.141814	1	1.068557
freetime	1.321398	1	1.149521
goout	1.496482	1	1.223308
Dalc	2.028513	1	1.424259
Walc	2.389544	1	1.545815
health	1.179266	1	1.085940
absences	1.256253	1	1.120827
G1	4.673491	1	2.161826
G2	4.409261	1	2.099824

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

```
(Intercept)  (Intercept)     schoolMS         sexM          age
-1.48232271   0.00000000   0.00000000   0.00000000   0.00000000
 Fjobhealth    Fjobother  Fjobservices  Fjobteacher   reasonhome
 0.00000000   0.00000000   0.00000000   0.00000000   0.00000000
    paidyes activitiesyes  nurseryyes    higheryes  internetyes
 0.00000000   0.00000000   0.00000000   0.00000000   0.00000000

    addressU      famsizeLE3       PstatusT          Medu         Fedu
  0.00000000      0.00000000     0.00000000    0.00000000   0.00000000
reasonother reasonreputation guardianmother guardianother   traveltime
  0.00000000      0.00000000     0.00000000    0.00000000   0.00000000

Mjobhealth    Mjobother  Mjobservices   Mjobteacher
0.00000000   0.00000000    0.00000000    0.00000000
 studytime     failures  schoolsupyes     famsupyes
0.00000000  -0.02616055    0.00000000    0.00000000
```

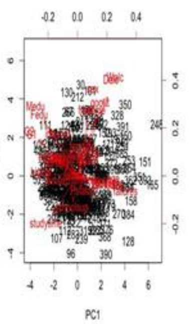
As shown above, lasso with cross-validation uses only one variables, failures. It is clear from the Test MSE of two methods that most of the predictors do not have significant impact on response since the lasso assumes that response is a function of a few variables.

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

```
> coef(regfit.full ,7)##full model
 (Intercept)         age Fjobservices      famrel        Walc    absences          G1
 -0.35159502 -0.20886536  -0.43873469  0.39194586  0.14492849  0.04124892  0.15972322
          G2
  0.98203474
> coef(regfit.fwd ,7)##forward selection
 (Intercept)         age Fjobservices      famrel        Walc    absences          G1
 -0.35159502 -0.20886536  -0.43873469  0.39194586  0.14492849  0.04124892  0.15972322
          G2
  0.98203474
> coef(regfit.bwd ,7)##backward selection
 (Intercept)         age Fjobservices      famrel        Walc    absences          G1
 -0.35159502 -0.20886536  -0.43873469  0.39194586  0.14492849  0.04124892  0.15972322
          G2
  0.98203474
> coef(regfit.ex,7)##exhaustive search
 (Intercept)         age Fjobservices      famrel        Walc    absences          G1
 -0.35159502 -0.20886536  -0.43873469  0.39194586  0.14492849  0.04124892  0.15972322
          G2
  0.98203474
```

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



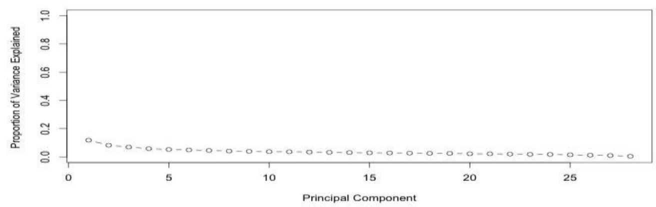
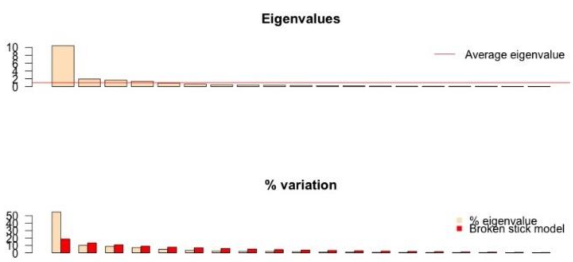


Figure: Proportion of variance explained. A plot is almost flat and each PCAs do not explain this data set so much.



**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.

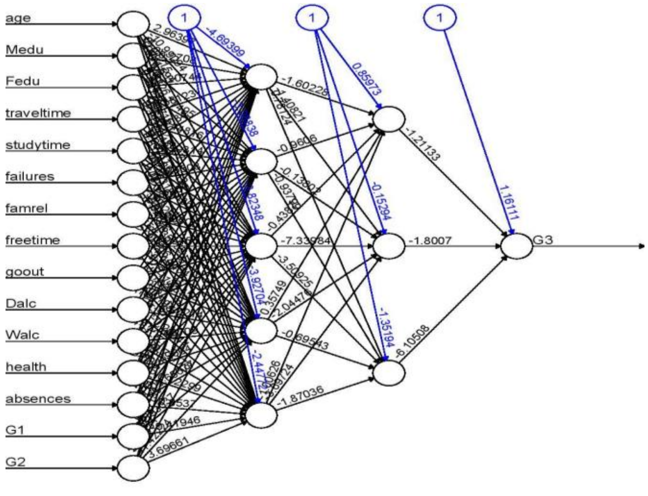


Figure: A plot of neural network with two hidden network, each has five and three neurons.

Test MSE

lm: 4.33279789720453   NN:2.2616425787414

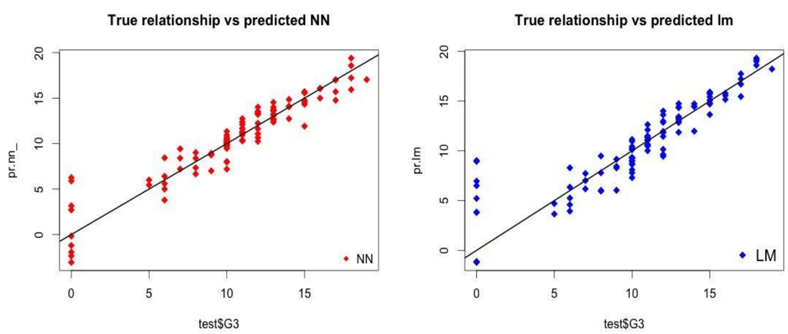


Figure:Plots of predicted neural network and predicted linear regression. Since this is a simulated data, a true line can be shown. (It is hard to measure the goodness of fit of two models since two plots are similar.)

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

```
> mean(cv.error)
[1] 2.136440248
```

```
> cv.error
 [1] 0.7693432519 3.3885045894 2.8441931934 3.0540547837 1.8935709844

 0.9205972652 2.6966543159 0.9679167976 2.3438944832 2.4856728192
```

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.

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

```
      Fail Pass
 Fail  167   19
 Pass   26  183
> 45/nrow(d1)
[1] 0.1139240506
```

```
      Fail Pass
 Fail  131   55
 Pass   38  171
> (55+38)/nrow(d1)
[1] 0.235443038
```

```
        0   1
FALSE 167  20
TRUE   19 189
> 39/nrow(d1)
[1] 0.09873417722
```

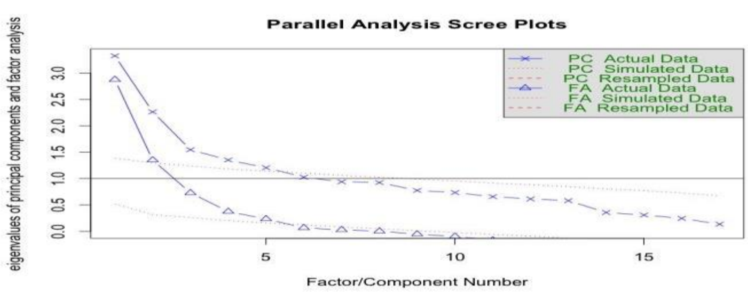
```
knnmod   1   2
     1 164 123
     2  45  63
> 168/nrow(d1)
[1] 0.4253164557
```

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




```
> fa.parallel(firstData)#Determine number of factors
Parallel analysis suggests that the number of factors =  5  and the number of components =  5
```

Kenta Suzuki
73616154

Figure: determinant of the number of factors

```
Call:
factanal(x = firstData, factors = 5, scores = "regression")

Loadings:
           Factor1 Factor2 Factor3 Factor4 Factor5
G1          0.91
G2          0.92
gradesplit -0.80
goout               0.52                    0.37
Dalc                0.57            0.35
Walc                0.91            0.37
Medu                        0.82
Fedu                        0.74
sex_num                             0.63
freetime                            0.25    0.55
age                        -0.22
traveltime                 -0.20
studytime                          -0.47
failures   -0.34           -0.28
famrel                                      0.29
health                              0.24
absences

               Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings       2.52    1.55    1.47    1.05    0.61
```

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.

```
Loadings:
           previousGrades alcoholConsumption Parent's education gender$alcohol freetime with friedns
G1          0.91
G2          0.92
gradesplit -0.80
goout                      0.52                                                0.37
Dalc                       0.57                                 0.35
Walc                       0.91                                 0.37
Medu                                          0.82
Fedu                                          0.74
sex_num                                                         0.63
freetime                                                        0.25           0.55
age                                          -0.22
traveltime                                   -0.20
studytime                                                      -0.47
failures   -0.34                             -0.28
famrel                                                                         0.29
health                                                          0.24
absences

               previousGrades alcoholConsumption Parent's education gender$alcohol freetime with friedns
SS loadings              2.52               1.55               1.47           1.05                 0.61
```

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.

Loadings:

	previousGrades	alcoholConsumption	Parent's education	gender$alcohol	freetime with friedns
G1	0.95				
G2	0.95				
gradesplit	-0.82				
Dalc		0.58		0.22	
Walc		1.01			
Medu			0.85		
Fedu			0.77		
sex_num				0.61	
freetime				0.22	0.62
age					
traveltime					
studytime				-0.44	
failures	-0.29		-0.22		
famrel		-0.23			0.32
goout		0.32			0.43
health				0.24	

	previousGrades	alcoholConsumption	Parent's education	gender$alcohol	freetime with friedns
SS loadings	2.63	1.60	1.45	0.81	0.79
Proportion Var	0.15	0.09	0.09	0.05	0.05
Cumulative Var	0.15	0.25	0.33	0.38	0.43

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.

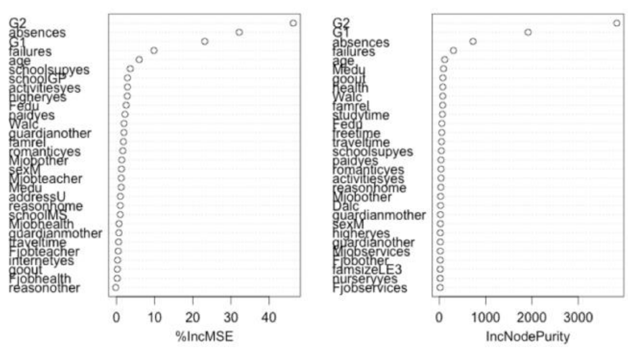


Figure: Plot for importance of variables. G2, G1, absences are quite important variables in both methods.

8. k-means clustering and hierarchical clustering

```
    1   2
1 174  35
2  59 127
301/nrow(d1)
1] 0.7620253165
```

```
cuthc   1   2
    1 207 183
    2   2   3
> 185/nrow(d1)
[1] 0.4683544304
```

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