STAT 448 Final Project Report

Predicting and Analyzing Student Academic Performance

Jian Zhang

Group 9

jzhng148@illinois.edu

University of Illinois at Urbana-Champaign

Introduction

This dataset was obtained from University of California Irvine Machine Learning Repository[1]. The variables include student grades, demographic, social and school-related features like alcohol consumption and number of class fail before. Two datasets are provided regarding the performance in Mathematics and Portuguese.

In this project, my goal is to build models on predicting whether a student will pass the final exam or not. We assume that difficulty of Mathematics and Portuguese is the same, so that we will combine two datasets together to get larger training set that might contribute to better classification results.

Noting that the full score in final exam is 20, and we consider that any student who receives more equal or more than 60% percent of overall score will pass the exam. That is to say, if a student gets a score lower than 12, he or she will be considered as fail the final exam. First, we combine two datasets into one big set and remove variable G1 and G3, since final score "G3" has already contained this information. Then, we add a new column "pass" followed the criterion discussed above.

Logistic regression and discriminant analysis are applied to predict final exam pass rate. The dataset will be split into training and testing set to avoid over-fitting. Classification rates of training and testing set will be provided for comparison and choose the best model based on testing classification rate.

Stepwise selection method is performed to the models in order to identify significantly important variables that have strong influence on student academic performance. We will also check the model performance after feature selection.

Then, we use principal component analysis on the dataset for the dimensionality reduction and refit model on the dimension-reduced dataset to see the impact on prediction power from information loss. All the analysis work will be done in R studio.

Data Exploration

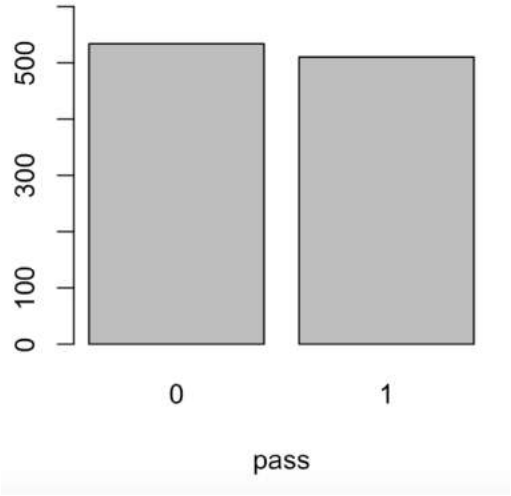
The combined dataset has 1044 observations with 32 variables. 17 of them are categorical variables and the rest 15 variables are numerical. Some interesting variables like alcohol consumption during workday and weekend are also included. Before applying statistical analysis, I personally think that these variables definitely have strong impact on student performance. Generally, if a student has large alcohol consumption, he/she has very low probability to perform well the exam. Other interesting features include in the dataset are that parent's jobs are provided. I am very curious to see whether parent's occupation will be related to passing the exam or not.

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2
2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1
3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1
4	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1
5	GP	F	16	U	GT3	T	3	3	other	other	home	father	1
6	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1

	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime
1	2	0	yes	no	no	no	yes	yes	no	no	4	3
2	2	0	no	yes	no	no	no	yes	yes	no	5	3
3	2	3	yes	no	yes	no	yes	yes	yes	no	4	3
4	3	0	no	yes	yes	yes	yes	yes	yes	yes	3	2
5	2	0	no	yes	yes	no	yes	yes	no	no	4	3
6	2	0	no	yes	yes	yes	yes	yes	yes	no	5	4

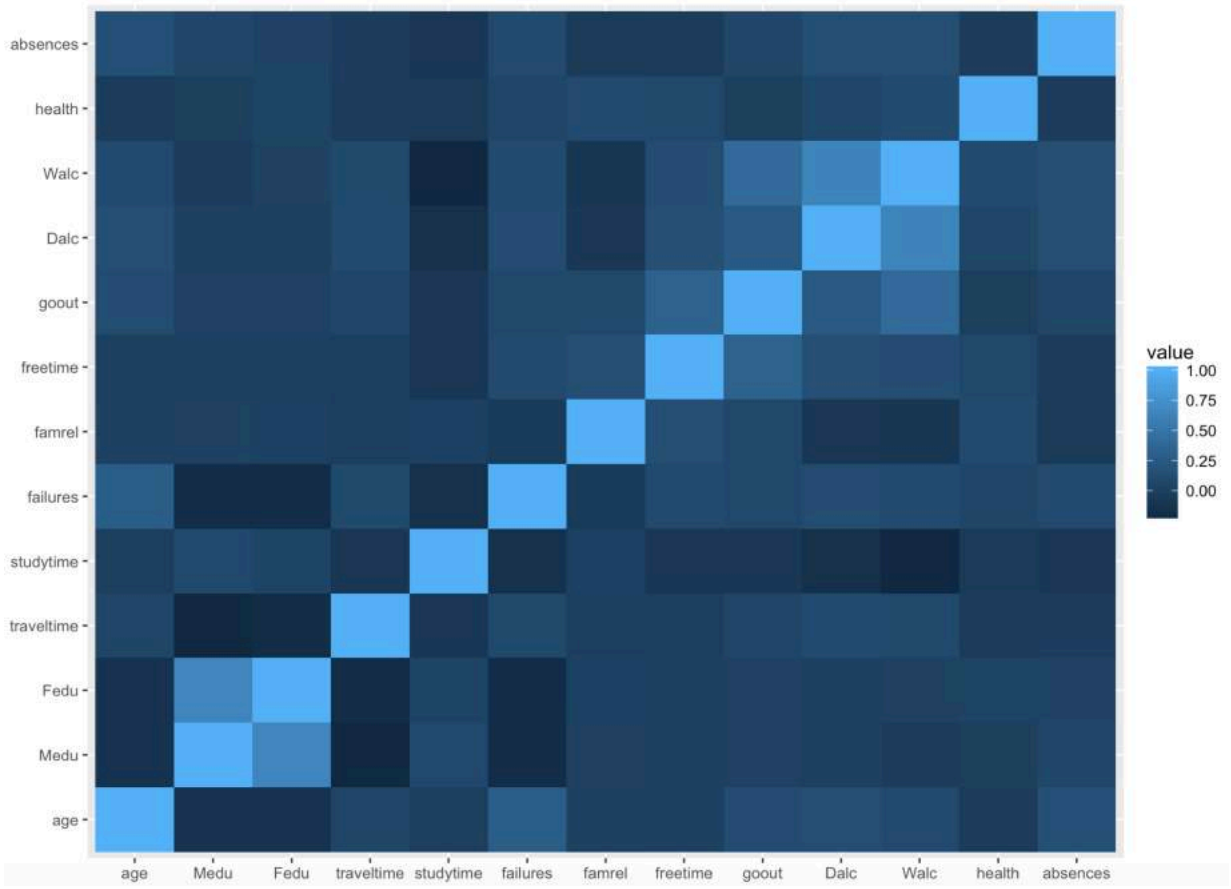
	goout	Dalc	Walc	health	absences	G3	pass
1	4	1	1	3	6	6	0
2	3	1	1	3	4	6	0
3	2	2	3	3	10	10	0
4	2	1	1	5	2	15	1
5	2	1	2	5	4	10	0
6	2	1	2	5	10	15	1

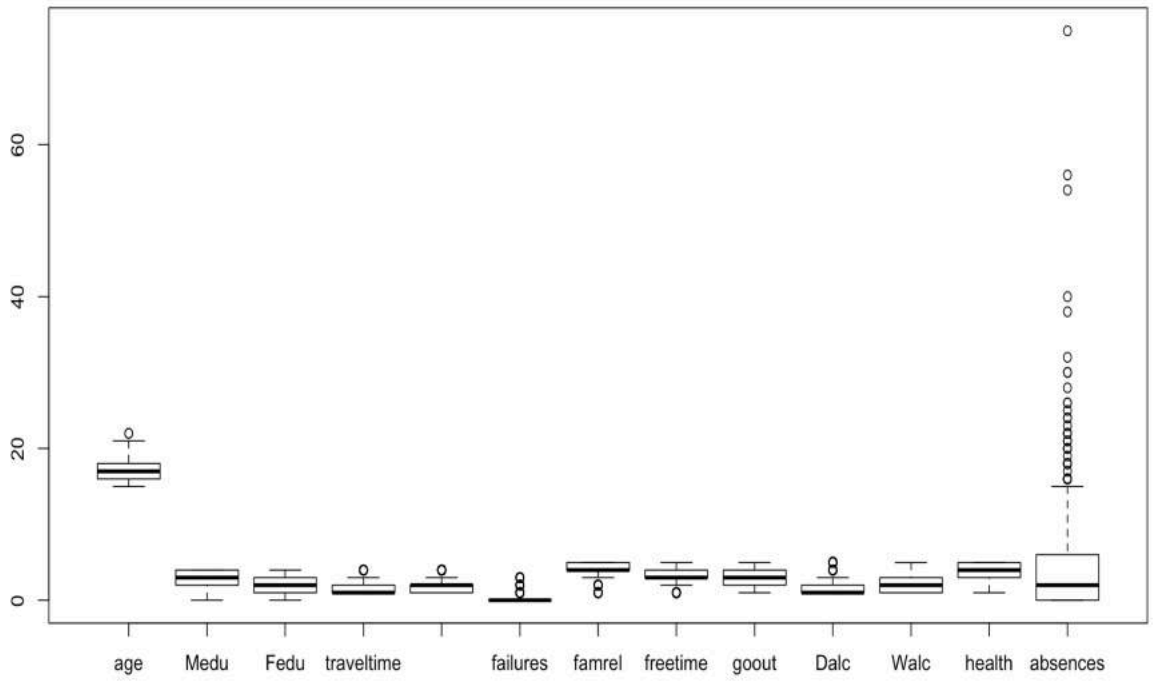
Let's take a close look at response variable. Since we convert final score into a pass or not, variable "pass" only contains 1 and 0, which 1 stands for pass while 0 stands for fail. Number of students who didn't pass the final exam is 534, while 510 students have successfully passed the exam. It is clear that we are dealing with a very balanced data set, so we can use classification rate as performance metric to evaluated model prediction power rather than F1 score that takes into account of both precision and recall.



Pass	0	1
Number	534	510

Then we produce a heatmap based on covariance matrix of 13 numerical variables. It is clear to see that all the variables have slightly different correlation with each other. Only a few pair of variables has very low correlations, while most of them have similar correlation. I am not sure what the result of covariance test will be simply by checking the plot. Further test will be performed since this step is crucial in choosing LDA and QDA.


absences
health
Walc
Dalc
goout
freetime
famrel
failures
studytime
traveltime
Fedu
Medu
age
age
Medu
Fedu
traveltime
studytime
failures
famrel
freetime
goout
Dalc
Walc
health
absences
value
1.00
0.75
0.50
0.25
0.00


0
20
40
60
age
Medu
Fedu
traveltime
failures
famrel
freetime
goout
Dalc
Walc
health
absences

Boxplots of these numerical variables are also provided above. Most of variables are ranging within similar region indicating that data are pretty stable. Age is one of the variables that significantly higher than others, however, it only shows that most students in this dataset are around 19 years old. One thing we should focus on is that there are a lot of outliers exist in absence, and absence definitely have strong impact on student's academic performance. Some students have extremely high absence number. There is very low probability that a student could pass the exam with high absences.

Methods

Before running different algorithms, we randomly split data into training and testing set, 80% of the data are assigned to training set and the rest 20% of data are testing set. This approach is to avoid over-fitting. By comparing training and testing error, we should choose the model with highest classification rate. In the meantime, by checking discrepancy between training and testing rate, we also could find how robust each algorithm is. Further statistical inferences are also performed to check significance of models and features.

Logistic Regression

First of all, we train the logistic regression on the training set with all the variables available to us. The summary of the model has been shown below. As we can see, there are only a few variables significant according to the Z-test. Hosmer and Lemeshow Goodness-of-Fit Test is also performed between fitted value and actual pass, the P-value is very close to zero. In this case, we should reject the null hypothesis and conclude that the logistic regression with full variables has problems in goodness-of-fit.

Our primary goal is to construct model accurately predicting which student is going to pass the final exam. Accuracy rates have been provided for both training and testing set. Training accuracy is 0.7413174 while testing accuracy is 0.6602871, indicating logistic regression is slightly over-fitting the data. We should also compare accuracy rates based on testing data, since it can represent how the model will perform in other datasets that are "unknown" to it.

Model	Logistic Regression
Training Rate	0.7413174
Testing Rate	0.6602871

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.052015   1.554749   0.033 0.973311
schoolMS       -0.639383   0.226748  -2.820 0.004805 **
sexM           -0.340218   0.191133  -1.780 0.075075 .
age            -0.021849   0.079495  -0.275 0.783433
addressU        0.236446   0.204676   1.155 0.248000
famsizeLE3      0.164828   0.187295   0.880 0.378835
PstatusT        0.035529   0.264081   0.135 0.892978
Medu            0.078018   0.118966   0.656 0.511952
Fedu            0.178924   0.105188   1.701 0.088946 .
Mjobhealth      0.690490   0.432435   1.597 0.110322
Mjobother       0.257879   0.249711   1.033 0.301738
Mjobservices    0.569129   0.296544   1.919 0.054958 .
Mjobteacher     0.010744   0.377047   0.028 0.977266
Fjobhealth      0.081480   0.571832   0.142 0.886694
Fjobother       0.079930   0.370285   0.216 0.829096
Fjobservices   -0.232617   0.385374  -0.604 0.546101
Fjobteacher     0.538985   0.515136   1.046 0.295424
reasonhome      0.170589   0.216565   0.788 0.430870
reasonother    -0.074407   0.304724  -0.244 0.807092
reasonreputation  0.183079   0.221474   0.827 0.408441
guardianmother -0.288787   0.207679  -1.391 0.164363
guardianother  -0.109863   0.453679  -0.242 0.808655
traveltime     -0.088066   0.124463  -0.708 0.479213
studytime       0.107330   0.106063   1.012 0.311564
failures       -1.446410   0.256993  -5.628 1.82e-08 ***
schoolsupyes   -1.420158   0.279746  -5.077 3.84e-07 ***
famsupyes      -0.141816   0.178287  -0.795 0.426357
paidyes        -0.754613   0.204324  -3.693 0.000221 ***
activitiesyes   0.321780   0.173411   1.856 0.063512 .
nurseryyes     -0.164825   0.213217  -0.773 0.439500
higheryes       1.173662   0.406320   2.889 0.003871 **
internetyes     0.190126   0.219158   0.868 0.385653
```

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  as.numeric(train$pass), lg_train_pred
X-squared = 3115.9, df = 8, p-value < 2.2e-16
```

Since the logistic regression has problems in goodness-of-fit and a lot of insignificant variables, we apply stepwise selection on features selection according to AIC. The summary of selected model has been shown below. As we can see, almost every variable is significant. Again, we check accuracy from both training and testing set. Training accuracy is 0.7101796 while testing accuracy is 0.6746411. Even though the model performance reduced in the training set, I am surprised to see that testing accuracy is actually increased. The reduced model outperformed the original model with even fewer variables, indicating it efficiently captures important features that explain and predict pass rate. Also, the goodness-of-fit test favors reduced model as well.

Model	Logistic Regression	Selected Logistic
Training Rate	0.7413174	0.7101796
Testing Rate	0.6602871	0.6746411

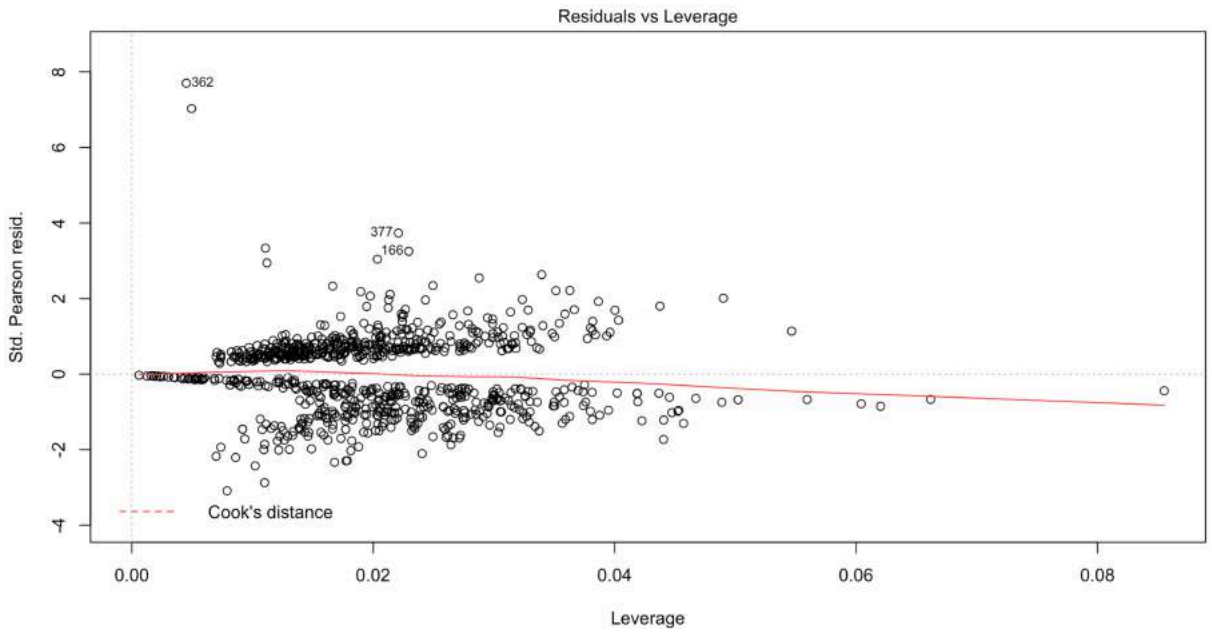
```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.16702  -0.92013  -0.07565   0.88026   2.86189

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.18689    0.61196  -1.939 0.052441 .
schoolMS     -0.70536    0.20251  -3.483 0.000496 ***
sexM         -0.29103    0.17645  -1.649 0.099079 .
Medu          0.19684    0.11364   1.732 0.083241 .
Fedu          0.21560    0.09744   2.213 0.026926 *
Mjobhealth    0.75317    0.40154   1.876 0.060698 .
Mjobother     0.68333    0.25005   2.733 0.006280 **
Mjobservices  0.96306    0.28908   3.331 0.000864 ***
Mjobteacher   0.24147    0.36401   0.663 0.507111
studytime     0.19291    0.10457   1.845 0.065073 .
failures     -1.39089    0.22683  -6.132 8.69e-10 ***
schoolsupyes -1.26477    0.27147  -4.659 3.18e-06 ***
paidyes      -0.71920    0.19888  -3.616 0.000299 ***
higheryes     1.41782    0.43814   3.236 0.001212 **
goout        -0.18969    0.07173  -2.645 0.008180 **
health       -0.12117    0.05920  -2.047 0.040677 *
absences     -0.06380    0.01564  -4.079 4.53e-05 ***


        Hosmer and Lemeshow goodness of fit (GOF) test

data:  train_pred, as.numeric(train$pass)
X-squared = -1368.2, df = 8, p-value = 1
```

The diagnostic plot doesn't show any high influential points, so we don't need to remove any outliers from the dataset.



Discriminant Analysis

In this part of analysis, we have tried linear discriminant analysis as well as quadratic discriminant analysis and compare with accuracy rate from both training and testing set. The table on the left-hand side is confusion matrix of training set, it is clear that QDA successfully predicts more students who eventually pass or fail the final exam compared to LDA. However, it is not the case in the testing set. LDA outperform QDA showing it has better capability of generalization from unknown dataset.

```
> table(lda_train_pred$class, train$pass)

    0   1
0 291  90
1 131 323
> table(qda_train_pred$class, train$pass)

    0   1
0 301  51
1 121 362
```

```
> table(lda_test_pred$class, test$pass)

   0  1
0 68 29
1 44 68
> table(qda_test_pred$class, test$pass)

   0  1
0 63 27
1 49 70
```

Model	LDA	QDA
Training Rate	0.7353293	0.794012
Testing Rate	0.6507177	0.6363636

Noting that QDA provides the highest accuracy rate among other models, however, it has pretty low testing accuracy. Obviously, it is over-fitting the model. According to testing accuracy, we should choose LDA to predict the pass rate.

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  student1[, 1:30]
Chi-Sq (approx.) = 1983.1, df = 465, p-value < 2.2e-16
```

Homogeneity of within covariance matrix test has been performed, and we can see that huge statistics reject null hypothesis and conclude that covariance within dataset is not the same. Theoretically, we should prefer QDA to continue analysis on the dataset. However, our main goal is to predict accuracy rate, in this case, LDA outperformed QDA in terms of accuracy rate. So, we continue developing model based on LDA.

Then we apply stepwise selection on Linear Discriminant Analysis based on accuracy. And we select 10 variables left in the final model. We refit the selected model on both training and testing set to get accuracy shown below. Similar to what we've seen in logistic regression, selected LDA delivers the highest accuracy among all discriminant analysis models. We can conclude that the dataset contains noise features that may decrease classification rate. Simply by remove insignificant features will actually contribute to model performance.

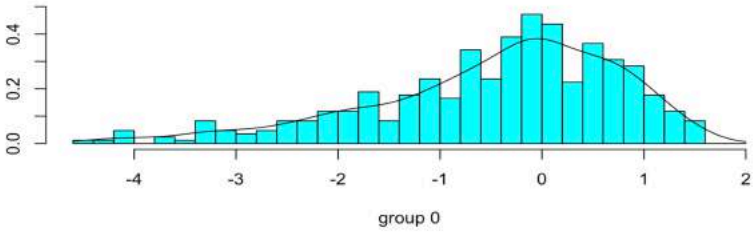
Model	LDA	QDA	Selected LDA
Training Rate	0.7353293	0.794012	0.6838323
Testing Rate	0.6507177	0.6363636	0.6602871

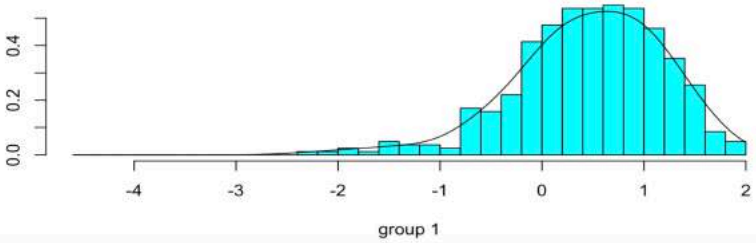
```
accuracy: 0.08687;  in: "failures";  variables (1): failures
accuracy: 0.11401;  in: "Medu";  variables (2): failures, Medu
accuracy: 0.13398;  in: "absences";  variables (3): failures, Medu, absences
accuracy: 0.14324;  in: "Dalc";  variables (4): failures, Medu, absences, Dalc
accuracy: 0.14952;  in: "studytime";  variables (5): failures, Medu, absences, Dalc, studytime
accuracy: 0.15385;  in: "health";  variables (6): failures, Medu, absences, Dalc, studytime, health
accuracy: 0.15727;  in: "traveltime";  variables (7): failures, Medu, absences, Dalc, studytime, health, traveltime
accuracy: 0.15837;  in: "goout";  variables (8): failures, Medu, absences, Dalc, studytime, health, traveltime, goout
accuracy: 0.15917;  in: "Fedu";  variables (9): failures, Medu, absences, Dalc, studytime, health, traveltime, goout, Fedu
accuracy: 0.15939;  in: "age";  variables (10): failures, Medu, absences, Dalc, studytime, health, traveltime, goout, Fedu, age

 hr.elapsed min.elapsed sec.elapsed
      0.000       0.000      17.309

method      : lda
final model : pass ~ age + Medu + Fedu + traveltime + studytime + failures +
    goout + Dalc + health + absences
```

Then, we visualize the result of selected LDA model. The plot displays histograms and density plots of each observation in the space of the first two discriminant functions. We can see two groups are largely overlapped on the right side in the plot, which makes it difficult for model to distinguish data.



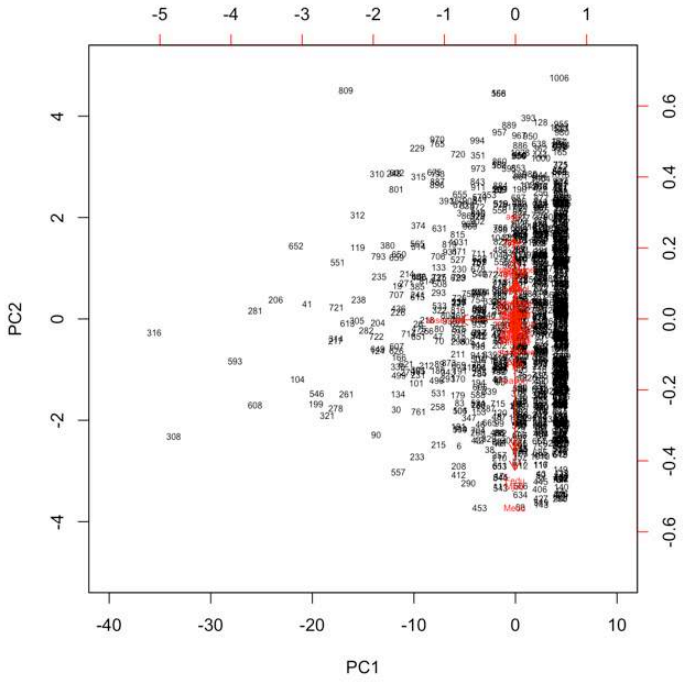


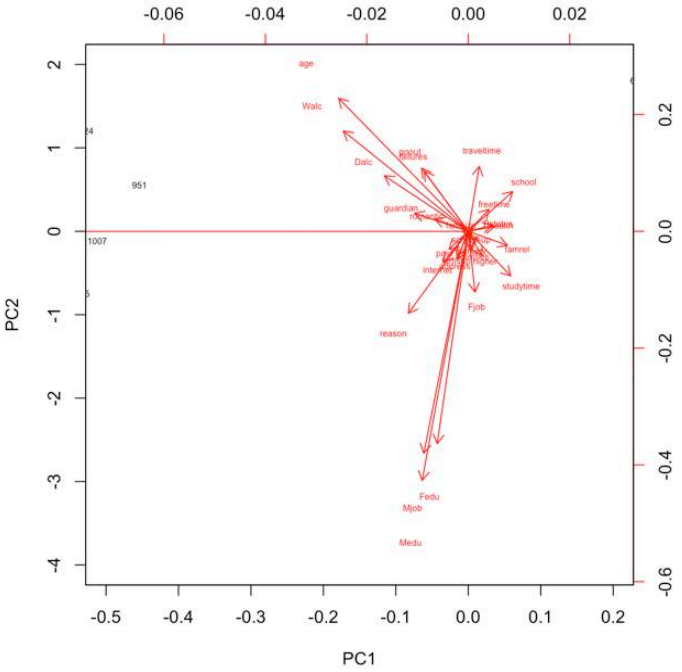
Principal Component Analysis

In this section, we applied Principal Component Analysis on the original dataset, since more than half of the data are categorical variables that cannot be directly computed by PCA, we hence convert those categorical variables into numbers. Then, we reduce dimensionality on the data based on PCA result and check whether there is a change in model performance after information loss.

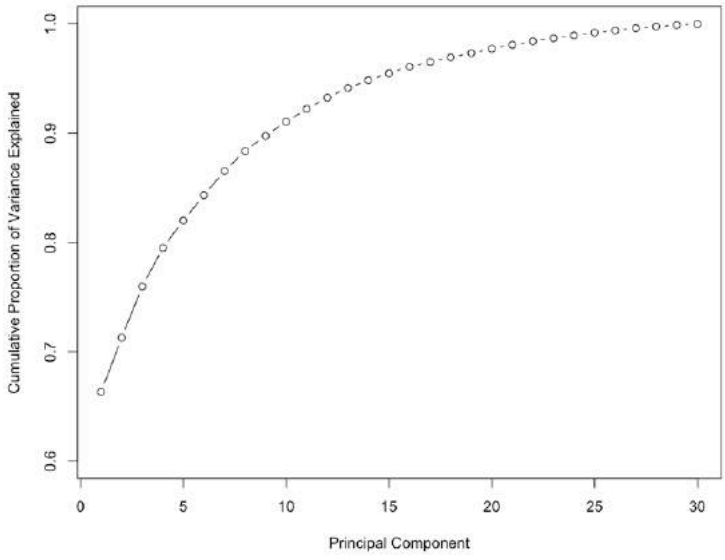
	PC1	PC2
school	0.0109486261	0.084921019
sex	0.0007953793	-0.011457443
age	-0.0319304200	0.284742017
address	-0.0026563402	-0.059283087
famsize	-0.0014099265	0.010315500
Pstatus	0.0058736530	0.012709121
Medu	-0.0113668774	-0.533249050
Fedu	-0.0076478022	-0.454403287
Mjob	-0.0109591615	-0.475218991
Fjob	0.0016516203	-0.130360424
reason	-0.0147728953	-0.175634299
guardian	-0.0132306495	0.038538981
traveltime	0.0027052918	0.138778598
studytime	0.0104518313	-0.095451982
failures	-0.0108513719	0.128584378
schoolsup	0.0005687625	-0.015092647
famsup	-0.0024013558	-0.051725713
paid	-0.0047014768	-0.038616246
activities	0.0007831176	-0.042583011
nursery	0.0007300922	-0.036528502
higher	0.0033373289	-0.052160342
internet	-0.0060454121	-0.065556389
romantic	-0.0082033187	0.026344905
famrel	0.0096379791	-0.030017423
freetime	0.0051101903	0.046960106
goout	-0.0114876378	0.135121911
Dalc	-0.0206665380	0.118735652
Walc	-0.0307963220	0.214623045
health	0.0065200840	0.010091518
absences	-0.9980305370	-0.003720495

The loading vectors of first two principal components are shown above. We can see that the first component picks up variable absences due to large negative value of -0.998, indicating that the students with high number of absences tend to have a smaller value. The first principal component might be presenting some variables that are positively related with final score. The second principal component picks up age and weekend alcohol consumption. In term of second PC, students tend to be grouped together by age and alcohol consumption, older students with large weekend alcohol consumption are expected to have a large value in this dimension.





We plot result from Principal Component Analysis shown above. Since most of variables were converted from categorical data, a lot of observations show no linear relationship in the direction of first principal component. In terms of PC2, all the data are pretty symmetric by zero. Again, absence has been picked up and highly negatively related to the first principal component. Later, we plot cumulative proportion of variance explained by each principal component, and select first 10 PCs that explain total 91.064% variation in the data. In the other words, these PCs contain over 90% of total information covered in the original dataset.



	PC10	PC11
Standard deviation	0.87270	0.81666
Proportion of Variance	0.01305	0.01143
Cumulative Proportion	0.91064	0.92207

Model	Logistic Regression	LDA
Training Rate	0.6467066	0.645509
Testing Rate	0.6794258	0.6889952

After dimension reduction, we refit logistic regression and LDA on the reduced dataset. Noting that we still split data into training and testing set to compare testing error. Surprisingly, both models have better performance in the testing set. Moreover, by comparing classification rate with the models on the full dataset, we can see that LDA on the reduced dataset still has the best performance among all the models. We can conclude that by dimensionality reduction, noisy data has been removed by PCA in this dataset. Generally, we believe that reduce data will result in loss of information, hence smaller dataset will decrease model performance. However, in this case, PCA did a great job in removing irrelevant variables that mislead classification.

Model	Selected Logistic	LDA	Selected LDA	Logistic Regression(PCA)	LDA(PCA)
Training Rate	0.7101796	0.7353293	0.6838323	0.6467066	0.645509
Testing Rate	0.6746411	0.6507177	0.6602871	0.6794258	0.6889952

Conclusion

Since we have applied different models to capture important feature from the data, we will draw conclusions from different models separately.

Logistic Regression

Based on result of stepwise selection, we find out that the most significant features are **failures**, **schoolsupyes** and **paidyes**. (1) Failures stands for number of past class failures for a student. It has negative value of – 1.55861, which has strongly negative relationship with final score. It also makes sense that a student with a large number of failure classes before has low chance of passing the final exam. (2) Schoolsupyes stands for extra educational support, since it is a categorical variable, we find that yes is converted to be 2 while no is assigned to be 1. The estimate of schoolsupyes is -1.55148 with very small P-value. So we can say that the student who need extra education support is the one who actually has difficulty in learning the material and keeps up with the class, hence, this student is very likely to fail the final exam. (3) Paidyes means that extra paid classes within the course subject, and yes is assigned to be 2 and no is converted to be 1. It also has negative relationship with final score estimated to be -0.79773. It is interesting to see that the student

whose tuition was paid by others has lower possibility of passing the exam by comparing to those who actually paid by themselves. There is no wonder that people will value more on thing that they actually paid for them.

Linear Discriminant Analysis

Stepwise selection on LDA chooses following 10 variables to predict pass rate. Surprisingly, Medu(mother's education) has the biggest impact on how students perform in the final exam. The higher education the student's mother has, the greater the positive impact will be on student's final exam. Similar influence has shown on studytime and Fedu(father's education), but the impact is not as big as Medu. To be mentioned that workday alcohol consumption has significantly negative impact on student performance. Travel time is another important feature that indicates student will fail. In all, all the selected features make logical sense in the real life situation.

```
Coefficients of linear discriminants:
                  LD1
age         0.09690145
Medu        0.30752056
Fedu        0.12018210
traveltime -0.22262707
studytime   0.22455026
failures   -1.04266023
goout      -0.12988665
Dalc       -0.20155038
health     -0.12125838
absences   -0.05476657
```

Principal Component Analysis

PCA provide variation contributed by each PC, and we find out that first 4 PCs explain about 80% of total variation. Like we explain before, the first PC picks up absence while the second PC picks up age and weekend alcohol consumption. In terms of third PC, it has large value on goout, Dalc and Walc, which indicating degree of distraction from studying. Data tend to be grouped by these variables and students who spend more time going out with friends and have large consumption on alcohol will have large value in this direction. The third PC might be an indicator capturing factors that have negative impact on student performance. The 4th PC picks up variable health, since it has large negative value of -0.8505. If a student has good health condition, this score is going to be big, so we can infer that this principal component capture factors that have negative impact on student performance just like the 3rd PC.

	PC1	PC2	PC3	PC4
school	0.0109486261	0.084921019	-0.005479969	0.0165673870
sex	0.0007953793	-0.011457443	0.118757091	-0.0067870814
age	-0.0319304200	0.284742017	0.062241192	0.1371522536
address	-0.0026563402	-0.059283087	0.006536767	0.0010801915
famsize	-0.0014099265	0.010315500	0.015575612	0.0105532046
Pstatus	0.0058736530	0.012709121	0.008661401	0.0009924282
Medu	-0.0113668774	-0.533249050	0.161020160	0.1030664367
Fedu	-0.0076478022	-0.454403287	0.168020067	0.0595915908
Mjob	-0.0109591615	-0.475218991	0.236980394	-0.0078393385
Fjob	0.0016516203	-0.130360424	0.072457539	0.0467602695
reason	-0.0147728953	-0.175634299	-0.085961574	0.2894446914
guardian	-0.0132306495	0.038538981	-0.002734951	0.0077468375
traveltime	0.0027052918	0.138778598	0.003192905	0.0149450633
studytime	0.0104518313	-0.095451982	-0.125052791	0.0216376912
failures	-0.0108513719	0.128584378	0.038765255	-0.0096896168
schoolsup	0.0005687625	-0.015092647	-0.017465499	-0.0112046771
famsup	-0.0024013558	-0.051725713	-0.005116041	-0.0087670392
paid	-0.0047014768	-0.038616246	0.010904217	0.0119139908
activities	0.0007831176	-0.042583011	0.027155334	0.0124627513
nursery	0.0007300922	-0.036528502	-0.006772404	0.0017495479
higher	0.0033373289	-0.052160342	-0.007710600	-0.0046108141
internet	-0.0060454121	-0.065556389	0.032190788	0.0300246580
romantic	-0.0082033187	0.026344905	-0.007724977	0.0026997552
famrel	0.0096379791	-0.030017423	0.028729762	-0.0870922208
freetime	0.0051101903	0.046960106	0.249702072	0.0341079851
goout	-0.0114876378	0.135121911	0.404676741	0.2742723837
Dalc	-0.0206665380	0.118735652	0.343854091	0.1367696996
Walc	-0.0307963220	0.214623045	0.599503038	0.2266144359
health	0.0065200840	0.010091518	0.365183293	-0.8505250409
absences	-0.9980305370	-0.003720495	-0.034488136	-0.0291596539

Appendices

All the analytics works are performed by R studio. Several additional packages listed below are required to install before running the codes to replicate our results:

(1) gplots – statistical package for plotting data in R

(2) ResourceSelection – required package to run Hosmer-Lemeshow Goodness-of-Fit Test

(3) biotools – call function boxM for homogeneity of within covariance matrix test

(4) MASS – necessary package required for running LDA and QDA

(5) klaR - call function stepclass for stepwise selection on LDA

All the R codes are contained in the file of Final Prject.R uploaded along with this document. Each step of analysis has been clear labeled, if you are interested in replicating my results, you could follow the comment and run the codes.

This project was planned to perform clustering analysis at the beginning, then we find out that R studio doesn't support CCC and Pseudo T plots very well, which makes it unclear to visualize how we choose the number of clusters. Further work should be modified and done under SAS rather than R Studio. It is better to use R Studio for data exploration and display basic statistics. SAS is better at providing comprehensive statistical test results and more detailed diagnostic plots. In the future, I will replicate our analysis in SAS if time permits.

Reference

[1] Student Performance Data Set, UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Student+Performance