Student Performance Analysis using Logistic Regression and Classification

By

Ashwin Kumar Vajantri

Computer Science , MS ,Clemson University

Introduction about the project

Education is a key variable influencing long haul financial advance. Achievement in the core languages give a semantic and numeric framework for different subjects later in students' scholarly careers. The development in school instructive databases encourages the utilization of Data Mining and Machine Learning practices to enhance results in these subjects by recognizing components that indicate failure. Anticipating results permits teachers to take restorative measures for underperforming students which will in turn mitigate the risk of failure.

This project focuses on the student performance analysis using logistic regression and cross validation. This data consists of student achievement information in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). There is a strong correlation between the attributes G1, G2 and the final grade G3. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful.

The data is downloaded from the University of California Irvine machine learning dataset repository. The file is a semi – colon separated file. I divide the datasets into test and training datasets. And then I introduce a new variable which specifies whether the result is pass or fail based on assumptions. In the earlier stages I perform exploratory data analysis to demonstrate the final grade distribution per the different factors. Following the exploratory data analysis, I use decision tree approach to determine the variables which have strong relationship with the grades. From the summary of the decision tree results I have found the features which are significant in deducing the final grade. Finally, I apply k-fold cross validation to calculate the error estimate.

Aim

To predict the student final grades using logistic regression and k-fold cross validation.

Insights into the data

There are two datasets, each for Portuguese and mathematics. The mathematics dataset has 395 records and 33 columns. The Portuguese dataset has 649 records and 33 columns as shown below –

There are no null values in any of the field as shown below –

```
> sapply(mat_data, function(x)all(is.na(x)))
    school        sex        age    address    famsize    Pstatus       Medu
     FALSE      FALSE      FALSE      FALSE      FALSE      FALSE      FALSE
      Fedu       Mjob       Fjob     reason   guardian traveltime  studytime
     FALSE      FALSE      FALSE      FALSE      FALSE      FALSE      FALSE
  failures  schoolsup     famsup       paid activities    nursery     higher
     FALSE      FALSE      FALSE      FALSE      FALSE      FALSE      FALSE
  internet   romantic     famrel   freetime      goout       Dalc       Walc
     FALSE      FALSE      FALSE      FALSE      FALSE      FALSE      FALSE
    health   absences         G1         G2         G3
     FALSE      FALSE      FALSE      FALSE      FALSE
>
```

As shown below in the screenshot , the dataset 33 fields and 395 rows. Str shows the internal structure of the dataset by indicating the different levels. For eg the field school has 2 levels as there are 2 schools mentioned.

```
> str(mat_data)
'data.frame':   395 obs. of  33 variables:
 $ school    : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
 $ sex       : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
 $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
 $ address   : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
 $ famsize   : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
 $ Pstatus   : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
 $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob      : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
 $ Fjob      : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3 5 3 3 ...
 $ reason    : Factor w/ 4 levels "course","home",..: 1 1 3 2 2 4 2 2 2 2 ...
 $ guardian  : Factor w/ 3 levels "father","mother",..: 2 1 2 2 1 2 2 2 2 2 ...
 $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
 $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
 $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
 $ schoolsup : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
 $ famsup    : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
 $ paid      : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 1 1 2 2 ...
 $ activities: Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
 $ nursery   : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2 2 2 ...
 $ higher    : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ internet  : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
 $ romantic  : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
 $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
 $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
 $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
 $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
 $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
 $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
 $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
 $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
 $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
 $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

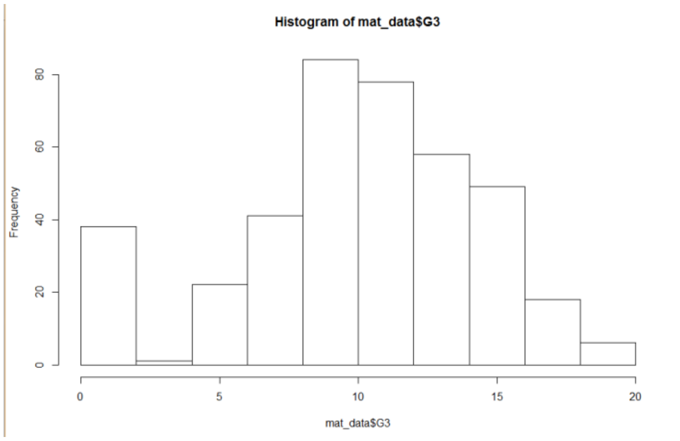
Exploratory Data Analysis

Both the data sets are loaded as shown in the screenshot below -

```
#Loading the Datasets for the performance in two distinct subjects - Maths and Portuguese

mat_data <- read.csv("student-mat.csv",sep=";")
port_data <- read.csv("student-por.csv",sep=";")
```

The below plot shows the distribution of the final grade G3 using a histogram –

hist(mat_data$G3)

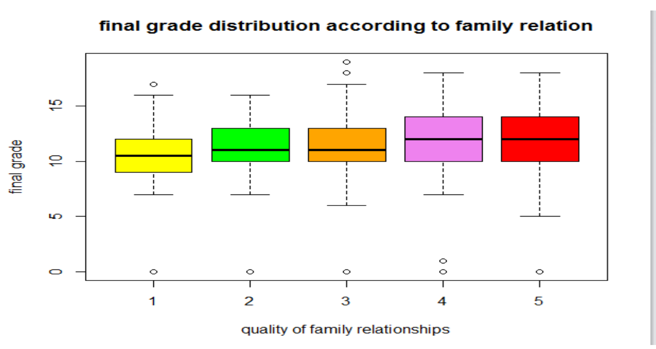


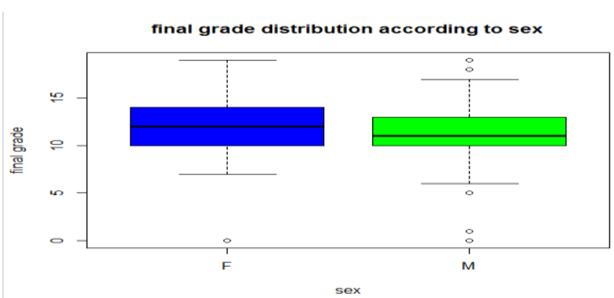
As seen in the above screenshot, the G3 grade is almost normally distributed except one end which is unreliable. I have done the analysis by considering the Portuguese dataset as the training data and the mathematics dataset as the test data. Hence to do prediction I introduce a new field called “res” (for the result) to indicate whether the student is a “pass” or fail”. In this case, instead of predicting the response directly I calculate the probability that the final variable belongs to some category.

```
port_data_train$res <-
  factor(ifelse(port_data_train$G3 >= 8, 1, 0),labels = c("fail", "pass"))
```

As seen “res” is the new variable for which I have set the threshold as 13 i.e above or equal 13 is considered “pass” and below 13 is considered as “fail”. I have done the analysis by considering the Portuguese dataset as the training data and the mathematics dataset as the test data in the further analysis.

Below are the box plots of the G3 against different parameters like G2, famrel and absences .





As shown in the first box plot of G3 vs famrel , students with good family relation (quality of family relationships) tend to score good grades .Also the variance is quite low for the same.

Determination of significant variables

To determine which features should be used for the prediction. Here I am considering the Portuguese dataset as the training dataset. I created a dataset of some important factors as shown below –

```
port_dataset <- select(port_data_train, school ,sex, G1, G2, Mjob, Fjob, goout,
                        absences, reason, Fjob, Mjob, failures, Fedu, Medu, res)
```

I have used the decision tree approach to determine the significant variables which I can apply in my modelling. I use decision tree because there too many variables in the data . To keep all the numeric variables on the same scale and to get accurate results I have normalized all the numeric data. The normalization will help in the situation when the end result is “yes” or “no” (in this case “pass” or “fail”) While applying the decision tree I excluded the “res” variable as this is my class variable.

```
Decision tree:

G2 > 0.4210526: pass (576/2)
G2 <= 0.4210526:
:...G2 <= 0: fail (7)
    G2 > 0:
    :...goout <= 0.75:
        :...reason in {course,home,reputation}: pass (43/6
        :   reason = other: fail (7/2)
        goout > 0.75:
        :...G2 <= 0.368421: fail (5)
            G2 > 0.368421:
            :...Medu <= 0.25: pass (4)
                Medu > 0.25: fail (7/2)


Evaluation on training data (649 cases):

            Decision Tree
          ----------------
          Size      Errors

             7   12( 1.8%)   <<


           (a)   (b)    <-classified as
          ----  ----
            22     8    (a): class fail
             4   615    (b): class pass


        Attribute usage:

        100.00% G2
         10.17% goout
          7.70% reason
          1.69% Medu
```

As shown above on the screenshot, the attribute usage shows that the variables G2 ,goout ,reason and Medu(Mother’s education). Also the error rate is 1.8%. The decision tree shows that there is a strong relationship between the final result and the G2 and also has some relationship between goout,reason and mother’s education as well.

Logistic Regression

I have applied generalized linear model on the variables derived from the decision tree approach . I created a glm keeping G2 ,gout ,reason and Medu as predictors. Below is the result –

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.87528   0.00438   0.02488   0.09700   1.73855

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       -8.1059     1.0966  -7.392 4.52e-13 ***
G2                28.5944     2.6483  10.797  < 2e-16 ***
goout             -2.5616     0.5887  -4.352 1.57e-05 ***
reasonhome        -0.3276     0.4441  -0.738 0.460980
reasonother       -1.3452     0.3730  -3.606 0.000334 ***
reasonreputation   1.5519     0.6840   2.269 0.023611 *
Medu              -0.3530     0.6140  -0.575 0.565550
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 0.2925161)

    Null deviance: 243.05  on 648  degrees of freedom
Residual deviance:  87.59  on 642  degrees of freedom
AIC: NA
```

Since Medu does not have a strong relationship as seen above , I created another model with G2 ,gout and reason as predictors .

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.90594   0.00440   0.02416   0.09017   1.70129

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       -8.2836     1.0640  -7.785 2.79e-14 ***
G2                28.6739     2.6701  10.739  < 2e-16 ***
goout             -2.6004     0.5920  -4.392 1.31e-05 ***
reasonhome        -0.3873     0.4347  -0.891 0.373373
reasonother       -1.3706     0.3741  -3.664 0.000269 ***
reasonreputation   1.5048     0.6867   2.191 0.028777 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 0.2974186)

    Null deviance: 243.045  on 648  degrees of freedom
Residual deviance:  87.686  on 643  degrees of freedom
AIC: NA
```

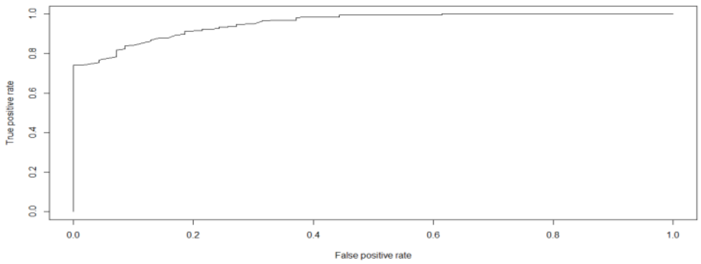
As seen above the 3 predictors have a strong relationship although the “reason” variable is only strong in one of the categories. I will be considering the above fit in my next stages of analysis. I selected this model based on the analysis of deviance table and on Pr value as shown below .

```
Analysis of Deviance Table

Model 1: res ~ G2 + goout + reason + Medu
Model 2: res ~ G2 + goout + reason
  Resid. Df Resid. Dev Df  Deviance     F Pr(>F)
1       642     87.590
2       643     87.686 -1 -0.096234 0.329 0.5665
>
```

To check the accuracy of the training fit on the test data (mathematics dataset), I have created a confusion matrix. I have followed the same approach where I have normalized the numeric data in mathematics dataset as well to keep all the numeric values on the same scale . The confusion matrix gives me 90.89% accuracy on the test data . Before building the confusion matrix I predicted the probability of the test data students using the training fit. The test error rate almost 10%.

```
Confusion Matrix and Statistics


pred_test fail pass
     fail   39    5
     pass   31  320

               Accuracy : 0.9089
                 95% CI : (0.8761, 0.9353)
    No Information Rate : 0.8228
    P-Value [Acc > NIR] : 9.737e-07

                  Kappa : 0.6342
 Mcnemar's Test P-Value : 3.091e-05

            Sensitivity : 0.9846
            Specificity : 0.5571
         Pos Pred Value : 0.9117
         Neg Pred Value : 0.8864
             Prevalence : 0.8228
         Detection Rate : 0.8101
   Detection Prevalence : 0.8886
      Balanced Accuracy : 0.7709

       'Positive' Class : pass
```

To check the model performance, I have plotted the graph of the true positive rate vs the false positive rate. This plot will demonstrate the area under the curve which is called AUC. Below is the curve which I got when I plotted the performance of the prediction. As shown below the area under curve is closer to 1 which is good.



I applied k-fold cross validation to calculate the error estimates. Since I have used glm , to predict the error estimate I have conducted the k-fold cross validation. The error estimate on test data was as shown below.

```
> cvfit  <- glm(res ~ G2 + goout + reason , family = quasibinomial, data = mat_dataset)
> cv.err.10 <- cv.glm(data = mat_dataset, glmfit = cvfit, K = 10)
> cv.err.10$delta
[1] 0.05308698 0.05281271
```

As shown above the two values for K=1 and K=10 are almost same and the error estimates are significantly small thereby suggesting that the model can perform in deducing the final grade.

References

1. https://archive.ics.uci.edu/ml/datasets/Student+Performance - University of California Irvine machine learning dataset repository – Student performance
2. https://www.rulequest.com/see5-unix.html - C5.0: An Informal Tutorial , March 2013, rulequest research.
3. http://sebastianraschka.com/Articles/2014_about_feature_scaling.html - About Feature Scaling and Normalization , Jul 11, 2014 by Sebastian Raschka.
4. https://qizeresearch.wordpress.com/2014/05/25/decision-tree-c5-0-example/ - Decision Tree C5.0 Example , May 25, 2014, Sailfish – Big Data Tech Blog.
5. http://gim.unmc.edu/dxtests/roc3.htm - The Area Under an ROC Curve
6. http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf - An Introduction to Statistical Learning -Gareth James , Daniela Witten ,Trevor Hastie , Robert Tibshirani