PREDICTING STUDENT OUTCOMES

Gwen Rino

October 4, 2017

INTRODUCTION

Education is a staggeringly large project. In the United States alone, over 55 million children are enrolled in grades preK-12, and over 20 million adults attend American colleges and universities (source: National Center for Educational Statistics, https://nces.ed.gov/). Millions more nontraditional students are enrolled in university-sponsored MOOCs and other online programs, such as those offered by Coursera, edX, Khan Academy, and Springboard. The sheer volume of learning going on every day is heartening to all who recognize education as the single best pathway to the full realization of a person's potential.

However, the project of education fails for too many students. Numerous reports show that over a million students drop out of American high schools each year. The cost to these individuals in lost opportunity, and the societal cost of increased poverty and alienation, are profound concerns. This is not just an American problem: Appendix A shows high school completion rates for all OECD countries.

Educational institutions and programs need to be able to identify those students most at risk of failure in order to intervene with appropriate support and remediation. Successfully predicting which students are at risk of failing in time to intervene on their behalf would greatly increase learning, graduation rates, and, ultimately, human well-being.

Experienced educators can attest to the fact that most students' achievement level is consistent throughout the term of a course. Students at or near the top of the class at midterm are very likely to receive top final grades; students with failing marks at midterm are at very high risk of failing the course. While this is the most common pattern, some students do not fit in. These are students who have low-fair marks at midterm but, for whatever reason, fail or drop out by the end of the term. These students are difficult for educators to identify in time to intervene because at midterm they appear to be on track to pass. Another challenge for educators is that most students who will ultimately fail the course are already failing by midterm. It would be a great help if these students could be identified for academic support earlier in order to afford them the best chance of passing.

This project explores these two practical questions:

1. *Can the methods of statistical learning predict which students are likely to fail by taking interim grades (among other variables) as predictors?*

2. *Can the methods of statistical learning predict which students are likely to fail before the term begins (i.e. without taking interim grades as predictors), or early in the term (i.e. without taking second period grades as predictors)?*

THE DATA

The University of California, Irvine, Machine Learning Repository includes student performance datasets collected at two secondary schools in Portugal in 2008. They record students' performance in math or Portuguese, and also thirty other attributes that vary from mother's level of education to weekend alcohol consumption. The datasets can be found at http://archive.ics.uci.edu/ml/datasets/Student+Performance.

The datasets include 33 variables. Five of these—G1 (first period grade), G2 (second period grade), G3 (final grade), school (Gabriel Pereira or Mousinho da Silveira), and absences (number of absences)—were gleaned from students' academic records, and the rest were collected through a questionnaire. A description/explanation of each variable can be found in Appendix B.

Citation: P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

DATA WRANGLING

The datasets were very clean to start with, in tidy format with no missing values. The main wrangling consideration was whether to combine the math and Portuguese datasets, and if so, how to do that (given that some students appeared in both datasets). Considerations included:

- Is a particular student's achievement in math a distinct piece of information from her achievement in Portuguese?
- Is the goal to predict failing students in general or failing students in particular subject areas?
- What would be gained and what would be lost by various approaches to merging these datasets?

Ultimately, I decided to add the two datasets together while preserving the course distinction (math or Portuguese) as a new attribute "course". I also reclassified many of the variables as factors and relabeled their levels for clarity.

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Conflicts with tidy packages ---------------------------------------------
-

## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
# Read in data
d_math <-
read.csv2("~/Desktop/Springboard_Data_Science_Intro/Capstone/Data/student-
mat.csv")
d_port <-
read.csv2("~/Desktop/Springboard_Data_Science_Intro/Capstone/Data/student-
por.csv")

# Tidy up variable types, labels of variables
d_math$studytime <- factor(d_math$studytime, labels = c("<2 hrs", "2-5 hrs",
"5-10 hrs", ">10 hrs"))
d_port$studytime <- factor(d_port$studytime, labels = c("<2 hrs", "2-5 hrs",
"5-10 hrs", ">10 hrs"))
d_math$failures <- factor(d_math$failures, labels = c("0", "1", "2", "3"))
d_port$failures <- factor(d_port$failures, labels = c("0", "1", "2", "3"))
d_math$famsize <- factor(d_math$famsize, labels = c(">3", "<=3"))
d_port$famsize <- factor(d_port$famsize, labels = c(">3", "<=3"))
d_math$Medu <- factor(d_math$Medu, labels = c("None", "Primary", "Middle",
"Secondary", "Higher"))
d_port$Medu <- factor(d_port$Medu, labels = c("None", "Primary", "Middle",
"Secondary", "Higher"))
d_math$Fedu <- factor(d_math$Fedu, labels = c("None", "Primary", "Middle",
"Secondary", "Higher"))
d_port$Fedu <- factor(d_port$Fedu, labels = c("None", "Primary", "Middle",
"Secondary", "Higher"))
d_math$traveltime <- factor(d_math$traveltime, labels = c("<15 min", "15-30
min", "30-60 min", ">60 min"))
d_port$traveltime <- factor(d_port$traveltime, labels = c("<15 min", "15-30
min", "30-60 min", ">60 min"))
d_math$famrel <- factor(d_math$famrel, labels = c("Very Bad", "Poor", "Fair",
"Good", "Excellent"))
d_port$famrel <- factor(d_port$famrel, labels = c("Very Bad", "Poor", "Fair",
"Good", "Excellent"))
d_math$freetime <- factor(d_math$freetime, labels = c("Very Low", "Low",
"Medium", "High", "Very High"))
d_port$freetime <- factor(d_port$freetime, labels = c("Very Low", "Low",
"Medium", "High", "Very High"))
```

```
d_math$goout <- factor(d_math$goout, labels = c("Very Low", "Low", "Medium",
"High", "Very High"))
d_port$goout <- factor(d_port$goout, labels = c("Very Low", "Low", "Medium",
"High", "Very High"))
d_math$Dalc <- factor(d_math$Dalc, labels = c("Very Low", "Low", "Medium",
"High", "Very High"))
d_port$Dalc <- factor(d_port$Dalc, labels = c("Very Low", "Low", "Medium",
"High", "Very High"))
d_math$Walc <- factor(d_math$Walc, labels = c("Very Low", "Low", "Medium",
"High", "Very High"))
d_port$Walc <- factor(d_port$Walc, labels = c("Very Low", "Low", "Medium",
"High", "Very High"))
d_math$health <- factor(d_math$health, labels = c("Very Bad", "Poor", "Fair",
"Good", "Excellent"))
d_port$health <- factor(d_port$health, labels = c("Very Bad", "Poor", "Fair",
"Good", "Excellent"))

# Create new variable "course" and fill in with "math" or "port"
course <- rep("math", times = length(d_math$school))
d_math <- cbind(d_math, course)
course <- rep("port", times = length(d_port$school))
d_port <- cbind(d_port, course)

# Combine datasets
d_total <- rbind(d_math, d_port)
# Convert variable "course" to factor
d_total$course <- factor(d_total$course, labels = c("Math", "Portuguese"))
# Tidy up
rm(course)
```

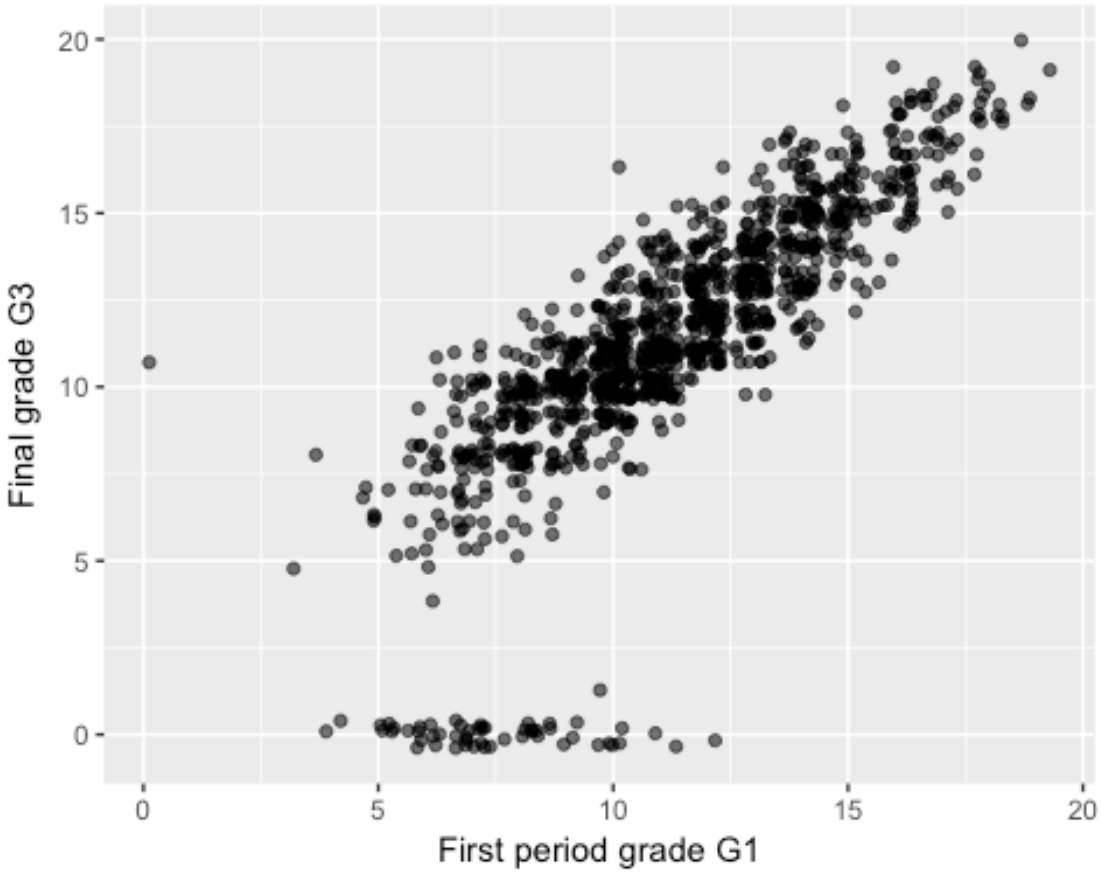
I also decided that, since my fundamental goal was to predict which students will fail, it would be helpful to recast the target G3 from a numeric to a factor variable "outcome" with two levels, "pass" and "fail".

```
# Create new variable "outcome" and define levels in d_math and d_port
d_math_cat <- d_math %>% mutate(outcome=ifelse(G3<8,"fail","pass")) %>%
  select(-G3)
d_port_cat <- d_port %>% mutate(outcome=ifelse(G3<8,"fail","pass")) %>%
  select(-G3)

# Make "outcome" a factor variable
d_math_cat$outcome <- as.factor(d_math_cat$outcome)
d_port_cat$outcome <- as.factor(d_port_cat$outcome)

# Combine d_math_cat and d_port_cat
d_total_cat <- rbind(d_math_cat, d_port_cat)
```

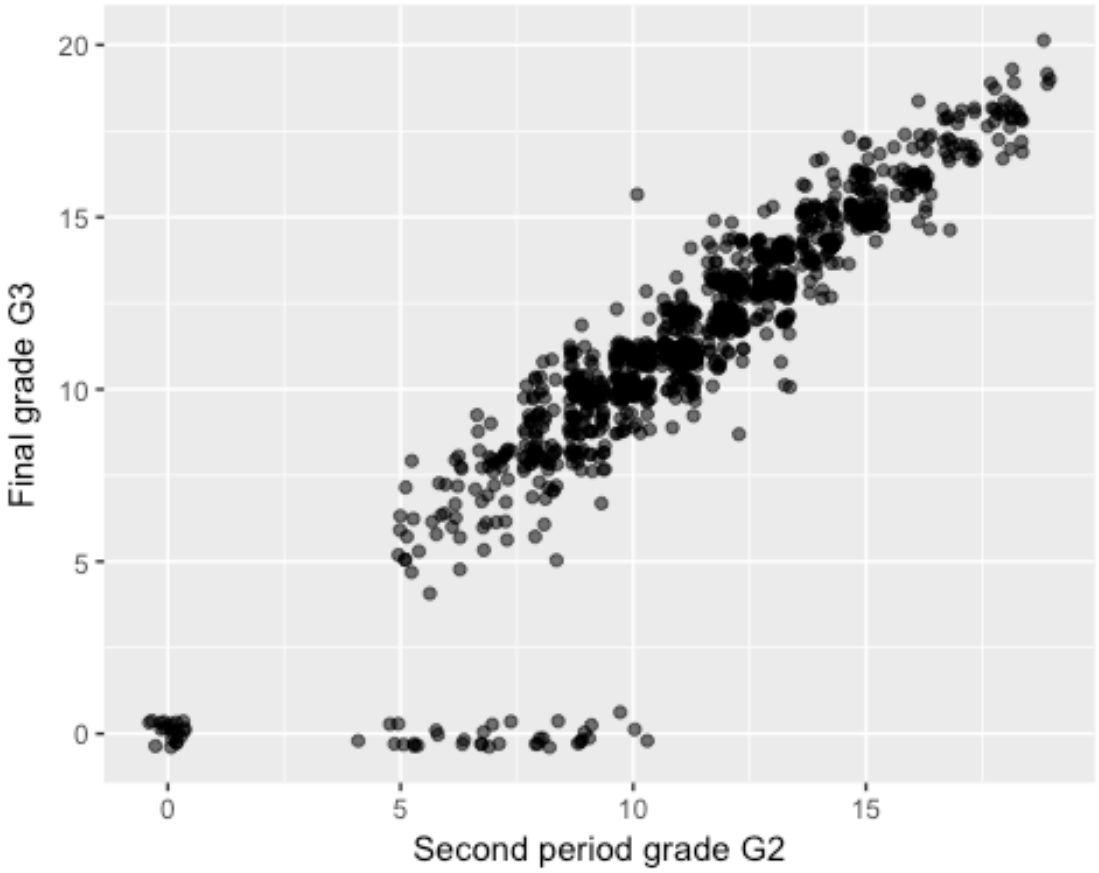
QUESTION 1: PREDICTING FAILURE USING INTERIM GRADES AS PREDICTORS

LINEAR REGRESSION

The first step is to explore the interim grades (G1 and G2) and final grade (G3) to confirm that, as expected, for most most students interim grades have a strong linear relationship to G3. If so, then we will explore whether this relationship is strong enough to build a good predictive model.



As expected, G1 appears to have a strong linear relationship with G3 for most students. There is one outlier who was failing at G1 but received a final G3 grade of 11, and there is a small group of outliers who had grades of 4-12 at G1 but had a final G3 grade of 0.



Like G1, G2 has a strong linear relationship with G3 for most students. There is a similar small group of outliers who had grades of 4-11 at G2 but had a final G3 grade of 0.

Given the outliers, is the linear relationship strong enough that a linear regression will be an effective predictive model?

```
# Split the dataset into a training set and a test set.
set.seed(123)
dt = sort(sample(nrow(d_total), nrow(d_total)*.7))
Train.reg <- d_total[dt,]
Test.reg <- d_total[-dt,]

# Build the model
lin.model <- lm(G3 ~ G1 + G2, data = Train.reg)
summary(lin.model)

##
## Call:
## lm(formula = G3 ~ G1 + G2, data = Train.reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9540 -0.4039  0.0648  0.8305  6.0460
```

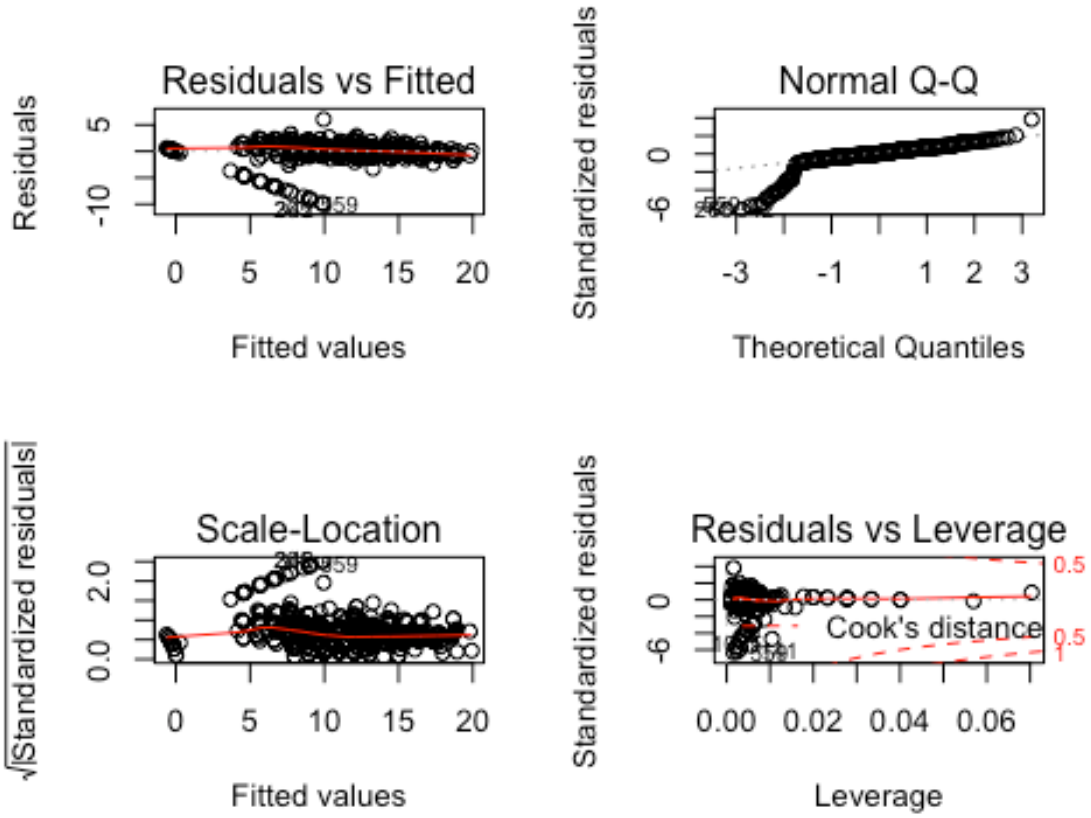
```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.12379    0.22368  -5.024 6.37e-07 ***
## G1           0.12656    0.03746   3.379 0.000766 ***
## G2           0.98121    0.03381  29.022  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.578 on 727 degrees of freedom
## Multiple R-squared:  0.8433, Adjusted R-squared:  0.8428
## F-statistic:  1956 on 2 and 727 DF,  p-value: < 2.2e-16
```

```
# Apply the model to test data
lm.pred <- predict(lin.model, newdata = Test.reg)

# Calculate R-squared
SSE = sum((Test.reg$G3 - lm.pred)^2)
SST = sum((Test.reg$G3 - mean(Train.reg$G3))^2)
1 - SSE/SST
```

```
## [1] 0.8004423
```

With adjusted R-squared values greater than 0.8, this model seems to be worth exploring further with diagnostic plots.

```
par(mfrow=c(2,2))
plot(lin.model)
```



The Residuals vs. Fitted plot shows the strong linear relationship for most students as the points gathered along the horizontal red line, and also the small group of outlier students as the group of points between the fitted values of 3-11 with negative residuals.

The Normal Q-Q plot shows that the residuals in the lowest quantile are not normally distributed. This also makes sense given the outliers who were passing at midterm but failed at G3.

The Scale Location plot also reveals the outlier group between the fitted values of 3-11 with residuals that are spread wide of the horizontal line.

The Residuals vs. Leverage plot shows no points outside Cook's distance, so there aren't any individual data points that are unduly influencing the model.

It makes sense that a linear regression model based on G1 and G2 would not be effective at predicting students whose final grades are very different from their interim grades. What percentage of the total fall into this category?

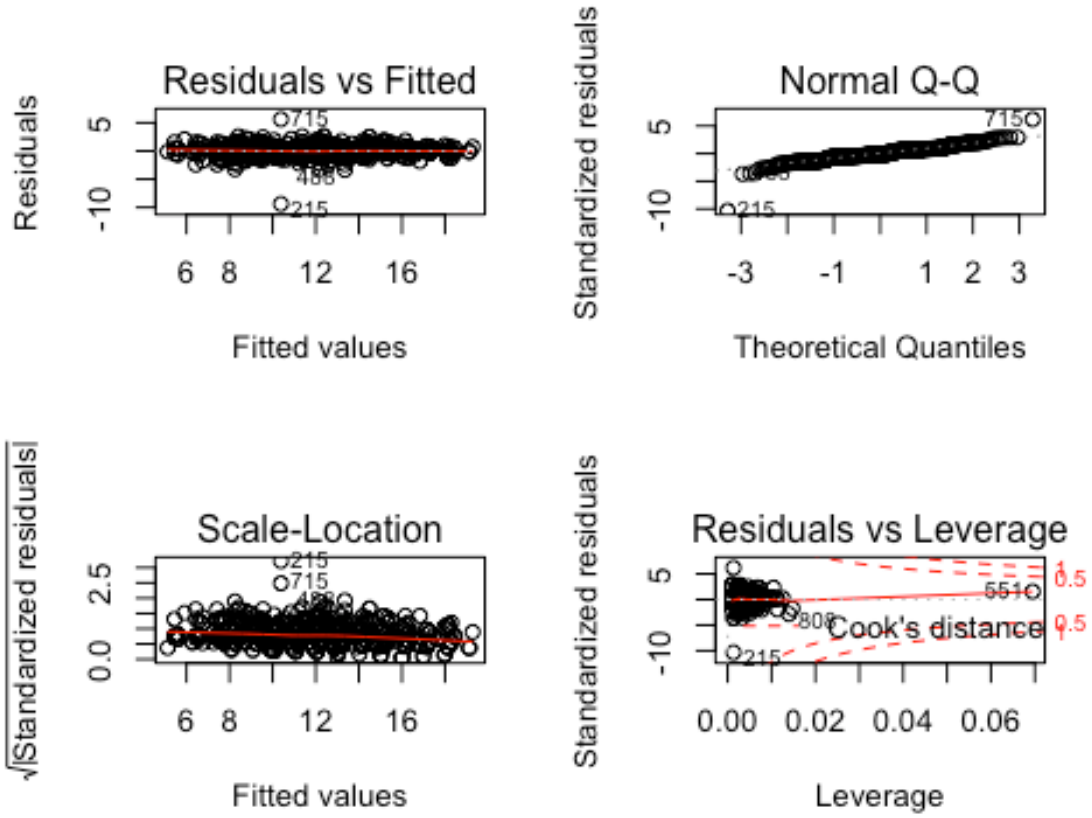
```
# Subset students for whom G3=0
zeroes <- subset(d_total, d_total$G3 == 0)
# Subset students among these for whom G1 and/or G2 > 0
dropouts <- subset(zeroes, zeroes$G1 != 0 | zeroes$G2 != 0)
nrow(dropouts)/nrow(d_total)

## [1] 0.05076628
```

About 5% of the total students had non-zero grades at G1 and/or G2 (typically in the 5-10 range), but received a final grade of 0. Presumably these students dropped out.

It seems, then, that a linear regression model should be highly effective at predicting the final grades of the 95% of students who do not drop out. To confirm this intuition, consider a linear regression model that eliminates the dropouts.

```
# Create dataset of all students except the dropouts.
d_adj <- anti_join(d_total, dropouts)

## Joining, by = c("school", "sex", "age", "address", "famsize", "Pstatus",
"Medu", "Fedu", "Mjob", "Fjob", "reason", "guardian", "traveltime",
"studytime", "failures", "schoolsup", "famsup", "paid", "activities",
"nursery", "higher", "internet", "romantic", "famrel", "freetime", "goout",
"Dalc", "Walc", "health", "absences", "G1", "G2", "G3", "course")

# Create linear model on this dataset.
lin.model.adj <- lm(G3 ~ G1 + G2, data = d_adj)
summary(lin.model.adj) # Adjusted R-squared:  0.9027

##
## Call:
## lm(formula = G3 ~ G1 + G2, data = d_adj)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3857 -0.4872 -0.1718  0.6143  5.6143
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.49811    0.12329   4.040 5.75e-05 ***
## G1           0.15767    0.02189   7.204 1.16e-12 ***
## G2           0.83109    0.02238  37.128  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9085 on 988 degrees of freedom
## Multiple R-squared:  0.9029, Adjusted R-squared:  0.9027
## F-statistic:  4592 on 2 and 988 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lin.model.adj)
```



As expected, all the anomalies in the original model exposed by the diagnostic plots are eliminated. A linear model is an excellent predictor of final grades for 95% of students.

Clearly, no linear model can catch the dropouts: by definition, these students' interim grades are not in a linear relationship with their final grades. However, the dropouts are among the students that educators most wish to predict, so we'll leave linear modeling at this point to seek other models that may be more effective in identifying the dropouts.

LOGISTIC REGRESSION

It is possible to build a good predictive model for the binary outcome "pass" or "fail" using logistic regression.

```
# Split training and test set
library(caTools)
set.seed(77)
split <- sample.split(d_total_cat$outcome, SplitRatio = 0.75)
Train.cat <- subset(d_total_cat, split == TRUE)
Test.cat <- subset(d_total_cat, split == FALSE)
```

```r
# Build model
log.model.1 <- glm(outcome ~ G1 + G2, data = Train.cat, family = binomial)
summary(log.model.1) # AIC: 190.57
```

```
##
## Call:
## glm(formula = outcome ~ G1 + G2, family = binomial, data = Train.cat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.89871   0.00740   0.04215   0.18568   2.11885
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.0121     1.3165  -7.605 2.85e-14 ***
## G1            0.2601     0.1439   1.807   0.0708 .
## G2            1.1598     0.1733   6.694 2.18e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 494.42  on 782  degrees of freedom
## Residual deviance: 184.57  on 780  degrees of freedom
## AIC: 190.57
##
## Number of Fisher Scoring iterations: 8
```

```r
# Apply to test data
log.pred.1 <- predict(log.model.1, newdata = Test.cat, type = "response")

# Examine confusion matrix
table(Test.cat$outcome, log.pred.1 > 0.5)
```

```
##
##        FALSE TRUE
##   fail    17    8
##   pass     2  234
```

At first glance, this model seems good. The predictive accuracy rate, calculated from the confusion matrix, is 96%. However, the specificity (the accuracy rate for the minority outcome, "fail") is only 68%, which drags down the balanced accuracy rate (the mean of the predictive accuracy for each class) to 84%. So this model is good at predicting who will pass, but not so good at predicting who will fail. And we are particularly interested in predicting who will fail!

The problem, in part, is that the dataset is quite imbalanced, with just over 90% of the outcomes "pass" and not quite 10% of the outcomes "fail". Balancing the dataset should improve the specificity, and also the balanced accuracy, of the model.

The package “ROSE” offers four different methods for balancing a dataset:

- oversampling the minority class
- undersampling the majority class
- both over- and undersampling
- creating synthetic data in the minority class

Experimentation revealed that the best balancing method for a logistic regression with this dataset was undersampling. (See Code lines 759-821 for details of this exploration.)

```
library(ROSE)

## Loaded ROSE 0.0-3

# Balance training set by undersampling "pass"
Train.cat.under <- ovun.sample(outcome ~., data = Train.cat, method =
"under", N=150)$data

# Build model with balanced data
log.model.1.under <- glm(outcome ~ G1 + G2, data = Train.cat.under, family =
binomial)

# Apply to test data
log.pred.1.under <- predict(log.model.1.under, newdata = Test.cat, type =
"response")

# Examine AIC and confusion matrix
summary(log.model.1.under) # AIC: 76.832

##
## Call:
## glm(formula = outcome ~ G1 + G2, family = binomial, data =
Train.cat.under)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.11529  -0.27476   0.00069   0.25720   1.94266
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 10.97903    2.04700   5.363 8.16e-08 ***
## G1           0.02146    0.22440   0.096    0.924
## G2          -1.28948    0.29495  -4.372 1.23e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 207.944  on 149  degrees of freedom
## Residual deviance:  70.832  on 147  degrees of freedom
```

```
## AIC: 76.832
##
## Number of Fisher Scoring iterations: 7
```

```
table(Test.cat$outcome, log.pred.1.under > 0.5)
```

```
##
##        FALSE TRUE
##   fail     1   24
##   pass   220   16
```

Balancing the dataset greatly improved this model's predictive accuracy. It lowered the AIC from 190.57 to 76.832. Calculating from the confusion matrix, the accuracy rate is now 93.5% (the balanced accuracy rate is even higher), and the specificity is 96%. This model only miscategorized one failing student and 16 passing students. This is an excellent and useful predictive model.

DECISION TREES

Decision trees offer another effective method for predicting the binary outcome "pass" or "fail"."

```
library(rpart)

# Build model
tree.mod.1 <- rpart(outcome ~ G1 + G2, data = Train.cat, method = "class")

# Apply to test data
tree.pred.1 <- predict(tree.mod.1, newdata = Test.cat, type = "class")

# Examine AUC of ROC curve and confusion matrix
roc.curve(Test.cat$outcome, tree.pred.1, plotit = FALSE)

## Area under the curve (AUC): 0.836

table(Test.cat$outcome, tree.pred.1)

##       tree.pred.1
##        fail pass
##   fail   17    8
##   pass    2  234
```

As with the logistic regression model on unbalanced data, this model predicts with a high accuracy rate of 96% but a modest specificity of 68%. The balanced accuracy rate is 84%. The model is good at predicting who will pass, but not so good at predicting who will fail.

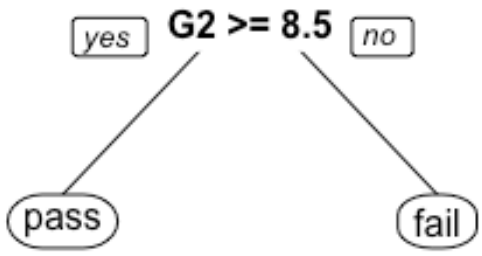
Also like the logistic regression model, the decision tree's predictive accuracy is greatly improved by balancing the dataset. This time, experimentation revealed that both over- and undersampling was the best balancing method. (See Code lines 920-968 for details.)

```r
# Balance the training set by over- and undersampling
Train.cat.both <- ovun.sample(outcome ~., data = Train.cat, method = "both",
p=0.5, N=783, seed = 1)$data

# Build model with balanced data
tree.mod.1.both <- rpart(outcome ~ G1 + G2, data = Train.cat.both, method =
"class")

# Apply to test data
tree.pred.1.both <- predict(tree.mod.1.both, newdata = Test.cat, type =
"class")

# Examine AUC of ROC curve and confusion matrix
roc.curve(Test.cat$outcome, tree.pred.1.both, plotit = FALSE)

## Area under the curve (AUC): 0.946

table(Test.cat$outcome, tree.pred.1.both)

##       tree.pred.1.both
##        pass fail
##   fail    1   24
##   pass  220   16
```

Balancing the dataset raised the predictive specificity to 96% and the balanced accuracy to 95%. It also raised the AUC of the ROC curve from 0.836 to 0.946. This model is as good as the logistic regression for prediction.

One of the advantages of decision trees over logistic regressions is that they are visual and easy to interpret. So what does this decision tree look like?

```
library(rpart.plot)
prp(tree.mod.1.both)
```



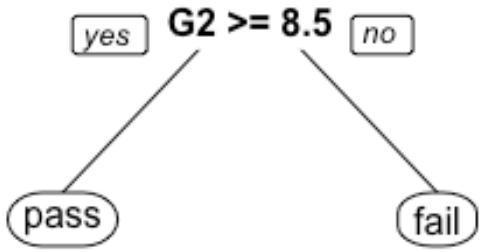
It may be a good predictive model, but it's not very interesting! It simply predicts that all students with G2 greater than or equal to 8.5 will pass.

Would the model be improved by considering other predictors besides G1 and G2? For this model, experimentation revealed that undersampling was the best balancing method. (See Code lines 1114-1140 for details.)

```
# Build model with balanced data
tree.mod.7.under <- rpart(outcome ~ ., data = Train.cat.under, method =
"class")

# Apply to test data
tree.pred.7.under <- predict(tree.mod.7.under, newdata = Test.cat, type =
"class")
```

Visually inspect and evaluate this model.

```
# Examine AUC of ROC curve and confusion matrix
roc.curve(Test.cat$outcome, tree.pred.7.under, plotit = FALSE)

## Area under the curve (AUC): 0.946

table(Test.cat$outcome, tree.pred.7.under)

##       tree.pred.7.under
##        pass fail
##   fail    1   24
##   pass  220   16

# Visual representation of tree
prp(tree.mod.7.under)
```



Interestingly, this model is identical to the decision tree model built with just the predictors G1 and G2. Adding more predictors did not improve the model at all, so for reasons of parsimony the previous model is preferable.

QUESTION 2 PART 1: PREDICTING FAILURE WITHOUT INTERIM GRADES AS PREDICTORS

I explored this question exhaustively, building dozens of linear and logistic regressions, decision trees, and random forests. I experimented with balancing the dataset by all the methods offered in the "ROSE" package or, in the case of random forests, by manipulating the classwt argument. I tried numerous combinations of independent variables as predictors, including those identified by forward- and backward- stepwise selection ("leaps" package) as the most likely combination of predictor variables for a good model. (See Code lines 546-1272 for the details of this investigation.)

In the end, I was forced to agree with the original researchers, Cortez and Silva, who wrote, "The obtained results reveal that it is possible to achieve a high predictive accuracy, *provided that the first and/or second school period grades are known*" (emphasis mine). Without some information about a student's interim performance, it is impossible to predict whether he will pass or fail from the information in the dataset.

QUESTION 2 PART 2: PREDICTING FAILURE WITHOUT G2 AS A PREDICTOR

If educators can't predict who is likely to fail before the term begins, they would at least like to know as early in the term as possible who is at risk and would benefit from academic or other interventions. Therefore, a model that could predict failure without G2 (i.e. as soon as the G1 grades are available) would be valuable to educators.

As in Question 2 Part 1, the regression models, decision trees, and random forests that excluded G2 as a predictor were not very effective in predicting outcomes. (See Code lines 1274 to the end for the details of this investigation.)

I did find one logistic regression worth discussing. It depends on the undersampling method to balance the dataset, and uses G1, previous course failures, and the course (math or Portuguese) as the predictors.

```
# Build the model
log.model.10.under <- glm(outcome ~ G1 + failures + course, data =
Train.cat.under, family = binomial)

# Apply it to the test data
log.pred.10.under <- predict(log.model.10.under, newdata = Test.cat, type =
"response")

# Examin the AIC and confusion matrix
summary(log.model.10.under) # AIC: 112.51
```

```
##
## Call:
## glm(formula = outcome ~ G1 + failures + course, family = binomial,
##     data = Train.cat.under)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.39888  -0.40223  -0.01687   0.51099   1.78288
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    7.79936    1.42623   5.468 4.54e-08 ***
## G1            -0.83277    0.15514  -5.368 7.97e-08 ***
## failures1      0.01664    0.60079   0.028   0.9779
## failures2     18.85914 1608.59846   0.012   0.9906
## failures3      0.42882    0.82892   0.517   0.6049
## courseport    -1.11280    0.51108  -2.177   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 207.94  on 149  degrees of freedom
## Residual deviance: 100.51  on 144  degrees of freedom
## AIC: 112.51
##
## Number of Fisher Scoring iterations: 17
```

```
table(Test.cat$outcome, log.pred.10.under > 0.5) # Confusion matrix
```

```
##
##        FALSE TRUE
##   fail     2   23
##   pass   200   36
```

This model successfully predicts 23 of the 25 failures, for a specificity of 92%. However, in order to correctly catch this many students in the "fail" class, the model miscategorizes 36 students who passed. The balanced accuracy of this model is 89%, not bad, but not nearly as good as the models that included G2 as a predictor.

SUMMARY AND IMPLICATIONS

1. *95% of students' final grades can be predicted directly from their interim grades*

For the vast majority of students, the linear relationship between G1/G2 and G3 is very strong. The very best predictor of most students' final grade is their grade at G2. For most students, G1 is also a very strong predictor for G3.

While this is not exactly breaking news to educators, it does offer a very simple predictor of a major group of students who are in need of intervention: Any student who is failing at the end of either the first or the second marking period is at serious risk and needs immediate attention. Furthermore, the lower G1 and/or G2 is for a particular student, the greater that student's risk of failing.

2. *Dropouts can be predicted using statistical learning techniques*

While the linear relationship between G1/G2 and G3 is a good predictor for most students, there is a group of outliers, about 5% of the total, whom I've called “dropouts.” These students earn grades usually between 5-10 at G1 and/or G2, but receive a final grade of 0. By definition, these students' final grades cannot be predicted by linear models. So how can educators guess who, among students who are earning low-fair interim grades, is at greatest risk of dropping out?

I have described a logistic regression model balanced by undersampling, and also a decision tree model balanced by both over- and undersampling, each of which has a high accuracy rate for predicting the students in both the "fail" group and the "pass" group.

For educators, these predictive tools hold promise for detecting difficult-to-identify students whose interim grades are not directly predictive of their final grades. Any student predicted to fail by either of these tools is at serious risk and needs immediate academic or other intervention in order to prevent dropping out.

3. *It is possible to identify students at high risk of failure before the second grading period*

The best predictive models require both first and second period grades as predictors. However, early intervention is highly desirable for best student outcomes, so it is helpful that it is possible to build a good logistic regression model without G2 as a predictor. This model, balanced by undersampling, has a high accuracy rate for predicting the students in the "fail" group, at a cost of misclassifying a larger number of students in the "pass" group. This means that the balanced accuracy of this model is not as high as for the models that include G2. However, given the very high value of catching students at risk of failing as early as possible, and also given the fact that those misclassified false "fail" students will likely benefit from (or at least won't be harmed by) extra academic support, this cost seems of small concern from the perspective of educators whose highest goal is to prevent academic failure.

4. *Demographic and social information about students is not predictive of their academic achievement*

As Cortez and Silva found, without the interim grades, it is impossible to build a predictive model from this dataset. While this fact frustrates the effort to anticipate students at risk of failure in advance of their first failing grade, in a sense this is great news for educators who are committed to the belief that a learner's prospects are not predetermined by her circumstances as understood in this dataset. No variable—not sex, nor parents' level of education, nor rural address, nor weekday drinking habits, nor any other demographic or social classification included in this dataset—contributes in a meaningful way to the

prediction of a student's final grade. Nothing in this dataset gives educators any reason not to approach each individual learner, regardless of background, as a potential top student.

RECOMMENDATIONS AND POTENTIAL NEXT STEPS

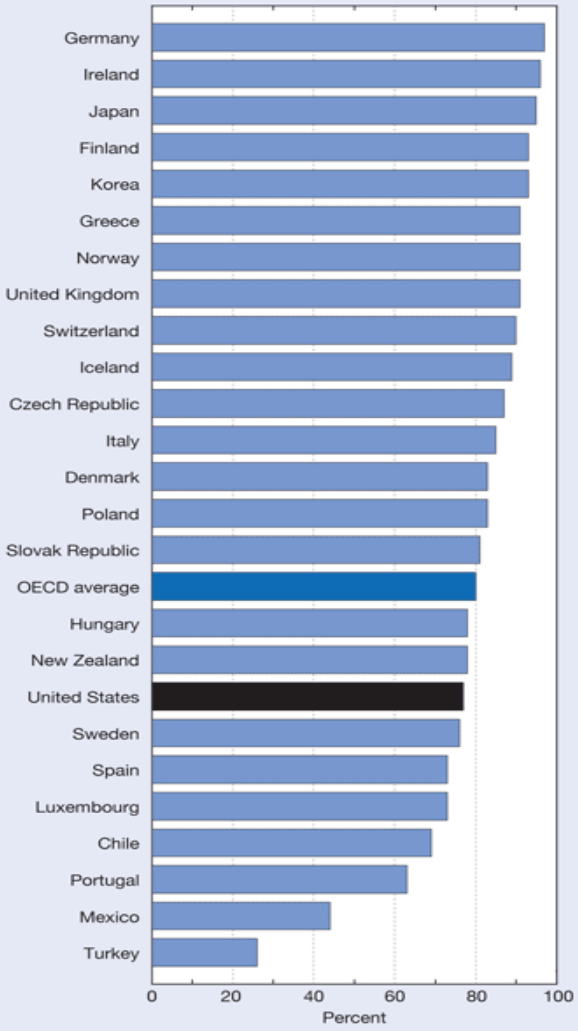
1. The fundamental takeaway is that students whose interim grades are low are at great risk of failing. This is true of both low-achieving students who stick through to the end of term (but fail) and also of low-achieving students who drop out (receiving a final grade of 0). With or without the tools of statistical learning, educators already have the information they need to to direct their attention where it's needed most. Any plan for academic intervention should focus on students who are not passing the class at the interim marking periods.

2. For the purposes of predicting which students are at risk of failing, the demographic and social information included in this dataset is not useful. This is not to suggest that demographic and/or social factors have no impact on students' academic success, but rather to challenge future researchers to collect and consider different, perhaps more relevant information. Some possibilities:

- family income
- food security
- stability of housing
- incarcerated parent
- juvenile infractions
- drug use/addiction
- pregnancy/parenting
- is the student employed
- experience of trauma
- citizenship status
- home language
- more information about student's health

APPENDIX A

Figure 1-14
High school graduation rates, by OECD country: 2008



OECD = Organisation for Economic Co-operation and Development

NOTES: High school graduation rate is percentage of population at typical upper secondary graduation age (e.g., 18 years old in United States) completing upper secondary education programs. OECD average based on all OECD countries with available data. To generate estimates that are comparable across countries, rates are calculated by dividing the number of graduates in the country by the population of the typical graduation age.

SOURCE: OECD, *Education at a Glance: OECD Indicators 2010* (2010).

Science and Engineering Indicators 2012

APPENDIX B

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets

Label	Attribute	Description
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 type	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 – 5th to 9th grade, 3 – secondary education or 4 – higher education
Fedu	father's education	numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education
Mjob	mother's job	nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other"
Fjob	father's job	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	binary: yes or no
activities	extra-curricular activities	binary: yes or no
nursery	attended nursery school	binary: yes or no
higher	wants to take higher education	binary: yes or no
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