Predictive Analytics Exam

Sample Project Report – Student Success

Note to Candidates: This report represents a high-quality solution. For this sample project, the solution is intended to be difficult, but not impossible to create in the time allotted. Some tables and graphs are deliberately presented in a rough form to illustrate exam expectations.

To: Steve Jones

From: Me

Date: Today

Title: Drivers of Passing Course Grades

Executive Summary

In today's dynamic economy, education is more important than ever. School administrators need to recognize those who are likely to succeed and those who may need some extra help. School Wiz has asked us to determine if we can identify a set of behavioral and demographic factors that predict which students will pass or fail so they can direct remedial services toward students most likely to benefit.

The most important factors are (with the direction that implies a higher chance of the student passing):

1. Parents' education (more)
2. Previous failures (fewer)
3. Going out with friends (less)
4. Family supplement (not present)
5. Internet access (present)

It is important to keep in mind the sensitive nature of some of the information contained and it may be unavailable to educators, inappropriate to use, or difficult to implement. These concerns are not addressed in this analysis but should be addressed before a final model is developed.

Because School Wiz is most interested in passing versus failing, we did not attempt to predict the score a student would earn, only if a student can earn a score of ten or more. To that end, we note the following:

- With no predictors available, the "best" prediction would be that everyone passes. Because 64% of the sampled students pass, this prediction is seen to be accurate 64% of the time.
- To test our model, we held out a sample of 140 students in which 64% also passed. Applying our model to these students, our prediction was accurate 73% of the time.

© 2018 Society of Actuaries

Should you agree to engage our consulting services, note that some of the variables we used may turn out to be unavailable (which would reduce the accuracy of our final model). However, with more time (and a larger dataset) we should be able to improve upon the models we built for this demonstration.

Data Exploration, Preparation, and Cleaning

The dataset provided included demographic, behavioral, and past performance records of 585 students. Some of the information contained within the dataset may not be available to School Wiz, and thus has not been used for our analysis, specifically absences, G1, and G2. There are other variables that may be inappropriate for an educator to know and/or take into the decision-making process; however, for the purposes of this analysis, all available variables have been used.

There is one target variable to consider (pass/not pass), but two separate ways to interpret the target:

1. Directly predicting pass/not pass (classification problem)
2. Predict grade, and then infer pass/not pass (regression problem)

Given that School Wiz is only interested in pass or fail, I have concentrated on Option 1 in this analysis. Given more time, Option 2 could be investigated to see if it performs better.

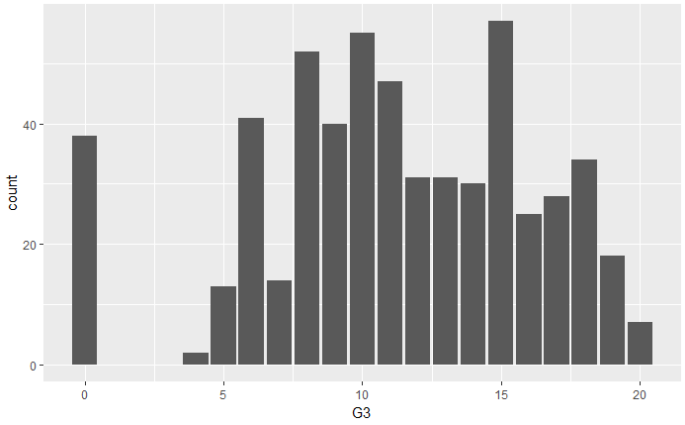
A summary of the dataset is provided in Appendix 1. In exploring the summary, we notice that no variables contain NA values, but G3 contains values both larger (>20) and smaller (<0) than could be obtained. These entries have been excluded from the analysis, and this reduced the dataset to 568 observations. An oddity was that 38 students had a grade of 0 while none had grades of 1-3. It could be that 0 means something other than a poor grade, but, for this preliminary analysis, these were retained. A failing grade is indicated by G3 < 10 and passing is all other grades.

The dataset contains a large number of variables compared to the number of individuals studied, with 568 rows and 29 predictor variables - one row for each student in the study. Each row contains information about the student including school, sex, age, address (urban or rural), family size and quality of relationships, parent and guardian status, mother & father job and education, motivation for school and higher education, travel time to school, amount of time studying, previous failures, school and family supplements, additional paid classes, relationship status, amount of free time, social life, alcohol consumption (both on weekdays, and weekends) and health.

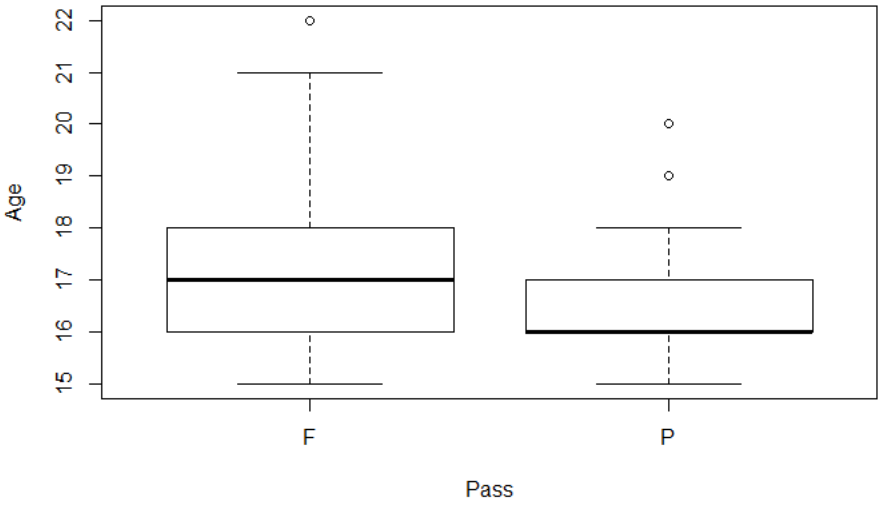
The following observations stand out

The target variable, G3 or final grade, is non-normally distributed, with a high weight at the zero point. The high proportion of zeros may make G3 a difficult response variable to model as a continuous variable, supporting the decision to model G3 as a pass/fail variable.

© 2018 Society of Actuaries

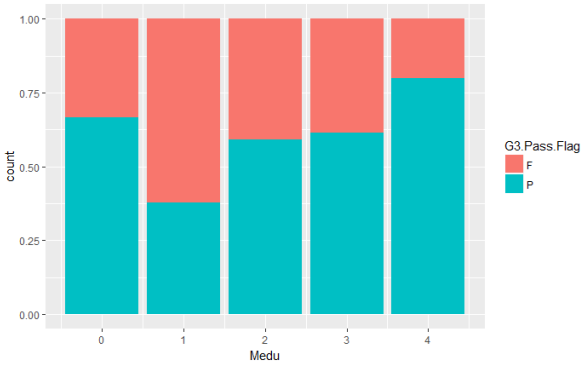


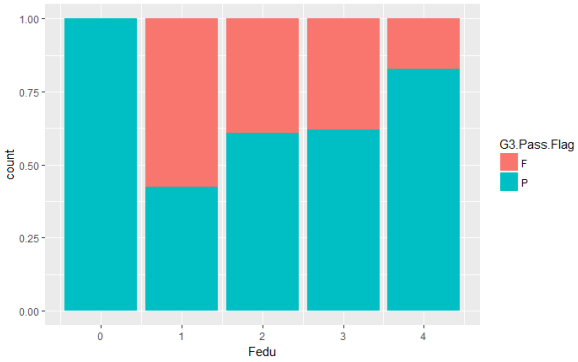
The age range is from 15 to 22. The high end of the age range seems old for high school. This would be an appropriate question to reach out for further explanation from School Wiz. Box plots of age (see below) indicate there is a relationship between age and passing and that there are few of these high age students. For this analysis no adjustments were made.



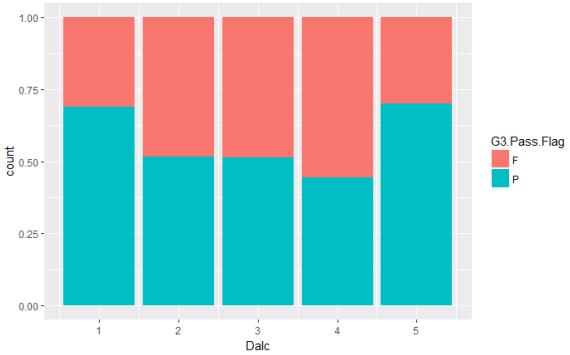
For the categorical variables (including those on a 1-5 scale) I plotted the proportion passing for each factor level. Three stood out. Plots of Medu and Fedu (mother's and father's education) against passing and failing produced an odd result. As expected, more education led to higher percentage passing (see the graphs below), except that those with education = 0 had higher pass rates. It is not clear what education = 0 means (the data dictionary says "none," which is unlikely). These five records were removed, reducing the total number of records to 563.

© 2018 Society of Actuaries





The graph for Dalc (weekday alcohol consumption) showed that those in the highest category were more likely to pass (see graph below). There were 10 students in this category and with no reason to question the validity of this result, no changes were made.



The other plots did not show strong relationships between the variable and passing. I also examined the correlations between grade (not passing and failing) and the numeric predictor variables. They indicated a poor relationship in most cases, with greater parent education being the most notable. I decided to let the modeling process perform variable elimination.

	age <dbl>	Medu <dbl>	Fedu <dbl>	traveltime <dbl>	studytime <dbl>	failures <dbl>	famrel <dbl>	freetime <dbl>
G3	-0.17	0.41	0.37	-0.06	0.03	-0.36	0.11	-0.07

goout <dbl>	Dalc <dbl>	Walc <dbl>	health <dbl>	G3 <dbl>
-0.18	-0.15	-0.25	-0.04	1.00

After cleaning the data, we observe that 64% of the sample passed. Thus, the accuracy measure for any model must exceed this value for the model to be an improvement over predicting a pass for every student.

© 2018 Society of Actuaries

Feature Selection

The following four new variables were created:

Variable Name	Transformation	Reason	New Column
Dalc & Walc	Dalc * Walc	Alcohol consumption may be compounding	combine.alc
Medu & Fedu	Medu * Fedu	Parents' education may be compounding	combine.education
Medu & Fedu	If Medu & Fedu both 4, then 1	Parents both attending college may signal higher grades	both.college
failures	flag for any failures existing	Failing previously is predictive of future failures	failures.flag

Multiplying variables together or creating flags allows algorithms to pick out the patterns and interactions much more easily than hoping that the algorithm finds them itself. For example, in a linear regression, the effect of moving from 0 to 1 failure is the same as moving from 1 to 2. The flag gives the potential for an extra boost (or penalty) for those with at least one failure.

Because of the high number of variables and the high correlations between some of the independent variables, using all the variables in a regression setting will lead to poor prediction. Feature selection can be conducted in several ways. We conduct a full analysis using all the variables and assess variable importance in the full model. We also use regularization to simultaneously fit the model and remove unimportant variables. These approaches are detailed in the modeling section.

Model Selection and Validation

A stratified random sampling approach was used to get the same proportion of variables in the train and test sets due to the imbalanced sample of passers (64%) to non-passers (36%). I placed 75% of the data in the training set and the remaining 25% in the test set.

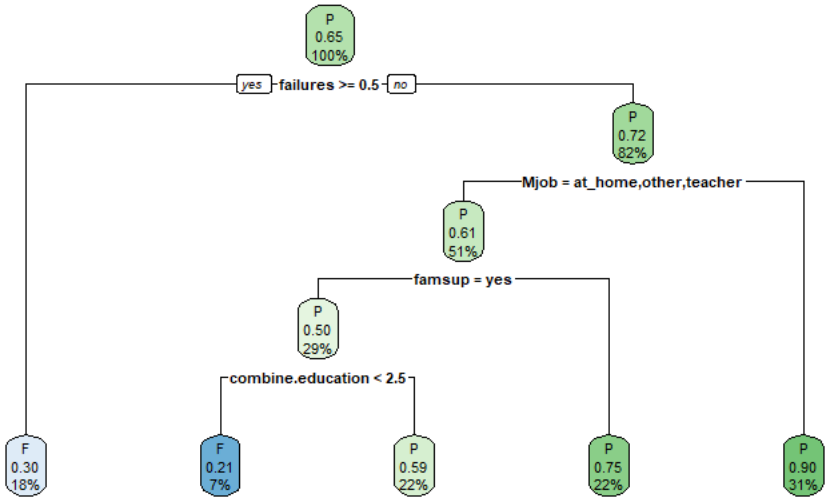
Model 1 – Decision Tree Model

The first model is a decision tree model, used to get a view on which variables are the most impactful in passing. The initial decision tree model contained too many splits and likely overfit the data. It did perform well with accuracy 86% on the training set and 70% on the test set. The reduction in performance on the test set supports the overfitting observation. For the training set, 265/423 = 63% were predicted to pass. As a result, the cutoff of 0.5 seems reasonable and will be used throughout. I modified the parameters to better prune the tree:

- minbucket: increase the minimum observations in a split from 5 to 10;
- cp: increase the complexity parameter from .001 to .02 to increase the required gain from splitting; and
- maxdepth: decrease from 20 to 10, to limit the depth of splitting.

Making these changes resulted in a less complicated model (shown below), but one which is only marginally better than guessing the majority class, with an accuracy of 75% (guessing the majority class would yield an accuracy of 64.5%) on the training set and 71% on the test set.

© 2018 Society of Actuaries



Decision Tree Model Confusion Matrix:

		Test - 25%, Blind		**Train - 75%**	
		Actual		Actual	
		Pass	Fail	Pass	Fail
Predicted	Pass	76	27	244	75
	Fail	14	23	29	75

While I could have used a formal pruning method, due to a lack of stability and low accuracy compared to the random forest model, I do not recommend using this model.

Model 2 - Random Forest Model

The second model I built is a Random Forest model to get an idea for which variables were the most impactful, and then using those insights to build a GLM. Random forest models also do a good job of picking out any nonlinear interactions within the data. Due to time constraints, I only used 50 trees in the forest. We may want to try this again with more trees if we win the contract.

This model achieved 100% accuracy on the training set and 79% accuracy on the testing set. Both 100% accuracy on the training set and the difference between the accuracy in the training and test set indicate the model is being overfit.

Random Forest Model Confusion Matrix:

		Test - 25%, Blind		**Train - 75%**	
		Actual		Actual	
		Pass	Fail	Pass	Fail
Predicted	Pass	74	13	273	0
	Fail	16	37	0	150

© 2018 Society of Actuaries

From the Variable Importance Plot in Appendix 2, we can see which variables are the most important. Those include, in descending order of importance: goout, combine.education, failures, Medu, faliures.flag, internet (yes), famsup (yes), Mjob (services), Fedu, and health.

The top predictive variables in the random forest model includes those in the decision tree model, so we feel confident building the GLM using the top variables from the random forest model.

GLM

For the third model (Model 3), a GLM was used because GLM coefficients are intuitive to understand and can be easily communicated. Using the variables from the random forest plot, I obtained an accuracy of 77% on the training set and 75% on the test set, values that indicate less overfitting compared to the random forest model.

The output was:

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.07296    1.01090   1.061  0.28851
goout             -0.49506    0.12377  -4.000 6.34e-05 ***
combine.education -0.05933    0.12157  -0.488  0.62553
failures          -1.38941    0.59305  -2.343  0.01914 *
Medu               0.49046    0.33501   1.464  0.14318
failures.flag      0.07083    0.82338   0.086  0.93145
internetyes        0.86130    0.32839   2.623  0.00872 **
famsupyes         -1.03837    0.26234  -3.958 7.56e-05 ***
Mjobhealth         1.42458    0.75994   1.875  0.06085 .
Mjobother         -0.36501    0.39250  -0.930  0.35240
Mjobservices       1.00028    0.46123   2.169  0.03010 *
Mjobteacher       -0.70048    0.53472  -1.310  0.19020
Fedu               0.34918    0.36666   0.952  0.34092
health            -0.15279    0.09750  -1.567  0.11710
```

Several variables were insignificant. I used the stepAIC procedure to see which ones should be removed. It removes variables in the following order: failures.flag, combine.education, and Fedu. For the one categorical variable with more than two levels (Mjob) the procedure removes all or none. To investigate the individual levels, it is best to ensure the base level has the most observations. It turns out services is the leading category. I next changed that to the base level and reran the GLM with variables retained by the AIC process.

The output is now:

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.49805    0.79386   3.147 0.001651 **
goout       -0.49500    0.12352  -4.008 6.13e-05 ***
failures    -1.35506    0.24151  -5.611 2.01e-08 ***
Medu         0.48625    0.14581   3.335 0.000853 ***
internetyes  0.87113    0.32650   2.668 0.007628 **
famsupyes   -1.03974    0.26144  -3.977 6.98e-05 ***
Mjobat_home -0.90170    0.45170  -1.996 0.045910 *
Mjobhealth   0.32993    0.67235   0.491 0.623637
Mjobother   -1.38851    0.35765  -3.882 0.000103 ***
Mjobteacher -1.74476    0.42517  -4.104 4.07e-05 ***
health      -0.14551    0.09622  -1.512 0.130479
```

© 2018 Society of Actuaries

The accuracy values are now 78% and 73%. I considered combining the insignificant level "health" with "services" but with only 26 observations in that category it is unlikely to make a difference.

GLM Model Confusion Matrix & Stats

		Test - 25%, Blind		**Train - 75%**	
		Actual		Actual	
		Pass	Fail	Pass	Fail
Predicted	Pass	73	21	238	58
	Fail	17	29	35	92

Penalized Regression Models

(Note to candidates: after the decision tree, random forest and GLM models, you have fulfilled the requirements as defined in the Sample Problem. This is an alternative approach for the GLM that could be used in place of AIC to remove variables.)

Next, because I have some extra time left, I will try penalized regression coefficient estimates rather than importance statistics. Regularized regression is an alternative to reduce the number of variables. Logistic regression is used to model the pass rate based on the additional variables. The hyper-parameter λ was tuned using cross validation and only the lasso penalty was used as it provides the most variable reduction.

During the model matrix creation process, all variables are standardized and thus interpretation of each variable can be less intuitive than under typical regression. This is a drawback of penalized regression models. Below are the confusion matrices for regularized regression; because accuracy of penalized regression was similar to that of the GLM, it was decided to go with the GLM. Reaching a similar result in two different ways gives me confidence that my methodology is reasonable.

The train set had an accuracy of 77% and the test set had an accuracy of 76%.

		Test - 25%, Blind		**Train - 75%**	
		Actual		Actual	
		Pass	Fail	Pass	Fail
Predicted	Pass	76	19	247	70
	Fail	14	31	26	80

I also note that the lasso retained most of the variables and so does not provide as much insight as the earlier GLM did.

Findings

Two distinct model frameworks were analyzed for predicting whether a student would pass the course at the end of the year, with different methodologies providing different results. Using a decision tree model (Model 1) to predict a passing grade results in an underfit model that has the property of being unstable. Using a random forest model (Model 2) to predict only a passing grade results in an overfit model but provides insight into which variables are most important. The insights from the decision tree

© 2018 Society of Actuaries

model were contained within the random forest model, and thus insights from just the random forest were used to inform the building of a generalized linear model (GLM).

From the GLM I determined:

1. Based on demographic and behavioral factors, a student can accurately be predicted to pass or not pass around 73% of the time, higher than the base assumption of 64% (assuming all students pass); and
2. The following variables have a statistically significant impact on predicting a student passing:
    a. Combined parents' education, particularly mother's education (positive)
    b. Social lives (going out) (negative)
    c. Mother with a job in services (positive)
    d. Receiving family supplements (negative)
    e. Having internet (positive)
    f. Previous failures (negative)

It is not surprising that having previous failures, spending excessive time with friends, and receiving family supplements is a marker for doing less well in class. This should not be confused with these factors *causing* students to do more poorly or that the programs do not work. But they can help identify students who are *likely* to do less well. Further, higher education for parents and access to internet makes sense in that they indicate an environment where passing is more likely. However, it seems arbitrary that having a mother in the services industry would relate to better student performance. Further investigation is needed into this matter.

Given more time, I would run a similar battery of models, but trying to predict G3 instead of G3.Pass.Flag. Changing the target variable would change the problem from a classification problem to a regression problem. Building a regression model for G3, it could easily be transformed to a pass prediction. There is the potential for a regression model to provide different results, and potentially a more accurate solution.

In using the predictive models developed, we can identify which factors are most likely to relate to a student failing a course, and direct further attention to those students. Further investigation is needed on certain variables to determine if they are viable to include into a model for School Wiz, as there are data and student privacy concerns, but we are confident an accurate model can be built to predict student failures.

© 2018 Society of Actuaries

Appendices

Appendix 1: Summary of Dataset

```
school    sex          age        address famsize   Pstatus      Medu            Fedu
 GP :483   F:305   Min.   :15.00   R:160   GT3:408   A: 46   Min.   :0.000   Min.   :0.000
 MHS:102   M:280   1st Qu.:16.00   U:425   LE3:177   T:539   1st Qu.:2.000   1st Qu.:2.000
                   Median :16.00                             Median :3.000   Median :3.000
                   Mean   :16.61                             Mean   :2.853   Mean   :2.668
                   3rd Qu.:18.00                             3rd Qu.:4.000   3rd Qu.:4.000
                   Max.   :22.00                             Max.   :4.000   Max.   :4.000

      Mjob           Fjob            reason      guardian     traveltime      studytime
 at_home : 73   at_home : 34   course    :222   father:135   Min.   :1.000   Min.   :1.000
 health  : 39   health  : 25   home      :166   mother:413   1st Qu.:1.000   1st Qu.:1.000
 other   :193   other   :291   other     : 50   other : 37   Median :1.000   Median :2.000
 services:184   services:176   reputation:147                Mean   :1.482   Mean   :1.988
 teacher : 96   teacher : 59                                 3rd Qu.:2.000   3rd Qu.:2.000
                                                            Max.   :4.000   Max.   :4.000

   failures      schoolsup famsup     paid     activities nursery   higher    internet  romantic
 Min.   :0.0000   no :527   no :264   no :346   no :271    no :107   no : 24   no : 98   no :359
 1st Qu.:0.0000   yes: 58   yes:321   yes:239   yes:314    yes:478   yes:561   yes:487   yes:226
 Median :0.0000
 Mean   :0.2855
 3rd Qu.:0.0000
 Max.   :3.0000

    famrel         freetime         goout           Dalc            Walc           health
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000
 Median :4.000   Median :3.000   Median :3.000   Median :1.000   Median :2.000   Median :4.000
 Mean   :3.935   Mean   :3.231   Mean   :2.997   Mean   :1.397   Mean   :2.174   Mean   :3.668
 3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000

      G3         G3.Pass.Flag
 Min.   :-19.00   F:204
 1st Qu.:  8.00   P:381
 Median : 11.00
 Mean   : 11.93
 3rd Qu.: 15.00
 Max.   : 92.00
```

Appendix 2: Random Forest Variable Importance Plot