Predictive Analytics Exam

Sample Project Report – Student Success

To: Steve Jones

From: Me

Date: Today

Title: Drivers of Passing Course Grades

Executive Summary

If School Wiz knows which students are likely to fail as well as which factors impact pass rates then they can use this information to help teachers to improve student pass rates. This improves remedial student outcomes by helping students and teachers to succeed.

Using predictive analytics we have found leading indicators of student failure. From the beginning of the school year we can predict which students will end up passing with an accuracy of 69%. This number may not seem that high, but this model actually does a better job of detecting which students are likely to pass. Of the students predicted to pass by the model, 93% end up actually passing.

The leading indicators of student success are

- Number of absences
- Number of past failures
- The highest level of education by either father or mother
- Their health status
- Whether or not their family supports them (extra family supplement)
- The amount of weekend alcohol consumption

After taking into account all other factors we can measure how each of these changes the pass rate. Students that fail tend to

- Have a greater number of absences, with each additional absence increasing the likelihood of failing
- Have a history of failing
- Are of lower health status on a scale from 1-5
- Do not have a mother or father with a high level of education.
- Consume more alcohol on the weekends

It would be a simplification to assume that doing any one of these activities alone is a reason for failing. For students over the age of 21, for instance, who are of a reasonable health status and do not miss class there is no harm in consuming alcohol. For students who are missing class for legitimate reasons such as a health-related issue, there is a smaller penalty by the model. Missing is between 5-10 days of class in a school year is less of an issue for students who are in good standing than for those who have a history of failing.

This analysis is based on 585 actual students and information related to their demographics, lifestyle choices, and academic standing. There were 29 predictor variables considered. The objective was to make a model that will be actionable for teachers to use at the beginning of the school year, and so only variables which can be found before the beginning of the year were used. The 1st and 2nd trimester grades were not used in predicting whether or not the student had at least a value of 10 on the third trimester grade.

There was a correlation of 0.2-0.3 between the parents education level (on a numeric scale from 1-4) and the students' grades (G1, G2, and G3). Because using correlated variables together in models can cause difficulties, the variables for the mother's education (Medu) and the father's (Fedu) were combined into a single variable which was the highest of the two.

There were 53 students marked as outliers and removed. There were several other students who appeared to have records which were either in error or who should not have been included in this analysis, but given the complex nature of the academic data I did not feel that I could interpret it well enough to tell apart real cases from erroneous ones. School Wiz can inform us of any additional records which should be removed. The 53 students were removed due to

1) Having a negative grade
2) Having a father or mother with no education
3) Having greater than 40 absences

There were a number of variables which appeared in the data that had very minimal impact on the final model. All variables were tested out initially.

Before doing any modeling, 25% of the data was set aside to serve as a holdout set. None of the models "saw" this data when being built and were then evaluated based on how well they predicted the pass rates based on the these "unseen" students. A stratified sample was used to make sure that both training and test sets had the same proportion of passing and failing students. The pass rates was about 65% for both groups.

The modeling process was threefold: 1) a decision tree 2) a random forest and 3) a generalized linear model. Each method has advantages and disadvantages. The single decision tree provides interpretable output for School Wiz. The random forest has higher accuracy but is less interpretable. The GLM is more interpretable, stable when retraining on new data, easier to interpret, and simpler to implement.

The decision tree noticed interaction effects between student health and absences and past failures and absences. This information was used in the GLM modeling section to craft interaction features. The random forest had the highest accuracy (69%) and incorporated all variables. This model was then used to generate the marginal effects plots in the Findings section. Finally, several different GLMs were fit using only the most predictive variables. This strikes a balances between using all of the predictive power of a machine learning model while keeping the simplicity of a linear model.

The outputs of these models are probabilities of passing, or pass rates, for individual students. The cutoff value is the range from 0-1 which determines if the prediction is a pass or a fail. In order to help School Wiz, the cutoff value was fine-tuned so that the most students who fail were predicted to fail. Note that this value could be adjusted easily if School Wiz decides that predicting more or less student failures is more desirable. If they could estimate a “cost” of having a student failure then this dollar amount could be optimized directly.

The sample size of less than 600 is small which means that the results will only apply to students who are very similar to these in the data. In order to use this model on other student populations, a retraining exercise would be needed in order to capture the nuances of these other students.

A number of the variables used are based on a score from 1-5. This is an arbitrary measure that is difficult to interpret. A higher score always means “more of something”, such as better health or more education for the students parent, but a score of 1 for “health” is very different than a score of 1 for “gout”. The data dictionary should be used to interpret these models.

This data did not have missing values explicitly, but there could be hidden missingness in the survey responses. The project description did not mention this. If certain fields are 0 because they indicate a missing record then this analysis would need to be updated.

Many of the fields in this data could be considered private and personal information which School Wiz will not have access to or students would not feel comfortable releasing. Fortunately, the GLM and decision tree use only a handful of inputs and so these models could be used in the event that all of these variables are not available for students.

Survey responses are an unreliable data source because of the potential for survey applicants to be dishonest. For instance, a student filling out a survey on alcohol consumption who is under the legal drinking age is unlikely to answer truthfully if they consume illegally. To the extent that these fields are unreliable the results of this analysis will not hold for these fields.

Data Exploration, Preparation, and Cleaning

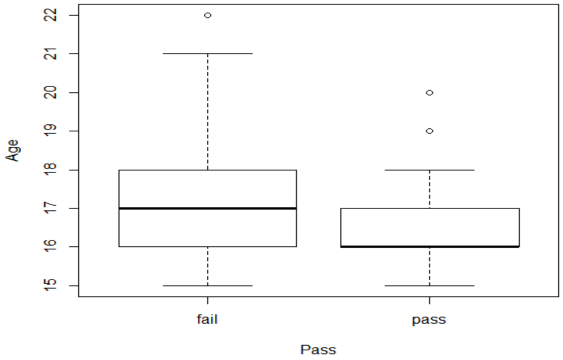
The data consists of 585 student profiles with information from 29 predictor variables on their school, age, gender, address, family size, lifestyle, health status, and prior test performance. The outcome is G3, the grade on the final trimester.

There were 41 records where G3 was either 0 or less than 0. These records were removed. A grade of 0 could indicate that these students were absent on the final semester, but this would be an assumption as the data dictionary does not provide this detail. To be safe, the 41 records were removed.

An indicator variable was created to determine if a student passed their final semester. This was 1 if their G3 grade was greater than 10 and 0 otherwise. The overall pass rate was 65%.

```
  pass_flag     n percent
1 fail        204   0.349
2 pass        381   0.651
```

There were a few outlying cases based on age. The outlying three points refer to students who have age greater than 18. Because these are remedial students it is expected that they are older than 18. Removing outliers could make the results of this analysis not apply to the most troubled, at-risk students who are older and continue to have failing grades. This could have significant impact given the small sample size of 585. To strike a balance between being fair to all students and having credible data, only students under the age of 20 were kept. This led to the removal of two students.



A summary of the numeric variables is below. I have checked that the ranges of each of these makes sense with the ranges from the data dictionary. For example, Medu, the years of mother's education is 0 – 4, indicating no education – secondary education. It did not make sense that a person could have *no* education and so these records were removed. The same was done for fathers with *no* education. This removed 5 records.

- Travel time has a max of 4, which means that a student is traveling at least 1 hour (not that they are traveling for 4 hours)
- Studytime has a minimum of 1, which means that all students are doing *some* studying
- There were three students with more than 40 absences which were removed

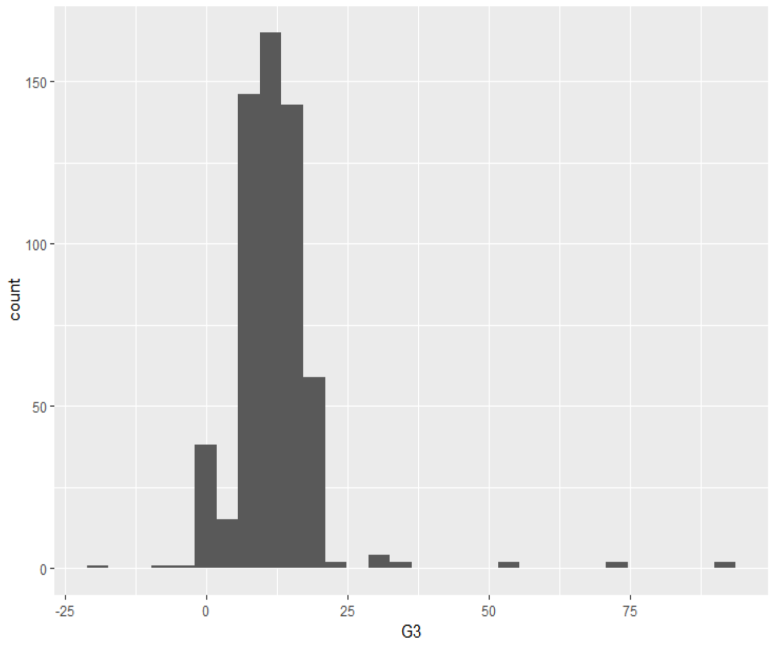
```
     age             Medu             Fedu          traveltime       studytime    
 Min.   :15.00   Min.   :0.000   Min.   :0.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:16.00   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000  
 Median :16.00   Median :3.000   Median :3.000   Median :1.000   Median :2.000  
 Mean   :16.59   Mean   :2.856   Mean   :2.674   Mean   :1.484   Mean   :1.991  
 3rd Qu.:18.00   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:2.000  
 Max.   :20.00   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000  
    failures          famrel         freetime         goout            Dalc      
 Min.   :0.0000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000  
 Median :0.0000   Median :4.000   Median :3.000   Median :3.000   Median :1.000  
 Mean   :0.2762   Mean   :3.931   Mean   :3.226   Mean   :2.993   Mean   :1.388  
 3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000  
 Max.   :3.0000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
```

```
      Walc          health         absences            G1              G2
 Min.   :1.000   Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00
 1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 1.500   1st Qu.: 8.00   1st Qu.: 8.00
 Median :2.000   Median :4.000   Median : 4.000   Median :11.00   Median :11.00
 Mean   :2.168   Mean   :3.674   Mean   : 6.292   Mean   :11.34   Mean   :11.09
 3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.: 9.000   3rd Qu.:14.00   3rd Qu.:14.00
 Max.   :5.000   Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00
```

To make sense of the percentages, the following table was created for each of the categories with values from 1-5. This shows that most students have very low alcohol consumption, do not fail often, have average family relationships, have parents with some college education, some free time, go out on the weekends, study between 2-5 hours per week outside of class, and tend to drink more on the weekends than during the week.

```
==========  ====  ====  ====  ====  ====  ====
feature        0     1     2     3     4     5
==========  ====  ====  ====  ====  ====  ====
Dalc        0.00  0.75  0.16  0.06  0.02  0.02
failures    0.83  0.10  0.04  0.03  0.00  0.00
famrel      0.00  0.02  0.04  0.20  0.50  0.25
Fedu        0.00  0.20  0.24  0.23  0.32  0.00
freetime    0.00  0.05  0.18  0.37  0.30  0.10
goout       0.00  0.07  0.31  0.31  0.21  0.11
health      0.00  0.11  0.09  0.22  0.19  0.40
Medu        0.01  0.16  0.21  0.22  0.40  0.00
studytime   0.00  0.29  0.50  0.14  0.07  0.00
traveltime  0.00  0.62  0.29  0.07  0.02  0.00
Walc        0.00  0.45  0.19  0.18  0.13  0.06
==========  ====  ====  ====  ====  ====  ====
```

The distribution of G3 grades is highly right skewed and centered at about 15. Even applying a log transform would not fully correct for this. Modeling for whether or not a student passes (has a higher grade than 10) leads to a more consistent measure of performance.



There is a strong correlation between the parent's education and the students grades. The correlation between Medu and Fedu and the G1,2,3 grades was between 0.2 and 0.3, which indicates that when the parents education is higher the students tend to perform better.

There was also a strong correlation between free time and the amount of alcohol consumption during the week and weekend.

In summary, there were 585 records in the original data set and 573 in after removing outliers.

Feature Selection

Because parent's education was so correlated with success, a new feature was created which is the max of the parent's education status. Because there are already so many features in this data, not a lot of time was spent in creating new features. 29 Predictors is already a lot to work with. If I had more time I could spend more time on this.

Model Selection and Validation

The data was split into 75% train and 25% test sets. The models were trained on the train set and not used on the test set until a final model was selected. To insure that both sets had the same number of good and poor students, stratified sampling was used with the pass flag variable (those students who had passed). The pass rates of the train and test sets were 65.4% and 65.5% respectively.

I noticed that some of the code was set up to train the model on the full data set. This was corrected so that the models were only trained on the training partition and then performance metrics were generated based on the test set.

The objective is to predict if a student passes their third trimester with a grade of at least 10. To make the results actionable for School Wiz, only the predictor variables which the school will know before the start of the school year were used. This means that in building the models, G1, G2, and G3 were excluded as predictors.

Decision Tree

The advantages to decision trees are that they

1) Detect nonlinearities
2) Detect interaction effects
3) Handing missing values
4) Are easy to interpret

The disadvantages are

1) Weak predictive power since each observation is an average over other terminal nodes
2) Can only predict within the range of training data, since eacy prediction is an average
3) Can easily overfit to the data

The goal of the tree was to produce interpretable insights. The random forest could be used for prediction. All features were included. The parameters for the minimum number of observations per node were set to 50 to reduce overfitting. The max depth was set to 5 to create a simpler tree.

Using the default cutoff of 0.5 meant that the sensitivity (true positive rate) was 63% which means that 63% of the time the model's predictions for which students would pass ended up passing. This is too low for School Wiz because it would result in a lot of student's with signs of failure going unnoticed.

I set the cutoff to 0.6 and the specificity increased to 0.17 which means that 17% of the students who would actually fail were predicted to fail. The results from scoring the model on the test set are below.

```
         Reference
Prediction  0  1
         0 17 75
         1 25 16
                                          
               Accuracy : 0.2481          
                 95% CI : (0.1774, 0.3304)
    No Information Rate : 0.6842          
    P-Value [Acc > NIR] : 1               
                                          
                  Kappa : -0.3176         
                                          
 Mcnemar's Test P-Value : 9.584e-07       
                                          
            Sensitivity : 0.4048          
            Specificity : 0.1758          
```

Random Forest

A random forest is a collection of bagged decision trees. Some of the advantages to using this model are

1) Greater predictive power
2) Handles more variables without overfitting because each tree can handle different variables
3) Detects non linearies
4) Detects interactions
5) Handles missing values

The main disadvantages are

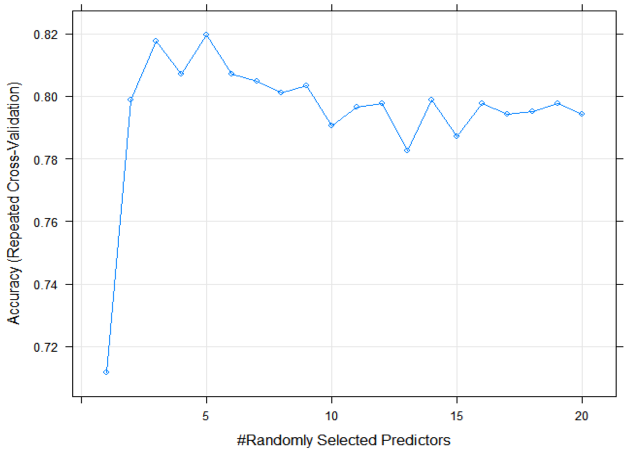
1) Is less interpretable
2) Is computationally difficult to implement
3) Takes longer to train

For the random forest model the same variables were excluded as in the decision tree. These are G1, G2, and G3. The target was the pass flag.

Cross validation that helps to make the most of the data. There are less than 600 observations, which means that it would be very easy to overfit with a flexible model such as a random forest. Cross validation was used to in order to produce realistic measures of the model's error rate. Using 5-folds, this splits the data into 5 groups of equal size and trains the model on 4 of the groups while using the 5th as a holdout set. Then this process is repeated for each of the five groups. Finally, this entire process was repeated twice. On a more powerful computer this should be increased to from 5 to further improve the training process.

The number of variables to consider at each split is an important hyper parameter for random forests. RF's perform best when the predictions from each tree is uncorrelated. If each tree were to be trained using the same features then this correlation would be close to one. By limiting the number of variables which are considered at each split to mtry the variance of the trees is increases thereby making the trees less correlated and improving the model.

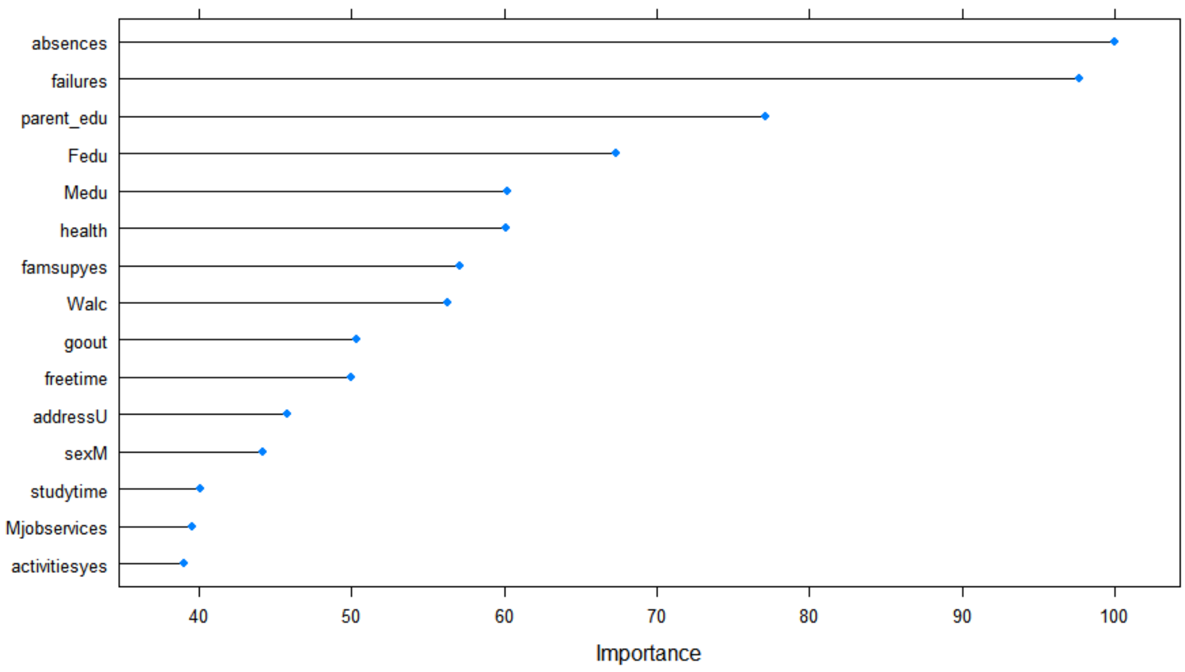
Below is the out-of-fold accuracy of different models against the value of mtry. The accuracy appears to be highest when mtry = 5. This is the value used in the final model.



The variable importance measures how much each variable changes across all trees changes the prediction of student failure. Because RF is based on decision trees, there are close connections with these importance rankings and the results from the single tree. Variables of higher importance appear higher up on the tree. The top two most importance features, absences and failure for example, are also the first two splitting points in the single tree.

The parent_edu feature is the feature which I created. This is the third most predictive feature.

Variable Importance of Classification Random Forest



The results from the random forest are much better than the single tree. See the output from the test s et below which uses the same cutoff value of 0.2. The overall accuracy is 69% as opposed to 25% and th e sensitivity is 93% as opposed to 41%.

```
          Reference
Prediction  0  1
         0 39 38
         1  3 53

               Accuracy : 0.6917
                 95% CI : (0.6058, 0.7689)
    No Information Rate : 0.6842
    P-Value [Acc > NIR] : 0.4674

                  Kappa : 0.4174

 Mcnemar's Test P-Value : 1.097e-07

            Sensitivity : 0.9286
            Specificity : 0.5824
```

GLM

Using all 29 features in a single GLM would likely lead to poor results and low interpretability. One of the key assumptions of GLMs is that the inputs are uncorrelated, which we have already seen not to be the case. A solution is to only use the most predictive features which have already been identified by the earlier steps. These are

- Number of absences
- Number of past failures
- The highest level of education by either father or mother
- Their health status
- Whether or not their family supports them (extra family supplement)
- The amount of weekend alcohol consumption

Note that it would be incorrect to include the parent_edu feature as well as Medu and Fedu because these would not form a linear independent set which would result in a rank-deficient fit on the GLM.

I tried out two different link functions, the logit and the probit. The logit is the default in R and the canonical link for the binomial response distribution. This lead to a test set accuracy of 17% and a sensitivity of 40%. I then tried the probit link, which is the inverse of the normal CDF. The accuracy improve to 21% and the sensitivity to 57%.

```
          Reference
Prediction  0  1
         0 24 87
         1 18  4

               Accuracy : 0.2105
                 95% CI : (0.1447, 0.2897)
    No Information Rate : 0.6842
    P-Value [Acc > NIR] : 1

                  Kappa : -0.2667

 Mcnemar's Test P-Value : 3.22e-11
```

```
            Sensitivity : 0.57143
            Specificity : 0.04396
```

Tree-based models automatically detect interaction effects. From the decision tree there were two key interaction effects. These were

1) Health status and the number of absences
2) Absences and the number of past failures

Both of these were added to the GLM and had improved the AIC from 387 to 388.

```
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)        0.4584180  0.4309559   1.064 0.287454    
absences          -0.0285347  0.0350100  -0.815 0.415046    
failures          -0.3774184  0.1852988  -2.037 0.041669 *  
health             0.0007162  0.0837963   0.009 0.993180    
parent_edu         0.1813228  0.0747759   2.425 0.015313 *  
famsupno           0.5701669  0.1530567   3.725 0.000195 ***
Walc              -0.0181638  0.0595542  -0.305 0.760370    
absences:failures  0.0122043  0.0199160   0.613 0.540017    
absences:health   -0.0142306  0.0086704  -1.641 0.100736    
```

The results on the test set are below. The cutoff was set turned up to 0.6 in order to increase the specificity. The accuracy is 23% and the sensitivity 45%. The specificity is the percentage of students who fail who are predicted to fail. If school Wiz wants to be able to intervene and reduce the number of failures then it needs a model which is optimized for specificity.

These results are worth than the random forest, but this is expected with a GLM. The advantages to using a linear model are that it is easier to interpret and easier to implement.

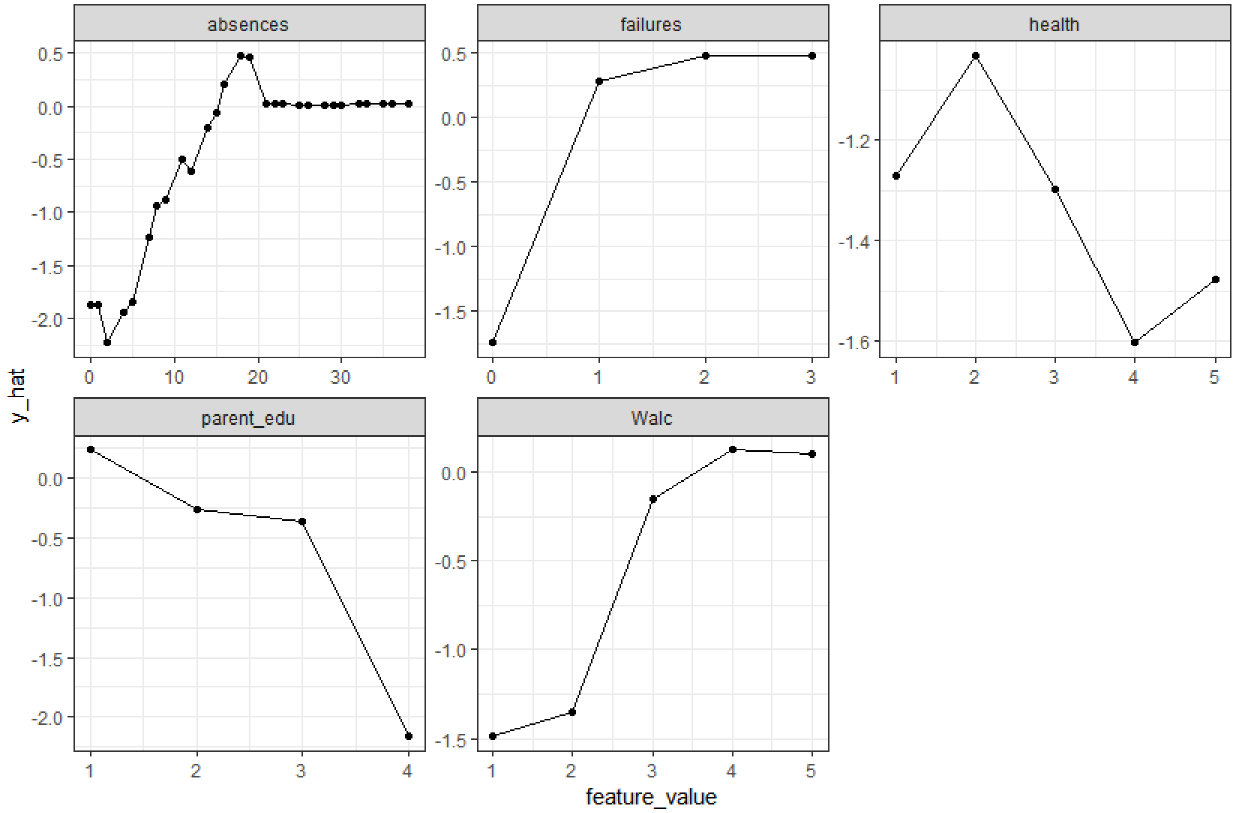
```
          Reference
Prediction   0   1
         0  71 323
         1  86  52
                                         
               Accuracy : 0.2312         
                 95% CI : (0.196, 0.2694)
    No Information Rate : 0.7049         
    P-Value [Acc > NIR] : 1              
                                         
                  Kappa : -0.2843        
                                         
 Mcnemar's Test P-Value : <2e-16         
                                         
            Sensitivity : 0.4522         
            Specificity : 0.1387         
```

Findings

We can measure how the probability of a student failing changes after into account all other levels in the model. The below results are from the partial dependence plots of the random forest model.

This shows that students who fail tend to

- Have a greater number of absences, with each additional absence increasing the likelihood of failing
- Have a history of failing
- Are of lower health status on a scale from 1-5
- Do not have a mother or father with a high level of education. Whether or not a student has a mother or father with an education does not make as much of a difference.
- Consume more alcohol on the weekends
- Whether or not their family provides supplemental support (not pictured)



* **Note** that the probabilities are not scaled and so produce values outside of 0 and 1 but that the relative change is consistent.

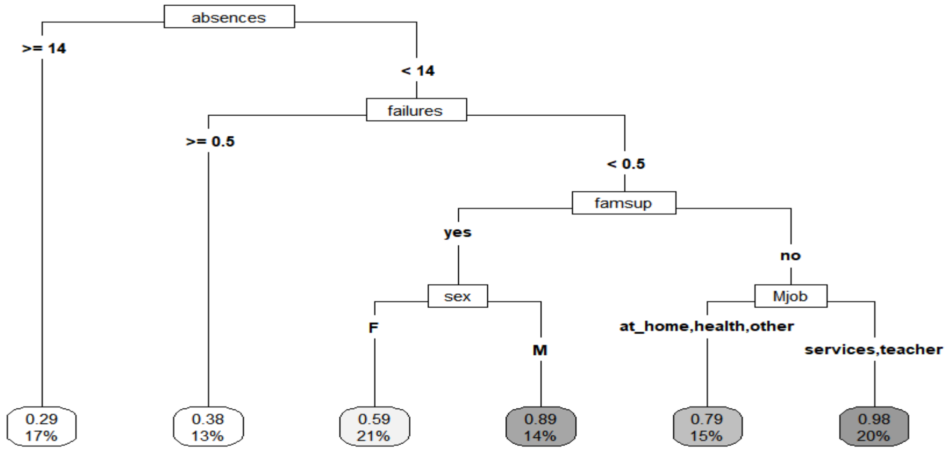
At the beginning of the school year we can predict which students will fail their third trimester by getting a grade less than 10 with 68% accuracy. This model can be used to support School Wiz faculty by helping them to flag students who are likely to fail and then take action to improve their grades.

We fit a decision tree model which can separate out passing students from failing students using simple yes/no question. The result of this model is that students who fail tend to

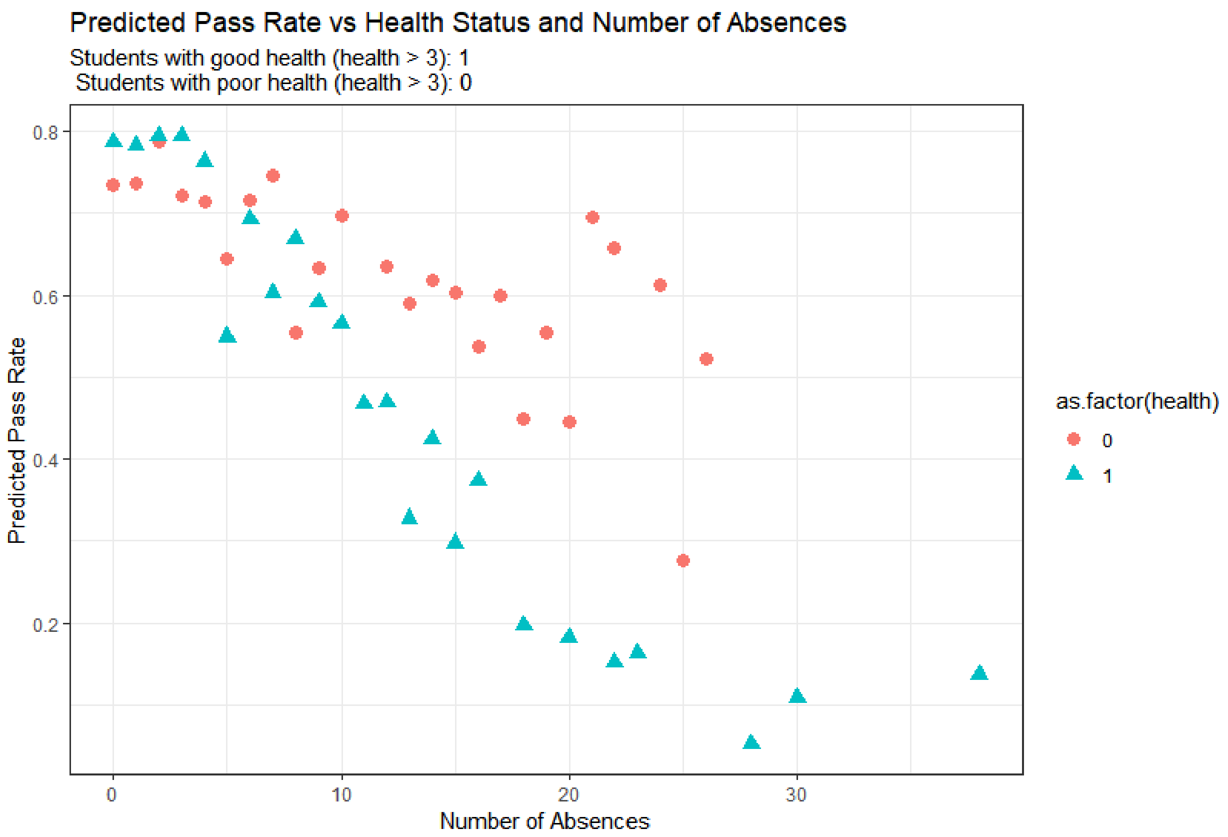
- Have more than 14 absences (17% pass rate)
- Have less than 14 absences but at least 1 past failed trimester (13% pass rate)
- Have less than 14 absences, no history of past failure, but have family support and are men. (14% pass rate).

Students who tend to perform better have

- Have less than 14 absences, no history of past failure, but have family support and are women (21% pass rate)
- Have less than 14 absences, no history of failing, no family support, but have a mother who is a teacher (20% pass rate)
- Have less than 14 absences, no history of failing, no family support, but a mother who is stay-at-home or other (15% pass rate)

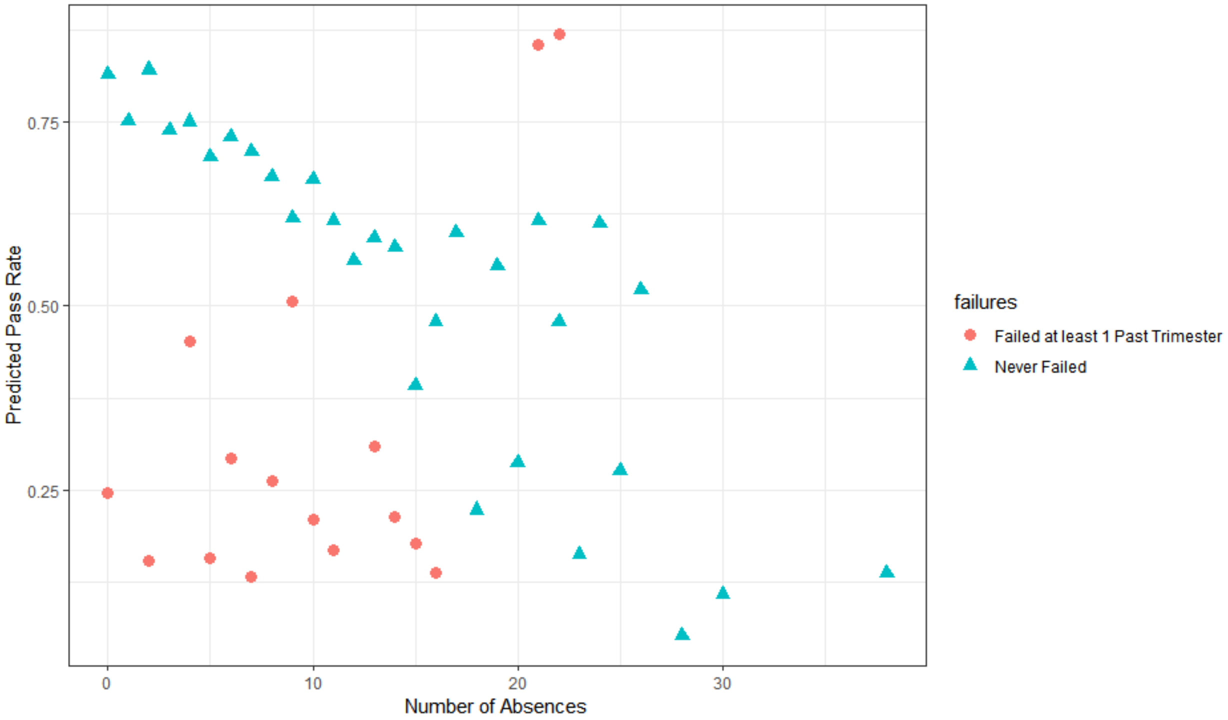


There are two complications to the above assumptions related to student health status, number of absences, and past failures. Students who have a high number of absences due to health-related issues still have a higher pass rate if they have missed school. If a student has a legitimate reason to miss class then they can still end up with a good grade. This is reflected in the graph below which shows the predicted pass rate form the GLM model.



The second interaction effect is around absences and the number of past failures. The trend for students who have never failed a past semester is different than those with a past of failing.

a. For students who have never failed (green triangles) the average pass rate decreases as they miss more days of class. For each 10 days of class missed the pass rate decreases by about 0.2.
b. For students who have failed before (orange circles), even missing a few days of class sets them on a course for future failure. They also have a lower pass rate overall, as can be seen by the group in the lower left corner.


Predicted Pass Rate
0.75
0.50
0.25
0
10
20
30
Number of Absences
failures
● Failed at least 1 Past Trimester
▲ Never Failed

Appendices