IST 707 - Applied Machine Learning

Student Grade Prediction

Mackenzie Houser, Blessy Thomas, Nathan Widlake

June 13th, 2023

Syracuse University

Introduction

Objective:

Two schools in Europe spent the last year collecting data on their students as well as their grades in a Math course. In order to discover who struggles the most on the topic, they hired a team of data scientists to model the data and discover how the school can better help their students in the future

Target audience:

School guidance counselors and teachers who interact with these students

Assumptions:

- Both schools have similar class sizes
- Both math courses have the same syllabus
- All classes designate the same amount of time per week

Business Understanding

Why do schools care about these grades?

With other schools competing for students, the school that can identify who might struggle and how to best adapt for them will receive the best reputation

How can we better understand our students?

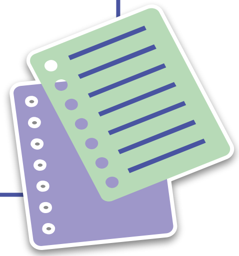
By collecting historical data about students, the school can identify trends and adjust teaching strategies to meet particular needs

What are the real world applications of this study?

A school is looked at by the grades it produces. With a successful study, the entire district can better prepare students for College while also benefiting their reputation which leads to increases in funding and more families desire to attend

Data Set

- The data was acquired from Kaggle
  - https://www.kaggle.com/datasets/dipam7/student-grade-prediction
- Created by University of Minho in Portugal
- A smaller dataset, with 33 variables and 395 rows
- There is a mix of Char, Int, Factor, Ordinal variables
- Our outcome variables is based on the final grade:
  - Those who have a final grade of an 11 or below (55%) 1 being True 0 being False
- Data is mostly clean but we'll look closely for pre-processing

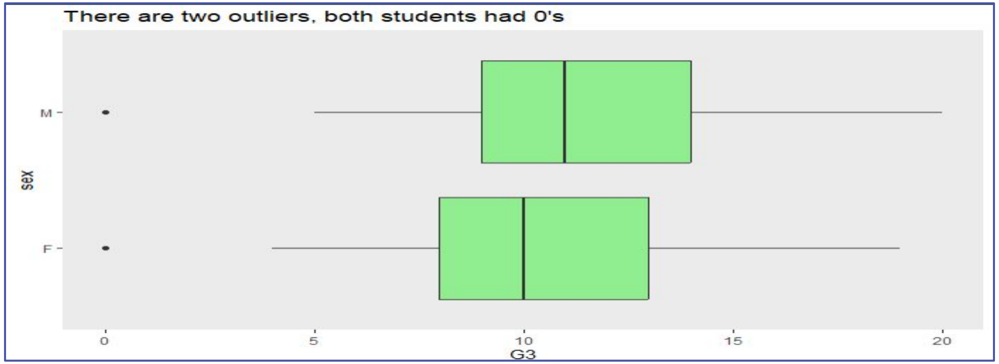


Data Dictionary

FIELD	TYPE	SAMPLE DATA	FIELD	TYPE	SAMPLE OF DATA
Age	Num	15 16 18 18 20 17 21 22	Paid	Factor	"Yes" "No"
Address	Factor	"R" "R" "U" "R" "U"	Activities	Factor	"Yes" "No"
Famsize	Factor	"GT3" "LE3" "LE3" "GT3"	Nursery	Factor	"Yes" "No"
Pstatus	Factor	"T" "T" "A" "A" "T"	Higher	Factor	"Yes" "No"
Medu	Ord.Factor	1, 2, 3, 2, 3, 2, 4	Internet	Factor	"Yes" "No"
Fedu	Ord.Factor	1, 2, 3, 2, 3, 2, 4	Romantic	Factor	"Yes" "No"
Mjob	Char	"At_Home" "Services"	Farmrel	Ord.Factor	1, 2, 3, 2, 3, 2, 4
Fjob	Char	"Teacher" "Health"	Freetime	Ord.Factor	1, 2, 3, 2, 3, 2, 4
Reason	Char	"Other" "Home" "Course"	Goout	Ord.Factor	1, 2, 3, 2, 3, 2, 4
Guardian	Char	"Mother" "Father"	Dalc	Ord.Factor	1, 2, 3, 2, 3, 2, 4
Traveltime	int	1 2 2 4 3 7 1	Walc	Ord.Factor	1, 2, 3, 2, 3, 2, 4
Studytime	int	1 2 0 0 8 4	Health	Ord.Factor	1, 2, 3, 2, 3, 2, 4
Failures	int	0 0 3 3 9 3 4 0	G1	Int	6 4 10 2 10
Schoolsup	Factor	"Yes" "No"	G2	int	6 5 15 21 18
Famsup	Factor	"Yes" "No"	G3	int	6 6 10 15 11
Sex	Factor	"M" "F"	Target	Factor	0 1 1 0 1 0 0
School	Factor	"GP" "MS" "MS" "GP"			

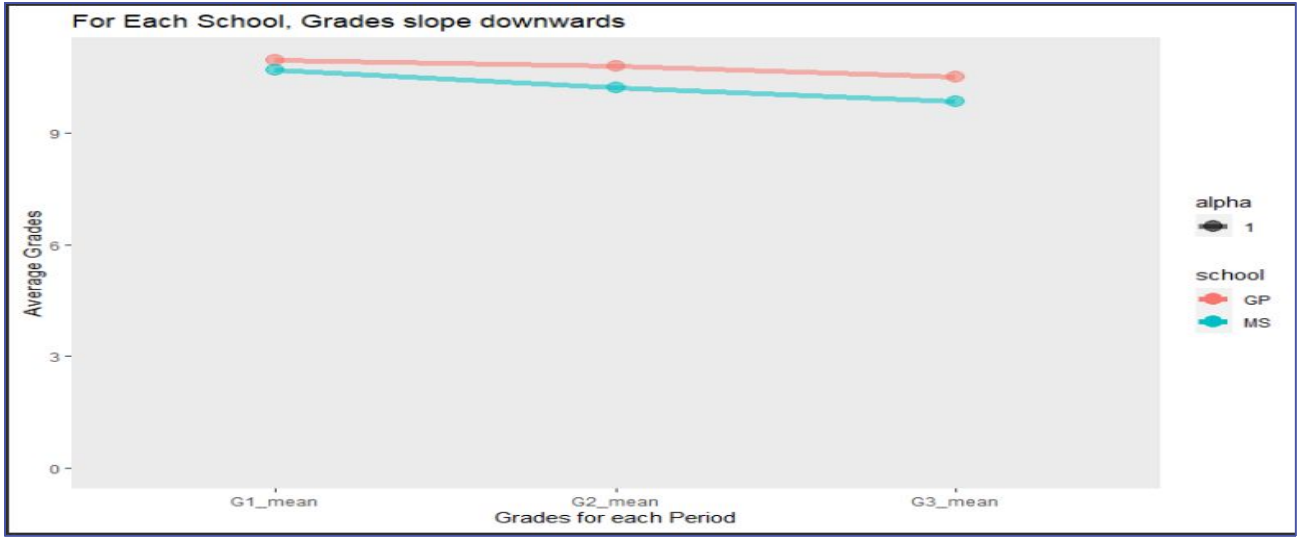
Data Pre-Processing

- Some outliers were discovered with students with a final grade of 0
    - Don't want to remove them since they are our target audience



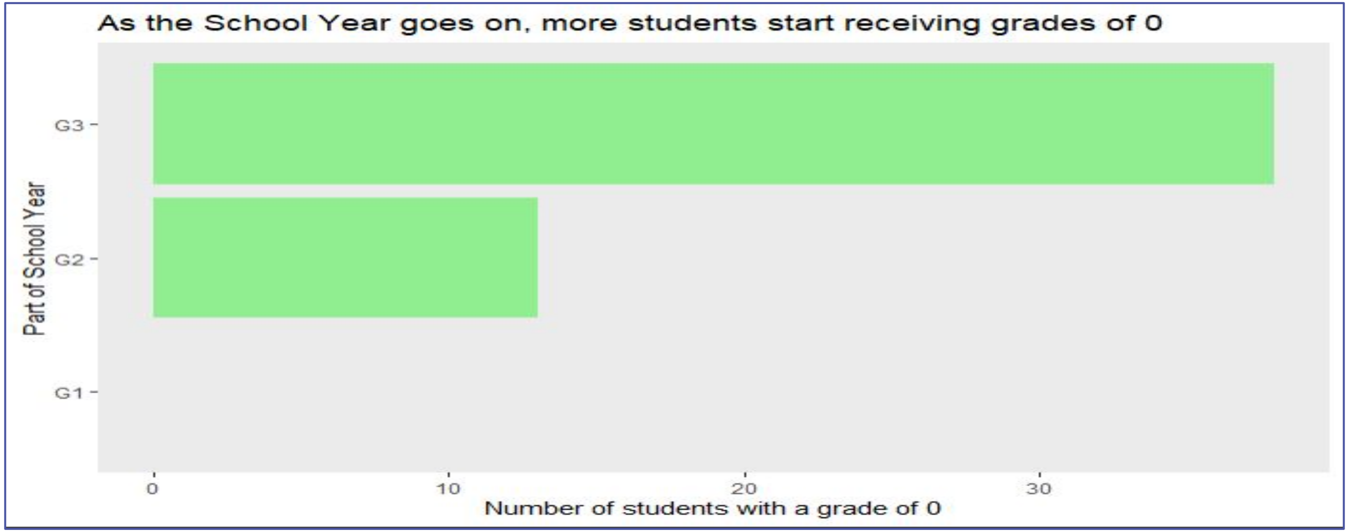
- No N/A data was found within the data set
- Data types were updated before the modeling to reflect each variables true nature

EDA



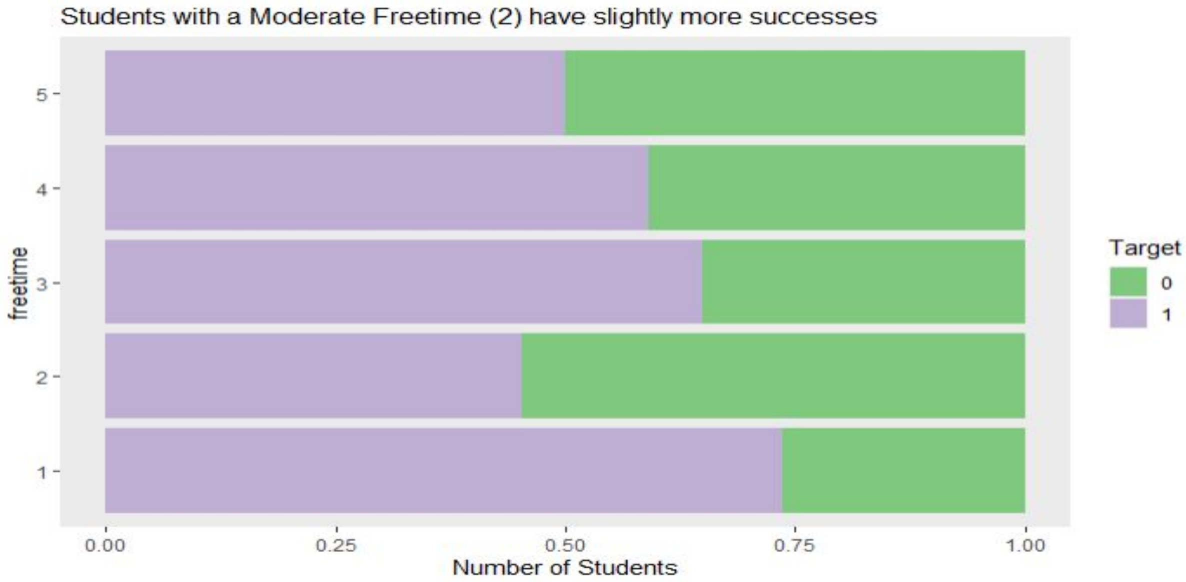
- Given that both the schools show a decrease in grades as the academic year progresses, they should consider trying to spread out the curriculum to try and get ahead of this dip

EDA



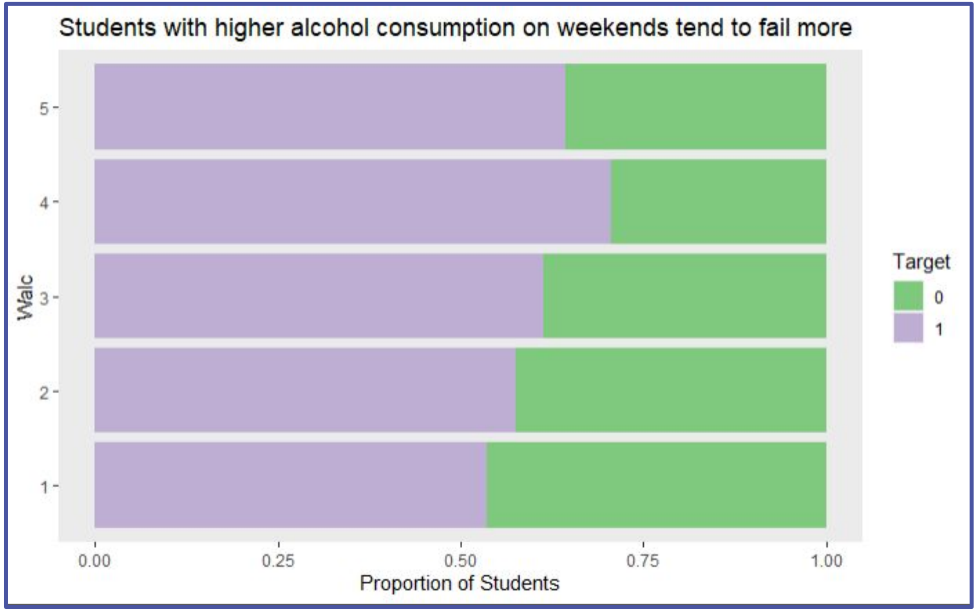
- Students who receive grades of 0 drastically increase as the year progresses
- This suggests students grasp the first part of the course well but later topics are the ones causing the challenge

EDA



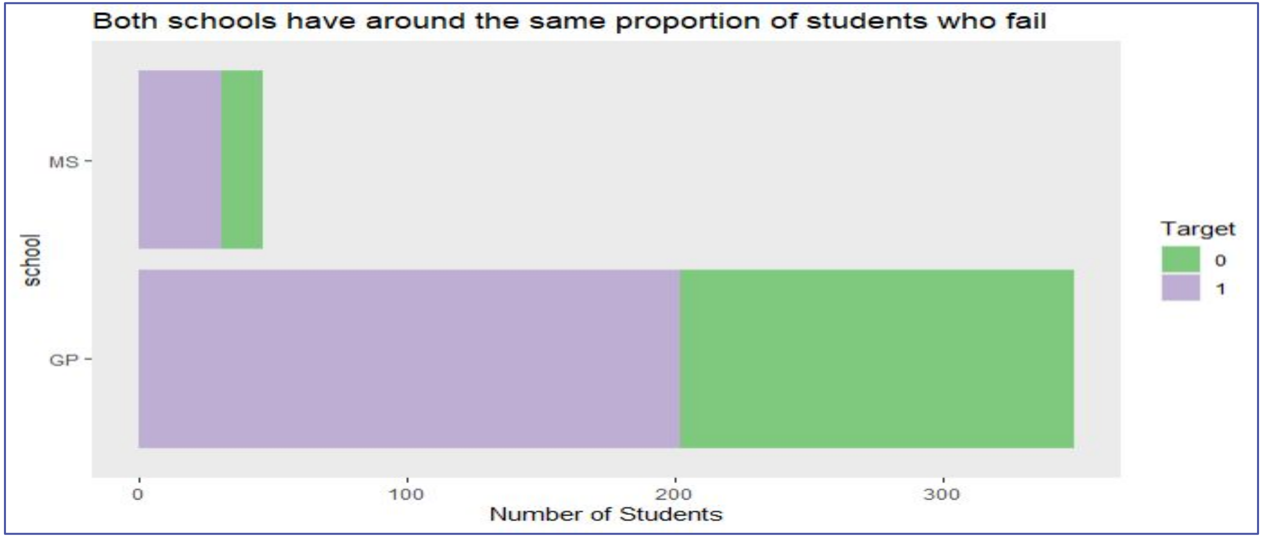
- People with more freetime most likely are the ones not taking the class seriously, we want our students to to have a moderate amount of time (2)

EDA



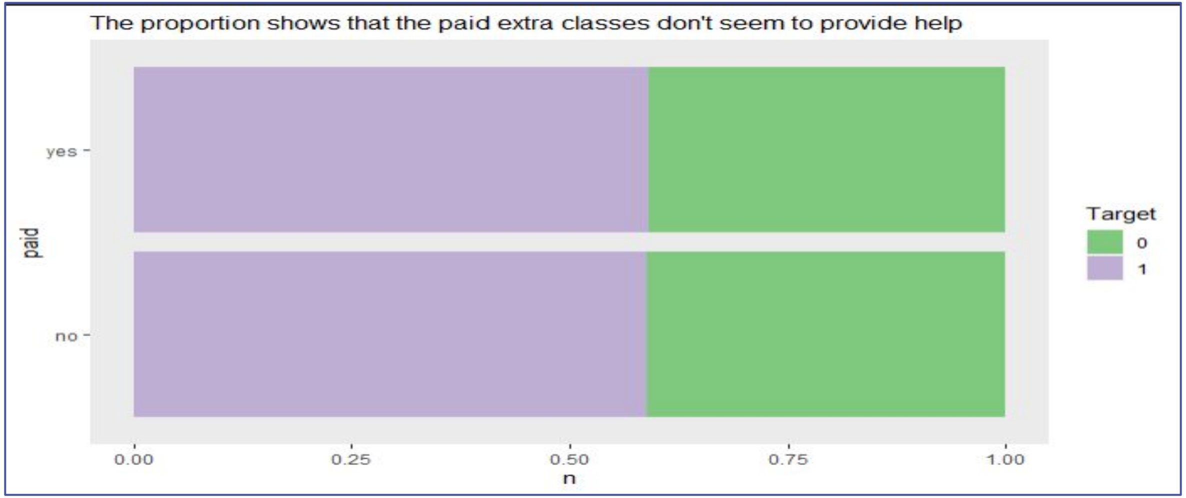
- We can see a direct increase in students who fail the class as their weekend alcohol consumption increases

EDA



- Both schools have around the same proportion of students who fail, suggesting that the environment isn't coming into play, the students are struggling with the course as a whole

EDA



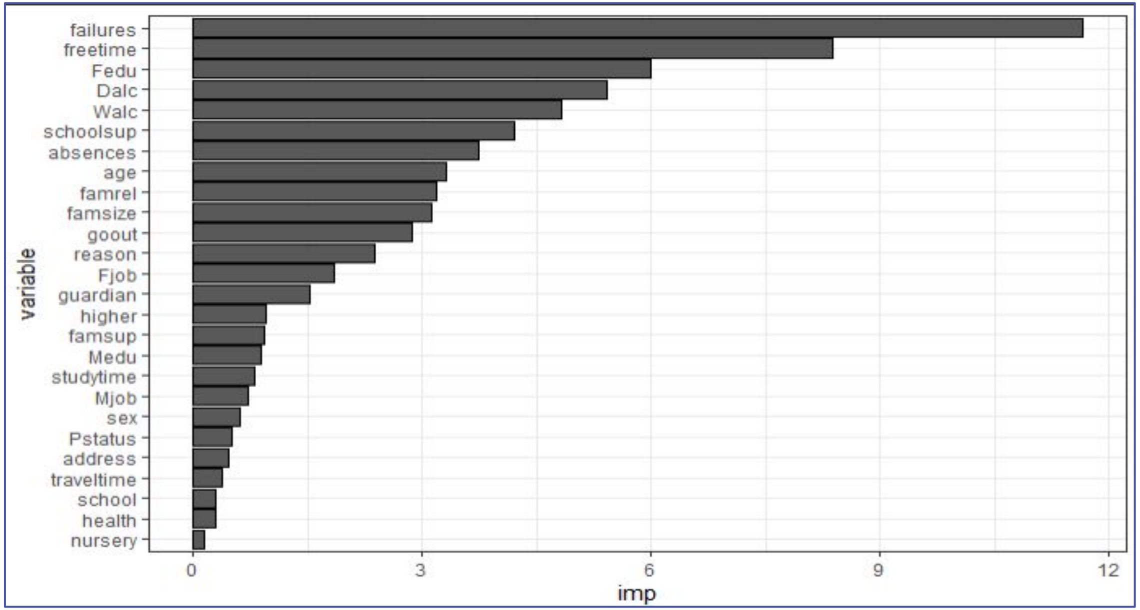
- There is a very bad ROI on the Paid help classes, with 0 improvement seen between those who do and don't pay for the support
- The schools need to optimize this program to actually benefit the students whose families pay for it

Initial Decision Tree Result

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 22 21
         1 24 52

               Accuracy : 0.6218
                 95% CI : (0.5284, 0.7091)
    No Information Rate : 0.6134
    P-Value [Acc > NIR] : 0.4653
```

The original unpruned tree was a messy visual and only barely beat the NIR (assuming every student failed)

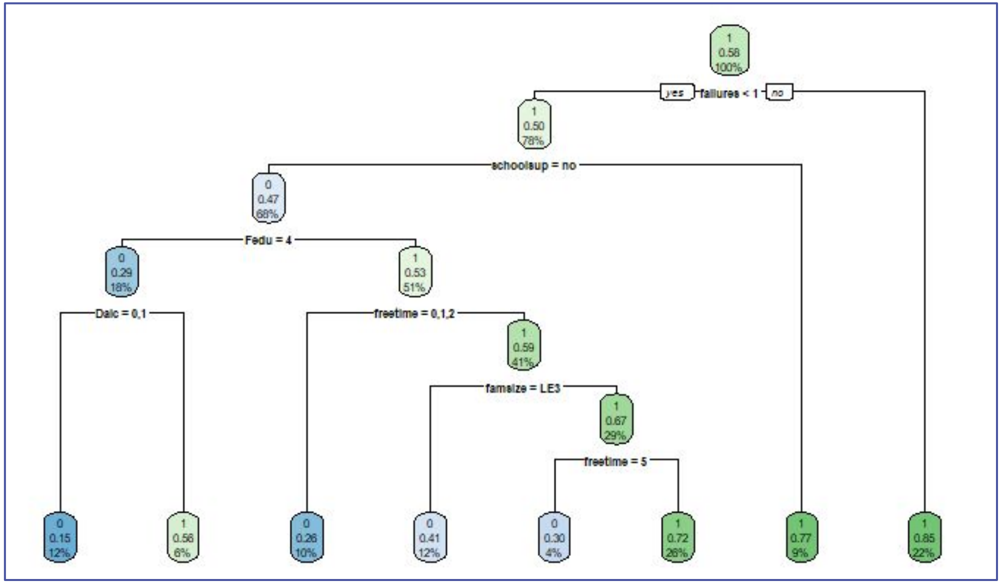
Check Feature Importance



To use:
Failures, freetime,
Fedu, Dalc, Walc,
Schoolsup,
Absences, Age,
Famrel, Famsize

The top variables importance from the unpruned tree shown here, we will return the model without the non-important variables with pruning

Final Decision Tree Result




```
          Reference
Prediction   0   1
         0  74  28
         1  42 132

               Accuracy : 0.7464
                 95% CI : (0.6908, 0.7966)
    No Information Rate : 0.5797
    P-Value [Acc > NIR] : 5.695e-09
```

After only using the top 10 variables from the last model, The model accuracy jumped by over 10% proving additional benefits from those variables

Logistic Regression

```
Call:
glm(formula = Target ~ ., family = binomial, data = test_student_1)
```

Most Significant / Best Predictors: Failures, schoolsupes, Freetime.C, Freetime.4

Accuracy: 71.74%

```
Call:
glm(formula = Target ~ Fedu + failures + schoolsup + freetime +
    Walc, family = binomial, data = train_student_1)
```

Most Significant / Best Predictors: failures, schoolsupyes

Accuracy: 69.93%

```
Call:
glm(formula = Target ~ failures + schoolsup, family = binomial,
    data = test_student_1)
```

Most Significant / Best Predictors: failures, schoolsupyes

Accuracy: 62.68%

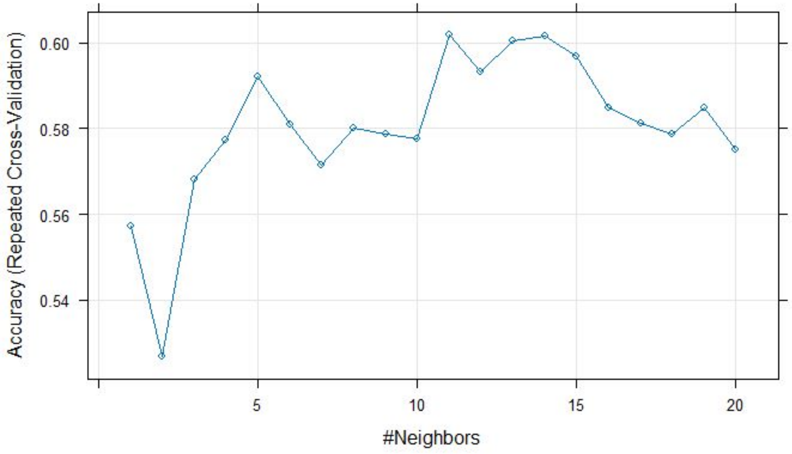
Logistic Regression

Most accurate model:

- Top 10 variables used
- Accuracy: 71.74%

```
Call:
glm(formula = Target ~ ., family = binomial, data = train_student_1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)   0.18636  176.56193   0.001 0.999158    
failures      1.37650    0.36597   3.761 0.000169 ***
freetime.L   -1.17747    0.65343  -1.802 0.071550 .  
freetime.Q    0.41363    0.55495   0.745 0.456060    
freetime.C   -1.38893    0.44043  -3.154 0.001613 ** 
freetime^4    0.80258    0.30510   2.631 0.008524 ** 
Fedu.L       -9.84347  558.29609  -0.018 0.985933    
Fedu.Q        7.42671  471.84633   0.016 0.987442    
Fedu.C       -4.87438  279.14817  -0.017 0.986068    
Fedu^4        1.67696  105.50843   0.016 0.987319    
Dalc.L       -0.59442    0.91504  -0.650 0.515943    
Dalc.Q       -0.76375    0.70588  -1.082 0.279256    
Dalc.C       -0.19206    0.72679  -0.264 0.791580    
Dalc^4       -0.12978    0.70220  -0.185 0.853367    
Walc.L        0.80760    0.73237   1.103 0.270153    
Walc.Q        0.18474    0.55029   0.336 0.737091    
Walc.C        0.05596    0.44279   0.126 0.899430    
Walc^4       -0.57787    0.37159  -1.555 0.119911    
schoolsupyes  1.80990    0.54057   3.348 0.000813 ***
absences      0.03500    0.02817   1.242 0.214121    
age           0.14541    0.13151   1.106 0.268865    
famrel.L      0.89688    0.69419   1.292 0.196361    
famrel.Q     -0.03045    0.60494  -0.050 0.959851    
famrel.C     -0.77025    0.59828  -1.287 0.197940    
famrel^4      0.61239    0.48418   1.265 0.205942    
famsizeLE3   -0.63690    0.33690  -1.890 0.058699 .  
```

kNN



- Accuracy was used to select the optimal model using the largest value
- The final value used for the model was k = 11
- Top 10 variables used

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  69  39
         1  47 121

               Accuracy : 0.6884
                 95% CI : (0.6301, 0.7426)
    No Information Rate : 0.5797
    P-Value [Acc > NIR] : 0.0001303

                  Kappa : 0.3544

 Mcnemar's Test P-Value : 0.4503513

            Sensitivity : 0.5948
            Specificity : 0.7562
         Pos Pred Value : 0.6389
         Neg Pred Value : 0.7202
             Prevalence : 0.4203
         Detection Rate : 0.2500
   Detection Prevalence : 0.3913
      Balanced Accuracy : 0.6755

       'Positive' Class : 0
```

Association Rule Mining

	lhs <chr>	<chr>	rhs <chr>
[1]	{age=[15,16), studytime=[2,4], schoolsup=yes}	=>	{Target=1}
[2]	{Medu=3, studytime=[2,4], schoolsup=yes}	=>	{Target=1}
[3]	{reason=home, nursery=yes, Walc=4}	=>	{Target=1}
[4]	{guardian=mother, goout=5, absences=[0,2)}	=>	{Target=1}
[5]	{schoolsup=no, goout=5, absences=[0,2)}	=>	{Target=1}
[6]	{Medu=1, Fedu=1, Mjob=other}	=>	{Target=1}
[7]	{sex=F, internet=no, absences=[2,6)}	=>	{Target=1}
[8]	{address=R, famsize=GT3, Walc=3}	=>	{Target=1}
[9]	{Fedu=1, romantic=yes, freetime=3}	=>	{Target=1}
[10]	{address=R, famsize=GT3, goout=4}	=>	{Target=1}

- Ten strongest rules according to lift for target audience
- Entire dataset used

SVM

```
Confusion Matrix and Statistics

polymodel1Pred   0   1
             0  89  23
             1  27 137

               Accuracy : 0.8188
                 95% CI : (0.7682, 0.8624)
    No Information Rate : 0.5797
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6265

 Mcnemar's Test P-Value : 0.6714

            Sensitivity : 0.7672
            Specificity : 0.8562
         Pos Pred Value : 0.7946
         Neg Pred Value : 0.8354
             Prevalence : 0.4203
         Detection Rate : 0.3225
   Detection Prevalence : 0.4058
      Balanced Accuracy : 0.8117
```

- Radial kernel
- cost parameter = 0.95
- The accuracy of the model is 81.88%
- This is a black box method which doesn't tell us the feature importance from the model, but given the accuracy, we can assume that the top 10 variables do play a large role in predicting our target variable

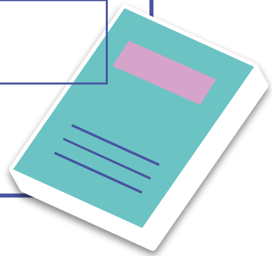
Random Forest

```
Confusion Matrix and Statistics

rfmodel3Pred   0   1
           0 115   1
           1   1 159

               Accuracy : 0.9928
                 95% CI : (0.9741, 0.9991)
    No Information Rate : 0.5797
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9851

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9914
            Specificity : 0.9938
         Pos Pred Value : 0.9914
         Neg Pred Value : 0.9938
             Prevalence : 0.4203
         Detection Rate : 0.4167
   Detection Prevalence : 0.4203
      Balanced Accuracy : 0.9926
```

- ntree parameter = 500 (Default)
- mtry parameter = 4
- The accuracy of the model is 99.28%
- Just using the 10 top variables
- Might be overfit based on the extreme accuracy level

Model Comparison

Model	Accuracy
No Information Rate	.5897
Decision Tree	.7464
Logistic Regression	.7174
kNN	.6884
SVM	.8188
Random Forest	.9928



Deployment

- Given the high accuracy levels, we recommend the Random Forest model after assuring that no overfitting has taken place
- This would be distributed to the schools for their incoming class to predict who might need assistance in Math
- Finally, with these top 10 variables identified, the schools survey to students can be drastically reduced which will promote better completion rates and accuracy