Midterm project: Student Performance in Math Class

Bin Yang

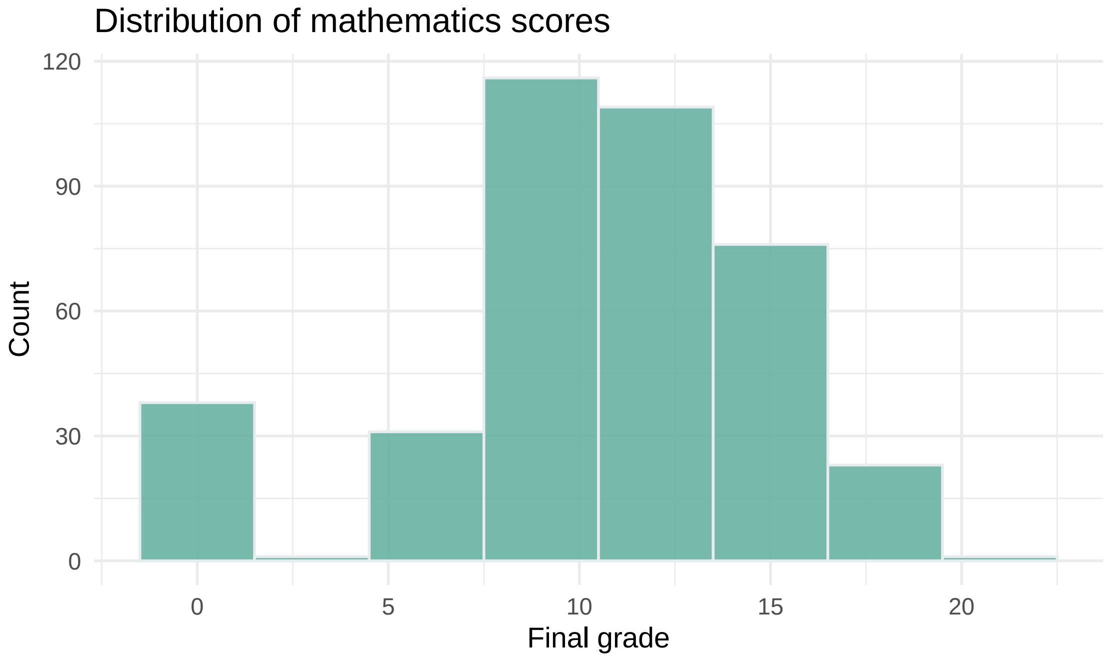
Introduction

Student performance can be determined by various factors. We are using the student performance data set to perform various machine learning techniques to better identify the association between different variables and grades. The student performance data set contains information regarding the student achievement in secondary education of two Portuguess schools. The data set includes student grade, demographic, social and school related features.

The data set contains 395 observations and 33 variables in total. The data set is complete with 0 NA value. Our interested outcome variables are first period grade (G1), second period grade (G2), and final grade (G3). There are 16 numeric variables and 17 character variables. To utilize machine learning techniques to analyze the data set, we convert the character variables to factors and randomly partition the data set into a test and a training set.

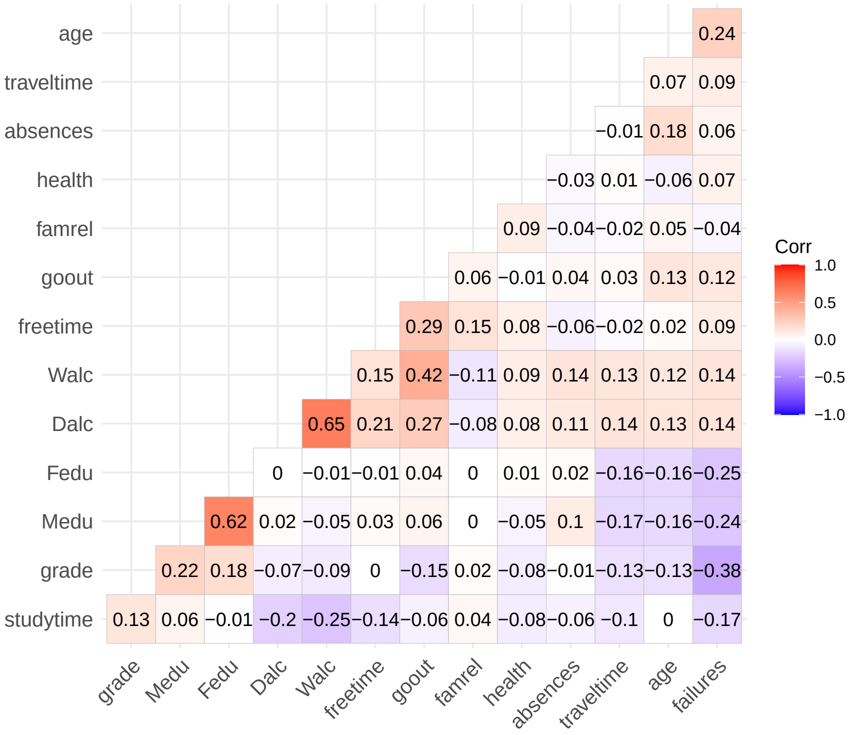
Exploratory analysis/visualization

We first purpose to use final grade (G3) as our only outcome variable and thus we use a bar plot to see the distribution of G3. The result shows that G3 is approximately normally distributed, however there are unusual amounts of 0 grade. With further inspection, we see that there are 38 0 grades, which is about 9.6% of the total sample. The proportion is relatively large considering our small sample size. Since we do not have further information regarding the 0 grades, we will keep these records in our analysis.



We then create a correlation plot to observe the correlation between variables. As we can see from the correlation plot, first period grade (G1), second period grade (G2) and final grade (G3) are highly correlated. Other numeric variables don't seem to have a strong correlation with the outcome. There are some other notable correlations between variables: Weekend alcohol consumption (Walc), workday alcohol consumption (Dalc), going out with friends (goout) are highly correlated; mother's education (Medu) and father's education (Fedu) are also highly correlated.

Considering the high correlation between first period grade (G1), second period grade (G2) and final grade (G3), and relatively low correlation between final grade(G3) and other predictors, we will combine first period grade (G1), second period grade (G2), final grade (G3) and take average for a better interpretation and capture the relationship between other variables and grade.



After combining G1, G2, and G3 and taking average, we further create a trellis plot to see the relationship between our outcome variable grade and other predictors. With close observation, we see that absences, age and grade seem to have a nonlinear relationship. We also notice that students who want to take higher education are more likely to obtain higher grades. Number of past failures have a negative relationship with grades, as the number of past failures increase, students tend to have lower grades.

Models

In order to identify the association between variables and grade and further make predictions, we are using all the predictors in our data set and utilizing various machine learning techniques including ridge regression, lasso regression, elastic net, principal component regression, partial least squares regression, generalized additive model, multivariate adaptive regression splines, regression trees and random forest. Considering the relatively small sample size, large number of predictors and also the multicollinearity and nonlinear relationship existing in the data set, we will not use a multiple linear regression but use the above methods that can better work with our data set. This section introduces the models we are using, compares the models, evaluates prediction accuracy as well as variable importance. The tuning parameters are selected using repeated 10 fold cross validation.

Shrinkage methods

As we have mentioned previously, our data set has some multicollinearity issues. We also have a relatively large number of predictors compared to sample size. Therefore we first consider regularized models including ridge regression, Lasso and elastic net. To perform these models, we assume linearity, constant variance, and independence. We also need to standardize our predictors by subtracting to their means and dividing by their standard deviation. We will use repeated 10 fold cross validation to select the tuning parameter.

As a result, the best tuning parameters obtained are $\lambda = 1.62$ for ridge regression, $\lambda = 0.141$ for Lasso, and $\alpha = 1, \lambda = 0.141$ for elastic net. Although the shrinkage methods can deal with the multicollinearity issue as well as the relative large number of predicors, they might not capture the non linear relationship in the data set.

Dimension Reduction Methods

Another method to deal with multicollinearity as well as a large number of predictors is the dimension reduction method including principal components regression and partial least squares. By using PCR, we assume that the directions in which the predictors show the most variation are the directions that are associated with outcome. We use repeated 10 fold cross validation to select the tuning parameter m, the number of partial least squares directions used.

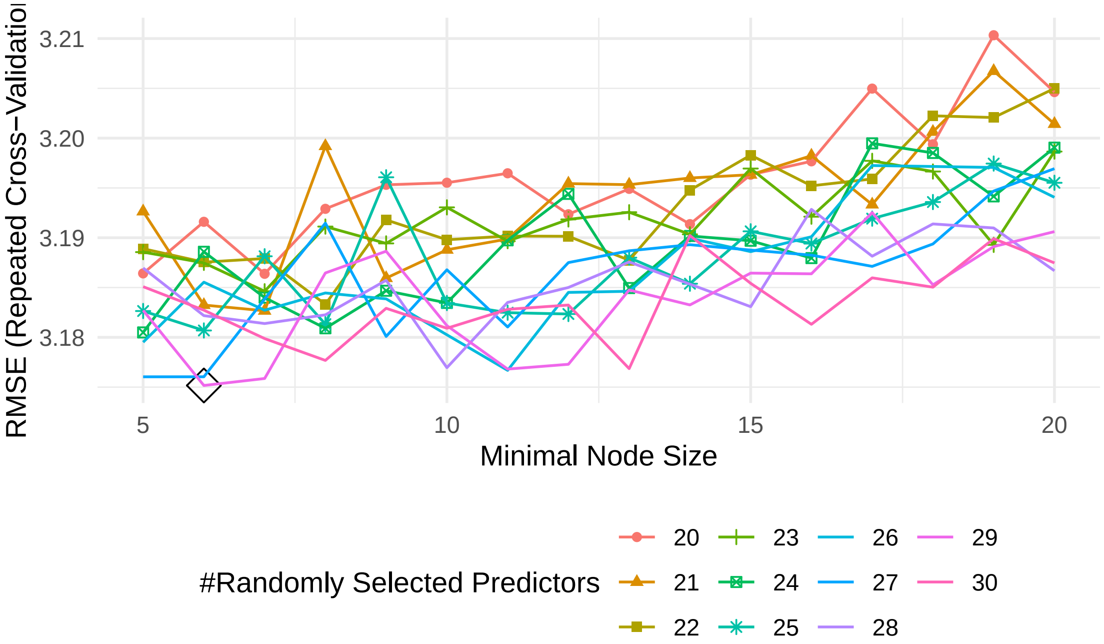
As a result, the best tuning parameters obtained are $m = 29$ for PCR model and $m = 3$ for PLS model. Similar to the shrinkage methods, the dimension reduction methods can not capture the non linear relationship in the data set.

Non linear models

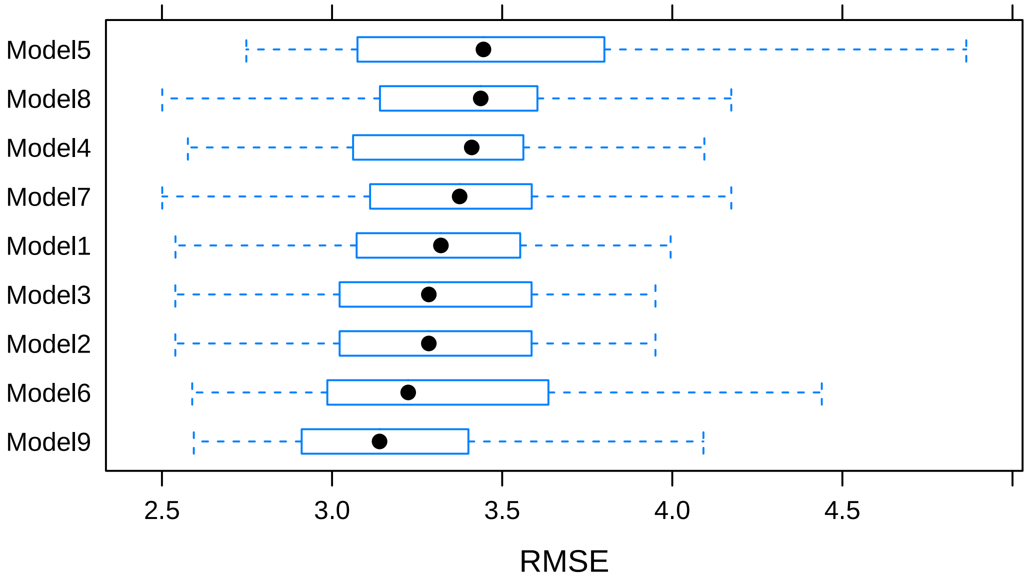
As mentioned previously, absences, age and grade seem to have a nonlinear relationship. Therefore we will also consider non linear models including generalized additive model and multivariate adaptive regression splines. We use repeated 10 fold cross validation to select the tuning parameters for the MARS model, which are the degree of features and the number of predictors, and as a result, we obtain 1 and 13 as best tuning parameters.

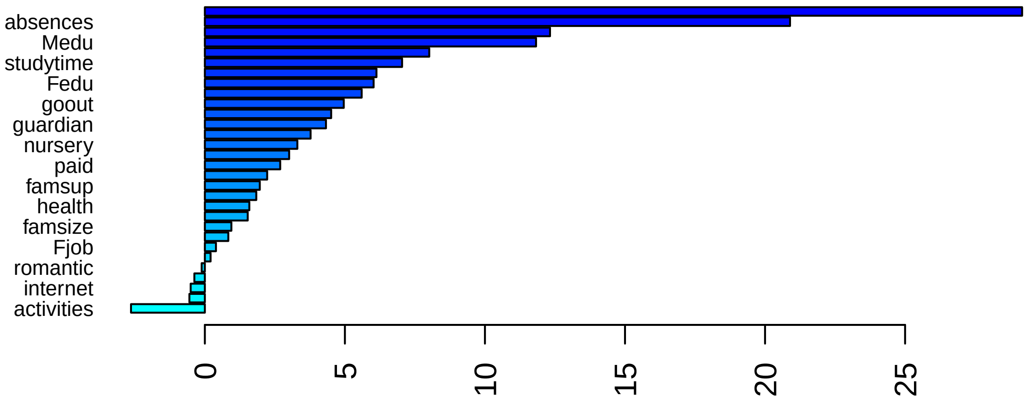
Tree based methods

We also use tree based methods for prediction, including a single regression tree, random forest. Regression tree is simple and useful for interpretations, but may not perform well in terms of prediction accuracy. Additionally, trees can be very non-robust. In other words, a small change in the data can cause a large change in the final estimated tree. By using random forest, the predictive performance of trees can be substantially improved and the variance can be reduced. We use repeated 10 fold cross validation to select the tuning parameters. As a result, we obtained complexity parameter of 0.0468 for the CART approach, mincriterion 0.969 for the conditional inference tree, and $m = 29$, minimal node size of 6 for the random forest.



Finally, by comparing the above methods using resampling results, the random forest model has the lowest median and mean RMSE. Therefore random forest is the best performing model. Using random forest on test set, we are able to obtain mean squared error of 11.26. We then inspect the variable importance. As shown in the plot, number of past class failures(failures), number of school absences(absences), extra educational support(schoolsup) are the top three most important variables.





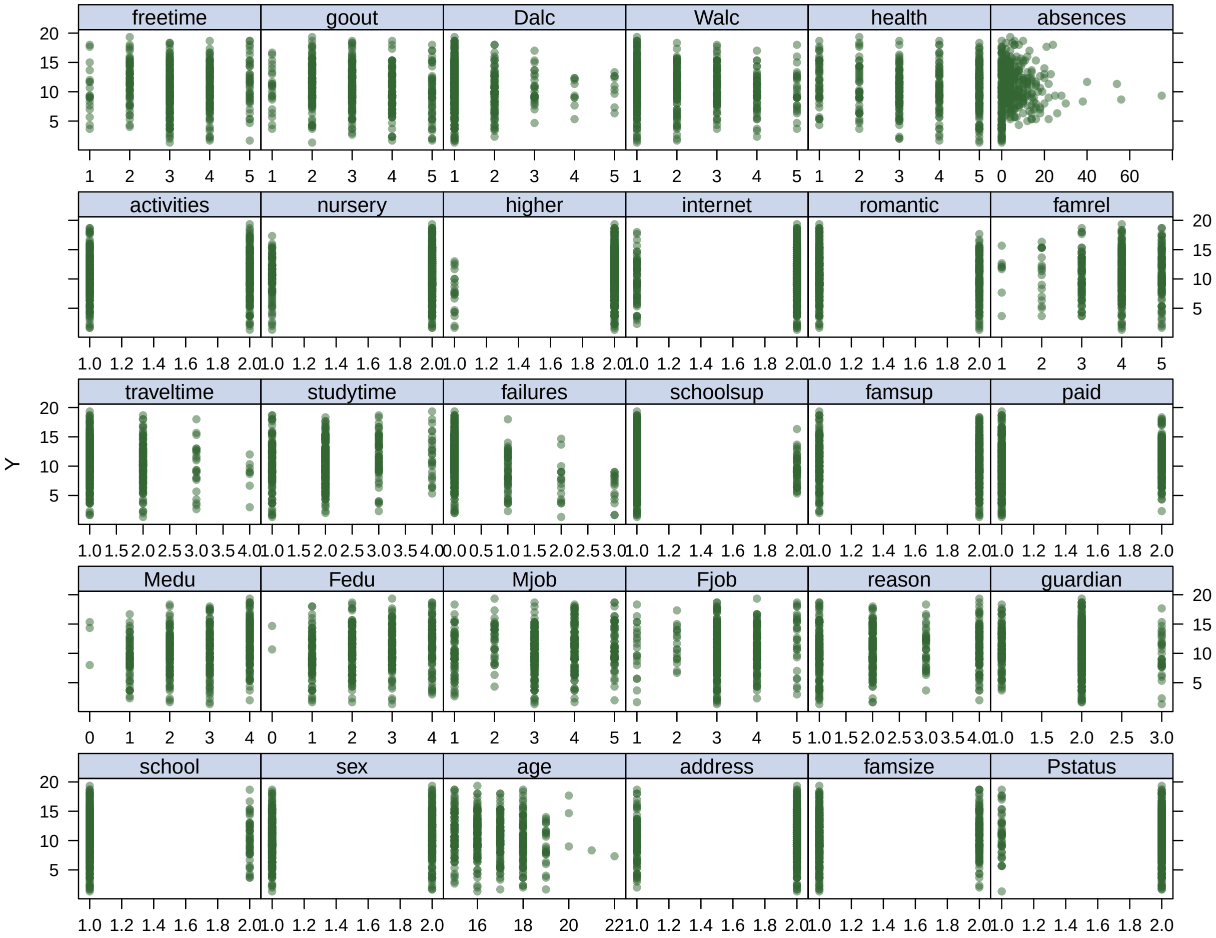
Conclusion

Student performance is a crucial topic in our society. In this project, we used several machine learning techniques to make predictions about students' grades using demographic, social and other school related data. We achieved best prediction performance using the random forest model and we concluded that number of past class failures(failures), number of school absences(absences), extra educational support(schoolsup) were the top 3 important variables in making predictions, which also aligned with our observation in the exploratory data analysis. Therefore, to advocate for a better education outcome, we could offer extra help to students who had more past failures, encourage students' attendance and offer extra educational support.

Our study also has several limitations. Although random forest model achieved great prediction performance, the R Squared for random forest is only 0.267, which means that the model does not explain the variation within the data set well. Moreover, our data set is pretty small, which contains only 395 observations and only collects information of students in a Portuguess school. In order to make meaningful conclusions about education, we will need a larger data set with more information.

appendix

Trelis plot for data set:


freetime
goout
Dalc
Walc
health
absences
activities
nursery
higher
internet
romantic
famrel
traveltime
studytime
failures
schoolsup
famsup
paid
Medu
Fedu
Mjob
Fjob
reason
guardian
school
sex
age
address
famsize
Pstatus
Y
20
15
10
5