Final Report: Student Performance in Math Class

Elaine Xu, Rio Yan, Bin Yang (Group 8)

Contents

Introduction

Our final project dataset comes from the University of California Irvine's machine learning repository. It contains information on student achievement in secondary education of two Portuguese schools. The dataset has 33 variables in total, including attributions like student grades on the mathematics subject, demographic, social, and school-related features. Grades play significant roles in students' academic life and as students ourselves, we want to perform various machine learning techniques to better identify what contributes to good grades in Portuguese students and how well the model performs. Therefore, we want to investigate potential factors that are associated with students in these two Portuguese schools' performances on the course subject, math.

The summary table output from the skim function shows that the dataset contains 395 rows and 33 columns, with 17 nominal variables and 16 numeric variables. The dataset is very clean with no missing or unreasonable values. We plan to use the math grade as the outcome variable and the data exists three grades, G1, G2, G3, each representing first, second, and final period of grade. According to the source website for the data, G1 and G2 are highly correlated, so there might be some multicollinearity problems. We examined the three grade variables in the skim table and found that their mean and standard deviation are similar, and the histograms also seem to be normally distributed. For simplicity, we took the mean of the three periods of grade to get only one continuous outcome variable, grade, in representing the overall grade performance for students. We also created a new binary variable called letter grade through stratification. If the grade is below 10, we marked it as "fail", and if the grade is equal to or above 10, we marked it as "pass".

Exploratory analysis/visualization

The summary table from the skim function outputs both categorical and continuous variables. From the categorical variables' summary table, we can see that there are more students from the GP school, more females, more students live in the urban area, more students have a family size bigger than 3, most of the students' parents live together, most of the students' parents have a job and work in civil services, most students' guardian is mother, most students have extra educational support, attended nursery school, want to take higher education, and have internet access at home. From the continuous variables summary table, we can see that the average student age is 16.7 years old, mothers' average education level is higher than father's, average home to school travel time is around 15 minutes, average weekly study time is around 2 to 5 hours, the average number of past class failures is less than 1, the average quality of the family relationship is high, average free time after school and going out with friends times are average, weekend alcohol consumption is higher than workday alcohol consumption, current health status is good, average school absences is around 5 times, and average math grade is around 10.68 and normally distributed. We then used the trellis boxplot to explore the relationship between math grade and other continuous variables. Most of the variables appear to have similar distributions, with weekly study time(studytime), number of past failures(failures), age, mother's education(medu), and father's education(fedu) appear to have stronger associations. We also used the trellis density plot to get a better visualization of how the data is distributed.

Before proceeding, it is important to assess crude correlation among relevant variables, in case issues of multicollinearity arise during model development. We used model.matrix to plot out the correlation graph by only including the continuous variables. However, none of the variables are highly correlated ($r < |0.70|$) with each other. This might be due to categorical data exclusion. If only focusing on the correlation values between each predictor and the outcome of interest, we might predict that number of school absences (absences), current health status (health), weekend alcohol consumption (walc), workday alcohol consumption (dalc), going out with friends (goout), quality of family relationships (famrel), number of past class failures (failures), home to school travel time (traveltime), and age have negative correlations with being able to obtain a passing grade. We might also predict that weekly study time (studytime), father's education (fedu), and mother's education (medu) have positive correlations with a passing grade.

Models

In order to perform the classification on the letter grade, we employed the following models including all predictors in the data set:

Linear methods: glm, penalized logistic regression, GAM, MARS

We first considered the linear methods for classification, including logistic regression, penalized logistic regression, generalized additive model, and multivariate adaptive regression splines. The logistic regression makes the assumptions that the observations are independent, the independent variables should not be highly correlated, linearity of independent variables, and log odds. Since our data set contains a relatively large number of predictors, we also performed penalized logistic regression and applied a penalization coefficient λ and mixing proportion coefficient α to control for the number of predictors. Using repeated cross-validations, we obtained the best tuning parameter of $\alpha = 0.4$ and $\lambda = 0.0987$ respectively. Due to the nonlinearity existing in the data set, we also considered using GAM and MARS models. For the MARS model, the best tuning parameters, the degree of interactions of 1 and the number of retained terms 9 were also selected using repeated cross-validations.

LDA/QDA/NB

In the beginning, we wanted to use discriminant analysis because we split the outcome into more than two categories. However, after we adjusted the outcome to binary, we decided to remain the discriminant analysis and test their fitness for this set of data. For discriminant analysis models, we selected linear discriminant analysis, quadratic discriminant analysis, and naive bayes models. Because of the size of our data and the number of predictors, we were not expecting discriminant models to outperform logistic regression. For the naive bayes model, we obtained best tuning parameters, $fl = 10$ for the "Laplace Correction" and $adjusts = 4$ for the adjust parameter controlling the bandwiths of the kernal density estimates. After we compared the results of all the models, LDA offers a better fit for the data among these three models and outperforms logistic regression. This illustrates that the Bayes decision boundary is more likely to be linear instead of nonlinear. The mean accuracy of QDA is only 64.08%, and for LDA is 67.45%.

Tree-based methods

We also employed tree-based methods including single classification tree, random forest, and boosting. Tree-based models allow for the learning of non-linear decision boundaries but are based on little theoretical basis and employ greedy search with no clear optimization formula. A single classification 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. Similarly, using boosting methods can also reduce the variances and improve the prediction accuracy.

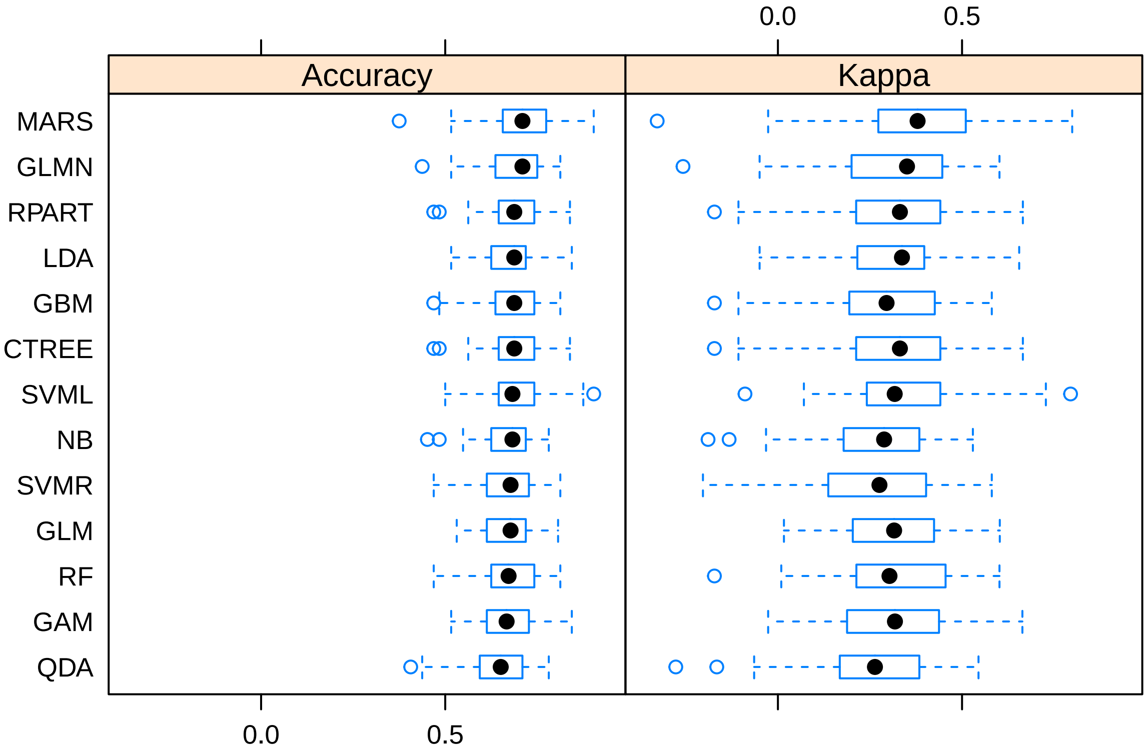
Consistent with the model tuning methods used prior,repeated cross-validation was used to tune for the best tuning parameters. For the classification tree using the CART approach, the complexity parameter $cp = 0.0574$ was selected. For the conditional inference tree model, the $mincriterion = 0.977$ which controlled for the splits was selected. For the random forest model, the number of variables selected at split $mtry = 20$, and the minimal node size $min.node.size = 4$ were selected. For the ada boost model, the number of trees $B = 2000$, the interaction depth $d = 2$, shrinkage parameter $\lambda = 0.001$ and minimal node size $n.minobsinnode = 1$ were selected.

Support Vector Machines

We further applied the support vector classifier model and the support vector machine with a radial kernel to the training data. SVMs are efficient learning algorithms for nonlinear functions and are equipped with computationally friendly quadratic optimization. Since our data is noisy with no clear decision boundary, the support vector classifier could maximize a soft margin that allows for some misclassification by applying a regularization coefficient, C. Since the decision boundary can hardly be linear, we also used the support vector machine with a radial kernel by expanding the original feature space. Using repeated cross-validations, we were able to obtain the best tuning parameters for the above models. For the radial kernel model, we employed the tuning method to tune over both C and σ by looking at which sigma on which cost curve has the highest accuracy. As a result, we obtained best tuning parameter $C = 0.0888$ for the support vector classifier and $C = 0.00483, \sigma = 1.249$ for the support vector machine model.

Model comparison

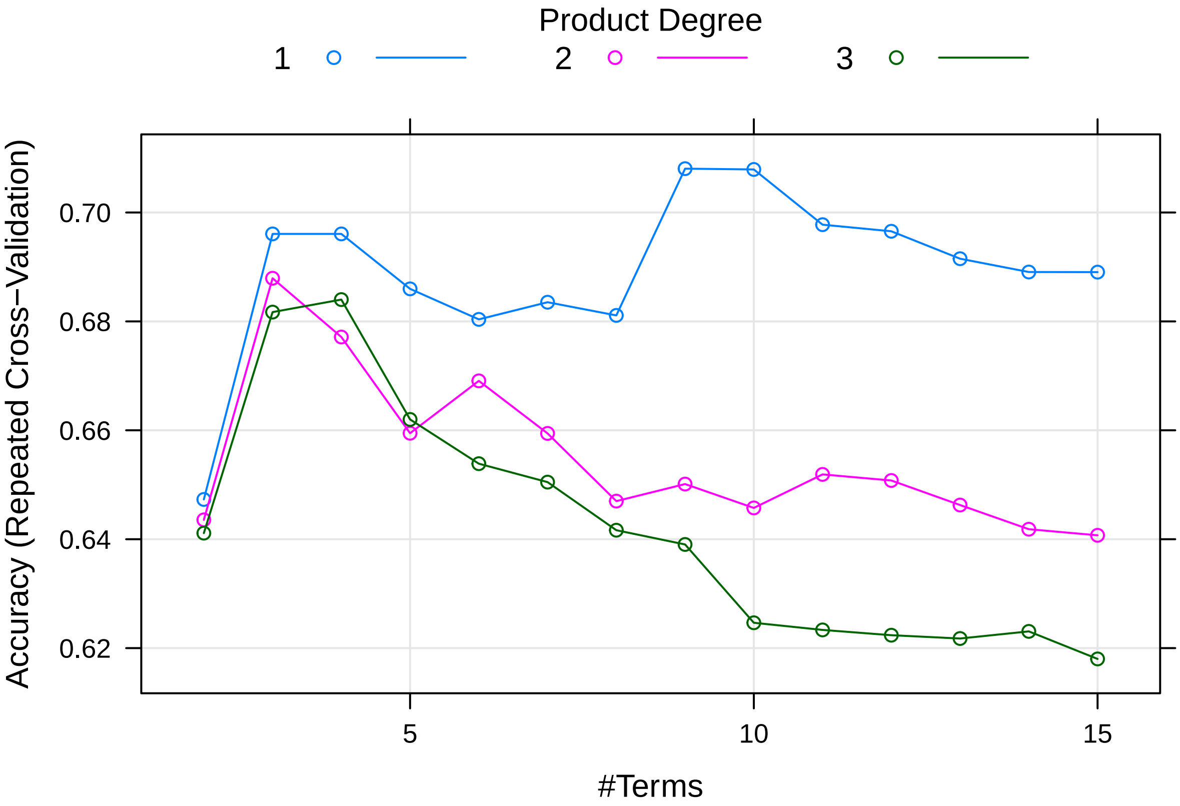
We compared the above models employed using the resampling accuracy and kappa as demonstrated by the below boxplot. We concluded that the MARS model was the best model with the best mean prediction accuracy.



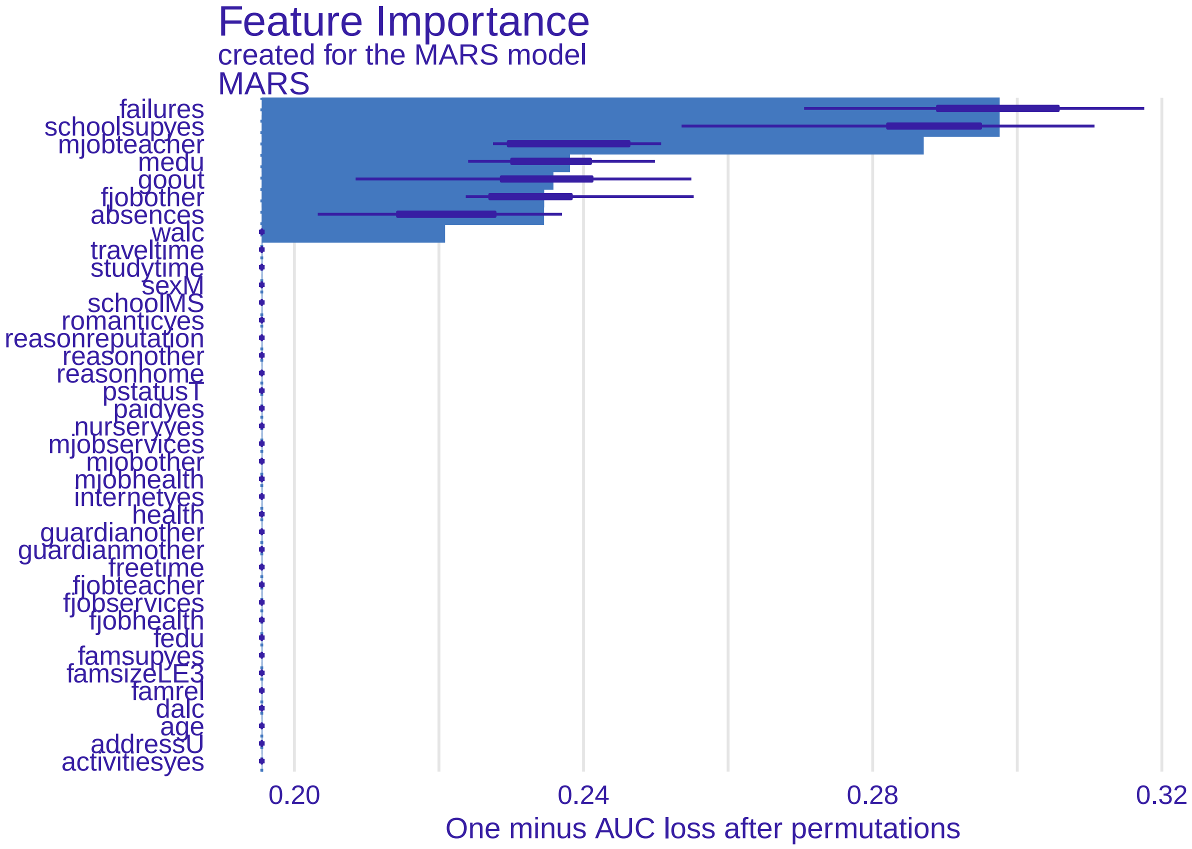
Since MARS was the best model with the best prediction accuracy demonstrated by the resampling comparison, we examined its test data set performance. We obtained a test accuracy of 62.82%.

As mentioned before,the best tuning parameters, the degree of interactions of 1 and the number of retained terms 9 were selected using repeated cross-validations. The final model obtained was:

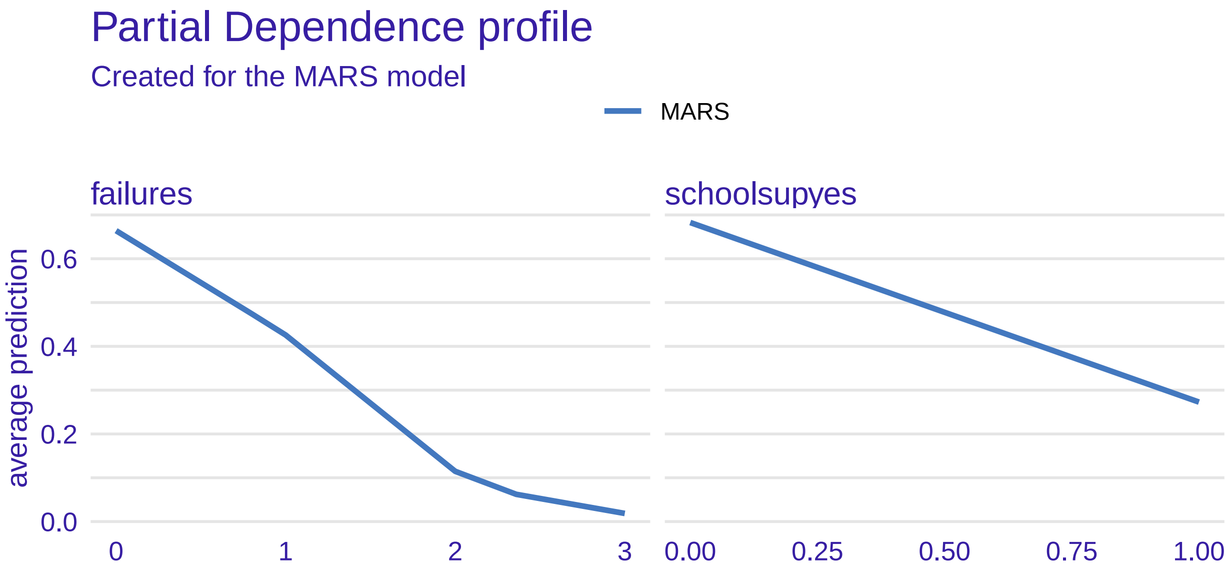
$$y = 0.713 - 2.078 \times h(failures - 1) + 1.202 \times h(1 - failures) - 2.135 \times schoolsupyes + \\ 0.549 \times h(4 - goout) - 0.423 \times h(2 - absences) - 0.699 \times h(3 - medu) - 1.443 \times mjobteacher \\ - 0.952 \times fjobother$$



We further examined variable importance, and we can see that “failures”, “schoolsupyes” and “goout” were the top three important features in making the prediction.



We also made partial dependence profile for “failures” and “schoolsupyes”. As demonstrated by the plots, the average probability of “pass” decreases as failures increases and for students without extra educational school support.



Conclusions