The Influence of Student Background on Final Grades

Aleksandra Budkina Rachel Han Sharon Yang Oliver Zhang

1 Introduction

Our group was interested in predicting the final grade of students given student drinking habits, and other life and environmental factors. To do so, we used the "Student Alcohol Consumption: Social, gender and study data from secondary school students" dataset from Kaggle:[1]. About 1000 students from a Math and a Portuguese class in a Portuguese secondary school were surveyed about the students' drinking habits, as well as their personal information, family and eco-social background.

2 Description of data, explanatory, and response variables

Results from each class were compiled into two datasets - one per subject. In total, 1044 students provided responses to all 32 questions (described in [1]). Moving forward, all explanatory variable names will be written in single quotes. We wish to predict 'G3', the student's final grade for a given subject, scored out of 20. Most explanatory variables are binary or categorical, as many of the questions asked to the students were yes/no or ranking questions.

In order to discover new relationships to final grades, we decided to exclude explanatory variables 'G1' and 'G2' (the first and second period grades respectively), since they directly affect the response variable 'G3'. After preliminary analysis of the variables via scatter plots, we also excluded the variables 'Mjob' (mother's job), 'Fjob' (father's job) and 'reason', as they had little influence on the response variable.

3 Initial hypothesis

Conway and DiPlacido J. [2], and Kelly et al. [3] cited alcohol consumption and the number of skipped classes as significant factors in predicting final grades of students; hence we expected similar results in our analysis. We also expected the amount of time spent going out with friends, the amount of study time, and participation in extra-curricular activities to be important explanatory variables. Family related variables (like family educational support, or quality of family relationships) are expected to be the least significant [3], since the mean student age is 16.7, suggesting some independence from family.

4 Methods

We compared the prediction power of various models and methods on our data by computing the MSPE from 50 runs of 5-fold cross validation on the dataset with 28 explanatory variables (labeled as Data1). We ran the following tree-based models: a pruned overfitted tree, bagging with 201 trees, and Random Forest (RF) with 201 trees. We also ran the following linear models: least squares, Ridge Regression (RR), LASSO and Elastic Net (EN) with $\alpha = 0.75$. After running all the models, we compared the MSPE, or the out-of-bag error rate (as an approximation of the MSPE for RF and bagging) between the tree-based and linear models using boxplots, and picked the best one based on lowest MSPE. We repeated the same steps for a reduced model, labeled as Data2. Details and motivations are addressed in the Results and Analysis section.

5 Results and Analysis

Comparing the MSPE of the models presented, the RF method (RFor) has the best predictive ability (Figure 1).

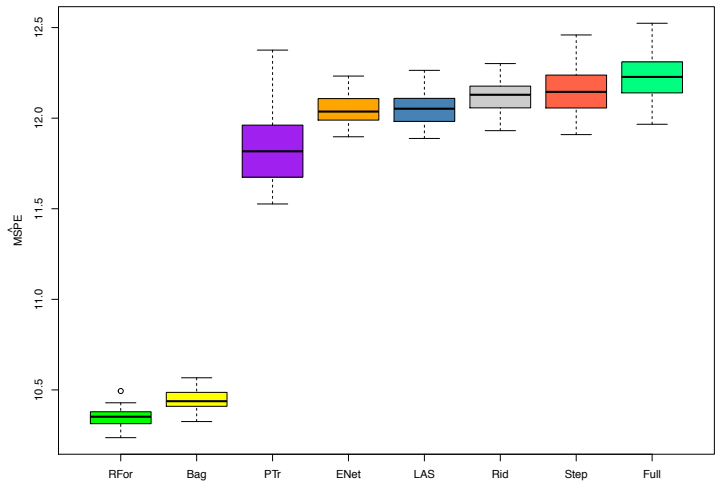

Figure 1: Boxplots of MSPE for various models ran on Data1.

Model	$\lvert\beta_{\text{`failures'}}\rvert$	$\lvert\beta_{\text{`higher'}}\rvert$
RR	1.3299	1.3108
$LASSO$	1.6948	1.3286
$EN_{\alpha=0.75}$	1.6838	1.3286

Table 1: Magnitude of the coefficient for 'failures', compared to the next significant coefficient 'higher'.

Figure 2: IncNodePurity for random forest (tmp.rf) and bagging (tmp.bag); 10 most significant predictors.

Linear models performed the worst, suggesting that the explanatory variables measures do not have a linear relationship with the final grade. For all linear models, 'failures', the number of previously failed classes, was influential in predicting 'G3', followed by 'higher', the intent to pursue higher education (Table 1).
Figure 2 describes the Inc Node Purity distributions for RF and bagging trees for the 10 most influential predictors. Again, 'failures' is an influential variable; 'absences', the number of absences from school, is the next significant variable.
To discover other relationships between student responses to final grades, we excluded the two most influential variables 'failures' and 'absences'. In doing so, we assumed that the other explanatory variables may directly contribute to the number of previously failed classes and the number of absences. Treating 'failures' and 'absences' as confounding variables, we retrained a new set of models on a reduced dataset Data2.

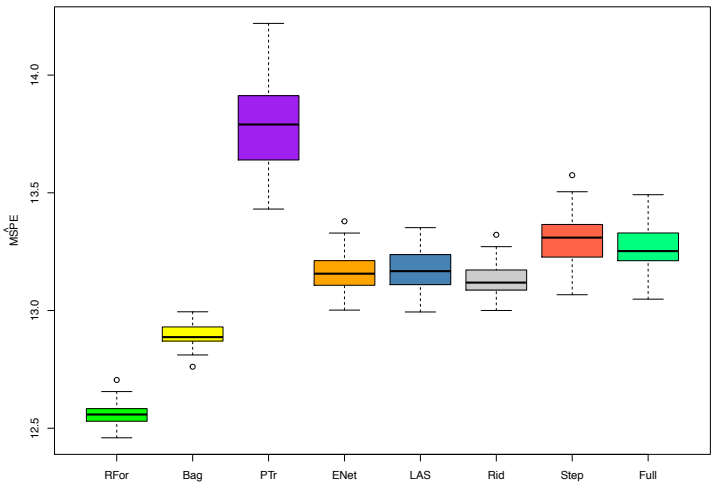

Figure 3: Boxplots of MSPE for various models ran on Data2.

Linear prediction models perform better after eliminating the potential confounding variables, and they also outperform pruned trees (Figure 3). Linear models describe the intent to pursue higher education ('higher') as the most powerful predictor, followed by 'schoolsup', describing whether or not the student receives extra educational support (Table 2). The pruned tree model also chooses the variable 'higher' as the first predictor to split on.

Model	$\lvert\beta_{\text{'higher'}}\rvert$	$\lvert\beta_{\text{'schoolsup'}}\rvert$
RR	1.7580	1.1071
$LASSO$	2.1289	1.1991
$EN_{\alpha=0.75}$	2.1303	1.2354

Table 2: Magnitude of coefficient picked for 'higher' in comparison to the next significant coefficient 'schoolsup'.

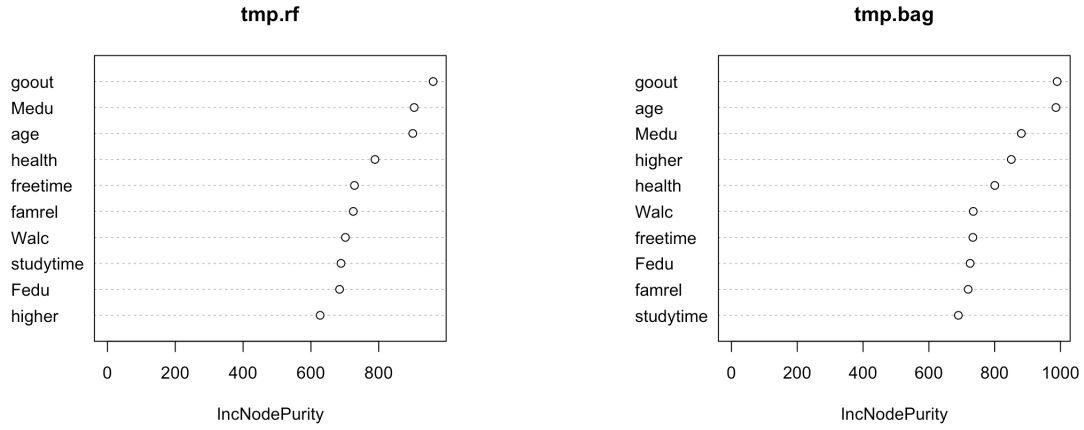

Figure 4: IncNodePurity for random forest (tmp.rf) and bagging (tmp.bag); 10 most significant predictors.

On the other hand, both RF and bagging indicate 'goout', the time spent going out with friends, as the most important predictor (Figure 4). Since the random forest model yielded the lowest estimated MSPE (Figure 3), we focus our remaining discussion on the results of the RF model only. This model finds 'goout', 'Medu', 'age', and 'health' to be the most influential in predicting final grades.

As expected, since the variable importance levels for the explanatory variables are high (Figures 2 and 4), removing two most influential predictors affected the predicting ability (and thus increased the MSPE) for the tree-based models. Figure 3 shows a larger increase in the MSPE for tree models from Data1 to Data2 compared to the linear models with 'failures' and 'absences' removed; in fact the pruned tree model's performance is the worst.

In both Data1 and Data2, the MSPE for all models is above 10. Since 'G3' is scored out of 20, this implies that the average error in prediction for any model can be estimated to be at least $\pm\sqrt{10}/20 \approx \pm 15\%$, which is very high variance for predicting final grades.

6 Discussion

Kelly et al. [3] previously found that "peer drinking networks are interconnected with educational outcomes." From our data, we can extend this notion to general peer networks, as the time spent with peers outside of school has the strongest positive prediction ability, if we ignore the number of absences or previously failed classes. For teenagers, the influence and opinion of their peers may be more important than other social or behavioural factors. An unexpected predictor 'Medu', indicating maternal level of education, was found. We hypothesize that mothers with higher education may be more involved in the education of their children, and provide a higher quality of education.
'age' is also significant in the RF model. Plotting a boxplot against the response variable 'G3' shows a significant drop in the performance of students over 20 years old.
We speculate that there is a somewhat negative correlation between 'health', the level of health, and the final grade. Perhaps less healthy students compensate for their lack of physical performance with increased effort to achieve better grades.
Surprisingly, 'Dalc' (daily alcohol consumption) and 'Walc' (weekly alcohol consumption) have smaller predicting influence for the given data, contrary to what was found by Conway and DiPlacido [2].
The main challenge we faced was the non-trivial relationship between the given explanatory variables and the response variable. Teenager psychology can also be quite complicated and studying academic performance may depend on other factors not taken into account in this study. Finally, the responses received from students may not accurately reflect the students' actual behaviours.
The overall high MSPE, even when including trivial variables, suggests a more complex relationship between various factors and students performance than the models presented here. However, given our results, it is likely that instructors and family members can help improve student grades by, for instance, increasing student involvement with after school activities, rather than leaving students with extra free time, and working closely with mature students, particularly those over 20 years old.

References

[1] Student Alcohol Consumption: Social, gender and study data from secondary school students — Kaggle. Retrieved October 15, 2017, from https://www.kaggle.com/uciml/student-alcohol-consumption.

[2] Conway J.M., DiPlacido J. "The Indirect Effect of Alcohol Use on GPA in First-Semester College Students." Sage Journals. 16 Mar 2015. DOI: https://doi-org.ezproxy.library.ubc.ca/10.1177/1521025115575705

[3] Kelly, A.B., O'Flaherty M., Toumbourou J.W., Homel R., Patton G.C., White A., Williams J. "The Influence of Families on Early Adolescent School Connectedness: Evidence That This Association Varies with Adolescent Involvement in Peer Drinking Networks." Abnormal Child Psychology. 12 Oct 2011. DOI: 10.1007/s10802-011-9577-4