Steven Brotz, Michael Li, Junno Tseng, Andrew Wang
Professor Colin F. Camerer
BEM/Ec 150
December 8 2017

Enabling Optimal Allocation of Educational Resources Through Early Detection of Students at Risk for Failure and Examination of Alcohol's Effects on Academics

Introduction

Our project was on secondary school students in Portugal. The data analyzed were various attributes of the students' demographic background, family, and habits. Also included were the students' academic performance. We had two main objectives. First was to develop an accurate model to predict a student's academic performance based on his or her attributes. Second was to examine if alcohol consumption was an important factor to academic performance, and if so, to what extent. The motivation for these objectives is to allow schools to identify at-risk students and help them succeed.

Dataset

Our dataset[2] consisted of two comma separated value files that were split based on two subjects, Portuguese and math. The data was collected from two Portuguese high schools through surveys and grade reports and includes various statistics including grades, demographic, social, and school features. Examples of attributes include parents' education, study time, travel time to the school, and whether or not they were in a relationship. There were 395 students surveyed in math and 649 in Portuguese, garnering 30 attributes for each student to predict on their three grade attributes: G1 (first period grade), G2 (second period grade), and G3 (final period grade). Many of the attributes were binary (i.e. sex) and categorical (i.e. alcohol), so we labeled and one hot vector encoded our data prior to testing. Additionally, about 10% (38 math, 65 portuguese) of students had missing grades data (grade of 0), presumably due to a students dropping classes or transferring to other schools. To impute these missing grade values, we implemented a weighted KNN approach. For each missing grade, we took the ten most similar students, in terms of Pearson correlation coefficient, and computed a weighted average of these other closely related students' scores to fill in the missing grade value. Another important thing to note about the dataset is that because some data was gathered by survey, we had a many variables that were qualitative and relative to the students' own perceptions. For example, for the

'Dalc' variable, students ranked their weekday alcohol intake ranging from 1 (very low) to 5 (very high). Of course, one student's 5 could be another's 1. We kept these limitations of our dataset in mind as potential causes for error.

Methods

We primarily used Python and R for this project. We used the numpy and pandas libraries in Python to manipulate our data. We used the matplotlib and seaborn libraries in Python along with the ggplot2 and waffle libraries in R to create data visualizations.

To determine a baseline for acceptable predictor performance, we started with the naive methods of using the mean/median value from the training set for each attribute to predict on the test set. The performance metrics we used to evaluate the models were R^2, mean squared error (MSE), and mean absolute error (MAE). The models we tested include KNN, SVM, linear regression, and random forests regression trees, all from the scikit-learn library. After tuning parameters through grid search paired with 5-fold cross-validation, the random forests model was found to perform best. Specifically for the random forests model, the parameters on which we grid searched were number of trees in the forest (n_estimators), as well as the regularization parameters max depth (max_depth), minimum number of samples required to elicit a split (min_samples_split), and minimum number of samples in a leaf node (min_samples_leaf). We left other parameters to their default values set by scikit-learn. The parameters for random forests model that yielded the best performance differed between the math and Portuguese datasets:

Subject	*n_estimators*	*max_depth*	*min_samples_split*	*min_samples_leaf*
Math	46	3	4	2
Portuguese	43	9	3	2

Results

Four models for predicting student performance were created. Residuals for each are shown in fig. 1 and the error measurements are shown in fig. 2. The feature importances for the top seven features in the tree model are shown in fig. 3.

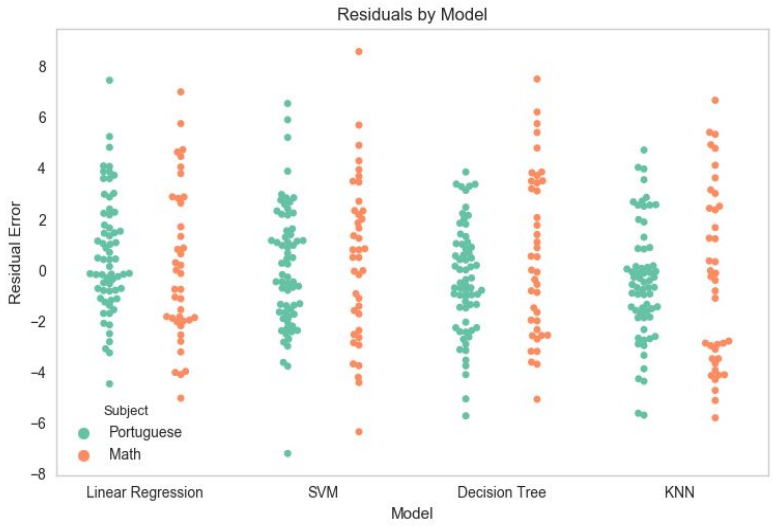

Fig. 1: Residuals by model for student academic performance

Model (Portuguese)	***R^2***	***Mean squared error***	***Mean absolute error***
Naive (mean/median)	0.00	7.35	2.18
KNN	0.05	6.53	2.05
SVM	0.11	5.72	1.88
Linear regression	0.16	5.40	1.82
Regression Trees	0.21	5.26	1.80

Math			
Naive (mean/median)	0.00	10.99	2.75
KNN	0.00	10.76	2.74
SVM	0.01	10.41	2.64
Linear regression	0.10	9.70	2.53
Regression Trees	0.12	9.42	2.53

Fig. 2: Error measurements for models

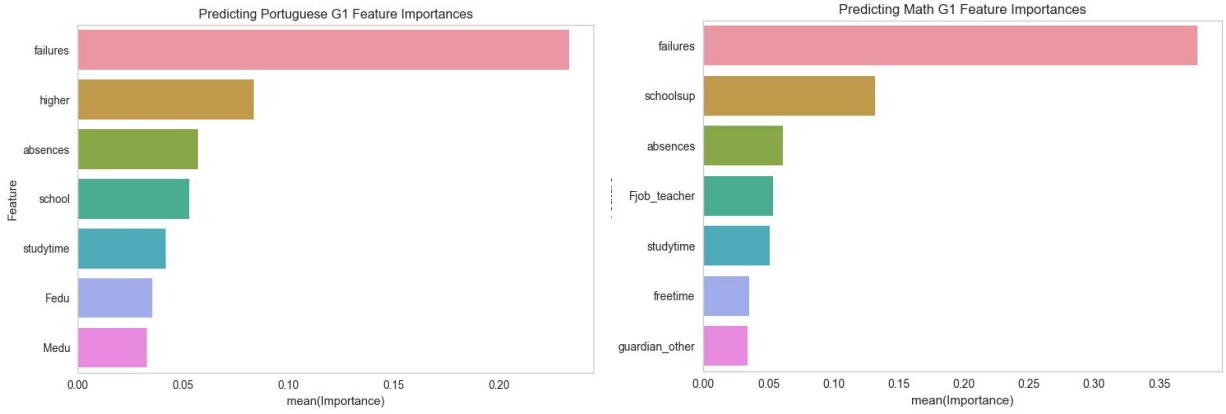

Fig. 3: Portuguese and Math G1 Feature Importances

Now, we transitioned into measuring the effect of alcohol on academics. Alcohol levels are reported from a scale of 1-5 with 1 being very low and 5 being very high. Fig 4. shows the distribution of weekend and weekday alcohol intake for all students. Fig 5 shows the combined weekly alcohol intake (weekday intake + weekend intake) vs G1 grades for both subjects.

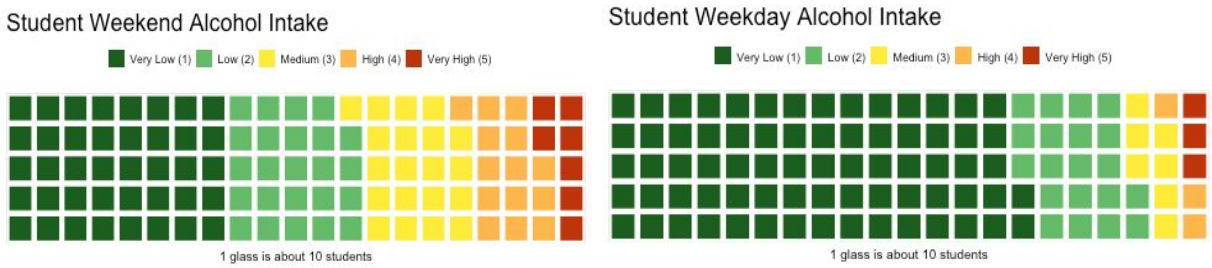

Fig. 4: Weekend and Weekday Alcohol Intake

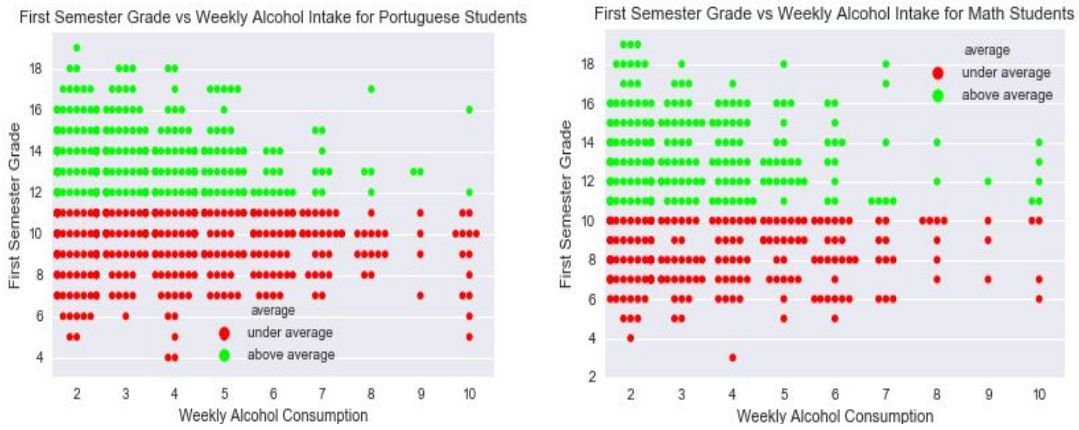

Fig. 5: G1 vs Weekly Alcohol Intake

Next, we compared weekday against weekend drinking for both subjects, the results of which are shown in fig. 6 (Portuguese) and fig. 7 (math).

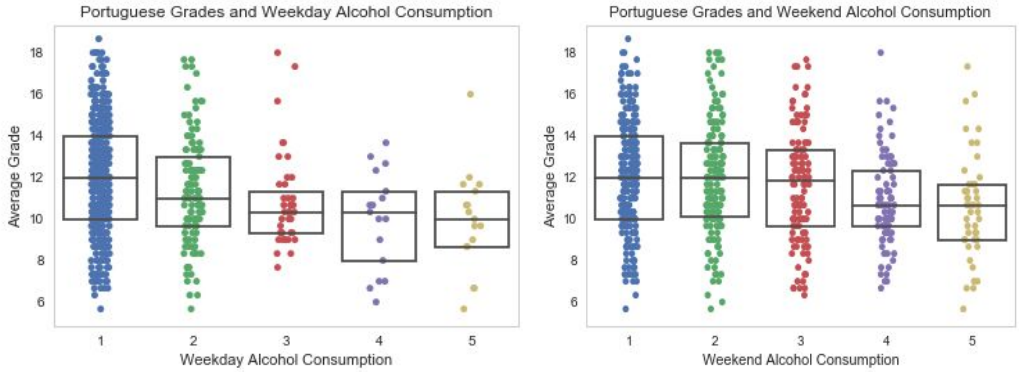

Fig. 6: G1 Portuguese vs Weekday and Weekend Alcohol Consumption

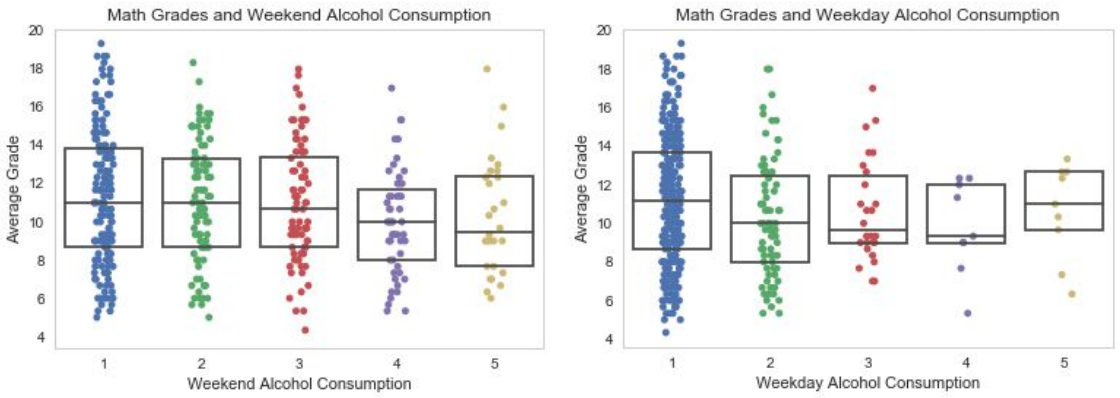

Fig. 7: G1 Math vs Weekday and Weekend Alcohol Consumption

The independence between G1 and alcohol consumption was tested using a Chi-square test of independence, the results of which are shown in fig. 8.

Model	*P-value*	***Independent***
Weekday Intake	0.01225	**No**
Weekend Intake	0.2391	**Yes**
Total Intake	0.2059	**Yes**

Fig. 8: Chi-Square Test of Independence of G1 vs Alcohol Consumption

Discussion

G1 Grade Prediction

In fig. 1, we plot the residual errors for each of our tested models. Just from eye's glance, we can easily see that trees did the best job of predicting G1 grades for Portuguese and that either trees or linear regression did the best job of predicting G1 grades for math. In order to solidify our qualitative predictions, we see actual quantitative error measures in fig. 2. For Portuguese, we see that trees did perform the best with the highest R^2, and lowest mean squared and mean absolute errors. Likewise, for math, we see the same results. Out of all the models we tested, trees did seem to be the most robust and with multicollinearity in our dataset, linear regression and SVMs did not fare as well. KNN fared the worst of our machine learning models, likely due to the limitations of our dataset. With so many categorical variables and very few

continuous variables, calculating a meaningful distance / similarity measure between different students proved to be quite problematic.

Now that we know that trees did the best job of regressing on G1 grades as compared to the other models, we look at the mean absolute error (MAE) to interpret just how well trees can predict G1 grades. Again, from fig. 2, we see that regression trees had a MAE of 1.80 for portuguese and a MAE of 2.53 for math. Given that the grading scheme in Portuguese secondary schools is a range from 1-20,[4] 1.80 points means 9% and 2.53 means 12.6%. Thus, on average, our model predicts portuguese grades and math grades within 9% and 13%, respectively, of the true grades. We also note that our tree model did the best job of predicting the very low and very high grades, giving the lowest mean squared errors (MSE) for both subjects. The low MSE bodes well for our goal of early detection of students at risk for failure as these students are often outliers of the typical distribution of grades. Overall, we are confident that even with the limitations of our dataset, we were able to produce a model that predicts G1 grades to an acceptable level of precision.

Another interesting observation is that portuguese grades appear to be more predictable than math grades, as evidenced by the higher errors for math. This is likely due to the fact that we had 649 portuguese students in our dataset compared to only 395 students for math; more data means more training examples which reduces variance.

Feature Importance based on Subject

Now that we have chosen the tree model, we look at which features are most important in predicting G1 grades. From fig. 3, we see that failures is by far the most powerful feature. This makes sense; students with a high number of previous class failures are more likely to fail in future classes than students with no past failures. In addition to failures, number of absences and amount of study time logically appear in the top features for both classes. Other than these features, there's a bit of a shakeup between top features for portuguese grade prediction and top features of math grade prediction.

We see that for portuguese, we have *Fedu* and *Medu* which are categorical variables that correspond to the level of education the student's father and mother have, respectively; neither of

these variables appear in the list of top features for math. Studies [1] have shown that the amount and quality of vocabulary a child is exposed to as an infant largely affects language development later in life. This explains why Fedu and Medu are strong predictors for portuguese grades; parent education levels are indicative of the amount and quality of vocabulary spoken in the household. For math however, most students go into primary school taking the same level of math, so this household effect is non existent.

Examining the top features for math, we have *Fjob_teacher* which means the student's mom's occupation is teacher, and *schoolsup* which stands for extra educational support. The presence of these two features in top features for math and absence of these two features in top features for portuguese give the notion that out-of-classroom tutoring is very important for math grades while it is not that important for portuguese grades. This explains why parents are more likely to have their kids tutored in math over other subjects, making math the most tutored subject worldwide.[3]

Effect of Alcohol on Academics

Moving onto the effect of alcohol on academic performance, we first examine the overall distribution of drinking for weekday and weekend and see from fig.4 that overall, students drink more on the weekends than they do on the weekdays.

Combining weekday and weekend alcohol consumption to get a measure of "weekly" alcohol consumption for fig. 5, we see that higher levels of weekly alcohol consumption correlate to higher proportion of students scoring under average (red) for both subjects.

Now, looking at weekday vs. weekend alcohol consumption for students in the portuguese subject in fig. 6, we notice that weekday alcohol consumption has a more drastic negative effect than weekend alcohol consumption. The drop in median of average scores from students with very low weekday drinking to those from students with very high drinking weekday drinking is about 2 points or 10%. The corresponding value for weekend drinking is only about 1.5 points or 7.5%. In addition, there tends to be more leeway for drinking during the weekend than the weekday. Weekend alcohol consumptions levels of very low, low, and medium almost have identical distributions of average grades while there is about a 0.6 point (3%)

decrease in average grade from a very low to low weekday drinking and another 0.3 point (1.5%) decrease in average grade from a low to medium weekday drinking level. The other interesting observation is the significant drop off in medium average grade from medium drinking to high drinking on weekends. This may be explained by the conjecture that students with high weekend drinking levels are more willing to drink on Sunday nights, which is obviously problematic given that the next day is a school day. We see similar trends in fig. 7 where we look at effects of alcohol on average math grades; students' academic performances are punished more severely by weekday alcohol consumption than weekend alcohol consumption and students have more leeway drinking on weekends than weekdays.

To further validate our findings, we performed a Chi-squared test of independence, where our null hypothesis was that G1 and alcohol intake are independent. The results of our test are found in fig. 8. At the 0.05 significance level, we reject our null hypothesis that G1 and weekday alcohol intake are independent while we fail to reject our null hypothesis that G1 and weekend alcohol or weekly alcohol intake are independent. This confirms our interpretation that weekday drinking is more detrimental to student grade than weekend drinking.

References

1. "Child Development and Early Learning." *National Center for Biotechnology Information SearchDatabase*, U.S. National Library of Medicine, 23 July 2015, www.ncbi.nlm.nih.gov/books/NBK310550/
2. Cortez, Paulo. "Student Performance Data Set." *UCI Machine Learning Repository*, University of California, Irvine, archive.ics.uci.edu/ml/datasets/Student+Performance.
3. Michaelidou, Anna. "The 2016 Top 5 Subjects Students Get Most Tutored For." *First Tutors*, 17 Nov. 2016, www.firsttutors.com/uk/blog/2016/11/the-2016-top-5-subjects-students-get-most-tutored-for/.
4. "The Portuguese Grading System." *Study in Europe*, www.studyineurope.eu/study-in-portugal/grades.