Catherine Beazley

SYS 6018: Data Mining

Professor Gerber

12/11/2018

What Predicts Student Performance in Math and Reading?

Current research shows that social and economic factors play a large role in student performance, with students of a lower socioeconomic status performing worse than students of a higher socioeconomic status (1). Researchers argue that understanding the factors related to student success is important for advancing sociological understanding of education and social problems for research and policy (2). While this reasoning makes sense, there are also a lot of other factors and surrounding events directly unrelated to the student that could also play a large part in student success like school funding, neighborhood factors, and current social and political events. The datasets describing the current educational situation are numerous. There is a lot of educational data on websites like the National Institute of Health and Data.Gov describing a wide range of educational aspects, like bullying, qualities of teachers, financial aid, and student performance.

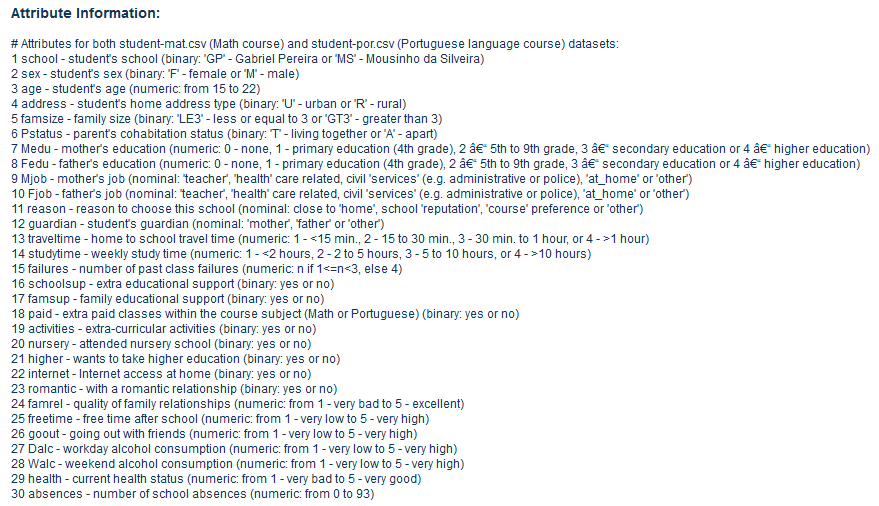
In an ideal world, education policy makers and educators would be able to address the factors this project finds to be related to poor academic performance so students are able to achieve their best, and so educators and educational policy makers would be the most interested in this project. By understanding the factors that affect academic performance in math and reading, an education policy maker could write policy that provides opportunity for students who may not have all the factors related to higher academic performance. Likewise, educators could understand the factors some students make be lacking to not only understand the whole picture of the student, but also to help the student by pointing them to appropriate resources, like school counselors, school breakfast programs, and school sponsored tutoring. After addressing the factors that relate to student success, more students would ideally be able to achieve their best and break out of feedback loops created by low socioeconomic status and poor academic achievement.

For this project I am trying to find that factors that most relate to and most accurately predict academic success in order for educators and educational policy makers to optimize academic success for all students despite their social and economic background. These academic changes can be measured by grades in math and reading. To show that I am using the best subset of regressors and the best model, I will use the validation error and importance of variables in the models.

A previous research project by the NIH investigated health and academic data to research whether the health and emotional distress of student played a role their secondary school performance. They applied logistic regression to the data with the response variable of failing one or more classes and found which variables in the data were most related with failing one or more classes. The evaluated their model using prediction accuracy. Though there has been a lot of research into this subject, there are so many social factors that can go into a student’s success and many combinations of social factors that may be present in the dataset or left out. Finding the true social factors and evaluating them, especially as time progresses and social and educational policy changes, keeps this problem a relevant problem to solve.

My hypothesis is that Random Forest will predict student performance in both Portuguese Language and Mathematics better than Ordinary Least Squares with optimal subset selection, Decision Trees, and Bagging. Additionally, I expected that, while the most important variables for each model may differ slightly for each model, there will be overlap among the most important features among all models, and the most important variables will be the same for predicting final exam grade in Mathematics as for predicting final exam grade in Portuguese Language.

The data used for the experiment come from the dataset “Student Performance Data Set” from UC Irvine Machine Learning repository and describe “student grades, demographic, social and school related features” (3). The student grades consist of first period grade, second period grade, and final grade in both Portuguese Language and Mathematics. The data from UC Irvine came from two datasets, one consisting of all the regressors and the three grades for Portuguese Language, and another consisting of the same regressors and the three grades for Mathematics. I merged two datasets to include only the students recorded in both, and removed the repeated regressor columns. Because the objective is to predict Mathematics and Portuguese Language grade separately, the merged dataset was split back into two datasets to separate language an mathematics for computational ease. The dataset notes that the period grades are highly correlated with the final grades, so each of the final two datasets were further split into a dataset including period grades and a dataset excluding them. This way it is possible to analyze the effect of the other regressors in predicting final grades without the effect of the period grades. Ultimately, four datasets were used: one dataset including period grades for predicting Language final, one dataset excluding period grades for predicting Language final, one dataset including period grades for predicting Mathematics final, and one dataset excluding period grades for predicting Mathematics final.

[Screenshot of Regressor Explanation excluding period grades (Source 3)]

Data Exploration showed that a number of the numerical variables were skewed right, but because of the nature of these variables, the skew makes sense since one would expect predominately lower values for some of the variables, with a lower bound of 0 and no upper bound. For example, one variable that is heavily skewed is *absences*, which makes sense because there is more likely that students have a low number of absences with a lower bound of 0, with several students having larger numbers of absences. The distributions of the response variables

are normal, with a mean of 12.52 for Language and 10.39 for Mathematics. Overall, the final Mathematics grades seem to be lower than the final Language grades. Data Exploration also showed that some of the categorical variables had a class imbalance. The variables *higher*, *schoolsup*, and *internet* all had an overwhelming majority of students fall under one category. Finally, plots of regressors against response suggested that some of the regressors had relationships with the response. The regressors failures, *traveltime*, *Dalc*, *Walc* seemed to have negative linear relationships with final grade, and the period grades had clear positive linear relationships with the final grade.

Methods:

The first step in the project was to split the four datasets into training and validation sets, and then train all the models on those training sets. After that, I could train the regression models on the data and ultimately test the performance of the models on the validation set.

I chose to apply Ordinary Least Squares because data exploration showed that some of the plots of response versus regressor suggested that there is a linear relationship between the regressors and the response. Additionally, though some of the regressor variables are skewed, the response variables seem to follow a normal distribution and there are enough observations to assume heteroscedasticity. With clues that a linear model may be correct and with the assumptions met, Ordinary Least Squares seems an appropriate regression model for the data.

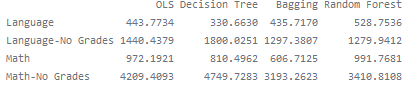
Because the datasets contain a lot of variables, with sets including period grades having 32 and sets without having 30, the number of varibales included in the final models were reduced by subset selection. To determine the best subset of regressors, all possible subsets were found for each of the four models, and then possible subsets were ranked by adjusted R^2, Mallow’s Cp statistic, AIC, and BIC. The best subset according to its adjusted R^2, Mallow’s Cp statistic, AIC, and BIC was used to fit an Ordinary Least Squares model for the dataset.

I chose to apply the tree-based regression models of Decision Trees, Bagging, and Random Forest to the data because of the high number of variables. With tree-based regression, the regressor space is constantly stratifies and segmented to address the high number of predictors and ultimately make an appropriate prediction. Additionally, the regressors consist of mostly categorial variables in addition to the numeric variables, which may make tree-based regression a better approach to prediction. I applied these three models to compare their results of their differences in node splitting.

After fitting all the models, I compared the Mean Square Error validation accuracy of them to see which model was the best predictor. I also compared the importance of the regressors in each of the models to see which variables were the most important in predicting Language and Mathematics grades.

Evaluation, Executed Approach, and Results:

Evaluation of the Mean Squared Validation Error showed that overall, one model did not perform better than the rest and instead, three of the four models showed to have the best performance for different datasets. For the dataset for Language, the Decision Tree had the best prediction accuracy. For the dataset for Language excluding period grades, Random Forest performed the best. For both datasets for Mathematics, bagging performed the best. Additionally, the MSEs for datasets containing the period grades were much lower than the MSEs of the datasets excluding them, which confirms UC Irvine’s note that the period grades are extremely correlated with the final grades and that prediction without the period grades is more challenging than with them.



Based on the importance of the variables in the tree-based methods and the varibales used in the subset of regressors in the OLS models, it seems that consistently, the most important variables, when excluding the period grades, are the number of absences and the number of previous failures for predicting Mathematics final grade, and the number of previous failures and whether the students intends to obtain higher education for predicting the Language final grade.

What I learned about the problem is that academic information predicts student success more than social and demographic variables describing a student. I also learned that there is not one regression model among the ones I tested that consistency predicts with the best accuracy, which means could mean that there are multiple ways to model this problem. What I learned about the data, methods, and evaluation is that the inclusion of highly correlated variables affects the role of other variables in the model. For instance, when including the period grades in the model, the relative importance of the other variables decreased for the tree-based methods and not as many other variables were included in the best subset in the OLS method. By removing the period variables and fitting regression models to those datasets, I was able to isolate the other variables to evaluate their importance.

Overall the results do not support the hypothesis. Rather than Random Forest performing better than the rest according to Mean Square Error, no model performed the best. Additionally, while there is overlap among the most important regressors in the models, it seems that the importance of regressors is different for the Language datasets and for the Math datasets.

My work moves us closer to ultimately finding the variables that relate to student performance by suggesting that academic variables, like the number of absences, number of previous failures, and whether the student wants to obtain higher education, are more important than social and demographic variables in predicting and explaining student performance. However, these academic variables are not exclusively academic, as the number of absences can be related to physical and mental health and family events at home, which are predominantly physical and social aspects of a student. Similarly, the number of failures and whether the student wants to obtain higher education could be related to other social and economic variables not included in the dataset. My work could also show that more or different regressors are missing from this dataset that could explain student performance, ultimately furthering understanding of the variables are and are not important in predicting and understanding student performance to advance sociological understanding of education and social problems for research and policy.

Bibliography

1. “Education and Socioeconomic Status.” *American Psychological Association*, American Psychological Association, www.apa.org/pi/ses/resources/publications/education.aspx.
2. “National Center for Biotechnology Information.” *Current Neurology and Neuroscience Reports.*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/.
3. UCI Machine Learning Repository: Flags Data Set, archive.ics.uci.edu/ml/datasets/Student Performance.