Title: Predict students' math class final exam score
Name: Xuancheng Tu

Students' academic performance is associated with many in-class and out-of-class factors. The goal of this project is to predict students' math class final score using factors that may be associated with academic performance. This might provide educators with insight about which factors have the most significant impact on students' academic performance and make more efficient education strategy based on that.

The dataset that is analyzed in this project is downloaded from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Student+Performance). The data is collected from two Portuguese secondary schools and contain the information of 395 students. The original dataset contains 33 columns that include information but not limited to personal information, family condition, extracurricular life, and academic record. 10 attributes are chosen for data analysis: (1) having extra paid math classes or not (2) having extracurricular activities or not (3) having Internet access at home or not, (4) with a romantic relationship or not, (5) quality of family relationship, (6) amount of free time after school, (7) frequency of going out with friends, (8) current health status, (9) first period grade and (10) second period grade. Among those columns, (1), (2), (3), and (4) contain non-numeric "yes or no" values. So "yes" is replaced by 1, and "no" is replaced by 0 for data analysis. (5), (6), (7), and (8) are numerical columns with scale of 1 to 5 (from 1 for very low/bad to 5 for very high/good). (9), (10) and the final grade are numerical with scale of 0 to 20. (Description of attributes are cited from original dataset website https://archive.ics.uci.edu/ml/datasets/Student+Performance).

Firstly, we analyze the relationship of period grade 1 and period grade 2 with final grade using Figure 1. Points in figure 1 are jittered in order to avoid overlapping. The color of points represents the final score. Darker color implies higher final score. Points in the lower left corner, which represent students with low period grade 1 and 2, have light color which represents low final grade. From lower left corner to upper right corner, we can see that points are getting darker, which means that higher period grade 1 and 2 are associated with higher final grade. Points on the same vertical line are often darker when they have higher y-axis coordinate, which means that for students with the same period grade 1, students who have higher period grade 2 tend to have higher final score. In comparison, the variation of darkness of points on the same horizontal line is less obvious, which means that the impact of period grade 1 on final grade for students with the same period grade 2 is less significant than the impact of period grade 2 on final grade for students with the same period grade 1. Also, for all students who get 0 point for period grade 2, they get 0 point for final grade as well, regardless of period grade 1.

Then, we run principal data analysis (PCA) to investigate the dimensionality of the data, as shown in Figure 2. When PCA is run on unscaled data, 5 variables can explain 95% of the variance. However, different variables are in different scales in the dataset. Period grades are in 0-20 scale; binary columns only contain zero-or-one value; other numerical attributes are in 1-5 scale. So in order to make columns comparable, a pipeline with (1) a StandardScaler and (2) a PCA is applied. The pattern of the curve of scaled data is close to a linear straight line. Thus, the dimensionality of the data cannot be reduced.

Finally, we perform linear regression in order to find the coefficients of different attributes. 75% of data is randomly selected for training regression model, and 25% of data is for testing. Both the unscaled and scaled version of data achieved 0.781 r2_score on test dataset, which shows that the linear regression model fits well. Based on the same reason as stated in discussion of figure 2, we choose to use the scaled data. The top five most significant attributes are the second period grade (3.65), first period grade (0.35), extracurricular activities (-0.21), romance relationship (-0.17), and health (0.13). Students who perform well in previous academic activities tend to have a high final grade, and period grade 2's influence on final grade is more significant than period grade 1's, which is consistent with the discussion of Figure 1. Students with better health also tend to perform better academically. In addition, Extracurricular activities, romance relationship negatively influences students' final grade. The factor with the least significant influence is whether students have internet access at home or not. Its coefficient is -0.001 which is invisible in the plot, which represents a weak negative influence.

In conclusion, Higher period grade 1 and 2, and better health status are associated with higher final grade for students. period grade 1 and 2 are the two most significant factors that predict students' final grade. Compared to period grade 1, period grade 2 has more significant impact on final grade. Other important and negative factors are having extracurricular activities and having romance relationship. Among the selected 10 attributes, none of them can be reduced if we still want to keep variance of data.

Figure 1: Period grade 1 and period grade 2's relationship with final grade

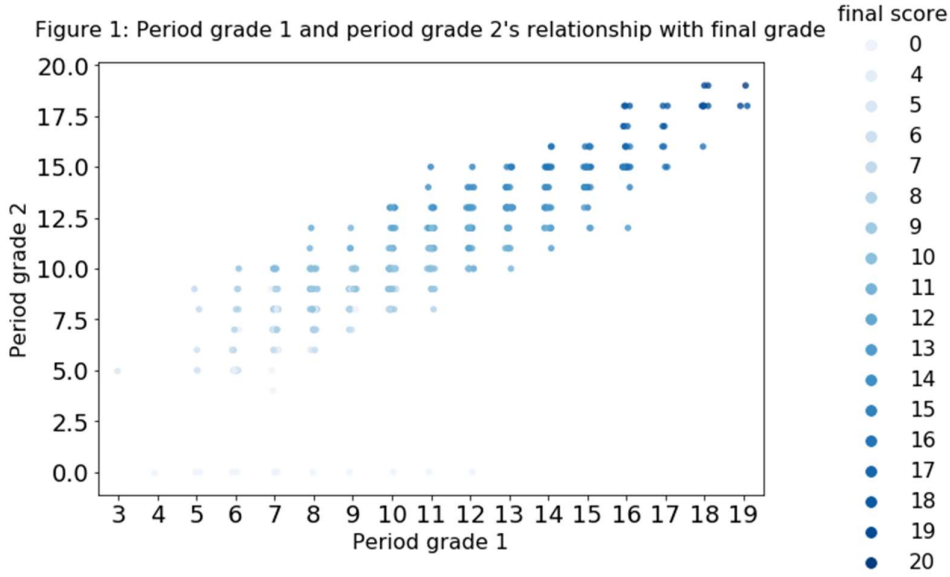

Figure 2: Principal Components of Breaks

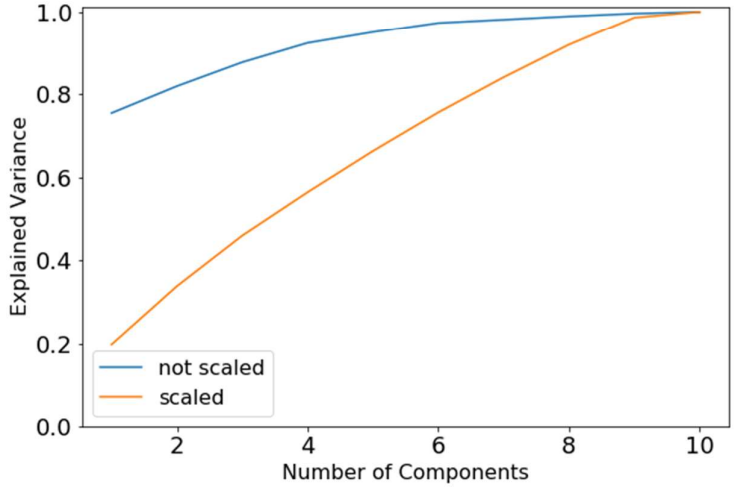

Figure 3: Linear Regression Coefficients