Project 4 Analysis:

The aim of our project is to uncover patterns in high school student success rate. We’ll examine relationships between class participation and study time, at home support and tutoring, extracurricular activities and finally gender and ethnicity derived from the data.

Using a data set with a large sample size of students in various grades, we were able to create a model that predicts students' grades based on these variables, and the results of those students actual GPA, showing how some positively correlate with school performance, while others can have negative impacts. These are the results of the model:

Mean Squared Error: 0.03421369163775982

R-squared: 0.8397896069525346

ParentalEducation -0.004164

StudyTimeWeekly   0.075879

Tutoring          0.057121

ParentalSupport   0.072350

Extracurricular   0.041709

Sports            0.046196

Music             0.024733

Volunteering     -0.001423

Log\_Absences     -0.413729

The results of our projected GPA based on actual GPAs in this sample size showed a MSE, which indicates the average squared difference between the observed actual outcomes and the values predicted by the model, of roughly .034. This number is relatively low, indicating our models’ predictions are fairly close to actual values.

R-squared measures variance, and the value our model gave is roughly 0.839, which interprets as roughly 83.9% of the variance in our dependent variable, which is GPA, can be explained by the independent variables used in the model.

Next, we listed our independent variables, with coefficients, showing correlation between each variable and predicted GPA outcome. First, we measure parental education, which surprisingly has a negative correlation with GPA outcome. The value is roughly -0.0042, which although is relatively small, still indicates higher parental education correlates with lower GPA for students.

The next variable we used was study time on a weekly basis. As expected, this shows a positive correlation with predicted GPA, with a value of roughly 0.076. This is a much higher value, meaning more hours of study time has a much larger positive impact relative to the negative impact of parental education.

From here we take a look at absences, which has the largest negative correlation with predicted GPA outcomes. This value is -0.41, showing missing classes has a big impact, and missing too many classes can be detrimental to success in school.

Tutoring and parental support both have positive impacts on predicted GPA. Although relatively small, with tutoring having a coefficient of 0.057 and parental support coefficient being 0.072, we can see these are both positively impactful for student success.

Finally, we take a look at extracurricular activities, as many kids in school participate in clubs and sports outside of the classroom, and we felt it would be important to include in the model to measure any impacts. Extracurriculars at a whole had a low, but positive impact with predicted GPA, having a coefficient value of .042. Surprisingly, when we looked deeper into specific activities, we see that sports have the best impact, with a coefficient of .046, while music has a smaller but positive impact with a coefficient of 0.025. Volunteering interestingly had a negative impact on predicted GPA. The value is miniscule, being only -0.0014, but still a value the group didn’t expect the model to have.