Dsga1001_capstoneProject_KDS

Marcus Choi [choim09], Yunseok Jang [yj2369], Yeong Koh [yk2678], Yoon Tae Park [yp2201]
December 21, 2021

1. Introduction

Our dataset is composed of 395 students taking math courses in Portuguese secondary schools. The dataset itself is extracted from Kaggle under the name of 'Student Alcohol Consumption.' (UCI ML) The rows and columns are provided as:

Row 1: headers
Row 2-396: responses from individual participants
Column 1: school - student's school (binary: 'GP' - Gabriel Pereira (0) or 'MS' - Mousinho da Silveira (1))
Column 2: student's basic information including sex, age, and address (urban (0) or rural (1))
Column 5-10: information about the student's family including family size, parent's cohabitation status, parents' education, parents' jobs
Column 11: reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
Column 12: guardian - student's guardian (nominal: 'mother', 'father' or 'other')
Column 13-14: time spent outside of school (traveltime - home to school travel time and studytime - weekly study time)
Column 15: failures - number of past class failures (numeric: n if 1<=n<3, else 4)
Column 16: schoolsup - extra educational support (binary: yes or no)
Column 17: famsup - family educational support (binary: yes or no)
Column 18: paid - extra paid classes within the course subject (binary: yes or no)
Column 19: activities - extra-curricular activities (binary: yes or no)
Column 20: nursery - attended nursery school (binary: yes or no)
Column 21: higher - wants to take higher education (binary: yes or no)
Column 22: internet - Internet access at home (binary: yes or no)
Column 23: romantic - with a romantic relationship (binary: yes or no)
Column 24: famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
Column 25: freetime - free time after school (numeric: from 1 - very low to 5 - very high)
Column 26: goout - going out with friends (numeric: from 1 - very low to 5 - very high)
Column 27-28: Dalc/Walc - daily/weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
Column 29: health - current health status (numeric: from 1 - very bad to 5 - very good)
Column 30: absences - number of school absences (numeric: from 0 to 93)
Columns 31-33: first period/second period/final grades (numeric: from 0 to 20)

As for the data preprocessing, we took care of the null and duplicate values. For the binary text features, we encoded it to binary dummy variables in terms of 0 and 1. Our main interest throughout this dataset is the relationship of a variety of variables affecting the grades of students.

2. What factor leads to academic success?

Given each student's characteristics, our team wanted to figure out which factors are important to get a good grade. Grades were not binary values, so our team conducted linear regression. In terms of features, there were 15 numerical columns, and 18 categorical columns. For categorical columns, our team conducted one hot encoding. We confined our target value to 'G1.'

A simple linear regression resulted in train_error and test_error of 4.6984 and 5.4468, respectively, with an r^2 score of 0.3740. For Ridge regression, the optimal result in terms of minimizing test_error and maximizing r^2 score was alpha equal to 1, train_error as 4.6987, test_error as 5.4355, and r^2 score as 0.3740. For Lasso regression, the optimal result was alpha as 1e-05, train_error as 4.6984, test_error as 5.4468, and r^2 score as 0.3740. Visualizing relationships between features and target value, there were some factors that seem to lead to academic success. Namely,

- No extra educational support leads to good grade
- School GP has a positive correlation with grade
- Willingness to take higher education is the strongest feature
- Mom's job as a teacher, long study time, and educated mom leads to a good grade

The first result came to pique our interest. Intuitively, it makes sense that extra educational support would help to get a good grade. So, we conducted a hypothesis testing to see if this feature difference actually has a difference given two student groups. Given the alternative hypothesis that students with extra support get better grades than those without and the null hypothesis that there is no difference, we obtained p-value 0.0487, so we rejected the null hypothesis.

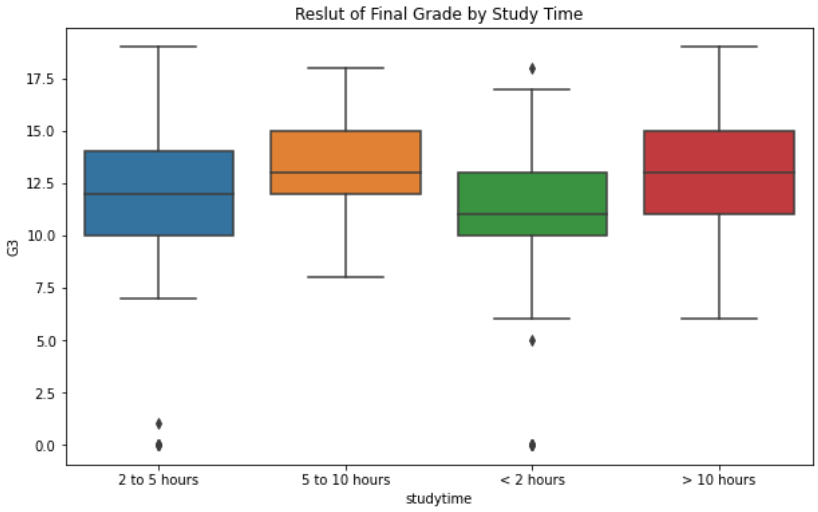
Since school GP has a positive correlation with grade, our team conducted decision tree modeling to make a model that can predict a school, given the same dataset. Then, we visualized the feature importance to see what is the most important feature that affects the model. Setting school as the target value, and other columns excluding grades as features, we were able to create a decision tree model that had an accuracy of 0.6821 and an auc score of 0.6440. Doing hyperparameter tuning using grid search cv, our model improved as shown by the increase in accuracy to 0.7026 and the auc score of 0.7159. After visualizing the feature importance, our team found an interesting fact that living in urban areas is the most important feature. This can be also checked by looking at the proportion of the students living in urban areas in each school. For school GP, the proportion was 81.56%, while school MS was 47.34%.

The schools' mean grades were also different (school GP: 11.99, school MS: 10.30), and our team checked to see if this difference was statistically significant. Given the alternative hypothesis that students in school GP received higher grades than the students in school MS, and the null hypothesis that there is no difference, we obtained p-value 7.119791239241908e-15, therefore rejected the null hypothesis.

3. Does the amount of time spent studying have an effect on students' final grades?

To determine the effect of study time on students' final grades, we first decided to take a closer look at the distributions of the categorical variable, which consisted of 4 different groups – students who studied 2 to 5 hours, 5 to 10 hours, less than 2 hours, and greater than 10 hours. Comparing the student's final grades based on the amount of time studied, we observed that the average value of the groups is highest for students who spent 5 to 10 hours of studying with a mean value of 13.23 and is lowest for students who spent less than 2 hours with a mean value of 10.85.

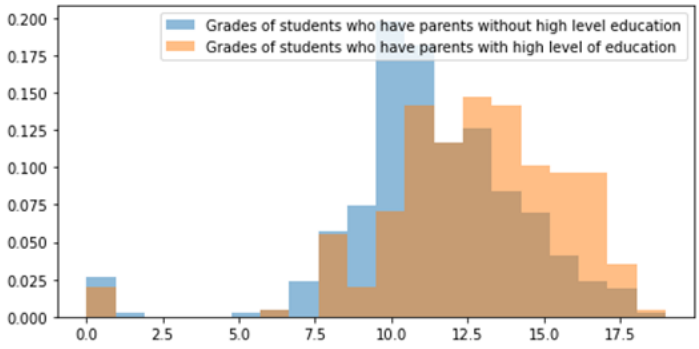
In order to determine if there is a significant difference between the time spent studying and the final grades of the students, a chi-squared hypothesis testing was conducted as this variable was categorical and not reasonable to reduce the values to the mean. The p-value of this test was 1.144e-4, which is less than the threshold of p-value 0.05. Therefore, we were able to reject the null hypothesis and conclude that there is statistical significance between the time studied and the results of the final grades of the students.



The p-value obtained from hypothesis testing was consistent with the values of the boxplot. However, one interesting thing to note was that, although there is statistical significance between the amount of time studied and the results of the final grades, it is not necessarily a linear relationship between these two variables. Students who spend more time studying do not particularly receive higher scores than students who study less as we can observe the mean scores are not higher for students studying more than 10 hours. In conclusion, this is an indication that the potential optimal time of studying would be between 5 to 10 hours as studying more efficiently would be more effective than studying longer hours.

4. Do parents' level of education affect students' performance in school?

To determine whether the parents' education level has any effect on the students' grades, our dataset was divided into two groups – 1) students who have at least one parent who has pursued education above secondary school and 2) students with parents, neither of whom pursued education above secondary school. A preliminary check on the distribution and the mean grades of the two groups showed that the grades are approximately normally distributed and the average grade of the first group is 12.9567, whereas the average grade of the second group is 11.4770. To determine whether the difference in the overall grades of the students in the two groups is significant, an independent samples t-test was conducted, which yielded a p-value of 2.30726e-08. The p-value that was obtained is approximately 0, which is smaller than the significance level of 0.05. Therefore, we reject the null hypothesis that the overall grades of students in the two groups are the same and conclude that the grades of students who have at least one parent with a high level of education is greater than the grades of students without a parent who has a high level of education.



The p-value was consistent with the above histogram for the overall grade distributions of students who have at least one parent with a high level of education and of students whose parents do not have a high level of education. The plot shows that the means of the two compared groups are different, and one would have been able to conjecture that the grades of the students who have parents with a high level of education are higher than those of students with parents without a high level of education.

Although most students in the dataset answered "yes" in response to the question in the survey whether they want to take higher education after high school, a much higher proportion of the students in Group 2 than the students in Group 1 said that they do not want to pursue higher education. Specifically, 67 out of the 407 (~16.5%) students in Group 2 do not want to take higher education as compared to only 2 out of the 206 (less than 1%) students in Group 1. This could be an indication that, rather than parents' level of education directly affecting the students' performance in school, the students' motivation to study is one of the confounding factors that influence the students' grades.

5. Summary and Conclusions

Our questions mainly orbited around finding a meaningful relationship between a variety of factors and academic achievements which were represented in terms of grades obtained. In terms of question 2, 'What factor leads to academic success', our team concluded that motivation of taking higher education is the strongest factor. Also, there were some interesting features that led to academic success such as no extra educational support and school GP's performance compared to school MS. While comparing the school's performance, we found that each school's grade difference was statistically significant evidenced by null hypothesis testing, having a p-value of 7.120e-15. Also, geological location of students was an important factor in each school's academic performance. Specifically, 81.56% of school GP's students lived in urban areas, while 47.34% of school MS's students lived in urban areas.

For question 3, we were interested in the effects of time spent studying on the results of the final grades for the students under the assumption of the null hypothesis. We decided to conduct a chi-squared hypothesis test as the variable of interest was categorical and not reasonable to reduce the values to the mean. The p-value was 1.144e-4, which is less than the threshold level of 0.05 and therefore, we concluded that there is a significant difference between the amount of time studying and the final grade scores for the students in this dataset. One interesting thing to note was that the limitation of this analysis is not a linear relationship as students who spent more time studying did not necessarily score higher on their final grades. One way this question could be answered better with more student data would be to examine the optimal time of studying.

An independent sample t-test led to the conclusion that there is a significant difference in the grades of students with parents who have a level of education and the grades of students whose parents do not have a high level of education. The test was performed under the assumption that the grades in each group were approximately normally distributed after plotting the distribution of the student grades on a histogram. Without the normal distribution assumption, Mann Whitney U-Test would have been also appropriate, which would have resulted in a p-value of 6.112e-11 and led us to the same conclusion. In answering this question, we realized that there could be many confounding variables that could have contributed to the students' grades.

The overarching conclusion of this project is that we were indeed able to find very likely factors that explain academic performance of students. However, there were some limitations as we proceeded in conducting data analysis. Namely, our dataset may not be representative of the entire population. Also, our underlying assumption of this project is that these features account for the majority of factors that likely influence academic performance. Lastly, an inherent issue of this dataset comes from its bulk. That is, the sample size can be considered not adequate, leading to susceptibility of noise and outliers. A bigger and more comprehensive dataset may provide more accurate results. Nonetheless, our team is confident that the result we found at least points several directions that may lead to finding the optimal method of education.