Aaron Drexler

DSC 680

Term Project Proposal 1

Topic: Personal Life Impact on Education

For my project, I wanted to delve into the educational field to discover what features of a student's life has an impact on his or her academic performance, both good and bad.

Business Problem

While I am getting my master's in Data Science right now, I am also a high school math teacher during the day. One thing that I deal with daily are poorly performing students in my class. There are many students in high school who regularly get low grades on a regular basis. Now, the cause of their bad grades usually has two different factors, classroom performance and home life. As a teacher, one of the main things that I do every day is analyze in class performance and then either create or adjust the educational plans or methods accordingly to compensate and try to improve student scores.

Then again, let's consider the second factor on a student's grade, the student's personal life. As a teacher, you find that there are some students that do not want or will not learn, no matter what the educational plan happens to be, due to either lack of desire or motivation. One of the biggest things experience teachers tell their younger counterparts (aka me) is that you do not know what is going on in students' personal life, and how it is affecting their education.

The problem is that we don't truly track and quantify factor two like we do with factor one. We don't see what personal issues are affecting students over a larger sample, only on an individual basis. That is my goal for this project. ***I want to predict what aspects of students' personal home life has the greatest impact on their school performance.***

Data Sets

I plan to use the following Kaggle data set: https://www.kaggle.com/datasets/larsen0966/student-performance-data-set This is a database that I found on Kaggle that gathered a large amount of secondary information about students' personal lives, habits, and backgrounds, while also sharing their mathematical performance. ”This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por).” The following are all the data points gathered are as follows:

- student’s school
- student’s sex
- student’s age
- student’s family size
- parental cohabitation status
- mother’s education
- father’s education
- mother’s job
- father’s job
- guardian
- home to school travel time
- weekly study time
- number of past class failures
- extra educational support
- family educational support
- extra-curricular activities
- attended nursery school
- wants to take higher education
- home internet access
- in a romantic relationship
- quality of family relationships
- free time after school
- going out with friends
- health status
- absences
- first period grade
- second period grade
- final grade

Methods

The focus of my project is going to be to study student performance and cross reference that with the attributes and descriptors about each student. With this study I want to find the specific areas where the relationship between personal life and school performance are most connected. The outcomes that I will be looking at is to find patterns in the data to find descriptors that have a high correlation to poor grade performance.

Using this data, I want to build an optimal profile, by determining, via correlation, which of these attributes have the greatest effect on their scores, while also determining which group of students are in the most need of assistance. I also plan to eliminate some features to focus exclusively on quantifiable personal life tracking. The focus of the model was looking at the Pearson Correlation Heat map to determine correlation between the features of the data. Using the correlations, I can create a linear regression model.

Ethical Considerations

There are a few ethical dilemmas when it comes to this data base. First, the accuracy of the data might not be reliable, as we are relying on the truthfulness of high school students, which is very suspect at best. Second, this study was conducted with Portuguese high school students. This may cause a drift or bias in the data that is unique to Portugal and does not apply to American Students. Third, the study has only 648 students. While that may appear to be a lot, I believe that it still will present as a small population creating bias in the data. Finally, this data is taken from just two schools, and bias occurs with specific schools. In short, the data base is lacking diversity and size, while relying on student honesty. So, while it may not affect the results, we must take the outcomes with a grain of salt.

Challenges/Issues

There are not too many Issues I believe I will run into. One issue that might occur is comparing categorical data with numerical data. Another issue is how the ethical issues with the population size will affect the correlation strength. Finally, being able to take the correlation data and turn it into a linear regression model might be a challenge.

References

https://www.kaggle.com/datasets/larsen0966/student-performance-data-set