Aaron Drexler

DSC 630

12/12/2021

**Milestone 2: Data Selection and Project Proposal**

When I was considering which direction I wanted to go with this project, there were several different directions that I could go. I have several interests that have multiple different avenues of research possible. However, I considered, what do I have experience with and what would I best be able to articulate in a data project? When looking in this direction, I chose to look in an area in which I both have field experience and have conducted and participated in other data projects, Education.

While I am getting my master’s in data science, I currently am working as a math teacher working at a high school in Florida. I teach two subjects entitled Liberal Arts Mathematics and Advanced Topics in Mathematics. While most do not know these subjects, they consist of a very wide range of topic coverage. One area of focus in teaching math is equity, in ensuring every student is receiving the best opportunities to succeed. The reason why I am bringing up both the subjects that I teach and the focus on equity is because I both personally use data in my instructional plans and I am also participating in an educational grant to study equity in mathematics.

For my project, I wanted to focus on the equity aspect of education as well as optimizing student performance given external factors. First, let’s focus on equity. The educational grant that I am participating in is studying the disparity in math performance. Minority communities, particularly as African American and Hispanic, consistently score lower on standardized testing in mathematics when compared to White and Asian students. While this grant is specifically studying methods of instruction that would improve minority scores, I want to see if I can duplicate their initial data, highlighting the concern present and score disparity, potentially refining and improving on their conclusions.

The next aspect of this how to improve and maximize math performance. I want to find the optimal student. I want to look beyond individual or racial data and look at a wide range of descriptors and data points about the students. I want to gather as many data points as possible in relation to the students. I am not looking at maximizing the student totals, but rather maximizing the number of comparable attributes and tendencies about each student that I can gather, then using them to find useful and actionable conclusions about education and mathematical testing.

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. So much of education is about bridging the educational divide, and with this study I want to find the specific areas where that bridge is pronounced. 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 math performance while also searching for an optimum educational situation. The reason to search for the optimal situation is to create the goal to reach for those who are educationally deprived.

The first and foremost concern with this project is data collection. When dealing with these large-scale comparisons, having a large enough population as well as enough data elements to compare is a struggle. Initially, I have one good database that I found. 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. The data points in this database are:

* 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
* 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
* math scores.

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 plan to run various models such as K-means, correlation matrices, and other correlation-based visualizations.

There are a few flaws with the data though. This data was gathered is not the best representation of the population under investigation (Portuguese town and ages 15-22). However, the data asked is universal and can be properly applied and compared to its American counterparts, if available. I also plan to reach out to the grant that I am working with to see if they will allow me access to their initial statistical findings.

The plan moving forward has two core elements to it, gather more data points, through request or research to cross reference with our current data, and create a code designed to identify most vital descriptors, then compare them to find both insufficient and optimal levels or outputs for that data category. The more data included, the more refined our parameters will become. The goal is to gather enough data to piece together the best student perdition possible.

* **Problem Statemen**t: Too many students are disadvantaged in education, which leads to inequity, both with testing performance and educational environment
* **Data and research scope**: I have one wide ranging data set that involves a great deal of additional descriptors. I plan to add more data set with the largest possible topic range.
* **Technical approach**: Most of the work will involve cleaning data, and applying them to graphics, visuals, and descriptors. I plan to either use ggplot2 in R or pyplot in Python.
* **Data Sources**: https://www.kaggle.com/janiobachmann/math-students/version/1
* **Expected Results**: I plan on creating a model of ideal student circumstances to achieve the best possible results while simultaneously highlighting specific attributes that have the strongest negative effect on performance.
* **Project Plan**:
  + Gather more data set, preferably high schools in America, to find more descriptors to compare and predict
  + Create a model and code the visuals and data to find actionable insights
  + Reach conclusion and create formal answer to optimal and deprived students
* **Risks**:
  + Database Bias
  + Lack of enough data
  + Obvious conclusions