Capstone Project #1 Proposal

Chris Malec

Predicting Student Outcomes:

A quality education is an important factor for long-term financial stability in the modern world. Providing as many people as possible with the opportunity for a high-quality education is critical to building a strong middle class and helping people reach their potential. There are many ways to measure success in school, not all of which correlate to earning potential and happiness, however dropping out of school results in several negative consequences, such as a lower median income,[1] and higher unemployment rate.[2]

It would be impossible for all students to get A's all the time, and academic excellence isn't necessary for individuals to live happy and productive lives, but achieving financial security without a High School diploma is difficult in today's world. As a nation, we should be able to set a goal to achieve as close to a 100% high school graduation rate as possible.

Many, many factors go into a student's decision to drop out of High School. Using data from the National Center for Education Statistics, we follow over 20,000 students on their journey from Freshman year in 2009 to graduation in 2013, and even beyond.

The goal of this capstone project is to use this data to predict whether a student will graduate or not from High School given a combination of school programs, living situation, and academic achievement.

The dataset is collected by the National Center for Education Statistics and a public use version, with certain personally identifying information removed, is available on their website. It contains a random sampling of high schools in the US and a random sampling of students within each school. Data was collected on students in 2009 in their freshman year, their junior year, their proposed graduation date along with their high school transcript, and several years post-graduation.

Weights are provided so that given a particular school and time period the data can be aggregated to reflect the US population. This data is ready for use by researchers and government workers, and documentation for all columns and labels is provided. Therefore the data will not require undo cleaning and wrangling in order to build statistical models.

---

[1] Digest of Education Statistics 2017, table 502.30
[2] Digest of Education Statistics 2017, table 502.80

The csv file is quite large, but not so large that it can't fit in memory, however, it fits in much more easily if we take away some of the columns that won't be useful to the analysis right away. These include columns that are suppressed for public use due to privacy concerns, columns that include statistical weights used to create accurate aggregates of the data, and columns indicating where values have been imputed.
The client for this project is a school district. A school district is highly incentivized to increase graduation rates with limited resources. District performance can be tied to state funding and individual school performance can affect district funding. There are a multitude of ideas on how to help graduation rates, by building statistical models, the benefit of those ideas for a particular student or school can be studied using them. This will help districts concentrate resources on programs and staff that will render the greatest benefit to graduation rates.

The deliverable will consist of all Jupyter notebooks I develop for this project, a final report detailing the work completed, and a presentation slide deck explaining the models and associated results.