Capstone Project #1 Milestone Report

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.

Wrangling the data centers around reading in a large csv file and then separating the columns by type and time. There are three different types of columns: data,

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

weights, and imputation flags. The data contains answers to questions sent to students over the course of the study. Since the students were not randomly selected, weights are included to convert the given responses to numbers that are representative of the US population. Weight columns can also be used to calculate variances in the counts and statistics calculated from the data columns. Finally, a variety of imputation columns indicate which answers were imputed in the data.

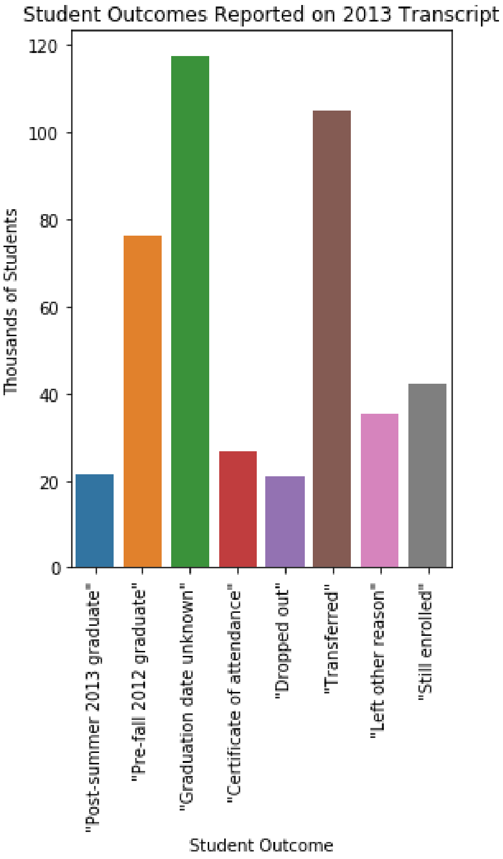
All negative numbers are different types of missing data, I converted these to nan values to more easily process them. I can go back to the original files if I need to know exactly why each value is missing.

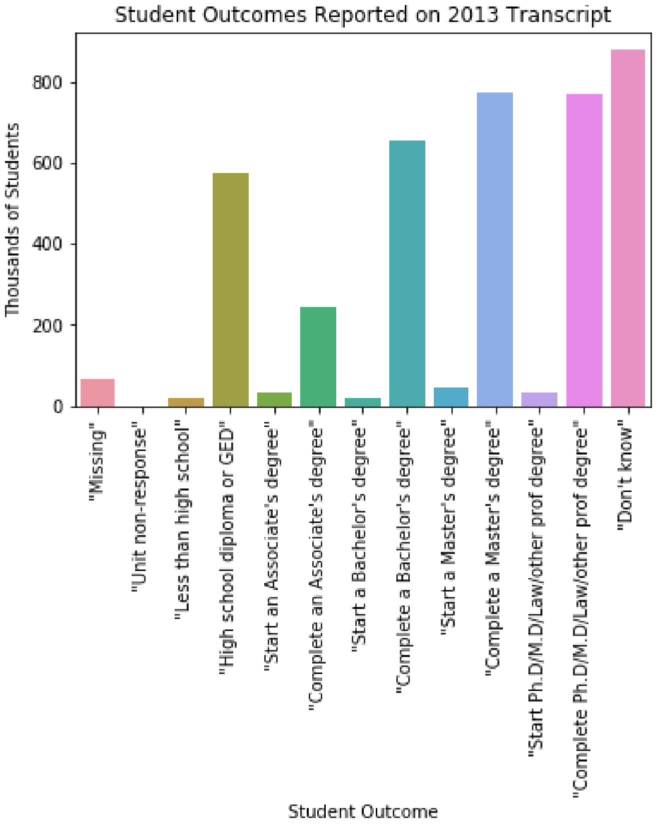
I also split the data into four different data sets for the four different time frames of the study. When I build my statistical model, and make comparisons between columns, I want to keep data that was collected at different times separate, particularly when it is time to select features for a statistical model.

I created a dictionary from the documentation so that I can type number_labels['S1ABILITYBA']['4'] and obtain "Definitely" for use as a label in a graph or figure. Similarly, I include the description of the variable under the key 'desc'.

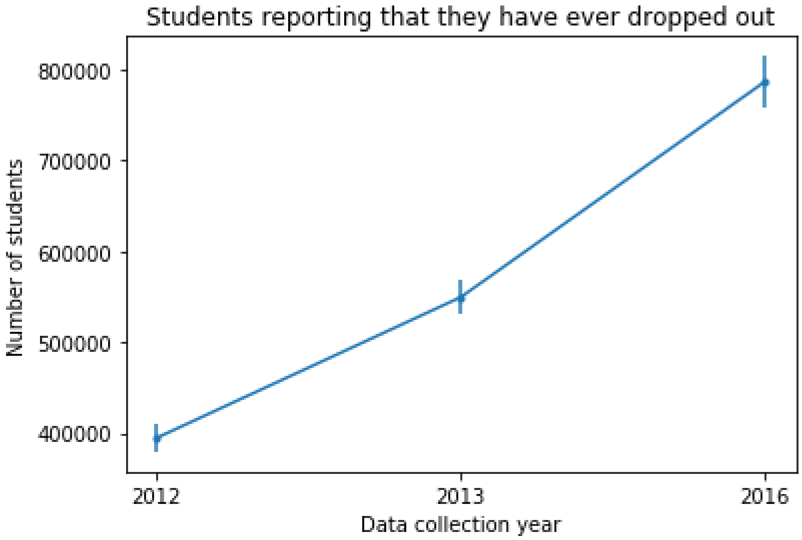
There are several variables that indicate the students' dropout status and whether or not they have ever dropped out. This variable was chosen as the label because it indicated the outcome of the student as stated on their transcript four years after the ninth grade study. Other columns may be removed later due to high (trivial) correlation with the label.

Looking at some initial findings I was looking at the breakdown of student outcomes, as well as the outcomes that students themselves predicted. The students who graduated are removed from the first plot, because, fortunately most students graduate, and the other categories become obscured. The next figure is what students believe their educational achievement will be. While a large number of students ‘don’t know’, most students believe they will complete high school, college, a masters degree, or even further. Few students believe they will complete less than high school, which implies that most students who drop out aren’t planning on it.

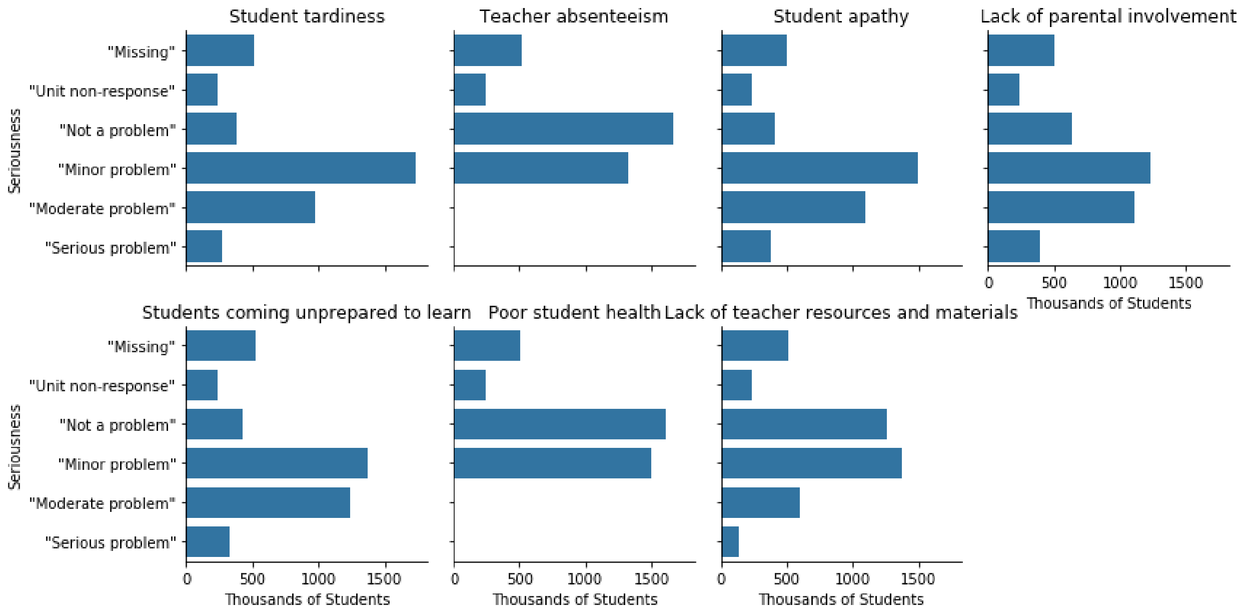




Students also may drop in and out several times, and not just once. This is evidenced by the fact that the number of students that has 'ever dropped out' is much larger than the numbers of students who are listed as dropping out on their transcripts. Over time, this number becomes very large by the end of the study, which is seven years after the freshman year of the students.



Finally, the principals at the schools involved in the study reported problems that they feel affect the school. Minor problems include 'student tardiness' and 'teacher absenteeism'. While 'students coming unprepared to learn' and 'Lack of parental involvement' warrant more serious problems as assessed by the principals.



Looking at statistical tests, I found several categories to test whether there were differences between drop out and non-drop out groups. Here, males are shown to be a much larger group than females within the dropout category. Males are 17 percentage points higher in the dropout category while females are 17 percentage points lower.

Moving onward, we look at the question ‘Do you feel safe at your school’. For this one test, all but one category had a difference between the proportion of students answering this question a certain way depending on if they were in the 'drop out' or 'non-drop out' category. The strongest effects were that those who 'strongly agreed' that they feel safe at the school were 14% percentage points less likely to be in the drop out category, and those who 'disagreed' that they feel safe at school were 14% percentage points more likely to be in the drop out category. The remaining categories were less than a 1 percentage point difference.

Drop out:

	category	estimate	s.e.
1	"Strongly agree"	0.163576	0.003473
2	"Agree"	0.600024	0.011062
3	"Disagree"	0.221125	0.006343
4	"Strongly disagree"	0.015275	0.001104

Non-drop out:

	category	estimate	s.e.
1	"Strongly agree"	0.304911	0.000508
2	"Agree"	0.595092	0.000496
3	"Disagree"	0.077045	0.000265
4	"Strongly disagree"	0.022952	0.000141

Stat summary:

```
   difference    t_statistic        p-value
1   -0.141335    -40.269371         6.541278e-64
2    0.004932    0.445407           3.284943e-01
3    0.144080    22.696372          0.000000e+00
4   -0.007677    -6.894834          2.465504e-10
```

Significance is indicated by p<0.0125

Next steps would be...