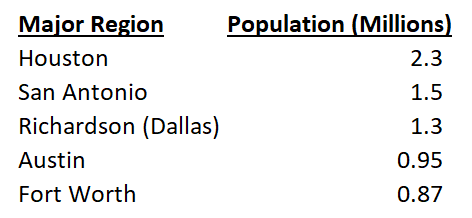
Forecasting Likelihood of College Graduation in 4-Years Based on School District Features

Capstone 1: Report

***“We Would Like Our Child to Attend College. Do You Know Which School Districts Improve the Likelihood of this Happening?”***

Many parents across the country find themselves asking each other this exact question. Their motive is simple, they are looking to provide their child with a quality education. You hear it all the time, education is key! It has been proven time and time again that a quality education can unlock doors for the average individual. It’s common belief that having your child attend college will lead to a more comfortable life, which is what all concerned parents want for their child transitioning into adulthood.

In this study I aimed to help parents considering the move to or within the “major regions” (listed below) of Texas. These particular regions were chosen as they represent the areas with the most economic opportunity, making them prime locations for families and talented educators. The information provided is strictly educational in providing parents with “food for thought” ahead of a potential move. Parents may even use the data to rule out a move entirely should their current school district be favorable among other options.



In order to achieve this goal, we first must address the common question from parents mentioned at the top of the page. While this question has good intentions, I don’t necessarily believe this is the right question to ask for your child. Parents need to ask “which school districts improve the likelihood that my child will be able to earn their degree in college” instead of simply “which school districts increase the likelihood of my child getting into college?”

The honest truth is that there are many colleges willing to accept your child even if they had poor college admission test results. Why is this the case? Colleges are operated as businesses. They will gladly collect expensive tuition checks until the student fails out (not prepared) or earns a degree (prepared student). Depending on the college, there’s often a considerable dropout rate for freshman who have come in unprepared.

In a 2011 Harvard study “Pathways to Prosperity”, the U.S. contained the highest college dropout rate among industrialized nations. Among four-year colleges, just 56% of students graduated within six years (not four). One must remember that more time spent at a college means more money spent or more debt accumulated. Financial pressure and academic disqualification remain the top two reasons why a student drops out of college.

To answer the college graduation question, historical data was collected on the percentage of students who were able to graduate college in four years’ time after graduating high school for each respective school district. School district features were also collected to measure their effect on the resulting college graduation percentage. Utilizing all the historical data for the major regions, I aimed to forecast each school district’s features/test results and the resulting likelihood of a student earning their degree in four years’ time. Later in the report I will go into more detail about the chosen school district features, but for now, I’ll provide a brief summary as to what they are in case the reader has not seen them before.

The SAT and ACT are both college admissions tests that indicate a student’s readiness for college. Thus, you can see their importance for this study. The SAT includes a Math and a Reading/Writing section that are both scored out of 800 points for a maximum possible score of a 1600. The ACT includes a Math, Reading, English, and Science section which are all scored out of 36. A student’s composite score on the ACT is the average score of all these sections.

AP exams are taken at the end of the school year for those enrolled in AP classes in high school. These AP classes are comparable to introductory college classes and cover several subjects like calculus, physics, literature, biology, and psychology, for example. At the end of the class, students have the option to take AP exams for college credit. The tests are scored out of five with most colleges nationwide accepting a passing score of a three. In this study we will use a three as a passing score translating into college credit.

Though it doesn’t represent our main concern in assessing the quality of a school district for a student, the percentage of students who enroll into college after graduation is an important feature to explore. Sure, part of this feature is influenced by the wealth of parents residing in the district, but it also can be influenced by impressive test scores that increase the amount of scholarship money offered by colleges to students.

A family’s financial state does not only affect college enrollment, but also its ability to reside in a particular school district. Housing prices/taxes in the respective school districts are not all created equal, as demand can be higher for certain areas. This demand is affected by several features, but a major one is certainly the quality of education. I acquired each district’s historical wealth per average daily attendance (“Wealth/ADA”) data for later analysis.

I find it important to clarify how college enrollment percentage and college graduation percentage is assessed for a particular year. It’s also important to note that this study strictly focused on Texas colleges and the statistics do not include students who enrolled into colleges outside the state. If we say the year is 2013, then the percentage of students who enrolled into colleges in the fall of 2013 is denoted as “Enrolled 4-Year (%)”. The percentage of those students who were able to graduate four years later is denoted as “Graduated 4-Year (%)” and represents the target variable I attempted to forecast.

Now that we understand the desired data for each school district, let’s go over how I collected and wrangled it for later analysis/modeling.

**Data Wrangling:**

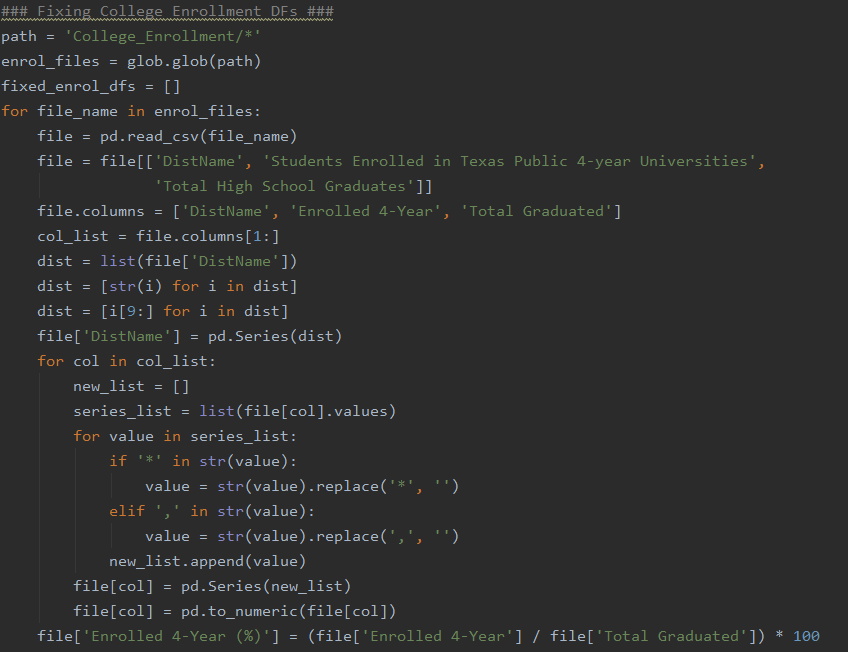
Yearly SAT, ACT, AP exam, and Wealth/ADA datasets were downloaded from the Public Education Information Management System (“PEIMS”) on the Texas Education Agency’s website. At the time of this project, the latest data out was from the class of 2017 and the earliest was from 2011, resulting in seven years of full historical data for each respective school district. Having the data from before 2011 would be nice, but I also do believe there’s a sense of “what have you done for me lately” that goes into assessing a school districts value to a student.

The datasets on college enrollment and college graduation were downloaded from the Texas Public Education Information Resource (“TPEIR”) website. It’s important to note that this data strictly focuses on Texas colleges.

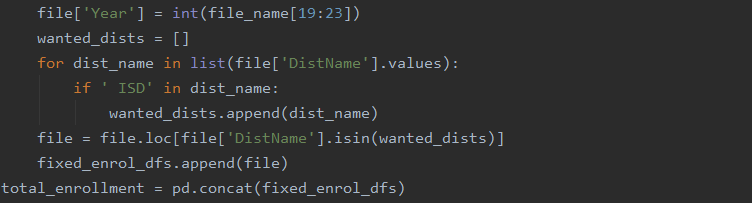
From there I needed to recognize what variables were important in each dataset and remove the unnecessary ones using my knowledge/experience of once being a high schooler in Texas. The ACT, SAT, and AP datasets all contained breakdowns by ethnicity for scores, but I filtered these to simply be “all students” for each respective district. The SAT and ACT datasets contained average scores (“SAT-Total”, ”ACT-Composite”) for each school district and participation data was also included in both. The AP dataset contained data on participation, amount of AP exams taken in each respective district, and the amount of passing exams (scored 3 or above).

For each dataset type mentioned above, I created a folder containing the separate yearly datasets with the goal of later creating one “Total” dataset (Ex: Total\_AP) containing respective district data for 2011 – 2017. With the different “Total” datasets in place, I could merge them into one final dataset containing all the district features. Before being able to merge all the data and perform analysis, the data needed to be cleaned and wrangled to make my life and the readers life easier later on.

First up were the yearly datasets on college enrollment. The main problems here were that district names contained an ID number and name in all caps (Ex: 4825170 KATY ISD). The ID number is not really necessary so I got rid of it to leave the District name in all caps. I also needed fix numerical data that contained “\*” (data not available) or was represented as a string with a comma (Ex: 1,244). From there I was able to calculate the percentage of graduating students who were able to enroll into a four-year college.

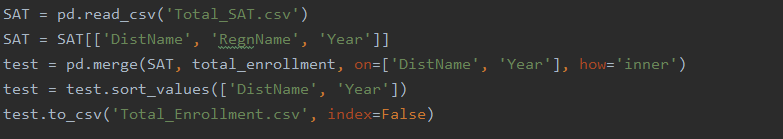


The respective years were added to each dataset and public districts were extracted using district names that contained “ISD” (Independent School Districts). The datasets included academies/prep schools, but I wanted to strictly focus on public school districts.

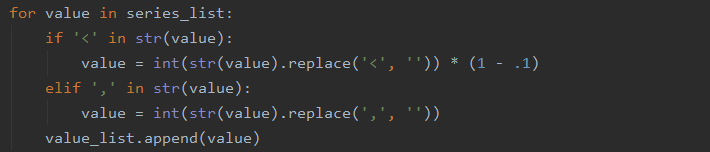


Once each year’s enrollment dataset was in the correct format, it was added to a list of enrollment DataFrames to later be concatenated into one total (2011 – 2017) DataFrame. To spare the reader from having to look at too much code, this relative process was repeating for each dataset type.

Unfortunately, the enrollment data also did not contain each district’s respective region name, which was important in meeting my desire to focus on the major regions. Luckily the SAT datasets contained this data and they were utilized to perform an inner merge to provide the enrollment dataset with respective region names.



The yearly AP datasets contained numerical approximations in the form of strings (Ex: <60) and some instances of the string with a comma problem. I decided to be consistent in decreasing the number by 10% for each of the “less than” cases.



Upon fixing the numerical data, I was then able to add the number of AP exams taken per student in each district for each year.



Besides filtering for public schools in the major regions, I also decided to remove districts that did not take more than 50 AP exams. The removed districts represent small ones with very limited data. Rather than play a guessing game for the features of these small districts (which has a large effect on statistics involving percentages), we can just exclude them from our analysis as families very rarely decide to move to these districts for educational purposes.



With all the DataFrames containing yearly data for public school districts in the major regions of Texas, I was then able to merge them all into one DataFrame containing seven years of full historical data for the features I deemed important.



Seven years of historical data for a school district was required as I wanted to forecast each respective district’s features for 2018, 2019, and 2020 as accurately as I could. With the forecasted features, I could then proceed in forecasting the target variable, the percentage of student who will graduate college in four years. This, of course, could only be done after gathering as much historical college graduation percentage data as I could.

For the college graduation data, I manually inputted all of the available district data for the years of 2011, 2012, 2013, and 2014. Remember, this is the percentage of students who went on to graduate college in four years’ time. Some districts did not contain data for all four of the above-mentioned years. That didn’t matter in this particular point in the study as I needed as much historical feature and target data as possible to train a predictive model.

Before getting into the machine learning section of this study, we can get to know the district features a bit better through some data analysis on the seven years of historical data.

**Data Analysis:**