Springboard DSC:

Capstone Project 1

*(Predicting the Percentage of Students Who Will Graduate College Within 4-Years, Based on the Features of the School District They Attended High School in)*

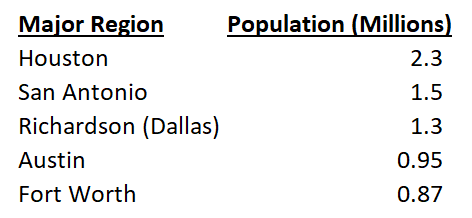
By: Rachid Rezzik

March 2019

***“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 a 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 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 should instead to ask “which school districts are proven to contain a higher percentage of its graduates earn a college degree within four years?”

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. This is why I chose to focus on four-year graduation, as that should be the goal.

To answer the college graduation question, historical data was collected on the percentage of students who earned a college degree within four years’ time after graduating high school from a particular school district. For each school district, their historical features were also collected to measure their influence on the resulting college graduation percentage.

Utilizing this historical data, I aimed to build a predictive model to estimate the percentage of students going to college (from a particular school district) that will graduate within four years’ time. Parents/students may use this model to explore the historical and predicted college graduation percentage for a particular district and class year, view different district results within a particular region they are considering a move to, or predict graduation percentage from a new school district who’s only been in existence for less than four years.

Before we move further, I find it important to clarify how college enrollment percentage and college graduation percentage is assessed for a particular high school class year and to note that this study strictly focused on Texas colleges. 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 (%)” and the number of students enrolling is denoted as “Enrolled 4-Year”. The percentage of those students who were able to graduate four years later (2017) is denoted as “Graduated 4-Year (%)”, the target variable I attempted to predict.

Below is a link to my project deliverables folder on GitHub that contains all the code utilized in this project.

(The “Data\_Wrangling”, “Data\_Analysis”, and “Machine\_Learning” jupyter notebooks)

Link: <https://github.com/RachidRezzik/Texas_GradPercent_Predictor/tree/master/Project_Deliverables>

It’s encouraged that the reader opens the respective notebooks to follow along throughout the report for more details.

**Data Wrangling**

SAT, ACT, AP exam, and wealth per average daily attendance (“Wealth/ADA”) datasets were all 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 the class of 2011, resulting in seven classes of full historical data for each respective school district. Having the data from before the class of 2011 would be nice, but parents also tend to have a “what have you done for me lately” approach when assessing a school district’s value. The datasets on college enrollment and college graduation were downloaded from the Texas Public Education Information Resource (“TPEIR”) website. Again, it’s important to note that this data strictly focuses on Texas colleges.

With all the datasets in place, I needed to recognize what variables were important in each 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).

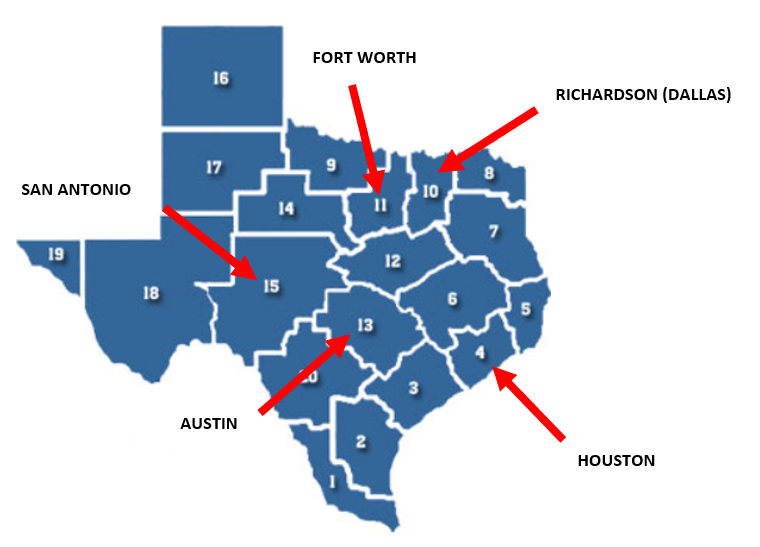
For each dataset type mentioned above, I created a folder containing the separate class years. I aimed to later create a “Total” dataset (Ex: “Total\_AP”, “Total\_SAT”) with all the respective district data for 2011 – 2017. With the different “Total” datasets in place, I could merge them into one final 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.

First up were SAT and ACT scores. “ACT-Composite” scores for each class year were suitable, but “SAT-Total” scores for the class years of 2011 – 2016 needed to be altered. CollegeBoard introduced a new scoring system out of 1600 (previously out of 2400) in 2016. Class of 2017 scores already included the new scoring system and did not need to be altered. Using CollegeBoard’s concordance tables, I adjusted the old SAT scores to be equivalent to the new as this would be crucial in using SAT scores as a feature to train a model that can predict the target.

Public school 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 as they are far more accessible to the average family. In order to maintain the focus on the major regions, the datasets were all sliced to only include the respective regions as well. After each class year dataset was cleaned, it was appended to a list to later be concatenated into one total (2011 – 2017) DataFrame. This process was used for all the dataset types (SAT/ACT/Enrollment/AP/Wealth/ADA).

Next, I needed to clean the datasets on college enrollment. The main problem was that district names contained an ID number and a name in all caps (Ex: 4825170 KATY ISD). The ID number is not 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 high school students who were able to enroll into a four-year college that fall.

I noticed that 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. To give the reader a visual perspective of the size of the major regions that districts belong to, I have provided the image below. This is how the Texas Education Agency splits up the educational regions for its data.



Luckily for me, the “Total\_SAT” dataset contained which region each district belonged to. I performed an inner merge with the “Total\_Enrollment” dataset to obtain the respective region names for each school district. This solution was also utilized for the “Total\_Wealth” dataset.

For the AP datasets, numerical approximations in the form of strings (Ex: <60) and some instances of the string with a comma problem were present. 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), they were excluded as families very rarely decide to move to these districts for educational purposes.

With all the respective DataFrames containing class of 2011 – 2017 data for public school districts in the major regions of Texas, I was then merged them all into one DataFrame (“Feature\_Target\_Data”). Seven years of historical data gives parents a good idea of what they can likely expect from a particular school district’s features going forward. For example, if parents want their child to earn as much scholarship money as possible, then they certainly should be interested in which school districts historically averaged the highest SAT and/or ACT scores.

For the college graduation data (the target), I manually inputted the number of students who were able to earn their college degree within four years of 2011 – 2014 (what was available). This can be read as the number of students who earned their college degree in 2015 – 2018 that belonged to the high school classes of 2011 – 2014.

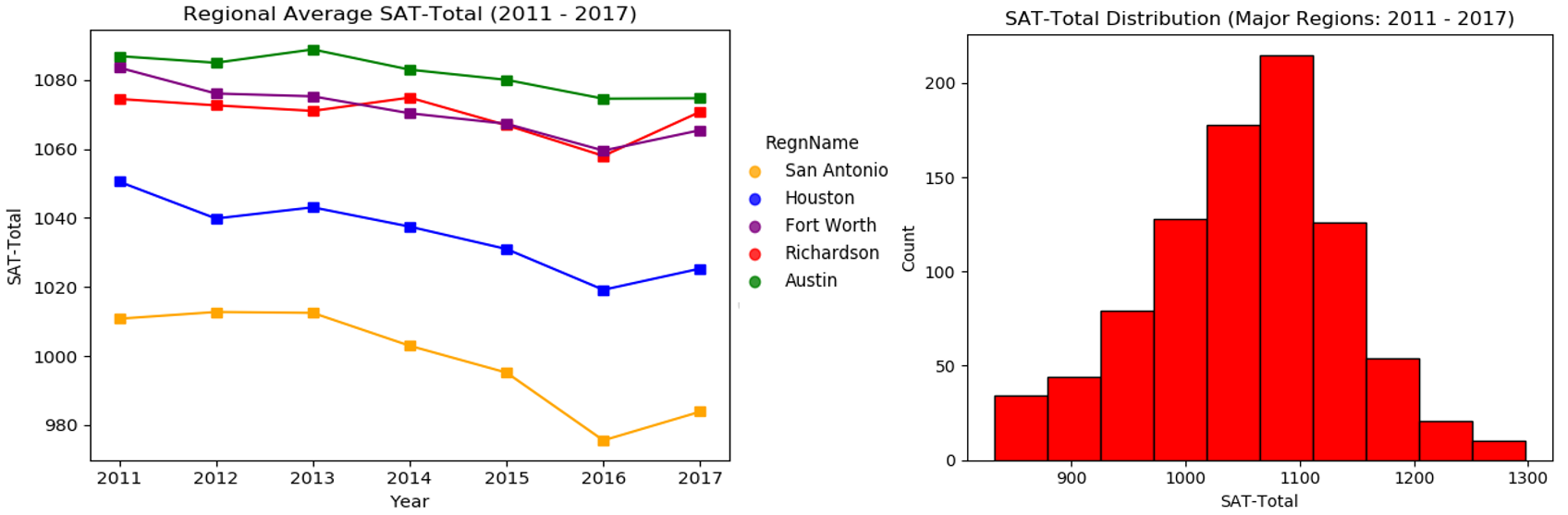
With the number of students graduating college within four years and the number of students who enrolled into college, I was then able to calculate the percentage of students who were able to earn their degree.

Following all this data wrangling/cleaning, I was left with a dataset that included all the feature and target data for the classes of 2011 – 2014. For 2015 – 2017, only the feature data was available and the target data (college graduation percentage in 2019 – 2021) is unknown. Later, I will attempt to predict the target for these class years after establishing a satisfactory machine learning model.

**Data Analysis: School District Features**

*(SAT)*

Let’s start of by viewing the regional averages for each class year and the distribution of district-level scores.

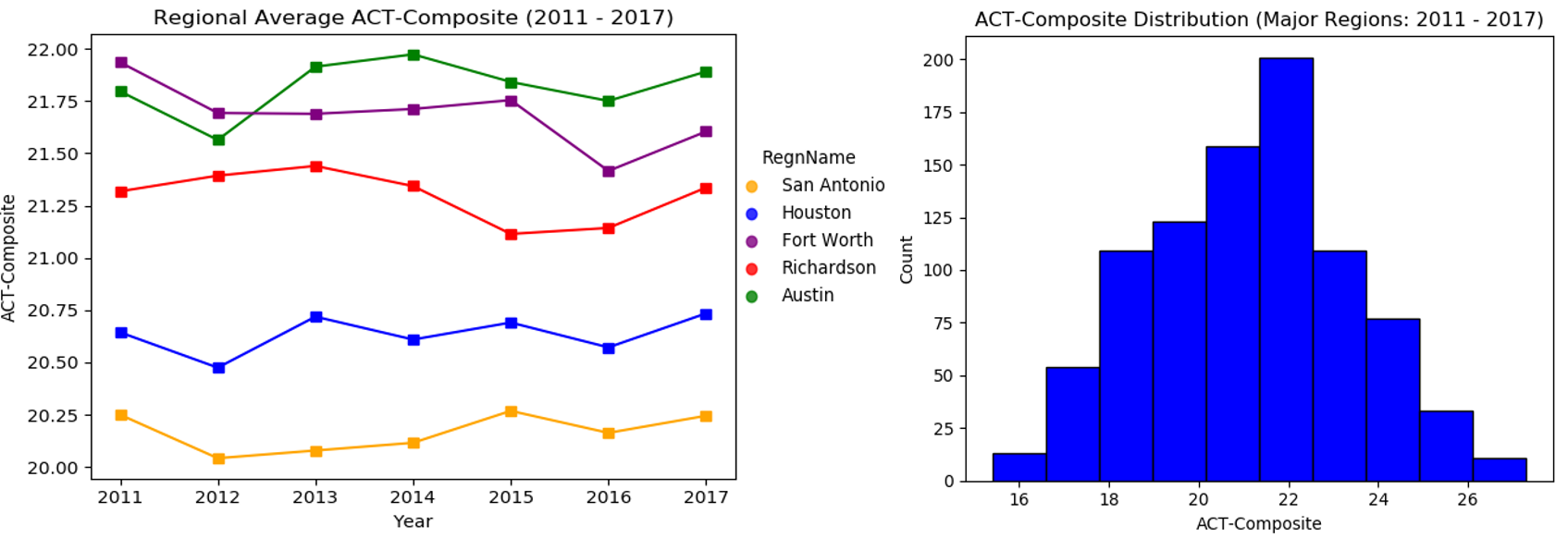


Mean District-Level SAT-Total (Out of 1600):

1052.8

We can see that Austin performed the best while Fort Worth and Richardson were neck and neck. The gap between the top region and San Antonio is quite large. This will, unfortunately for the region of San Antonio, remain a common theme throughout this analysis.

*(ACT)*

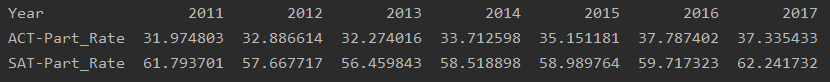


Mean District-Level ACT-Composite (Out of 36):

21.1

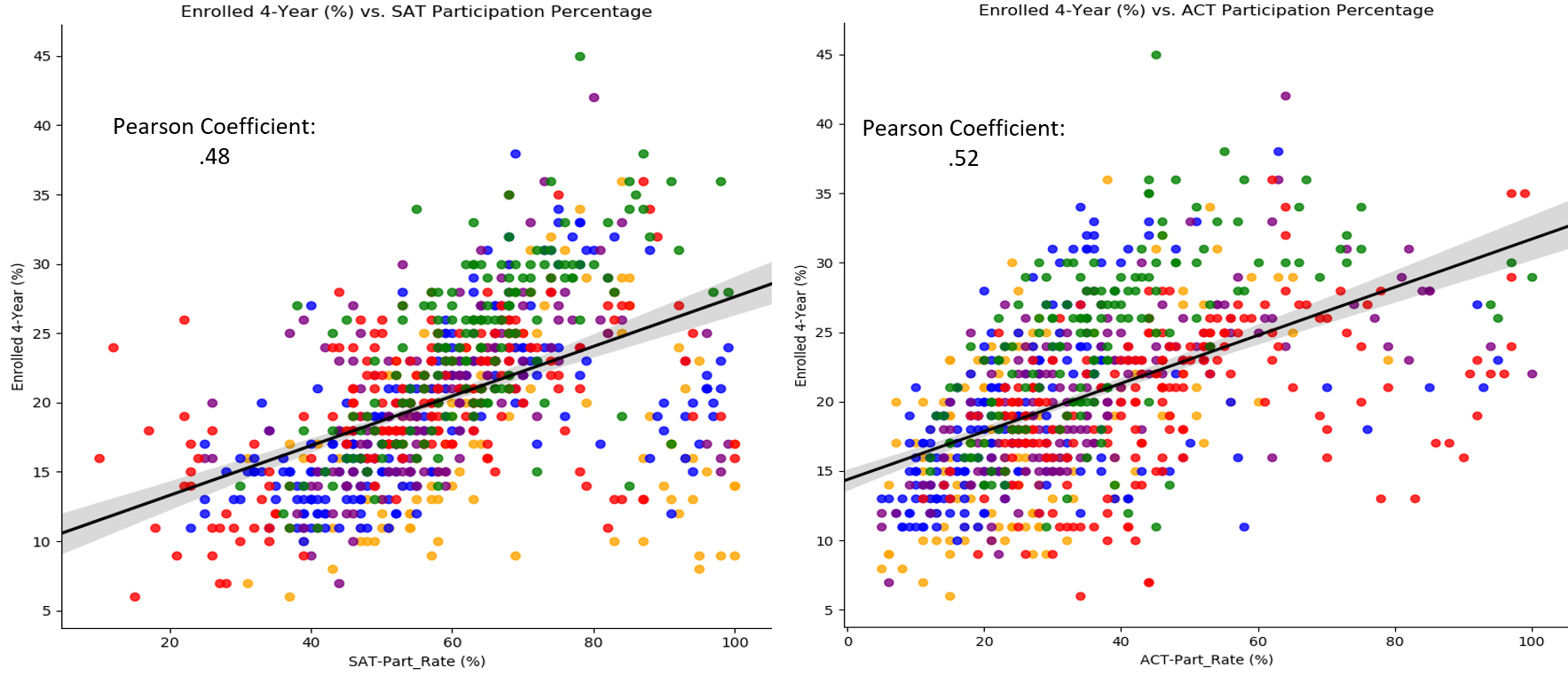
It appears that the districts located in Austin and Fort Worth achieved the best average scores on the ACT with Richardson not too far behind. As with SAT scores, there appears to be a slight uptick in the average scores for the class of 2017 (improved educators/smarter kids?).

*(SAT/ACT Participation %)*



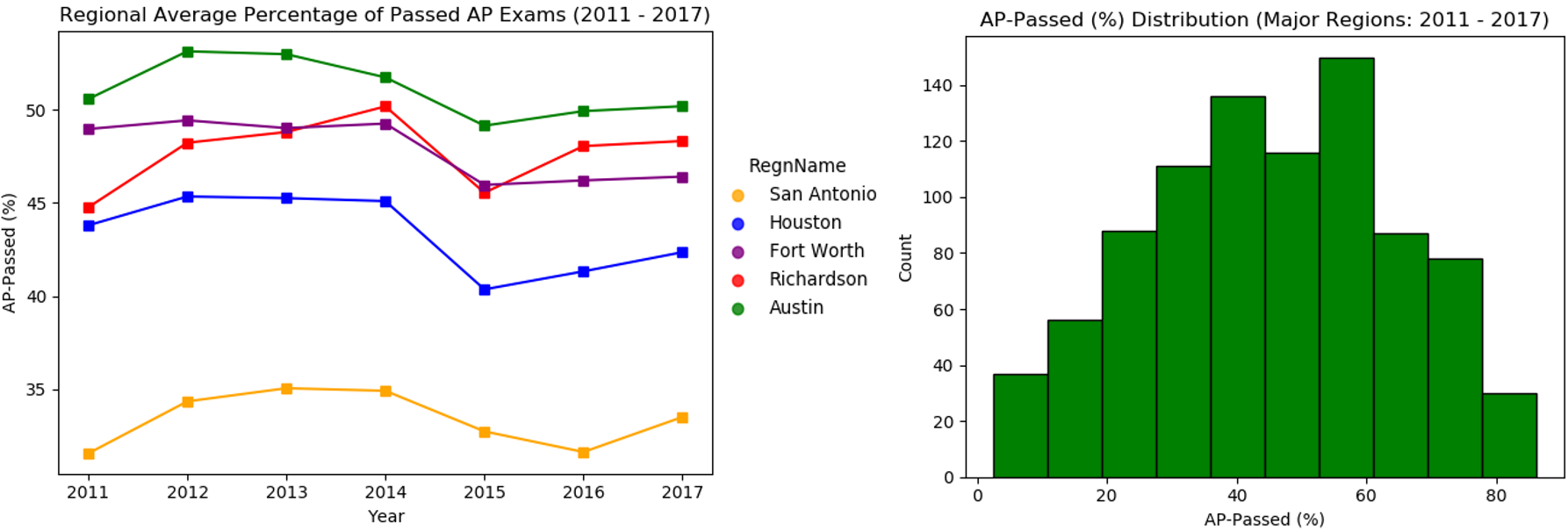
Something interesting I discovered is that more students consistently choose to take the SAT than the ACT. Why is this? Well… it’s very tough to know for certain without surveying high school students across Texas, but I’ll provide a possible theory. The SAT is the test students hear the most about while growing up from friends, parents, and teachers. You could say it’s the “standard test” students feel they have to take in high school.

It’s good to see in the later years that ACT participation grew, but it’s still quite concerning that students are not more encouraged to take the ACT as well as the SAT. Taking two tests instead of just one immediately increases the chance of a student doing well on at least one. Doing well on one of the college admission tests is all it can take to get into college and earn scholarship money. With scholarship money, enrolling into college becomes more likely/possible for a student. As you can see below, both SAT and ACT participation rates contain similarly positive correlation with Texas college enrollment percentage. Why not take both??



*(AP Exams)*

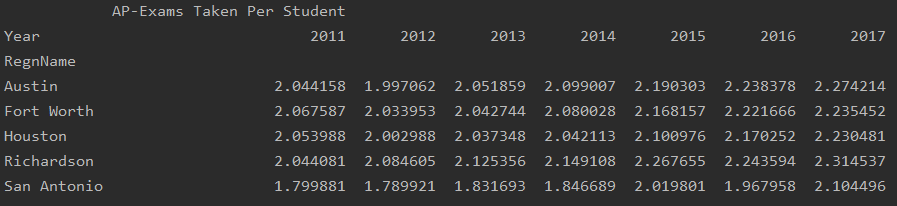
Note: In this study, a score of three or above on an AP exam is considered passing. Most colleges accept this score.



Mean District-Level AP-Passed (%):

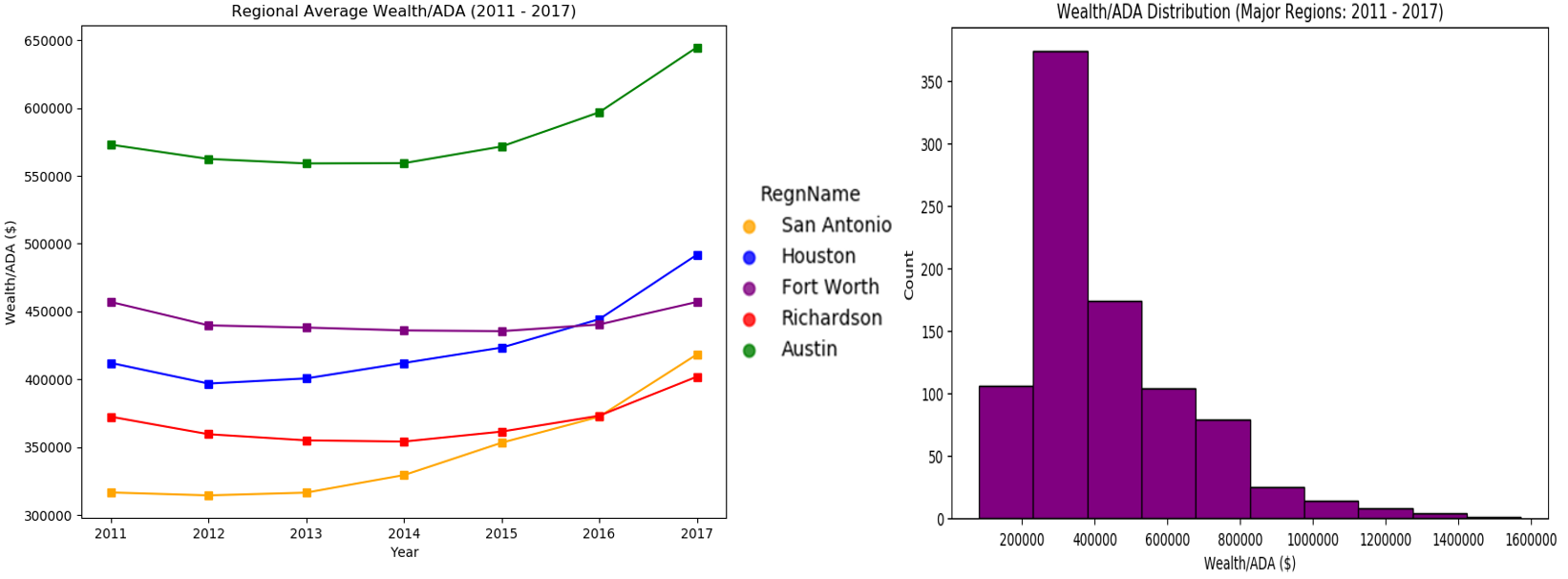
45.1

The regions of Austin, Richardson (Dallas), and Fort Worth appear to contain the best passing percentages. I found it interesting to also take a look at the availability of AP classes to students. One could argue that more availability to AP classes would result in a student being able to take more exams and earn more college credit/gain more college-level exposure.



From the standpoint of AP class/exam availability, students attending high school in Richardson (Dallas) had the most opportunity to students to earn college credit. Austin tends to offer the second most opportunity, which is quite impressive when also taking into account that Austin contained the best passing percentage.

*(Wealth/ADA)*



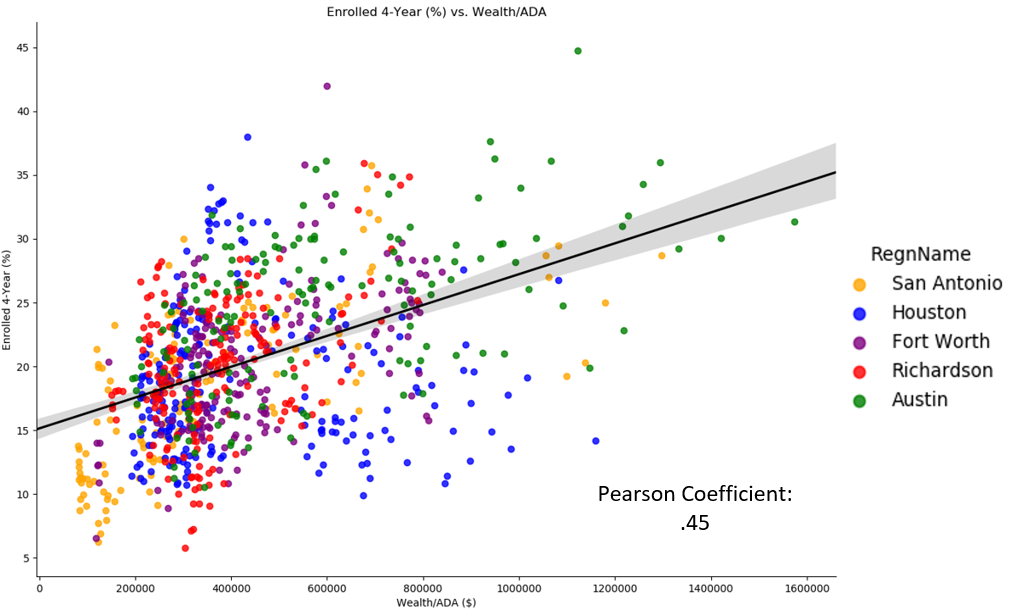
Mean District-Level Wealth/ADA ($):

431,718

Austin held a healthy lead over its nearest competitors in Fort Worth and Houston. Overall, it appears that Wealth/ADA has been increasing in recent years. Something we could later choose to explore is the average property tax for homes in each region.

*(Wealth/ADA Effect on College Enrollment)*

It should be interesting view how Wealth/ADA correlated with college enrollment percentage for the different class years as well. It would not be unreasonable to hypothesize that a student’s ability to enroll into college increases with wealth.

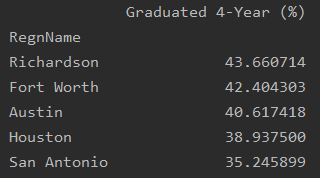


Though it certainly isn’t the only factor influencing college enrollment percentage, Wealth/ADA does indeed contain a positive correlation with college enrollment percentage. As mentioned before, this can partially be attributed to “wealthier” students having more access to college. Let’s say a student has poor test scores, but comes from a family with money. There are several colleges that are still willing to accept this student. Even if the student ends up failing out, the college will still be entitled to tuition payments.

With that being said, we still have to acknowledge that there are plenty of students in the districts containing higher Wealth/ADA that are taking advantage of the quality education available to them in testing well and earning scholarships/being more attractive to colleges. Though I can’t necessarily prove it with data right now, it’s not egregious to assume that many of the parents in these “wealthier” areas are well educated and therefore, encourage/press their children to do well in school as they know firsthand what education can bring to an individual’s life.

**Data Analysis: Percent of Students Earning a College Degree Within Four Years**

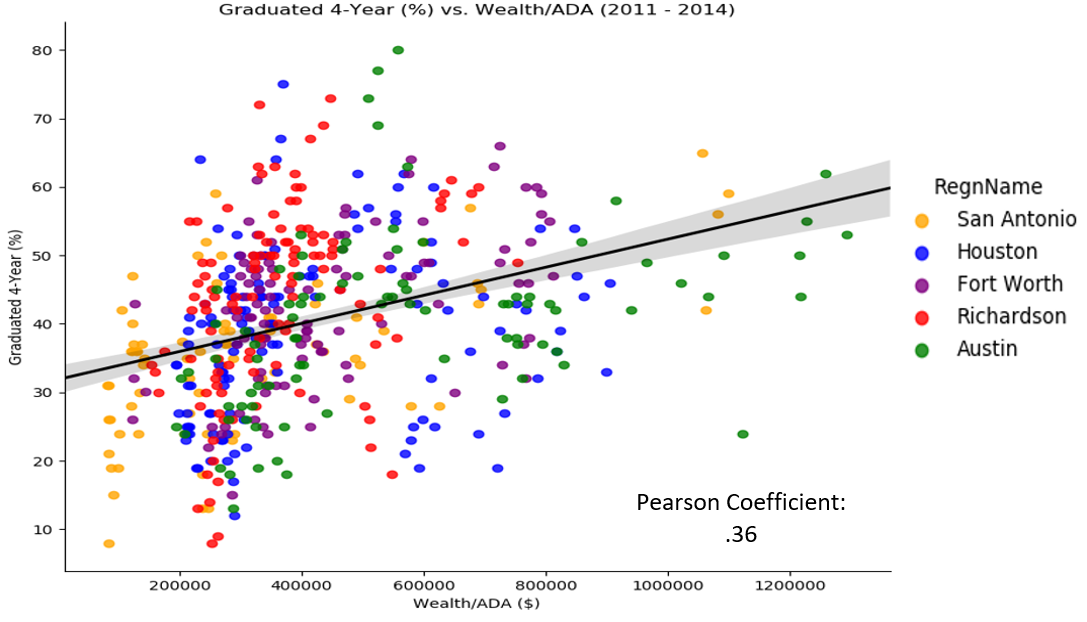
For the classes of 2011 – 2014, let’s look at the average percentage of students who were able to earn their college degree within four years by region.



Richardson (Dallas) contained the highest percentage of students who were able to earn their college degree within four years. There’s roughly an 8% difference between Richardson (Dallas) and San Antonio, which isn’t too much of a surprise as San Antonio, on average, contained poor school district features.

*(Wealth/ADA Effect on College Graduation %)*

As an exercise, let’s assume that most of the students from the wealthier school districts were able to enroll into college simply because they can afford it (there’s a college out there willing to accept them for their money, despite low scores). For the less fortunate areas, we’ll say the students who enrolled into college most likely earned scholarship and therefore, had a greater chance of earning their degree (more prepared). With these assumptions in place, we could then expect to see a very poor correlation between Wealth/ADA and those who graduate from college within four years. To test our assumptions, let’s take a look at the actual data from the classes of 2011 – 2014 (college graduation year: 2015 – 2018).



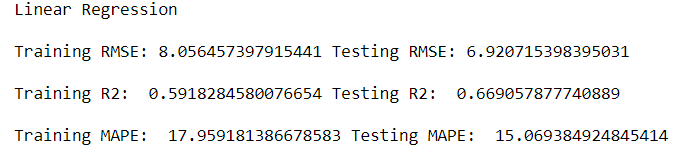
The correlation may be lower than the one for college enrollment percentage (pearson coefficient: 0.45), but the difference is not as extreme as one would expect from the assumptions made in the exercise above. As with college enrollment percentage, Wealth/ADA is not the only feature determining the percentage of students earning their degree within four years. We’ve already been presented with several other features that contain their own respective influence on college graduation percentage. These features will work together, with some having more influence than others, in predicting our target variable.

It’s also important to consider that our school district features contain some influence on each other. This could present some multicollinearity issues in building a regression model to predict our target. Some of the included features are a bit redundant or were used to engineer other features (mentioned in the data wrangling section). Despite some of these features being dropped in building my final model, they are all included in the complete feature-target dataset as information for those who are curious.

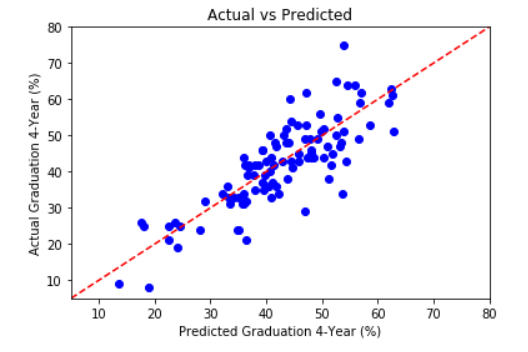
**Machine Learning:**

(*Model 1: Linear Regression*)

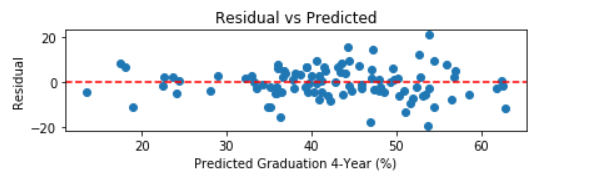
Utilizing sklearn’s Linear Regression, I trained the model with an 80 – 20 split (my dataset was on the smaller side with 508 records) and a random state equal to 42 (will keep this consistent for model performance comparison). To evaluate how the model performed on the training and test splits, I printed out the respective r-squared (“R2”), root-mean-squared-error (“RMSE”), and mean-absolute-percentage-error (“MAPE”) experienced.

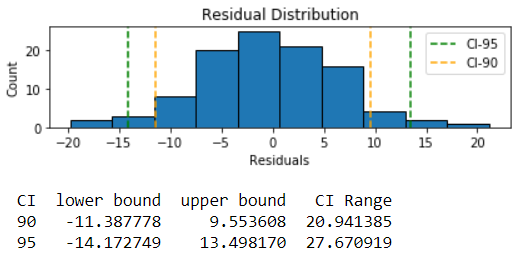


With the performance of the test set, I didn’t deem it necessary to include a Ridge or Lasso regression as the model did not appear to be overfitting the training data. To get a better visualization of my base model’s performance on the test set, I printed out an “Actual vs. Predicted” scatterplot.



Continuing my performance evaluation on the test set, I also printed out a “Residual vs Predicted” scatterplot and the residual distribution. The distribution includes confidence intervals of 90% and 95%, with upper and lower bounds being specified in the table provided below. With these plots and table, I was able to get a better feel of the residuals I could anticipate when utilizing this particular model in predicting the target.



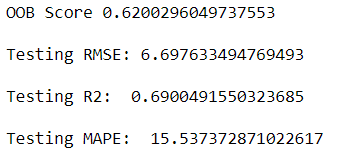


(*Model 2:* *Random Forest Regressor, Without HyperParameter Tuning*)

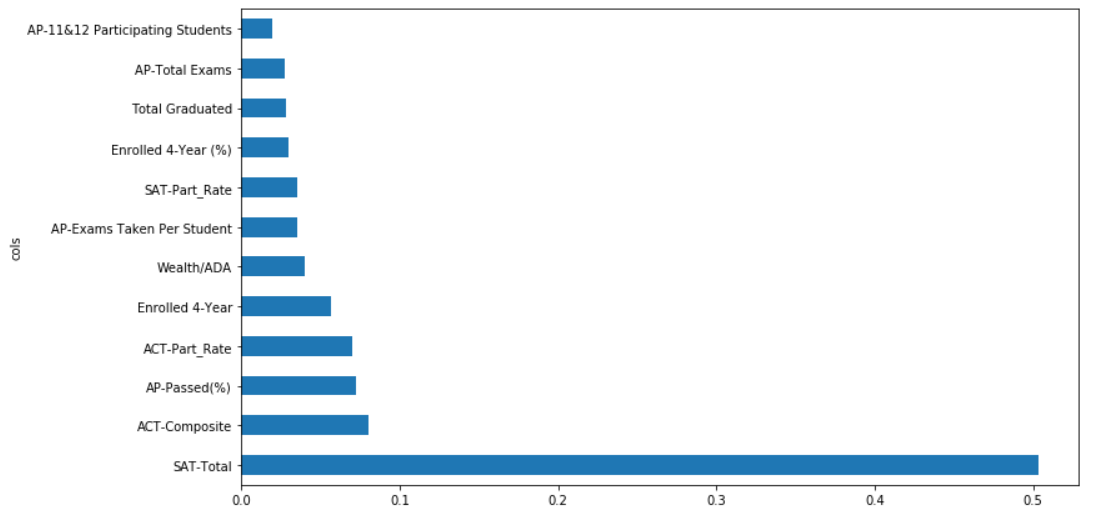
The Linear Regression model above served as a base model that I aimed to improve on. To do so, I chose to utilize a Random Forest Regressor (“RFR”) and evaluate the resulting performance in comparison to the base model. I believed the RFR would help solve my concerns with the size of the dataset when generating splits, adding a further element of randomness to prevent overfitting.

The number of features that can be split on at each node is limited to some percentage of the total, which also helps to ensure that the ensemble model doesn’t rely too heavily on any individual feature. This makes fair use of all potentially predictive features and helps the issue of multicollinearity.

In training my first RFR, I did not tune any parameters and used all features. I also specified the use of 1000 estimators (number of decision trees in the random forest). The resulting RMSE, R2, and MAPE on the test set are provided below along with the out-of-bag (“OOB”) score. In the implementation of the RFR algorithm, each tree is trained on roughly 2/3 of the total training set. As the forest is being constructed, each tree can then be tested on the data not used in building that tree. This results in an OOB score that we can use as another comparison metric for model performance between RFRs.



In previously stating concerns about redundant features that don’t add much value to our model, I thought it would be a good idea to view the resulting feature importance plot after implementing the base RFR model.



In the next model I trained (Model 3), I dropped the bottom four features shown in the figure. These particular features are indeed redundant with some of the other features.

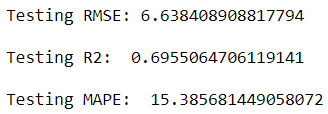
(*Model 3:* *Random Forest Regressor, HyperParameter Tuning & Feature Selection*)

In addition to the feature selection mentioned above, I also decided to tune the number of trees included in the random forest (“n\_estimators”), the maximum number of features to consider for splitting a node (“max\_features”), and the minimum number of data points allowed in a leaf node (“min\_samples\_leaf”) to improve on Model 2’s performance. I performed four-fold cross-validation on the training set and utilized GridSearchCV to provide me the best estimator (based on best average R2 score for all possible hyperparameter combinations). Below are the hyperparameters of the best estimator found.



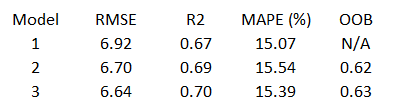
Utilizing the above hyperparameters, the model had the following OOB score and performance on the test set.





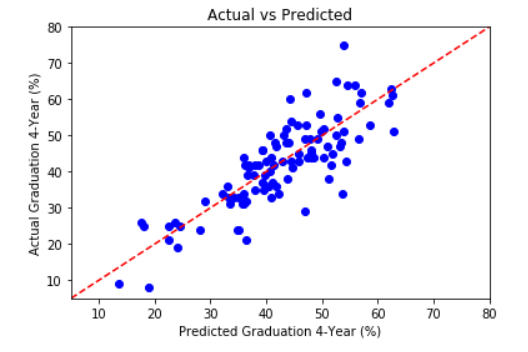
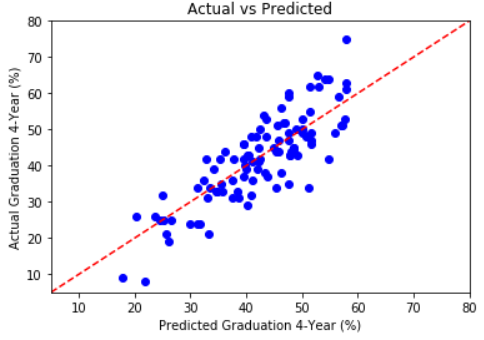
*(Model Selection – Summary)*

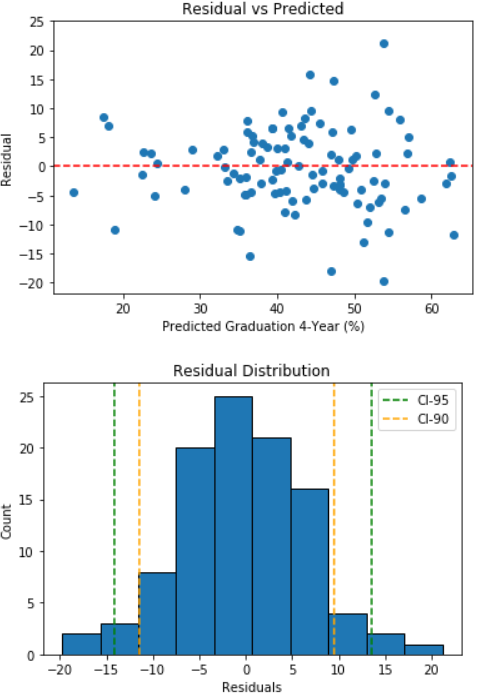
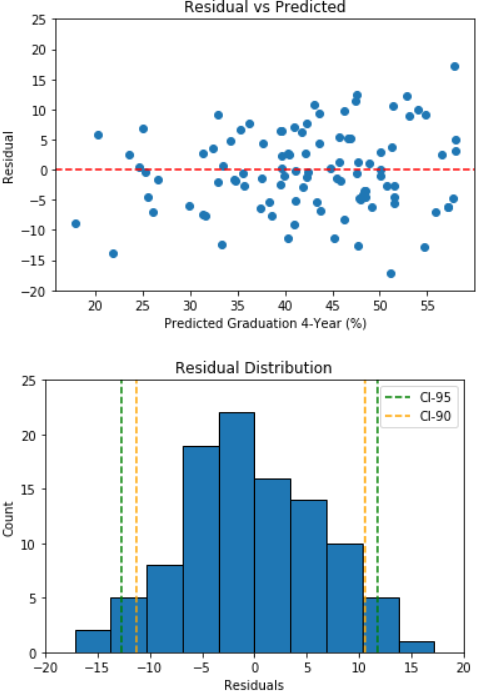
To make things a bit easier for the reader, I have summarized each model’s performance on the test set.

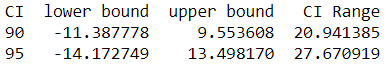
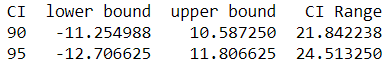


While these performance results led me to favor Model 3, it would also be a good exercise to compare its “Actual vs Predicted”, “Residual vs Predicted”, and residual distribution with the base model (Model 1).

Model 1 Model 3

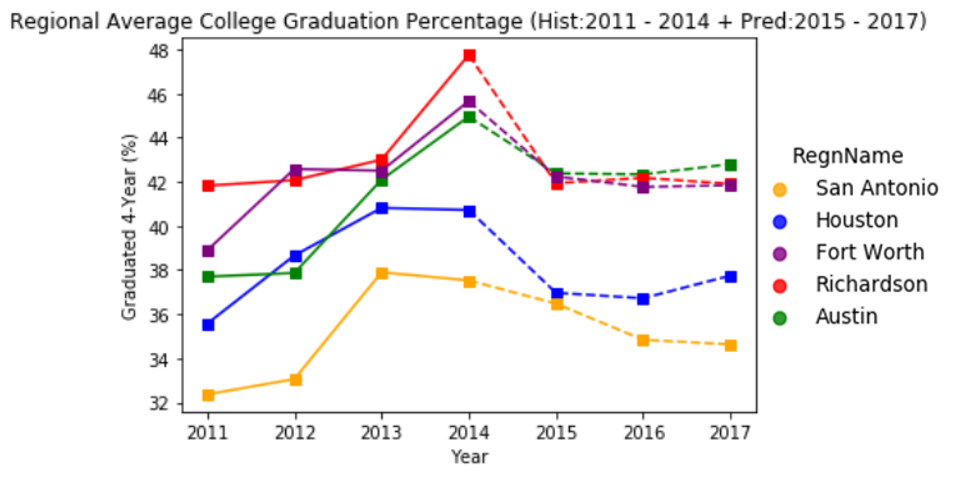
 

In attempting to predict the percentage of students who graduated within four years, I favor Model 3’s performance in decreasing the range of residuals, tightening the CI-95 range, and containing a more symmetric CI-90 lower/upper bound pair. While both models tend to underestimate the actual college graduation percentages a bit, it was more severe with Model 1. The range of residuals in Model 1 indicates to me that there are certain features/combinations of features that can cause greater residual. This is also apparent in the different R2 values for the two models, with Model 3 explaining a better proportion of the variance caused by the features in the regression model. These cases where the features/combinations of features increased the residual also helps explain why the 95% confidence interval range is larger compared to Model 3. While Model 1 does contain a slightly tighter CI-90 range, it is not as symmetric as Model 3’s.

I favor having the tighter confidence interval range for 95% of the residuals than 90%. With Model 1, the select problem cases mentioned above could misinform parents, swaying their interest towards or away from interest in a particular school district based on the poorly predicted college graduation percentages. The less symmetric distribution of residuals could also cause a larger difference in the prediction of two school districts with one being underestimated (which is more likely for Model 1, as shown by the left skew in its residual distribution & CI-95 range) and one being overestimated, unjustly influencing parents to favor one school district over the other.

*(Conclusions)*

If you remember from the data wrangling section, I collected school district features for the classes of 2015 – 2017, but not the actual target (2019 – 2021 college graduation percentage). This gave an opportunity to put my selected model to use in predicting all these missing graduation percentages (summarized below).



Looking at the high school class of 2015, we can see a predicted drop in college graduation percentage for all the major regions. When analyzing the previously provided feature importance plot, the top three features were “SAT-Total”, “ACT-Composite”, and “AP-Passed (%)”. For the class of 2015, regional-average SAT scores all worsened, with Richardson suffering the greatest decrease from its peak with the class of 2014. For ACT scores, the regions of Fort Worth, Houston, and San Antonio all improved while Richardson and Austin worsened. All regions also experienced a drop in the percentage of their students who passed their AP exams, with Houston and Richardson suffering the largest drops.

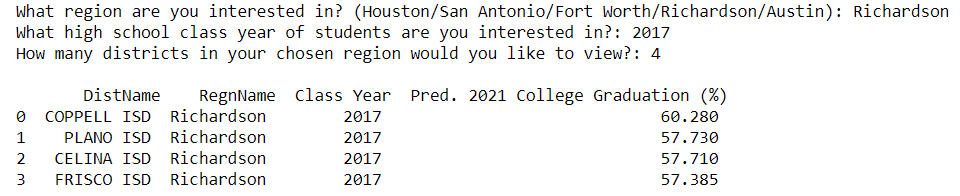
Also interesting is that the gap between Houston and San Antonio got tighter in the prediction for the class of 2015. Again, taking a feature importance point of view, the SAT score decrease was similar, but Houston suffered a much steeper drop in AP passing percentage and San Antonio also had a better improvement in their ACT scores. Does all this warrant the gap closing as much as predicted? I guess only time will tell when the data is released, but I feel it’s a tad bit overestimated. It’s possible that features I’m not considering/don’t have access to in this study helped the historical gap. Many things can occur on a student’s path to a college degree that can influence things one way or another. For the classes of 2016 and 2017, the gap in percentage is predicted to open up again.

When the actual high school class of 2015 graduation percentage data comes out, this will give me more historical data to train my model on, leading to better predicting power on unseen data. With time I also might be able to find features not previously consider or think about other features I could engineer that could add some predictive power. I have also not yet utilized a neural network, which could also improve predictive power in combination with more historical data. This project could also be replicated for other states as well, introducing some interesting results when comparing different states across the country. One would just have make sure to be consistent in only considering students who went to college within the state they attended high school in.

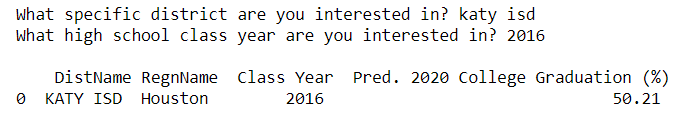
(Recommendations)

As a parent or student curious interested in the historical data or predicted data included in this study, I would explore the following use cases of the final “Hist\_Pred\_Data” dataset to gain a better understanding/make better inferences about what school district may provide some value.

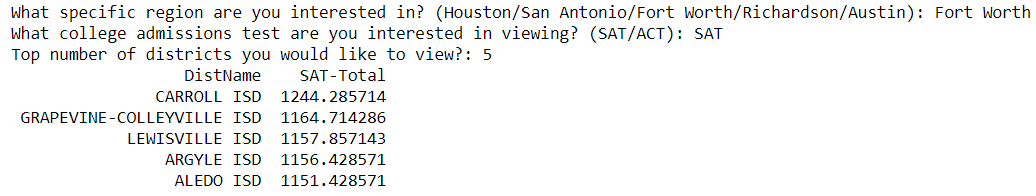
* *Explore the districts with the highest predicted college graduation percentage for a certain region and high school class year*

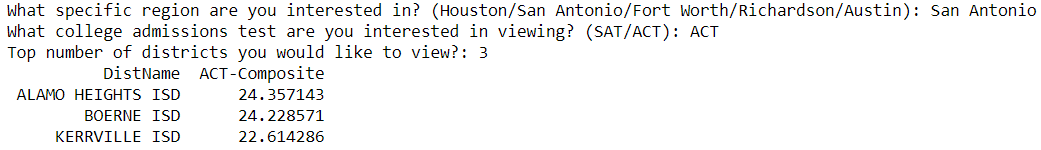


* *Focus on a particular school district you currently live in or are strongly considering*

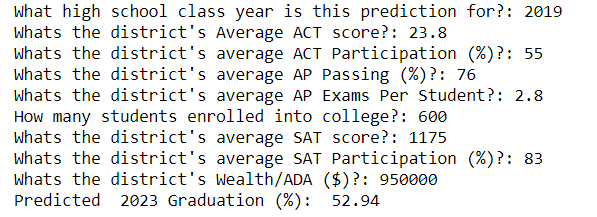


* *View which districts historically performed the best on certain test like the SAT or ACT for a certain region. These tests can earn a student valuable scholarship money!*





* *Let’s say there’s a new school district (population growth is very apparent in the major regions of Texas) that’s been in operation for less than four years. If we have the school district’s test results/features from its first graduating class, we can predict the percentage of those students who will earn a college degree within four years after enrolling into a Texas college*



With the new district’s predicted graduation percentage, one could compare it with the other districts in that particular region or all the major regions. Depending on the results, a family may find the new district to be their most attractive option.

*(Consulted Resources)*

Packages:

Pandas, Numpy, Sklearn, Matplotlib, Glob, Functools

Internet Resources:

TowardsDataScience, Package Documentation, StackOverflow, DataCamp