Springboard DSC:

Capstone Project 1 – Texas Education

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

By: Rachid Rezzik

March 2019

"Which School Districts are Proven to Contain a Higher Percentage of its Graduates Earn a College Degree Within Four Years?"

Many parents across the country find themselves looking for an answer to this 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 graduate from college will help them to live a more comfortable life, which is what all concerned parents want for their child transitioning into adulthood.

In this study, the above question was presented to me by parents considering the move to or within the "major regions" (listed below) of Texas. These parents also noted that they would like their child to attend college within Texas. These particular regions were chosen by the clients as they represent the areas with the most economic opportunity, making them prime locations for their families and talented educators. The information provided is strictly educational in providing clients with "food for thought" ahead of a potential move. Clients may even use the data to rule out a move entirely should their current school district be the most favorable option.

Major Region	Population (Millions)
Houston	2.3
San Antonio	1.5
Richardson (Dallas)	1.3
Austin	0.95
Fort Worth	0.87

When comparing the clients' question with the more popular one amongst parents of "which school districts improve the likelihood my child will get into college?", one may notice their intentions are much more specific. They have stated that gaining college admission is not the end goal, earning a degree is. The clients are looking to avoid a situation in which their child takes more time to earn their degree (more tuition money spent) or even fails out (worst case scenario with minimal return on investment). In strictly focusing on Texas colleges, the clients are also looking to avoid expensive out-of-state tuition.

I approached the problem presented to me by collecting historical data 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, its 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 specific school district) that will graduate within four years. Clients 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

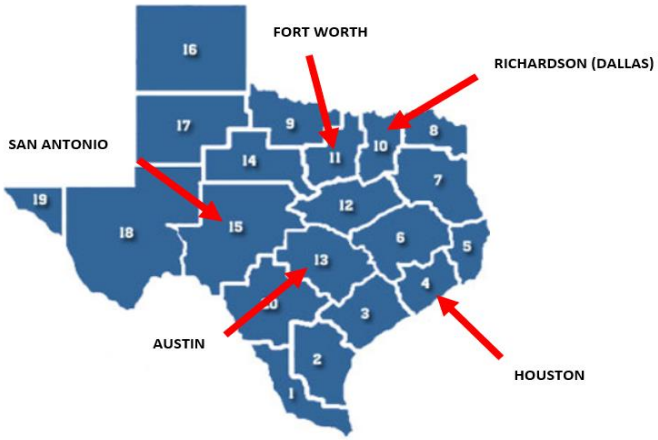
particular region they are considering, or predict graduation percentage from a new school district that has only been in existence for less than four years.

The code I developed for this project is available within my GitHub repository [1].

Data Wrangling

Before we move further, I find it important to clarify how college enrollment and college graduation percentages are assessed for a particular high school class year. If we say the year is 2013, then the percentage of students who enrolled into college in the fall is denoted as "Enrolled 4-Year (%)" and the number of students enrolling is denoted as "Enrolled 4-Year". The percentage of enrolled students who were able to graduate four years later (2017) is denoted as "Graduated 4-Year (%)", the target variable I attempted to predict.

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 [2]. The datasets on college enrollment and college graduation were downloaded from the Texas Public Education Information Resource's ("TPEIR") Website [3]. To give the reader a visual representation of the major regions, I have provided the figure below.



To clarify for anyone unfamiliar with the American education system, The SAT and ACT are college admission tests, while AP exams offer a student the chance to earn college credit. Wealth/ADA is simply the property value of each school district divided by its average daily attendance. The property value comes from the Texas state comptroller and is the basis for each school district's local property tax collections. At the time of this project, the latest data

[1] https://github.com/RachidRezzik/Texas_GradPercent_Predictor
[2] https://tea.texas.gov/
[3] https://www.texaseducationinfo.org/

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.

From the SAT and ACT datasets, I extracted the average scores ("SAT-Total", "ACT-Composite") and participation percentages (percentage of available students who took each respective test) for each school district. From the AP exam datasets, I extracted the total number of AP exams taken, and the amount of passing exams (Note: A score of a 3 or above was considered passing), and participation percentage for each respective district. The number of students who graduated high school ("Total Graduated") and the number of students who enrolled into a four-year college that fall ("Enrolled 4-Year") were taken from each of the respective enrollment datasets. Wealth/ADA was straightforward in just providing the figure for all districts.

For each dataset type (SAT, ACT, AP, Enrollment, Wealth/ADA), I sliced the datasets to only include public school districts (those containing "ISD" in the district name) within the major regions. Some of the dataset types required further cleansing or provided room for feature engineering, which I have outlined below. Once the cleansing of every class year was finalized, it was appended to a list to later be concatenated into one total DataFrame (Ex: "Total_SAT") containing all of the data for the classes of 2011 - 2017.

Further Cleansing:

- For the classes of 2011 – 2016, I adjusted "SAT-Total" scores to be equivalent to CollegeBoard's new scoring system out of 1600 (previously out of 2400, new scoring introduced in 2016) using CollegeBoard's concordance tables [4].
- The enrollment datasets contained district names that included an ID number (Ex: 4825170 KATY ISD). The ID number was not necessary, so I got rid of it to leave the district name in all caps. I also needed fix numerical data that was represented as "*" (data not available) or contained a string with a comma (Ex: 1,244). From there I was able to calculate the percentage of graduating high school students who were able to enroll into a four-year college that fall ("Enrolled 4-Year (%)").
- 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. I was then able to calculate the average number of AP exams taken per student ("AP-Exams Per Student") and the passing percentage ("AP- Passed (%)") for each district.
- For the enrollment and Wealth/ADA datasets, each district's region was not provided. Using the "Total_SAT" dataset, I performed an inner merge to obtain the respective region names for each school district.

[4] https://collegereadiness.collegeboard.org/pdf/higher-ed-brief-sat-concordance.pdf

With all the respective DataFrames containing class of 2011 – 2017 data for public school districts in the major regions of Texas, I then merged them all into one DataFrame ("Feature_Target_Data").

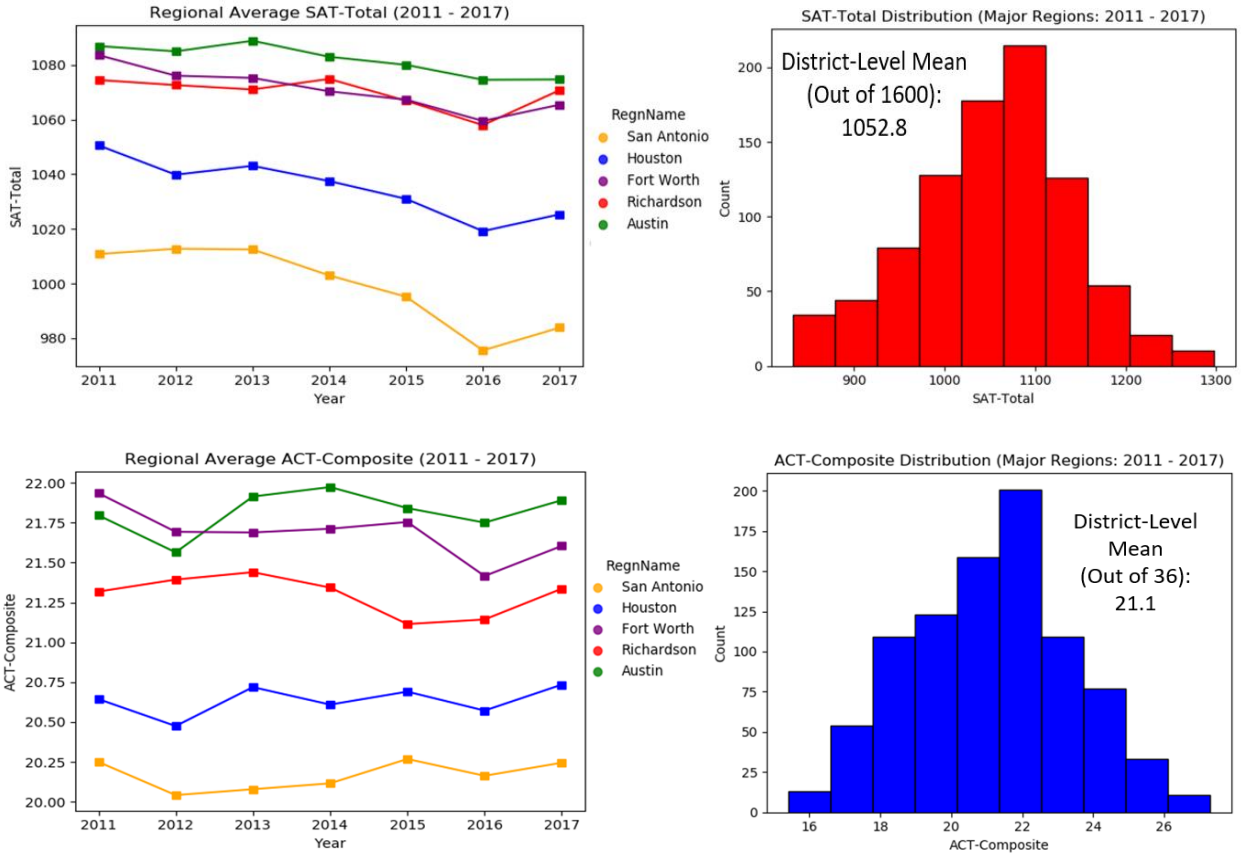
For the college graduation data (the target), I manually inputted the number of students who were able to earn their college degree within four years of 2011 – 2014. This can be read as the number of students who earned their college degree in 2015 – 2018 that belonged to the high school classes of 2011 – 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 ("Graduated 4-Year (%)").

I was then 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) was unknown. Later in the report, I will attempt to predict the target for these class years after establishing a satisfactory machine learning model.

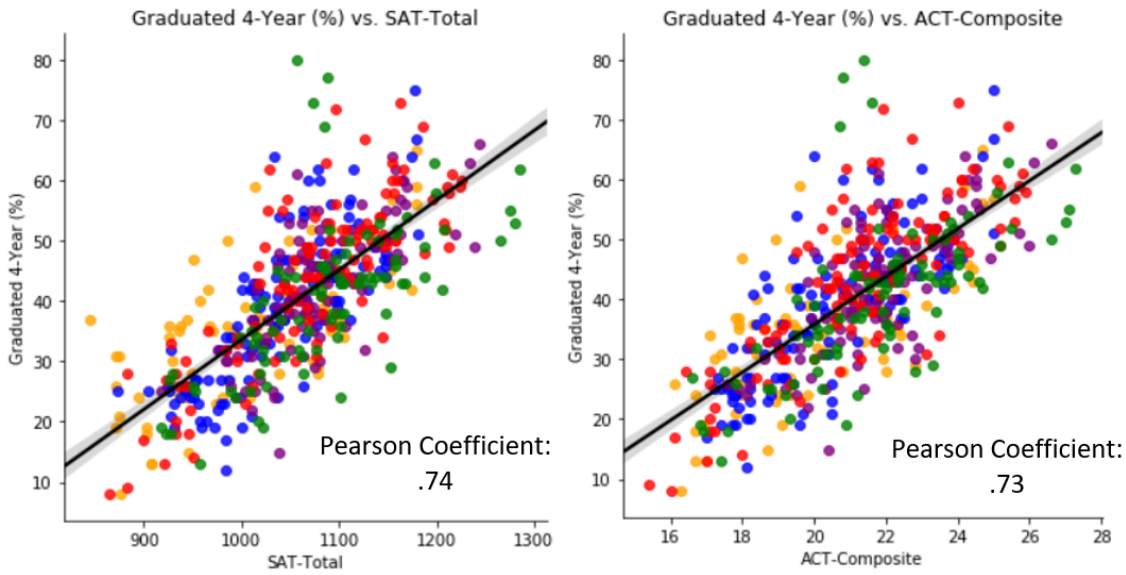
Data Storytelling / Inferential Statistics

(SAT & ACT)

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



We can see that Austin performed the best while Fort Worth and Richardson were close (especially for the SAT). The gap between the top region and San Antonio is quite large for both tests, this will remain a common theme for the other features as well. Now that we have seen how the different regions performed, let's see how increasing district-level scores affected the college graduation percentage for the classes of 2011 – 2014.

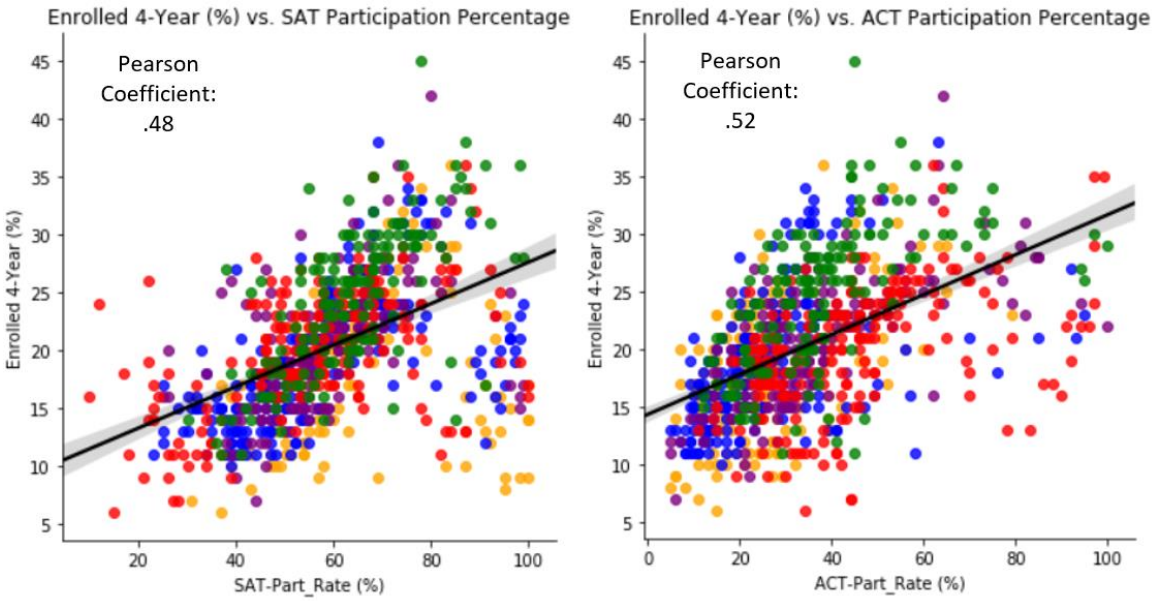


The SAT and ACT both contained a strong positive correlation with college graduation percentage. The correlations make complete sense as colleges use these tests to identify students that they believe are more likely to do well at their institution.

(SAT/ACT Participation %)

Year	2011	2012	2013	2014	2015	2016	2017
ACT-Part_Rate	32.007874	32.937008	32.346457	33.771654	35.204724	37.826772	37.409449
SAT-Part_Rate	61.866142	57.771654	56.511811	58.559055	59.039370	59.755906	62.267717

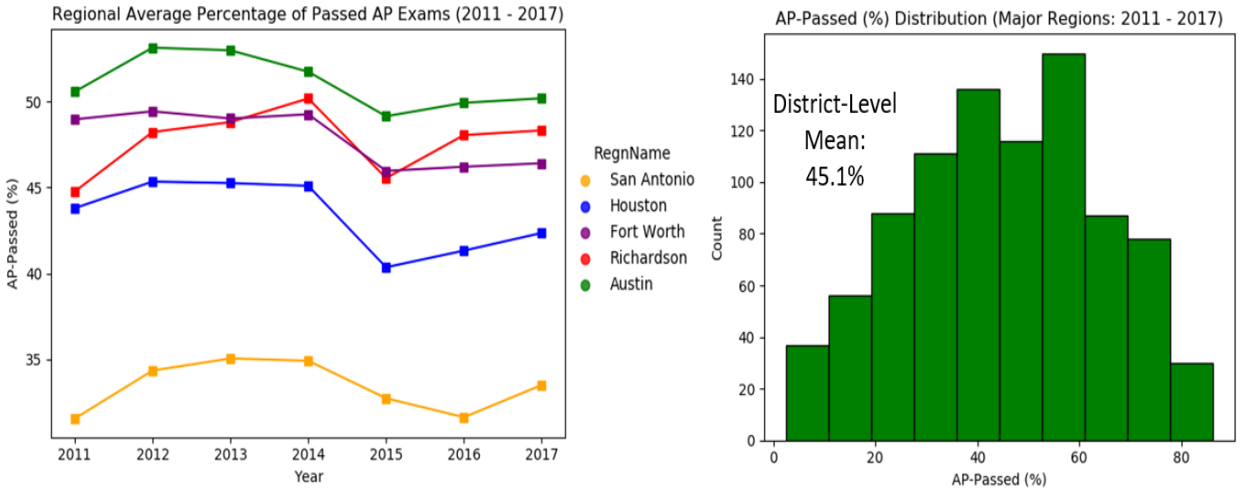
When looking at the classes of 2011 – 2017, I found that a greater percentage of students consistently chose to take the SAT compared to the ACT. Why is this? Well, we could see if choosing to take the SAT historically contained a stronger correlation with college enrollment percentage. If it turns out that this was indeed the case, this trend wouldn't be very surprising. Let's take a look.



Interestingly enough, we see that SAT participation did not contain a stronger correlation with college enrollment than ACT participation. The correlations were actually quite similar with ACT participation even containing a slightly stronger correlation.

As it relates to the client's question/goal, the first step towards earning a college degree is being admitted into college. The above figures indicate that taking both of the college admission tests would be a good idea in helping a student achieve college admission.

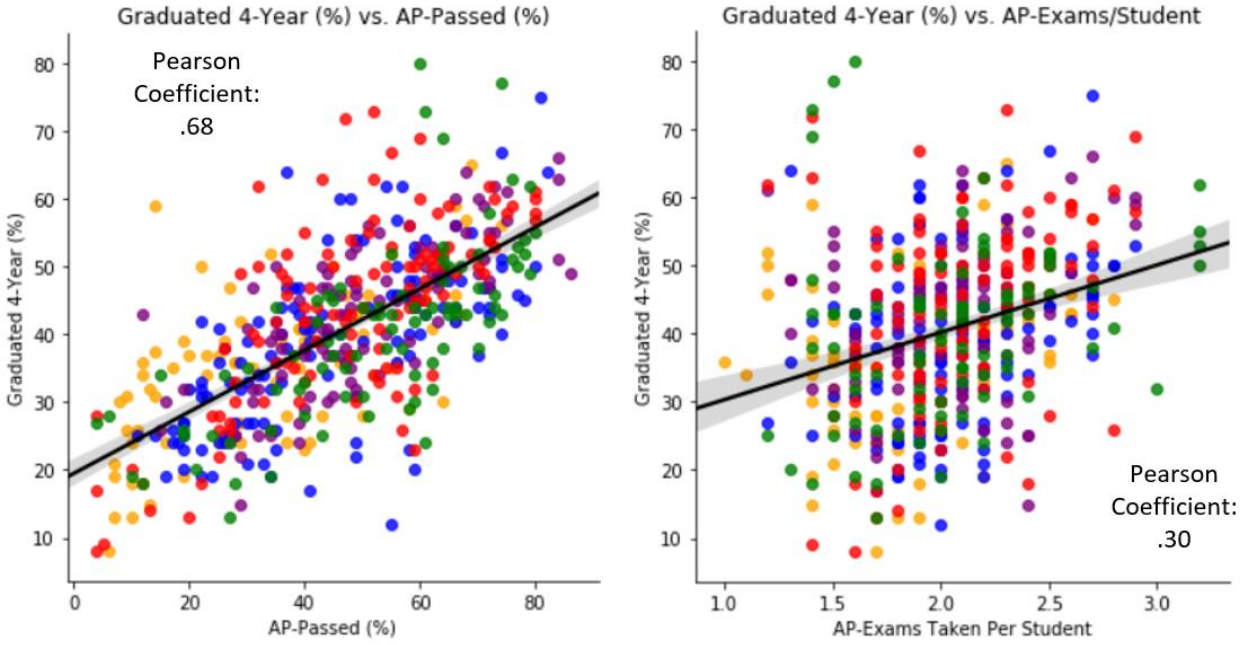
(AP Exams)



The regions of Austin, Richardson (Dallas), and Fort Worth appear to have contained 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.

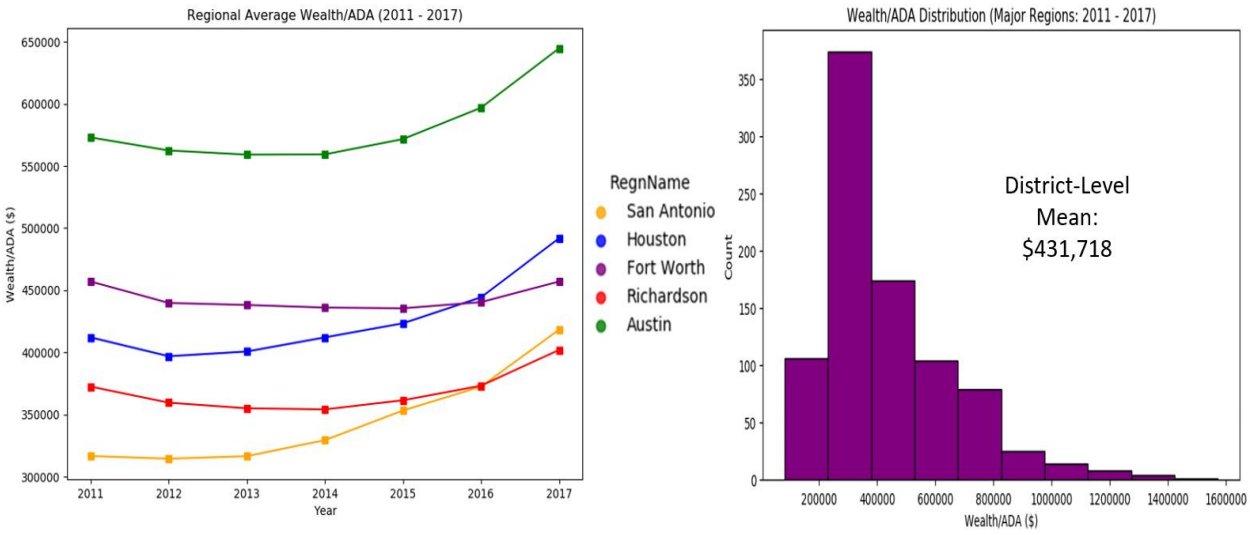
	AP-Exams Taken Per Student						
Year	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
RegnName							
Austin	2.043478	1.990909	2.050000	2.081818	2.200000	2.240909	2.271429
Fort Worth	2.065385	2.034615	2.038462	2.084615	2.161538	2.226923	2.238462
Houston	2.065625	2.009375	2.043750	2.050000	2.103125	2.178125	2.228125
Richardson	2.035714	2.082143	2.128571	2.142857	2.264286	2.242857	2.307143
San Antonio	1.811111	1.794737	1.831579	1.842105	2.005263	1.968421	2.105000

From the standpoint of AP class/exam availability, students attending high school in Richardson (Dallas) had the most opportunity to gain exposure to college-level courses and earn college credit. For the two AP exam – related features mentioned above, let's view their respective correlations with college graduation percentage.



When logically thinking about the two features, it's no surprise that the passing percentage contained the stronger correlation with college graduation percentage. What good is having access to more AP classes/exams if the student isn't prepared to prove they contain a college-level understanding of the material? Though it contains a weaker correlation with the target variable, the amount of AP exams taken per student still has some importance to the clients as it allows them the opportunity to further save money on college while giving their child more college-level exposure.

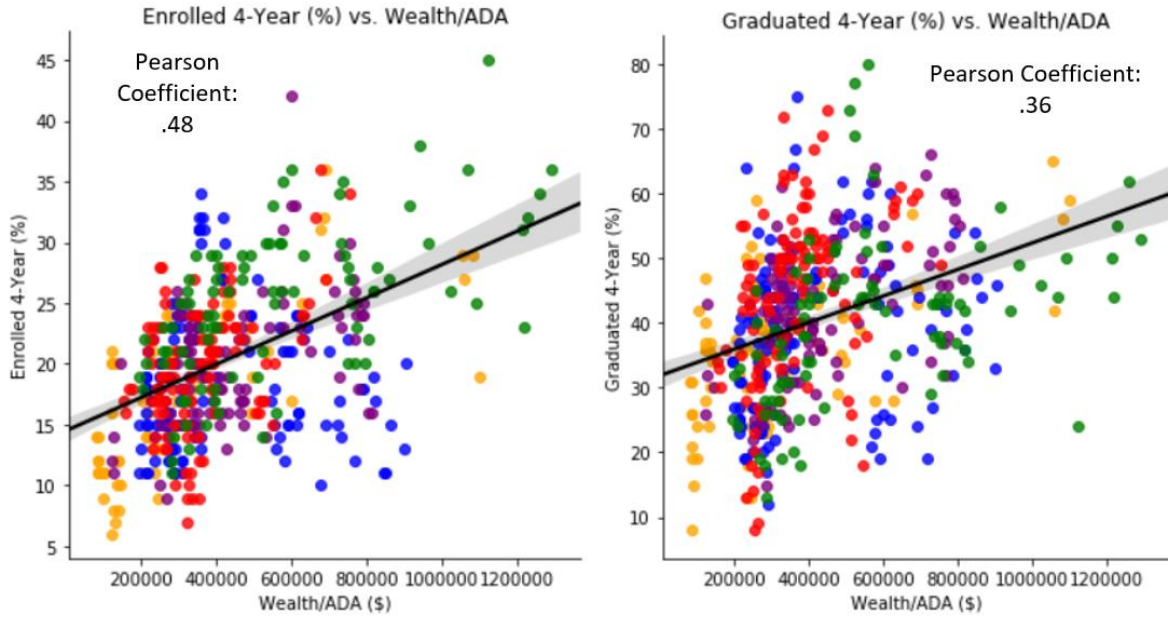
(Wealth/ADA)



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's Influence on College Enrollment (%) Versus College Graduation (%))

Did Wealth/ADA contain a greater positive correlation with college enrollment percentage or college graduation percentage? Let's see what we can extract from the classes of 2011 – 2014.



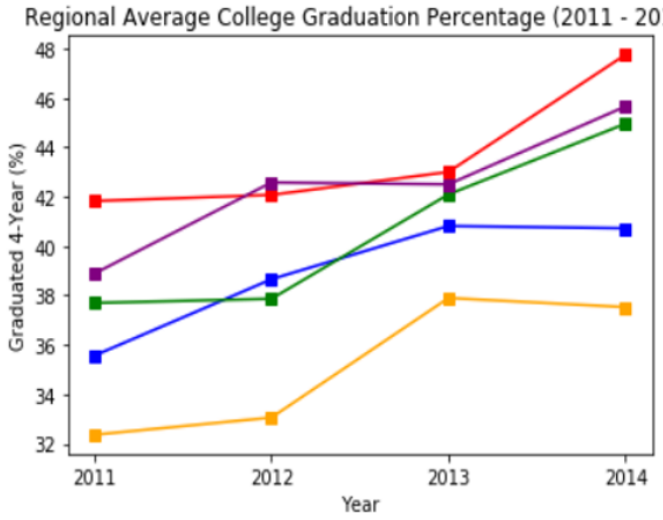
Above, we can see that the pearson correlation between Wealth/ADA and college enrollment percentage was roughly 0.48, indicating a moderate positive relationship between the two variables. Though it certainly isn't the only variable influencing college enrollment percentage, wealth does increase an individual's ability to afford college tuition.

The pearson correlation between Wealth/ADA and college graduation percentage was roughly 0.36, indicating that the wealth of the district a student attends high school in has less influence on college graduation percentage than college enrollment percentage.

When it comes to the more common question among parents about their child getting into college, these statistics tell us that wealth will contribute more to their desired outcome. Maintaining focus on the client's specific question, Wealth/ADA contains less influence in predicting college graduation percentage.

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

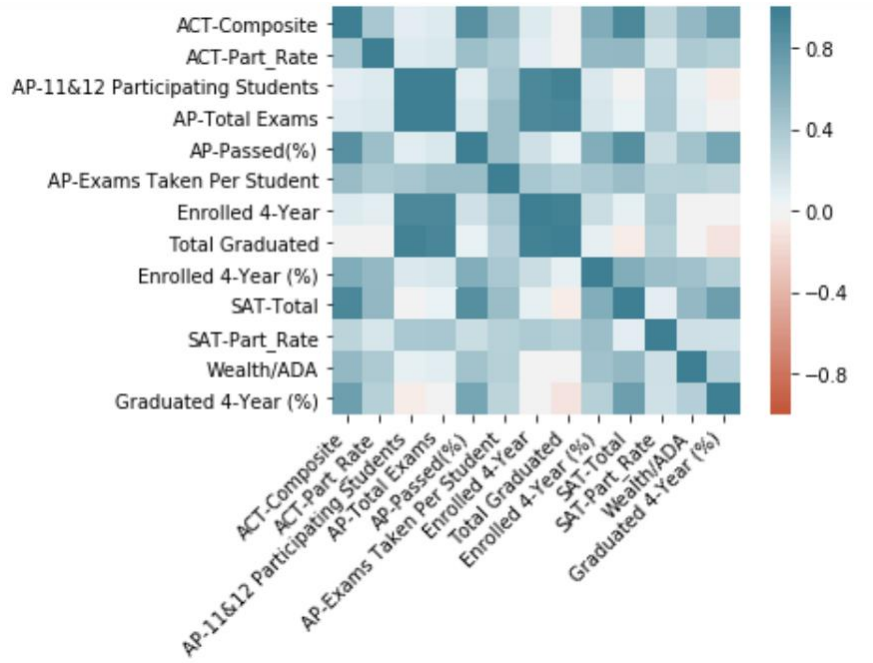
For the classes of 2011 – 2014, let's look at the average percentage of students who were able to earn their college degree within four years by region.



RegnName	Graduated 4-Year (%)
Richardson	43.660714
Fort Worth	42.404303
Austin	40.617418
Houston	38.937500
San Antonio	35.245899

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.

It's also important to consider that the school district features contain some influence on each other. Some of the features above are a bit redundant as they were used to engineer other features (mentioned in the data wrangling section). This could present some multicollinearity issues in building a regression model to predict our target. Let's explore the respective correlations among all the variables in the next figure.



The features that are prime candidates to be dropped are "Total Graduated", "AP-11&12 Participating Students", and "AP-Total Exams". These features contained slightly negative correlations / no correlation with college graduation percentage and strong correlations with features they helped engineer. In the machine learning section, I'll further explore feature importance to determine if dropping certain features will aid model performance.

Machine Learning:

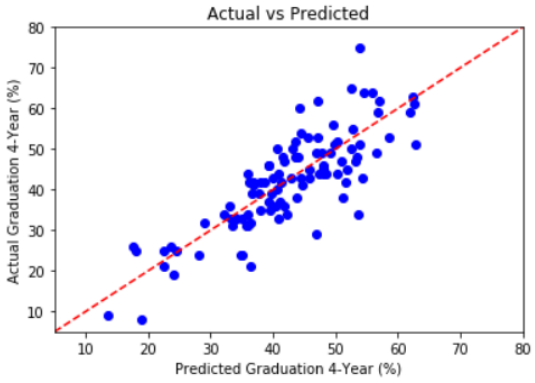
(*Model 1: Linear Regression*)

Utilizing sklearn's Linear Regression [5], I trained the model with an 80 – 20 split (my dataset was on the smaller side with 508 records). 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.

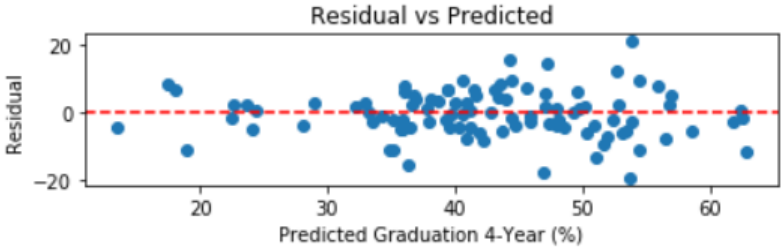
```
Linear Regression

Training RMSE: 8.056457397915441 Testing RMSE: 6.920715398395031

Training R2:  0.5918284580076654 Testing R2:  0.669057877740889

Training MAPE:  17.959181386678583 Testing MAPE:  15.069384924845414
```

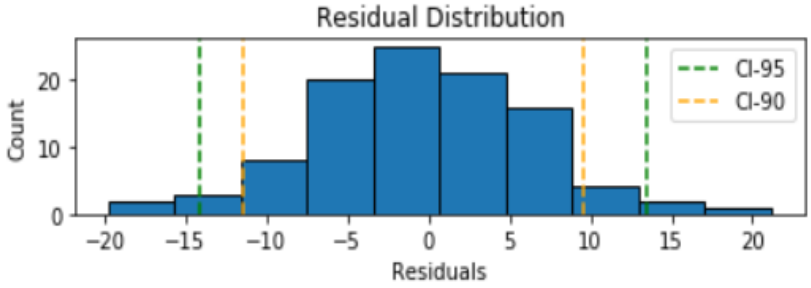
With the performance of the test set, I didn't deem it necessary to include a Ridge or Lasso regression as the model did not appear to be overfitting the training data. To get a better visualization of my base model's performance on the test set, I printed out an "Actual vs. Predicted" scatterplot.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html



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.





CI	lower bound	upper bound	CI Range
90	-11.387778	9.553608	20.941385
95	-14.172749	13.498170	27.670919

(*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 sklearn's Random Forest Regressor ("RFR") [6] and evaluate the resulting

[6] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

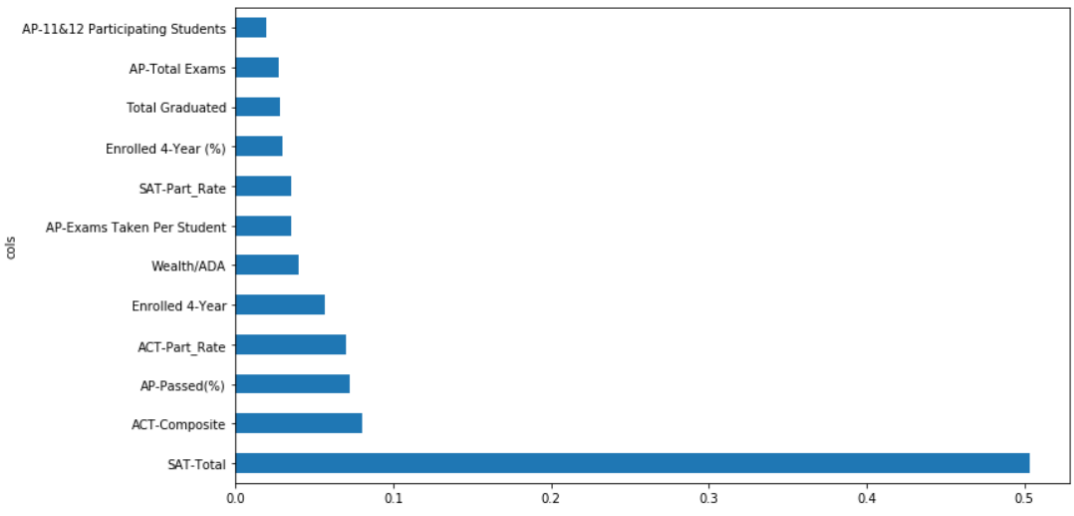
performance in comparison to the base model. I believed the RFR would help solve my concerns with the size of the dataset when generating splits, adding a further element of randomness to prevent overfitting.

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 can be used as another comparison metric for model performance between RFRs.

```
OOB Score 0.6200296049737553

Testing RMSE: 6.697633494769493

Testing R2:  0.6900491550323685

Testing MAPE:  15.537372871022617
```

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': [100, 200, 600, 1000]), the maximum number of features to consider for splitting a node ('max_features': ['sqrt', 'log2', 0.5, None]), and the minimum number of data points allowed in a leaf node ('min_samples_leaf': [5, 3, 2, 1]) 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.

```
Best Params: {'max_features': 'log2', 'min_samples_leaf': 1, 'n_estimators': 200}
```

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

```
OOB Score:  0.6285669955280373

Testing RMSE: 6.638408908817794

Testing R2:  0.6955064706119141

Testing MAPE:  15.385681449058072
```

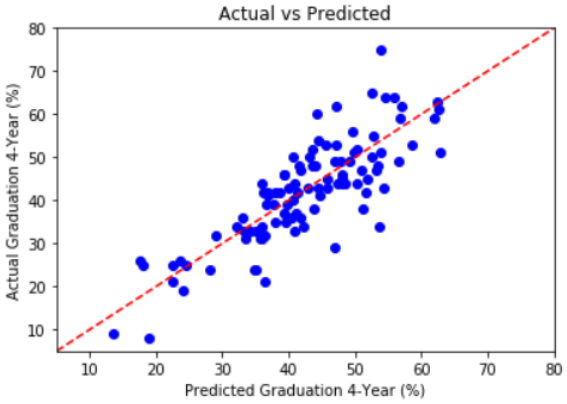
(Model Selection – Summary)

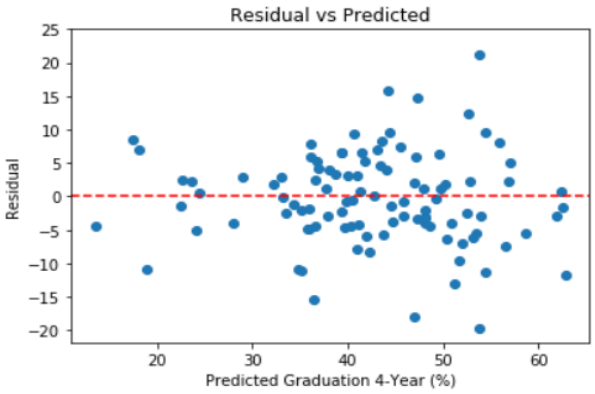
To make things a bit easier for the reader, I have summarized each model's performance on the test set.

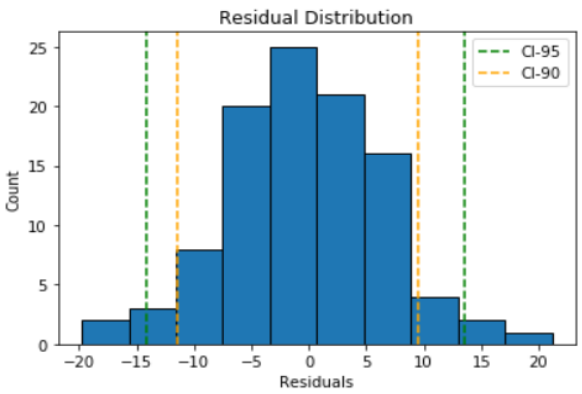
Model	RMSE	R2	MAPE (%)	OOB
1	6.92	0.67	15.07	N/A
2	6.70	0.69	15.54	0.62
★ 3	6.64	0.70	15.39	0.63

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

Model 1

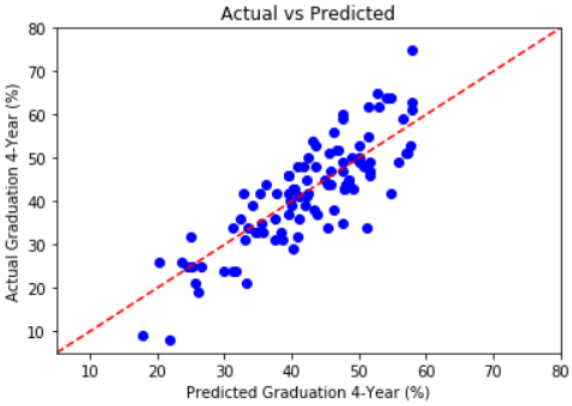


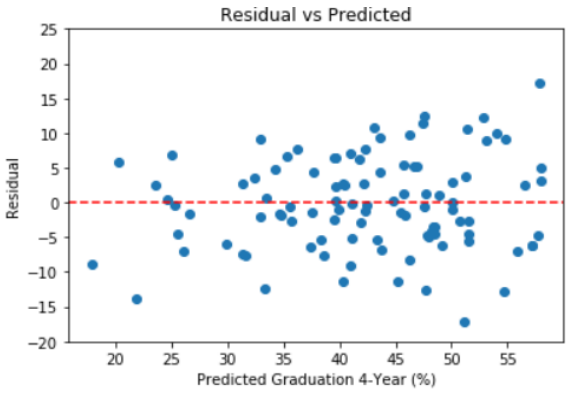


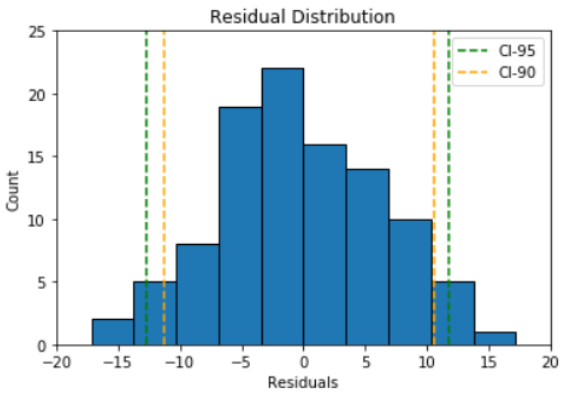


CI	lower bound	upper bound	CI Range
90	-11.387778	9.553608	20.941385
95	-14.172749	13.498170	27.670919

Model 3







CI	lower bound	upper bound	CI Range
90	-11.254988	10.587250	21.842238
95	-12.706625	11.806625	24.513250

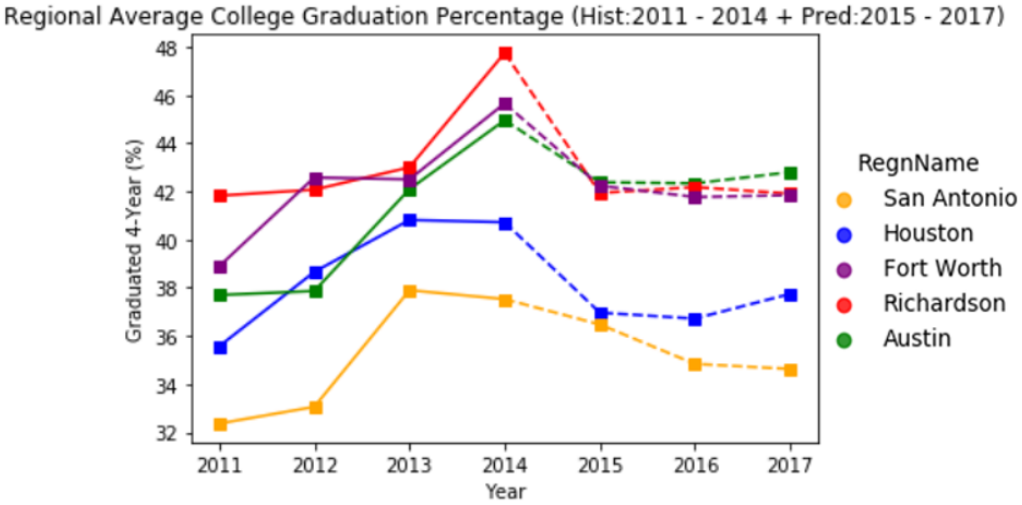
In attempting to predict the percentage of students who graduated within four years, I favor Model 3's performance in decreasing the range of residuals, tightening the CI-95 range, and containing a more symmetric CI-90 lower/upper bound pair. While both models tend to underestimate the actual college 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 select cases where the features/combinations of features increased the residual help 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 clients, swaying their interest towards or away from 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 the actual target (2019 – 2021 college graduation percentage) data was not available. This gave me an opportunity to put my selected model to use in predicting all the 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 or don't have access to in this study previously helped maintain 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.

(Future Work)

- When the 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 predictive power on unseen data. With time I could also possibly find/engineer new features that could also add some predictive power. A neural network was also not utilized in this study. Implementing one could lead to a better performing model 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. A client may not only be considering Texas as a state they would like to live in. I would just need to be consistent in only considering students who went to college within the state they attended high school in.
- Should data come out for students who were able to graduate within four years from a specific Texas college (Ex: "University of Texas" / "Texas A&M") after graduating high school from a particular Texas school district, I could build a model for clients that already have that certain Texas college in mind for their child.

(Recommendations to the Clients)

Utilizing the "Hist_Pred_Data" dataset in my Github repo and Model 3, I have outlined some useful ways the client can approach finding the school district that best meets their requirements. (The use cases below can be found at the end of my "Machine_Learning" notebook)

- Let's start of by assuming that the client is simply considering a move to or within the major regions and has not yet settled on a particular region. The client simply wants to focus on the top five districts that are historically proven and predicted (classes of 2011 – 2017) to average a greater college graduation percentage than its competitors. The top five options are presented below. (Top 20 options printed in notebook)

DistName	RegnName	Graduated 4-Year (%)
LA GRANGE ISD	Austin	62.973895
FRIENDSWOOD ISD	Houston	61.648571
CARROLL ISD	Fort Worth	58.885714
LOVEJOY ISD	Richardson	58.777857
GRAPEVINE-COLLEYVILLE ISD	Fort Worth	58.410714

- Any of these top five districts would be of value to the client's goals for their child. Should housing prices, employment opportunities, commute, etc. in the above districts not be favorable for the client, increasing the number of top districts in consideration or narrowing focus to a particular region would be a good choice. Below I have presented the count of districts in each region that are predicted to maintain an average college graduation percentage above 50% for the classes of 2011 – 2017.

RegnName	
Austin	4
Fort Worth	5
Houston	2
Richardson	10
San Antonio	1

- From the viewpoint of quality options, the client would be advised to explore school districts in Richardson (Dallas). For any region the client has chosen to focus on, they can provide the amount of top options they'd like to view. The client could choose to maintain focus on the average graduation percentage for the classes of 2011 - 2017, but some may prefer focusing on the class of 2017 due to the recency provided.

```
What region are you interested in? (Houston/San Antonio/Fort Worth/Richardson/Austin): Richardson
What high school class year of students are you interested in?: 2017
How many districts in your chosen region would you like to view?: 4
```

	DistName	RegnName	Class Year	Pred. 2021 College Graduation (%)
0	COPPELL ISD	Richardson	2017	60.280
1	PLANO ISD	Richardson	2017	57.730
2	CELINA ISD	Richardson	2017	57.710
3	FRISCO ISD	Richardson	2017	57.385

- Many clients are already be living in a school district within the major regions. After taking advantage of the strategies mentioned above, the client could choose to do a comparison between their current district and the one they are now strongly considering. Moving can be a great hassle, so they want to see some data that will convince them it may be worth it. Below I have provided an example of a client who is currently zoned to Round Rock ISD (Austin) and is strongly considering moving to Eanes ISD (Austin).

School District the Client Already Resides In:

```
What specific district are you interested in? round rock isd

*Predicted Average Graduation % for Classes of 2011 - 2017 + Rankings *

        DistName  Graduated 4-Year (%)  Ranking(All Regions)  Ranking(Austin)
 ROUND ROCK ISD                 49.47                    24                5


*Predicted Graduation % for Class of 2017 + Rankings *

        DistName  Pred. 2021 Graduated 4-Year (%)  Ranking(All Regions)  Ranking(Austin)
 ROUND ROCK ISD                            48.205                   29                7
```

School District the Client is Considering:

```
What specific district are you interested in? eanes isd

*Predicted Average Graduation % for Classes of 2011 - 2017 + Rankings *

   DistName  Graduated 4-Year (%)  Ranking(All Regions)  Ranking(Austin)
 EANES ISD             56.716429                     8                2


*Predicted Graduation % for Class of 2017 + Rankings *

   DistName  Pred. 2021 Graduated 4-Year (%)  Ranking(All Regions)  Ranking(Austin)
 EANES ISD                            58.21                     4                1
```

In the above example, the predicted data backs the client's opinion that moving to Eanes ISD would better prepare their child for graduating college within four years. For certain pairs of districts, the outcome could be the opposite in convincing the parent that their current district is favorable.

- As mentioned before, the clients are all concerned with saving money on college tuition. Therefore, it would also be in their interest to view which districts historically performed the best on certain tests like the SAT or ACT for a certain region. These tests can earn a student valuable scholarship money!

```
What specific region are you interested in? (Houston/San Antonio/Fort Worth/Richardson/Austin): Fort Worth
What college admissions test are you interested in viewing? (SAT/ACT): SAT
Top number of districts you would like to view?: 5
                DistName    SAT-Total
             CARROLL ISD  1244.285714
 GRAPEVINE-COLLEYVILLE ISD  1164.714286
          LEWISVILLE ISD  1157.857143
              ARGYLE ISD  1156.428571
               ALEDO ISD  1151.428571
```

- 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 use Model 3 to predict the percentage of those students who will earn a college degree within four years after enrolling into a Texas college.

  Note: This same process will essentially be used when the class of 2018 school district data comes out to predict 2022 college graduation percentages, giving the clients more data to consider.

```
What high school class year is this prediction for?: 2019
Whats the district's Average ACT score?: 23.8
Whats the district's average ACT Participation (%)?: 55
Whats the district's average AP Passing (%)?: 76
Whats the district's average AP Exams Per Student?: 2.8
How many students enrolled into college?: 600
Whats the district's average SAT score?: 1175
Whats the district's average SAT Participation (%)?: 83
Whats the district's Wealth/ADA ($)?: 950000
Predicted  2023 Graduation (%):  52.94
```

  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 - https://pandas.pydata.org/docs/
- Numpy - https://numpy.org/doc/
- Sklearn - https://scikit-learn.org/stable/
- Matplotlib - https://matplotlib.org/3.2.1/contents.html
- Seaborn - https://seaborn.pydata.org/
- Glob - https://docs.python.org/3/library/glob.html
- Functools - https://docs.python.org/3/library/functools.html

Internet Resources:

- Texas Education Agency - https://tea.texas.gov/
- Texas Education Info - https://www.texaseducationinfo.org/
- TowardsDataScience - https://towardsdatascience.com/
- StackOverflow - https://stackoverflow.com/
- DataCamp - https://datacamp.com/