CSE 487 - Data Intensive Computing

2021

Analysis of 2016 NYC DOE Highschool data based on median income and NYC boroughs

Contributors: Takhim Haque
Mustafa Raja

Abstract

Yearly, the statistics released for high school graduation and college enrolment in New York City show that graduation is a big hurdle for most high-school students. While the challenges may vary, these statistics pale in comparison to NYC's status as one of the world's wealthiest cities. This report details our study into the blurry lines behind these statistics. With the aid of various Machine Learning models, we were able to test several hypotheses to understand the reality of these challenges and draw up possible ways to solve them. Our findings point out social mobility as one primary reason for the challenges, and we believe that the provision of incentives can positively contribute to solving these challenges.

Introduction

"Every year, thousands of students in New York City struggle to graduate high school."

The above statement may be subject to varying debates for and against the perception that graduating high school in New York City (NYC) is a challenging ordeal. To get a clear picture of the reality of thousands of students in New York City, we decided to study the available facts. We focused on getting the exact figures, and we explored how they correlate with the challenge highlighted in the above quote.

Hence, this report covers our research and analysis of the graduation rate in New York City on a per-school basis, and it simultaneously covers our analysis of the graduation rates within each Borough.

We believe that this approach will help us and the readers understand any school's performance within a given Borough. In addition to this, we also took an interest in comparing different Boroughs' performance in terms of high school graduation and college enrollment. Moreover, we hope that our results will highlight the problems affecting students' graduation rates within the Boroughs where they are located.

Inspiration for the project: https://opendata.cityofnewyork.us/projects/data-visualization-of-nyc-high-schools-college-enrollment

Name of our data:

- 2016 NYC High school directory
- 2016 NYC High school performance
- 2016 income by zip code for NYC

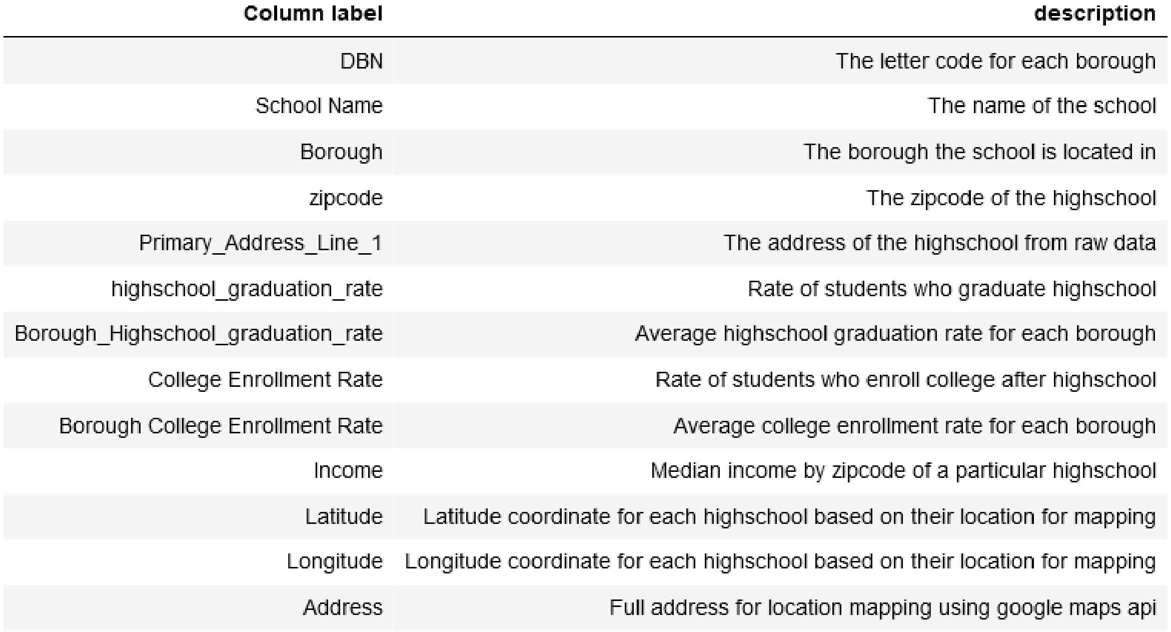
Background:

- Source 1: https://data.world/
- Source 2: https://guides.newman.baruch.cuny.edu/nyc_data/nbhoods
- Source 2: https://data.census.gov/cedsci/
- Time Period: 2016
- Scope (if there is, please specify): New York City

- Data Unit: Statewide

Hypothesis and conclusion

- The Borough of Manhattan will perform better in terms of graduation rate than other Boroughs.
- The Borough of Bronx will perform worst in terms of graduation rate than other Boroughs.
- A higher graduation rate varies directly to a wealthy neighborhood, i.e., a wealthy neighborhood will have a higher graduation rate, while a poor neighborhood will have a low graduation rate.
- There is a direct relationship between geographical locations and their high school graduation rate.
- There is a direct relationship between geographical locations and their college enrollment rate.



Column label	description
DBN	The letter code for each borough
School Name	The name of the school
Borough	The borough the school is located in
zipcode	The zipcode of the highschool
Primary_Address_Line_1	The address of the highschool from raw data
highschool_graduation_rate	Rate of students who graduate highschool
Borough_Highschool_graduation_rate	Average highschool graduation rate for each borough
College Enrollment Rate	Rate of students who enroll college after highschool
Borough College Enrollment Rate	Average college enrollment rate for each borough
Income	Median income by zipcode of a particular highschool
Latitude	Latitude coordinate for each highschool based on their location for mapping
Longitude	Longitude coordinate for each highschool based on their location for mapping
Address	Full address for location mapping using google maps api

Libraries used

- Pandas
- Numpy
- Seaborn
- Matplotlib
- Folium
- Requests: HTTP for Humans™
- Google maps sdk

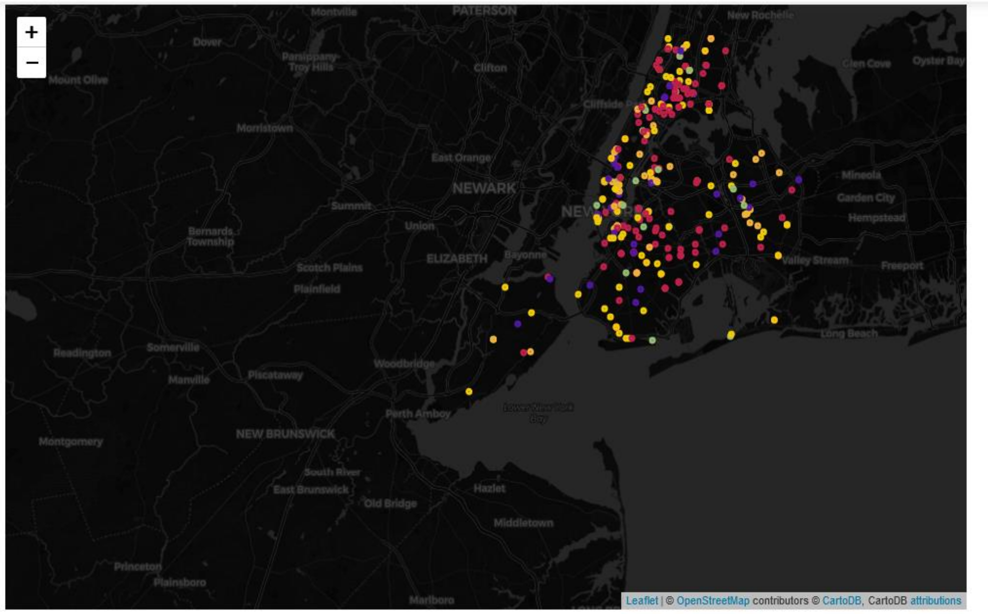
- Pycaret (ML Models)

How did we clean our data?

First, we merged data from the 2016 NYC High school directory and the 2016 NYC High school performance; then, we extracted the combined data that matched our set metrics. We deleted schools with empty values for college enrollment, Borough college enrollment, high school graduation and Borough high school graduation. We deleted them so we could avoid getting skewed data results when we try to access non-existing elements. Likewise, based on the zip code of a school, we matched it with the median income data gathered from available US census data. Also, we renamed some of the columns to help our readers fully grasp the data.

Exploratory analysis

We had to sift and separate best performing Boroughs from the worst-performing Boroughs for both high school graduation and college enrollment rates for our analysis. We further divided schools in each Borough among the best and worst high school graduation and college enrollment rates. Each of the Borough's median income data was studied to gain insights into the income disparity. Then, we compared each high school's graduation rate with their Borough's graduation rate and each high school's college enrollment rate to their Borough's college enrollment rate. Eventually, we mapped all the high schools in each Borough based on their location, then we highlighted them with different colors based on their college enrollment rate.



Key

- College Enrollment >= 90%: Green
- 75% <= College Enrollment < 90%: Blue
- 65% <= College Enrollment < 75% : Purple
- 50% <= College Enrollment % < 65%: Yellow

Experiments – ML MODELS

Hypothesis: A higher graduation rate varies directly to a wealthy neighborhood.

ML Algorithm: Method 1 - Extreme Gradient Boosting

We tested our first hypothesis with the Extreme Gradient Boosting (Xgboost) algorithm. The test was designed to confirm any relationship between high school graduation rates and median income. Our data was subjected to 17 different models, but Extreme Gradient Boosting would provide the 2nd best initial results.

During our initial comparison of different models, we used Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), R-squared (R2), Root Mean Squared Log Error (RMSLE), and Mean Absolute Percentage Error (MAPE) as our metrics.

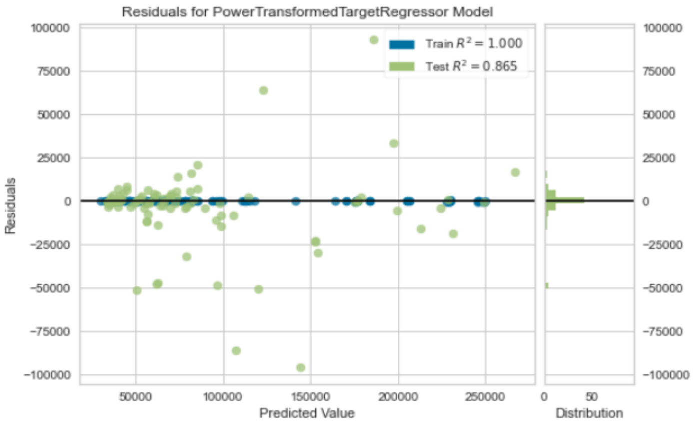
We shared our data in percentages of 70, 20, and 10 to train the model, test and validate data, and as unseen data to evaluate the final results produced by the Extreme Gradient Boosting (Xgboost) algorithm, respectively. The main parameters used for our evaluation were the RMSE, R2, and MAPE metrics. We chose RMSE to achieve a well-balanced model. R2 since it directly measures the goodness of fit in capturing the variance while training the data. Additionally, we picked MAPE because its properties are essential to our understanding of the results' accuracy.

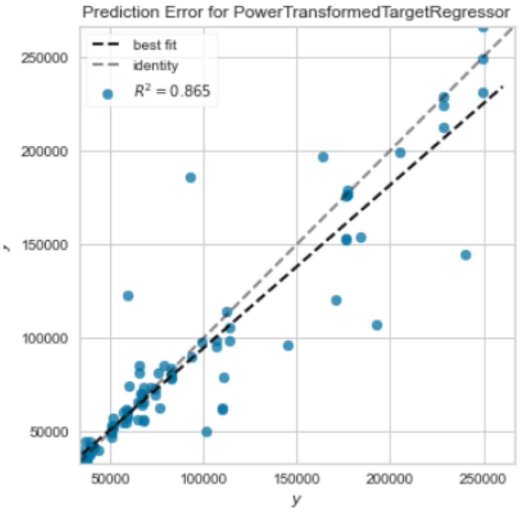
We tuned our RMSE parameter to produce better results; however, R2 and MAPE were not tuned since they over fit the results. Essentially, our algorithm attempts to predict how our target variable median income was achieved via merging estimates from a set of simpler and weaker models.

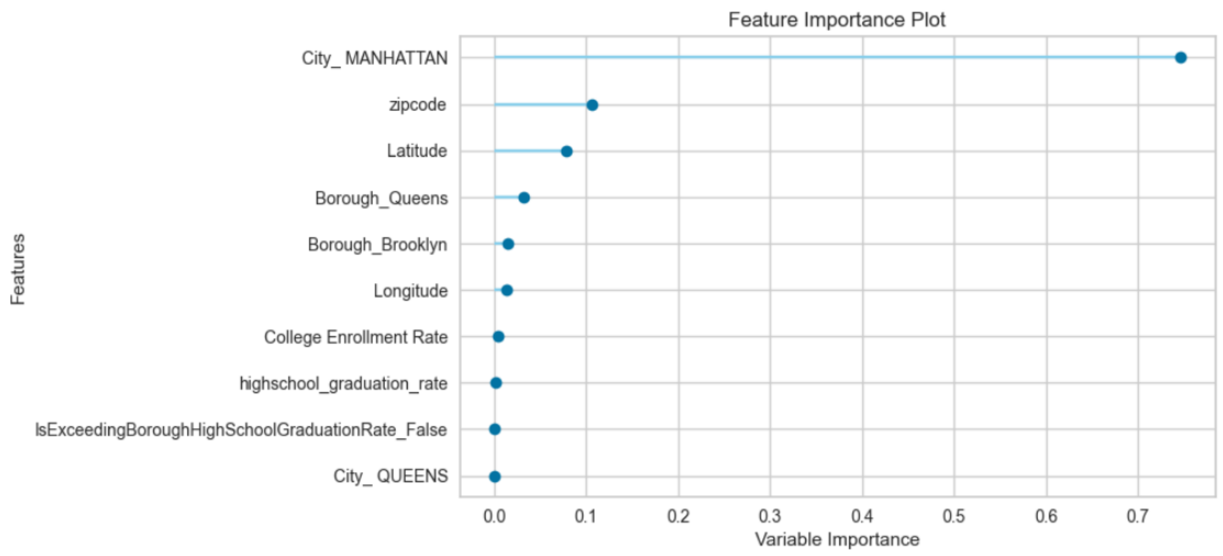
Our initial RMSE value was 20400.0885, R2 value was 0.7559 and MAPE value was 0.1067. After tuning and optimizing, our RMSE value was recorded as 17562.1210, R2 was 0.9148, and our MAPE value was recorded as 0.0734 from the test data. Our final RMSE value for the unseen data was 255050868.1769, R2 value was 0.9237, and our MAPE value was 0.0847.

Do these values provide an answer to our initial hypothesis? Yes, they show that our hypothesis was not entirely correct as there is no direct relationship between graduation rates and median income. Also, we noticed that when we used graduation rate as our only variable, our model failed to yield insightful results due to the very low value of R2. We had to support it with more variables such as Borough, Zip code, its relative Borough high school graduation rate and college enrollment rate to predict median income.

Our most important feature plot here was centered on the Borough of Manhattan. Manhattan also had the third-highest number of high schools among the Boroughs featured in our data. Considering our prediction on test data and our plot for R2 below, our algorithm performed relatively well at predicting median income due to its high R2 value. These results leave room for further testing with our next algorithm, the Gradient Boosting Regressor (GBR).







ML Algorithm: Method 2 - Gradient Boosting Regressor

We decided to test further to see any relationship between all graduation rates and median income. For this test, however, we used Gradient Boosting Regressor (GBR) to verify our assumption. We ran our data against 17 models, and the Gradient Boosting Regressor would provide the best initial results.

During our initial comparison against different models, we used Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), R-squared (R2), Root Mean Squared Log Error (RMSLE) and Mean Absolute Percentage Error (MAPE) as our metrics.

The Gradient Boosting Regressor (GBR) algorithm works by calculating the difference between the current prediction and the correct target value; therefore, we needed to maintain consistency on our GBR test, as this will allow us to compare it directly with Xgboost. 70% of our data went to training the model. 20% of our data was used for data testing and validation, while the remaining 10% data was treated as unseen data to evaluate final results produced by the Extreme Gradient Boosting algorithm.

The main parameter used for our evaluation were RMSE, R2, and MAPE. We chose RMSE to achieve a well-balanced model. R2 since it directly measures the goodness of fit in capturing the variance while training the data. We also picked MAPE because its properties are essential to our understanding of the accuracy of the results.

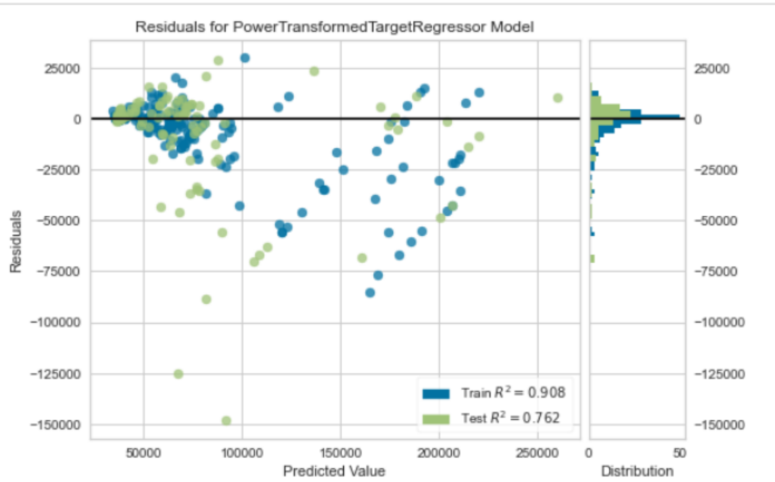
We tuned our RMSE parameter to produce better results; however, R2 and MAPE were not tuned since they over fit the results.

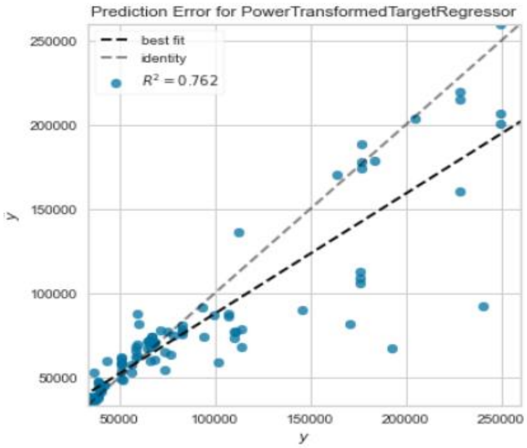
Our initial RMSE value was 20270.0617, R2 value was 0.7559, and MAPE value was recorded as 0.1268. When we tuned and optimized our GBR results, our RMSE value became 17562.1210, R2 became 0.9148 on test data, and MAPE value became 0.0734. These results agreed with the Xgboost results after they had been tuned and optimized.

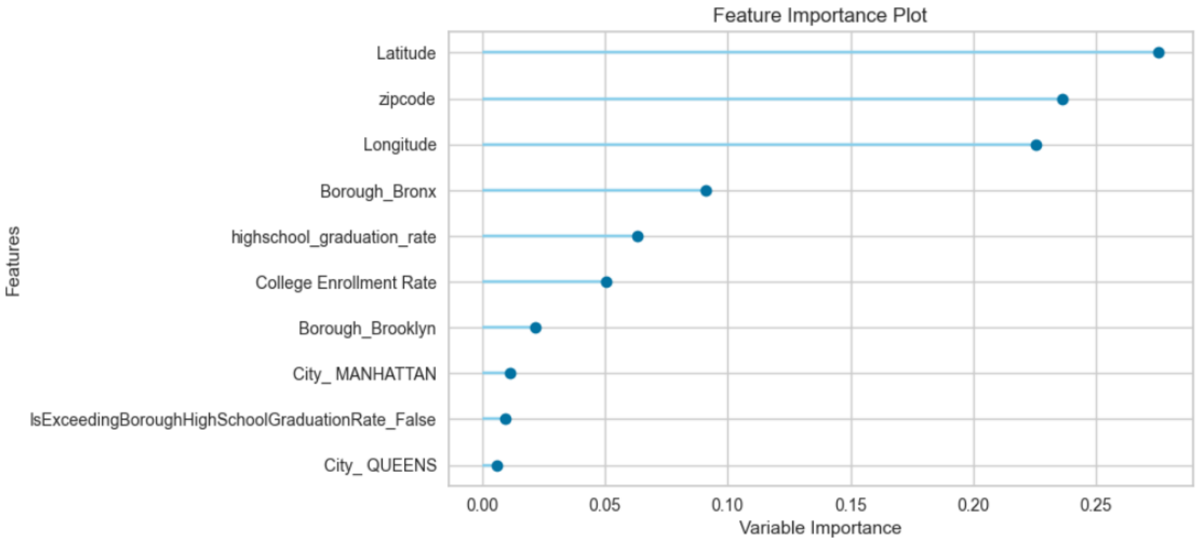
We recorded similar values on our Xgboost results as we did for our final GBR values. RMSE values on unseen data were 255050868.1769, R2 value was 0.9237, and our MAPE value was 0.0847. Again, this further proves that our hypothesis was not entirely correct as it also confirms that there is no direct relationship between graduation rates and median income.

Interestingly, when we used graduation rate as our only variable, our model does not produce insightful results due to having a meager R2 value, and this also agreed with the Xgboost results. We had to include more variables such as Borough, Zip code, its relative Borough high school graduation rate and college enrollment rate to predict median income.

Our most important features plot on our GBR test are the geographical variables such as Latitude, Longitude, and Zipcode. Even though Gradient Boosting Regressor had the best initial results, after tuning and optimizing our parameters, the results agree with the same results we got with Xgboost.







Hypothesis: There is a direct relationship between geographical locations and their high school graduation rate.

ML Algorithm: Method 3 - Bayesian Ridge

For this hypothesis, we tested our data to confirm if there was any relationship between geographical locations and high school graduation rate. We ran our data against 17 models, and the Bayesian Ridge algorithm would yield the 2nd best initial results.

When we carried out initial comparison against different models, we used Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), R-squared (R2), Root Mean Squared Log Error (RMSLE) and Mean Absolute Percentage Error (MAPE) as our metrics.

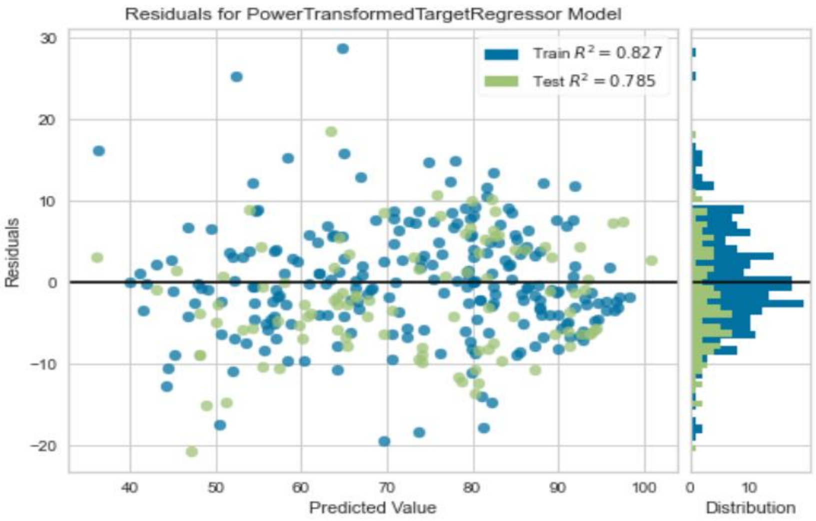
The Bayesian ridge algorithm computes the model by regularizing parameters and adapting them to the data at hand. 70% of our data was expended on training the model. 20% of the data was used for running data test and validation, and the remaining 10% data was used as unseen data to evaluate final results produced by the Bayesian Ridge algorithm.

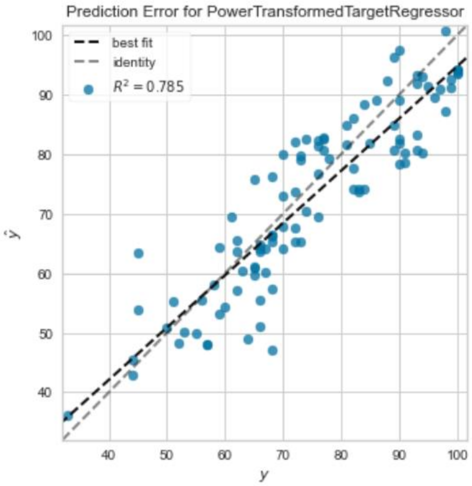
The main parameters we used for evaluation were MSE, RMSE, and R2. MSE was selected to gain appreciable insights into the differences between original values and predicted values. We also chose RMSE to achieve a well-balanced model. Finally, we chose R2 since it directly measures the goodness of fit in capturing the training data variance. We tuned our RMSE parameter to produce better results.

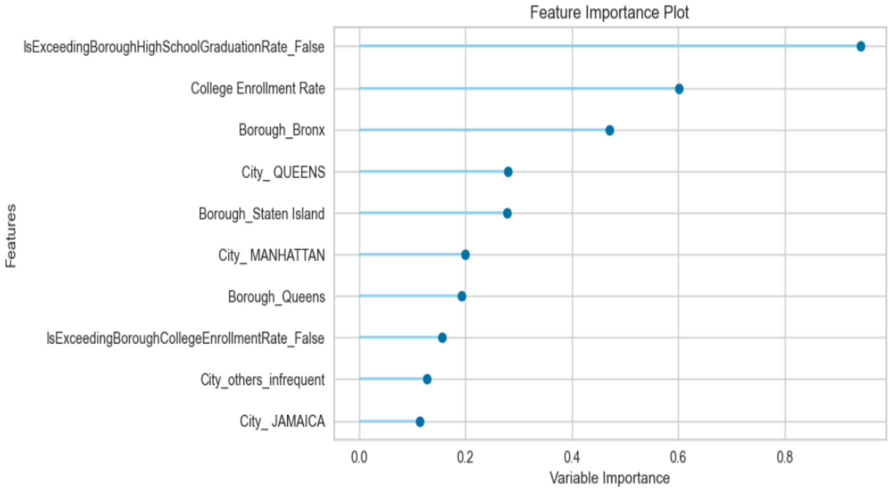
Our initial MSE value was recorded as 54.4903, RMSE value was 7.1431, R2 value was 0.7799. Our MSE value was recorded as 45.2488 after tuning and optimizing, our RMSE value was 6.7267, and R2 was 0.8089 on test data. Our final MSE value recorded for our unseen data was 45.2488, RMSE value was 6.7267, and our R2 value was 0.8721. What do these results say about our initial hypothesis on the relationship between geographical location and high school graduation rates?

Our results show that there is undoubtedly a relationship between geographical variables and high school graduation rates. However, it is not a direct linear relationship, so our hypothesis was only partially correct.

We can observe the impact of these variables in our feature plot. To calculate the high school graduation rate in a meaningful manner, we need to add more variables such as relative graduation rate to other variables such as Borough and college enrollment rate to get insightful results.







ML Algorithm: Method 4 - Ridge Regression

We employed the Ridge Regression algorithm to test the relationship between geographical locations and high school graduation rates. As with others, we also ran our data against 17 models, and we found that Ridge Regression would provide the best initial results.

During our initial comparison against different models, we used Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), R-squared (R2), Root Mean Squared Log Error (RMSLE) and Mean Absolute Percentage Error (MAPE) as our metrics.

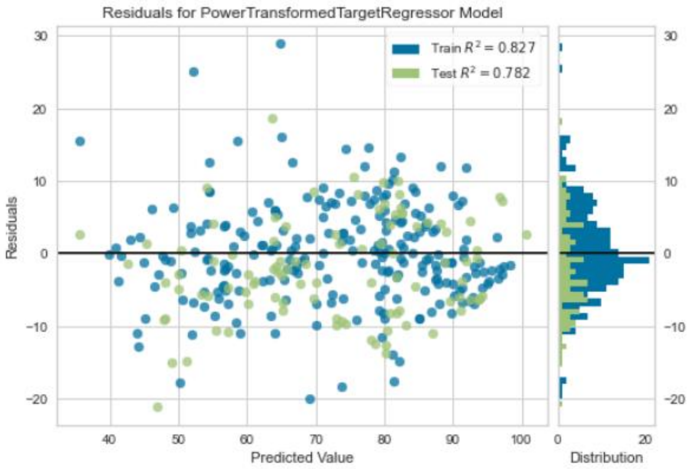
Ridge regression uses multiple regression models to produce the best results. Comparing results for the Bayesian Ridge and the Ridge Regression algorithms required that our data input remained consistent so, we used 70% of our data on training the model, 20% of our data was used for data testing and validation, while the remaining 10% of our data was used as unseen data to evaluate the final results produced by the Ridge regression.

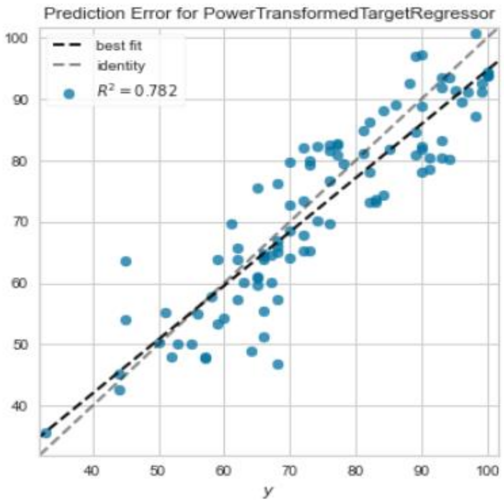
The main parameters we used for evaluation were MSE, RMSE, and R2. MSE was selected to gain appreciable insights into the differences between original values and predicted values. We also chose RMSE to achieve a well-balanced model. Finally, we chose R2 since it directly measures the goodness of fit in capturing the training data variance. We tuned our RMSE parameter to produce better results.

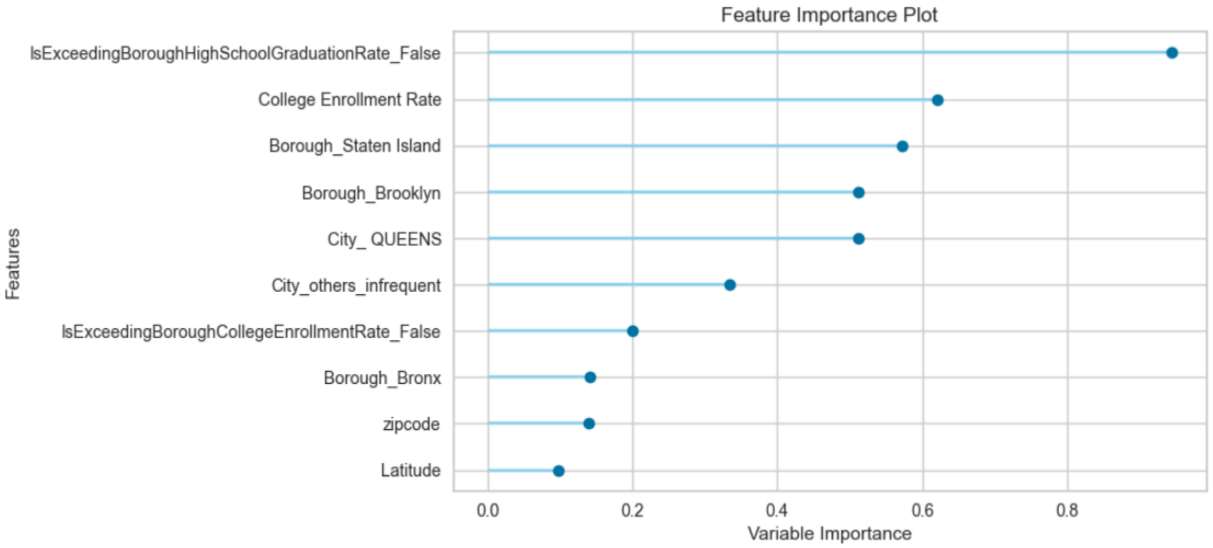
Our initial MSE value was recorded as 54.4455, RMSE value as 7.1406, R2 value was recorded as 0.7801. Our MSE value was recorded as 45.2488, RMSE value as 6.7267, and R2 was recorded as 0.8089 on test data after tuning and optimizing. Our final MSE value on our unseen data was recorded as 45.2488, RMSE value as 6.7267, and our R2 value was recorded as 0.8089.

Again, this further proves that there is only a partial connection between geographical variables and high school graduation rates based on our data.

Also, considering our most important feature plot, our top two features match up with the Bayesian Ridge plot. Additionally, geographical variables such as Borough and zip code also exhibited impacts on our results. To summarize all our findings, we see that the relationship between geographical locations and high school graduation rates are not absolute, i.e., we need to add more variables to produce better results.







Hypothesis: There is a direct relationship between geographical locations and their college enrollment rates

ML Algorithm: Method 5 - Linear Regression

Our hypothesis suggested that there was a relationship between geographical locations and college enrollment rates. We, therefore, decided to test this using the Linear Regression algorithm.

We ran our data against 17 models, and our estimate showed that Linear Regression would provide the 3rd best initial results. We could not choose our first (Bayesian Ridge) or second (Ridge Regression) best algorithm because they had been used previously to predict high school graduation rates.

For our initial comparison against different models, we used Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), R-squared (R2), Root Mean Squared Log Error (RMSLE) and Mean Absolute Percentage Error (MAPE) as our metrics.

70% of our data went to training the model, 20% of the data were used for data testing and validation, while the remaining 10% were regarded as unseen data to evaluate the final results we get from Linear regression.

The main parameters we used for evaluation were MSE, RMSE, and R2. MSE was selected to gain appreciable insights into the differences between original values and predicted values. We also chose RMSE to achieve a well-balanced model. Finally, we chose R2 since it directly measures the goodness of fit in capturing the training data variance. We tuned our RMSE parameter to produce better results.

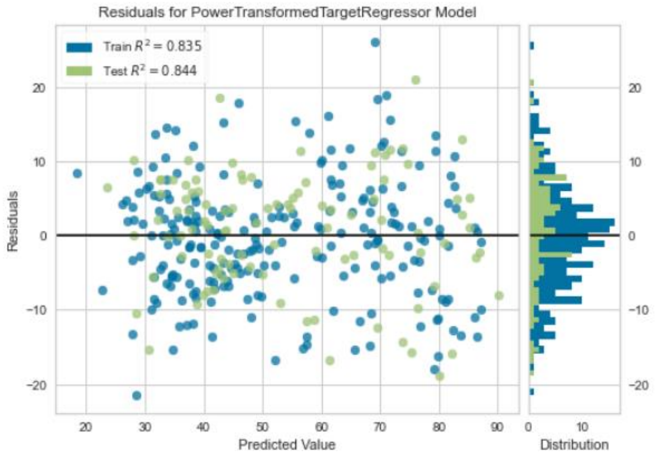
Although Bayesian Ridge and Ridge regression produce better results, we chose to test this hypothesis with linear regression to observe and learn how different ML models work.

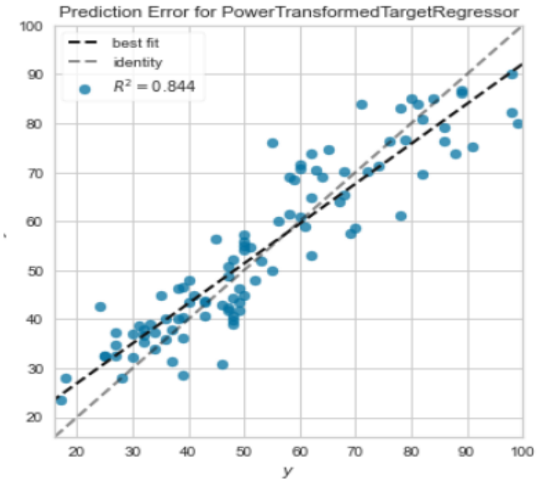
Our initial value recorded for MSE was 71.5546, RMSE value was recorded as 8.3997, R2 value as 0.7867. After tuning and optimizing, our MSE value was recorded as 55.2024, RMSE value was recorded as 7.4298, while R2 was recorded as 0.8583 on test data. Our final MSE value recorded for our 10% unseen data was 55.2024, the RMSE value we got was 7.4298, and our R2 value was 0.8632.

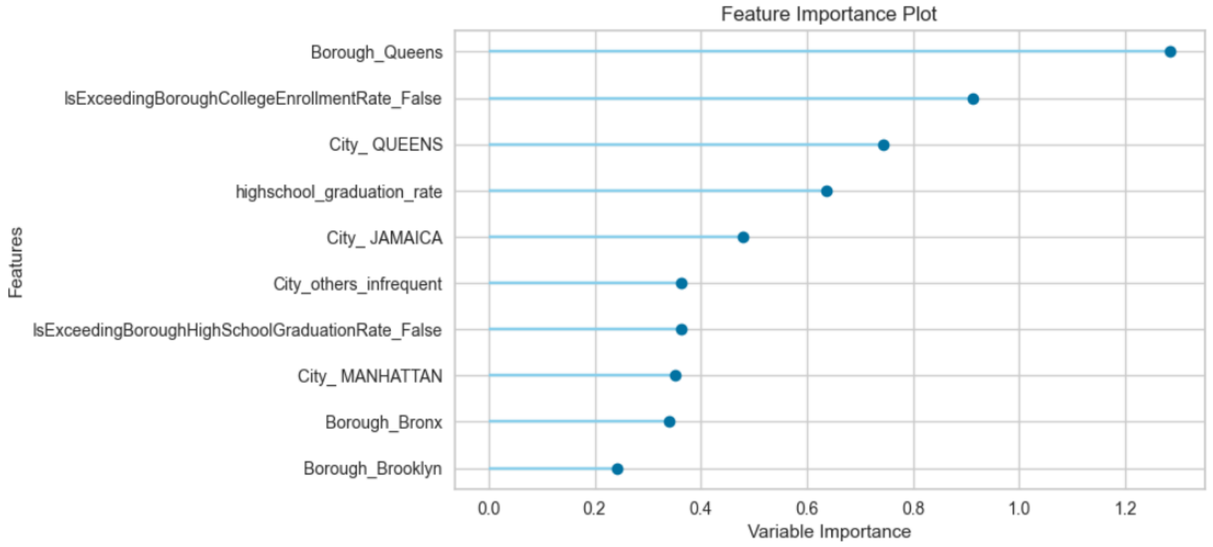
What does this say about our initial hypothesis? Are there any relationships between geographical locations and college enrollment rates?

There is undoubtedly a connection between geographical variables and college enrollment rates. However, geographical variables and college enrollment rates do not share direct links. The relationship between geographical variables and college enrollment rates is similar to the relationships shared between geographical variables and high school graduation rates.

A careful examination of our most important feature plot shows us that the geographical location with the most impact is the Borough of Queens, represented by 67 schools. However, the geographical location is insufficient as we need to add more variables such as our relative Borough college enrollment rates and high school graduation rates to get more insightful results.







ML Algorithm: Method 6 - Huber Regressor

For Method 6, we checked to see using the Huber Regressor algorithm to confirm any relationship between all the variables available and college enrollment rates.

We ran our data against 17 models, and we saw that Huber Regressor would provide the 4th best initial results. We could not choose our first (Bayesian Ridge) or second (Ridge

Regression) best algorithm because they had been used before to predict high school graduation rates.

During our initial comparison against different models, we used Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), R-squared (R2), Root Mean Squared Log Error (RMSLE) and Mean Absolute Percentage Error (MAPE) as our metrics.

It was important to maintain consistency even as we compared results between Linear Regression and Huber Regressor so; we used 70% of our data to train the model, 20% of our data was used for data testing and validation, while the remaining 10% of our data was used as unseen data to evaluate the final results produced by the Huber Regressor.

The main parameters we used for evaluation were MSE, RMSE, and R2. MSE was selected to gain appreciable insights into the differences between original values and predicted values. We also chose RMSE to achieve a well-balanced model. Finally, we chose R2 since it directly measures the goodness of fit in capturing the training data variance. We tuned our RMSE parameter to produce better results.

Although Bayesian Ridge and Ridge regression produced better results, we chose to test our hypothesis using Huber Regression to observe and learn how different ML models work.

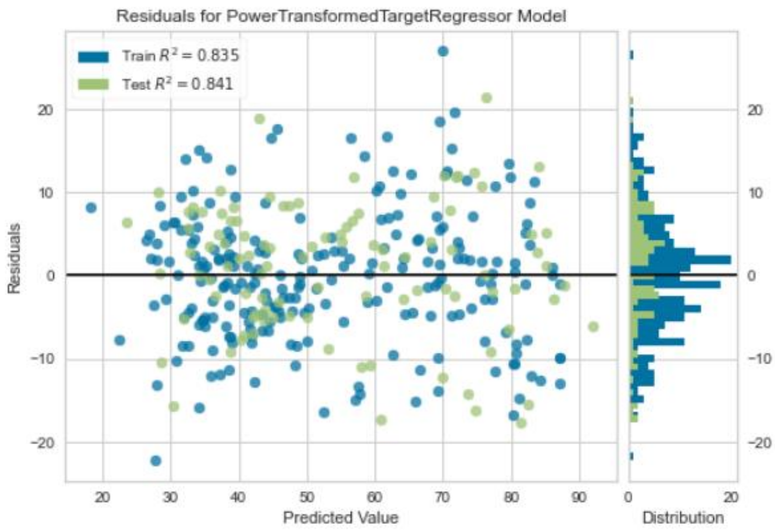
Our initial MSE value was recorded as 71.7504, RMSE value was recorded as 8.3835, while R2 value was recorded as 0.7856. After tuning and optimizing, our MSE value was recorded as 55.2677, RMSE value was recorded as 7.4342, and R2 as 0.8581 on test data. Our final MSE value recorded for our unseen data was 55.2677, the RMSE value was recorded as 7.4342, and our R2 value was recorded as 0.8632.

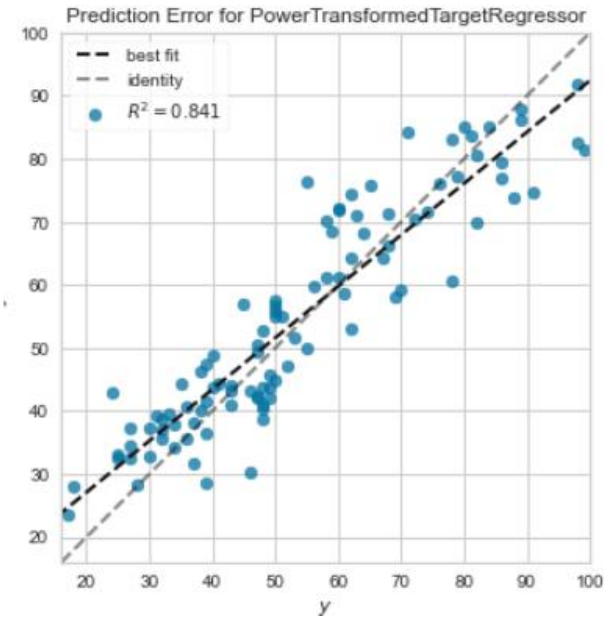
What do these values mean for our hypothesis? Furthermore, what do they say about the relationship between geographical locations and college enrollment rates?

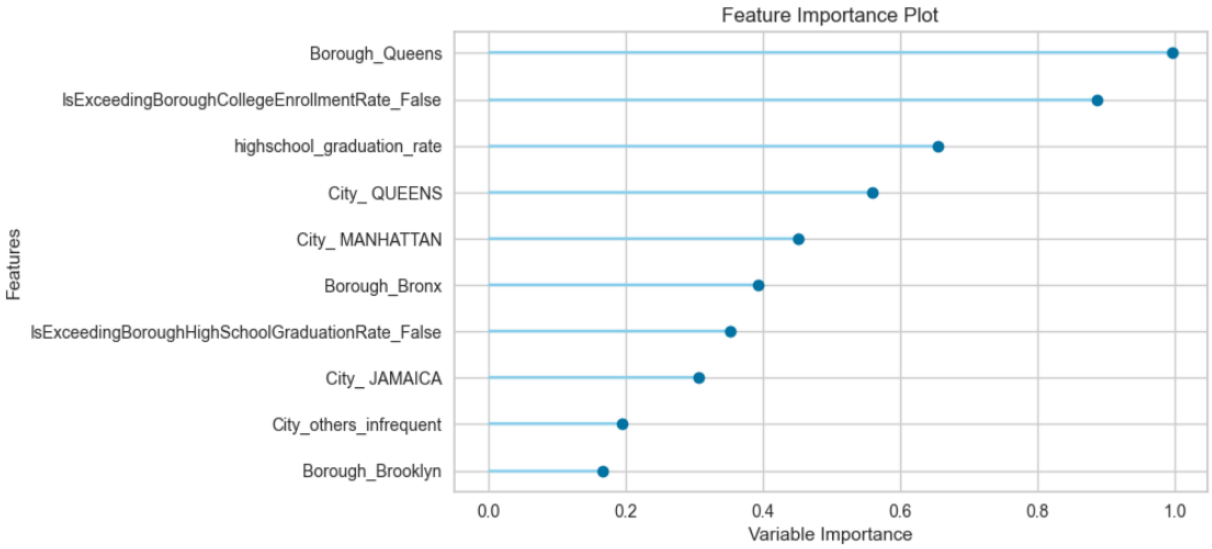
Again, the results produced by the Huber Regressor algorithm is very similar to that of the Linear Regression algorithm. Our Huber Regressor algorithm results also indicate that there is only a partial connection between geographical variables and college enrollment rates.

The results' similarities are even evident when comparing our important feature plots for the Huber Regression algorithm and the Linear Regression algorithm.

For Method 6, the Borough of Queens is the most important feature, and it has 67 representative schools. The variables tested are, however, insufficient for us to reach this verdict. We need to add more variables such as our relative Borough college enrollment and high school graduation rates to get more insightful results.







Conclusion & Recommendation

From our EDA Analysis, we can infer that our hypothesis about the Borough of Manhattan will perform best in terms of high school graduation rates, and this will remain true irrespective of the Borough under comparison to it. However, our hypothesis that suggests that the Borough of Bronx will perform worst in terms of graduation rates was wrong.

The proviso for these inferences rests heavily on our skewed data, which does not feature graduation rates for all schools in each Borough. Some Boroughs have better representation than others from the data we gathered. Brooklyn, for instance, was represented by 107 schools, the Bronx was represented by 99 schools, while 86 schools represented Manhattan. Sixty-seven (67) schools represented Queens, and ONLY TEN (10) schools represented Staten Island. If we had an equal amount of distribution for each of the Borough, we could have produced better results.

Our hypothesis about higher incomes and their effects on higher graduation rates was not wholly correct. There is no direct relationship between graduation rates and median income. Besides, our model does not yield insightful results when we use graduation rate as our only variable. We had to add more variables such as Borough, Zip code, and its relative Borough high school graduation rates and college enrollment rates to get good results.

Our hypothesis about the direct relationship between a geographical location and high school graduation rates is not entirely correct. There is undoubtedly a relationship between geographical variables and high school graduation rates; however, this relationship is not a direct linear relationship. As a result, our hypothesis was only partially correct. We can also observe the impact of these variables in our important feature plots. To get the best results for high school graduation rates, we need to add more variables such as relative graduation rates to the Borough and college enrollment rates to get insightful results.

Our hypothesis about the direct relationship between a geographical location and their college enrollment rate is not correct. There is undoubtedly a connection between geographical variables and college enrollment rates; however, we cannot perfectly say it is a direct relationship in the same guise as high school graduation rates. If we consider our important feature plot, the most impacted geographical location is the Borough of Queens, which 67 schools in our data represented. To get a more accurate result, we need to add more variables such as our relative Borough college enrollment rates and high school graduation rates to get more insightful results.

NYC is one of the wealthiest cities in the world. According to Business Insider, there are 113 billionaires in NYC. If one is part of the top 1%, there is a good chance that they are not sending their kids to public schools. Our data only represents NYC public schools.

One factor that is also strongly limiting is social mobility. A poor person will have a hard time getting into the middle class. A person in the middle class will have an even harder time getting to be part of the wealthy class. Now, amongst the wealthy class, there are different levels also. For example, a person who is making 100k a year can be considered wealthy; likewise, a person who makes $1 billion in one year is also wealthy.

If we examine our mapped data for college enrollment, the schools that are in the poverty line have a hard time with college enrollments. The rich here also have low college enrollment figures. One can infer that since they are already wealthy, going to college does not look attractive because they know it will not help that much, but that will only form another hypothesis for us to test. What we can confirm is that most middle-class students are also striving to go to college.

The government can provide incentives for poor and middle-class areas so that students can succeed no matter their financial background. We can achieve this by taxing the ultra-rich and those who make the most money in NYC but do not pay the most tax ratio. If we can do that, then we can provide many incentives to help poor students and middle-class students to increase the high school graduation and college enrollment rates for their locality.

Work Cited

Hoffower, Hillary. “The Top 10 Cities in the World for Billionaires, Ranked.” Business Insider, Business Insider, 2 July 2020, www.businessinsider.com/where-do-billionaires-live-top-cities-worldwide-ranked-2019-5.