Predicting High School Graduation Rates

Samantha White, Alex Gieger, Brett Bastianelli, Maria Ng

Professor Lando

IST 718 – Big Data Analytics

9/25/2022

Table of Contents

Executive Summary

Descriptive Summary

According to the US Census Bureau, published in 2017, “Between 2000 and 2017, the percentage of all people aged 25 and older who had not completed high school decreased by more than one-third, dropping from 16 percent to 10 percent”. While this seems like a trend in the positive direction, this statistic only views the entire population as a whole and not the subsets of individuals with various underlying factors. Throughout this project, our team's goal is to analyze these various factors and predict the academic success based on different variables such as income, divorce, and living conditions.

As the project progressed and our group developed varying models and analytical tools (Expanded upon more below), two key notes stood out as the most important. The first note is that certain factors such as divorce rate, state, median age, median family size, population, and the percentage of people not in the labor force have an impact on the high school graduation rate. This may seem like a rather simplistic observation, but it is important to note because it further backs that each variable has a direct impact on the graduation rate. Also, it validates that our project will provide insight to which variable has the largest impact and will be the most beneficial to make a recommendation on. The second note that our group came across was median age was the leading indicator for graduation, followed by the percentage of people not in the labor force. This was rather interesting to our group as we had prior expectations of different variables having a larger impact but through our analysis that we will disclose in the coming sections, median age, and people not in the labor force has a large influence upon the data.

Based on our data analysis and findings, our team developed two key recommendations that would look to benefit those impacted by low academic success due to these underlying variables. The first would be to develop tax incentives leading to a growth in high-paying jobs in the area and the second recommendation would be to host government-sponsored job fairs within the school. Both recommendations center around providing individuals who have faced hardship and may not have garnered a high school degree but would still provide value to the work force and enrich their lives. Both the tax incentive leading to high-paying careers and government-sponsored job fairs would ultimately combat both the median age and individuals not in the work force by providing labor opportunities to individuals who may not have completed high school.

Specification

With the average graduation rate for the United States in 2021-2022 sitting at 88%, it is important as ever to look at why that is. An increase in education, with high school graduation alone, individuals are provided with further opportunities to earn a higher income, gain a better living condition, and overall increase the health of the general.

Census data will be used to analyze the effects of different variables such as divorce rate, state, homeownership, and family size to predict academic success, outlined by the rate of graduation. By analyzing these variables, we will look to understand the longstanding effects and outline potential programs to help students who are suffering academically, to increase the graduate rate.

The census data was curated using over 12,000 different files spread across multiple folders buried deep in the backend of the census website. The backend of the census website can be found via the link www2.census.gov. To pull each of these files web scrapping algorithms were developed. In total three distinct categories of files were pulled. The first was the Gazetteer files, the second was the actual census data and the last set of data was the column names associated to the census data which had no column information. Once all three categories of data were extract they were combined to create the census data file.

The census dataset was combined with additional data. The data was cleaned by setting all NAs to 0, and duplicates were removed by grouping the data frame by zip code taking the median of each column. Each row in the dataset represents one US zip code.

Observation

When exploring the data, the following observations were made:

1. Figure 1 shows a boxplot representing the distribution for our variable of interest, high school degree rate. We can see that the median falls around 0.9. The distribution is left skewed with many outliers below 0.6. We do not want to remove these outliers since they are essential to understanding what contributes to low high school degree rates.

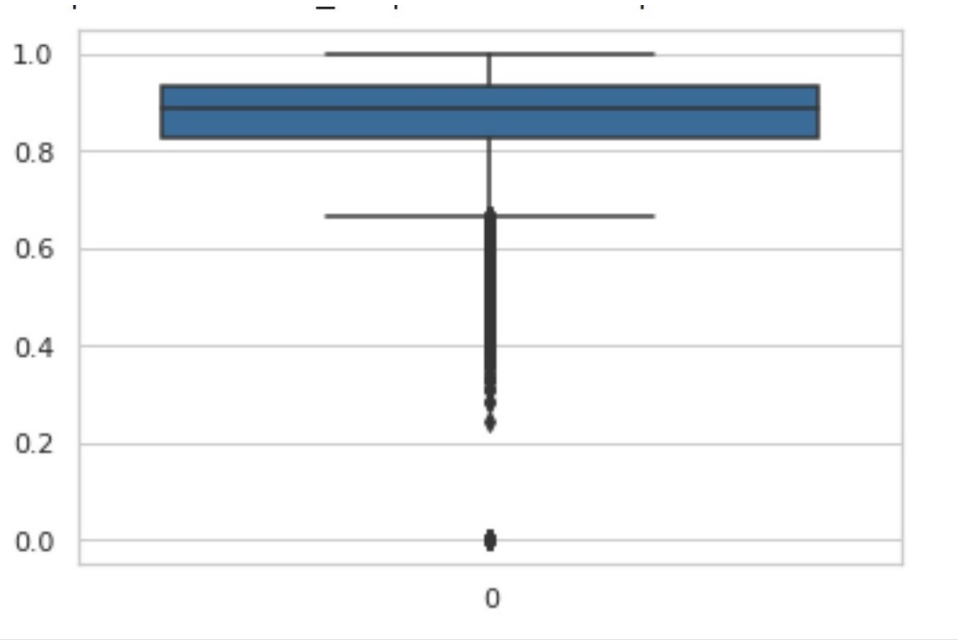


Figure 1: Boxplot of high school degree rate

2. When diving deeper into the zip-codes with a high school degree rate under 0.6, we see that California, Hawaii, Rhode Island and Texas have the highest number of zip-codes normalized by population that fall into that category.

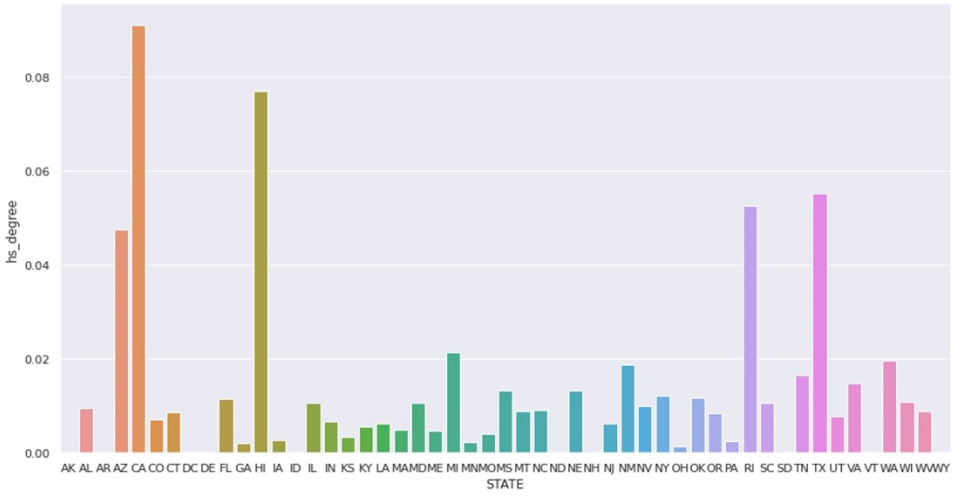


Figure 2: Bar chart of zip-codes by state that have a high school degree rate of lower than 0.6 normalized by population

3. Figure 3 shows us the correlation between each of the desired variables for the model. We can see that rent, age, family size, and elderly have the highest correlations with high school degree. We can also see that “prep”, “n2”, “numdep”, “total vita”, “pac” and “elderly” are highly correlated with each other, which means that we can only accept one in the model. Since “elderly” has the highest correlation with high school degree, we will use it in the model.

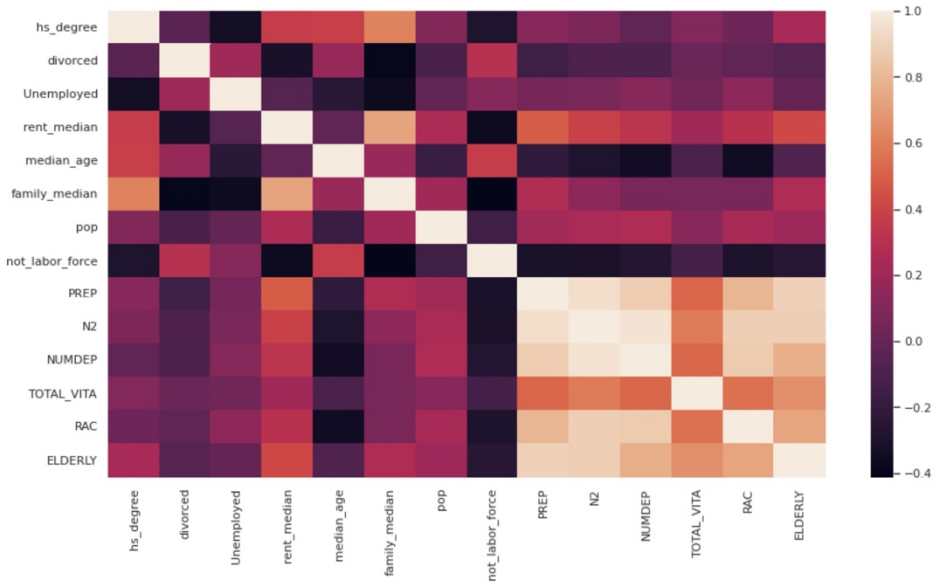
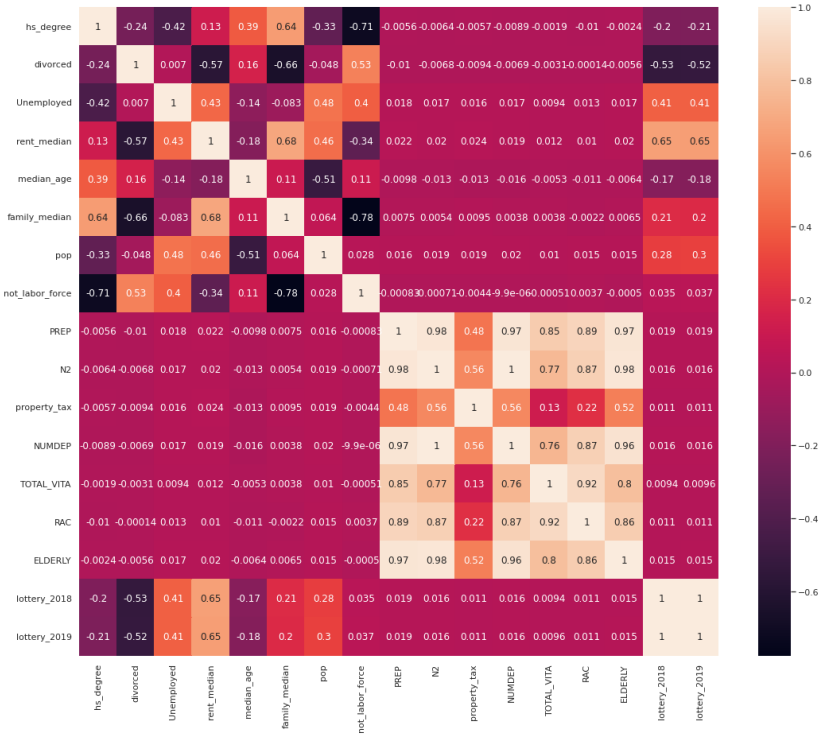


Figure 3: Correlation matrix to better understand how correlated each variable of interest is with hs_degree and each other

4. Figure 4 shows us the correlation with added variables, property tax and lottery. We can see that there is no strong correlation between high school graduation rates.



Analysis

Three models were used to predict the high school graduation rate of a zip-code. The first was an OLS regression model that took the following input variables: divorced, median age, family median, population, state, not labor force and elderly. To produce a more accurate model, the high school degree variable was squared. The model resulted in an Adjusted R-squared of 0.623.

The random forest model used all numeric input variables to predict the high school degree rate. The random forest model had a RMSE of 0.00047

One model was generated to predict the high graduation rate using dataset group by state. The OLS regression model was constructed based on the following variables: divorce, median age, not labor force, unemployed and property tax. The model resulted with an adjusted R-square of 0.746 with property tax was not significant with a p-value of 0.196.

Advanced Modeling

Introduction

The advanced modeling section utilized the Microsoft LGBM models to predict the high school graduation rate of a particular tract. Census tracts instead of zip codes were used in the advanced modeling section because they are smaller than zip codes. On average, every zip code has 3.792 unique census tract. The zip code which has the highest number of census tracts is Costa Mesa in California which is located right outside of Long Beach California. In total there are 43 unique tracts located inside of Costa Mesa.

Model Validation

The model validation technique was K-Fold cross validation. K-Fold cross validation is when the data is separated into equally sized segments called folds. After, one of the n folds is held and used for testing while the other n-1 folds are used for training. In total n models are iteratively trained and saved to a model's data structure. The reason K-Fold cross validation was utilized instead of a train test split was to ensure all the data was used for validation.

Model Results

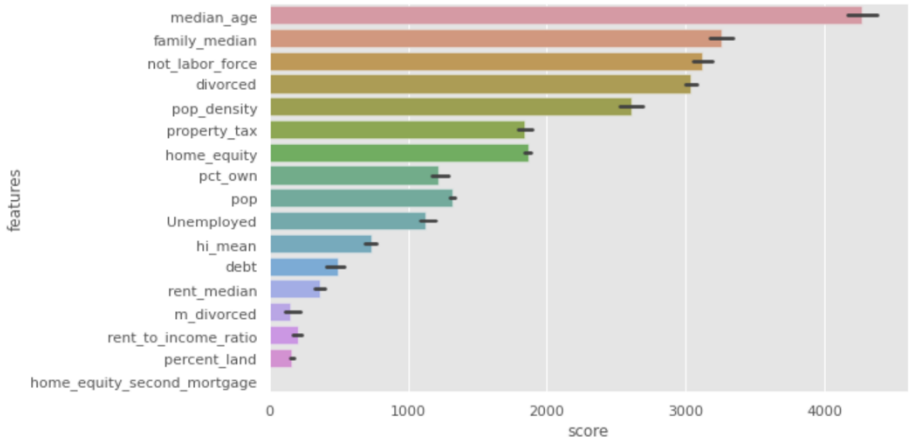
In total 5 models were trained one for each of the five folds. Of these five models, each was evaluated using several different metrics. These metrics which were utilized to evaluate the training and testing dataset included mean average error, mean percent error and root mean square error. The average score for each of the five metrics is provided below.

data	fold	mae	mape	rmse
test	3.0	0.044430	0.058082	0.062608
train	3.0	0.043125	0.056209	0.060254

Feature Insights

Every model that had been trained had a unique score for each feature in the model's feature set. Since multiple models had been trained, one for every fold this created a difficult problem for the visualization of the results. Therefore, the seaborn boxplot definition was utilized which statistically encapsulated the multiple values for each features score in a beautiful visual format.

The visualization speaks to how the leading indicators of the model are the median age followed by the median family income. From the analysis of the data, it is evident that the median age and high school graduation rates are positively correlated. Therefore, as the median age increases the graduation rate increases as well. In addition, the family income is also positively correlated to high school graduation rates. Therefore, as the median income increases so does the high school graduation rate.



Generating Predictions

Since there were multiple models, one for each of the folds an ensemble approach was utilized to produce a single prediction. An ensemble in Machine learning when multiple models are combined to produce a single more optimal solution. Since the model consisted of a continuous value the average of all models were taken to produce a definitive answer.

Recommendation

Introduction

For the project, the team wanted to go beyond predictive statistics and look to produce recommendations to help struggling communities increase their graduation rates. First, the team needed to identify a specific community struggling with getting students to graduate. The census tract the team had Identified was in Salians City California which had one of the lowest graduation rates in the country coming in at 28%.

Methodology

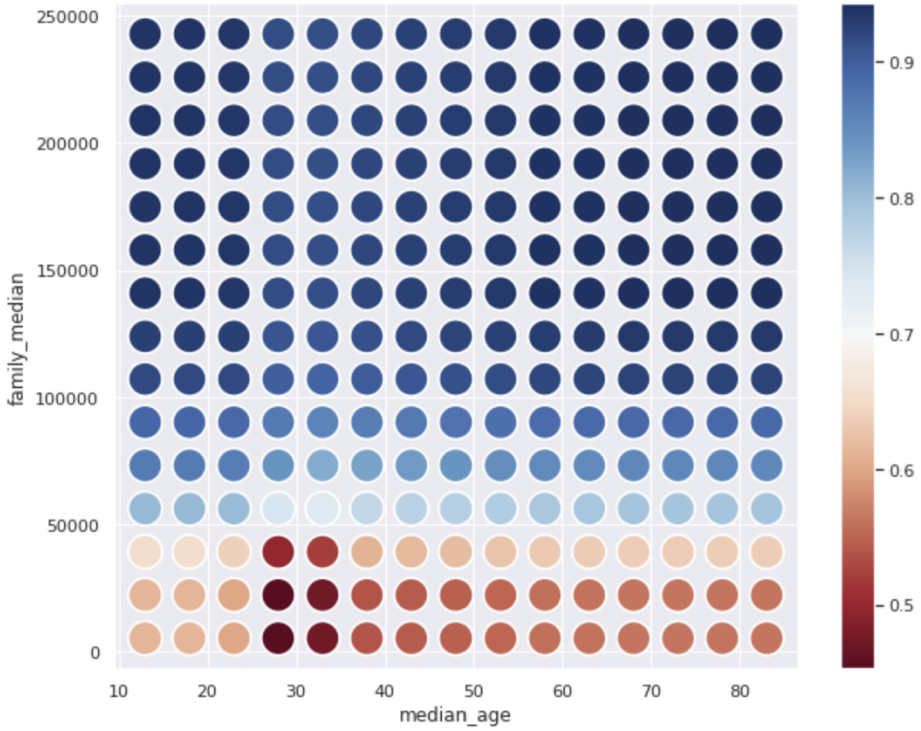
To inspire recommendations for the struggling community the team augment the features of the community and evaluate the impact of the augmentation on the community using the trained models. For instance, if the community were to increase its family income what impact would that have on high school graduation?

Specifically, the analysis that was conducted performed bivariate sensitivity analysis to quantify how changing two variables impacted the high school graduation rate. In short, a two-dimensional grid was generated. On the x-axis was variable 1 and, on the y-axis was variable 2. The value of the element corresponding to variable 1 and variable 2 was the high school graduation rate. All other variables or features in the model that were not variable 1 and variable 2 were kept static.

Results

The following vitalization depicts the impact of augmenting the family median income and the median age of residents of Salians City in California. From the analysis the median family income is an important driver of high school graduation rate. Therefore, the team recommends a job-fair to bring high paying jobs to the local area thereby increasing the graduation rate and solving this systemic problem. The job-fair would allow for individuals who may not have completed their education to see the many different career opportunities that are available. A large part of the education and financial disparity in America, at the moment, seems to be the lack of available information. Individuals may not know of what career paths would best fit and take the first paying opportunity to support themselves and their family. This career fair would break that cycle by providing information to what professional endeavors are available and how to apply for them. It appears in this instance that money may just solve the problem, this time.

Along with the recommendation of a job fair, the opportunity of developing tax incentives leading to a growth in high paying career opportunities in the area also shows promising signs. This recommendation would take more planning as we would be looking to directly influence the government rather than hosting a job fair, but it could lead to positive results for those impacted. If large companies were to receive tax credits by hiring individuals without a high school degree, it would be a mutual benefit. Individuals who were negatively impacted and could not complete their education due to extenuating circumstances such as a parental divorce, would receive a well-paying job that fits their skill set while the global company would receive a tax break at the end of the year. Along with a well-paying job, these individuals who are a part of the tax incentive program could also receive further funding or help with completing their GED to help make them more of a well-rounded employee and open to further advancements within the company.



References & Appendices

1. https://nces.ed.gov/fastfacts/display.asp?id=805
2. https://www.census.gov/data.html
3. https://www2.census.gov/geo/docs/maps-data/data/gazetteer/2016_Gazetteer/
4. https://www2.census.gov/geo/docs/maps-data/data/
5. https://www.taxpolicycenter.org/statistics/lottery-revenue
6. https://www.irs.gov/pub/irs-soi/18zpallagi.csv