Predicting High School Graduation Rates

By

Kevin Bryson

8/14/2018

Introduction

Across America school districts are examining ways to increase the graduation rates of their students. Money is being invested and programs are being developed to better service student populations. Data is more important than ever in understanding students and how to adjust programs to maximize outcomes. Schools have to know which variables are critical to the development of applicable and successful programs and to predict future graduation rates.

Data Sources

Data was collected from the following sources:

* California Department of Education Website
* [www.kidscount.org](http://www.kidscount.org)

Data Management

Pandas was used to load the datasets into a Jupyter Notebook. All data was cleaned to eliminate any NaN values and the datasets were then merged together into one matrix (m5).

Data Evaluation

The most correlated variables were calculated against the target variable of Cohort Graduation Rate using spearmanr correlation coefficient. Female, poverty and insurance coverage all had the highest positive correlation. Male had a high negative correlation.

Each of the variables were then graphed using seaborn boxplots and swarmplots respectively.

The higher the poverty level the higher the mean for the group. This would support the correlation value of 0.25.

There is a slight correlation between students with insurance coverage and the graduation rate. The spearmanr value for insurance is .25 which is the same as the poverty number.

The largest correlation is the gender column. Females have the highest correlation with a spearmanr value of 0.38. Males have a negative correlation value of -0.36.

Some counties have health centers in place. They seem to have a very small impact on graduation rates with a spearmanr value of 0.09.

Machine Learning

Once the matrix was created it was time to set up a machine learning model that could make accurate predictions. The variable being predicted is not one that is binomial in nature. It is a value that has more than two outcomes. Linear Regression and Lasso were the two machine learning models I chose to run. I first had to put my data with object values into numbers using Label Encoder. I then set up 2 pipelines for each model using the parameters of intercept for linear regression and alpha for lasso. I then used a dictionary to set up parameters that could be passed through GridSearchCV. I used the scoring method of r2 to find which method and which parameters had the highest prediction rate. The results showed that using Lasso with an alpha value of 0.05 gave an r2 score of .90 or 90% accuracy. I created a table to show the test value vs the predicted value.

Recommendations

There are several recommendations that can be made using the results of this information:

1. There needs to be more emphasis on programs designed to help male students be successful. These programs need to be focused on the importance of a high school diploma using community role models and mentors. Further research could include the number of male students that come from a single parent household.
2. Health centers need to offer basic medical services to their student populations including check ups and physicals. A doctor or someone with the authority to write basic prescriptions need to be on staff. Further research could include absentee numbers per year per student and how that affects graduation rates.
3. There needs to be an emphasis at universities to recruit and hire male teachers. These teachers need to represent all grade levels of education. A marketing campaign at schools that emphasis educational studies needs to be developed. Further research would include the number of male teachers per school compared to test results.
4. Teachers, parents, administrators and community leaders need to be directly involved with stimulating learning. There needs to be activities that begin during the day schoolwide and then continue that same evening. An example of this could be a Leonardo Davinci Day where during the day lessons are focused on his contributions to each discipline and then a family night that can include activities for parents, students and the community.