Predicting High School Graduation Rates

By: Kevin Bryson

Springboard

Introduction

One of the biggest challenges faced by school districts across the country is trying to increase high school graduation rates. There are many factors that can influence whether or not a student graduate’s or not. This report looks at some of those variables and tries to determine which ones are the most important and then design a model that can be used to predict graduation rates.

Target Audience

The information that will be presented in this study will be helpful to school districts, government agencies and any business that advises districts or creates programs for at risk students.

Data and Data Wrangling

Data was found on the internet from two sites. One of the sites was The California Department of Education and the other was from the San Bernardino County website. I picked San Bernardino County because I was a teacher in the county for 10 years. The data was loaded into a jupyter Notebook and then processed using pandas library. All null values were removed and the dataframes were combined on the year column and given the name m5. I then created a set of dummy variables that could be used in a regression learning model. I chose Linear Regression and Lasso Regression and set up a pipeline for each. Then each pipeline was passed through GridSearchCV to obtain an R2 score and the best parameters.

Data Analysis

I used spearman correlation to determine the 10 best variables. I then eliminated any variable that was a specific ethnicity and was left with gender and number of insured children in the county. The spearman value for females was 0.38 and for insurance it was 0.24. Males had a negative correlation of -0.36. I then used seaborn to graphically represent the correlation using a scatter plot to represent graduation rate vs. Insured rate and a boxplot to represent gender and graduation rate

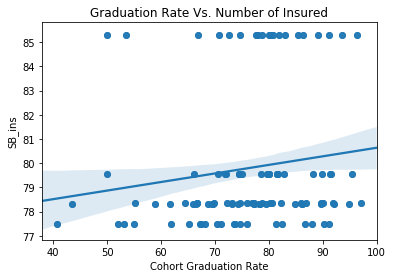
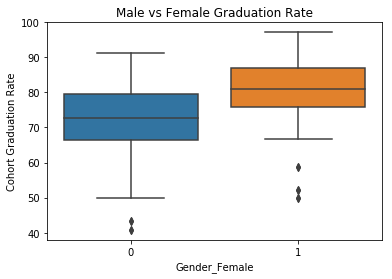


Figure represents graduation Rate vs Insurance Rate



Boxplot of Gender and Graduation Rate

After evaluating the graphs, each regression model was pipelined and then ran through GridSearchCV where an R2 score was calculated. Linear regression had the highest R2 score of 0.893.

Recommendations

1. Each school district needs to create a program for its male students that provides positive mentors. The program needs to be diverse in nature and community involvement would be ideal. The program needs to start at lower grade levels to be successful and continue through graduation.
2. Each district should have a health center for its students that provide a licensed doctor, nurse practitioner or physicians assistant. Basic healthcare needs will be addressed so each student can be in school as much as possible. The initial cost for the program could be absorbed through grants provided by the federal government or private entities.
3. An effort needs to be made to recruit and hire male teachers at all grade levels and train those teachers so they can better service their male students.

Further Investigation

There are more variables that could cause high school graduation rates to fluctuate. The number of students receiving free or reduced lunch is one. An examination of discipline by ethnicity could be important. Finally, parent’s education level could be examined to see its effect on graduation.