PPOL564 Final Project: Predicting Election Results

Abigail Paterson

Word Count: 2675

Introduction

For this project, I chose to investigate the effect of certain school conditions on student emotional well being. I will use data from California to investigate if student gang involvement, bullying, and truancy have an affect of student feelings of depression. I am using feelings of depression as a measure of emotional well being. In this report I will explain my data and how I gathered, cleaned, and organized it. I will then discuss the features of my data and how the characteristics of my data may impact my results. Lastly, I will create and explain a machine learning model to investigate how each of my features can predict student emotional well being. I also intend to discuss how my results vary on certain subsections of my data

Background

I have prior experience working with data about children, such as investigating maternal mortality rate and demographics, and how COVID-19 impacted early childhood education.

Due to this background, I was interested to investigate something related to child wellness, and this time, I decided to focus more on emotional well being than physical well being.

There is a tremendous amount of existing research on what factors influence childhood well being. Factors such as poverty level, housing situation, cultural background, environmental risk, and more have been proven to be predictive of childhood mental health and wellness. It is also well known that school has a significant impact on mental health in all ages. This is partially related to the stresses of being a student, such as grades and homework, but also because the school environment is home to many unique social and environmental factors that have an impact on emotional wellness.

The data I used originally comes from the California Healthy Kids Survey[3]. This survey has a subsection on school climate. This identifies five most important areas for school social climate assessment. These factors are student connectedness, school culture, school safety, physical and mental-well being, and student supports. This survey is designed to provide data that will improve school population resilience and has found that these factors heavily affect student outcomes in school related areas. Positive academic environment in general has been linked to positive psychosocial and health outcomes in youth. This study looks at all of its factors as a holistic measure of a school, intended to provide guidance for school staff or policy makers. I decided to look more closely at the individual features and see how the factors that impact school quality overall affect each other.

Data

Data Source

To accomplish my goal, I looked for data pertaining to student emotional health and any other factors that I could use to predict emotional health. I took my data from the website KidsData[1]. This organization is a program of the Population Reference Bureau[2] that focuses on promoting the well being of children in California through providing data on the topic. I decided to use this source because it has many different data topics all collected in a relatively consistent way.

Data Collection Issues

I ran into numerous struggles in gathering my data. The very first problem was that I could find very little data that was taken in comparable years or location. There were many more factors I originally wanted to look at, but I could find no data that was taken on the same scale as my other factors. This caused me to focus on school data as I did.

I initially found the website that contained my data, and from checking the "Inspect html" functionality of a web browser, I determined that the data was indeed formatted into tables. However, when I tried to use `read_html` to pull my data, I was unsuccessful. The results from `read_html` did not recognize any tables in the scraped html. I took a closer look at this problem, and by reading through the entire website code that BeautifulSoup scraped, it became clear that the website did not load the way it appeared in "Inspect html." I am not completely sure why this is, but the conclusion I came to was that the website was loaded

"dynamically," meaning that the html scraped from the website would not be the same as I was able to inspect

In order to solve this problem, I attempted to use Selenium. Selenium is a Python library that is used for automating webtasks, and it can be used to access html that is otherwise inaccessible. Using selenium, I opened what is called a "webdriver," which is what allows your code to run web browser tasks. Once this driver is open, it functions much in the same way as BeautifulSoup, and `read_html` should be able to now detect the tables I was not able to find initially. This did seem to work, at first. Using Selenium, I did successfully scrape the tables from my website. However, the tables were formatted in a very odd way that I could not understand. I still do not know exactly what was the issue, but `append` was unable to append the tables to each other, and I was not able to convert the tables to a DataFrame in any other way. The error I received was that the tables were three-dimensional, and only a two-dimensional object was allowed, but I was not able to successfully isolate the two-dimensional form of the tables.

After this, I was left to download my data from the website. For my three independent variables and one dependent variable, I individually downloaded the data for each year available. I then was able to continue the rest of the project.

Data Format

My data covers each county in California, and pertains to children in 7th Grade, 9th Grade, 11th Grade, and non-traditional schooling. The data was taken over the course of three 2-year periods. 2011-2013, 2013-2015, and 2015-2017. To simplify analysis, I labeled each

group by its end year. The overall unit of analysis is one county per year per grade level. This format is at risk of both temporal and geographical autocorrelation, so it is likely that my data will not have as large of a sample size as it appears to.

Data Cleaning

All of my variables came in its own .csv file. One file for one factor for one year. Because of this each variable had to be cleaned and then joined together. Each variable had the same steps for cleaning. The first step was to create a column in each DataFrame for the year in which the data was taken. After that I joined the three yearly columns of each variable together which created one long DataFrame for each variable.

The next step after this was to determine which columns I needed to keep. An initial confusion I ran into was that each variable had a column for any positive responses and one for any negative responses to the question being asked. First I dropped all the rows holding the negative response data. I chose to focus just on the positive responses. Then I dropped the negative column so it would not be providing irrelevant 0's. I also dropped a column indicating what the location type is, as all of the data was taken by counties; what the data type was; and I dropped the original time-frame column which I had replaced with my year column.

The next step was to create dummy variables for my categorical variables. I created dummy variables for each county and for year grade level. I also renamed the columns of each variable's data, as in the original data, all of these columns were named "Data," I needed the names to be discrete and also more descriptive. Once all of this had been done, I merged

all of the variables together into one large DataFrame. The last step in my data cleaning process was to remove all the missing values. This considerably shrunk my dataset, cutting it almost in half. There were two values indicating missingness in my dataset, "S" and "NA." I had to remove these values as the final step of my data cleaning process to ensure that each row matched, and there was no data in one variable that was not in the others.

Variables

The four variables of interest in my DataFrame are:

Estimated percentage of public school students in grades 7, 9, 11, and non-traditional programs who were bullied or harassed at school for any reason in the previous year.[4]

Depression related feelings as measured by the estimated percentage of public school students in grades 7, 9, 11, and non-traditional programs who, in the previous year, felt so sad or hopeless almost every day for two weeks or more that they stopped doing some usual activities[5]

Number of K-12 public school students reported as a truant during the school year, per 100 students[6]

Estimated percentage of public school students in grades 7, 9, 11, and non-traditional programs who consider themselves gang members[7]

All of these measures were taken from the same source survey. I have included a histogram of each of the variables of interest as you can see in Figure 1

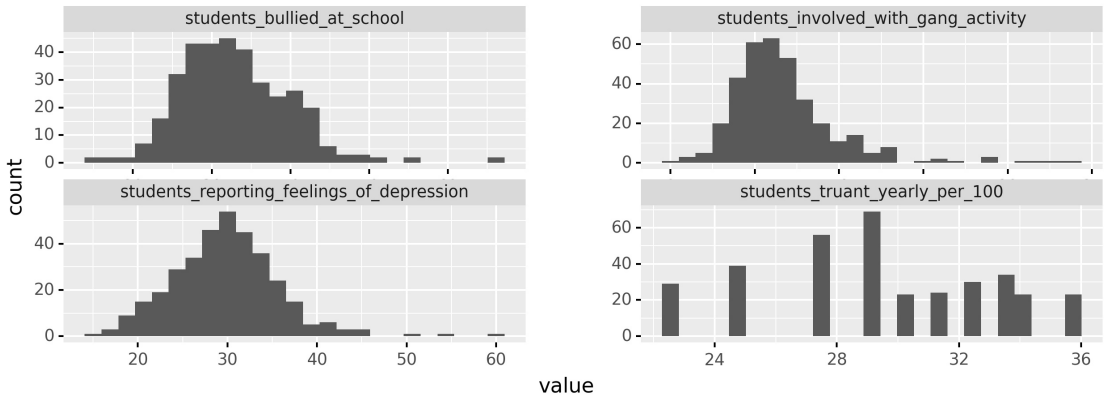

Figure 1: Feature Histogram

As you can see in these pictures, three out of my four variables have quite normal distributions, but Truancy is very staggered and not normally distributed.

My data is geographical, so I also created some visualizations to show the data by each California county. I did this with Geopandas and with a Plotly package called Figure Factory. These visualizations do have some limitations, because each FIPS code location can only have one values, so I had to isolate grade level and year for these visualizations. Below are maps of each variable by county.

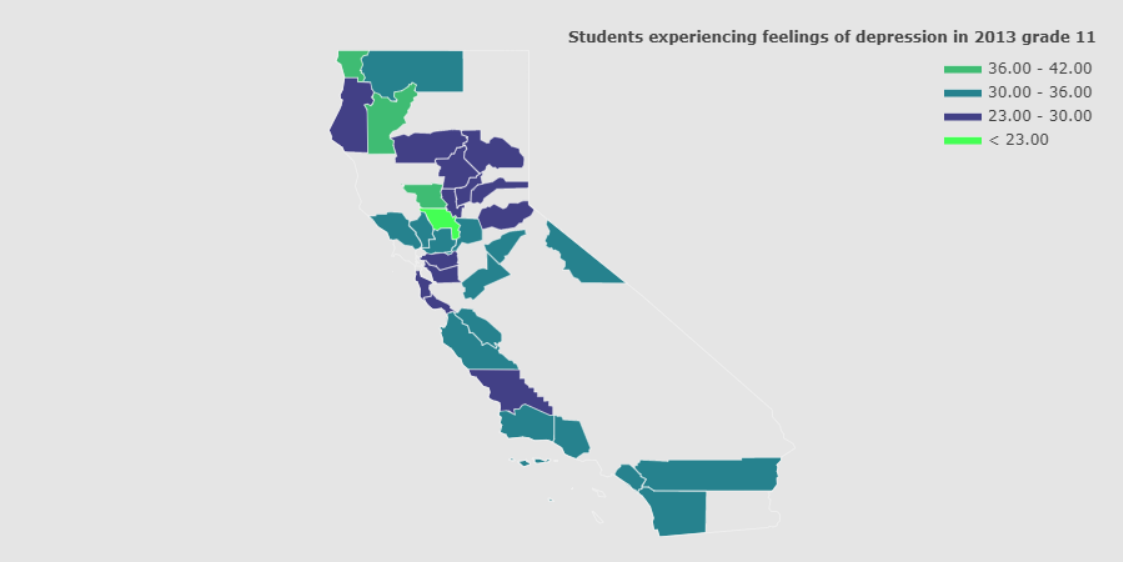

Figure 2: Students Experiencing Feeling of Depression in grade 11 in 2013

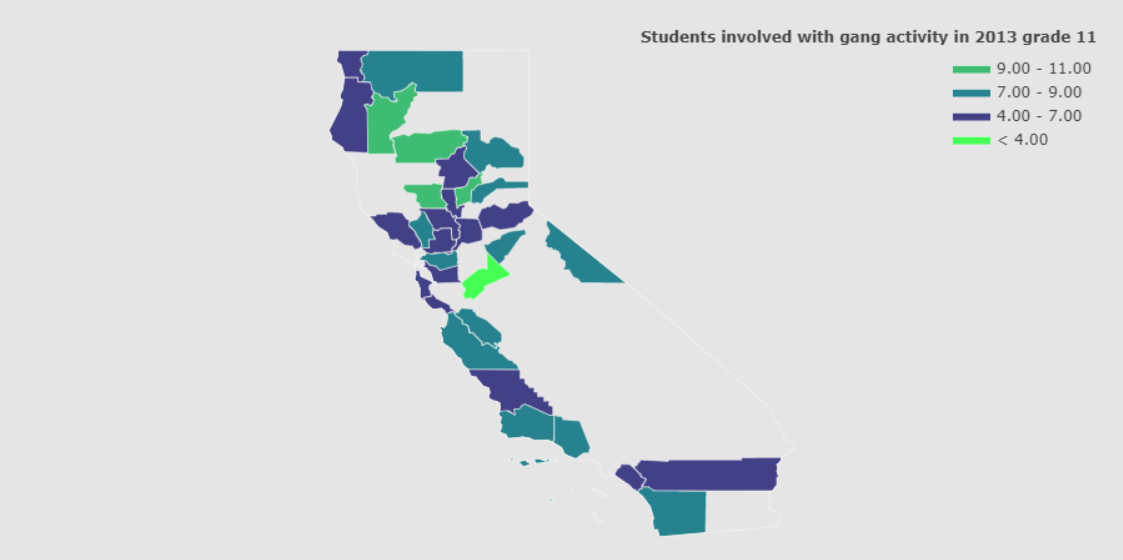

Figure 3: Students Involved with Gang Activity in grade 11 in 2013

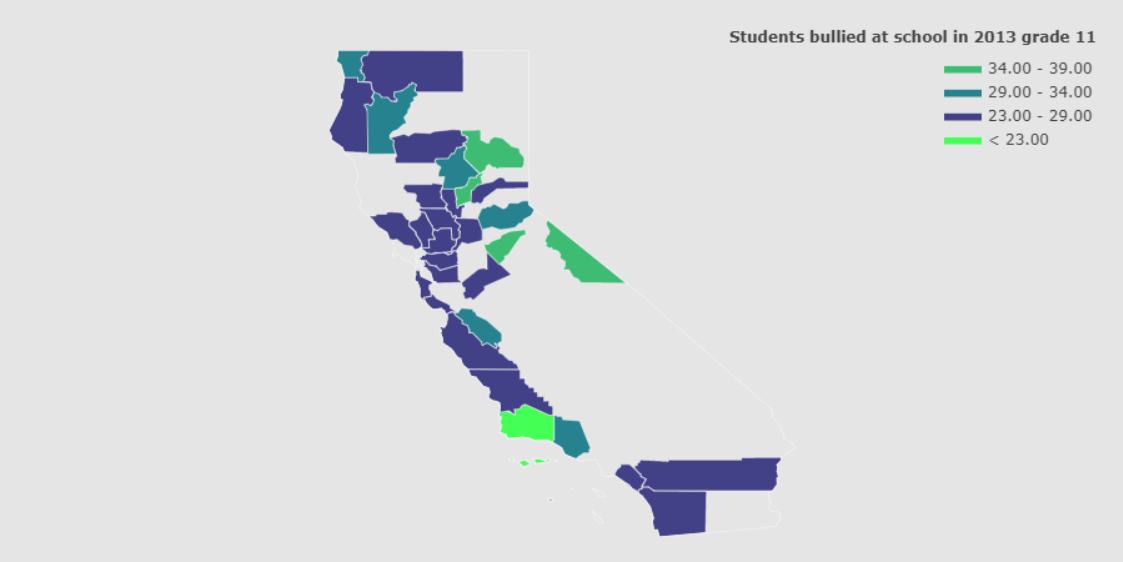

Figure 4: Students Bullied at School in grade 11 in 2013

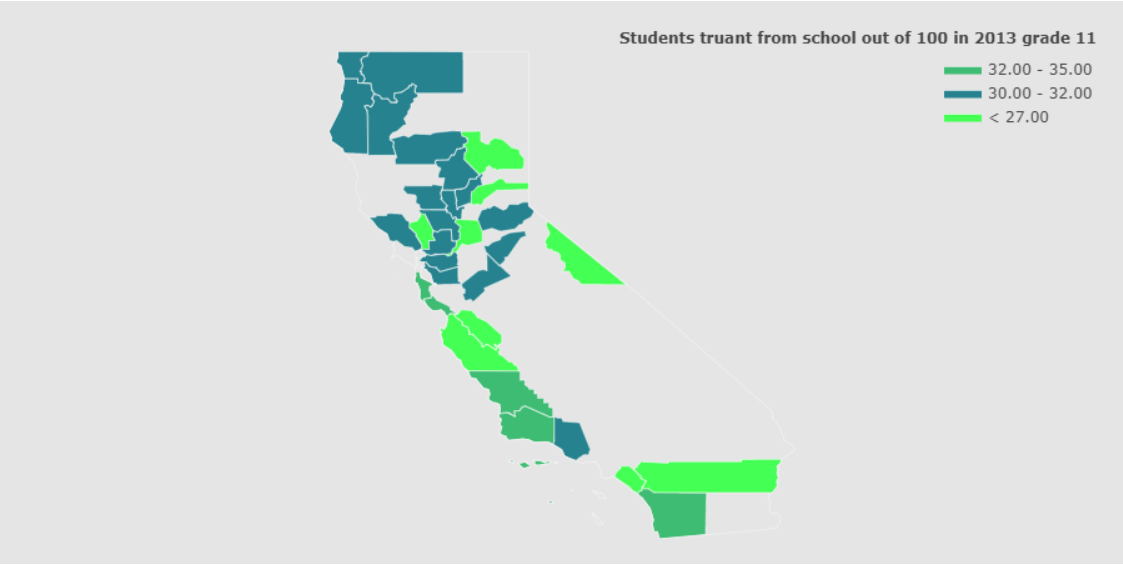

Figure 5: Students Truant from School per 100 in grade 11 in 2013

As we can see in all of the figures above, many counties are missing data all together. These counties are primarily the large counties in the Central Valley. This missingness in itself is an interesting way of analyzing the data. Why are certain counties more likely to be missing? This is a question to look at in more depth in a different project.

Analysis

My data predicted a continuous variable, so my models had to be able to do Regression and this mean that I had to use a scoring metric that functioned with regression. Therefore, I included a Linear Model, a Bagging Model, K-Nearest Neighbors, Decision Trees, and Random Forests in my search space. The scoring parameter I decided to use is `neg_mean_squared_error` because it is relatively simple to interpret - the lower the score, the better the fit, and the square root of this number will approximately represent the average distance between the predicted values and the actual values of the dependent variable. Because of my relatively low sample size, I split my data into training and testing data with 80% of the data going into the training set and only 20% in the testing set. I ran my analysis on two subsets of my data. First I ran the analysis on just the complete dataset, and then I ran it on a section that was just data points from year 2013 and grade 11. To run both of these analysis, I followed the same steps. First I split my data into X and Y subsets. The Y, the dependent variable, is my variable students_reporting_feelings_of_depression. My X data was all of my independent variables, this included year, grade level, county, and my predictors of interest. After I did this, I created a machine learning pipeline. This pipeline ran on a Grid Search, so I split my data into 10 folds to be compared. Then the pipeline performed the pre-processing step of scaling my data. Scaling data allows the variables to be better compared with each other. Next, the GridSearchCV tested each of the models to find the best one.

Results

As far as best model, my analysis found that the Linear Regression model was most predictive for my data. This does not terribly surprise me because this is the model that is best suited for regression out of the models I ran. The overall results from my analysis were not extremely predictive, but overall I am pleased with them. For my complete data, the negative mean squared error was -16.69, and the square root of this is around 4. This means that on average my line of best fit was 4 points away from the true data. Given that my dependent variable is measured in percent, this is not a bad scale for the error to be at. This analysis produced an r squared of .67, which is similarly not bad. The results were similar for the subsection of grade 11 and year 2013. This analysis had a negative mean squared error of 20.64, which indicates an average distance from the true value of Y of 4.5. This is slightly worse than the first analysis, and this likely comes from there being fewer data points.

Variable Importance

I calculated the variable importance for my model of the complete dataset. The results of this showed the three most important variables to be grade 7, grade 9, and grade 11. After this, students_bullied_at_school was the next most predictive. After this, the importance was taken over largely by county variable. Gang involvement was 6th most important, and truancy was 27th most important. Below in Figure 6, you can see the variable importance of each variable. The only of these variables that is highly predictive is the bullying variable, and this graph shows that an increase in bullying is strongly correlated with feelings of depression. Interestingly, increased gang involvement is correlated with less feelings of depression. The

same is true to with truancy, but the slope of the relation is much flatter.

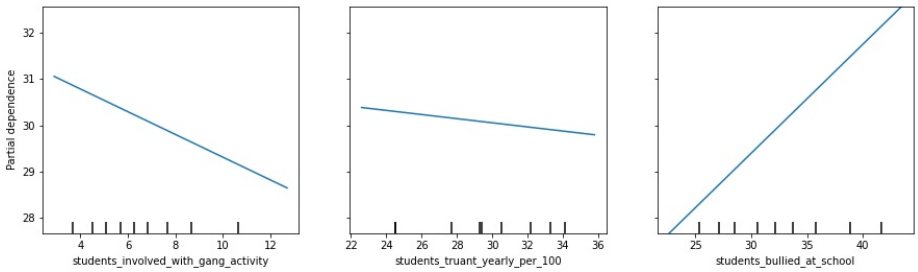

Figure 6: Students Truant from School per 100 in grade 11 in 2013

Figure 7 shows the variable importance of the grade level variables. This shows that being in 7th grade significantly decreases feelings of depression. I hypothesize this is because of age. Younger children often face fewer life stresses in general, and additionally, many of them may lack the vocabulary and understanding to fully discuss mental health.

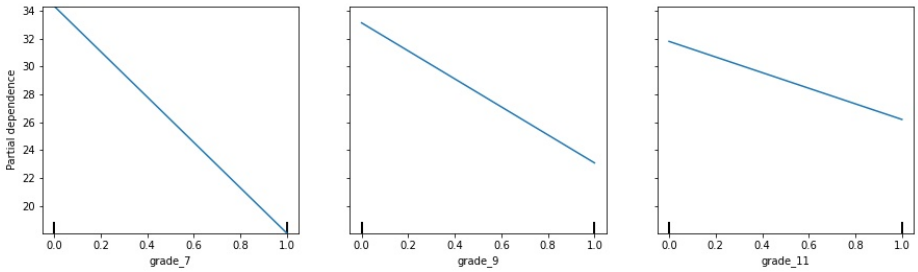

Figure 7: Students Truant from School per 100 in grade 11 in 2013

Discussion

Overall, I believe I only partially reached my own definition of success with this report. This is primarily because I encountered issues with data collection that I did not anticipate at the time of writing my project proposal, if I had known the limitation of the data I was looking for at the time, I believe I would have focused less on the data gathering part of

my definition of success. However, despite this, I am very happy with my analysis results. I was able to acquire decent predictive results. It is not surprising to me that grade level and bullying contributed most to feelings of depression in students. I also find it very interesting that, despite not being a very predictive, involvement with gang activity was correlated with less feelings of depression. I imagine that this either has to do with demographic features of areas linked to high gang involvement, and how that might correlate with views on mental health.

If I were able to go more in depth in this analysis, I would like to be able to do two related things. I would like to have data from the entire US, that way I could have a reasonable sample size without having to use time-series data, as this can create issues with statistical significance. I also would like to have data for poverty level and racial demographics for this project. Poverty level and race have been shown to correlate with well being in children and adults, and I would like to be able to include that in my report. I was unable to include this because I could not find comparable data on those subjects which I could easily access. However, counties can be seen as a stand in for some of this demographic data, and it is interesting to note that Del Norte and Trinity counties were highly predictive and these are two counties which experience much more poverty than many other counties in California.

Overall, I am pleased with this project. I think I picked a topic with difficult data, so while I would have liked to be able to gather more complete data and be able to webscrape successfully, I think I produced interesting regression results and learned a lot from this project.

Works Cited

[3]https://calschls.org/about/the-surveys/chks

[1]https://kidsdata.org/

[2]https://www.prb.org/

[4]https://www.kidsdata.org/topic/621/bullying-grade/tablefmt=874loc=2,127,347,1763,331,348,336,171,321,345,357,332,3

[5]https://www.kidsdata.org/topic/662/depressive-feelings-grade/tablefmt=943loc=2,127,347,1763,331,348,336,171,321,345

[6]https://www.kidsdata.org/topic/493/truancy/tablefmt=2392loc=2,127,347,1763,331,348,336,171,321,345,357,332,324,36

[7]https://www.kidsdata.org/topic/668/gang-grade/tablefmt=950loc=2,127,347,1763,331,348,336,171,321,345,357,332,324,