Katrina Greene
CSC 680
Final Project
7th December, 2022

Identity and Waste:
Applying Machine Learning Methods to Predict Superfund Site Presence from Socioeconomic Data

Introduction

Environmental Justice is the framework that recognizes how environmental hazards have a disproportionate impact on people of color. It refers to the institutional rules, regulations, or policies that deliberately target certain communities for undesirable land uses and lax enforcement of zoning and environmental laws [1]. As a result, certain communities are disproportionately exposed to toxic and hazardous waste. The 1987 research project *Toxic Waste and Race in the United States* is a landmark research paper that was the first to identify race as the most significant variable among variables tested in association with the location of commercial hazardous waste facilities. Using the 1980 Census data and a public dataset of hazardous waste facilities, the researchers found that this association between race and hazardous waste facilities is a national trend. They found that in communities with one commercial hazardous waste facility, the average minority percentage of the population was twice the average minority percentage of the population in communities without such facilities- 24% to 12% [2].

For this project, my goal is to see if I could see these national trends in data that I gathered myself from the 2019 American Community Survey and publicly available data on superfund site locations. Superfund sites are identified by the EPA as polluted locations in the United States requiring a long-term response to clean up hazardous material contaminations. For this project I gathered socioeconomic data on race, education levels, and poverty levels in each county in the US using the Census API. I then web scraped a Wikipedia page of a list of superfund sites, and then joined the two tables on county and state. My final table is composed of observations on each county in the US with socioeconomic data and a binary variable on whether there is a superfund site in that county. I did some data visualization to see whether the trends from 1980 are visible in this data, and I trained Random Forest, SVM, and MLP Classifier models on this data to see if the variables I collected have predictive power on the presence of a superfund site.

Data Collection

The first dataset I collected was socioeconomic data from the 2019 American Community Survey. The US Census provides an API with which people can query specified datasets. I used this API to query from the 2019 ACS, and I selected codes for the total population, total white people, total black people, and levels of education, unemployment, healthcare coverage, and poverty. The API allows you to drill down to the county level, so the resulting dataset from my API request was a dataset with the listed features and each county in the U.S. as an observation. The total number of rows of this dataset is 3220.

The second dataset I collected was web scraped using the Python library BeautifulSoup from a Wikipedia list of superfund sites. Each state had its own Wikipedia page, so I wrote a script that looped through each state's page, retrieved the html that corresponded to a table, and added the table to a final table. The result of this script was one table with each state's County name, State name, and the id of the superfund site.

Data Cleaning and Feature Engineering

I performed some cleaning and feature engineering to make the ACS table more appropriate for national comparison. The first thing I did was rename the columns to be something more descriptive than the Census table codes. I then created new columns to find the proportions of the features I got from the API. I wanted to get the proportions of the socioeconomic variables because the raw numbers that I got from the API would not be very useful in training models since each county has a different population. As such, the total amount of white people in one county would not be comparable to the total amount of white people in a smaller county. I did this by dividing the target variable with the corresponding total of the table the variable is from- for example, the race data I got were from a table in the ACS called RACE, so I divided the total_white column by the total_race column.

After calculating these new features, I joined the ACS table with the superfunds table with "County" and "State" as the primary keys. The resulting table had duplicate rows where counties had multiple superfund sites, so I dropped those duplicates so my final table only counts counties once, and the target binary variable is 1 if there is 1 or more superfund sites present in that county.

Data Visualization

For some exploratory data analysis, I first made a bar plot to compare racial makeup of counties with superfund sites vs. counties without them. I did so by filtering the data by the target variable, and coloring the bars by whether or not the percentage of non white people in that county is higher than the national average of 24.2%.

If the findings from *Race and Toxic Waste* were to be visible here, we would expect the orange column on the right to be high. This plot is not terribly informative since there are many more counties without superfund sites than there are with them, so it is hard to compare the two. As such, I produced some pie charts in order to better compare the two categories:

I produced these pie charts by manually calculating the average proportion of only white people in counties with superfund sites (0.81) and those without (0.82). The pie chart on the left is the average with superfund sites, and the one on the left is the average without. Interestingly, the averages are not very different from each other, and they are both higher than the national average of white people that comes up when I Google it, which is 75.8%. This leads me to believe that I might have made an error in my data collection and feature engineering. Given that these proportions are not very different from each other, I do not expect the features related to race to have very much predictive power in a machine learning model. However, the other socioeconomic features may be significant in training the models.

Race and Toxic Waste called out South Carolina to be the most explicit in the disproportionate exposure of minority groups to toxic waste, so I decided to drill down the pie charts I did above to only South Carolina. After getting the means again on the filtered dataset with just South Caroline, I got these pie charts:

Here it is a little more clear that the average proportion of white people in counties with superfund sites is lower than those without. Since this trend is more obvious in South Carolina I though about training the models on just that subset of data, but there were not very many observations so I stuck with the full national dataset.

Methods

The first model that I trained is a Random Forest Classifier. A Random Forest is an ensemble of Decision Trees, which classify a new instance by starting at a root node and following nodes composed of feature tests until a leaf node is reached, specifying the prediction class. The test data is split between the nodes based on the highest information gain, which indicates the highest purity of instances of a class in a split. The purer the split, the more beneficial for classification. For my random forest classifier, I first initialized a simple model from the sklearn library and conducted 5 fold cross validation with my training dataset. The cross validation runs the model 5 times with different subsets of the training data held out for verification each times, with the average of the results returned. My initial random forest model produced an ROC AUC score of 0.72. ROC curves plot the true positive rate against the false positive rate, so we want the area under that curve to be as close to one as possible. An ROC AUC score of 0.5 is equivalent to a random guess. As such, a score of 0.72 is better than no model at all, but it is not ideal. To improve it, I tried a grid search with different numbers of estimators and max features to tune the hyperparameters. After running the training data through this new model with the best selected hyperparameters, the ROC AUC score was a perfect 1 and the confusion matrix showed no misclassifications. While this is a great score, it is suspiciously too good and may be the result of overfitting.

The second model I used is a Support Vector Machine for linear classification. This classifier defines a linear separation between two classes with the widest possible margin between the classes. I first tried training a simple linear support vector classifier with the sklearn library, and the ROC AUC score from this model with 5 fold cross validation was 0.54. This is not a great score, so like with the random forest classifier I ran a grid search for the best hyperparameters. After running the model again with the tuned hyperparameters, the AUC ROC score was even lower at 0.5. The confusion matrix shows that both models only predicted positives.

The last model that I trained and tested was a Multilayer Perceptron neural network for classification. I first wanted to try a straightforward model, so I just instantiated an MLP model with one ReLU hidden layer and a sigmoid output layer. Since this is a binary classification application, I had only a single output neuron using the logistic activation function to interpret the estimated probability of the positive class. The estimated probability of the negative class is calculated by subtracting 1 from that number. When I compiled the model, I used a binary cross entropy loss function as well. When I fit the model with the training data and 10 epochs, the output from each epoch shows that the categorical accuracy is 1. This is the highest possible

accuracy and means the model is classifying superfund sites perfectly, so I am interested in seeing how this model performs with the test data and whether it is overfitted to the training data.

Results

After training the models with the training dataset, I then ran them again with the held-out test set. For the random forest model, I was correct in that the tuned model was overfitted to the training set because the test ROC AUC was very low at 0.54. The SVM classifier also did not improve with a ROC AUC of 0.5. The MLP model is interesting because, like with the training set, it also got an accuracy score of 1. This is interesting because, even though it performed better than the previous two methods, I would not expect perfect accuracy from a dataset with the problems that I pointed out earlier.

Conclusions

For this project, I tried to gather my own socioeconomic and toxic waste location data and identify trends between them. With the dataset resulting from the Census API and my superfund site web scraping, I was unable to directly see any significant national association between race and superfund site locations. This lack of association was also visible when other socioeconomic data such as poverty, education, and unemployment levels were included. The random forest and support vector machine classifiers did not have strong predictive power on the test set. The MLP classifier had a perfect accuracy score and as such performed the best out of the three models that I trained and tested.

References

[1] *Environmental Justice & Environmental racism*. Greenaction for Health and Environmental Justice. (n.d.). Retrieved from http://greenaction.org/what-is-environmental-justice/

[2] *Toxic Wastes and Race in the United States*, Commission for Racial Justice, United Church of Christ, 1987 https://www.nrc.gov/docs/ML1310/ML13109A339.pdf

[3] https://en.wikipedia.org/wiki/List_of_Superfund_sites

[4] https://api.census.gov/data/2019/acs/acs5/variables.html

[5] https://medium.com/analytics-vidhya/web-scraping-a-wikipedia-table-into-a-dataframe-c52617e1f451

[6] https://www.census.gov/content/dam/Census/data/developers/api-user-guide/api-guide.pdf