Predicting Varicella Outbreaks With CDC Vaccination Coverage Monitoring And Census Data

Laura McCallion
Brown University
https://github.com/lmccalli/1030_Project.git

Introduction

This project aims to predict the number of reported *Varicella* cases using vaccination coverage data from the CDC National Immunization Survey as well as data used from the 2000 and 2010 Censuses. Although the rate of *Varicella* morbidity has steadily decreased since the introduction of the vaccine in 1995, monitoring of the disease remains relevant for a number of reasons. First, if there are trends linking outbreaks to more general factors within the population, as might be found within the census data, then more resources could be allocated to preventing disease within that particular group. Second, given the recent trend of “anti-vaccination,” not only is it important to have rigorous evidence of the efficacy of the vaccine, but it is also critical to monitor the effects of a potential decrease in the proportion of vaccinated individuals on future outbreaks.

EDA

Data Gathering

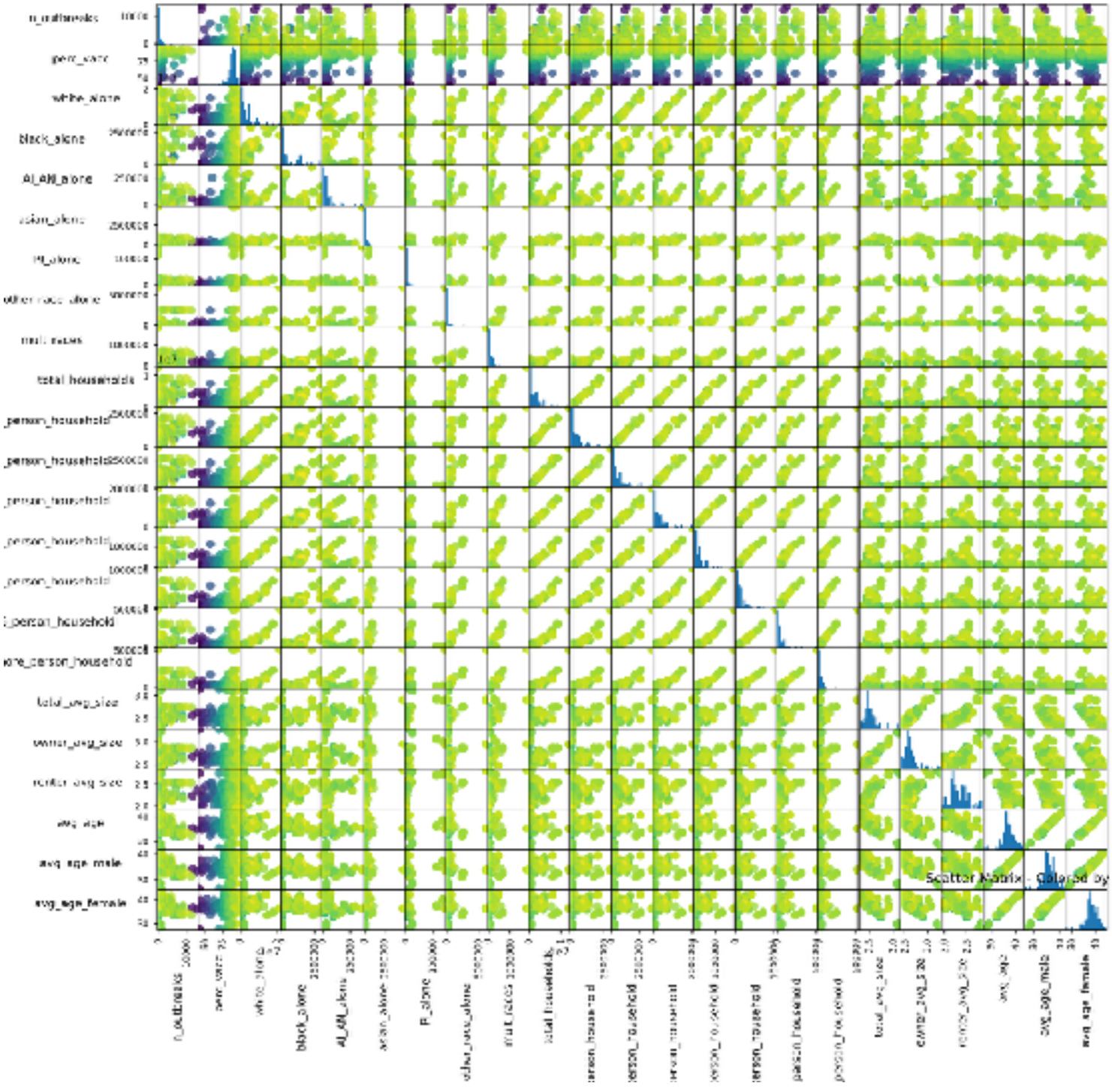
Annual data from the National Notifiable Diseases Surveillance System (NNDSS) has been available from 1993. However, multiple factors complicate the use of this data. First, the vaccine was only available inn 1995, so years before then were not used in the prediction. Second, various states may or may not report *Varicella* cases depending on both the state and the year — data were not available at all from 1994-1998, and in 2002. When

available, the data were separated into individual states and years.
Census data was chosen from just the standard census reports, and narrowed further based on which data sets were available in both 2000 and 2010. Data sets with different columns from 2000 to 2010 were used when possible, but in 2 cases data had to be excluded due to inadequate expertise in the subject matter that would be necessary to combine the columns in the data. Census data was assigned to outbreak and vaccination data based on the year that most closely matched the year of the census. In a previous version of the project, more columns were available from census data.

Missing Data
The only missing data present in any of the relevant datasets was in the target variable — the number of reported cases. Therefore, it was not possible to attempt to replace any of the missing values. Rows with missing data were dropped.

After initial cleaning and processing, the data contained 575 observations with 22 columns.

Figures


white_alone
black_alone
AI_AN_alone
avg_age_male
avg_age_female
Scatter Matrix - Colored by Percent Vac

As all the data used was numeric, most of the relationships can be seen clearly in a scatter plot. Some variables have relatively linear relationships, and these tend to be predictable. For example, the average age for men, women, and all sexes look fairly linear, which is intuitive. A similar relationship exists between the number of people in each household.

Another interesting facet of this plot is the relationships with percent vaccination. The scatter plots within the matrix are colored based on the percent vaccination value of the points, and while most plots show relatively little relation to the value, the number of outbreaks is significantly more correlated. This is indicated further in plots within the results section.

Methods

All data in the set was numeric, and the standard scaler was used to preprocess the data.
Both Lasso and Ridge regression were tested in order to find an effective method for regression, using mean squared error as a scorer. This was used as the data was numeric, and had a wide spread. Scoring using MSE allowed for the model to be fit in a way that best fit this type of data.

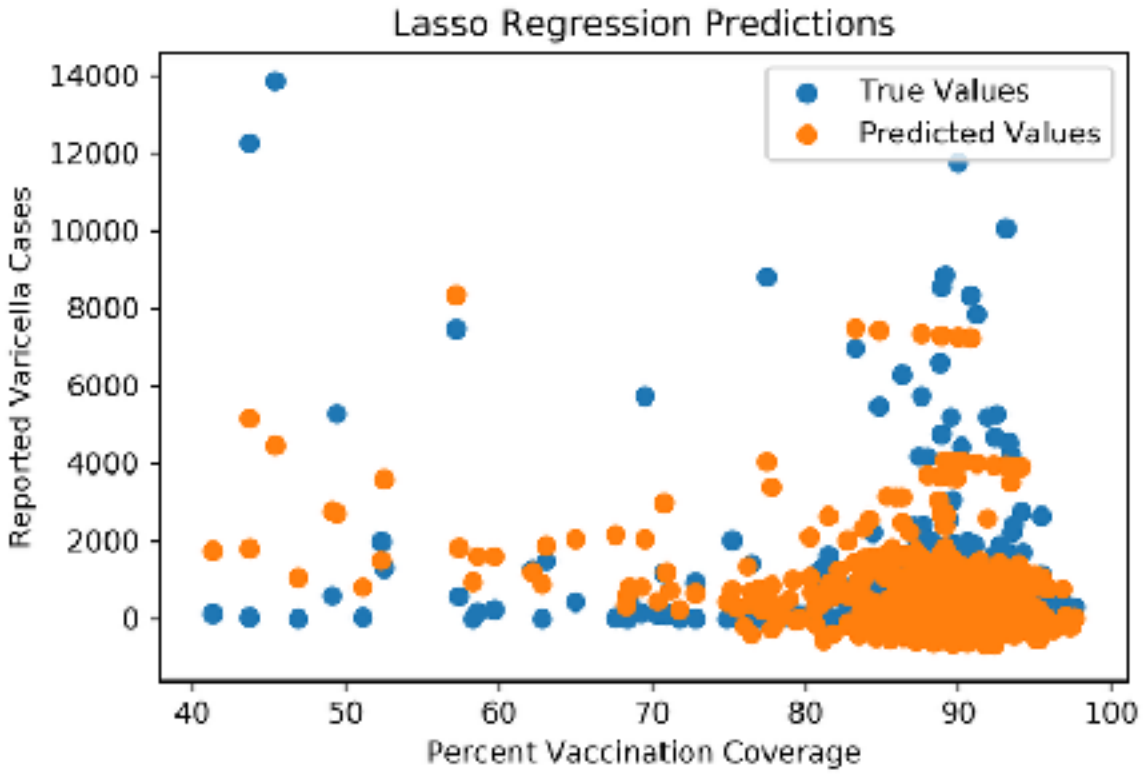
Notably, Lasso regression failed to resolve with a reasonable number of iterations. As the number of possible iterations was increased, the score of the lasso regression grew closer to the score of the ridge regression, but never surpassed it. Therefore, ridge regression was used, as it was a better functional match for the data, as well as being significantly faster.

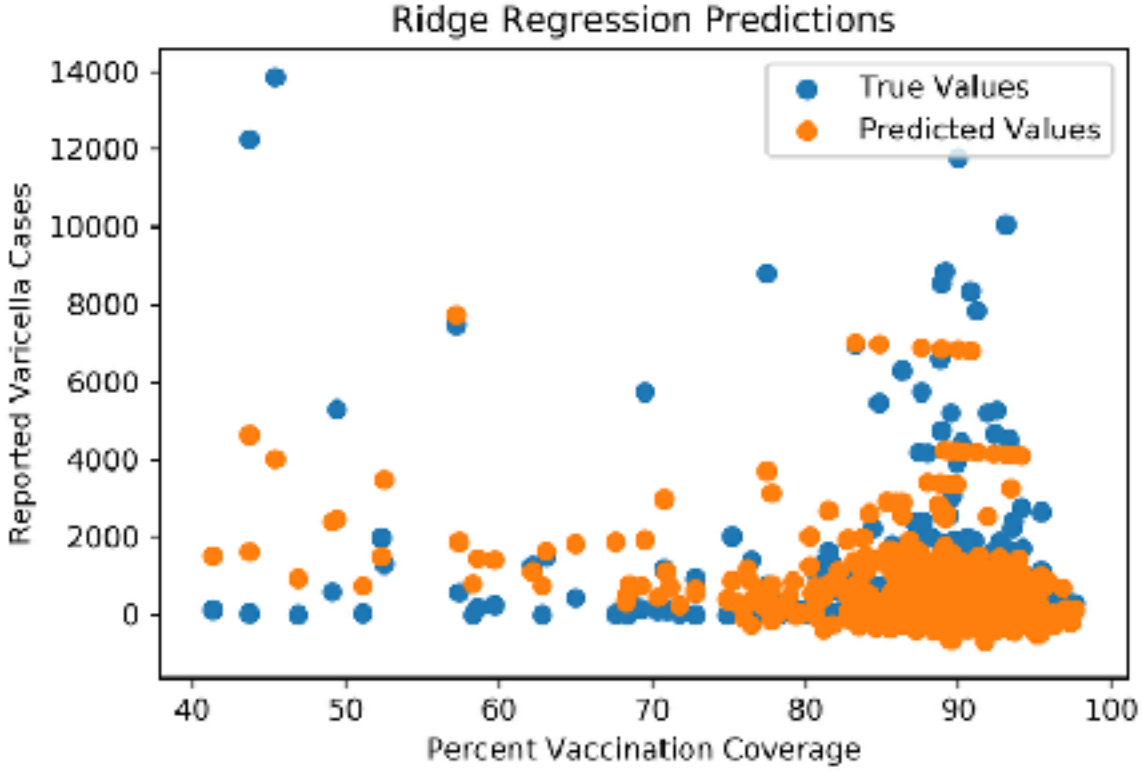
In addition, a classification method using SVC was also tested. This method only assigned values of 0 or 1 to the target variable — 0 if there were no reported cases, and 1 if there were. With this

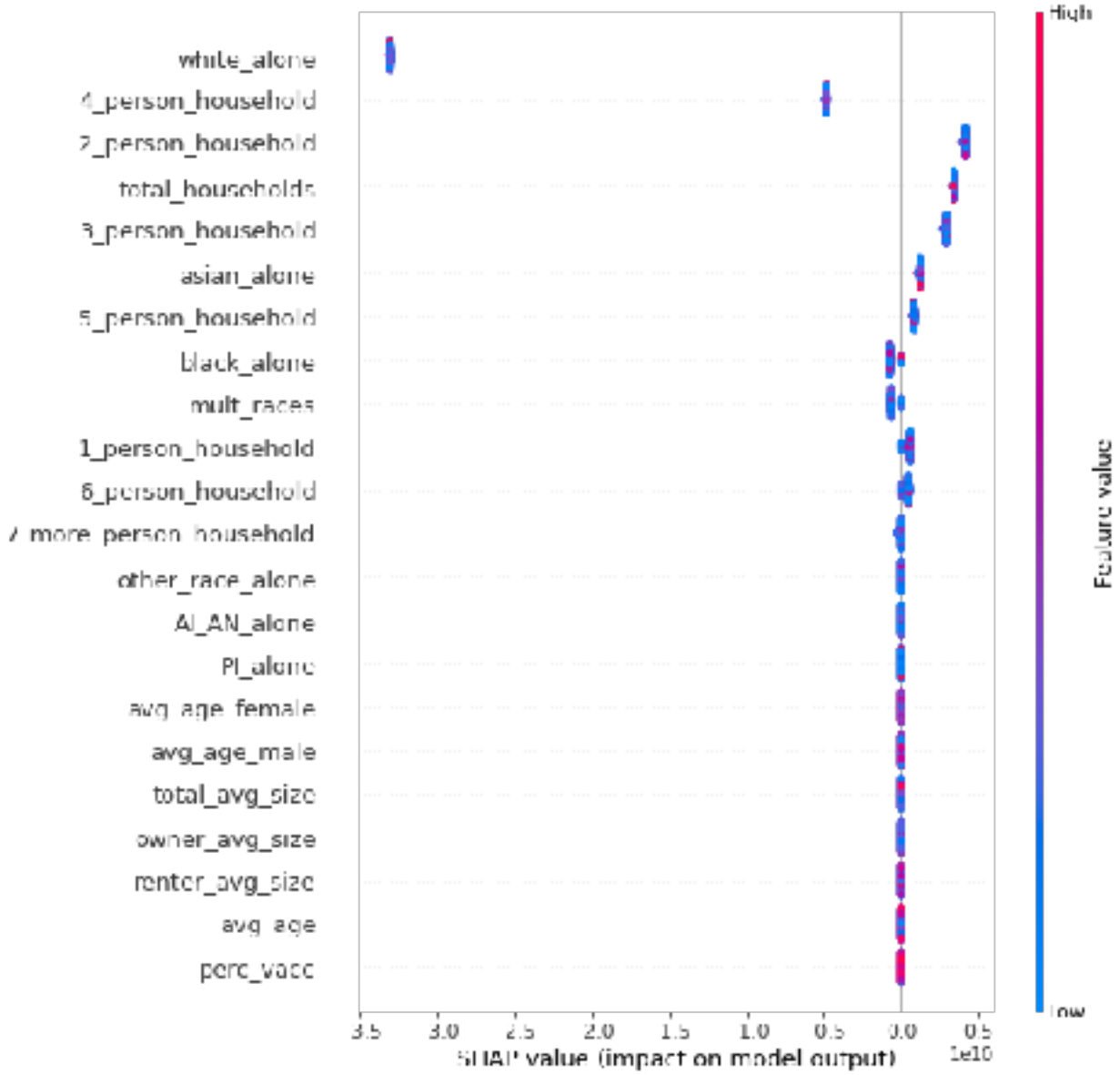
classification method, roughly 89% of the points were in class 0, making the dataset notably unbalanced.

All of these methods were used with data split 20% to testing data, as well as 5-fold cross-validation and a grid search to find the best parameters for the models.

Results


Lasso Regression Predictions
True Values
Predicted Values
Reported Varicella Cases
Percent Vaccination Coverage
0
2000
4000
6000
8000
10000
12000
14000
40
50
60
70
80
90
100


Ridge Regression Predictions
True Values
Predicted Values
Reported Varicella Cases
Percent Vaccination Coverage
0
2000
4000
6000
8000
10000
12000
14000
40
50
60
70
80
90
100



First, the results of the SVC model that was considered should be discussed. The accuracy of the model was roughly 90%. While this may seem impressive, the target variable is 89% class 0, meaning that this only represents a 1% improvement over the baseline. Combined with the fact that this classification prediction offers significantly less information on potential outbreaks than regression, the results allow us to discard this model entirely.

By contrast, the ridge regression model shows promise. The R2 value is roughly 0.53, which, while not remarkably close to the maximum of 1, is a relatively significant improvement over the base R2 value of 0. This is further supported by the scatter plots used to examine the data. While the data itself is too high

dimensional to easily view, the comparison of vaccination rate to predicted case count shows that the predicted values match the pattern of the true values closely. This also indicates that the vaccination coverage may be a significant factor in the prediction. Lasso regression, though it was not chosen for the final regression, shows a nearly identical pattern. The plots also show a relationship between vaccination coverage and reported cases — as the percent coverage increases, outbreaks tend to be smaller.

The SHAP plot, curiously, indicates a relatively low SHAP value for the percentage vaccination. It also indicates that the amount of individuals in a household is important to the model. Perhaps this is an indication that feature reduction using similar variables in the census data could yield a more informative indication of which variables effect the prediction, or even a more accurate prediction. While this might indicate that household size is important to predicting outbreaks, household sizes could also correlate with other cultural or social trends that influence the spread of illness.

Outlook

In the future, a larger scale project might be able to use more granular data. For this project, states were used for data points. However, both Census and CDC data also covers smaller regions — as an example, the state of New York was largely excluded, as within the CDC data, Upstate New York almost never reported *Varicella* outbreaks. However, New York City reported regularly. A more thorough search of all potential data points could yield a better predictor — not only in that the volume of data might increase, but in that the census data of smaller areas might have more predictive power for predicting outbreaks than larger states.

Further investigation into the relative importance of variables might also make the model more clear. It is possible that the

exclusion of the years with no vaccines undervalued the importance of vaccination in the model. It would also be useful to check the specific methods for gathering the data — in short, the vaccination rate is likely a summary metric, rather than the number of outbreaks, which is a number that follows a specific criteria when it is reported. The methods by which the data are collected could influence the best model to use for prediction.

It could also prove useful to compare historical events to outbreak data. For example, the number of reported varicella cases rose in Texas from 5464 in 2003 to 11768 in 2006 — this is unusual, given that disease rates generally declined after the introduction of the vaccine. This level of morbidity is also the highest of any reporting state in 2006. Finding out the cause of patterns such as this may indicate a new variable that could better allow the model to handle outlying data.

References

Census Data: American Fact Finder
https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml
2010 SF1 100% Data
2000 SF1 100% Data
2010 SF2 100% Data
2000 SF2 100% Data

MMWR: Summary of Notifiable Infectious Diseases
https://www.cdc.gov/mmwr/mmwr_nd/index.html

1996 through 2017 Childhood Varicella Vaccination Coverage Trend Report
https://www.cdc.gov/vaccines/imz-managers/coverage/childvaxview/data-reports/varicella/trend/index.html

Nationally Notifiable Infectious Diseases and Conditions, United States: Annual Tables
https://wonder.cdc.gov/nndss/nndss_annual_tables_menu.asp