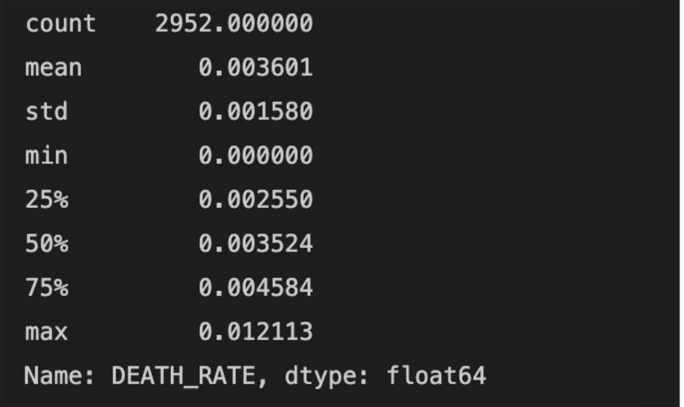
Classifying High COVID Fatality Rates

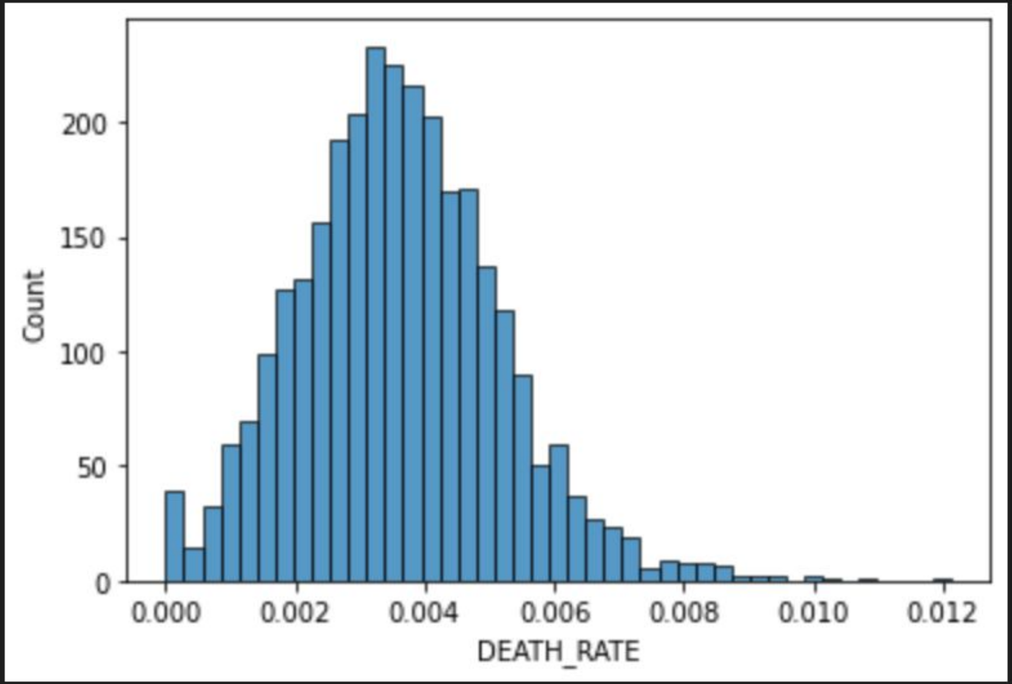
Casey Durfee
<casey.durfee@colorado.edu>

June 20, 2022

There is a huge variance in the USA county-level COVID death rate.



```
count    2952.000000
mean        0.003601
std         0.001580
min         0.000000
25%         0.002550
50%         0.003524
75%         0.004584
max         0.012113
Name: DEATH_RATE, dtype: float64
```



This variance is a tragedy.

1 std = .001580 * 330 Million people = 521,000 Lives = 10 football stadiums.

What might be behind the variance?

- Income inequality in a community is a strong predictor of COVID fatality rates [Tan 2021]
- Broadband access appears to be a strong predictor [Lin 2022]
- Political orientation of a county wasn’t a strong predictor of COVID deaths at first [Gao 2021]. That has changed since vaccines [NPR 2021].

My Approach: As Broad As Possible

- Pull in as much data as possible about US counties (100 metrics total)
  - Examples: infant mortality rate, ethnic makeup, homicide rate, median household income, access to healthcare
  - Drawn from countyhealthrankings.org, vaccinetracking.us, US Census, the CDC’s Social Vulnerability Index, MEDSL voting data
  - Based on 2018/9 (pre-pandemic) data
- Determining which factors most strongly predict high COVID fatality rates by fitting machine learning models to the data.

Data Details

- All metrics with > 5% missing values were tossed.
  - Median value was used for imputation
  - Many metrics were only missing a handful of values
- 2919 US Counties x 70 metrics
- A handful of metrics are very similar (ex. different estimates of household income and life expectancy - not obvious which is best source)
- Also tried just larger counties (>50K population). 932 counties x 81 metrics.

Model Details

- Predict when a county will have a higher than average COVID fatality rate
- Two models: for the first year of COVID (before vaccines) and the second year (after vaccines). This is because the demographics of COVID deaths changed after vaccines [NPR 2021]
- Multiple supervised learning models: SVM, AdaBoost, RandomForest, Logistic Regression

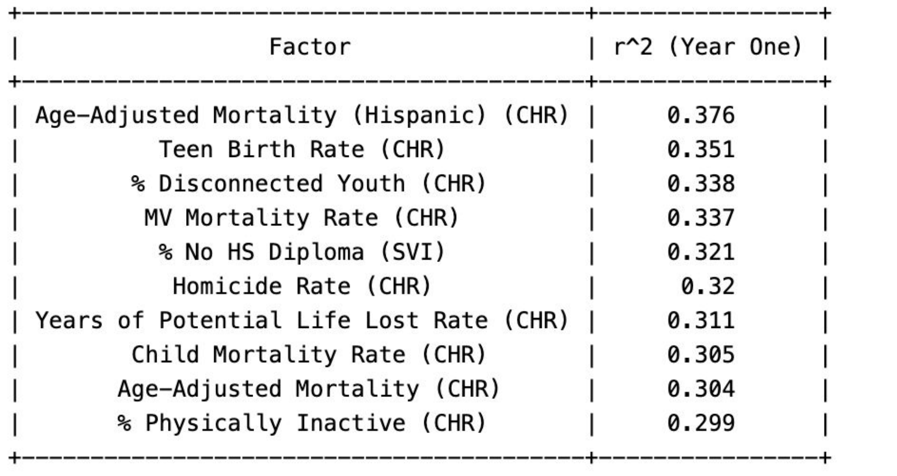
Evaluation - what does success look like?

- Statistically: How accurately can we predict high fatality rates from pre-pandemic county data?
- What statistically significant factors were correlated with changes in fatality rates year over year?
- Conceptually: Can we understand at least some of the variance?

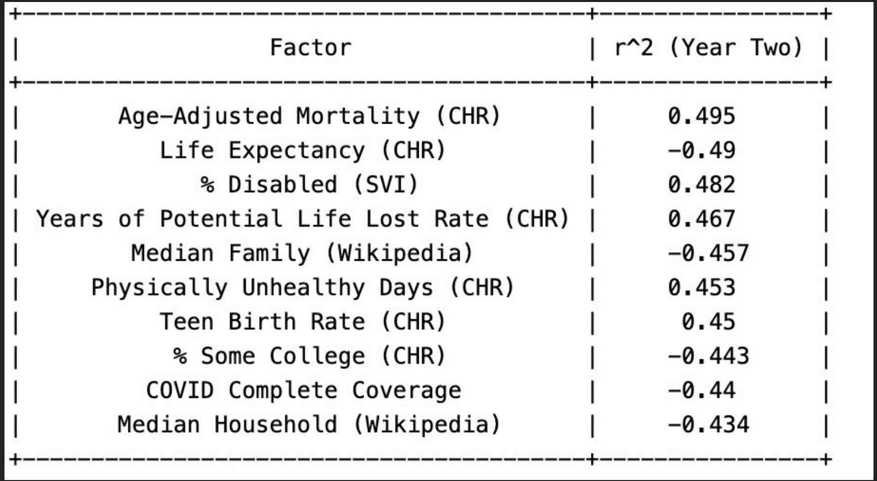
Evaluation

- Use accuracy score and 5-fold cross validation to decide best model for classifying higher than average fatality rate counties
- Determine most important factors with permutation importance
- Find top 10 factors that predicted high COVID fatality rates for year 1 and year 2 (separately)
- Re-run model on only larger counties

Correlations - Year One vs Year Two

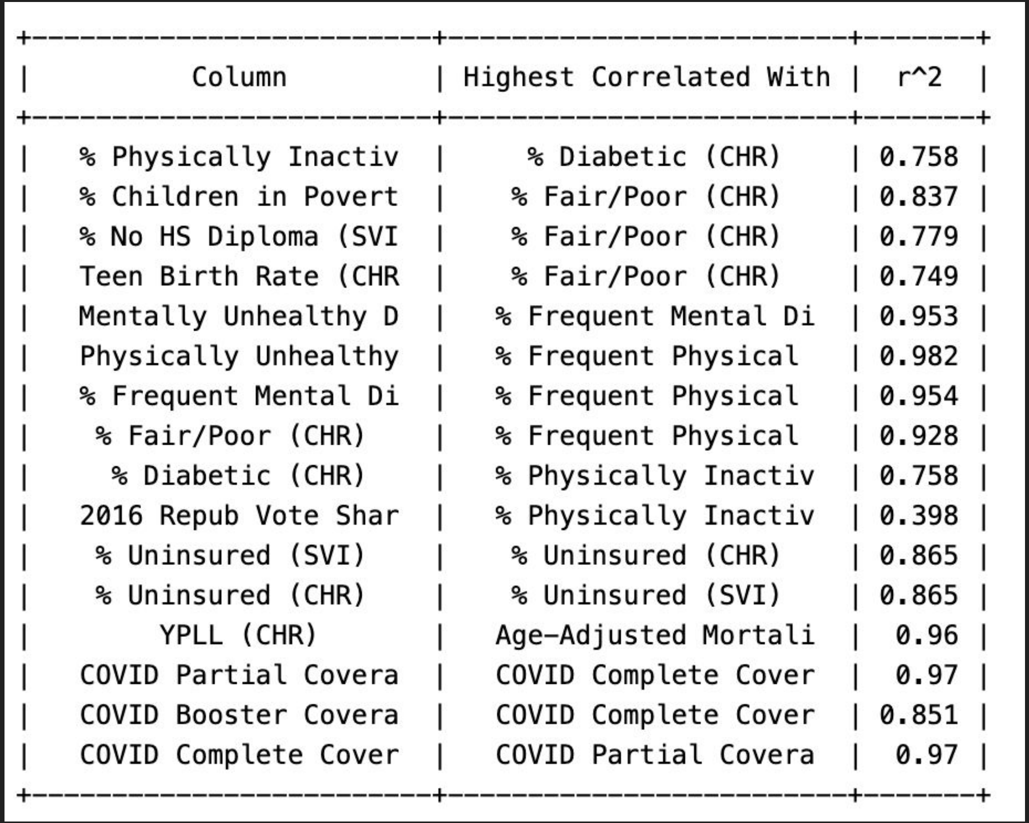


```
+-----------------------------------------+----------------+
|                  Factor                 | r^2 (Year One) |
+-----------------------------------------+----------------+
| Age-Adjusted Mortality (Hispanic) (CHR) |     0.376      |
|          Teen Birth Rate (CHR)          |     0.351      |
|        % Disconnected Youth (CHR)       |     0.338      |
|         MV Mortality Rate (CHR)         |     0.337      |
|          % No HS Diploma (SVI)          |     0.321      |
|           Homicide Rate (CHR)           |      0.32      |
| Years of Potential Life Lost Rate (CHR) |     0.311      |
|        Child Mortality Rate (CHR)       |     0.305      |
|       Age-Adjusted Mortality (CHR)      |     0.304      |
|       % Physically Inactive (CHR)       |     0.299      |
+-----------------------------------------+----------------+
```



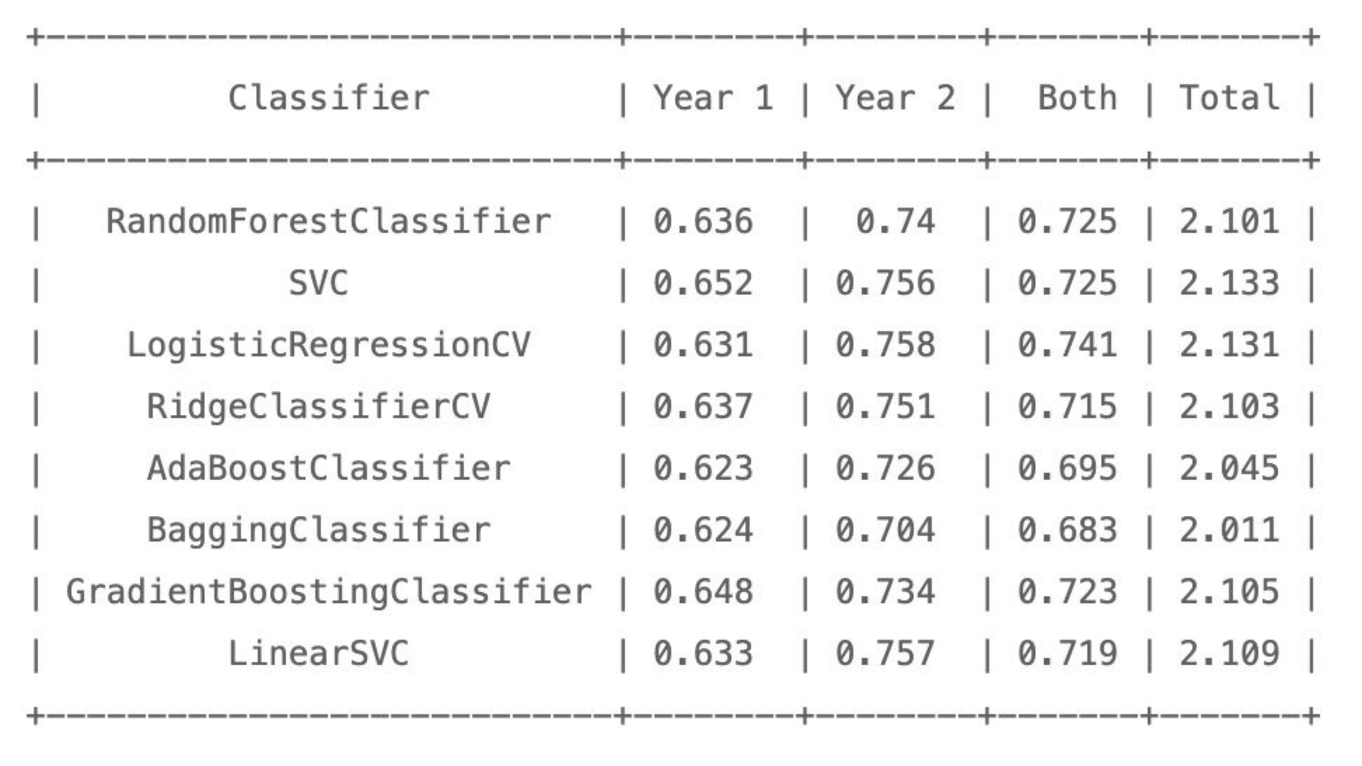
```
+-----------------------------------------+----------------+
|                  Factor                 | r^2 (Year Two) |
+-----------------------------------------+----------------+
|       Age-Adjusted Mortality (CHR)      |     0.495      |
|          Life Expectancy (CHR)          |     -0.49      |
|             % Disabled (SVI)            |     0.482      |
| Years of Potential Life Lost Rate (CHR) |     0.467      |
|        Median Family (Wikipedia)        |     -0.457     |
|     Physically Unhealthy Days (CHR)     |     0.453      |
|          Teen Birth Rate (CHR)          |      0.45      |
|           % Some College (CHR)          |     -0.443     |
|         COVID Complete Coverage         |     -0.44      |
|       Median Household (Wikipedia)      |     -0.434     |
+-----------------------------------------+----------------+
```

Strong correlations between some factors:



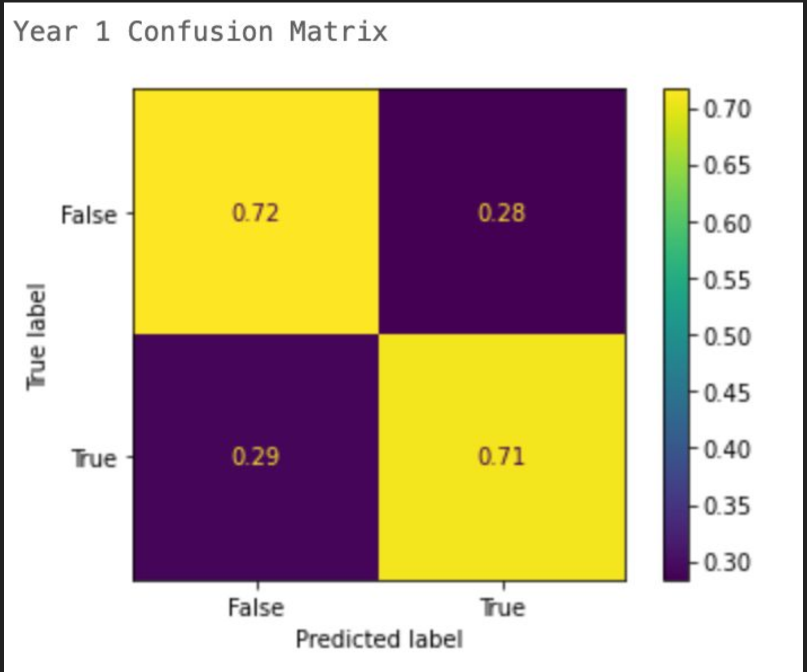
```
+-------------------------+-------------------------+-------+
|          Column         | Highest Correlated With |  r^2  |
+-------------------------+-------------------------+-------+
|   % Physically Inactiv  |     % Diabetic (CHR)    | 0.758 |
|   % Children in Povert  |    % Fair/Poor (CHR)    | 0.837 |
|   % No HS Diploma (SVI  |    % Fair/Poor (CHR)    | 0.779 |
|   Teen Birth Rate (CHR  |    % Fair/Poor (CHR)    | 0.749 |
|   Mentally Unhealthy D  |   % Frequent Mental Di  | 0.953 |
|   Physically Unhealthy  |   % Frequent Physical   | 0.982 |
|   % Frequent Mental Di  |   % Frequent Physical   | 0.954 |
|    % Fair/Poor (CHR)    |   % Frequent Physical   | 0.928 |
|     % Diabetic (CHR)    |   % Physically Inactiv  | 0.758 |
|   2016 Repub Vote Shar  |   % Physically Inactiv  | 0.398 |
|    % Uninsured (SVI)    |    % Uninsured (CHR)    | 0.865 |
|    % Uninsured (CHR)    |    % Uninsured (SVI)    | 0.865 |
|        YPLL (CHR)       |   Age-Adjusted Mortali  |  0.96 |
|   COVID Partial Covera  |   COVID Complete Cover  |  0.97 |
|   COVID Booster Covera  |   COVID Complete Cover  | 0.851 |
|   COVID Complete Cover  |   COVID Partial Covera  |  0.97 |
+-------------------------+-------------------------+-------+
```

Training on Y2 death rates were more accurate than Y1

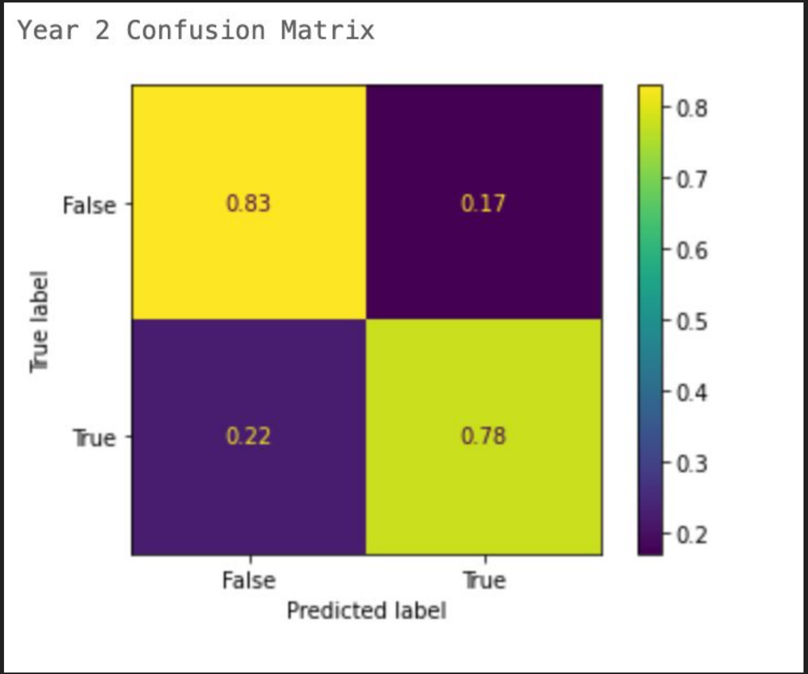


```
+----------------------------+--------+--------+-------+-------+
|         Classifier         | Year 1 | Year 2 |  Both | Total |
+----------------------------+--------+--------+-------+-------+
|   RandomForestClassifier   | 0.636  |  0.74  | 0.725 | 2.101 |
|            SVC             | 0.652  | 0.756  | 0.725 | 2.133 |
|    LogisticRegressionCV    | 0.631  | 0.758  | 0.741 | 2.131 |
|     RidgeClassifierCV      | 0.637  | 0.751  | 0.715 | 2.103 |
|     AdaBoostClassifier     | 0.623  | 0.726  | 0.695 | 2.045 |
|     BaggingClassifier      | 0.624  | 0.704  | 0.683 | 2.011 |
| GradientBoostingClassifier | 0.648  | 0.734  | 0.723 | 2.105 |
|         LinearSVC          | 0.633  | 0.757  | 0.719 | 2.109 |
+----------------------------+--------+--------+-------+-------+
```

Best classifier - LinearRegressionCV

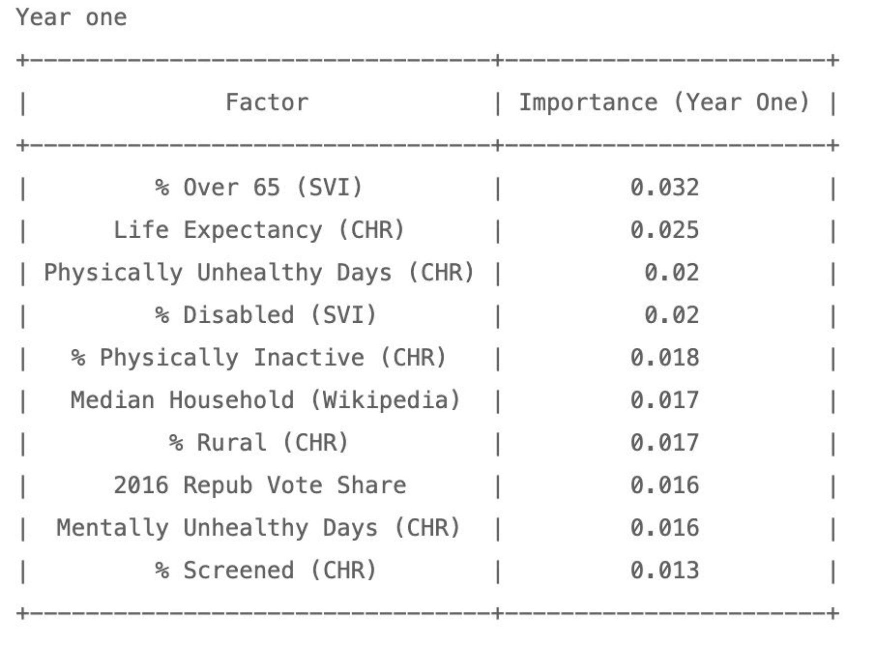


Year 1 Confusion Matrix

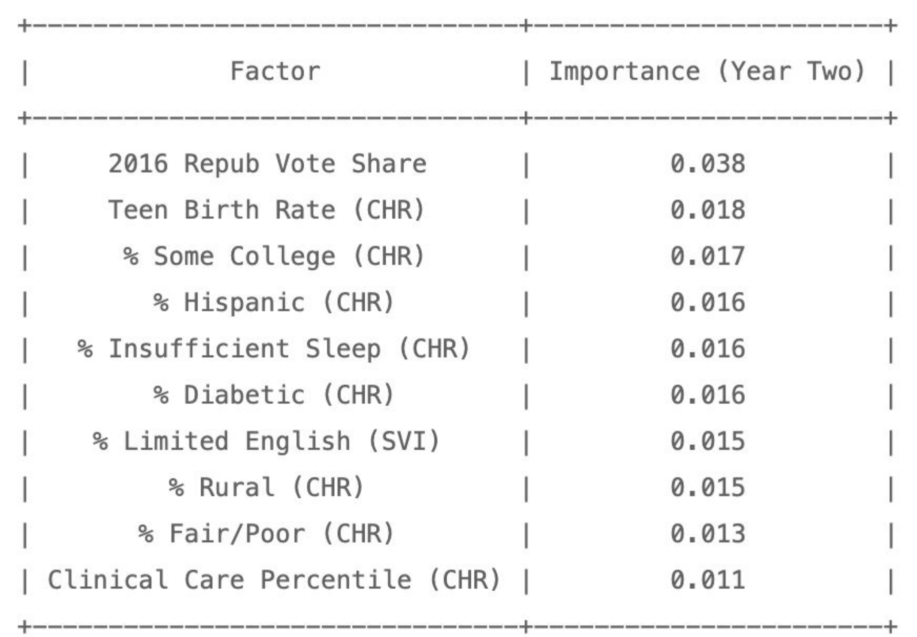


Year 2 Confusion Matrix

Permutation - what matters to the classifier?

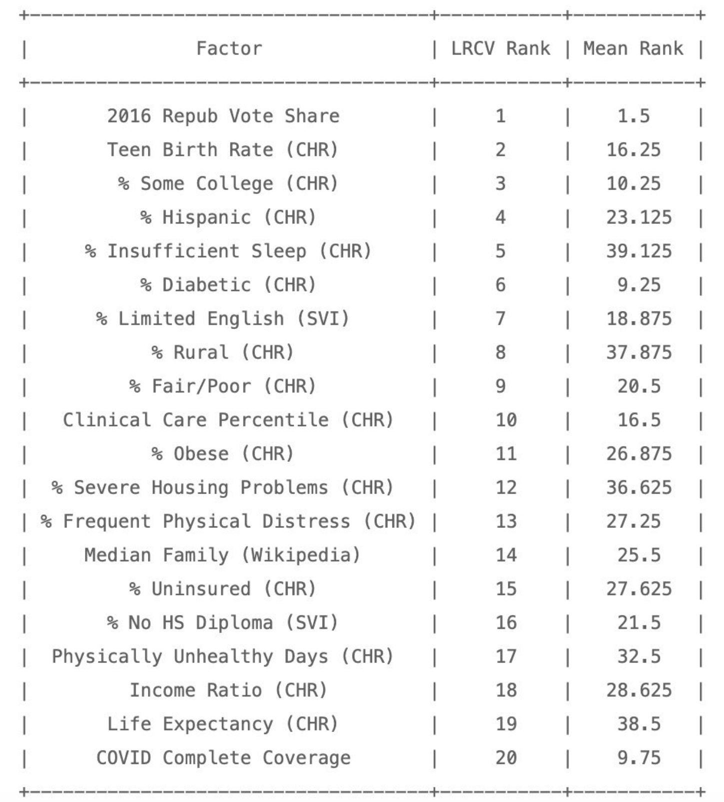


```
Year one
+---------------------------------+-----------------------+
|             Factor              | Importance (Year One) |
+---------------------------------+-----------------------+
|         % Over 65 (SVI)         |         0.032         |
|      Life Expectancy (CHR)      |         0.025         |
| Physically Unhealthy Days (CHR) |          0.02         |
|        % Disabled (SVI)         |          0.02         |
|   % Physically Inactive (CHR)   |         0.018         |
|  Median Household (Wikipedia)   |         0.017         |
|          % Rural (CHR)          |         0.017         |
|      2016 Repub Vote Share      |         0.016         |
|  Mentally Unhealthy Days (CHR)  |         0.016         |
|        % Screened (CHR)         |         0.013         |
+---------------------------------+-----------------------+
```



```
+--------------------------------+-----------------------+
|             Factor             | Importance (Year Two) |
+--------------------------------+-----------------------+
|     2016 Repub Vote Share      |         0.038         |
|     Teen Birth Rate (CHR)      |         0.018         |
|      % Some College (CHR)      |         0.017         |
|        % Hispanic (CHR)        |         0.016         |
|   % Insufficient Sleep (CHR)   |         0.016         |
|        % Diabetic (CHR)        |         0.016         |
|    % Limited English (SVI)     |         0.015         |
|         % Rural (CHR)          |         0.015         |
|       % Fair/Poor (CHR)        |         0.013         |
| Clinical Care Percentile (CHR) |         0.011         |
+--------------------------------+-----------------------+
```

What do other classifiers say?



```
+------------------------------------+-----------+-----------+
|               Factor               | LRCV Rank | Mean Rank |
+------------------------------------+-----------+-----------+
|       2016 Repub Vote Share        |     1     |    1.5    |
|       Teen Birth Rate (CHR)        |     2     |   16.25   |
|        % Some College (CHR)        |     3     |   10.25   |
|          % Hispanic (CHR)          |     4     |   23.125  |
|     % Insufficient Sleep (CHR)     |     5     |   39.125  |
|          % Diabetic (CHR)          |     6     |    9.25   |
|      % Limited English (SVI)       |     7     |   18.875  |
|           % Rural (CHR)            |     8     |   37.875  |
|         % Fair/Poor (CHR)          |     9     |    20.5   |
|   Clinical Care Percentile (CHR)   |     10    |    16.5   |
|           % Obese (CHR)            |     11    |   26.875  |
|  % Severe Housing Problems (CHR)   |     12    |   36.625  |
| % Frequent Physical Distress (CHR) |     13    |   27.25   |
|     Median Family (Wikipedia)      |     14    |    25.5   |
|         % Uninsured (CHR)          |     15    |   27.625  |
|       % No HS Diploma (SVI)        |     16    |    21.5   |
|  Physically Unhealthy Days (CHR)   |     17    |    32.5   |
|         Income Ratio (CHR)         |     18    |   28.625  |
|       Life Expectancy (CHR)        |     19    |    38.5   |
|      COVID Complete Coverage       |     20    |    9.75   |
+------------------------------------+-----------+-----------+
```

- Ran LinearSVC, GradientBoost, Ridge and RandomForest on data
- “2016 Repub Vote Share” was highly ranked by all models
- Rankings may be muddied by very similar factors (eg Median Family vs. Per Capita Income)

references

[Tan 2021] https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2779417

[Gao 2021] https://pubmed.ncbi.nlm.nih.gov/34226856/

[Lin 2022] https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2789619

[NPR 2021] https://www.wbur.org/npr/1059828993/data-vaccine-misinformation-trump-counties-covid-death-rate