kcoop610 / **capstone-pandemic-us-health-inequities**

Flatiron School data science bootcamp capstone project exploring inequities in the COVID-19 pandemic impact across US counties

1 star 0 forks

Unstar Unwatch

Code Issues Pull requests Actions Projects Wiki Security

main

kcoop610 minor updates to README and presentation, deleted outd... 1 minute ago 16

View code

README.md

Social Determinants of Health & the COVID-19 Pandemic

Supervised Learning Classification Model Predicting the Spread & Mortality of COVID-19, Extensible to Future Pandemics/Epidemics

Kristin Cooper | July, 2021 | DTSC-FT-022221 | Instructor: James Irving

Social Determinants of Health
Healthy People 2030 // U.S. Department of Health and Human Services, Office of Disease Prevention and Health Promotion _____

Purpose:

The purpose of this report is to study social, economic, and health inequities in the US and predict how those inequities influenced the spread and mortality of COVID-19 in a community/region. Predictions should be used to inform public policy and preparation for future pandemics/epidemics.

My original intent was to perform this analysis at the county level. However, finding reliable county-level COVID-19 case and death reports proved difficult. 23 states had a total of ~83k cases and ~1700 deaths that were unattributable to a county, based on differences in reporting structure and unclear reporting protocols between where the case was treated vs where the patient lived.

Results described below are currently at the state level, with an intent to revisit the county-level data in the future.

Data:

A variety of social determinants of health and demographic data from the CDC and University of Wisconsin Public Health Institute was combined with COVID-19 case, death, and vaccine measures.

Features:

COVID-19 Stats	Economic Measures	Health Measures	Social Measures	Dem
• Case count • Death count • Vaccine hesitancy • CVAC vaccine rollout concern	• Per capita income • Median household income • Income inequality • Poverty rate • Unemployment rate	• Life expectancy and premature deaths • Smoking and excessive drinking • Obesity • Poor health days (physical & mental) • Physical inactivity • Preventable hospital stays	• Housing • Internet access • Vehicle access • Food environment and food insecurity • Education • Air and drinking water pollution	•

		• Ratio of population to primary care physicians • Flu vaccinations • Uninsured population		

Sources:

- CDC's Social Vulnerability Index
- CDC's Vaccine Hesitancy
- The University of Wisconsin Population Health Institute's County Health Rankings
- New York Times COVID Case and Death Counts

Approach:

This is a multinomial classification problem with an engineered 3-class target.

Feature Engineering:

From the source data, the following initial features were engineered:

- Population density = population / area (sqmi)
- Cases per 100k = (cases / population) * 100,000
- Deaths per 100k = (deaths / population) * 100,000
- Percent of cases resulting in death, or 'deathrate' = (deaths / cases) * 100
- Impact points = ((1 * cases per 100k) + (3 * deaths per 100k)) * deathrate
- Impact category:
  - High = impact points > mean + (.8*(standard deviation))
  - Low = impact points < mean - (.8*(standard deviation)
  - Average = impact points within mean +/- (.8*(standard deviation))

Modeling Techniques:

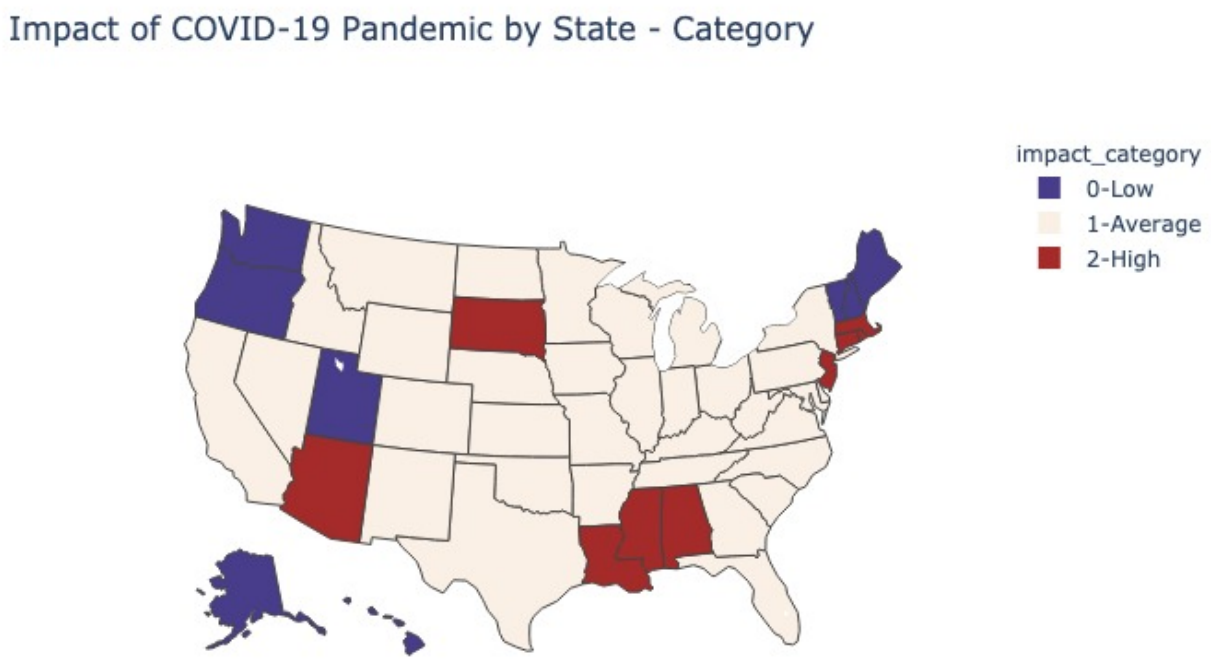
1. **Multiclass regressors**
    - Tree-based models (decision tree, random forest)
    - K-Nearest Neighbors
    - Logistic Regression with `multi_class='multinomial'`
2. **One vs Rest (OVR) Meta-Classifiers**
    - Gradient Boost Classifier
    - linear Support Vector Machine
    - Logistic Regression with `multi_class='ovr'`
3. **One vs One (OVO) Meta-Classifier**
    - Support Vector Machine

Evaluation Metrics:

Metrics for multiclass models could be calculated as either a macro average or a micro (weighted) average. Because training data is balanced using sklean.oversampling.SMOTE, I know my classes are balanced and can safely use the macro-average.

- **Overfitting** - the difference between the accuracy score on the training data and the accuracy score on the testing data
- **Accuracy** & **AUC** - measures of how many predictions the model gets right
- **F1 Score** - the harmonic mean of precision (how many predicted positives are true positives) and recall (how many actual positives were correctly predicted)

Results

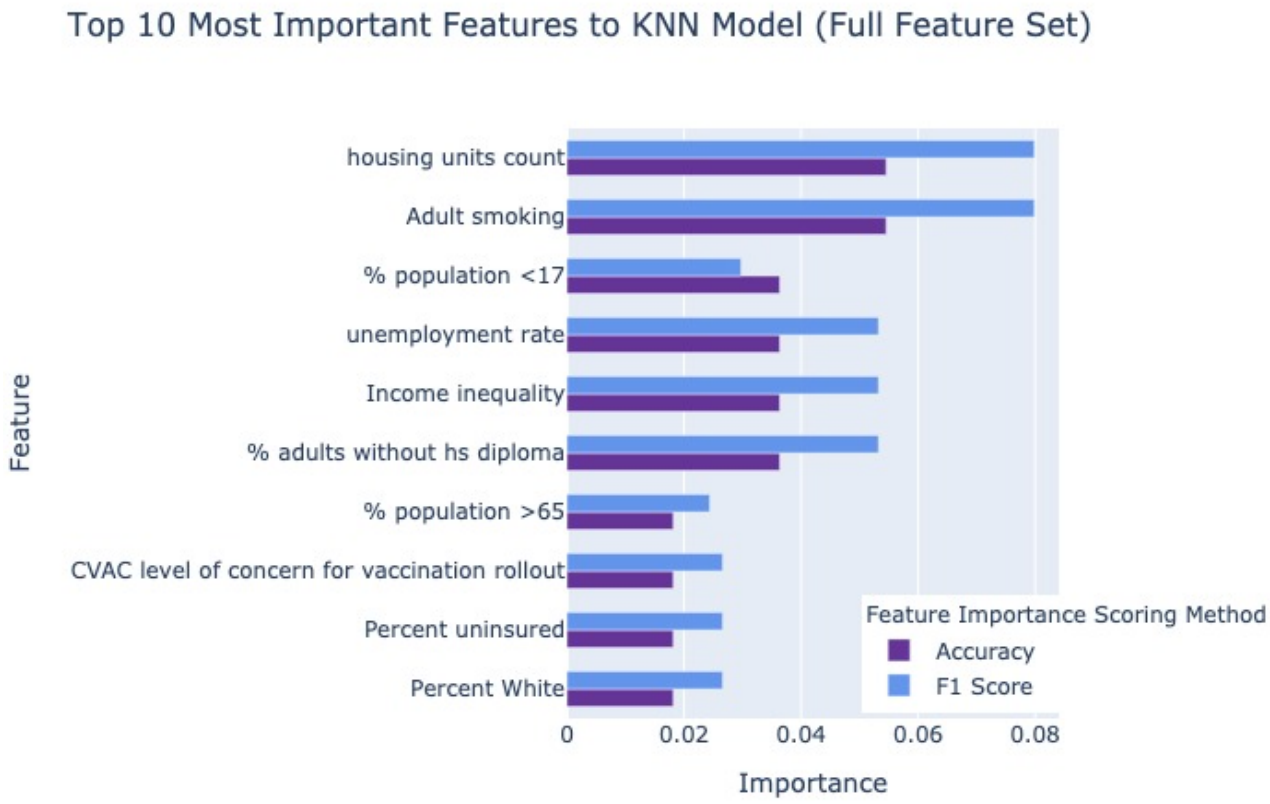


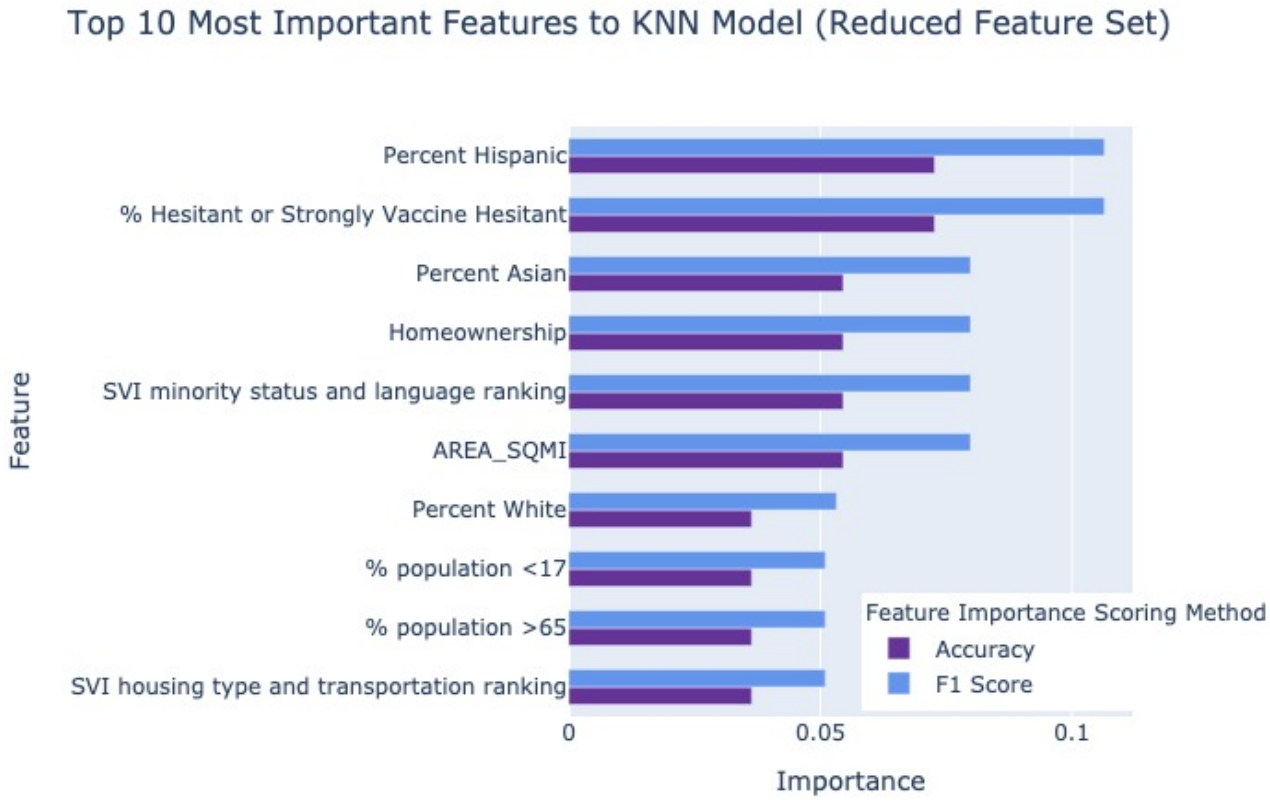
Based on population-controlled cases, deaths, and mortality rate of COVID-19, the following states experienced "high" impact of COVID-19: Arizona, South Dakota, Louisiana, Alabama, Mississippi, Pennsylvania, New Jersey, Connecticut, Rhode Island, and Massachusetts.

Given the health, social/demographic, economic, and COVID-19 measures, a K-Nearest Neighbors with optimal k=1 model predicted 91% of the validation dataset accurately.

When trained on the full dataset and a limited dataset excluding health health measures, the following features were considered the most important to making accurate predictions:

Economic Measures	Health Measures	Social & Demographic Measures
• Per capita income • Income inequality • Unemployment rate • Homeownership	• Smoking • Vaccine hesitancy	• Racial breakdown • Child population • Education • Area (sqmi)


Top 10 Most Important Features to KNN Model (Full Feature Set)
housing units count
Adult smoking
% population <17
unemployment rate
Income inequality
% adults without hs diploma
% population >65
CVAC level of concern for vaccination rollout
Percent uninsured
Percent White
Feature
0
0.02
0.04
0.06
0.08
Importance
Feature Importance Scoring Method
Accuracy
F1 Score


Top 10 Most Important Features to KNN Model (Reduced Feature Set)
Percent Hispanic
% Hesitant or Strongly Vaccine Hesitant
Percent Asian
Homeownership
SVI minority status and language ranking
AREA_SQMI
Percent White
% population <17
% population >65
SVI housing type and transportation ranking
Feature
0
0.05
0.1
Importance
Feature Importance Scoring Method
Accuracy
F1 Score

Recommendations

- **Measure what matters.**
  States should have accurate, frequent measurement plans in place for each of the features shown here to increase their community's vulnerability to extensive spread of illness and mortality. What you measure, you can manage.

- **Extend the reach of health budgets to invest in socioeconomic and access barriers.**
  When planning public health budgets, consider what social determinants - unemployment, education, income inequality, etc. - may play a role in health outcomes.

- **Ensure epidemic procedures at all levels of government enable resources to be allocated based on vulnerability.**
  In the case of a pandemic, organizations should be able to quickly decide and allocate emergency funds equitably - according to vulnerability - rather than equally - according to population - in order to achieve the most benefit from limited resources.

Future Enhancements

- Analyze and model at the county level of detail to capture more specific community-level disparities
- Conduct more rigorous feature selection to narrow down the features that most influence pandemic vulnerability
- Incorporate ICU/hospital capacity and economic measures (job loss, change in economic activity, bankruptcy, etc) into Pandemic Impact calculation

For further information:

Please review the full EDA notebook here and modeling notebook here, or review the non-technical presentation

For any additional questions, please contact kcoop610@gmail.com.

Repository Structure:

```
├── README.md
├── 1-EDA.ipynb
├── 2-model.ipynb
├── presentation.pdf
├── data
    └── source data files
├── data_compressed
    ├── merged_state_level_dataset.csv
    └── compressed data files
├── images
    └── images
└── styles
    └── custom.css
```

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%