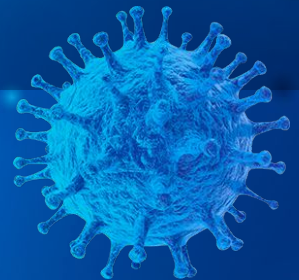
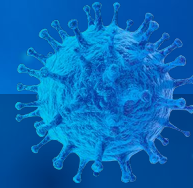


# Curbing the Curve:

Assessing Socioeconomic Factors in COVID-19 Risk Rates for the Purposes of Increased Resource Allocation

Joe Sanders



“

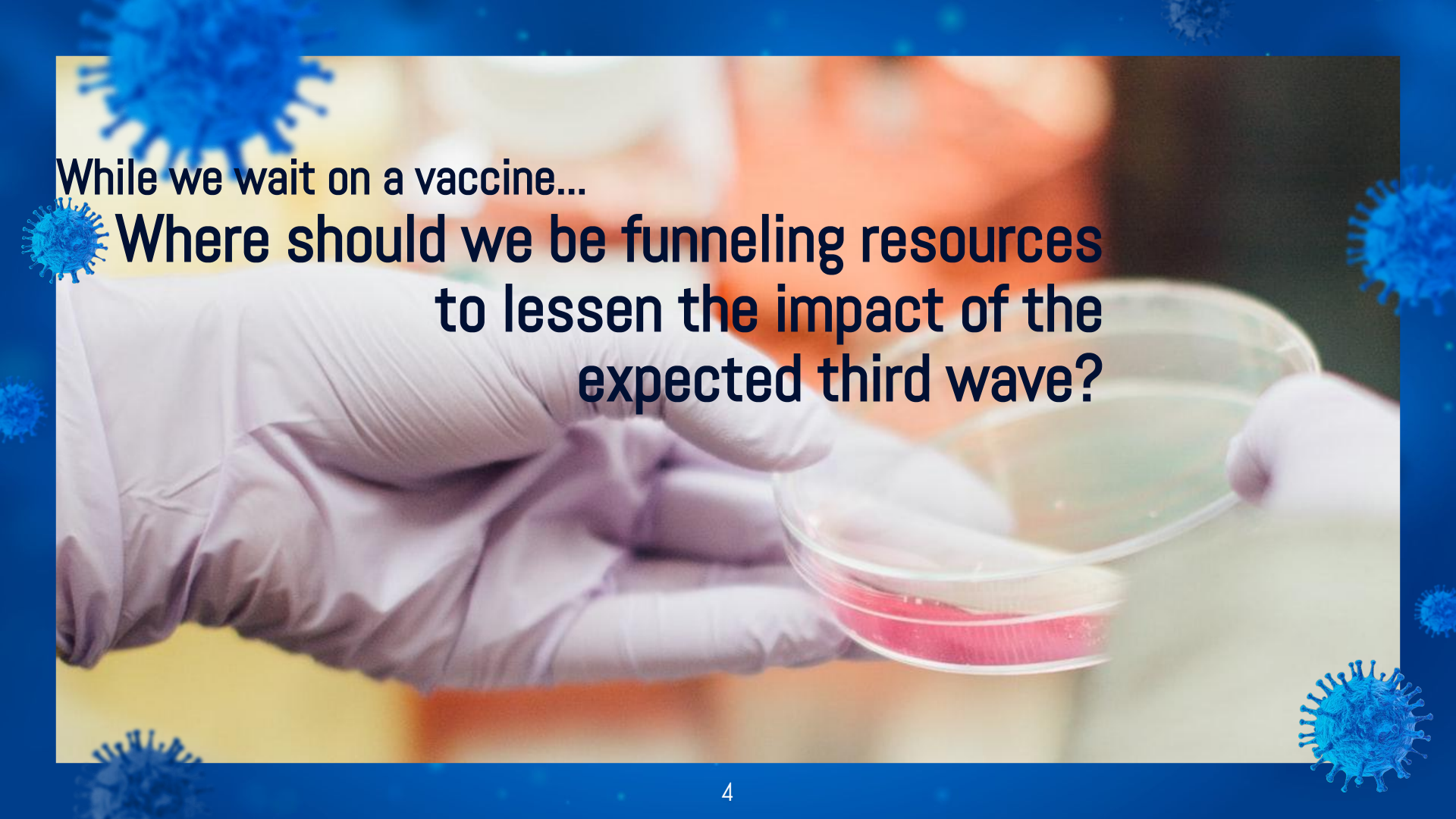
“The calvary is coming but don't put your weapons down, you better keep fighting because they are not here yet. ”

Dr. Anthony Fauci



The Roche Data Science Coalition (RDSC) is requesting the collaborative effort of the AI community to fight COVID-19, in what they are calling the UNCOVER challenge.

UNCOVER, which stands for **U**nited **N**etwork for **C**OVID Data **E**xploration and **R**esearch, in part seeks to identify which populations are at the greatest risk for COVID19.



While we wait on a vaccine...

**Where should we be funneling resources  
to lessen the impact of the  
expected third wave?**



The background of the slide is a solid blue color with several stylized, glowing blue virus particles scattered across it. These particles have a spherical core with many small, protruding spikes, resembling coronaviruses. They are positioned at various depths, creating a sense of a microscopic environment.

# Understanding the Current State

The median impact of COVID-19, and understanding where in the the country is being impacted the most

# 1 in 50

is the median infection rate\* across all counties in the United States

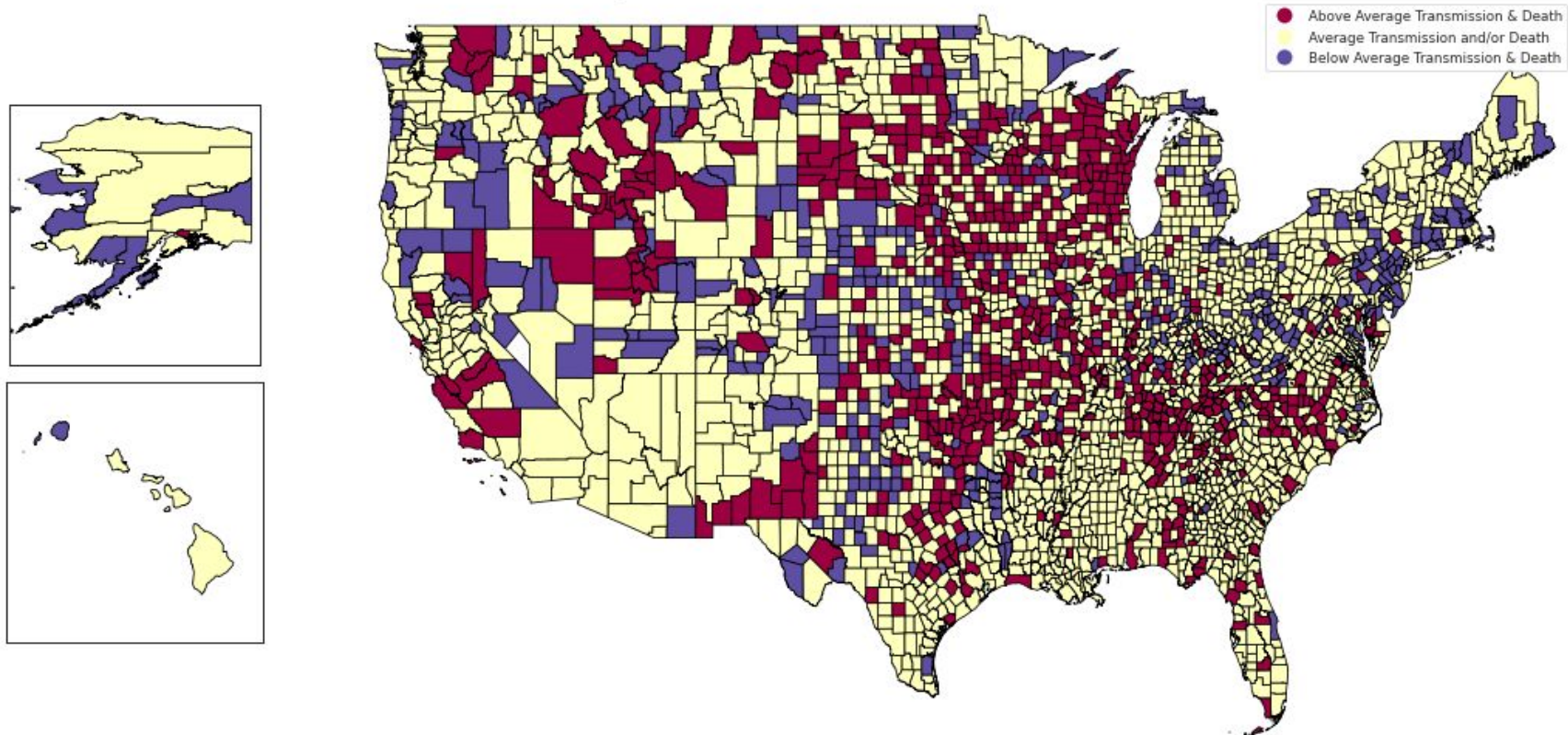
# 1 in 2000

is the median death rate\* across all counties in the United States



*\*accurate at time of data collection on October 24, 2020*

## Average COVID-19 Infections and Deaths in the United States

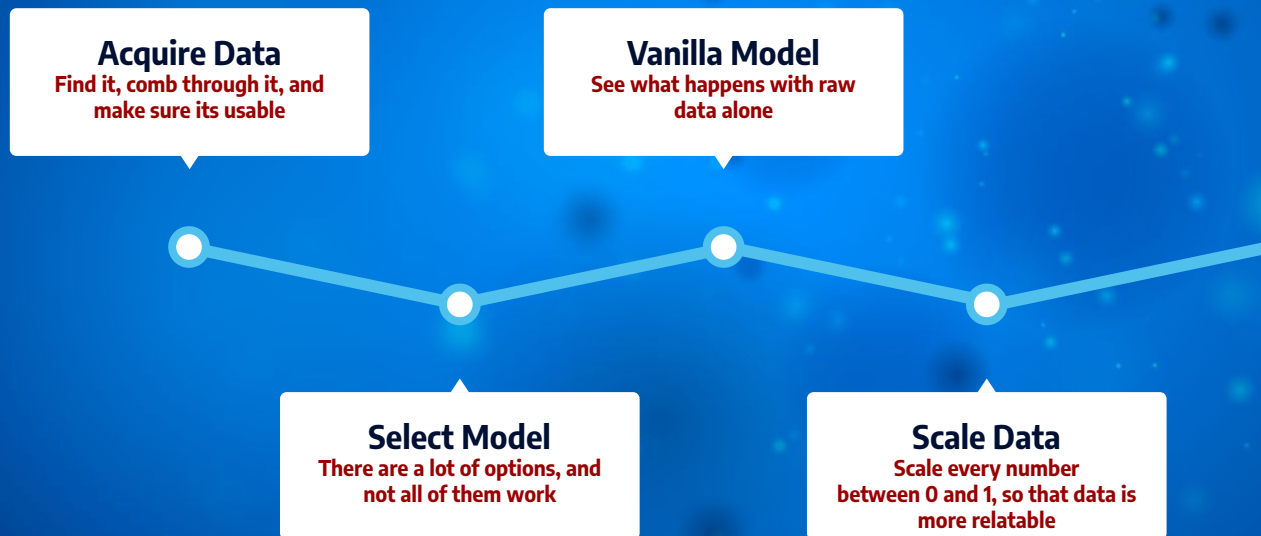


# The Model

Identify the protective factors and risk factors behind COVID-19 rates, using logistic regression with cross validation



# The Modeling Process



# The Modeling Process

## Polynomial Features

Factors A and B are good alone,  
but will they be great when  
multiplied together?

## Variance Threshold

When a factor doesn't vary  
much within itself, it generally  
isn't a good predictor

## L1 Regularization

Removes factors that are  
non-essential to model results  
(also known as LASSO)

## Parameter Tuning

Find the individual settings that  
will produce the best model  
result

# Understanding Model Outputs

## Low Risk Rates: Precision

What percent of “Negatives” are actually negative. The model will incorrectly identify a county as “At Risk” 7% of the time.

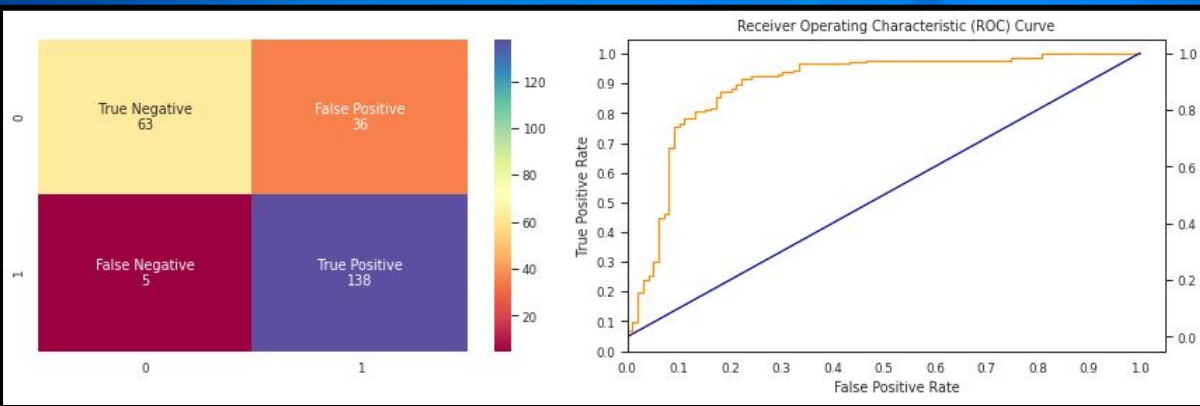
## High Risk Rates: Recall

What percent of “Positives” are correctly identified? The model will correctly identify “At Risk” counties 97% of the time.

	precision	recall	f1-score	support
0	0.93	0.64	0.75	99
1	0.79	0.97	0.87	143
accuracy			0.83	242
macro avg	0.86	0.80	0.81	242
weighted avg	0.85	0.83	0.82	242

**R-Squared: 83.1%**

**16.9% of the difference between counties could NOT be explained by the data provided.**



## ROC Curve

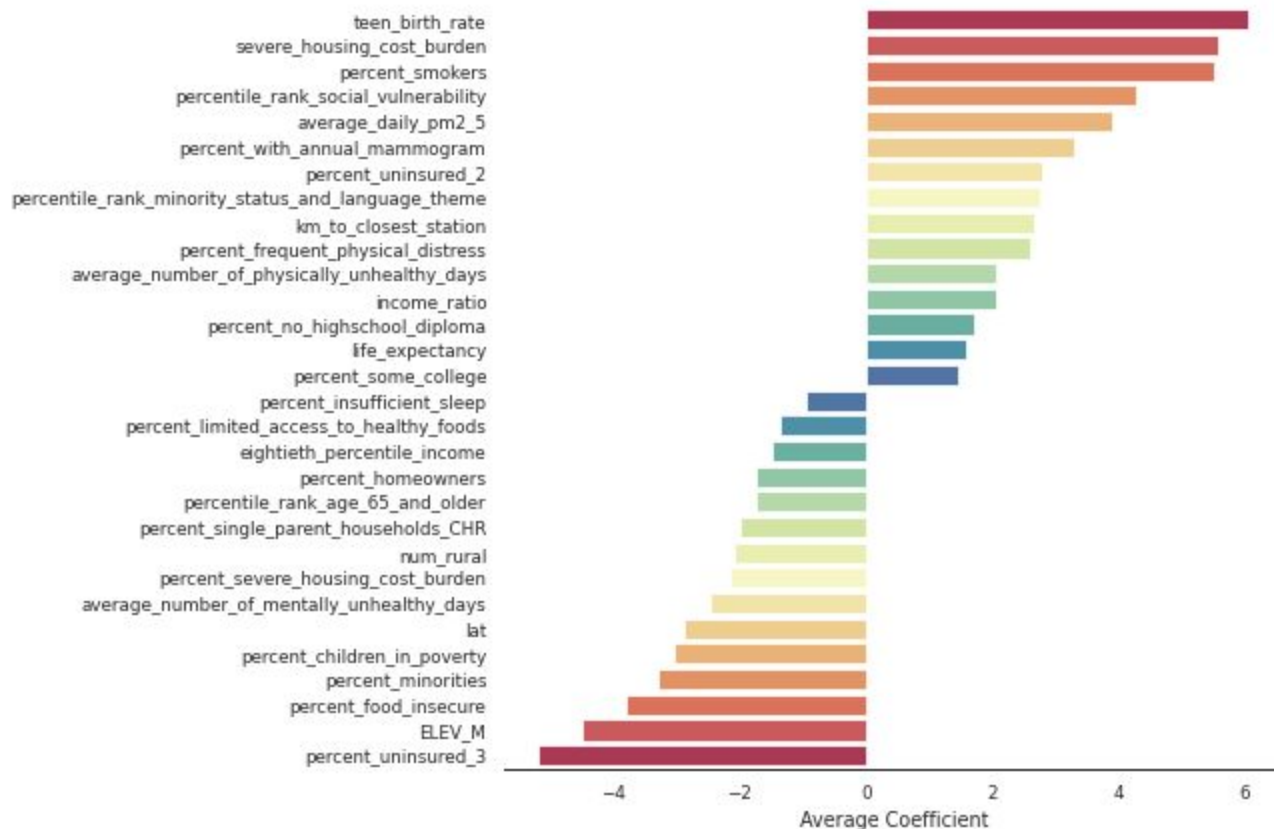
A visual way to understand model accuracy. Ideally, this curve will hug the left- and top-walls. A model that follows the blue line is effectively a guess.

# Analysis of the Results

What factors did the model identify as important, and where should we allocate resources?



## Top 30 Features with the Highest and Lowest Average Coefficient



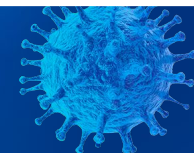
### Understanding Coefficients

Coefficients show the impact of a measure on the model.

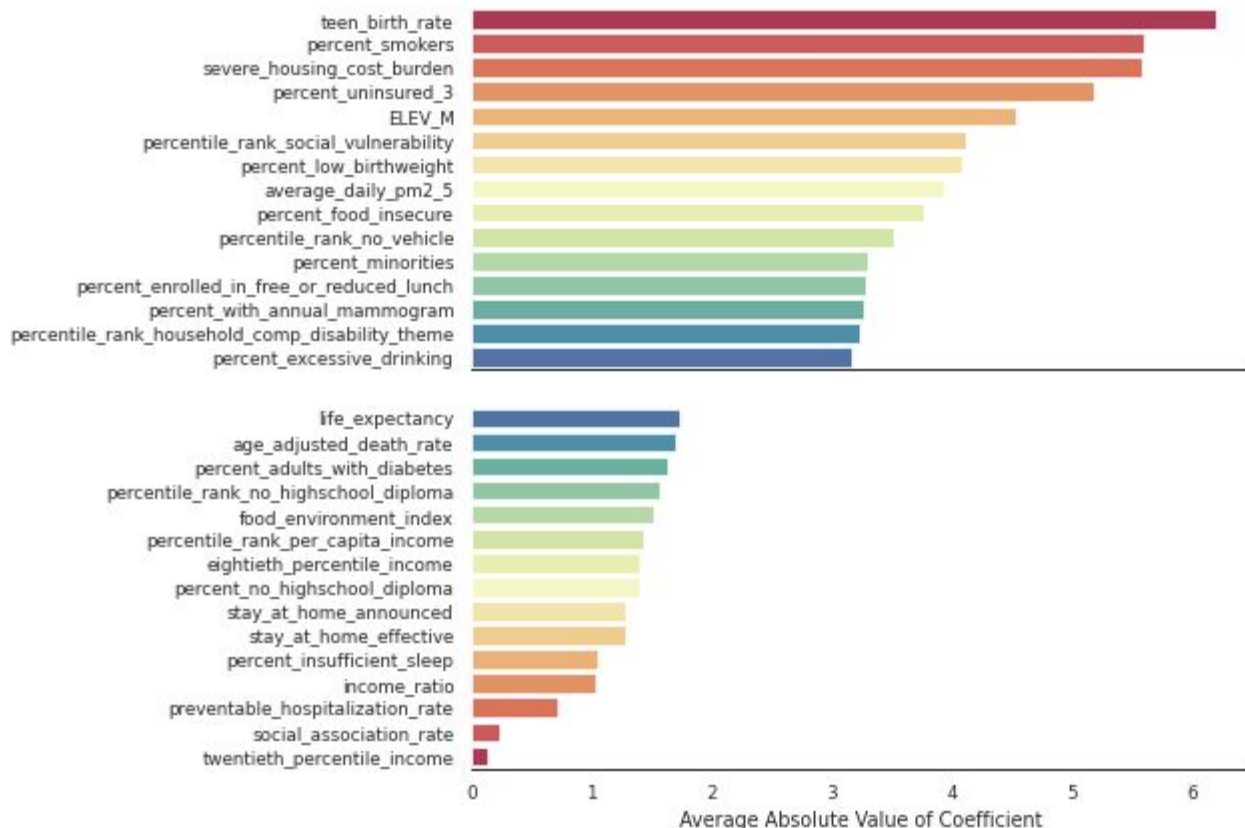
A positive coefficient indicates that as the value of the feature increases, the average risk also tends to increase.

A negative coefficient indicates that as the value of the feature increases, the average risk tends to decrease.

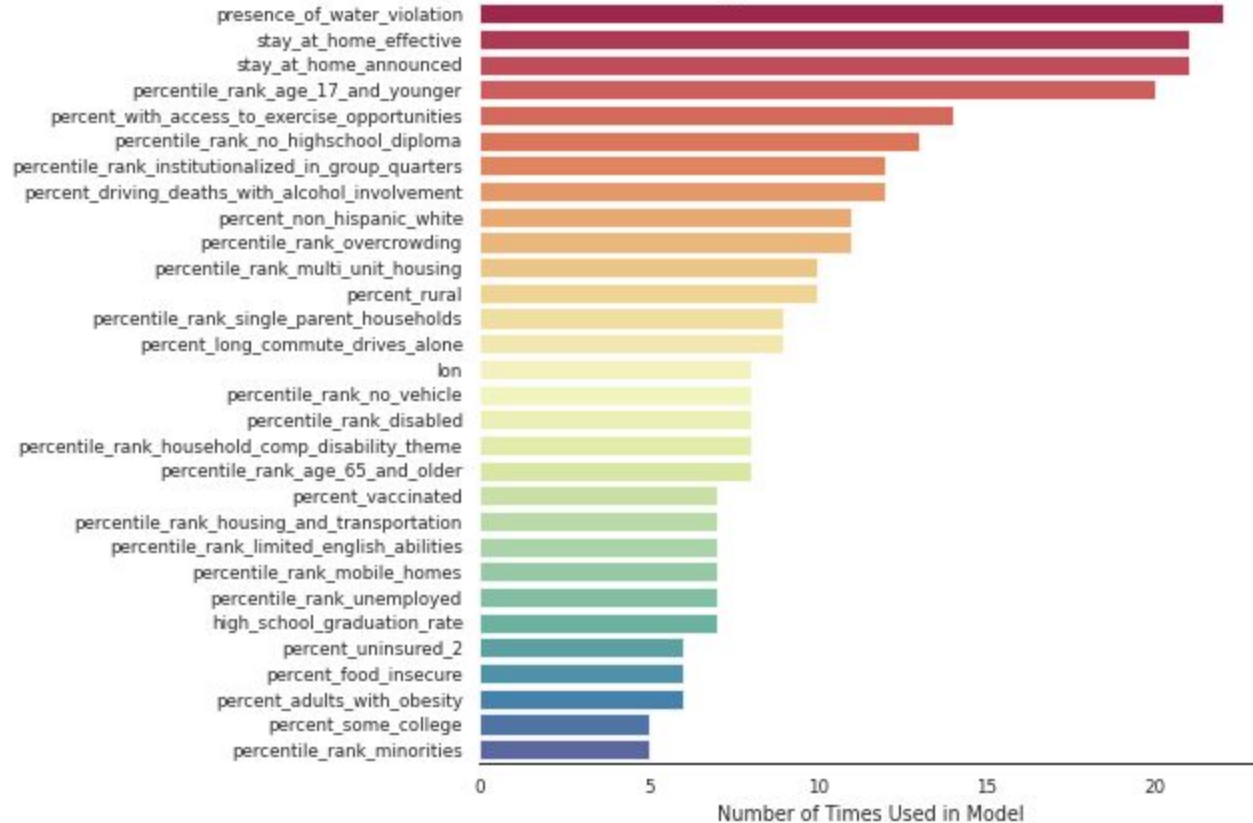
The further from zero a coefficient is, the stronger the impact on risk rate.



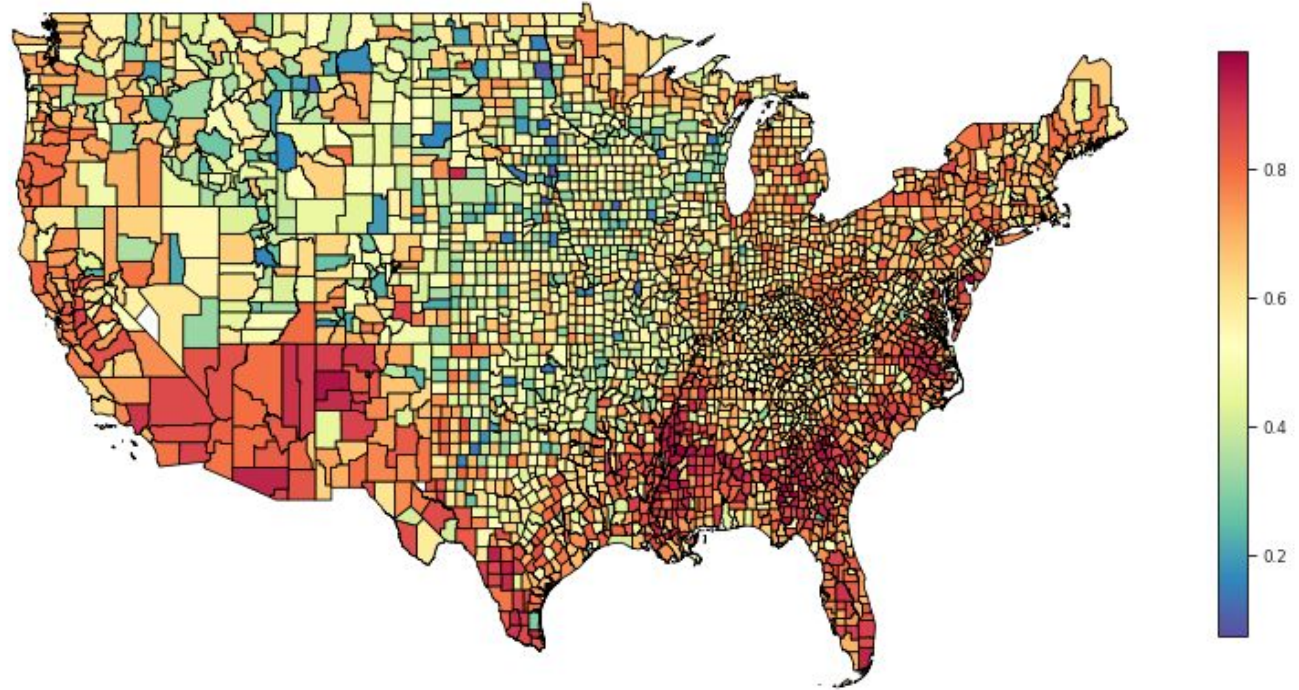
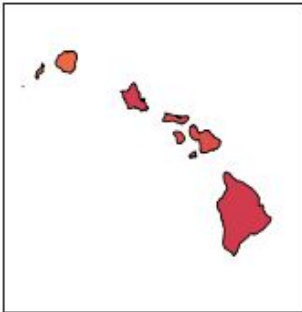
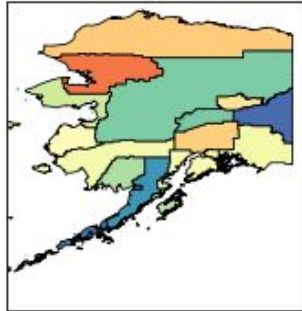
## Top 30 Features with the Highest and Lowest Average Absolute Value of Coefficient



## Top 20 Features By Number of Times Used in Model

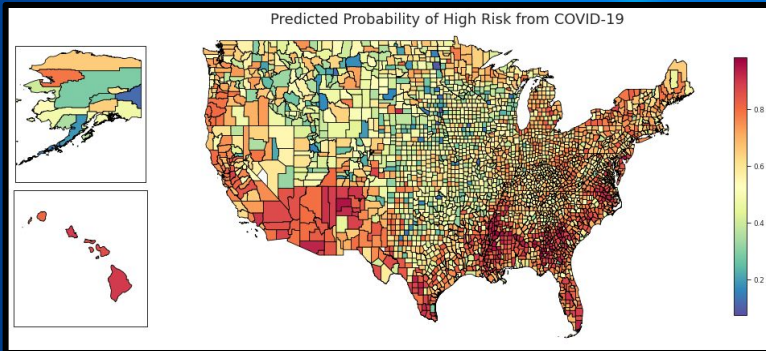
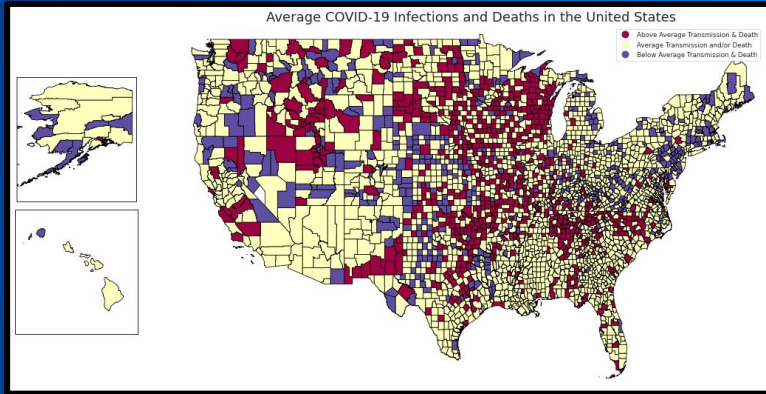


Predicted Probability of High Risk from COVID-19





# Finding: Don't Underestimate the Impact of Partisanship



## Politics may play a larger-than-expected role in COVID-19 risk rates

- ▶ The analysis fails to catch the current prevalence of the virus in the midwest and plains states.
- ▶ In the 2020 election, many of these states leaned or heavily favored the Republican party.

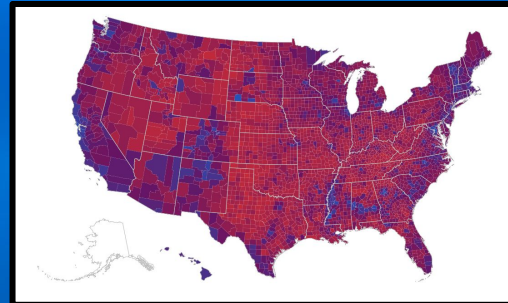
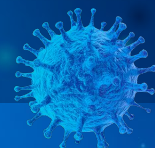
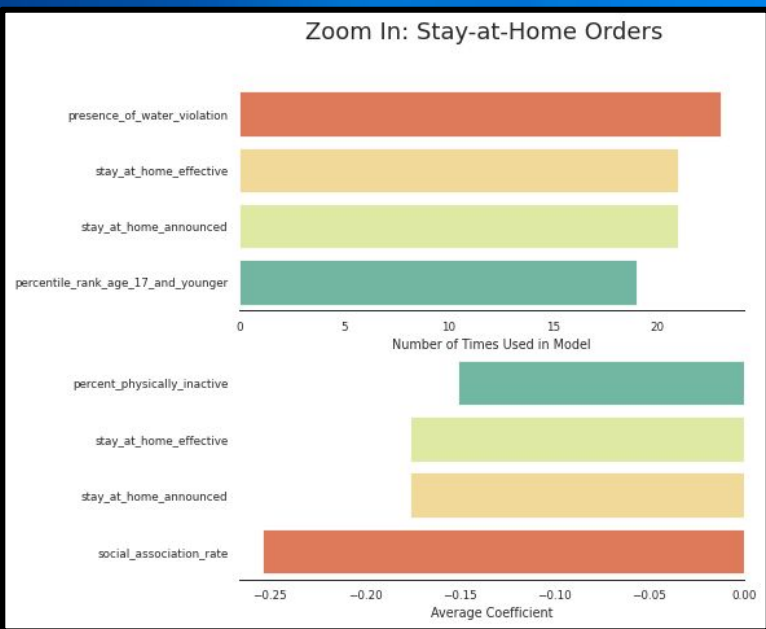


Image courtesy of Massachusetts Institute of Technology, USA TODAY analysis of Associated Press results data. Accurate as of 11/18/2020 at 12:05 pm EST.

# Finding: Issue Stay-at-Home Orders



Zoom In: Stay-at-Home Orders

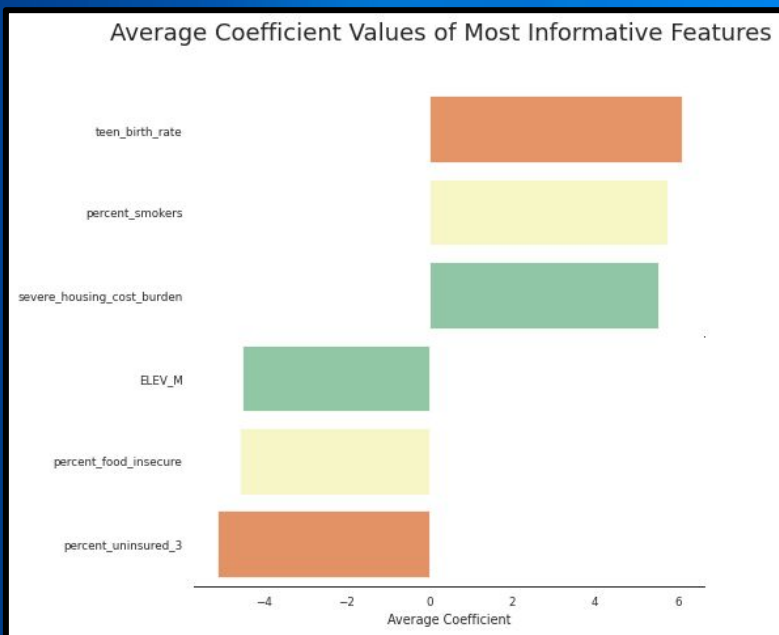


## Failure to Issue Stay-at-Home Orders Negate Protective Factors

- ▶ Stay-at-home orders are amongst the highest used features
- ▶ The average of coefficient is close to zero, meaning these features rely on feature interactions for its impact in the model
- ▶ A negative coefficient means impact on protective factors
- ▶ Stay-at-home orders are binary 0/1 scores

		Protective Factor
0	Stayed at Home Not Issued	Negated ( $X*0 = 0$ )
1	Stay at Home Issued	Maintained ( $X*1 = X$ )

# Finding: Focus on Areas with Extreme Poverty



## Many underlying risk factors are positively correlated with poverty

- ▶ **Teen Birth Rate:** Percent of women aged 15-19 who give birth in a calendar year
- ▶ **Percent Smokers:** Percent of residents 18 years or older who smoke cigarettes or use tobacco products
- ▶ **Severe Housing Cost Burden:** Citizens who pay more than 30% of their income for housing
- ▶ **Percent Food Insecure:** Percent of residents whose food intake is disrupted do to lack of money or resources
- ▶ **Percent Uninsured 3:** Percent of residents who are uninsured (3 month rolling average)

# Considerations for Future Research



## How do politics impact risk rates?

Analysis should be conducted on the impact of political affiliation, given the president's divisive rhetoric about the pandemic. Does this explain certain regions being impacted more than we would expect?

## How does climate impact risk rates?

Analysis should be conducted to see if certain climate and/or weather patterns either increase or decrease risk rates. While this was originally in scope for this project, due to high quantities of missing data, it was eventually excluded.

## What are the best interventions for high-risk areas?

Analysis should be conducted to determine what might most effectively mitigate identified risk factors: increased access to health resources, increased access to virus/pandemic education, or improved social safety net programs. Answering the question of “where” resources are allocated doesn’t answer “how.”





# Thank you

Want to continue the conversation?

**Github:** joesanders1010

**Medium:** @joesanders1010

**Facebook:** joesanders1010

**Twitter:** @joesanders1010

**LinkedIn:** /in/joesanders