PHYS378 Final Project

Vivian Vasquez, Kentaro Matsuoka

Problem: Since there has been much debate about how white poverty and frustration with the establishment drove Trump's unexpected victory in 2016, we would like to explore various socioeconomic and Quality of Life (QOL) measures like education level, income inequality, and life expectancy at the county level that may explain the 2016 election results. We plan to use a regression model to predict the ① county-level election results in 2012, 2016, 2020 from the ② County Health Rankings data set, which contain numerous socioeconomic and QOL variables at the county level. Although many studies have analyzed these relationships at the state and national levels, little work has been done at the county level. Thus, by increasing the number of observations from 50 to 3,000 (# of US counties), we believe that new trends may be revealed. Some cluster analyses with the ② dataset may also be revealing given its amount of data.

Data Source:

① County Presidential Election Returns 2000-2020 (12 Variables, 72617 Observations; MIT Election Data and Science Lab; 6/22/21)

"This dataset contains county-level returns for presidential elections from 2000 to 2020."

② 2021 County Health Rankings (~690 Variables, ~3196 Observations (per year, available from 2010-21); U of Wisconsin Population Health Institute)

This data file includes county level metrics of various demographic and health-related data. There are a few that also measure education levels and income inequality, which we believe would be interesting to explore.

Where can I find the data set?

① https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ

② https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation

Where can I find the column information?

① https://github.com/MEDSL (Since dataset is election results, column information is straightforward; however, this GitHub repository contains additional relevant info.)

② https://www.countyhealthrankings.org/sites/default/files/media/document/Trends%20documentation%202021.pdf (contains detailed information about all metrics/indices)

Relevant Papers: https://electionlab.mit.edu/research#reports (working papers and projects from MIT Election Data + Science Lab)

What Predicted Voting in the 2020 Election Cycle

Kentaro Matsuoka, Vivian Vasquez

Final Project Presentation

The life cycle of knowledge mining

Idea 1
Define a problem
Generate hypothesis

Data 2
Collect data

Data preparation 3
Data cleaning
Feature extraction
Data visualization
Sample selection

Model 4
Classification
Regression
Clustering
Density estimation

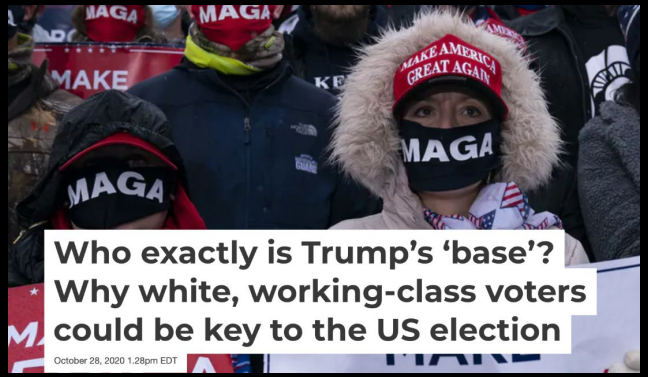
Evaluation 5
Visualization
Goodness of the fit
Outlier detection
Hypothesis testing

Knowledge 6
Interpretation
Discovery

Idea
Define a problem
Generate hypothesis
1

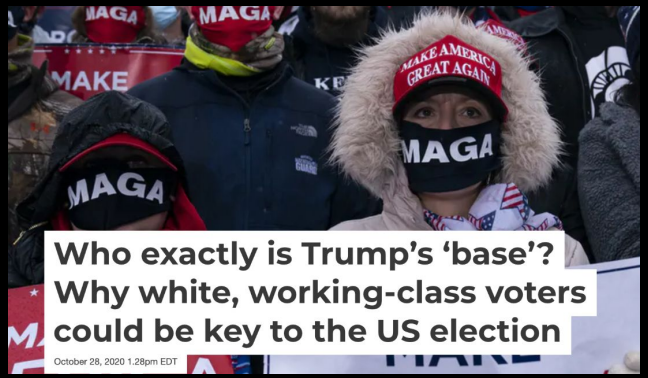
1. Idea: Background

- Debate about Trump's victory in 2016
  - How much of election results could be attributed to frustration with the establishment and white poverty?



1. Idea: Hypothesis

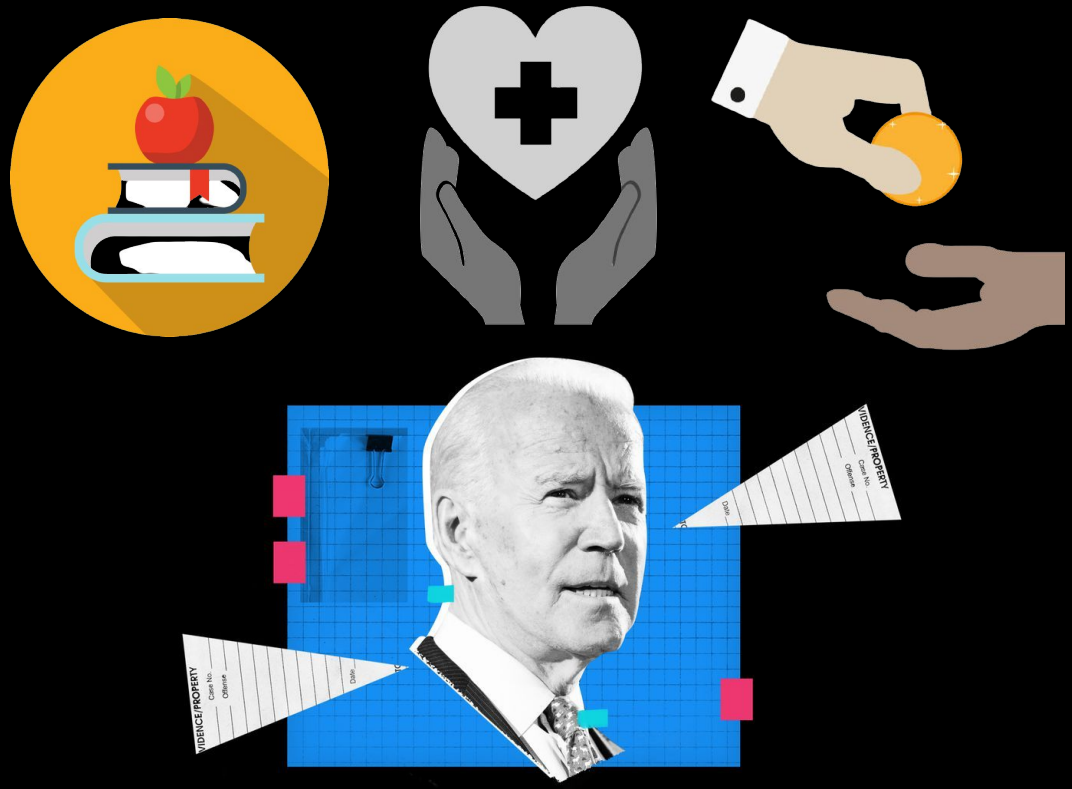
- 2020 Election Results will be best predicted again by race, poverty, education levels, and related socioeconomic indicators especially amongst Trump-supporting Republican voters.



Idea
Define a problem
Generate hypothesis
1

1. Idea: Question

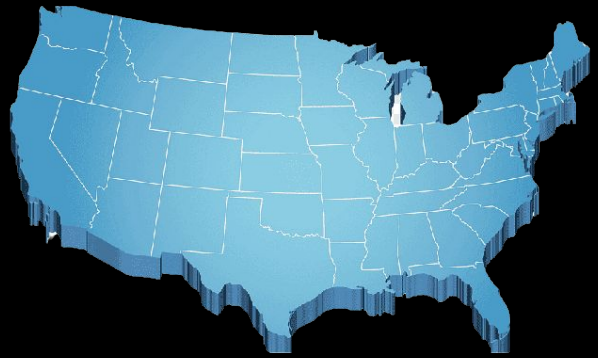
How did quality of life (QOL) measures like socioeconomic status, income inequality, and education at the <u>county level</u> contribute to the 2020 election results?

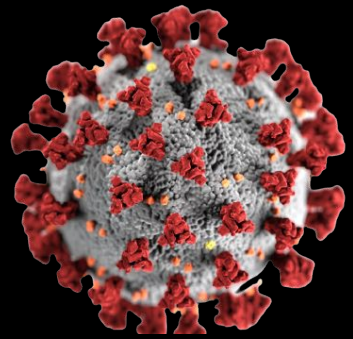


1. Idea: Importance of this Question

BLUEBONNET DATA

- Novelty of health data analysis at county level during COVID-19
- Find what metrics are related to voter behavior
- Mail in voting upsurge in 2020
- Personal interest





2. Data Source

- County Presidential Election Returns 2000-2020 (12 Variables, 72617 Observations, 6/22/21)
  - From MIT Election Data and Science Lab
  - “contains county-level returns for presidential elections from 2000 to 2020.”
- County Health Rankings (~690 Variables, ~3196 Observations (per year, available from 2010-21))
  - From University of Wisconsin Population Health Institute
  - County-level metrics of various demographic and health-related data

Where can you find the data:

[County Presidential Election Returns 2000-2020 - US Presidential Elections]

[County Health Rankings Data]

MIT ELECTION DATA + SCIENCE LAB

University of Wisconsin Population Health Institute
SCHOOL OF MEDICINE AND PUBLIC HEALTH

County Health Rankings & Roadmaps
A Healthier Nation, County by County

3. Data Preparation: Data Cleaning

- Problem: REALLY MESSY DATASET!! (esp voting df)
- #socialscienceproblems
- Goal: Get one clean DataFrame by merging the voting data with the health data via FIPS code (unique county identifier)

3. Data Cleaning: War with NaN

- Voting data = df_voting
- A LOT of NaN problems
- D.C. FIPS code NaN for some reason, so set manually to correct FIPS (11001)
- Many other rows with NaN FIPS
  - From clerical errors, redundancy, etc.
  - Especially with 2020 voting changes (mail-in, early voting, etc.), data needed to be further tweaked
  - Many FIPS codes were NaN due to special voting measures implemented in 2020
  - But! Able to identify causes of all but one of the FIPS NaN, so essentially dropping just one significant observation…should not affect results.

53321	2020	DISTRICT OF COLUMBIA	DC	DISTRICT OF COLUMBIA	NaN	PRESIDENT	DONALD J TRUMP	REPUBLICAN	18586.0	344356.0	20210622	TOTAL
53321	2020	DISTRICT OF COLUMBIA	DC	DISTRICT OF COLUMBIA	11001.0	PRESIDENT	DONALD J TRUMP	REPUBLICAN	18586.0	344356.0	20210622	TOTAL

3. Data Cleaning: San Joaquin Example

- Other df_voting issues… San Joaquin County example
  - Predict based on county voting history
  - Likely caused by how votes for 3rd party candidates are tallied

	year	state	county	fips	candidate	party	candvotes	totalvotes	mode
52780	2020	CALIFORNIA	SAN JOAQUIN	6077	JOSEPH R BIDEN JR	DEMOCRAT	161137.0	NaN	TOTAL
52781	2020	CALIFORNIA	SAN JOAQUIN	6077	OTHER	GREEN	1064.0	NaN	TOTAL
52782	2020	CALIFORNIA	SAN JOAQUIN	6077	JO JORGENSEN	LIBERTARIAN	2929.0	NaN	TOTAL
52783	2020	CALIFORNIA	SAN JOAQUIN	6077	OTHER	OTHER	NaN	NaN	TOTAL
52784	2020	CALIFORNIA	SAN JOAQUIN	6077	DONALD J TRUMP	REPUBLICAN	121098.0	NaN	TOTAL
	year	state	county	fips	candidate	party	candvotes	totalvotes	mode
52780	2020	CALIFORNIA	SAN JOAQUIN	6077	JOSEPH R BIDEN JR	DEMOCRAT	161137.0	286228.0	TOTAL
52781	2020	CALIFORNIA	SAN JOAQUIN	6077	OTHER	GREEN	1064.0	286228.0	TOTAL
52782	2020	CALIFORNIA	SAN JOAQUIN	6077	JO JORGENSEN	LIBERTARIAN	2929.0	286228.0	TOTAL
52783	2020	CALIFORNIA	SAN JOAQUIN	6077	DONALD J TRUMP	REPUBLICAN	121098.0	286228.0	TOTAL

3. Data Cleaning: Other Issues

- Many other extraneous observations
- Where totalvotes = 0, drop, since not adding any new info to model
- Rename columns for clarity and conciseness
- NaN in candidate votes column
    - Drop these observations, since cases where the 3rd party candidate (usually the Green party) was not officially on the ballot due to different election laws across states to get on ballot
    - Likely just caused most would-be Stein, (Gary) Johnson, etc. voters to vote for them via a write-in candidate
    - Caused a simple spillover of the 3rd party candidate votes to votes for the 'OTHER' candidate, which is kept in df
    - Thus, dropping these rows also have no material effect on the model

3. Merge DFs via FIPS Code! (finally)

- Change FIPS from floats/strings to ints to join df_voting to df_health by ‘fips’ key
- Exclude FIPS not in both data sets, i.e. take intersection of the 2 sets
- Create new indicator variable (the party that got plurality votes in each county)
  - Was either Republican or Democrat in each county in 2020
  - No 3rd party got plurality in any county
  - So either 0 (for Reps) or 1 (for Dems)

3. Some Final Cleaning

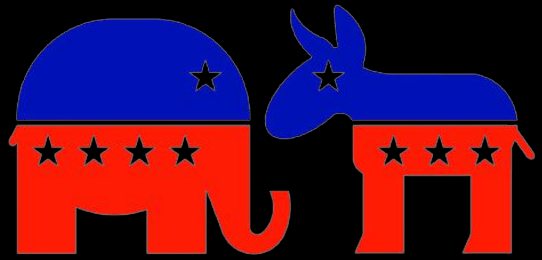
- Check that df_voting and df_health were joined correctly
- Several insignificant mismatches
  - Most due to simple differences in spelling, which can be ignored
  - Like La Porte County in Indiana, spelled La Porte in df_health but LAPORTE w/o the space in df_voting
- 2 major mismatch errors
  - 1. FIPS code 36000
    - Correctly linked to NY statewide data in df_health (note the 000 ending reserved for states)
    - Erroneously linked to Kansas City, Missouri in df_voting
    - No county of that name, and if results for Kansas City, MO, no city-level data exists in df_health
  - 2. The Alaskan Issue (AK)
    - In df_voting, votes tallied by State Congressional district instead of actual Alaskan Districts (counties)
    - So FIPS codes used in df_voting are incorrect: they simply tack on State Congressional district number (represented in 3 digits) to 02, the first 2 digits of the AK state FIPS code (e.g. 02016 16th district)
    - Thus, FIPS codes don't match for most cases, except when they coincidentally do 3 times, but the areas do not overlap or resemble each other
    - Solution: drop these data points, and replace them with one statewide observation
    - AK is one of the least populous states, so substitution appropriate with little effect on model

3. Final tally

- 3,115 observations, a.k.a counties and county-equivalents
- 3,143 was final count in 2020
- 3,143 - 3,115 = 28 !
- 557 input variables (down from 690 in original health df due to many extraneous blank or NaN columns)
- 3,115 x 557 = ~1.7 million data points

4. Model Classification and why we chose our model

- Y: Winning Party
- X: Health and Income Data
- Predicting class (Republican or Democrat) based on health and income data



	model_name	accuracy
0	KNN	0.849117
1	Decision Tree	0.796148
2	Random Forest	0.841091
3	XGBoost	0.852327
4	SVC	0.844302

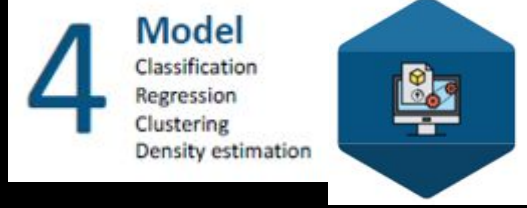
- Highest Accuracy : XGBoost
- Least Accuracy: Decision Tree

4 Model
Classification
Regression
Clustering
Density estimation

4. Model Classification and why we chose our model

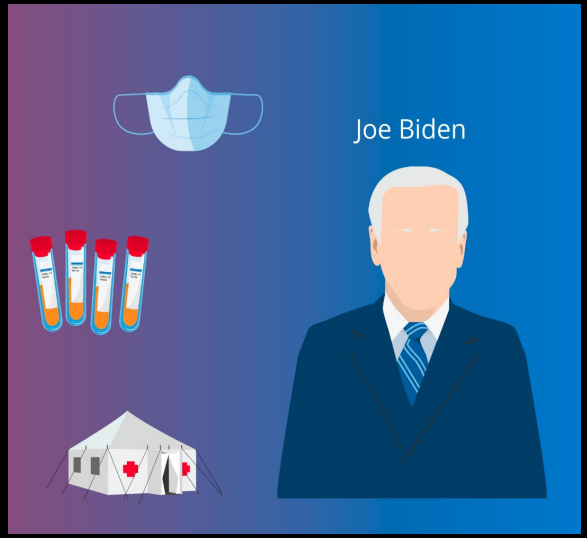
- XGBoost Model
  - Maximizes execution speed and model performance compared to Random Forest, Decision Trees, etc.
  - Uses a gradient boosting decision tree algorithm to minimize loss when adding new models (why Decision Tree viz looks bad)
  - Works with regression and classification predictive modeling / binary classification problems like ours

4 Model
Classification
Regression
Clustering
Density estimation

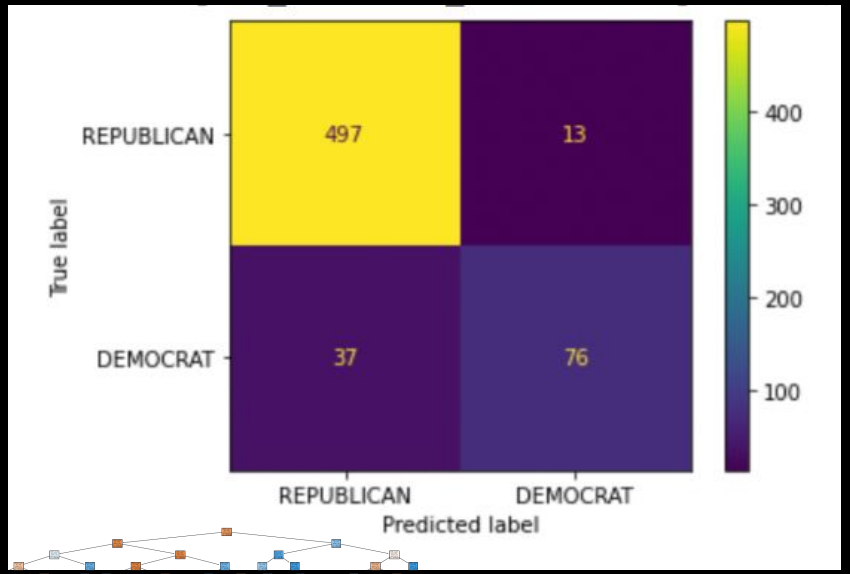


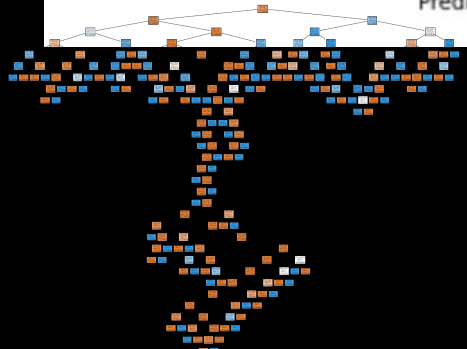
4. Sample Selection

- Filtering to only 2020 Election Data and Health Data
- Given COVID-19 pandemic and potential unique health outcomes
- Scrutiny of data analysis in 2016 given to 2016 surprise Trump win



5. Evaluation: Confusion Matrix





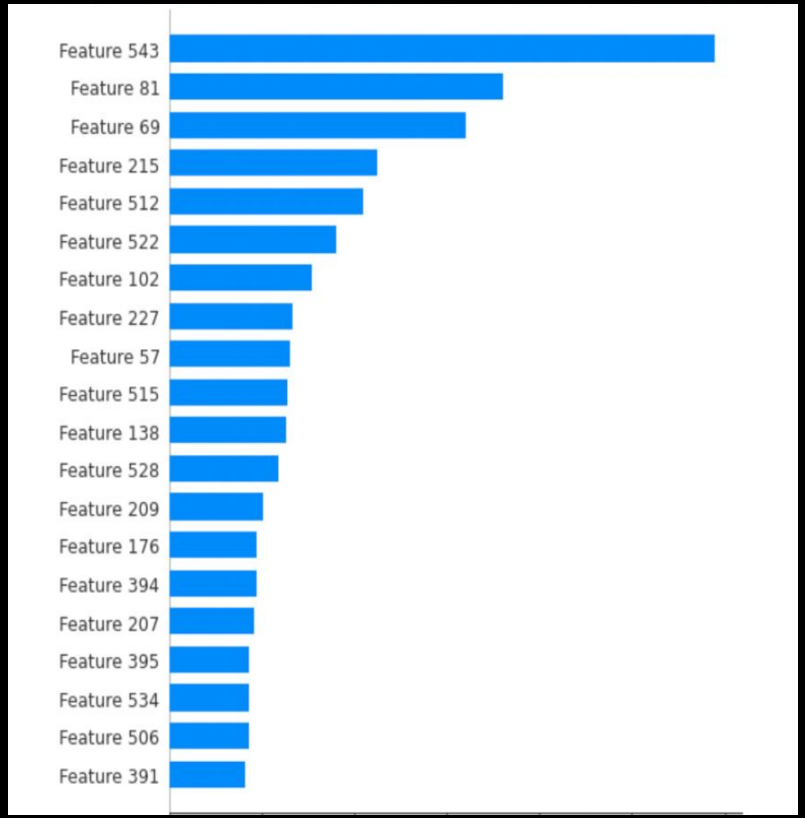
- Relatively low levels of False Positives and False Negatives (False Democrat and False Republican) Note: 1=Dem, 0=Rep
- Many more Republican counties than Dem counties
- Trump 2,496 (84%) vs Biden 477 (16%) (source: AP)

- Decision Tree too convoluted

5 Evaluation
Visualization
Goodness of the fit
Outlier detection
Hypothesis testing

5. Evaluation: SHAP Feature Importance Plot

- Tells you top 20 features of importance:
- TOP 5
  - Feature 543: % Non-Hispanic White raw value
  - Feature 81: Sexually transmitted infections raw value
  - Feature 69: Physical inactivity CI high
  - Feature 215: Percentage of households with high housing costs CI low
  - Feature 512: Severe housing cost burden raw value

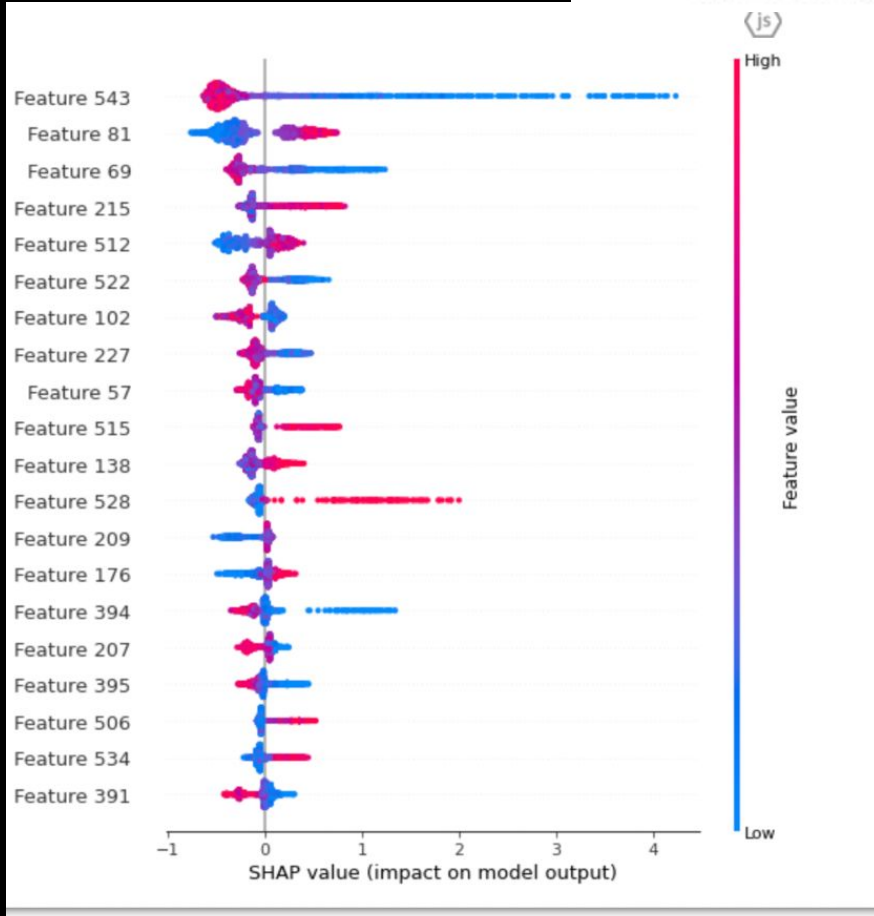


5. Evaluation: SHAP Summary Plot

5 Evaluation
Visualization
Goodness of the fit
Outlier detection
Hypothesis testing

Y: 1 = Dem (Biden), 0 = Rep (Trump); feature important + effects

Feature 543: % Non-Hispanic White raw value
Feature 81: Sexually transmitted infections raw value
Feature 69: Physical inactivity CI high
Feature 215: Percentage of households with high housing costs CI low
Feature 512: Severe housing cost burden raw value
Feature 522: % below 18 years of age raw value
Feature 102: Teen births CI low (White)
Feature 227: Driving alone to work CI high
Feature 57: Adult smoking CI low
Feature 515: Severe housing cost burden CI low
Feature 138: Flu vaccinations (White)
Feature 528: % Non-Hispanic Black raw value
Feature 209: Homeownership denominator
Feature 176: Children in single-parent households raw value
Feature 394: Uninsured children CI low
Feature 207: Air pollution - particulate matter raw value
Feature 395: Uninsured children CI high
Feature 506: Traffic volume raw value
Feature 534: % Asian raw value
Feature 394: Uninsured children raw value



6. Knowledge: Interpretation

- Feature 543: Percent White ⇒ Counties with a higher percentage of white people are correlated with more Trump voting
- Not riveting information, but useful in continuity from 2016 election in terms of what seems to be associated with Republican voting

6. Knowledge: Interpretation

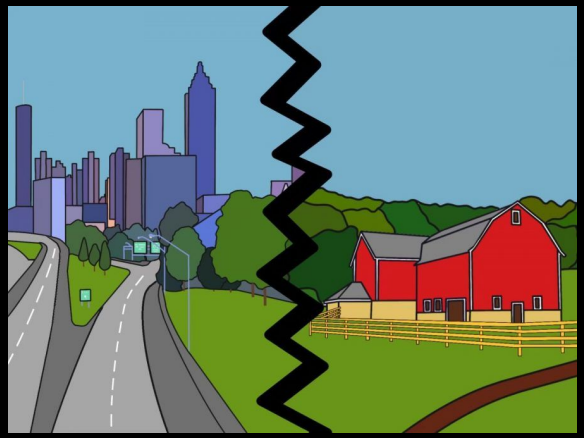
- Feature 512: Housing cost burden ⇒ high housing cost burden is associated with higher likelihood to vote for Biden
- Biden voters as the liberal elite?

6. Knowledge: Interpretation

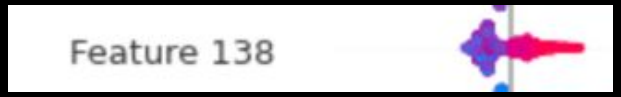
- Feature 69: Physical inactivity levels ⇒ Low physical inactivity i.e. high physical activity is correlated with more Democratic voting
- Conjecture: Time spent online during COVID -19 associated w/ Trump voting

6. Knowledge: Interpretation

- Feature 102: White Teen Births ⇒ higher white teen births correlated with Trump voting
- Less access to reproductive health resources (contraception, abortion) in rural areas with more strict laws on abortion



6. Knowledge: Interpretation



- Feature 138: White vaccination rate ⇒ high white vaccination rate in a county is correlated to higher voting for Biden

6. Knowledge: Interpretation

- Feature 528: Percent Black ⇒ Counties with a higher percentage of black are correlated with more Biden voting
- Again: Not riveting information, in line with Democratic voting trends since 1920s New Deal

6. Knowledge: Interpretation

- Feature 394: Uninsured children ⇒ higher rate of uninsured children is correlated with higher Trump voting
- Continuities from Boston Review article from Race, Politics, and Law class
- Conjecture: higher free healthcare access amongst minority, Democrat families

RACE

Dying of Whiteness

State policies shaped by white supremacy increase mortality rates in much the same way as other manmade health risks, such as pollution.

JUNE 27, 2019

JONATHAN M. METZL

Image: Fibonacci Blue

In early 2016 I met Trevor, a forty-one-year-old uninsured Tennessean who drove a cab for twenty years until worsening pain in the upper-right part of his abdomen forced him to see a

6. Knowledge: Interpretation

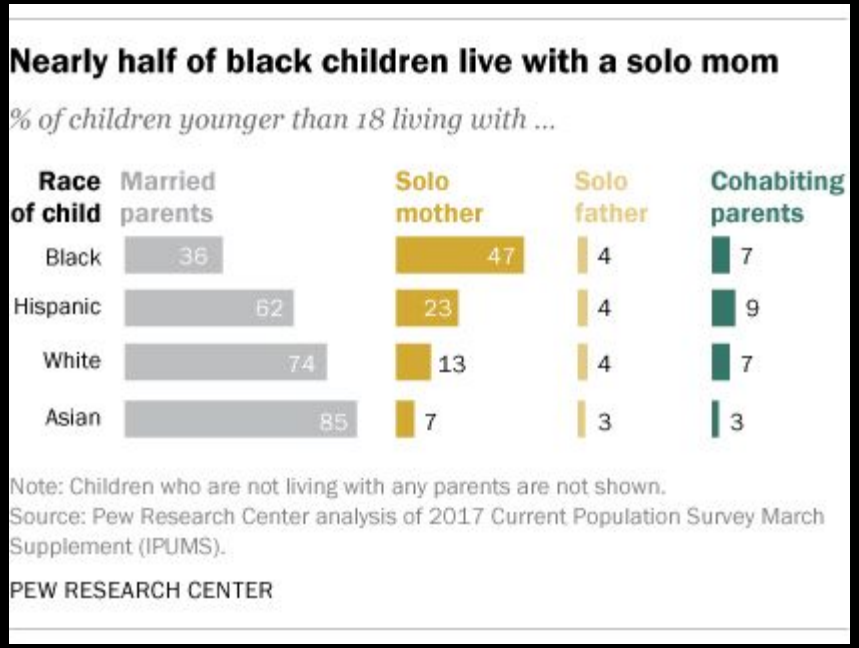
- Feature 57: Adult smoking ⇒ Higher adult smoking correlated with Trump voting
- Contrary to Gallup Findings on Political Viewpoints and Smoking Cigarettes in early 2000s

Political Viewpoints

Democrats are somewhat more likely than Republicans to smoke cigarettes. One in four Democrats (25%) smoke cigarettes, while 21% of Republicans do. Twenty-eight percent of independents smoke.

6. Knowledge: Interpretation

- Feature 176: Children in single parent households ⇒ More single parent children a county has more likely to be a Biden voting county
- May be associated with previously known voting trends: Black and Latinx households are single parent households at a higher rate + Black and Latinx voters tend to lean Democrat



6 Knowledge
Interpretation
Discovery

6. Evaluation of our Hypothesis

- Correct hypothesis: race, poverty (as associated with housing prices and costs) seem to be the features most taken into account by model
- In line with Pew Research's "Important Issues in the 2020 Election" article

Top issues for Trump supporters are economy, crime; Biden supporters prioritize health care, coronavirus

% of registered voters saying each is 'very important' to their vote in the 2020 presidential election

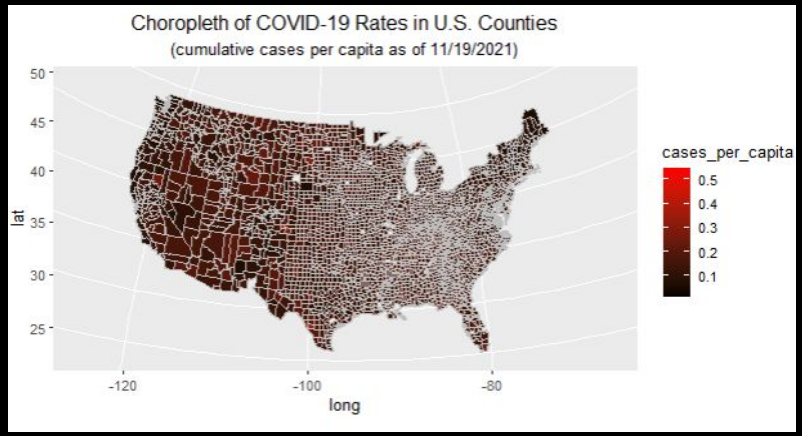
	Biden supporters	Trump supporters	Total
Economy	72	88	79
Health care	48	84	68
Supreme Court appointments	61	66	64
The coronavirus outbreak	39	82	62
Violent crime	46	74	59
Foreign policy	57	57	57
Gun policy	50	60	55
Race and ethnic inequality	24	76	52
Immigration	46	61	52
Economic inequality	28	65	49
Climate change	11	68	42
Abortion	35	46	40

Note: Based on registered voters.
Source: Survey of U.S. adults conducted July 27-Aug. 2, 2020.

PEW RESEARCH CENTER

6. Knowledge: Potential Future Research

- Latitude/Longitude County Map (potential choropleth visualization we could do from mapping health data)
- Extend model beyond 2020: “This data file contains constituency (state-level) returns for elections to the U.S. presidency from 1976 to 2020.”



Thank you!