FROM MASKS TO LIBERTY

COVID-19 Vaccination
Machine Learning Capstone Project

Hiywete Solomon
Rachael Wright
Rhonda Schwartz
Tyler Cooper

Project Overview

- Disparities in attitude of the vaccine exist by race/ethnicity as well as political position which may impact COVID-19 vaccination rate.
- **Original Hypothesis:** Can we predict when certain states will achieve herd immunity for COVID-19 (70%+) using machine learning
- **Modified Hypothesis:** Can COVID-19 vaccination rates determine political affinity of a county?
- To test this hypothesis, machine learning algorithms were used to demonstrate that racial disparities impact vaccination rate and that we can predict which counties are democratic or republican based on the number of people vaccinated.

Assumptions and Constraints

- COVID-19 vaccination data updated daily
- Not all counties are submitting timely CDC vaccination data
- Data timeframes differ based on source
- 2021 County Health Rankings Flu vaccination features are based on 2018 data
- County political designation from presidential election assumes no changes since November 2020

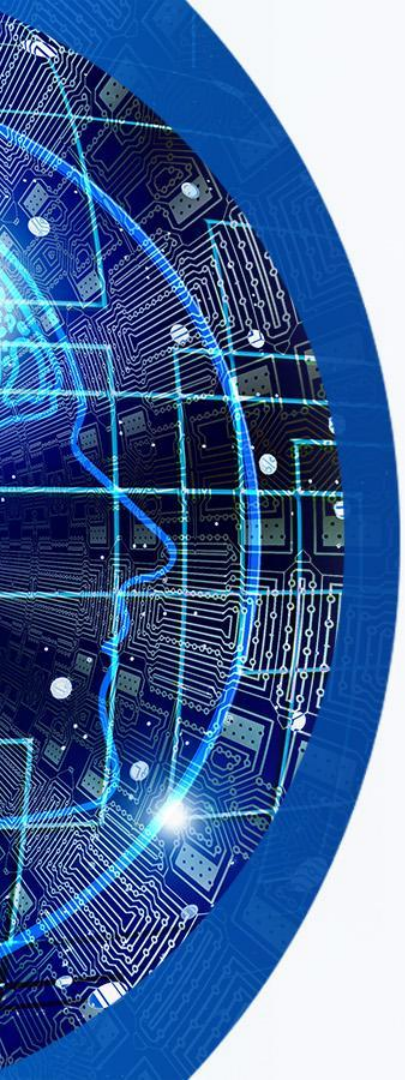
Data Sourcing, Ingestion, Storage

- CDC COVID-19 Vaccination Data for U.S. Counties - as of 6/16/21
- County Health Rankings and Roadmaps Data
- Kaggle - 2020 U.S. Presidential Election County Data

- Uploaded .csv files to an Amazon Web Services (AWS) - Simple Cloud Storage (S3) Bucket
- Python via Jupyter Notebook used to connect to S3 bucket, wrangle data, feature engineer, and build classification models on the county-level data

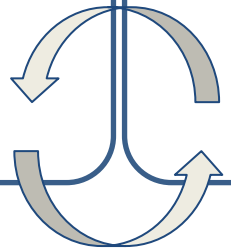
Data Wrangling and Feature Engineering

- Within election/political data we feature engineered a new column to capture percent of the county that voted for Joe Biden.
  - Counties that voted 50% or more for Biden were labeled ‘Democrat’ (0) and counties that voted less than 50% for Biden were labeled Republican (1)
- Left joined county election data to CDC vaccination data, and county demographic/health data
  - Removed records with no county election data
- Imputed mean values into missing features
- Final Data Set: 3,048 instances and 33 features



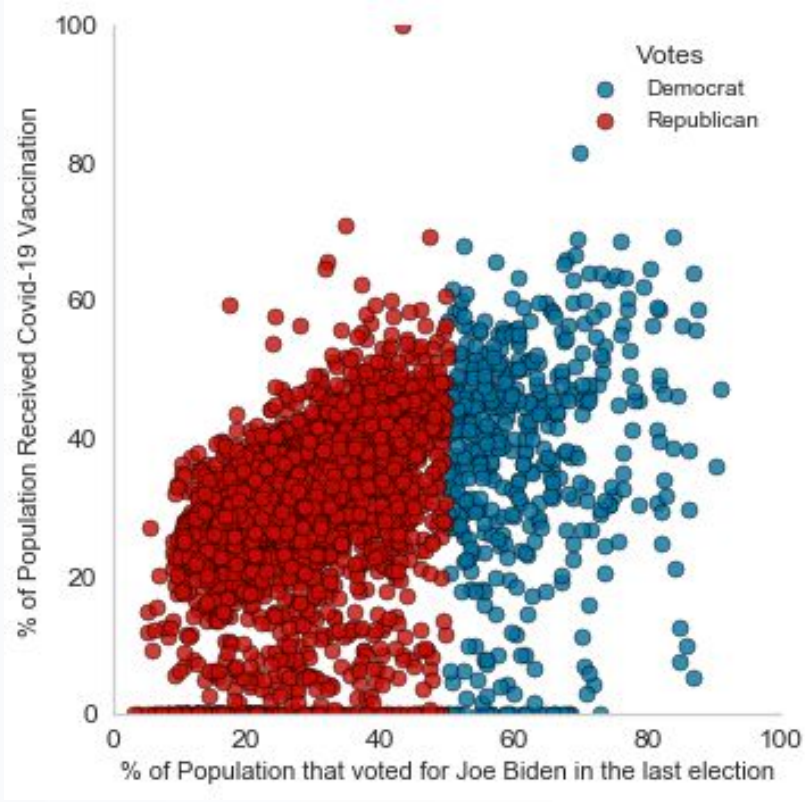
Machine Learning Pipeline

Data Ingestion	Data Wrangling	Feature Engineering	Modeling	Visualizations
CDC Vaccination Data County Health Ranking Data Presidential Election Vote Data	Clean data Consider Incomplete data (impute or remove) Join/merge data on county/state or FIPS	Creating binary target variable from voter data	Apply the Supervised Classification Techniques and Models	EDA Feature Importance Model Evaluation



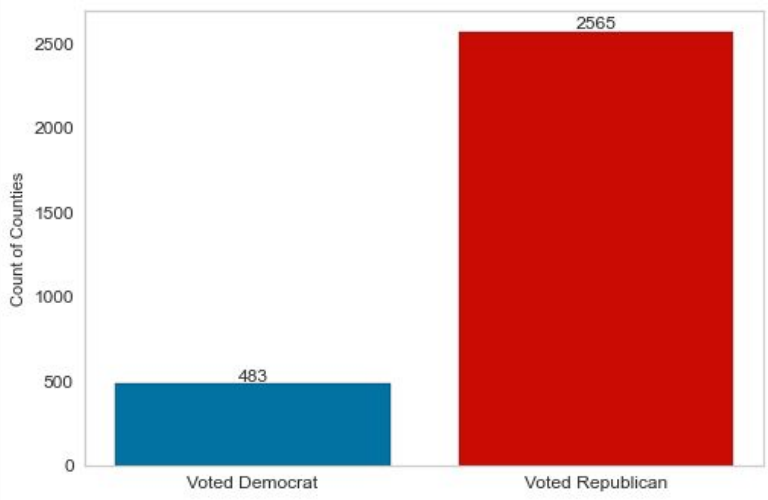
Exploratory Data Analysis

Percent of the County Pop. Vaccinated by Votes



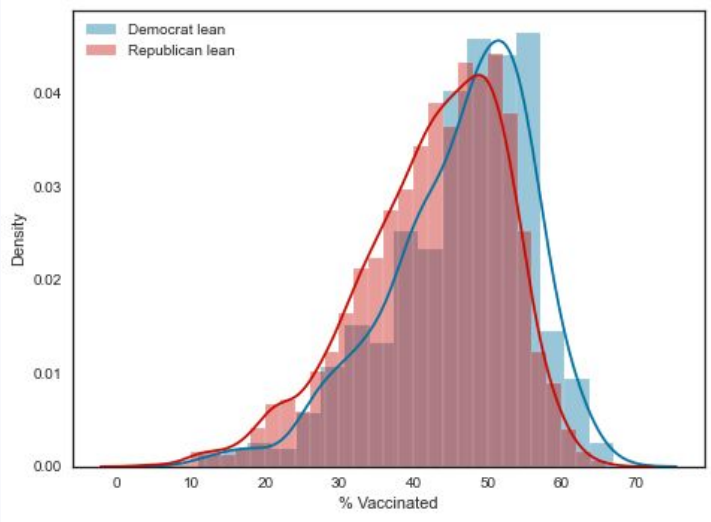
Class Imbalance

- Target:
  - Voted Democrat: 15.8% (483/3,048)
  - Voted Republican: 84.2% (2,565/3,048)
- Yellowbrick Class Balance was used to visualize the imbalance in the classes
  - For the purposes of this project we maintained the imbalance in our training/test splits and did not down sample or weight any of our features

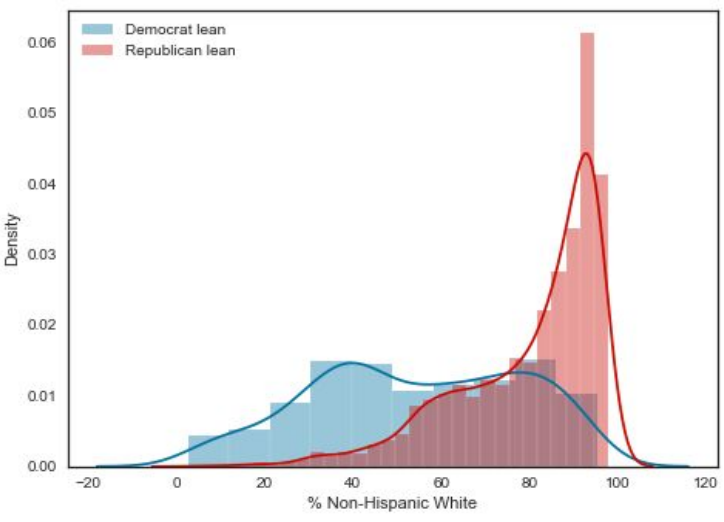


Density Plots/Histograms

Percent of County Pop.
Flu Vaccinated

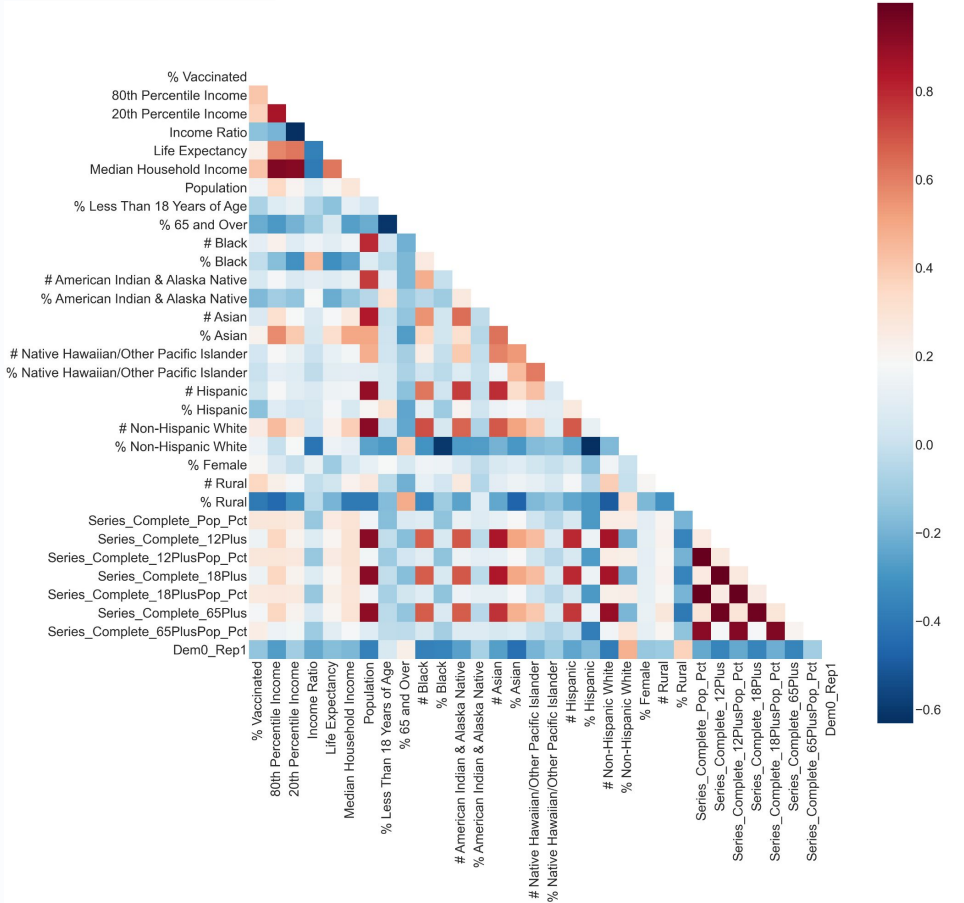


Percent of County Pop.
Non-Hispanic White

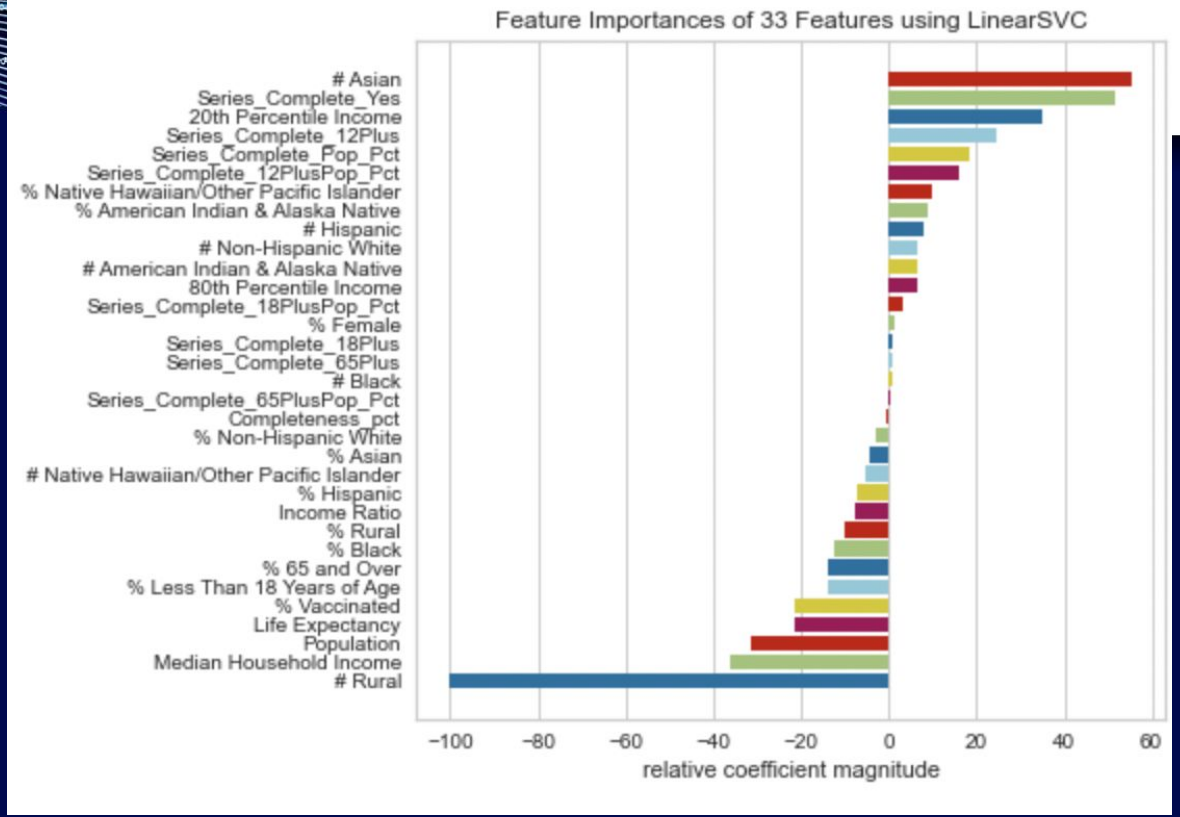


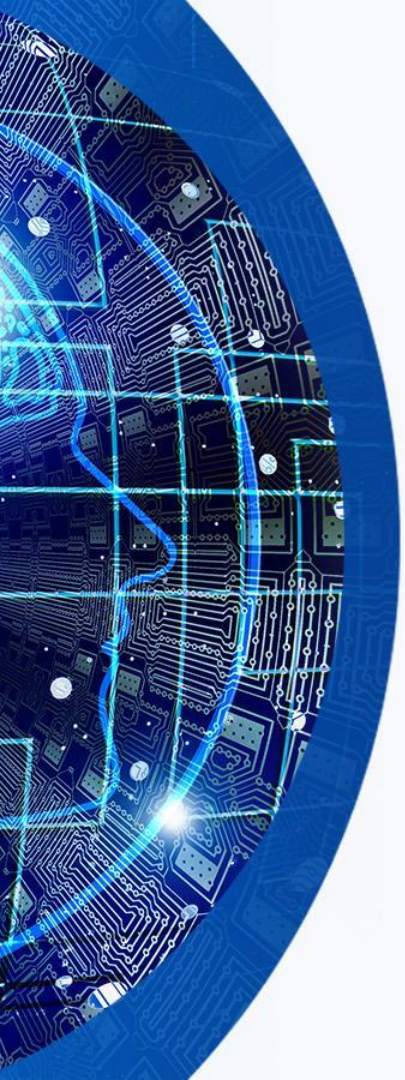
Pearson Correlation Matrix

- Multicollinearity among COVID-19 Vaccination features & some of the demographic features
  - addressed during feature importance
- Correlation between target and individual features not strong



Modeling


Feature Importances of 33 Features using LinearSVC
Asian
Series_Complete_Yes
20th Percentile Income
Series_Complete_12Plus
Series_Complete_Pop_Pct
Series_Complete_12PlusPop_Pct
% Native Hawaiian/Other Pacific Islander
% American Indian & Alaska Native
# Hispanic
# Non-Hispanic White
# American Indian & Alaska Native
80th Percentile Income
Series_Complete_18PlusPop_Pct
% Female
Series_Complete_18Plus
Series_Complete_65Plus
Black
Series_Complete_65PlusPop_Pct
Completeness_pct
% Non-Hispanic White
% Asian
# Native Hawaiian/Other Pacific Islander
% Hispanic
Income Ratio
% Rural
% Black
% 65 and Over
% Less Than 18 Years of Age
% Vaccinated
Life Expectancy
Population
Median Household Income
Rural
−100
−80
−60
−40
−20
0
20
40
60
relative coefficient magnitude



Initial Scores for Models

Model	F1 Score	Recall for Democratic
SVC	0.945	0.455
Linear SVC	0.967	0.798
SGD Classifier	0.953	0.717
K Neighbors Classifier	0.954	0.667
Logistic Regression	0.969	0.788
Logistic Regression CV	0.968	0.778
Bagging Classifier	0.958	0.717
Extra Trees Classifier	0.959	0.717
Random Forest Classifier	0.957	0.747

How to Judge our Models Performance

Why Recall?

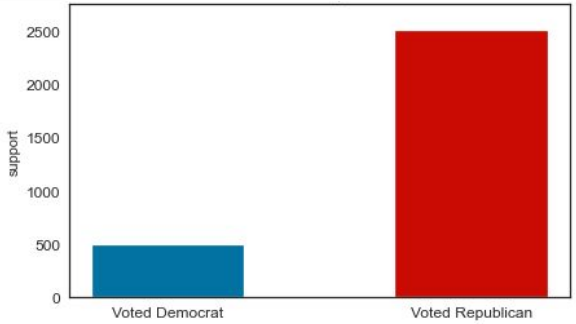
- Recall = True Positives/(False Negatives + True Positives)
  - If the model simply always guesses Republican, then Democratic Recall will be low

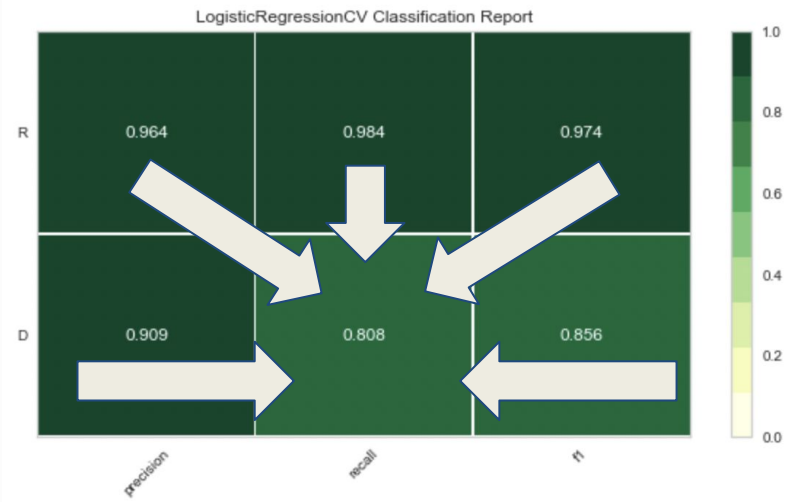
Counties Classified Democrat that were Democrat

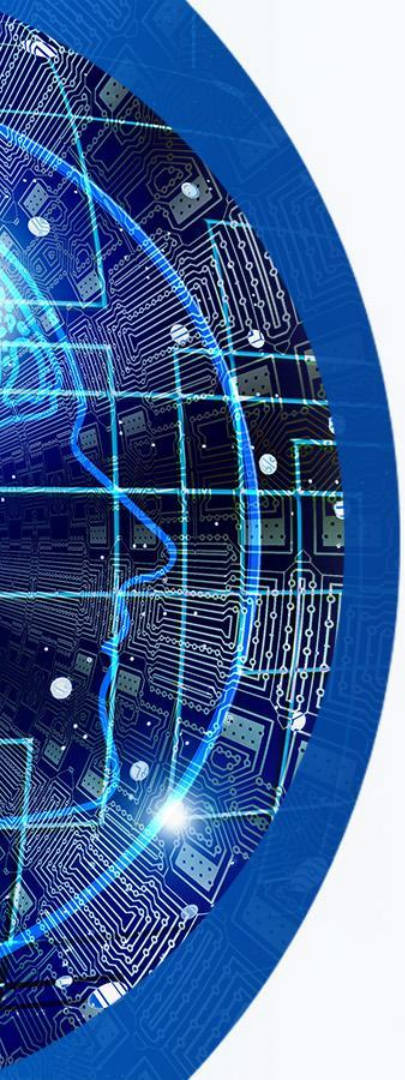
Counties Classified Democrat that were Democrat

+

Counties Called Republican but actually Democrat







Scalers Result Summary

After running our initial 9 models, we decided to experiment with different scaler options to understand their effects on our results

Before choosing which models to tune, we ran all nine models again with different scalers and compared the results to our original version, which simply utilized the Robust Scaler for both integers and floats.

RobustScaler:

Original Scaler Used on all Integers and Floats

MinMaxScaler:

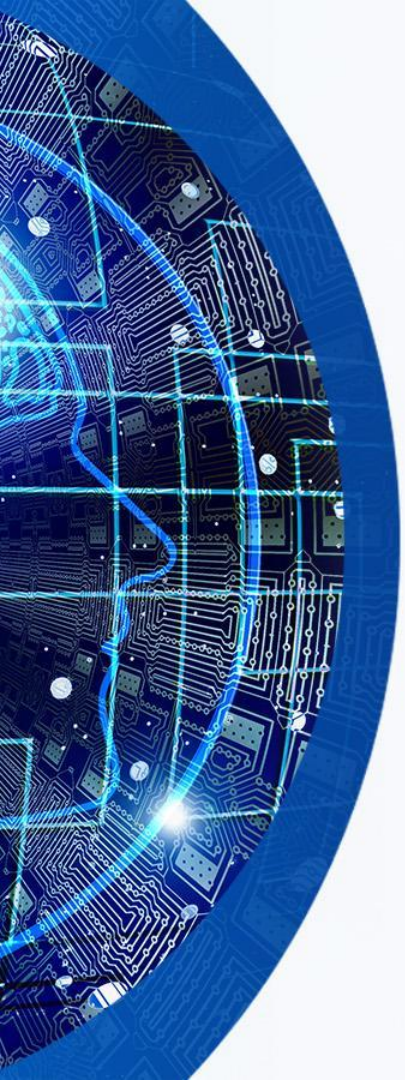
Improved LSVC D Recall: .701 to .785

QuantileTransformer:

Improved LogisticRegressionCV D Recall: .757 to .794

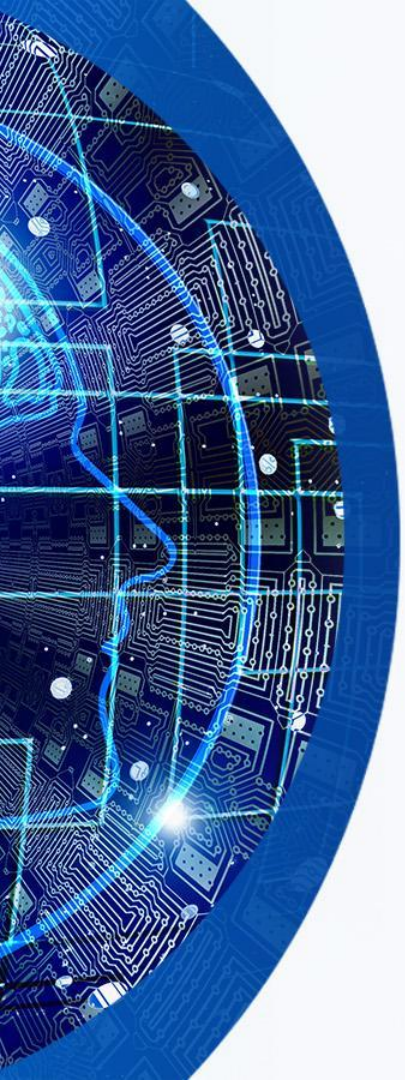
PowerTransformer:

Improved the scores of LSVC, but had no change in D Recall



Logistic Regression (CV=?)

Before Feature Reduction		After Feature Reduction
F1 SCORE LogisticRegressionCV: 0.9690522243713734 0.9475409836065574	(cv=2),	0.973862536302033
F1 SCORE LogisticRegressionCV: 0.9690522243713734 0.9475409836065574	(cv=3),	0.9718719689621726
F1 SCORE LogisticRegressionCV: 0.968054211035818 0.9459016393442623	(cv=4),	0.9718719689621726
F1 SCORE LogisticRegressionCV: 0.968054211035818 0.9459016393442623	(cv=5),	0.973862536302033
F1 SCORE LogisticRegressionCV: 0.9690522243713734 0.9475409836065574	(cv=6),	0.9718719689621726
F1 SCORE LogisticRegressionCV: 0.968054211035818 0.9459016393442623	(cv=7),	0.9718719689621726
F1 SCORE LogisticRegressionCV: 0.9671179883945841 0.9442622950819672	(cv=8),	0.973862536302033
F1 SCORE LogisticRegressionCV: 0.968054211035818 0.9459016393442623	(cv=9),	0.973862536302033
F1 SCORE LogisticRegressionCV: 0.9690522243713734 0.9475409836065574	(cv=10)	0.9718719689621726
F1 SCORE LogisticRegressionCV: 0.9690522243713734 0.9475409836065574	(cv=20)	0.9718719689621726
F1 SCORE LogisticRegressionCV: 0.968054211035818 0.9459016393442623	(cv=50)	0.9718719689621726

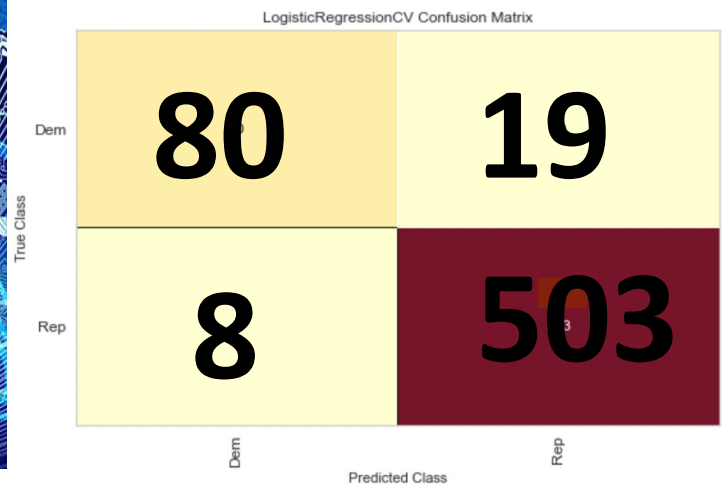


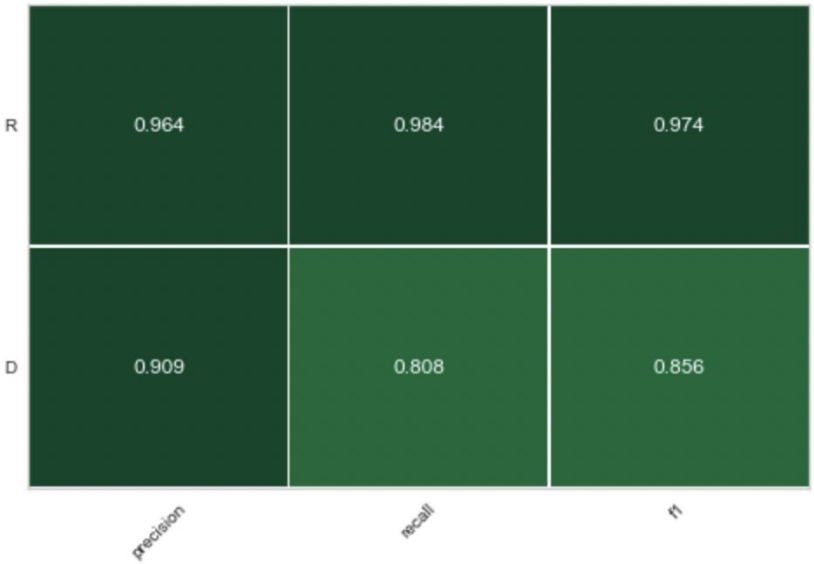
Linear or Radial Basis Function?

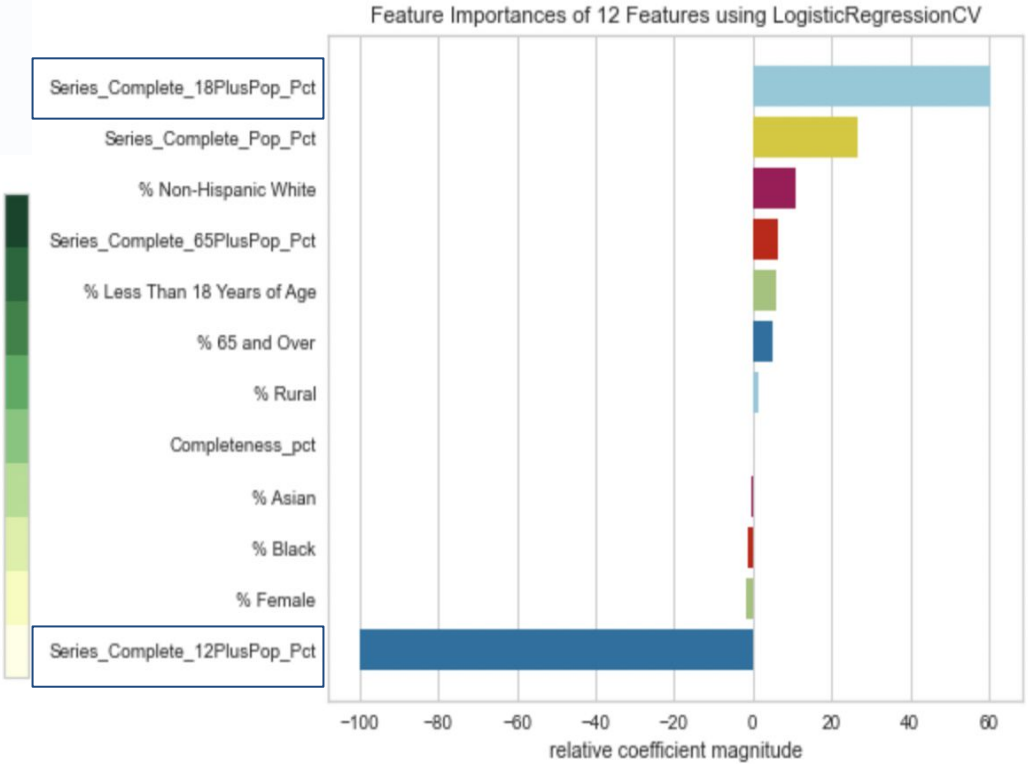
Using Gridsearch delivered that RBF would be better, but upon closer inspection, a linear kernel delivered better Recall for Democratic Counties.

Running Gridsearch after our feature importance now suggested a linear kernel, but offered different parameters for C, which we again ignored because we tuned specifically for Democratic Recall, C=1 was better by .02.

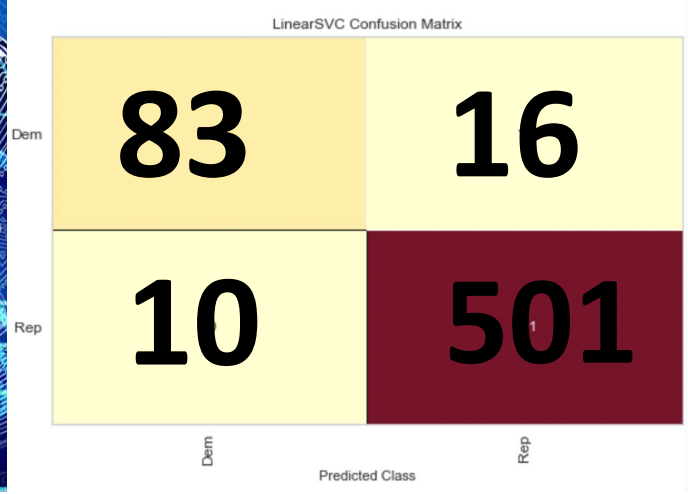
Logistic Regression CV

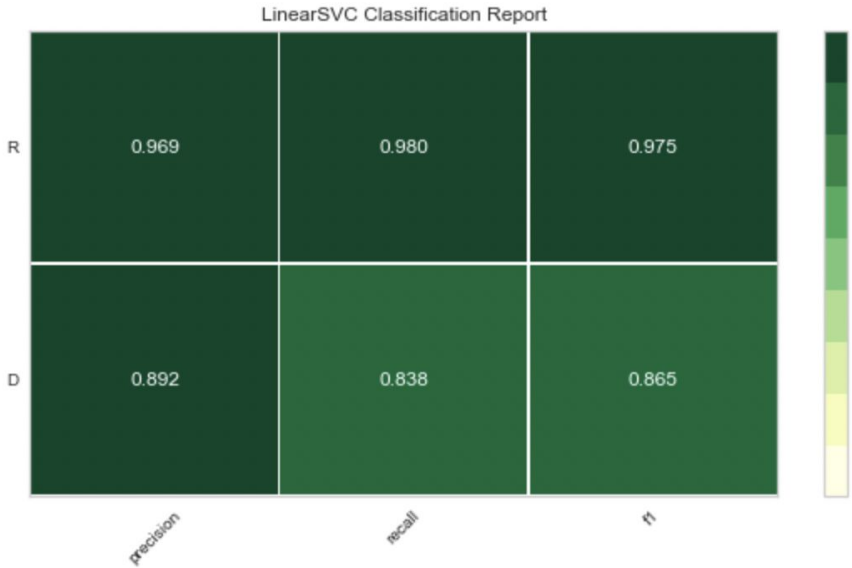


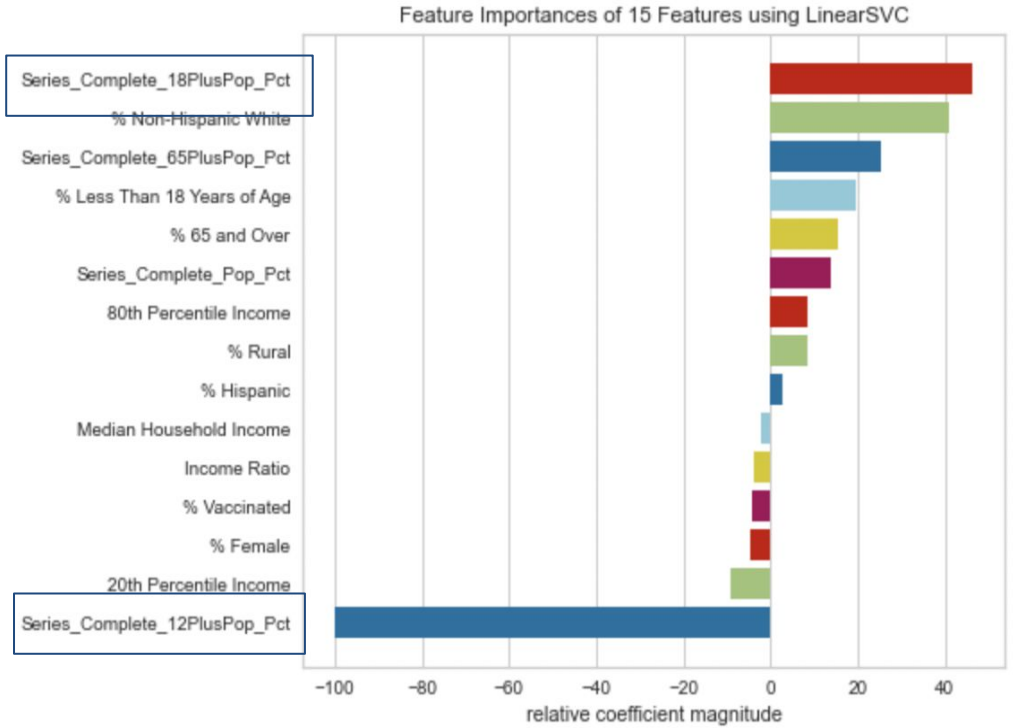


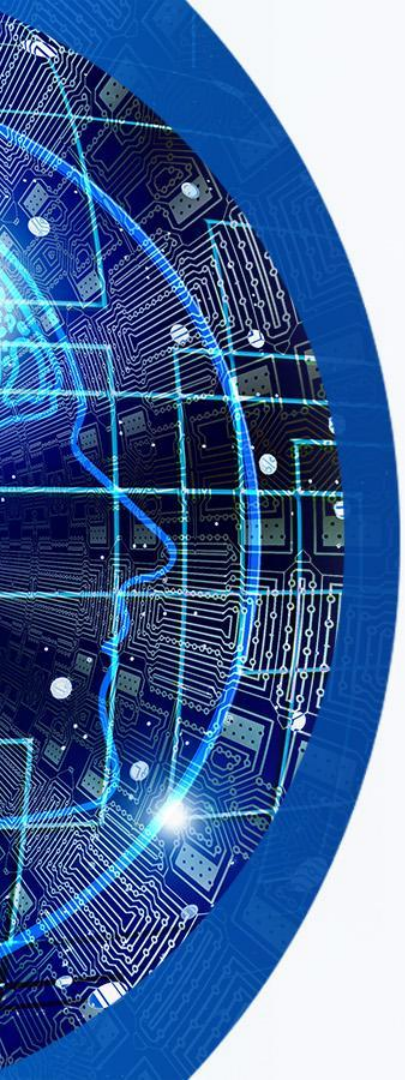


Support Vector Machine Classification









Feature Importance

Step 1: Separate features with both percent and integer collinear variations into separate models, and compare their performance

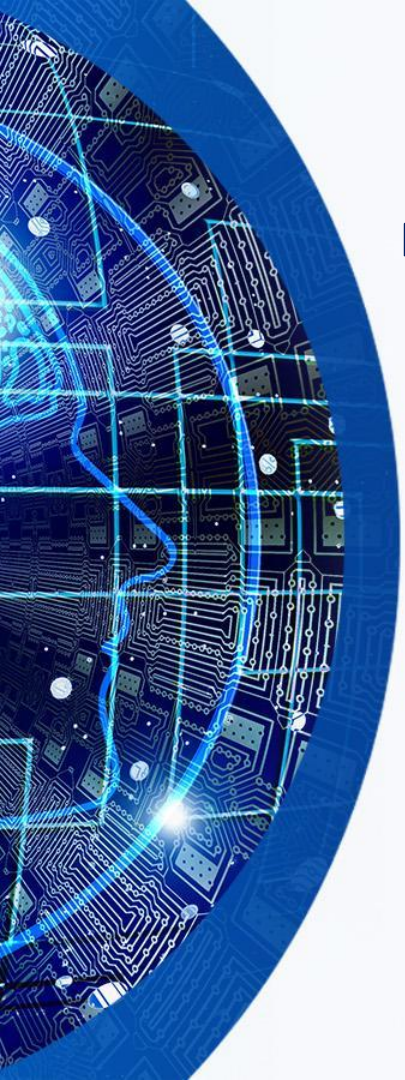
- In both cases, we kept the percent calculations, removing 11 features and improving performance.

Step 2: We broke out our features into three categories by datatype/source, to test how groups of features were performing

- Census data
- Census Racial Data
- Covid-19 Vaccination Data

Step 3: We analyzed every individual features effect on the model, one at a time.

- Guided by YellowBricks Classification Report and Feature Importances tools, we cut features that hurt Democratic Recall, and kept all features that even marginally improved that score



Feature Importance Conclusions

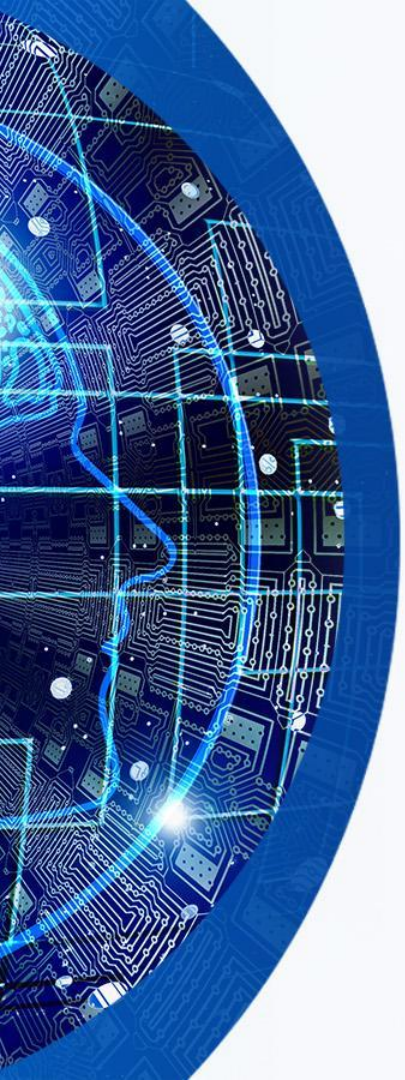
For both models, our two most important features were:

1. Percent of the Counties population over 12 who received the vaccine
2. Percent of the Counties population over 18 who received the vaccine

- While it is true that 12 to 18 year-olds did not vote in 2020, we assume that they also are not the ones making vaccination decisions, rather it is more likely their parents or guardians.

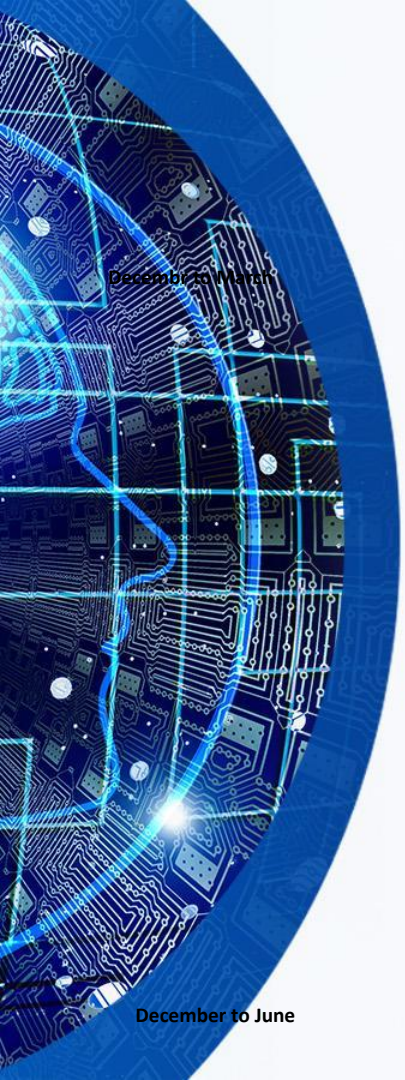
- Because both our models weighted Covid Vaccination Data as more important than other features we included, and our models performed reasonably well on F1 scores including Demcratic Recall, we concluded that it is reasonable to assume that politics is influencing Covid vaccination because we can use County vaccinations to classify 2020 election results.

Conclusion & Next Steps



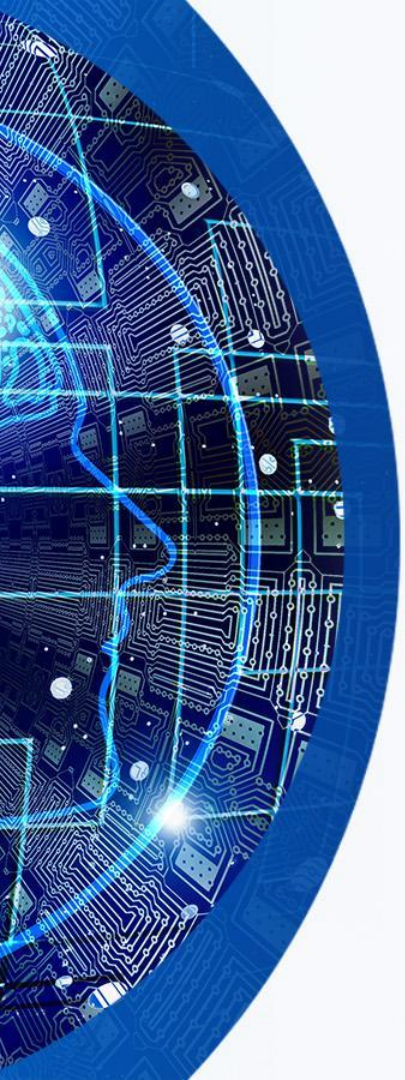
Conclusion

- Using machine learning models, we can successfully classify political affiliation based on Covid-19 vaccination data
- Percent of counties vaccinated are higher in democratic counties than republican
- From the original 33 Features, we were able to identify 15 features as potentially useful when classifying whether a County voted Democrat or Republican in the 2020 election
- Of these 15 features, Covid-19 vaccination data, measured as a percent for 12+ and 18+, was weighted as being the most important when classifying whether a County voted Democrat or Republican in 2020.



Lessons Learned

- Machine learning requires consideration for tools complexity as well as data available
  - Time Series was beyond our scope and will be easier to utilize with a few years of data, rather than a few months
- Utilizing machine learning is way easier than preparing data for your models.
- Visualizations dramatically help in understanding data and relationships among features, and interpreting your data.
- Code Together = Learn Together



Next Steps

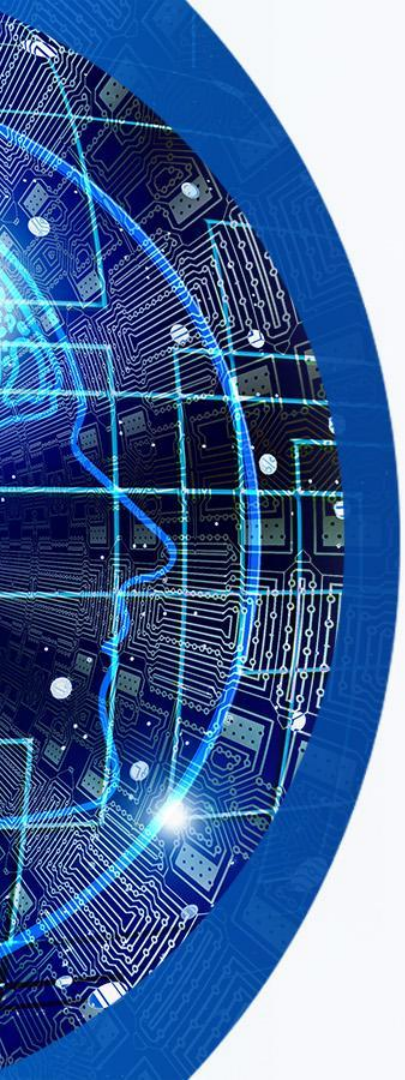
- Automate our wrangling process to provide daily updates to our EDA, modeling, and analysis.
- Utilize 2021 census data with 2021 vaccination data
  - wait and continue analysis with complete data rather than ‘current’ data
- Include infection rate in our analysis
- Compare State and County level herd immunity calculations to help define threshold limits and projections
- Construct different engineered features to include in our analysis
  age groupings 12-18; 18-65, 65+
- Dive deeper into EDA and start to analyze policy decisions and trends as they relate to our analysis on a local level

THANK YOU

REFERENCE SLIDES

Times Series Data

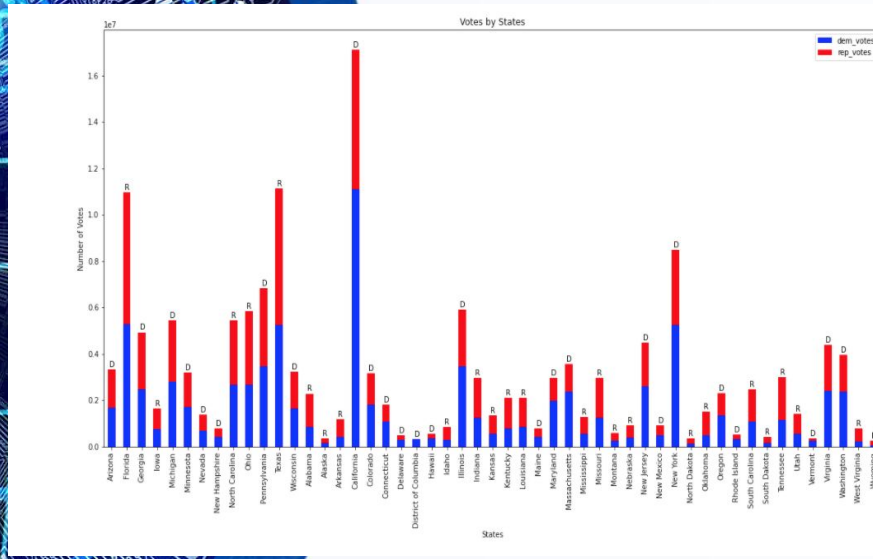
EDA from original hypothesis

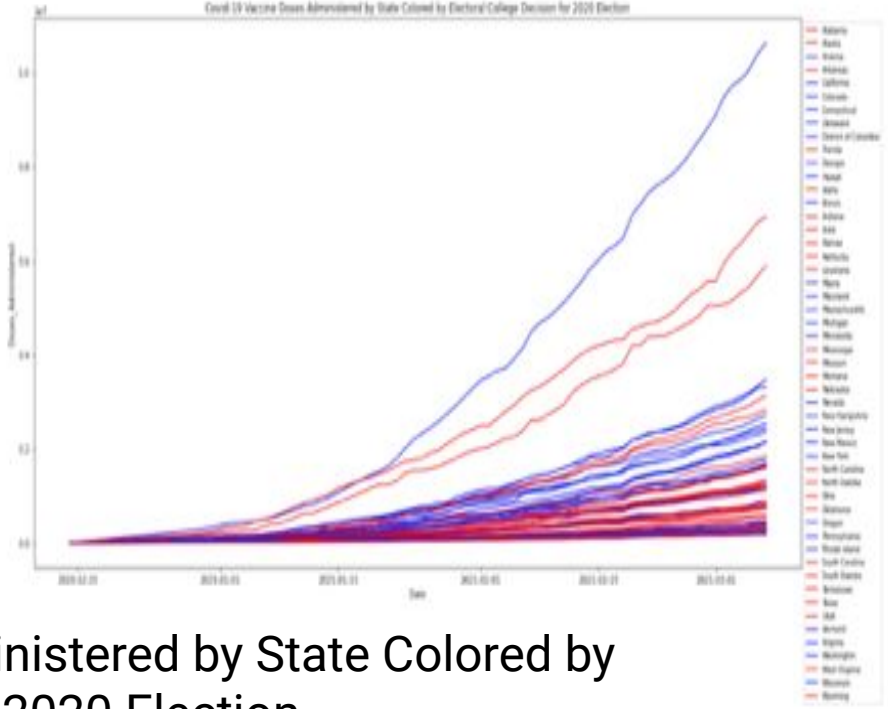


Original Hypothesis

Hypothesis: Classify states into red and blue, relating to 2020 election results, and determine if we can find data points or trends that show correlation between political stance and data relating to Covid-19 vaccination. In other words, can we predict if classifying data into Democrat or Republican will have an impact on the vaccine administered data?

Dataset Joins

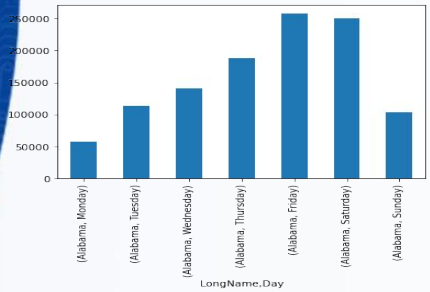


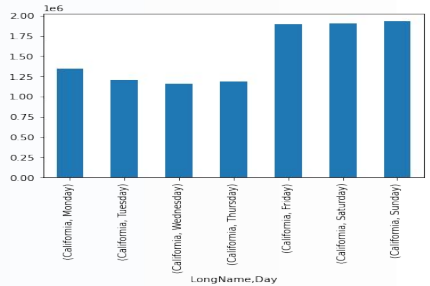


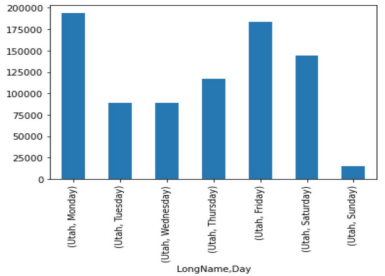
Covid-19 Vaccine Doses Administered by State Colored by Electoral College Decision for 2020 Election

Which day of the week is preferred for vaccinations?

By splitting the distribution, administration, and percent of the population vaccinated from the CDC data we had available to us (December 15 - March 31st) into weekdays and then analyze each State individually



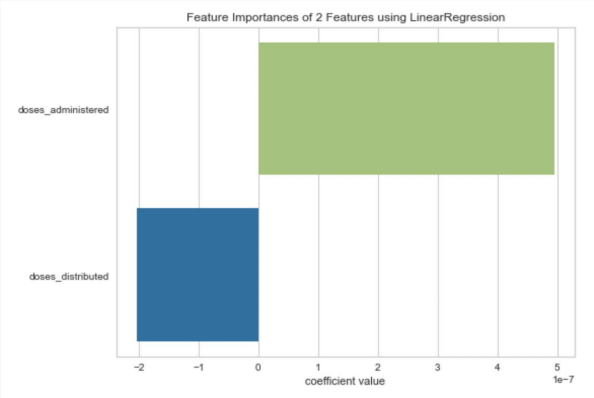


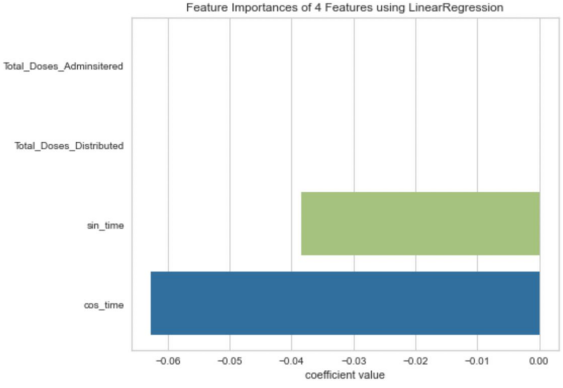


Value of Feature Engineered Day of the Week Vaccine Administration

Linear Regression models for each of the 50 States: our models performed considerably better when we added a sin/cosine cyclical wave function to account for day of the week, such that every day was the same distance from the day before and next day (rather than 0-6), and every weekday had the same value

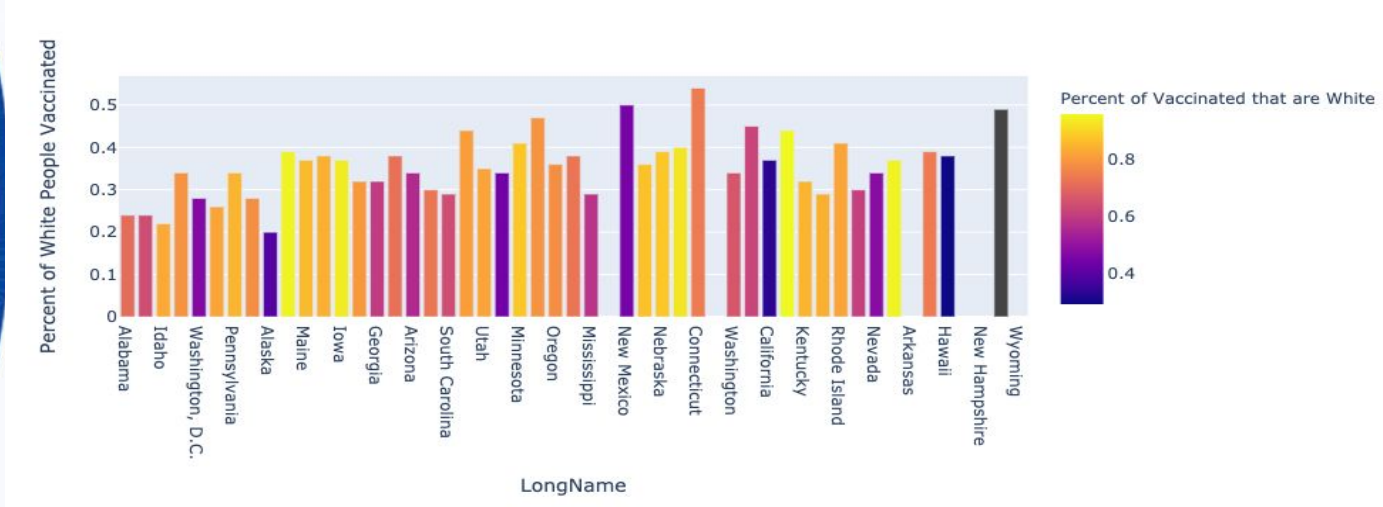
→ highest weighted feature





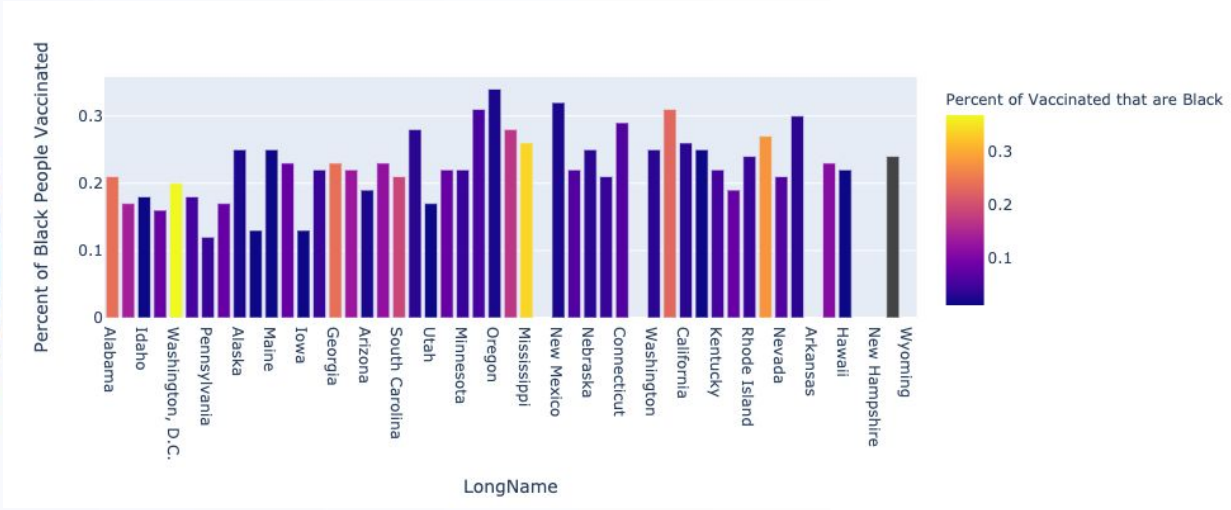
Doses Distributed vs Doses Administered Over Time in Different Racial Populations

Percent of White Population Covid Vaccinated



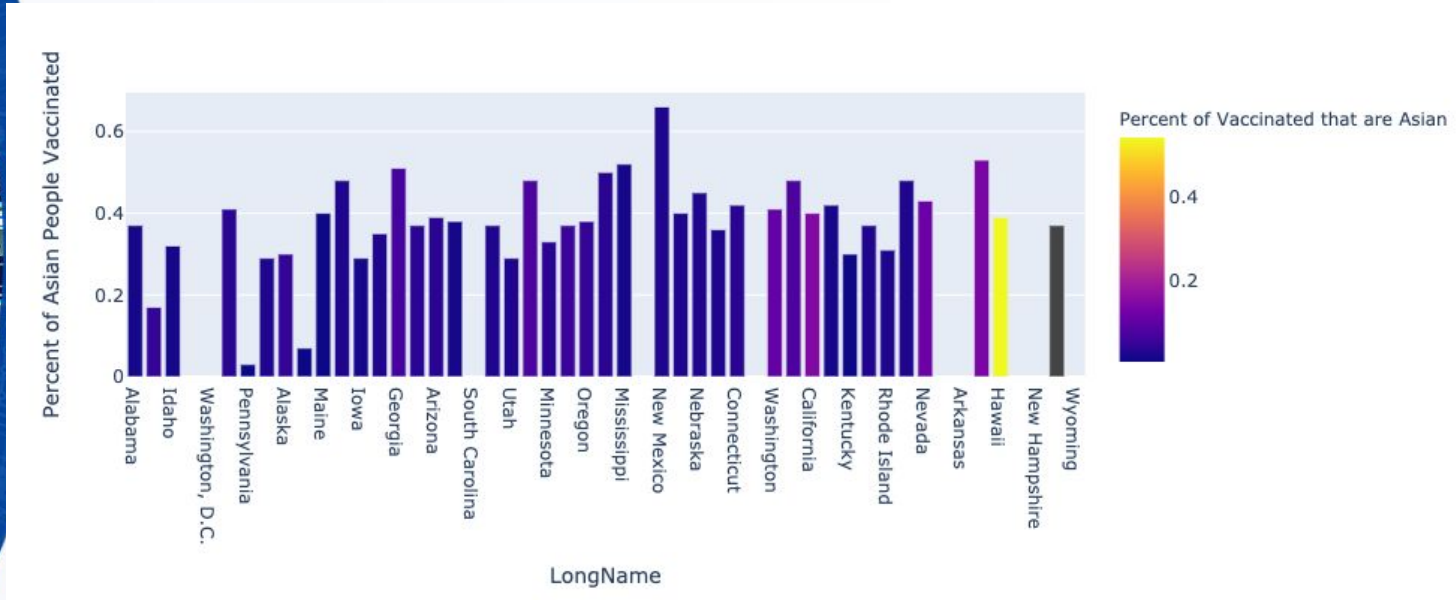
Percent of Vaccinated that are White was highest in South Dakota, and West Virginia however the states with the largest percent of White people vaccinated were Connecticut and New Mexico.

Percent of Black Population Covid Vaccinated



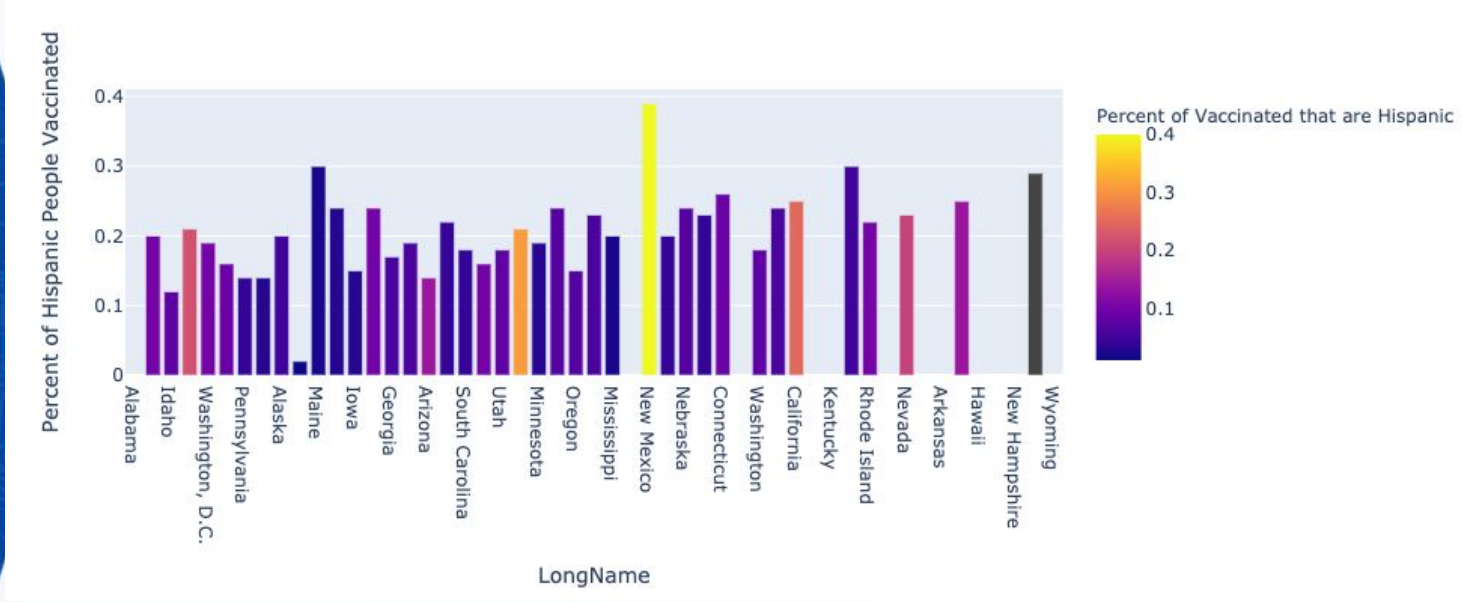
Percent of Vaccinated that are Black was highest in Washington, D.C. and Mississippi however the states with the largest Percent of Black people vaccinated were Oregon and New Mexico.

Percent of Asian Population Covid Vaccinated



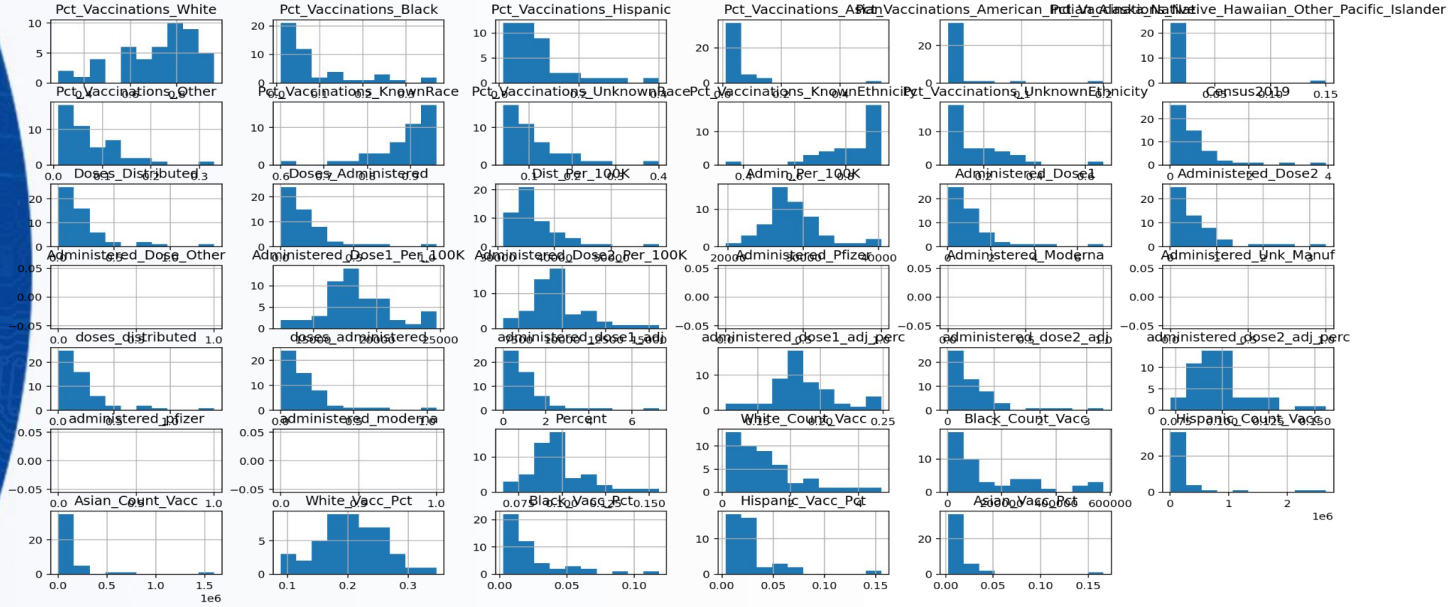
Percent of vaccinated that are Asian was highest in Hawaii, and California however the states with the largest percent of Asian people vaccinated were New Mexico and New York.

Percent of Hispanic Population Covid Vaccinated



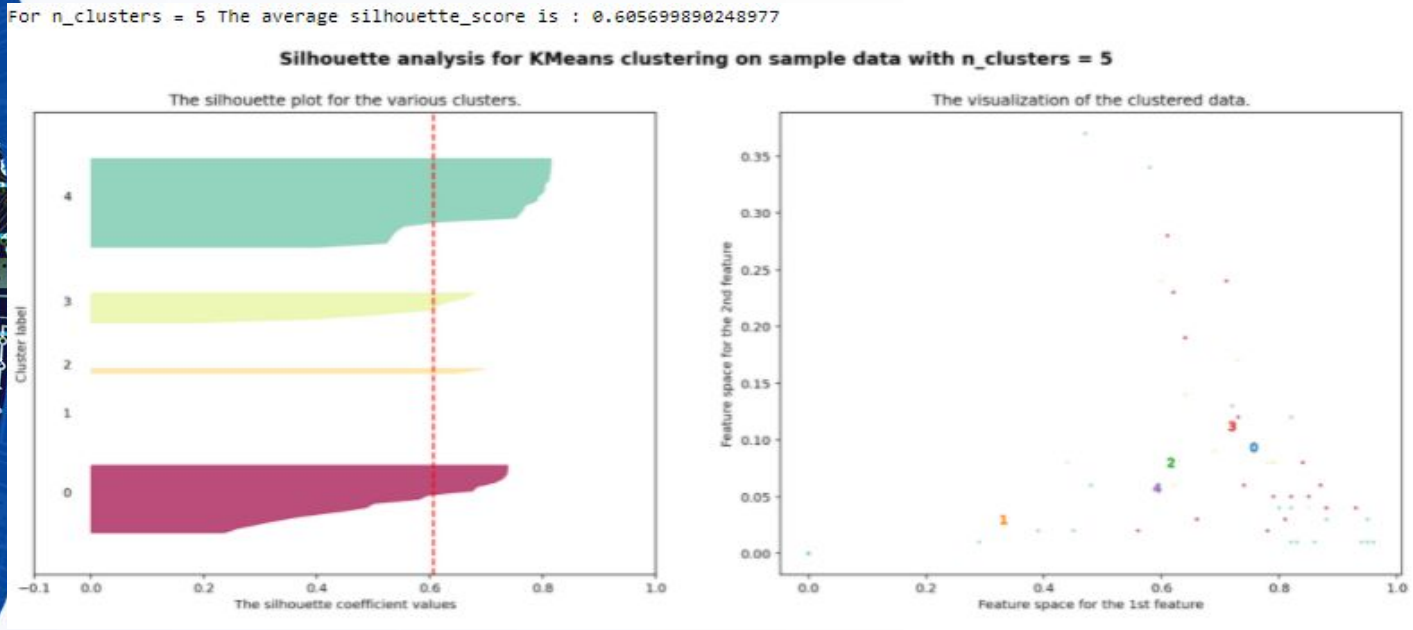
Percent of vaccinated that are Hispanic was highest in New Mexico, and Texas however the states with the largest percent of Hispanic people vaccinated were New Mexico, Maine, and Missouri.

Distribution of Different Features by Count of States



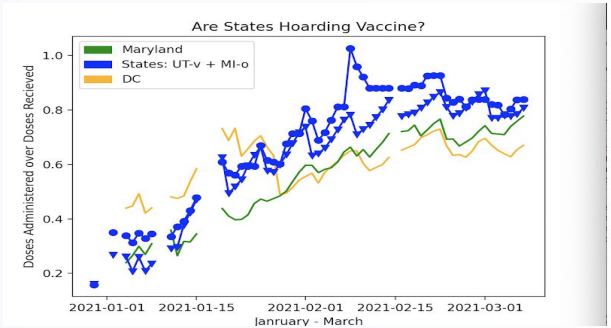
Percent of vaccinations among all non-White race/ethnic groups tends to skew to the left indicating that based on the snapshot data most non-White race/ethnic groups were not being vaccinated at the same rate as White persons.

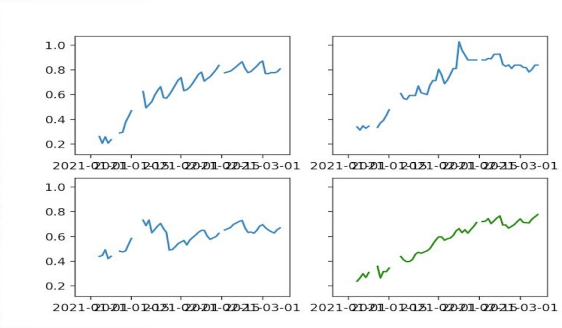
Silhouette Analysis for KMeans on Sample Data with n_clusters = 5

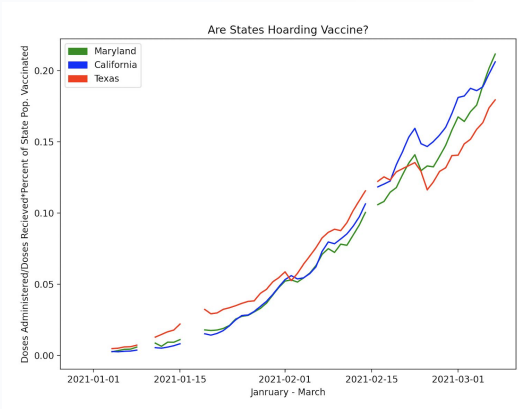


Initial K-means clustering of state level output. The n_cluster of 5 provided the best silhouette score, clusters with the most separation. Cluster 2 appeared to only have one state clustering alone. With only 50 states in the data set, clustering was limited.

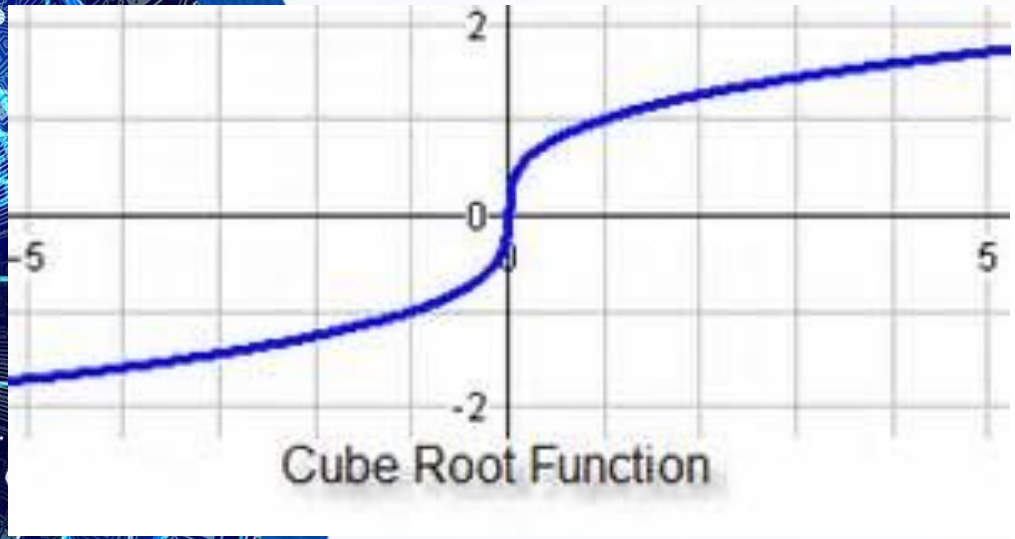
Debunking Vaccine Hoarding



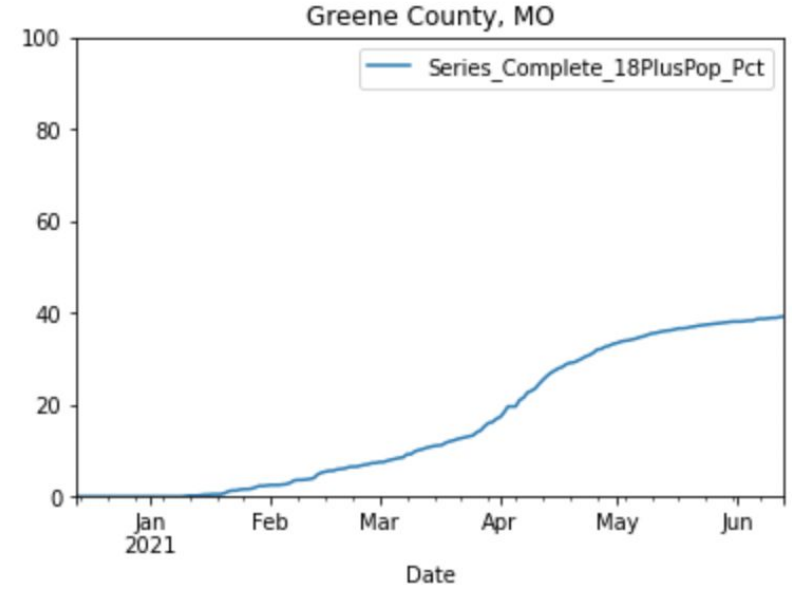




Times Series Consideration

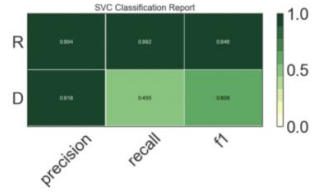


Prediction

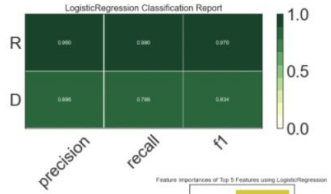


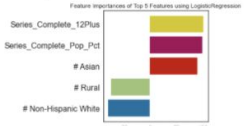
Reality

F1 SCORE SVC: 0.9458955223880596
0.9491803278688852

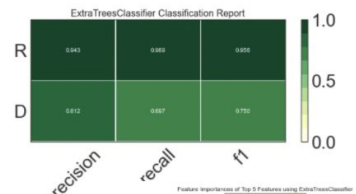


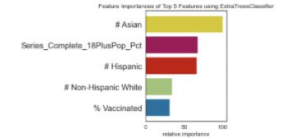
F1 SCORE LogisticRegression: 0.9699303194578897
0.9491803278688524



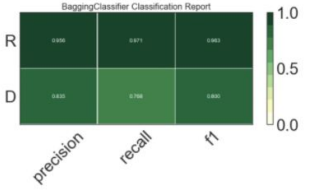


F1 SCORE ExtraTreesClassifier: 0.9651162790697674
0.9409836065573777

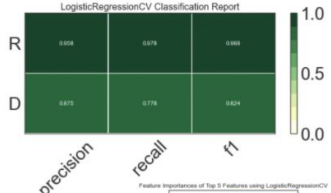


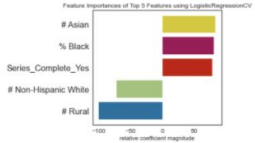


F1 SCORE BaggingClassifier: 0.9552529182879377
0.9245901639344263

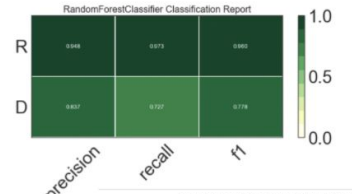


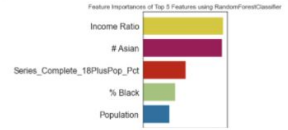
F1 SCORE LogisticRegressionCV: 0.9680542211035818
0.9459016393442623



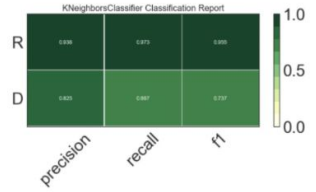


F1 SCORE RandomForestClassifier: 0.9582118561710399
0.9295081967213115

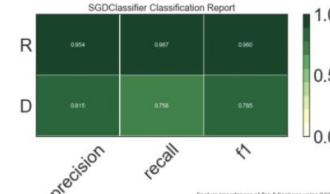


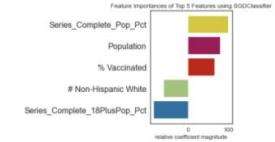


F1 SCORE KNeighborsClassifier: 0.9548511047070125
0.9229508196721311

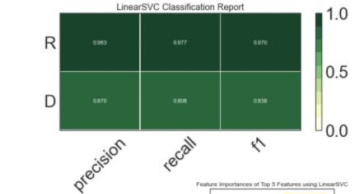


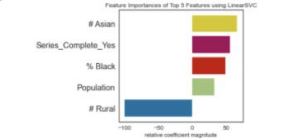
F1 SCORE SGDClassifier: 0.953125
0.921311475409836

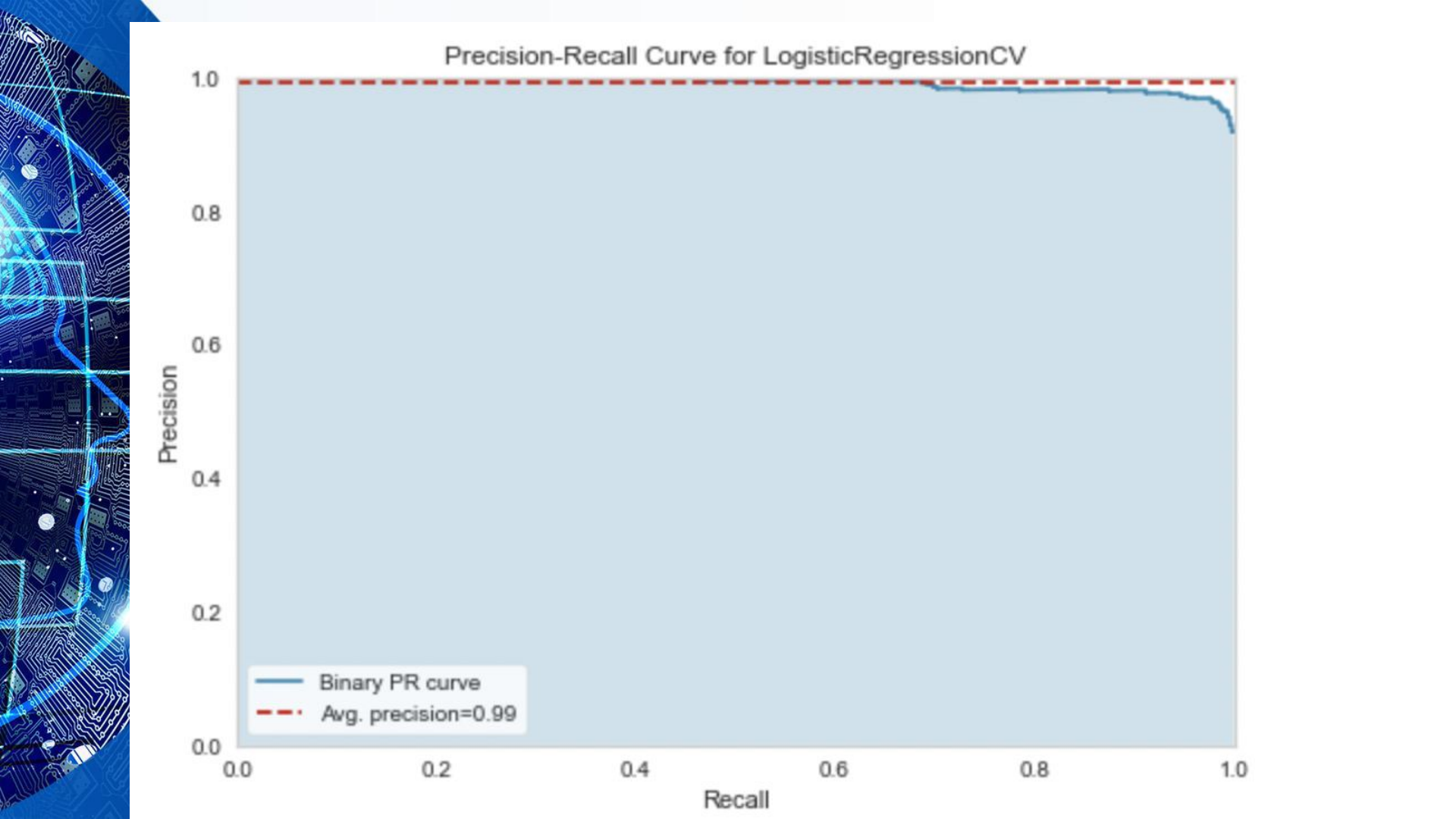




F1 SCORE LinearSVC: 0.9682386910490857
0.9459016393442623






Precision-Recall Curve for LogisticRegressionCV
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Binary PR curve
Avg. precision=0.99