Final Project Report COVID19

Authors: Paul Fentress, Nicole Ni, Jonathan Rubalcava

Address Design Doc Feedback

Feedback on the hypothesis:

- Our previous hypothesis lacks some depth in investigating the data. We only included features that could be done with basic EDA and don't require any extensive analysis. In our improved hypothesis, we decided to investigate more features that involve time periods and geographical locations.
- For our previous experimental design, our sample size was too small (only included 50 states), but now we changed from state-level to county-level which increased the sample size to 3141 samples.

Problem

Hypothesis:

We suspect that Democratic voting counties have better COVID prevention behaviors than Republican voting counties. Based on this we wanted to see if we can accurately predict a county's 2020 presidential vote based on features such as vaccination rates and mask use (COVID prevention behaviors)?

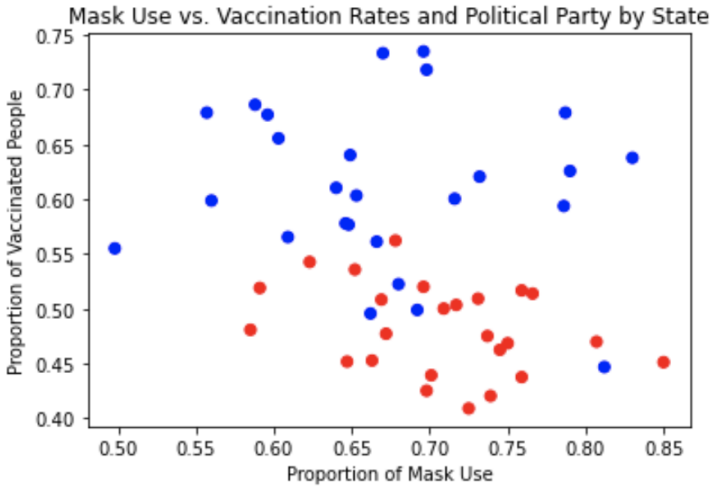
We can create models that take in features given by existing datasets to make predictions about the county's voting preference. If we import more datasets (getting more features) we will be able to generate a more accurate result. In this part of the project we have imported datasets that have more detailed information about vaccination rate per county, and the voting result in 2020 for each county. After training the model with given features we are able to make predictions for party preference of the counties, and then compare with the true voting result.

We believe that COVID prevention behavior and voting preferences are correlated, so the hypothesis can be confirmed with a model that includes features accounting for the COVID prevention behaviors (mask-use, vaccination rate, etc).

Answer

Since we believe that COVID prevention behavior and voting preferences are correlated, then we can use a binary classifier model to predict the county's vote. We can classify a county as Democratic if the county has features such as high vaccination rate, frequency mask use, faster response to vaccine recommendations, etc, and a county Republican if the county is slow in getting vaccinated and not wearing masks.

Below is a graph that shows each state as a data point, how they voted (color), their proportion of mask use (ALWAYS+FREQUENTLY) on the x, and the proportion of vaccinated people by state on the y.



This plot shows that this problem is clearly linearly separable which leads to our Model selection process.

Modeling

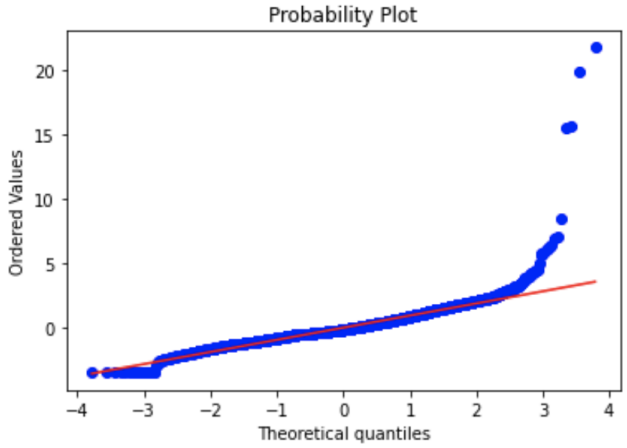
Model Selection

We iterated through 9 models before landing our final model, which was Logistic Regression with standardization for data pre-processing. We used KFold cross-validation with k=5 to assess initial model performance for each. We tried using SVM, Logistic regression, Logistic Regression with Normalization, Logistic Regression with Standardization, Random Forest, K-Nearest Neighbors, and Naive Bayes. Below are the initial models and their corresponding accuracy.

:

	Model Accuracy
Support Vector Machine	0.847857
Logistic Regression + Standardization	0.847500
Logistic Regression Model	0.845000
Random Forrest	0.840714
Logistic Regression + Normalization	0.822500
K-Nearest Neighbors	0.822500
Naive Bayes	0.786071

Our initial model had 4 inputs ["NEVER","ALWAYS", "CAPITA", "Percent_increase"] and a single binary output which was a classification for Republican or Democrat presidential vote prediction. We used Standardization because our features followed a normal distribution which was verified using a quantile quantile plot:



A QQ plot that has the points scattered straight along the line means that the sample data does follow the distribution you have specified (In our case we wanted to see if our data follows a normal distribution.). If your data diverges from the straight line that means your data does not follow the distribution you have specified. A QQ plot compares the quantiles of a distribution to our sample data and sees how correlated they are for all data points. If our sample data is highly correlated with the target distribution, the data will fit the target distribution line, otherwise, our sample data will diverge from the target line. As we can see our features data follows the line almost perfectly all the way up until the right tail, which is a visual confirmation that the majority of our features follow a normal distribution. When the data follows a normal distribution it is ideal to standardize the data.

Logistic Regression with Standardization received the 2nd highest initial accuracy with 84.79%. SVM has the highest initial accuracy; however, the difference was only .0000357 so it was very

small. Due to the fact we haven't learned SVM in class and logistic regression only underperformed by .0000357 accuracy, we chose Logistic Regression + Standardization.

We also tried Grid search CV to find optimal solver and penalty parameters for our Logistic Regression model. For solvers we tried: "newton-cg", "lbfgs", "liblinear", "sag", "saga". We found that all of the solvers performed equally well, and therefore the default was performing as the other options, so we did not get a boost in accuracy for changing our solver. We use "saga" as an arbitrary choice, because they all performed equally well. We found the best penalty was "l2", however, this was also the default, so we found our solvers and penalty parameters were not having an impact on the initial model performance.

Inputs/ Features & Explanations

inputs of our model are the features based on the information given by dataframes.

- **NEVER**: percent of people in the county that never uses masks
- **ALWAYS**: percent of people in the county that always uses masks
    - Above 2 are mask use information from given dataframes, indication of COVID-prevention behaviors.
- **CAPITA**: cases per capita per county on 9/12/21
- **Percent_increase**: cases increase from 8/12/20(when vaccine is generally available) to 1/10/21 in terms of percentage
    - The larger the increase, the more proactive the people are in terms of getting the vaccines (people who aren't anti-vax should already got theirs by Jan. 2021)
- **Series_Complete_Pop_Pct**: percent of people in each county who has fully vaccinated on 12/6/21
    - Another COVID-prevention behavior
- **Vax_percent_diff**: vaccination rate percent difference in each county from 4/11/21(according to the graph that's when the two parties have split in vaccination rate) to 12/6/21
- **Vax_percent_diff_window_A**: vaccination rate percent different in each county from 4/1/21 to 5/11/21 (according to the graph the first time democrat and Republican diverge)
- **Vax_percent_diff_window_B**: vaccination rate percent different in each county from 4/1/21 to 5/11/21 (according to the graph the second time democrat and Republican diverge)
    - Add those 3 feature to try to better classify/differentiate counties with different party affiliations according to current data

*The features below are for the county's one neighbor (found based on the closest latitude), and we will exclude them in our final model.

- **neighbor_NEVER**:
- **neighbor_ALWAYS**:
- **neighbor_CAPITA**:
- **neighbor_Percent_increase**:

- **neighbor_Series_Complete_Pop_Pct**:
- **neighbor_Vax_percent_diff**:
- **neighbor_Vax_percent_diff_window_A**:
- **neighbor_Vax_percent_diff_window_B**:

Outputs:

The model outputs a prediction of “DEMOCRAT” or “REPUBLICAN” for the county.

Model Evaluation and Improvements

After expanding the dataset from states to counties, we still have the problem that the features of the model are too surface-level. Before adding new features we only have COVID information at a certain point in time but not investigating the change over time. Not including a time-based feature might result in neglecting trends in the observations, which is an essential part of the observed data. We imported external datasets and were able to find out the vaccination rate increase percent difference from a time window. The vaccination rate increase over time tells information about how responsive and proactive people are with the COVID prevention policies such as getting vaccinated. If more people are getting vaccinated over time, it suggests that people in that county are generally more likely to accept the vaccine, and vice versa. The time-based features add more depth and dimensions to the model.

Adding Feature #1:

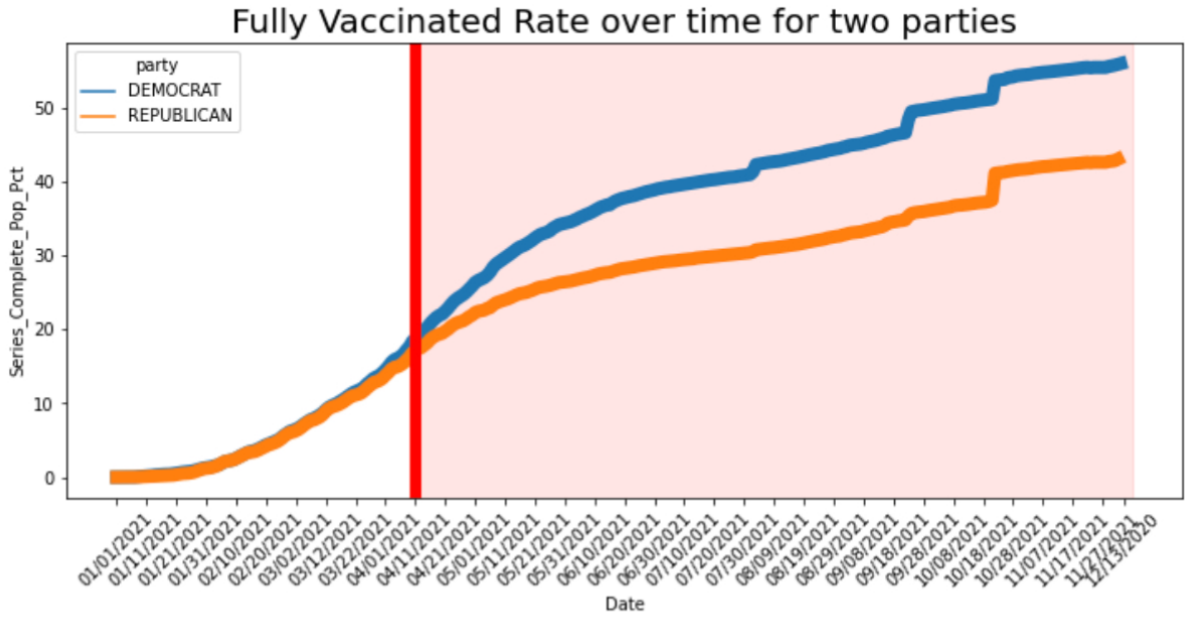
Improving our current model, we first introduced a data set that contains vaccination rates per county. We chose a recent date (12/06/2021) to focus on because we figured that if someone is not vaccinated at this point after the vaccine has been made available for this long, it must be that they are either anti vaccination or not vaccinated for religious reasons, which can be a distinct feature when categorizing if a county is Democratic or Republican. At this point, after running our Logistic Model on a total of 5 features, we result in an 86.71% accuracy.

Adding Feature #2:

We then implemented 2 more features to try and increase the accuracy of our model, which was Series_Complete_Pop_Pct and Vax_percent_diff(which is explained above). The reason we included Series_Complete_Pop_Pct as a feature was because we wanted to choose a starting date where there is a clear distinction of fully vaccinated parties. The following image is a visual representation of the vaccination percentage by county of parties that have been fully vaccinated across time (NOTE: we created bins for the dates on the x-axis to make the dates more visible and represent the vaccination rates through time):

Plot 1:

Title: fully vaccinated rate overtime for two parties
X-axis: Dates ranging from Jan. 2021 to Dec 2021
Y-axis: average percentage of people who are fully vaccinated among two parties
Key: blue for democrat, orange for Republican
Takeaway: when the vaccine is first available we can see that two parties roughly the same fully vaccinated rate, but at april there is a clear diverge between the two parties vaccination behavior, so we want to investigate what's happening there and maybe consider adding it as a feature when try to classify the county's party preference.
Connection: This graph motivated us to add a new feature that accounts for the vaccine rate increase from the splitting point to current date. Since there is a clear difference in percent increase between two parties, we think that it might contribute to training a better model.



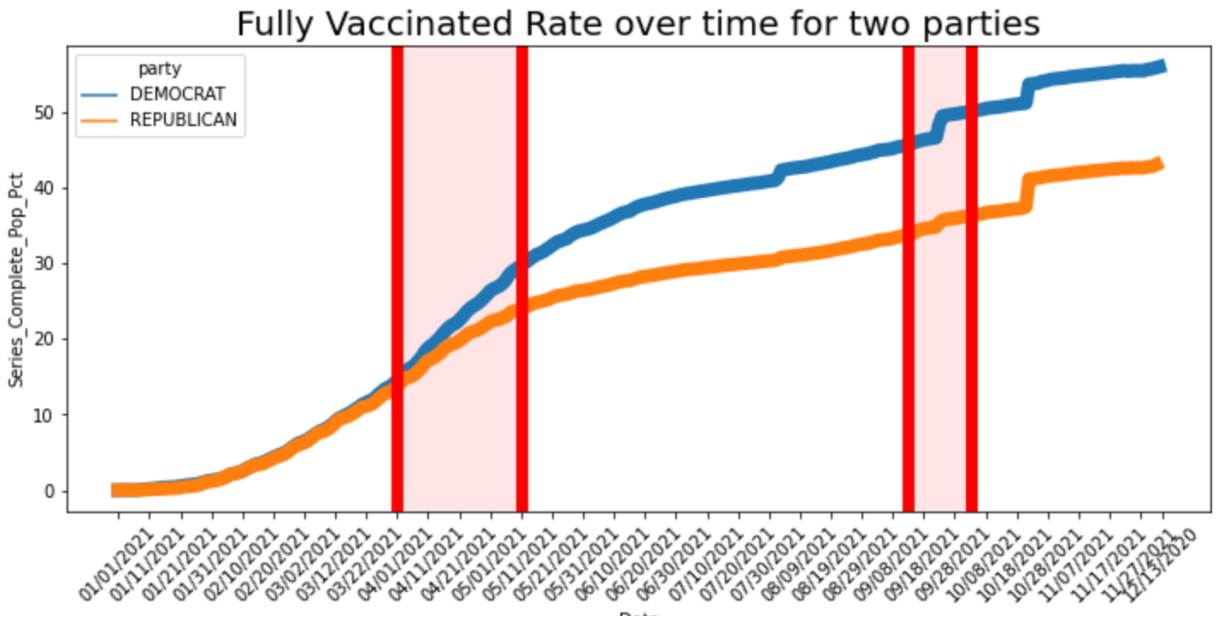
As we can see above, the split between democrat and Republican vaccination rates begins to part at the red line at about 4/11/2021. We then calculated the Series_Complete_Pop_Pct feature by taking the vaccination percentage beginning on 4/11/2021 all the way up to 12/06/2021 so that we can have a better classification of categorizing a county as Democrat or Republican. After running the logistic regression with this new included feature, we were able to achieve a 87.36% accuracy.

Adding Feature # 3 and #4:

As we added 2 more features, we were able to increase our accuracy to 87.39% which is only a 0.02% increase. We included these 2 features, Vax_percent_diff_window_A (4/1/21 - 5/11/21) and Vax_percent_diff_window_B (9/8/21 - 9/28/21), because we wanted to include

specific dates where parties of counties diverged the most. The following visualization highlights 2 areas where the counties diverge the most:

Title: fully vaccinated rate overtime for two parties
X-axis: Dates ranging from Jan. 2021 to Dec 2021
Y-axis: average percentage of people who are fully vaccinated among two parties
Key: blue for democrat, orange for Republican
Takeaway: the highlighted areas are the areas of interest where we noticed that both parties diverge the most compared to the rest of the graph
Connection: This graph motivated us to add 2 new features that can allow us to better distinguish parties. When determining whether these training a better model.

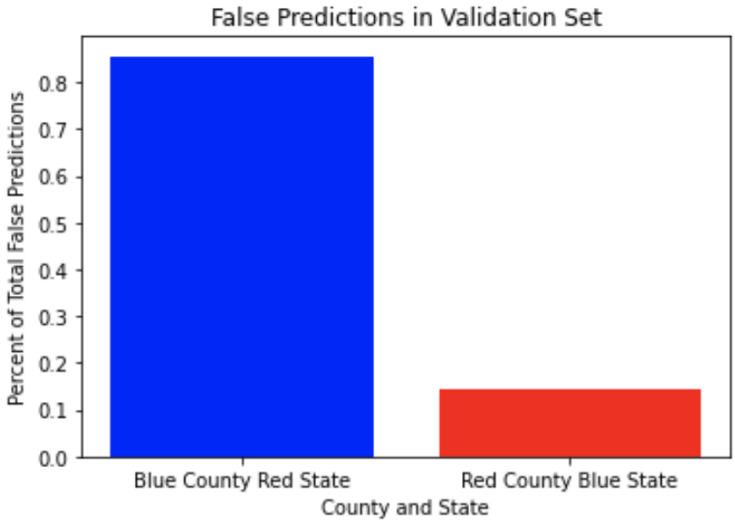


Adding Feature # 5:

We wanted to explore the relationship between neighbor counties, and see if this could improve our model. In order to do this we sorted by latitude, then shifted the column up 1 row. And added the table that has been shifted to the original table with all the features. This makes it so each row now has all the features of its neighbor also. As a result, adding this feature decreases our accuracy to 82.75%. To see why this was the reason we created a visualization that determined the amount of false predictions when classifying as Democratic in a Republican state or Republican in a Democratic state.

Title: False predictions in validation set
X-axis: county and state with its counter political preference
Y-axis: percentage of total false predictions
Key: blue for democrat, orange for Republican

Takeaway: We can see that the majority (85%) of false predictions were blue counties in red states. Next came red counties in blue states (15%) of false predictions. There were no blue-county in blue-state, or red-county in red-state false predictions.
Connection: Connecting this back to why the accuracy went down, we believe the accuracy went down because there can be blue counties nestled inside a state which is predominantly red. In cases like these, according to our hypothesis that mask use and vaccine rates are related to which party the county voted for, then the mask use of a blue neighbor will be high, while its neighbors might be low. This could lead to less accurate predictions. Lets look more into this and see which counties were predicted false.



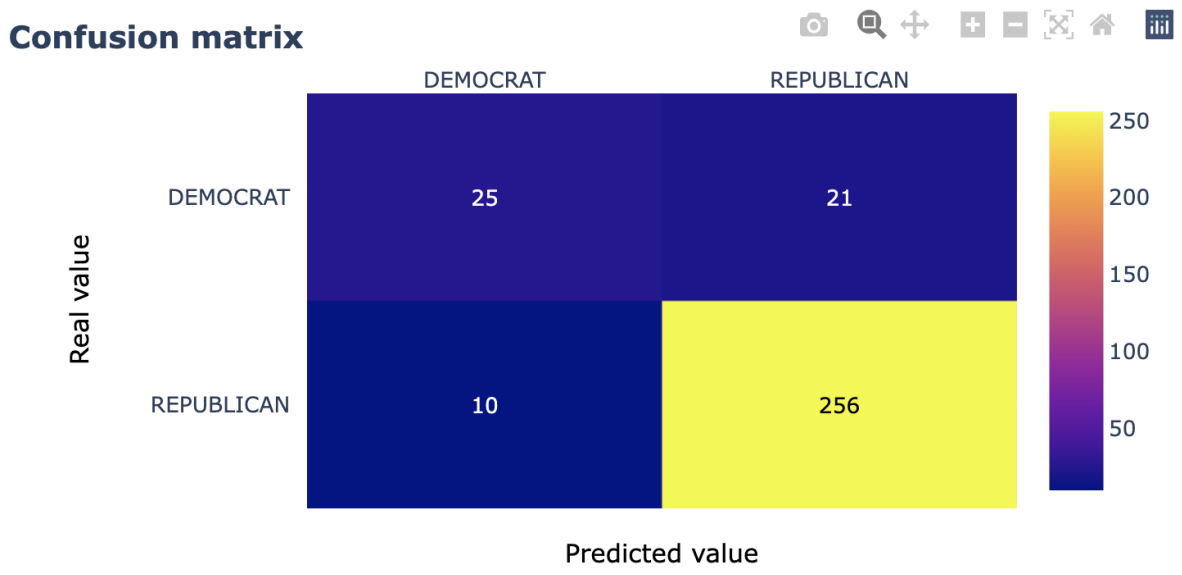
As a result, for our final model we are removing the neighbors feature because it decreases our model's accuracy. We believe that using the neighbors as a feature decreases the accuracy because due to politics neighboring states will have different levels of mask use and vaccination rates, and therefore looking to see what the neighbors' vaccination rates and mask use rates is not a good way to tell how a county voted. For example a blue county could have a high vaccine rate and mask usage, while the corresponding red neighbor will have low vaccine and mask usage. It becomes harder to distinguish between counties when using the neighbor for the same reason.

Evaluation Result For Model

Our final model (binary classifier) accuracy is 90% on the test set, which is significantly better than a baseline random guesser which has 50% accuracy for predicting which party a county voted for. This accuracy is fairly good because using our model we can almost predict all counties' political preferences.

After feature engineering and parameter tuning on our validation sets, we achieved 90% accuracy on the test set, therefore depending on an accuracy threshold, we can accurately

predict how a county voted in 2020 based on features such as mask use and vaccination rates. This depends on what level of accuracy is to be considered an accurate model.



Future Work

For future work, we can try to predict the voting preference for the neighboring counties of a given county. Currently, we only predict the voting result for a given county.

From plot 2 (False Predictions in Validation Set), we can see that the errors of our model come from the case which the county actually votes differently than the state, due to the fact that in some areas of a state the voting preference could be mixed due to the demographics and social-economic status of the area. If we can predict the neighboring county's voting behavior, we can have a better understanding of that specific part of the state.

This new direction could be interesting because it will help us understand how COVID affects people in different areas and prevent over-generalization. It will also help us discover and analyze how the COVID policy changes or reinforces people's political preferences, or how the party affiliation affects one's belief about COVID. That information could be useful for the next election.

We could do an analysis on the news and media over 2019-2021, and use more targeted windows of time to choose our vaccination rates over time features. This could be scraping tweets related to covid safety and analyzing their sentiment.

An interesting take away from this project is considering that features such as mask use and vaccination could be used and most likely will be included when creating a 2024 presidential presidential election.

External Dataset Citation

Dataset: "Popular vote backend - Sheet1 (1).csv" (2020 election results by states)

David Wasserman, S. A. (n.d.). *2020 popular vote tracker*. Cook Political Report. Retrieved December 14, 2021, from https://cookpolitical.com/2020-national-popular-vote-tracker.

Info: This data set contains the 2020 popular vote by state. We used this dataset in our initial logistic regression model when predicting how a state voted based on mask use and vaccination rates.

Dataset: “COVID-19_Vaccinations_in_the_United_States_County.csv” (Vaccination data by counties)

Centers for Disease Control and Prevention. (n.d.). *Covid-19 vaccinations in the United States,county*. Centers for Disease Control and Prevention. Retrieved December 14, 2021, from https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh.

Info: This dataset contains vaccination information by county in the US. We used this to merge vaccination info to our county dataframe with the FIPS value as the key. This dataset is also sorted by day, so we used this data to calculate vaccination rates over time. We used this in our updated model.

Dataset: “countypres_2000-2020.csv” (2020 election results by counties)

MIT Election Data and Science Lab. (2021, June 22). *County presidential election returns 2000-2020*. Harvard Dataverse. Retrieved December 14, 2021, from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FVOQCHQ.

Info: This dataset we used to see how counties voted in the 2020 presidential election. This dataset contained a FIPS code we could use as the key, and we used the information in this table to create labels for our updated county-level model.