Open Ended Modeling Report

Rachel Seong, Nianxu Wang, Max Kim

Dec. 13, 2021

Initial EDA Analysis (From Part 1)

The Data

Given Datasets

Initially, we were given four datasets: counties, vaccination, cases, and mask_use. The counties dataset provided population estimates of each county by FIPS code needed to calculate the COVID-19 cases per capita for each state. The vaccination dataset contained data of the number of fully and partially vaccinated individuals by FIPS code over time; this allowed us to generate the proportions of fully and partially vaccinated individuals in each state over time. The mask_use dataset provided the frequency of never, rarely, sometimes, frequently, and always wearing a mask by FIPS code based on roughly 250,000 survey results collected from Dynata during July 2-14, 2020. Lastly, the cases dataset provided the actual number of COVID-19 cases over time by FIPS code.

EDA Analysis

Guided EDA

From the guided EDA, we visualized the relationship between COVID-19 cases per capita and the frequency of never, rarely, or sometimes wearing a mask in the 4 counties located in Alabama, Texas, Illinois, and California (Figure A). We noticed that Mobile, AL and Tarrant TX, the two counties with the highest frequency of never, rarely, or sometimes wearing a mask also had the highest COVID-19 cases per capita among the four counties; the opposite would hold true for the other two counties.

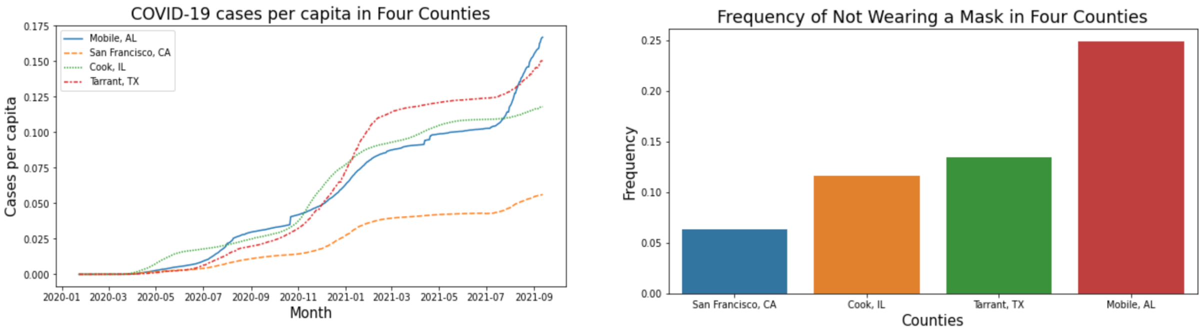
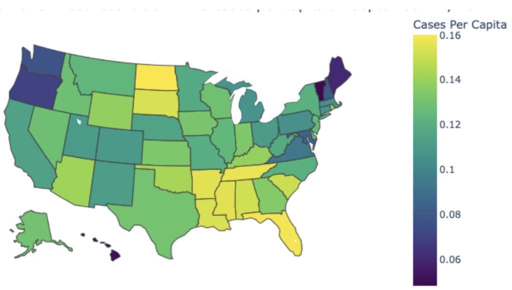


Figure A - COVID-19 cases per capita and Frequency of Not Wearing a Mask in Four Counties

Additionally, after plotting a choropleth map of the U.S. of the most recent COVID-19 cases per capita (Figure B) and a choropleth map of the U.S. of the frequency of never, rarely, or sometimes wearing a mask (Figure C), there were many trends indicating that states with higher frequencies of never, rarely, or sometimes wearing a mask also likely ended up having higher COVID-19 cases per capita; the opposite also generally would hold true.



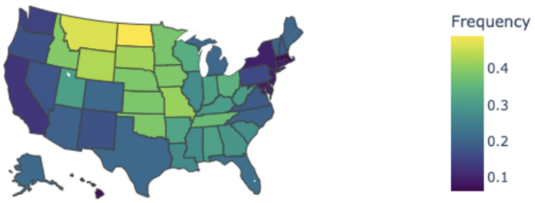


Figure B - Number of most recent COVID-19 cases per capita on September 12, 2021

Figure C - The frequency of never, rarely or sometimes wearing a mask

From these observations, we began to ask questions: *what else can we say about the relationship between safety protocols such as mask-usage and vaccinations and COVID-19 cases per capita? How else can we visualize this relationship in a more direct way?*

In other words, from our guided EDA, we began to see how potentially important safety protocols such as mask-usage might be for preventing the spread of COVID-19. Then, we began to wonder: *what factors determine whether the people residing in a particular state follow safety protocols? In a given state, is there a relationship between whether a high proportion of people follow one safety protocol such as mask-usage and another such as vaccinations?*

Guided Unsupervised EDA (From Part 1)

In our Unsupervised EDA, we began to explore many of these questions by generating our own visualizations. In order to directly visualize the relationship between COVID-19 cases per capita and mask-usage frequency, we generated a scatter plot to measure the relationship between the average mask-usage frequency by state during early July 2020 and most recent COVID-19 cases per capita by state as of mid-September 2021 (Figure D). From this visualization, we noticed that mask-usage frequency rates in July are moderately negatively correlated with recent COVID-19 cases per capita. Conspicuous outliers on the right portion of the plot may explain why we cannot observe a strong correlation. For instance, although Hawaii and Rhode Island both had high mask-wearing frequency, covid cases per capita of Rhode Island (0.143) is three times greater than that of Hawaii (0.049). These outliers indicate that there may exist confounding variables that might influence the mask-wearing frequency or covid cases per capita. In particular, a possible confounding variable could be the political inclination of each state. For instance, states where a greater portion of people frequently wear masks—Hawaii, DC, Maryland, and more—are all democrats. In contrast, states where a lesser portion of people frequently wear masks—Montana, North/South Dakota, Wyoming, and Oklahoma—all are republican states. Such a pattern may imply that political ties of the state may play a role in whether or not people choose to wear masks or not.

Visualization #1

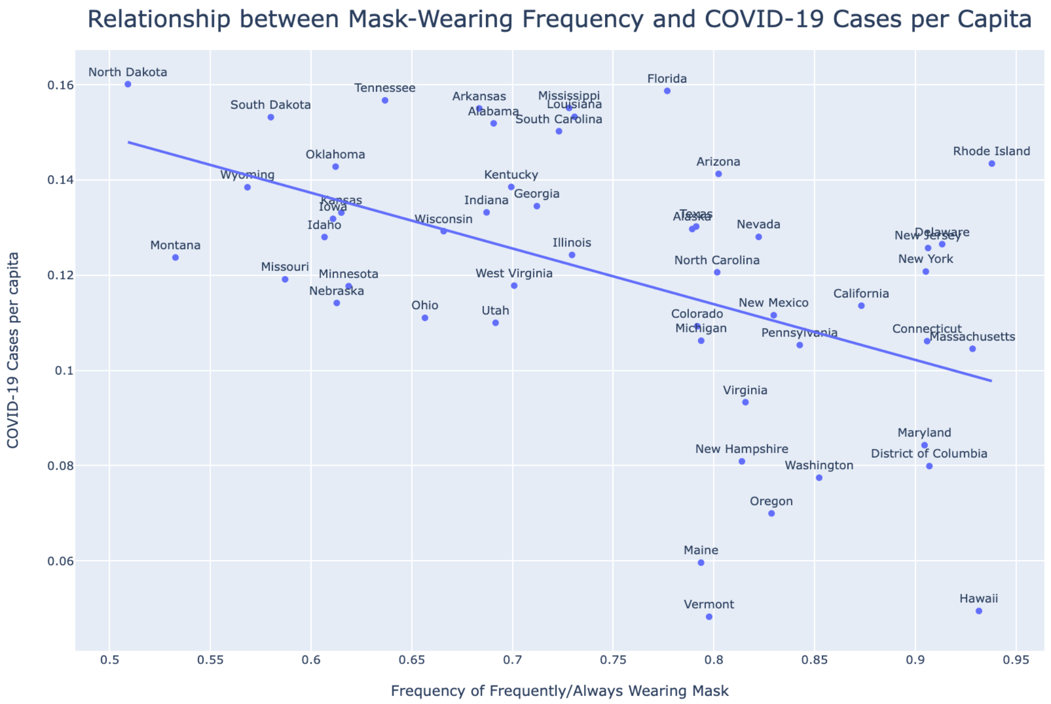


Figure D

Visualization #2:

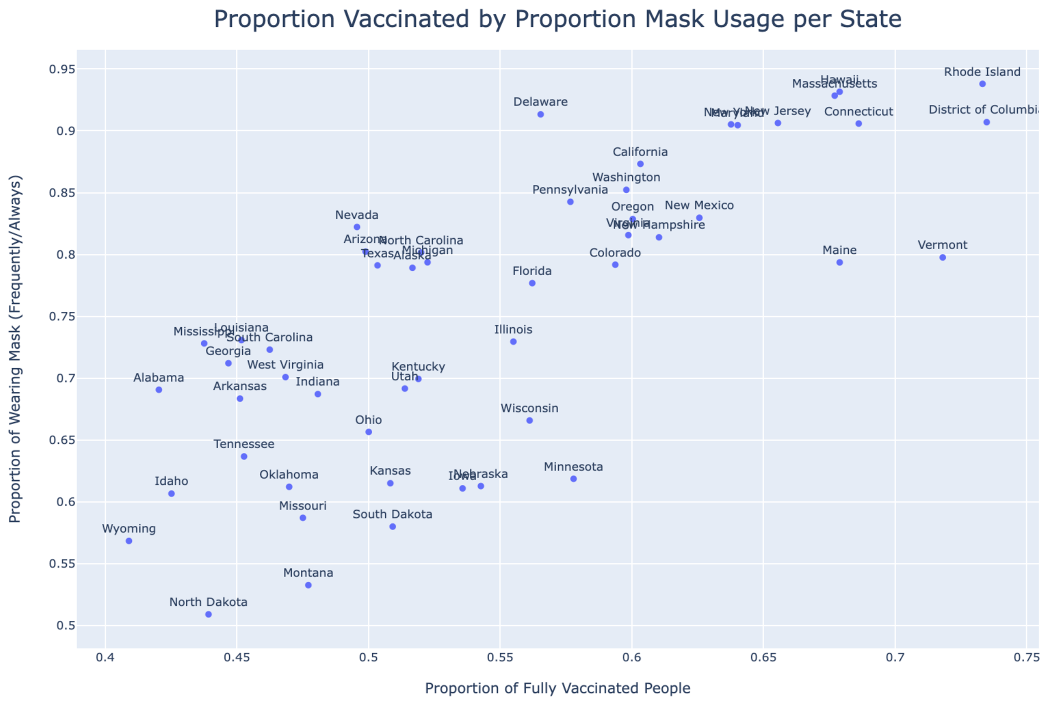


Figure E

What we are addressing here is investigating the relationship between vaccinations and mask usage. If we can see whether states that have a higher proportion of mask usage tend to also have higher vaccination rates, it will paint a better picture of whether people are following one safety protocol more than the other. This helps to identify issues where people are not abiding by one or both safety protocols.

What we found is that generally, states with a higher proportion of masked people also have higher vaccination rates. There is also a political divide: red states tend to have both lower masking and lower vaccination rates, blue states the opposite. In general, there is a slightly higher rate of mask usage versus fully vaccinated. This is easily explained by i) it's likely easier to put on a mask than to go and get vaccinated, and ii) most vaccines require two doses with a month in-between, so it takes longer to get fully vaccinated.

Because of the fairly strong positive correlation between masked percentage and vaccination percentage, we may only need to use one of these features, or combine these two in a certain way into one feature, in order to have a higher rank closer to the number of columns/features, and not have repeats.

In summary, with the importance of safety protocols in preventing the spread of COVID-19 and the association of the proportion of people following mask-usage and vaccination protocols highlighted in our unsupervised EDA, we began to wonder: *In a given state, what sort of factors may affect the proportion of people that follow safety protocols? Could one of such driving factors be political?*

These questions ultimately led us to our hypothesis, exploring the problem of whether the extent to which the people in states are following safety protocols may be associated with the state's political ties.

Hypothesis and Model (Updated for Clobber)

Hypothesis

Safety protocols, specifically the vaccinations daily rate per capita and tests daily rate per capita, are positively correlated with whether a given state is Republican, and thus, the extent of the **impact of**

COVID-19, measured by the cases and deaths daily rate per capita, are negatively correlated with whether a given state is Republican.[1]

We plan to extract "presidential_elections.csv" data from homework 10 containing data on how the 50 states plus Washington D.C. voted in presidential elections between 1972 to 2016 and adding the presidential election data from 2020. Additionally, we'll be using the vaccinations data, cases data, and importing two more external datasets: *https://github.com/nytimes/covid-19-data* for the time-series death data and *https://github.com/govex/COVID-19/tree/master/data_tables/testing_data* for the time-series total tests done data. We utilized each of the datasets to generate the 5 features mentioned in our hypothesis. Ultimately, we plan to validate or invalidate our hypothesis by reading our model weights and odds ratios and using them to accept or reject the hypothesis based on their sign. We will accept the hypothesis if all 5 model weights and odd ratios align with our hypothesis and reject the hypothesis if at least one of them does not align with our hypothesis because this contradicts our narrative about safety protocols and the political ties of a state.

Answer

Although we chose to **accept** our hypothesis based on the signs of the model coefficients and odds ratios from our model, we recognize the limitations of our hypothesis because we chose to leave out the magnitudes of these ratios in assessing our hypothesis since we didn't hypothesize about the strength of the correlations between our features and the political ties of a state. The model coefficients and odds ratios come from our equation: $f(x) = 1 / (1 + e^{-(.70 - 24.35x_1 - 20.82x_2 + 362.90x_3 + 8.92x_4 - 173.16x_5)}$). After the intercept, from left to right, the order of our features were: people_fully_vaccinated, people_partially_paccinated, daily_case, deaths_diff_per_capita, and the tests_diff_per_capita. After extracting the signs of the coefficients and odds ratios of our model, we concluded that the results align with our hypothesis.

[1] These inputs are further defined in the "Model" Section

Model

With our model, we plan to use a logistic regression model to use the vaccination per capita daily rate (i.e. difference between each day count divided by POPESTIMATE2020 for both partially and fully), the cases per capita daily rate (i.e. difference between each day divided by POPESTIMATE2020), tests per capita daily rate (i.e. difference between each day divided by POPESTIMATE2020), and the deaths per capita daily rate (i.e. difference between each day divided by POPESTIMATE2020) as our inputs to “predict” whether a given state is Democratic: 0 or Republican: 1 as our output. For context, the last two features were added after our baseline model suffered from low training and validation accuracy; we, thus, pivoted our hypothesis to be slightly more specific.

We chose to use a logistic regression model because we’re predicting discrete values to understand the possible relationship between safety protocols and COVID-19’s impact and whether a given state is Republican. We chose to use the default, cross-entropy loss, as we’ve seen the infeasibility of using L2 loss/MSE for logistic regression models and for our model. We split our mask-usage data into training and test sets. Then, we use our training data when designing our model and use cross-validation to test generalization.

Model Evaluation and Analysis

We first measured model performance comparing the ROC curves of the baseline and the improved model.

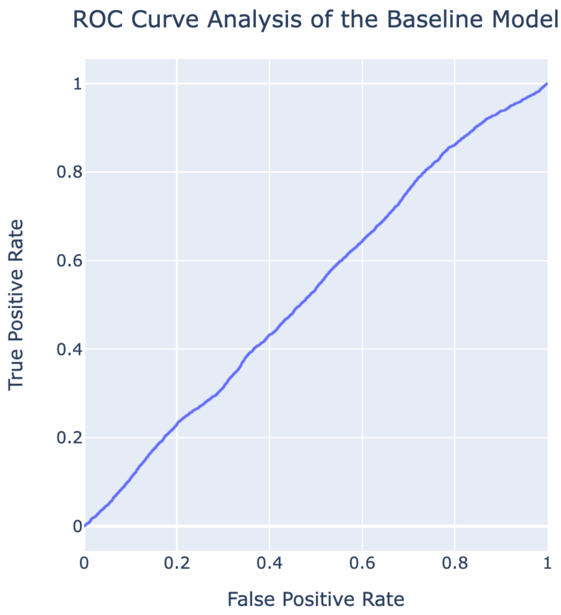


Figure F

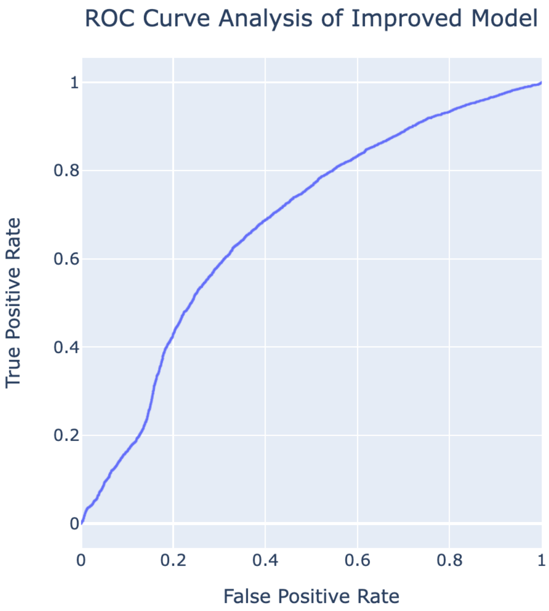


Figure G

We trained the logistic regression model, and our resulting accuracy was 65%, relatively higher than the baseline logistic regression model which would attain 51%. We created a ROC Curve plot to analyze and compare the predictive performance of the two models. As shown in figure F, the ROC curve of the baseline model is closer to a constant line of slope 1; it shows that our baseline model is rather a random classifier, with no predictive value. In contrast, as shown in Figure G, the ROC Curve of the improved model is smoother where the curve is closer to the top-left corner. This indicates that it has a better performance compared to the baseline, because the curve shows that the false positive rate is lower and the true positive rate is higher. The precision of the improved model was 0.628, implying that 62.8% of the positive predictions were correct predictions. The recall of the improved model was 0.71, implying that 71% of the positive predictions of all observations were correct. Although it is evident that the improved model performs better than the baseline model, the ROC curve looks less like what the ideal, well-performing model would look like. Thus, the model result isn't necessarily good, and this motivates a further model improvement in the future.

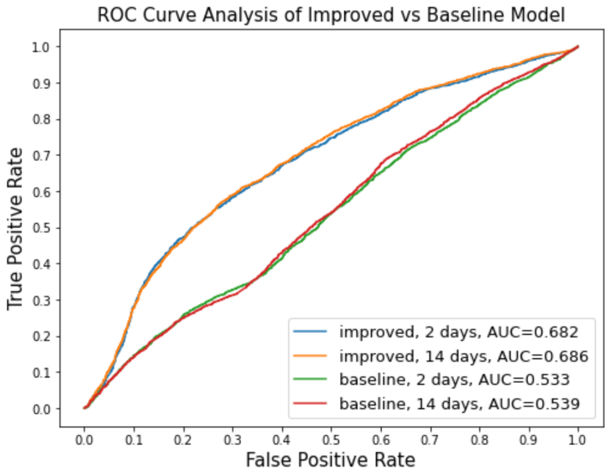


Fig H: Short Term vs Long Term Prediction (Different K-Values)

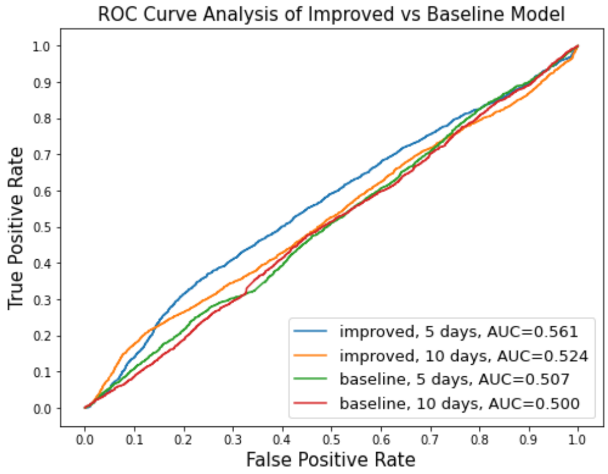


Fig I: Explicit Future Predictions (5 & 10 days ahead)

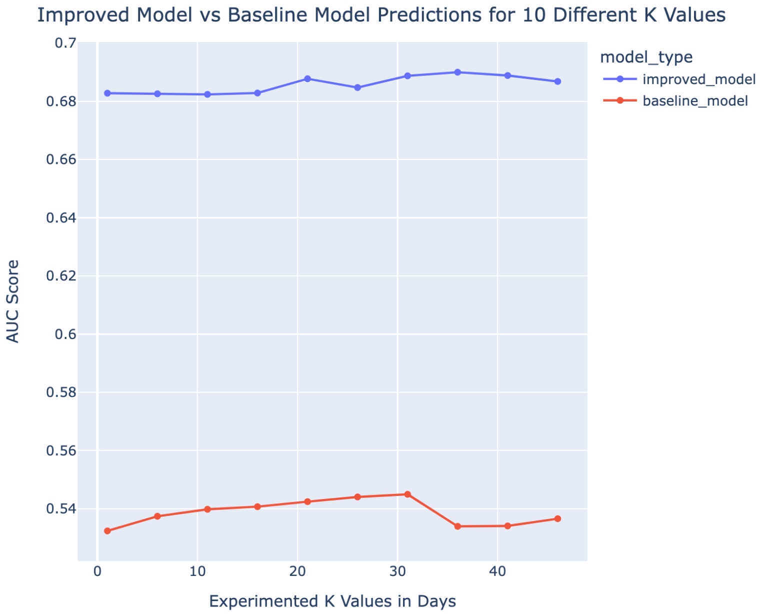


Fig J

Moving on, we wanted to experiment with the predictive power of our improved model and compare such with the baseline model. In particular, we trained the two models to cumulatively predict the next day upto 2 days and 14 days. As shown in Figure H, AUC score of the improved model is 0.682 (2 days prediction) and 0.686 (14 days prediction), higher than that of baseline (0.533, 0.539 for 2 days and 14 days, respectively). Given that the AUC score measures the ability of the model to distinguish between classes, we could interpret that the improved model correctly labeled between democrats and republicans 68% of the time, while the baseline model was still close to a random guesser. It was interesting to see that the AUC score for the 14 days prediction was slightly higher than that of the 2 days prediction; this could imply that our model is slightly better at long term prediction.

Extending on the experiment done as shown in Figure H, we wanted to try different k-values (1 to 50 days, incrementing by 5) and see how our model would predict larger k-values. Here, we could confirm that the aforementioned insight of our model was correct; our model makes better long term cumulative predictions, even up to 50 days. As shown in Figure J, the AUC score of the improved model increased when we tested 25 - 50 days, while that of the baseline model decreased starting from 30 days and on. This shows that not only our improved model predicts better than baseline model, but also suggests that it makes better long term predictions than baseline.

After exploring different k-values, we were interested to investigate whether our model was also capable of making explicit predictions instead of cumulative predictions of adjacent days. As shown in Figure I, the AUC score of the improved model for both 5 & 10 days prediction plunged to 52% ~ 56% although it was still slightly higher than that of baseline model. This indicates that our model is incapable of making explicit future predictions when the training data for each of previous days are not fed. However, it is worthwhile noting that the improved model's AUC score for 5 days prediction is conspicuously higher than that of 10 days prediction. This indicates that although our model may be incompetent at making explicit predictions, it still has better short term predictability than long term. Result shown in Figure I particularly motivates us to further improve our model, because we could observe that our model struggles to make explicit predictions, which was contrary to our expectation after experimenting with cumulative predictions.

Model Improvements

Preface: Applying 5-fold cross validation

Problem: The baseline model suffered from fairly low training accuracy; we figured that we needed more confidence in the performance of our model during the design process.

Solution & Result: We chose to apply 5-fold cross validation to get a better sense of the performance or accuracy of our logistic regression model. Additionally, we wanted to ensure that we were properly testing generalization (underfitting/overfitting) for our model. So, instead of just the one-time split of the training and validation sets, we also used cross validation to get a better estimate of our model's performance accuracy going into the design process. After 5-fold cross validation, we ultimately noticed an increase in our model's performance accuracy, where the output accuracy was much closer to the validation accuracy from the initial split.

Improvement 1: Adding death daily rate per capita as a feature

Problem: The baseline model suffered from both low training and validation accuracy; our baseline model only had three features: partially vaccinated per capita daily rate, fully vaccinated per capita daily rate, and daily cases rate. Thus, we figured that the three features we had weren't representative enough of the safety protocol/measures taken by each political party and the extent to which COVID-19 may have impacted each political party differently.

Solution: We chose to include an additional feature: death daily rate per capita. This way, we introduce more data to our model and thus would hope to improve the performance accuracy. We chose to include this feature because we felt that this statistic aligned with assessing the extent of COVID-19's impact, and we also noticed that the daily rate of deaths per capita was comparatively higher for a particular party after plotting this feature and comparing the rates of the two parties over time (Figure N); we also took into account the coefficient (8.92) and the odd ratio of this feature when deciding to include the feature into our model.

Result: We noticed a considerable increase in both training and validation accuracy, so we concluded that our intuition may have likely been correct since we considered accuracy as our metric.

Further Explanation: Below, you'll find a plot for each of our 5 features explained in our model section over time by party. We utilized these plots to help validate the inclusion of the new features but also to validate our current features.

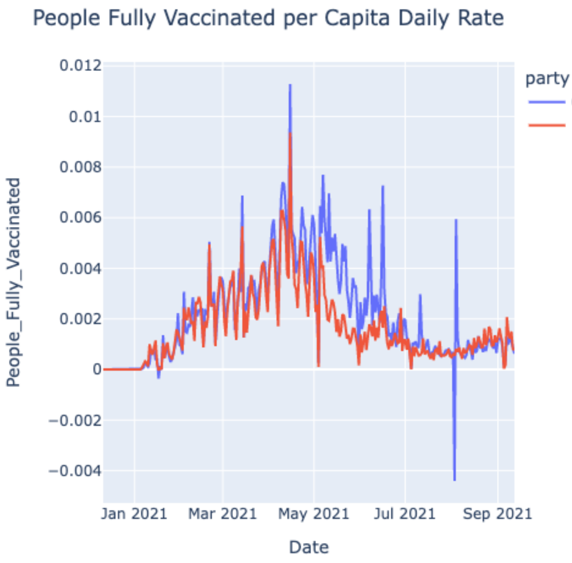


Figure K

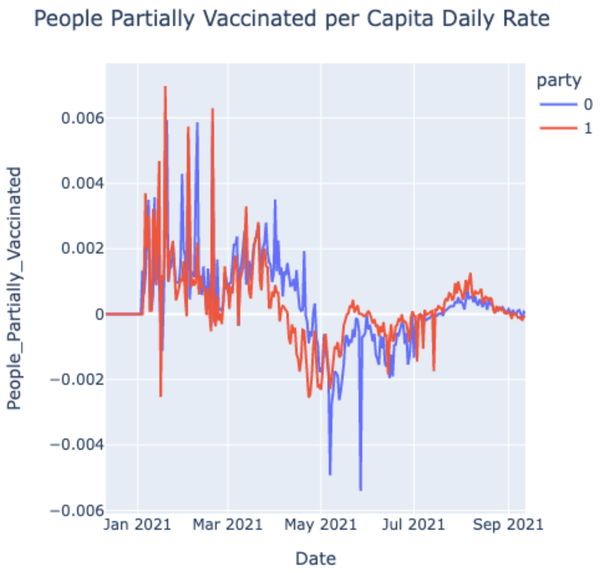


Figure L

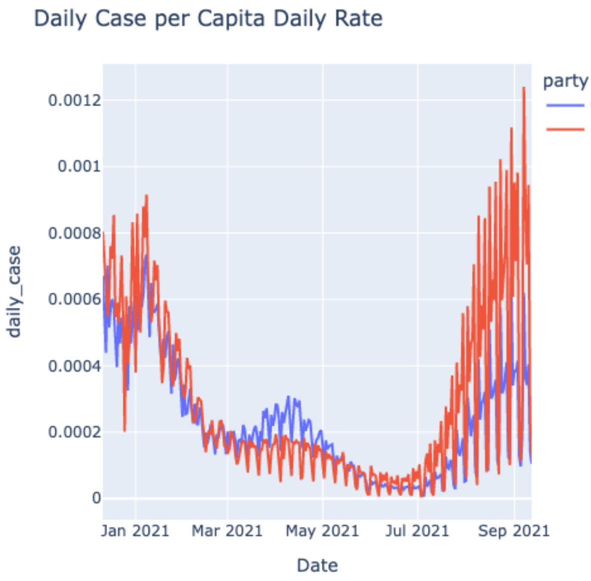


Figure M

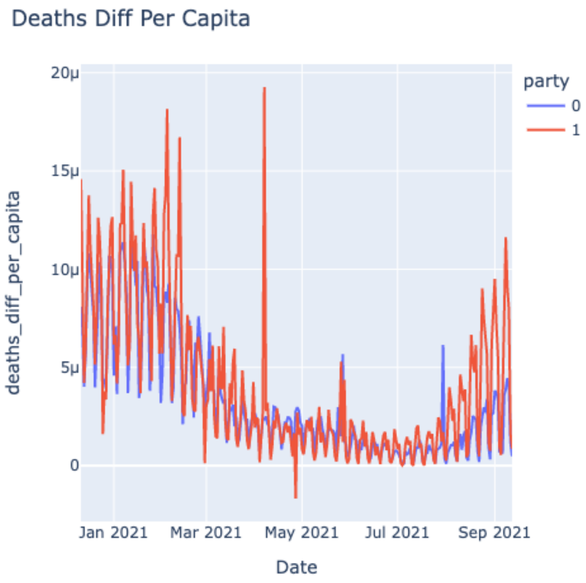


Figure N

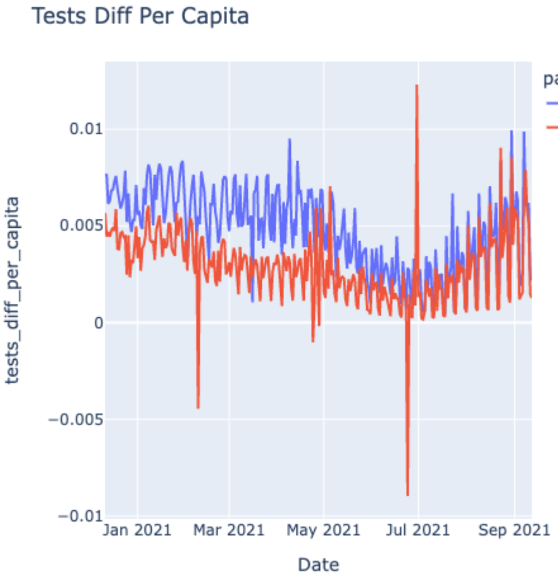


Figure O

Improvement 2: Adding testing daily rate per capita as a feature

Problem: The baseline model still suffered from fairly low performance accuracy; our baseline model now had four features: partially vaccinated per capita daily rate, fully vaccinated per capita daily rate, and deaths per capita daily rate. Thus, we figured that adding an additional relevant feature that aligned with our narrative of safety protocols would help improve our model's performance accuracy.

Solution: We chose to include an additional feature: COVID-19 daily tests rate per capita. This way, again, we introduced more data to our model, and thus would hope to improve the performance accuracy. We chose to include this feature because we felt that this statistic aligned with safety protocols, and we also noticed that the daily rate of tests per capita was comparatively higher for a particular party after plotting this feature and comparing the rates of the two parties over time (Figure O); we also took into account the coefficient (-173.16), and the magnitude of the odd ratio of this feature when deciding to include the feature into our model.

Result: We noticed a fairly large increase in both training and validation accuracy, so we concluded that our intuition may have likely been correct. We improved the training accuracy by 14% (baseline: 51%, improved model: 65%).

Final Thoughts for Future

In the future, our group wanted to further explore other possible COVID-19-related features that may be correlated with whether a given state is Republican. Currently, our model includes safety protocols such as vaccination and testing rates, but we wanted to explore the possible relationships between other safety protocols such as mask-usage and the political ties of a given state. However, for the mask-usage for example, we were unable to find time-series data for mask-usage frequency rates and only had the mask-usage data given from part 1 of the project. This data only consists of mask-usage statistics based on a survey conducted from July 2 to July 14 in 2020. If given more time, we wanted to possibly find other time-series data for mask-usage or if unavailable, look for other possible COVID-19 features related to safety protocols to include in our model to better predict the political ties of a given state. This would be interesting to explore, since it may not only improve our model's overall predictive accuracy, but also improve the ability to make explicit short and long term predictions. Additionally, we would also have liked to include features related to COVID-19's impact to improve the accuracy of our model. Then, with these features, in the future, we would like to also look more deeply into the **level** of correlation for each of these features with the political ties of a state.

Ultimately, since our experiment aimed to explore the possible relationships between a state's political ties and what that might mean for the safety of the individuals within that state during a pandemic such as COVID-19, we think there are many ways we could improve our model to help better strongly establish the existence of these relationships.