Charles Reinertson
Data 512
Dec 6, 2022

Canonical Written Report

1. **Introduction**: Why is this analysis interesting or important (to people besides you)? Does it solve a real problem or tackle an unresolved research question?

    - For the **common analysis** component, I asked the question how did masking policies change the progression of confirmed COVID-19 cases from February 1, 2020 through October 1, 2021 in Wayne, Michigan. This is an important question in determining the future of mask-wearing and masking mandates. At the start of the COVID pandemic, many people chose to wear masks because in theory, they would help stop the spread of COVID. However, with years of data behind mask mandates and the effect of those mandates on the spread of COVID, it is reasonable to assume people would require near-conclusive analysis that points towards masks slowing the spread of COVID if those same people are going to adopt mask-wearing again. This analysis works towards solving this real problem for Wayne, Michigan.

    - For the **extension plan** I built a model to forecast mask compliance in Wayne, Michigan to allow people to benefit from the Mask-Wearing Survey in perpetuity. This human-centered task allows those living and traveling to Wayne, Michigan to learn the rates of mask usage without necessitating updated mask compliance surveys. Henceforth, residents and visitors of Wayne can properly calculate their risk tolerance for COVID given their age and preexisting conditions. Mask compliance surveys are expensive, hard to implement and require public participation which decreases as people care less about COVID. Through the use of publicly available data such as daily COVID case count and daily mask mandates, we can forecast mask compliance without necessitating these surveys.

2. **Background/Related Work**: What other research has been done in this area? How does this research inform your hypotheses, your analysis, or your system design? What are your hypotheses or research questions? For

these COVID related questions there may not be peer-reviewed publications that are directly related to your hypothesis. There may be anecdotal claims in the popular press (blogs, newspapers) related to your analysis.

- Our **common analysis** touches on the effect of masks (specifically mask mandates) on COVID cases. Many research papers look at this important human-centered question or something very similar. The paper *Unmasking the mask studies: why the effectiveness of surgical masks in preventing respiratory infections has been underestimated* touches on the empirical evidence that indicates masks prevent disease transmission: "The framework demonstrates that masks can have a disproportionately large protective effect and that more frequently wearing a mask provides super-linearly compounding protection." Another study that looks at a question more directly related to the common analysis is *Face Mask Use in the Community for Reducing the Spread of COVID-19: A Systematic Review*. The author did a systematic review and meta-analysis to investigate the efficacy and effectiveness of face mask use in a community setting. Their findings support the use of face masks in a community setting to limit the spread of COVID. My analysis centered around the hypothesis that mask mandates (present or not) in Wayne, Michigan correlate with COVID cases. These studies indicate that my hypothesis of mask mandate correlation with COVID cases is reasonable.

- My **extension plan** hypothesizes that using daily COVID cases and whether or not mask mandates are present will allow me to accurately predict mask compliance in Wayne, Michigan. The paper *Optimizing LSTM for time series prediction in Indian stock market* optimized a Long Short Term Memory (LSTM) for time series data prediction of the Indian stock market. Although this is a very different research question than my own, the characteristics of the data I use and the data used in this paper are similar. Both datasets contain sequences of time series data with long-term trends, seasonality, cyclical fluctuations and random noise. This paper informed my research design to use an LSTM for regression to predict mask compliance. Many other papers dive into forecasting with COVID data. However, the papers I found look at forecasting COVID case counts or hospitalization censuses such as *Direct Multi-Step Forecasting with Multiple Time Series Using XGBoost: Projecting COVID-19 Positive Hospitalization Census for a Southern Idaho Health System*. I take a similar approach with a different

goal of forecasting the mask compliance of residents in Wayne, Michigan. However, the fact that others were able to perform forecasting with similar COVID datasets leads me to believe that my analysis is achievable and the hypothesis is reasonable.

3. **Methodology**: Not just your analytical methods, but also, why you chose them, and how human-centered considerations such as ethics informed the way you designed your study.

   - For the **common analysis** you can see my methodology in part_1.ipynb. I first clean the datasets to get new COVID cases per day and mask mandates (yes or no) per day. However, COVID case numbers are not always reported for individual days and the count is sometimes added to subsequent days. Henceforth, I smoothed the data by taking the average of a three-day window for each day. I then took the gradient of the COVID cases per day and made a graph plotting these gradients and colored points yellow where face masks were required and blue when mask mandates were not in effect. It is hard to discern from this plot whether masking policies changed the derivative function in a meaningful way so I performed further analysis. I used Facebook Prophet (for modeling time-series) and used the changepoint detection component to see if COVID case changepoints line up with mask mandates. This methodology showed that mask mandates (present or not) in Wayne, Michigan correlate with COVID cases. I incorporated the human-centered data science principle UX for ML in designing my common analysis. As stated above, the graph comparing the gradient of COVID cases to mask mandate status was hard to comprehend. This goes against the human-centered data science principle UX for ML because users will gloss over my analysis as it is hard to understand. Therefore, I went a step further in utilizing Facebook Prophet for a better visualization experience. This graph is more readable and follows the human-centered data science sub-principle of explainability.

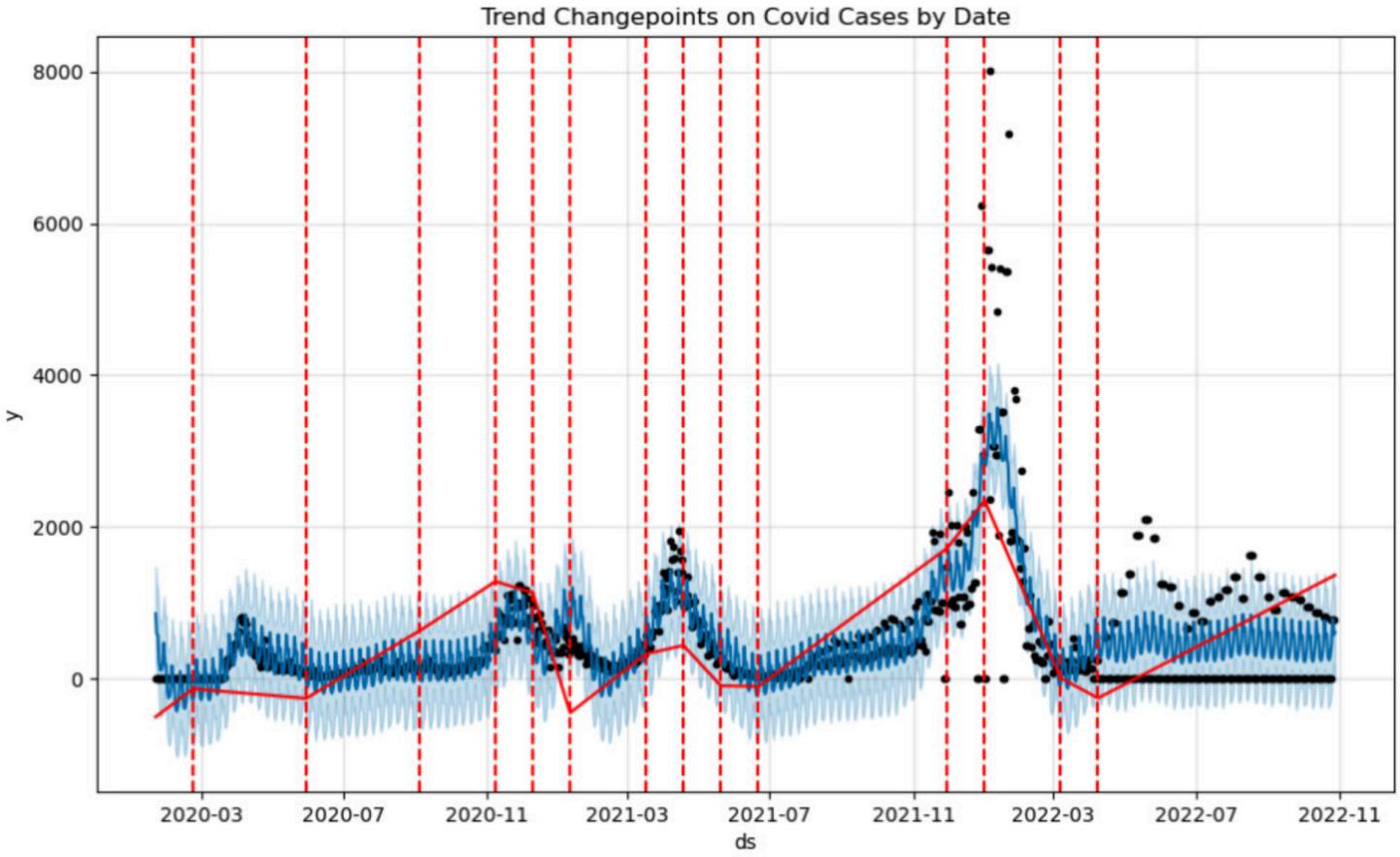
   - For the **extension plan** I first preprocess the data read in from the three datasets of the README section Data Descriptions. This step-by-step processing can be seen in step_1_clean_train_data.ipynb. At the end of this notebook, I save the final dataframe containing the columns state, county, date, daily case count, boolean daily face mask required, and mask use. Mask use is calculated by using the NY Times Mask-Wearing Survey Data.

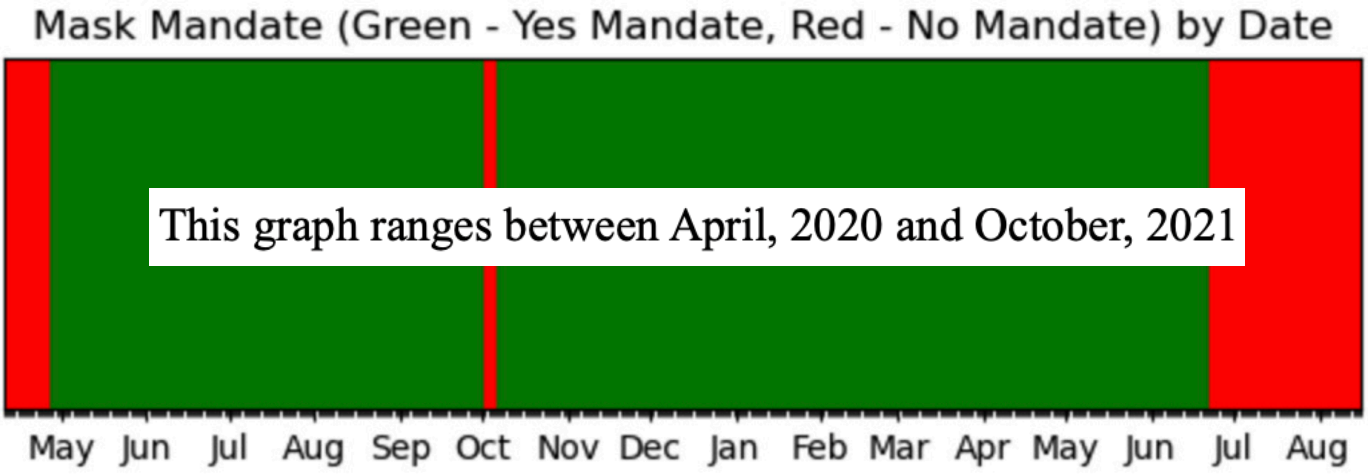
The survey asks “How often do you wear a mask in public when you expect to be within six feet of another person?” and fills in the percentage of people who fall into each of these five categories - NEVER,RARELY,SOMETIMES,FREQUENTLY,ALWAYS. These five columns can be represented by one numerical column by multiplying the percentage in each column by 0 for "NEVER, 0.25 for "RARELY", 0.5 for "SOMETIMES", 0.75 for "FREQUENTLY", and 1 for "ALWAYS". Finally, adding these results together will give us a nominal continuous representation of mask usage between 0 and 1. This single mask usage variable is much easier for the model to learn to predict in step_2_train_model.ipynb using daily case count and boolean daily face mask required as features. I held out counties named Montgomery, Wood, King, Columbia, Washington, Jefferson and Wayne from the training data to test if the model can generalize to unseen data. This guarantees the model is not trained on data from Wayne, Michigan (the county I base my hypothesis on). The model is then saved in the folder models as model.pt to be used in step four. That way, I do not have to train a new model every time I want to make a prediction. After training the model, the next step is the notebook step_3_clean_prod_data.ipynb which does the same data processing as step one but for an extra year long time window and only for counties named Wayne (most importantly Wayne, Michigan). This notebook simulates a data pipeline that can be fed into a pre-trained model in production. Finally, in step four step_4_prod_model.ipynb, I load in the pertained model and run the cleaned data from step three through this model. I visualize the predicted mask usage compared to the true mask usage in graphs and save these graphs to outputted_images. The predicted mask usage extends from April 2020 to April 2022, farther in time than the true mask usage that is between April 2020 and April 2021. The predicted mask usage allows users of this system to get accurate mask usage data without having to rely on the outdated NY Times mask usage survey. I forecast a year out from the point where I assume that the actual mask usage is not accurate (too much time has passed since the NY Survey was conducted). I considered ethics in designing my study by not considering data such as the gender or race of people living in a county. That way, race and gender biases are not unintentionally learned by my model in predicting mask usage per county. I also noticed that some counties’ mask usage forecasts were incorrect. However, the benefit of having somewhat flawed forecasts rather than no forecasts outweighs the human-centered ethical issue of publishing incorrect

forecasts that lead people to believe counties are more or less COVID-safe than in actuality.

4. **Findings**: What did you find? Use words and figures, don't just point to code.

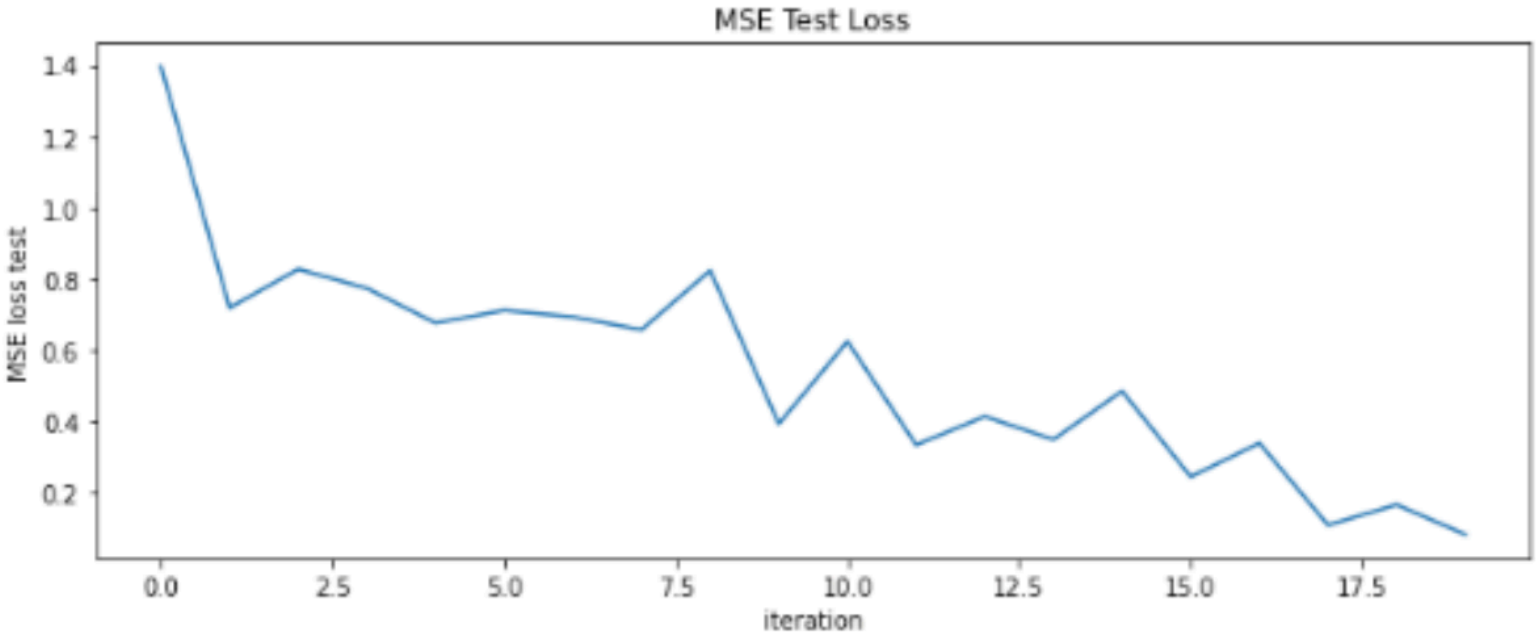
   - For the **common analysis** I found that masking policies change the progression of confirmed COVID-19 cases from February 1, 2020 through October 1, 2021 in Wayne, Michigan. We can see from the following graphs





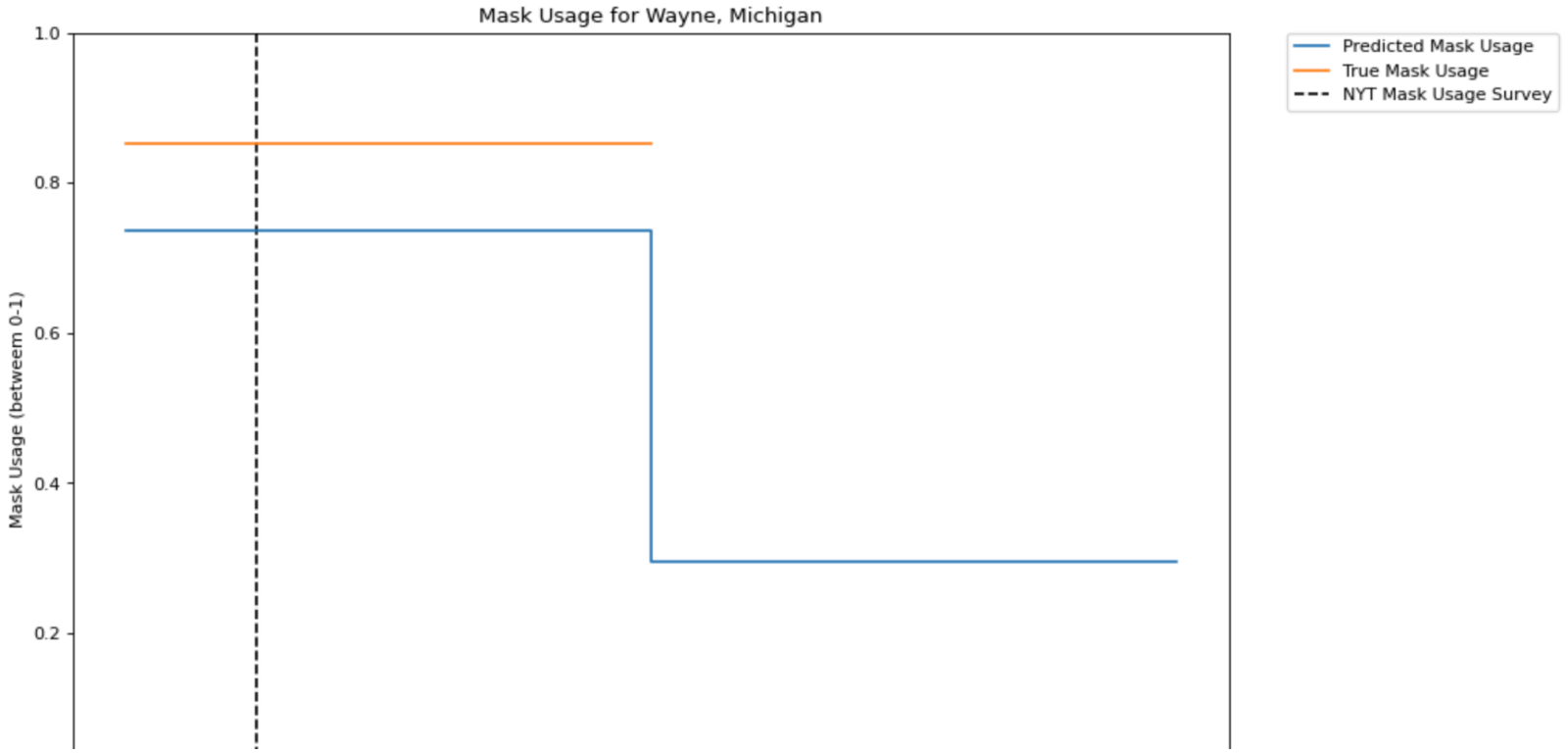
this effect. In October 2020 the mask mandate was removed in Wayne, Michigan. A changepoint occurred in the new COVID cases per day graph a month after the mask mandate was removed and subsequently COVID cases spiked. In May of 2020, Wayne county went from a no-mask mandate to a mask mandate and we see a changepoint occur right around this time as COVID cases per day dropped. Finally, the mask mandate was dropped right before July 2021 which corresponds to a changepoint in the new COVID cases graph and a large rise in COVID cases. In summation, each mask mandate change corresponds to a changepoint in the COVID cases per day graph and a subsequent rise or fall in COVID cases. Henceforth, masking policies change the progression of confirmed COVID-19 cases from February 1, 2020 through October 1, 2021 in Wayne, Michigan.

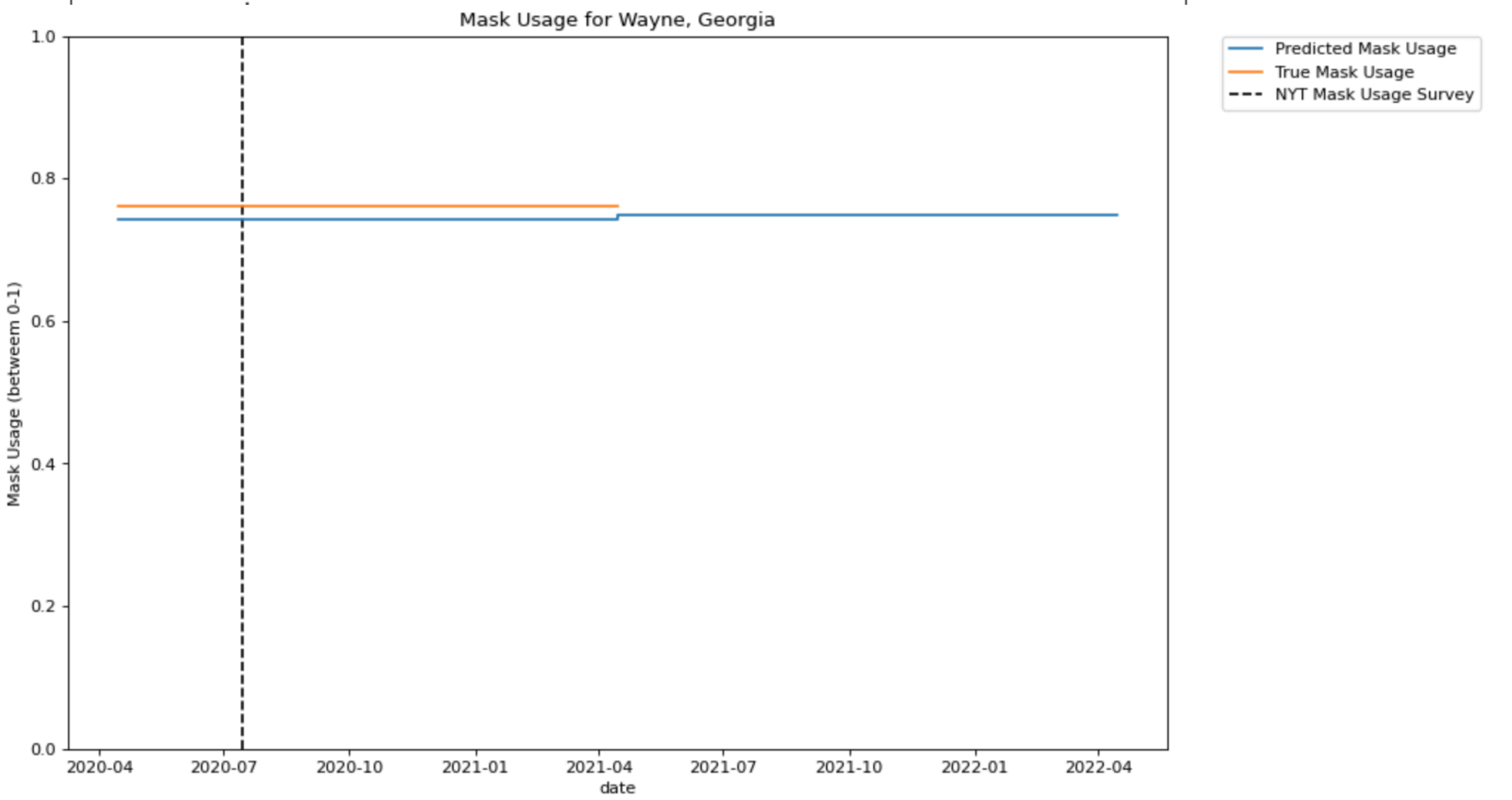
- For the **extension plan** I measured how well the model learned by plotting the mean squared error of the true mask compliance variable compared to the predicted mask compliance variable for the following held-out counties from the training set: All counties named Montgomery, Wood, King, Columbia, Washington, and Jefferson. We can see that the model

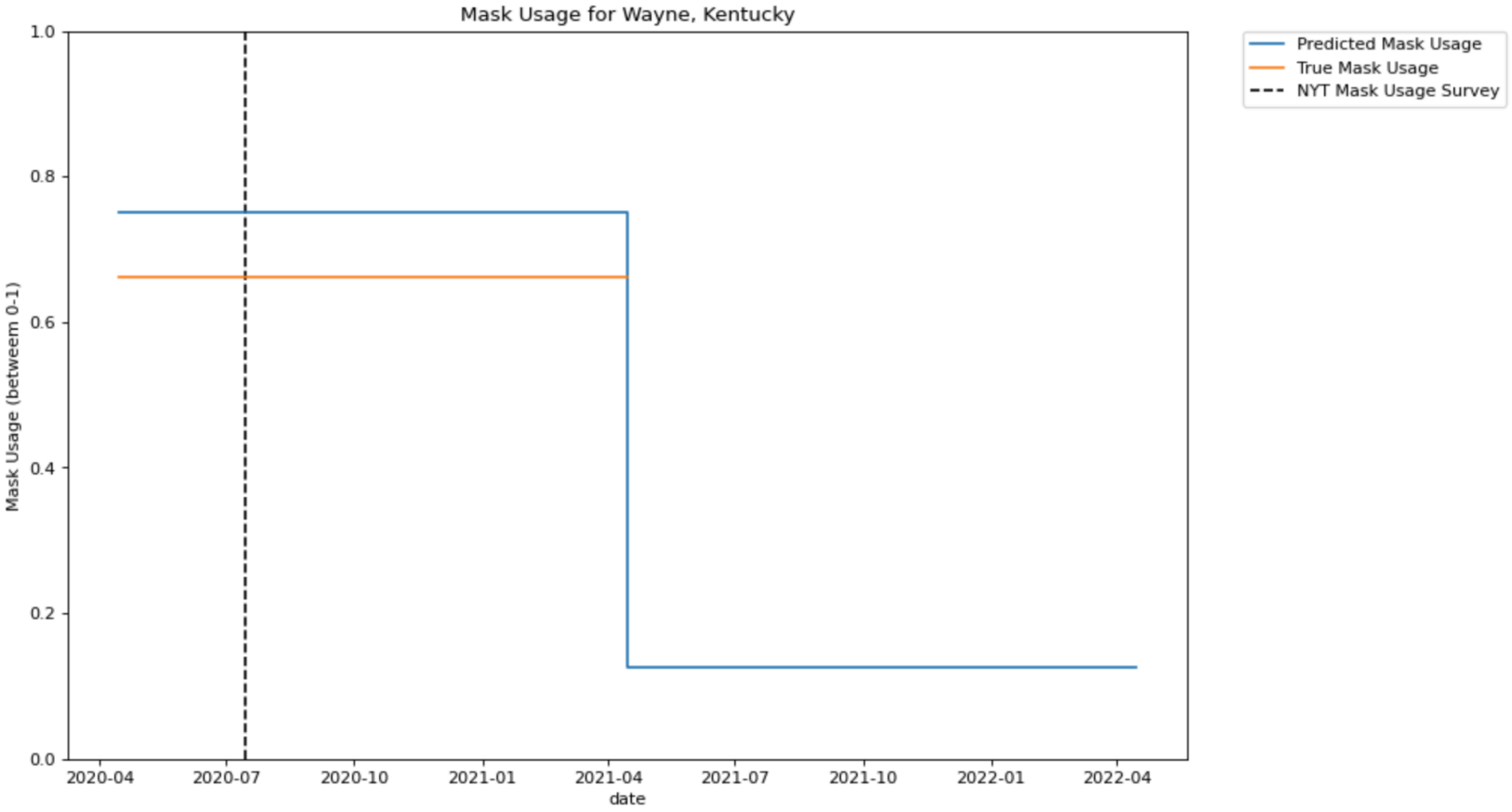


generalized to be able to predict the mask usage of these unseen counties during the training process. I also qualitatively evaluated the model's performance by comparing all counties named Wayne's true mask

compliance to their predicted mask compliance. All counties named Wayne were held out from the model's training dataset. The following graphs and the prior Mean Squared Error analysis prove my hypothesis that using daily COVID cases and whether or not mask mandates are present allows for a relatively accurate prediction of mask compliance in Wayne, Michigan.



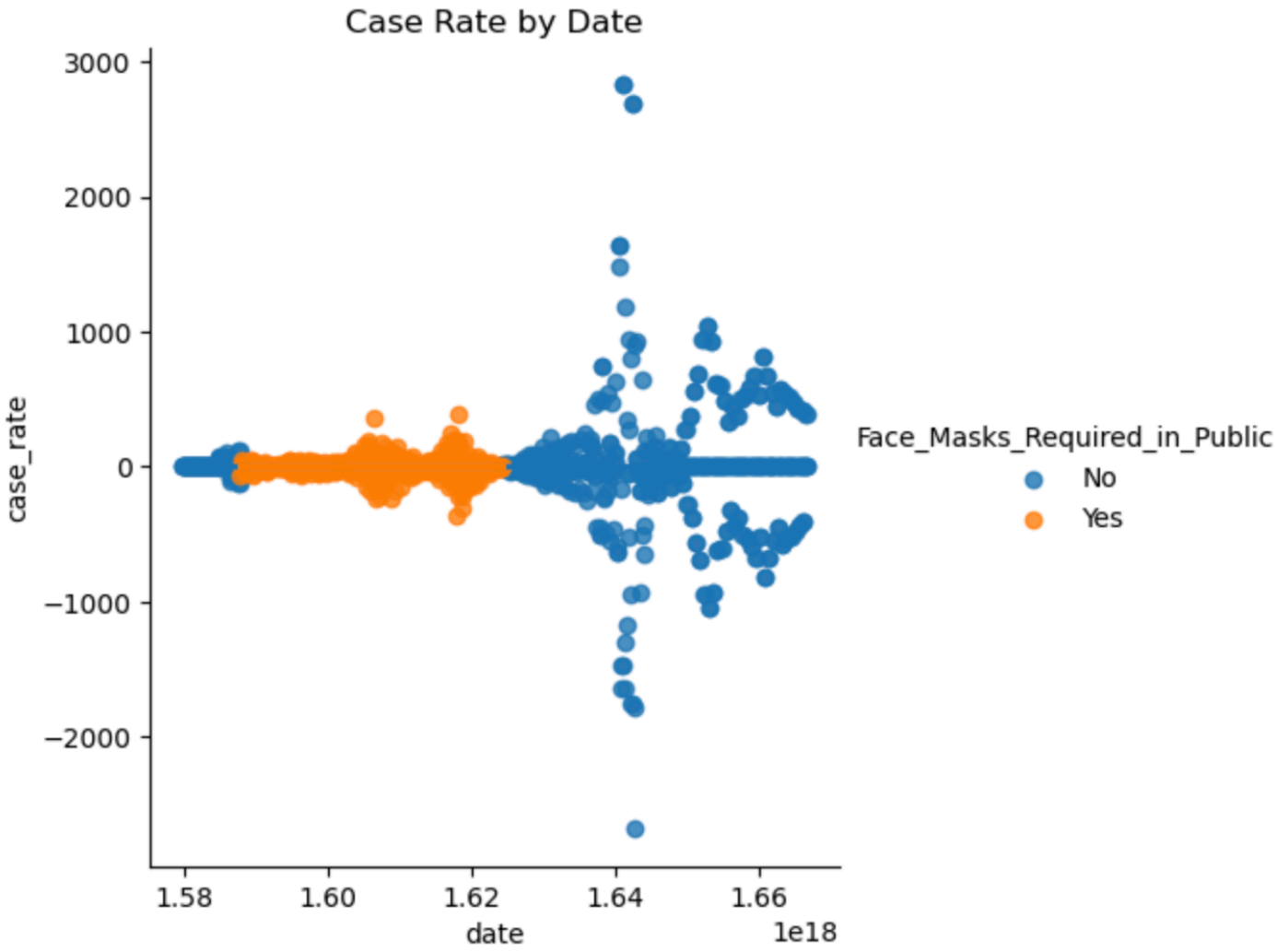




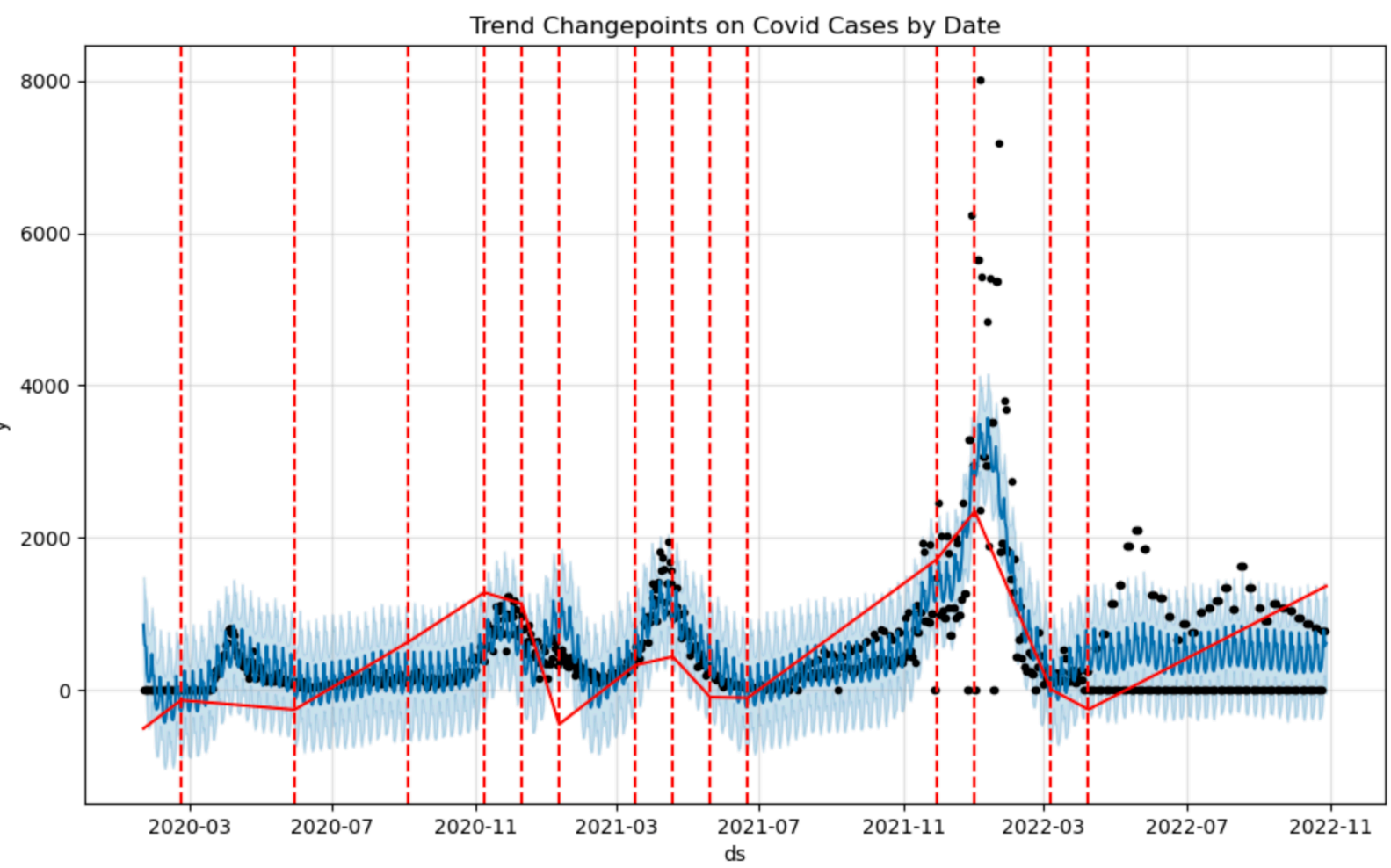
5. **Discussion/Implications**: Why are your findings important or interesting; How could future research build on this study? This section should include a thoughtful reflection that describes the specific ways that human centered data science principles informed your decision-making in this project.

   - For the **common analysis** I found that masking policies change the progression of confirmed COVID-19 cases from February 1, 2020 through October 1, 2021 in Wayne, Michigan. This finding is important in determining the future of mask-wearing and masking mandates. At the start of the COVID pandemic, many people chose to wear masks because, in theory, they would help stop the spread of COVID. However, with years of data behind mask mandates and the effect of those mandates on the spread of COVID, it is reasonable to assume people would require near-conclusive evidence that points towards masks slowing the spread of COVID if those same people are going to adopt mask-wearing again. My common analysis builds towards showing that masks slow the spread of COVID. My finding is based on visually comparing changepoints in new COVID cases to masking mandates. It is important in future research to find the statistical significance of a mask mandate's impact on new COVID cases to build a more compelling argument. It is also important for future studies to look at

other features than mask mandates to verify that there are no confounding variables. I incorporated the human-centered data science principle UX for ML in designing my common analysis. I first tried plotting the gradient of new cases against date and highlighted which days the mask mandate was present. This is the graph that I outputted from this analysis. This graph has



little comprehensibility which goes against this human-centered data science principle as people will gloss over my analysis because it is harder to understand. Therefore, I went a step further in using Facebook Prophet to analyze trend changepoints compared to new COVID cases. This graph is much more readable and follows the human-centered data science sub-principle of explainability. This readable graph is below.



- For the **extension plan** I built a model to forecast mask compliance in Wayne, Michigan to allow people to benefit from the Mask-Wearing Survey in perpetuity. This human-centered task allows those living and traveling to Wayne, Michigan to learn the rates of mask usage without necessitating updated mask compliance surveys. Henceforth, residents and visitors of Wayne can properly calculate their risk tolerance for COVID given their age and preexisting conditions. Mask compliance surveys are expensive, hard to implement and require public participation which decreases as people care less about COVID. Through the use of publicly available data such as daily COVID case count and mask mandates, we can forecast mask compliance without necessitating these surveys. Future research could be done in improving the ability of my model to forecast mask usage. I only use the features mask mandate and new covid cases as input to my model. If more features are used such as the percent of population vaccinated and the average age of population, it is likely the forecasting Mean Squared Error will decrease. However, I was careful in my modeling to follow the human-centered data science principle of algorithmic fairness and transparency. I did not want to incorporate the features race or gender into my model. This way, race and gender biases are not unintentionally learned by my model in

predicting mask usage of a county. This way algorithmic fairness is guaranteed for those of different races and genders. I am also open in my documentation regarding the features I used as well as the architecture of my model. That way, critiques of my design can easily point to my approach and methodology. I also considered the human-centered data science principle risk/benefit trade-off. There are predicted mask usage for several counties that are incorrect. However, the benefit of having somewhat flawed forecasts rather than no forecasts outweighs the human-centered ethical consideration of publishing incorrect forecasts that lead people to believe counties are more or less COVID-safe than in actuality.

6. **Limitations**: This is a required section for your report. You should prioritize and list the ones that are most likely to have a significant impact on your results. Specific license issues could be a limitation, depending on what data sources you used. Flaws in the data, data cleaning techniques, potential assumptions and/or how you handled missing values could be a limitation. Statistical techniques often have specific assumptions and preconditions; if you're not certain all of the data meets those requirements - this is a good place to make that clear.

    - For the **common analysis** there are a few key limitations. A big limitation in comparing daily new COVID case changepoints to mask mandates in Wayne, Michigan is there are only four points where mask mandates are implemented or reversed. Although these mask mandates do line up with the COVID changepoints, a sample size of four is too small to be certain about the analysis. I also had to smooth the data by taking a rolling average over a three-day window because on some days COVID case counts were not reported. Although this mitigates this data error, my analysis would be more accurate if COVID case numbers were accurately reported.

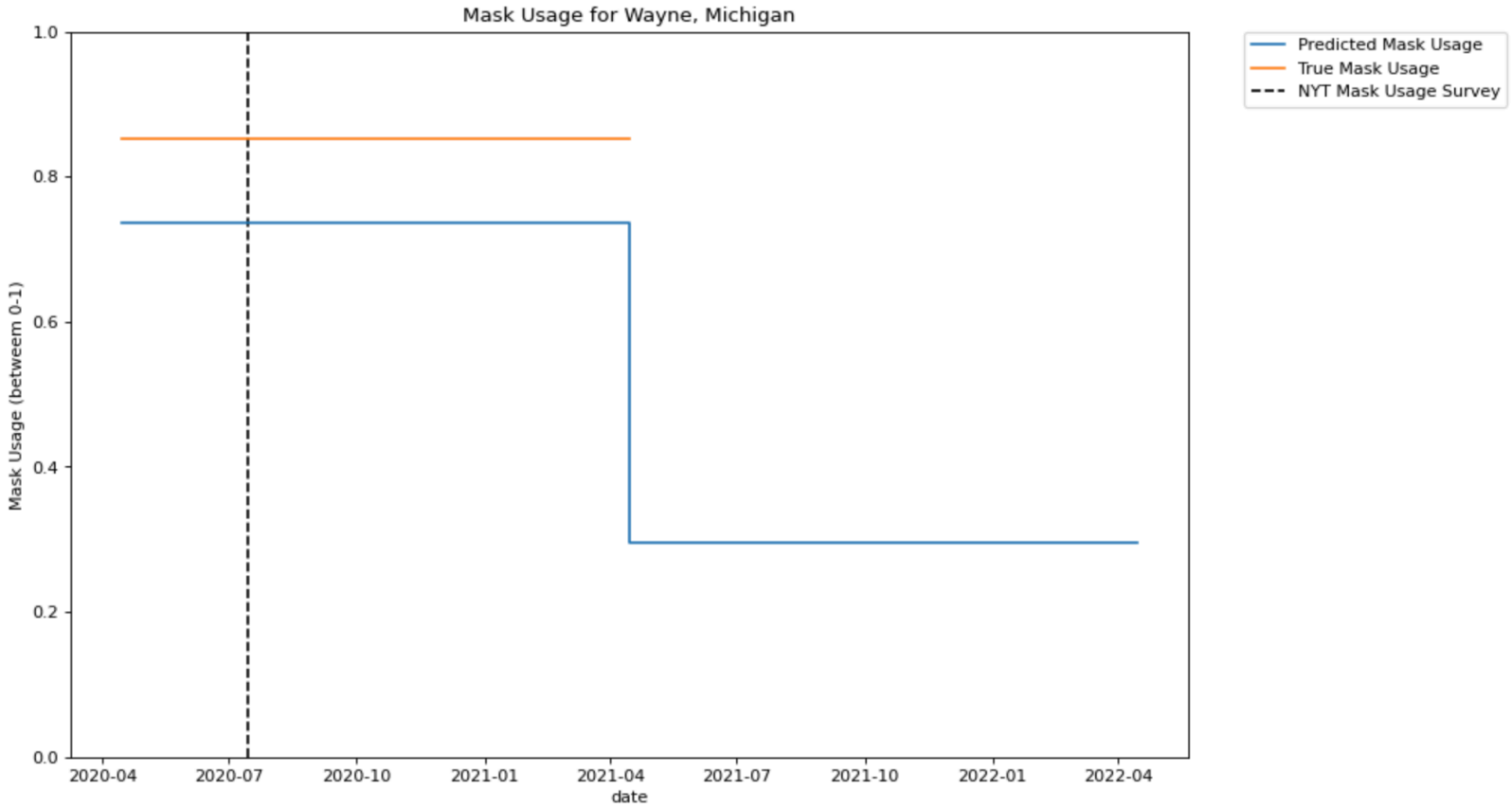
    - For the **extension plan** I assumed that the Mask-Wearing Survey was relevant for an entire year. This assumption is assuredly problematic as mask-wearing habits can change drastically within a years time. However, this assumption made the analysis simple as I could train the model on a year of data, then forecast a year out from the last day the model was trained. Another limitation of my analysis is although measuring the model success using Mean Squared Error from a held-out validation set of counties shows generalizable learning, I have no idea how small a Mean Squared

Error is necessary for the model to be publicly useful. Users of my model need to have trust in it, and if my model says everyone is wearing face masks when in reality nobody is, that is problematic for the users of my model who expect it to help keep them safe from COVID. Lastly, there is assuredly data drift when we are dealing with forecasting. Although signal from my features for one year helped accurately predict mask usage, the same signal for the following year may tell us something different about mask usage. It is impossible to verify if my forecasts are accurate as the New York Times did not do a repeat of the mask-wearing survey.

7. **Conclusion**: Restate your research questions/hypotheses and summarize your findings. Explain to the reader how this study informs their understanding of human centered data science.

   - For the **common analysis** component, I asked the question how did masking policies change the progression of confirmed COVID-19 cases from February 1, 2020 through October 1, 2021 in Wayne, Michigan. I found that masking policies change the progression of confirmed COVID-19 cases between these dates in Wayne, Michigan. When mask mandates were lifted, a changepoint in daily new COVID cases occurred within a reasonable amount of time and COVID cases increased. When mask mandates were introduced, a changepoint in daily new COVID cases occurred within a reasonable amount of time and COVID cases decreased. This analysis has a real-world human-centered impact on determining the future of mask-wearing and masking mandates. At the start of the COVID pandemic, many people chose to wear masks because, in theory, they would help stop the spread of COVID. However, with years of data behind mask mandates and the effect of those mandates on the spread of COVID, it is reasonable to assume people would require near-conclusive analysis that point towards masks slowing the spread of COVID if those same people are going to adopt mask-wearing again. My finding reassures people that mask mandates truly work in the prevention of COVID in Wayne, Michigan. This finding was only possible and presentable due to the data science principle UX for ML. As explained in the methodology section, I threw out my original analysis as the results were not clear and instead went with Facebook Prophet as a more explainable tool to showcase my results.

- For the **extension plan** I built a model to forecast mask compliance in Wayne, Michigan to allow people to benefit from the Mask-Wearing Survey in perpetuity. I hypothesized that using daily COVID cases and whether or not mask mandates were present would allow me to predict mask compliance in Wayne, Michigan. The model I built showed that it could generalize to data unseen during training as the Mean Squared Error decreased over epochs. Finally, when the data for Wayne, Michigan is run through the model, which was a county that was held out of the training dataset, we get an accurate prediction of mask compliance and a reasonable forecast. The below image is this result. This study informs the reader's



understanding of human-centered data science because it followed several important human-centered design choices. The model followed the design principle algorithmic fairness and transparency. I did not want to incorporate the features of race or gender into my model. That way, race and gender biases are not unintentionally learned by my model in predicting mask usage per county. This way algorithmic fairness is guaranteed for those of different races and genders. Second, the documentation design was purposeful to

explain features and the architecture of the model. That way, critiques of my design can easily point to my approach and methodology. Lastly, I considered the human-centered data science principle risk/benefit trade-off. Several counties have incorrectly predicted mask usage. However, the benefit of having somewhat flawed forecasts rather than no forecasts outweighs the human-centered ethical issue of publishing incorrect forecasts that lead people to believe counties are more or less COVID-safe than in actuality.

8. **References**: A list of publications (blogs, articles, research papers) that you refer to in your text.

- Kollepara PK, Siegenfeld AF, Taleb NN, Bar-Yam Y. Unmasking the mask studies: why the effectiveness of surgical masks in preventing respiratory infections has been underestimated. J Travel Med. 2021 Oct 11;28(7):taab144. doi: 10.1093/jtm/taab144. Erratum in: J Travel Med. 2022 Sep 17;29(6): PMID: 34490465; PMCID: PMC8499874.

- Coclite D, Napoletano A, Gianola S, del Monaco A, D'Angelo D, Fauci A, Iacorossi L, Latina R, Torre GL, Mastroianni CM, Renzi C, Castellini G and Iannone P (2021) Face Mask Use in the Community for Reducing the Spread of COVID-19: A Systematic Review. Front. Med. 7:594269. doi: 10.3389/fmed.2020.594269

- A. Yadav, C.K. Jha, A. Sharan Optimizing LSTM for time series prediction in Indian stock market, Procedia Computer Science, 167 (2019) (2020), pp. 2091-2100, 10.1016/j.procs.2020.03.257

- Direct Multi-Step Forecasting with Multiple Time Series Using XGBoost: Projecting COVID-19 Positive Hospitalization Census for a Southern Idaho Health System Drake Anshutz , Andrew Crisp , James Ford , Onur Torusoglu, Justin Smith https://ceur-ws.org/Vol-2884/paper_108.pdf

9. **Data Sources**: A list of links to the relevant data sources that you used.

- The RAW_us_confirmed_cases.csv file from the Kaggle repository of John Hopkins University COVID-19 data. This data is updated daily. You can use any revision of this dataset posted after October 1, 2022. Liscense Attribution 4.0 International (CC BY 4.0) which means we are free to share

and adapt this data. (https://www.kaggle.com/datasets/antgoldbloom/covid19-data-from-john-hopkins-university?select=RAW_us_deaths.csv)

- The CDC dataset of masking mandates by county. Note that the CDC stopped collecting this policy information in September 2021. (https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i)

- The New York Times mask compliance survey data. The New York Times Company is providing this database under the following free-of-cost, perpetual, non-exclusive license - https://github.com/nytimes/covid-19-data/blob/master/LICENSE. (https://github.com/nytimes/covid-19-data/blob/master/mask-use/mask-use-by-county.csv)