Predicting Social Distancing in the Age of COVID-19

Introduction to Data Science — Final Project

Frances Chen

05.20.2020

Contents

Introduction 1
Problem Statement and Background 3
Data 4
Methodology 5
Model Analysis 8
Results 9
Discussion 12
References 14
Appendix 15

Introduction

Social distancing has played a crucial role in the global response to the Coronavirus (COVID-19) pandemic. Minimizing social interactions and mobility reduces the rate at which the infection spreads. Reducing mobility helps "flatten the curve" so that healthcare systems do not become overwhelmed by high case numbers in order to better treat infected individuals.

However, it remains unclear how well the American public has responded to social distancing policies and guidelines across states. Understanding the predictors for the reduction in social contact and travel is critical to measuring the effectiveness of the policy —especially as these policies remain in effect for an extended period of time and citizens begin to relax adherence to these precautions. Further, population perception and behavioral changes may be helpful to include in epidemiological models in predicting and planning for future COVID-19 infections.

Ultimately, this project aims to use machine learning methods to explore which political attitudes or risk perceptions best predict mobility reduction within states across February and April 2020. Understanding these determinants of mobility reduction will help policy-makers and public health officials to determine which states may be more at risk of high COVID-19 cases due to deviance from social distancing guidelines. My goal is to build a model that can explain the majority of the variation in state mobility reduction. The success of the project will be determined whether the best performing model can determine the top variables of importance to predict mobility using the data between February and April 2020.

This report will first discuss background on the U.S. COVID-19 pandemic leading up to the enforcement of Social Distancing Guidelines on March 15th, 2020 and detail the data sources used in this analysis. I will then present the methodology used to construct a time-variant dataset and de-meaning process used to partial out state fixed effects. Using data before April 11th as training data, I test various supervised machine learning regression models —including the linear probability model, k-nearest neighbors, classification and regression tree, and random forest— to determine which model has the strongest performance to best predict mobility in the last two weeks of April. To conclude, I will discuss the successes and shortcomings of this project, as well as opportunities for improving and expanding the analysis in the future.

Problem Statement and Background

The U.S. response to the Coronavirus disease (COVID-19) pandemic has been poor and chaotic, in part due to poor leadership. President Trump's responses to the pandemic have been riddled with misinformation and false claims, which have contributed to immense delays and confusion in instituting wide-spread preventative measures and testing across the country. National Social Distancing Guidelines were not announced until March 16, 2020, after many states already started enforcing local restrictions. While mobility declined by 48.77% across the United States before and after the Social Distancing Guidelines were in place, mobility has varied greatly across states (Broniatowski, Dredze, and Xu, 2020a).

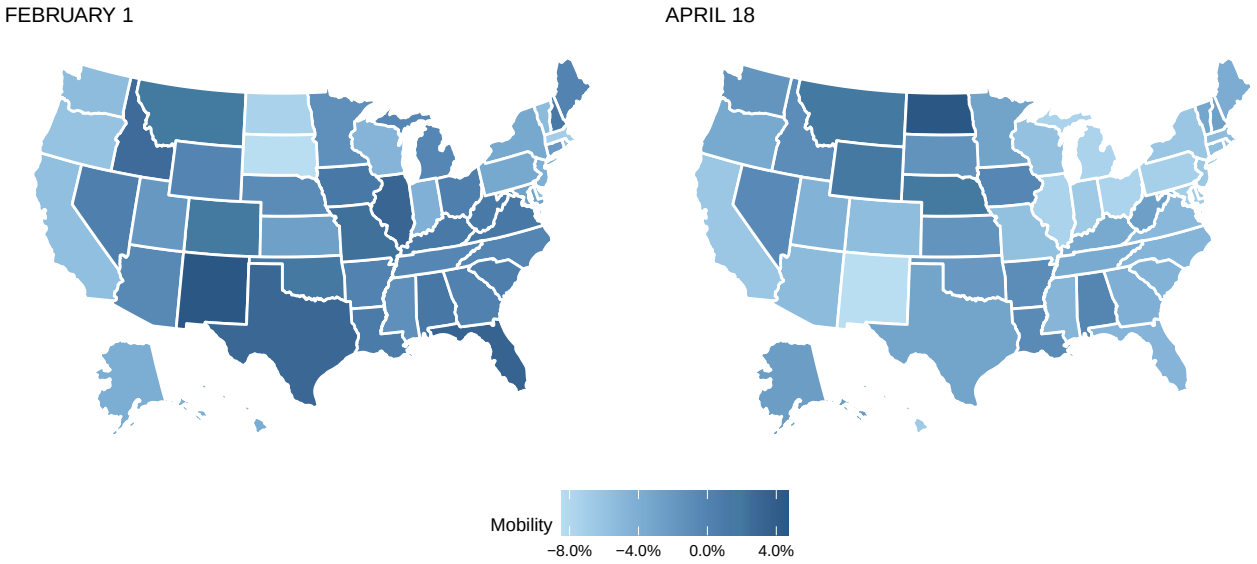

Figure 1: High variation in mobility is evident across states

By initially downplaying the severity of the pandemic, Trump may have imparted a false sense of security and influenced citizens who hold his opinions in high regard to downplay the risks, while citizens who distrust him would be more inclined to follow advice provided by public health officials. This divide in risk perception may be most evidently captured by how closely citizens adhere to social distancing guidelines as measured through their mobility. Given the high variation within state-level mobility across time, it is evident that adherence to social distancing varies greatly across the country.

Capturing population mobility in real-time is a challenging task, especially during an ongoing pandemic. Perfect surveillance is practically and ethically near impossible to achieve. However, in recent years, public social media and mobile device data have been more widely utilized to proxy for mobility measures in public health (Dredze and Paul, 2017). In an effort to track the extent of social distancing practices worldwide, Google's "COVID-19 Community Mobility Reports" use Google data to "chart movement trends over time" across countries (Google, 2020). Unacast's "Social Distancing Scoreboard" utilizes millions of proprietary mobile phone records to track mobility (Unacast, 2020). In a similar vein using Cuebiq's proprietary cell phone data, the New York Times conducted a mobility analysis that showed state populations under state-mandated stay-at-home orders significantly reduced travel, while populations in states that waited to enact restrictions continued to travel frequently (Glanz et al, 2020).

Efforts to capture mobility are critical to assessing the nation-wide preventative practices during this pandemic in addition to understanding the factors that affect Americans' travel decisions and risk assessments. Given the immense health and safety risks associated with mobility during the age of COVID-19, this project aspires to determine which attitudes, perceptions, or other factors best predict citizen mobility within states.

Data

The data in this project is constructed using five main sources:

- **The Twitter Social Mobility Index**: Measure of mobility, social distancing, and travel based on the standard deviation of users' geolocated Tweets each week between January 1, 2019 and March 20, 2020, aggregated over state-levels (see Appendix 1 for detail on how the Twitter Social Mobility Index calucates reductions in mobility). (Broniatowski, Dredze, and Xu, 2020b)

- **Google Trends** `gtrends` package: Weekly number of hits of mobility, political, and

COVID-19-related search terms between February and April 2020. (Massicotte & Eddelbuettel, 2020)

- **The COVID Tracking Project**: Weekly COVID-19 tests, confirmed cases, hospitalizations, and patient outcomes from every US state. (The Atlantic, 2020)
- **United States Department of Labor, Employment & Training Administration**: Unemployment Insurance weekly claims data in each state. Initial claims measure emerging unemployment and continued weeks claimed measure the number of persons claiming unemployment benefits. (U.S. Department of Labor, 2020)
- **COVID Exposure Indices**: Indices derived from anonymized, aggregated smartphone movement data provided by PlaceIQ which describe (potential) exposure varying across locations and time within the United States. (Couture, Dingel, Green, Handbury, and Williams, 2020)

The unit of analysis for this project is state-week, and the main variable of interest is state-level mobility (standard deviation) as measured by the Twitter Mobility Index and re-weighted using mobility and COVID-19 related search term interest between February 1st and April 18th. As the original mobility measure in the Twitter Mobility Index only accounted for mobility until March 29th, an initial machine learning model was used to construct time-variant mobility estimates using risk perceptions captured through mobility-relevant Google search term queries across time.

Methodology

In the intial stages of this project, I intended to use a static model to determine which state-level characteristics most strongly influenced reductions in mobility; however given small sample size concerns, an alternative methodology was developed (see Appendix 2 for more details on the initial methodology). In an effort to address the small sample size, static

nature, and selection bias concerns, an initial machine learning model was used to construct time-variant mobility estimates using risk perceptions captured through mobility-relevant Google search term queries across time. Twenty-five search terms were queried with the `gtrends` package between February 1st and April 18th on a weekly basis (Massicotte & Eddelbuettel, 2020).

Google Search Terms Used to Re-Weight the Mobility Outcome

Transportation and COVID-19 related terms				
coronavirus	cdc	hand sanitizer	shortness of breath	vaccine
social distancing	fever	mask	unemployment	epidemic
hand washing	testing	covid19	stay at home	isolation
quarantine	pandemic	bus schedule	movie times	flights
travel	traffic	hotel	airport	car rental

Figure 2: Terms used to re-weight mobility estimates

A random forest model was then trained on pre- and post-social distancing search interest frequencies and mobility index measures. Predictions of temporally variant measures of mobility were generated using this established model fit and then the estimated measues were inverted to reflect social mobility. Finally, to account for state fixed effects, the mobility estimates were de-meaned by subtracting each state-week observation by the respective state's average mobility across the two-month time period.

The following plot illustrates the resulting time-variant estimates created. As can be observed, there is a drastic downward shift in mobility following the Social Distancing Guidelines announcement on March 15th, as indicated by the dotted line. However, the spread of the gray lines that represent individual states indicate that despite the general decline in mobility, there is significant variation in mobility between states.

After constructing the time-variant dependent mobility variable, the independent variables were compiled and similarly de-meaned in order to account for state fixed effects. The COVID case and device exposure data were imported from respective websites as .csv files. As these datasets were originally collected on a daily basis, I summarized the values on a

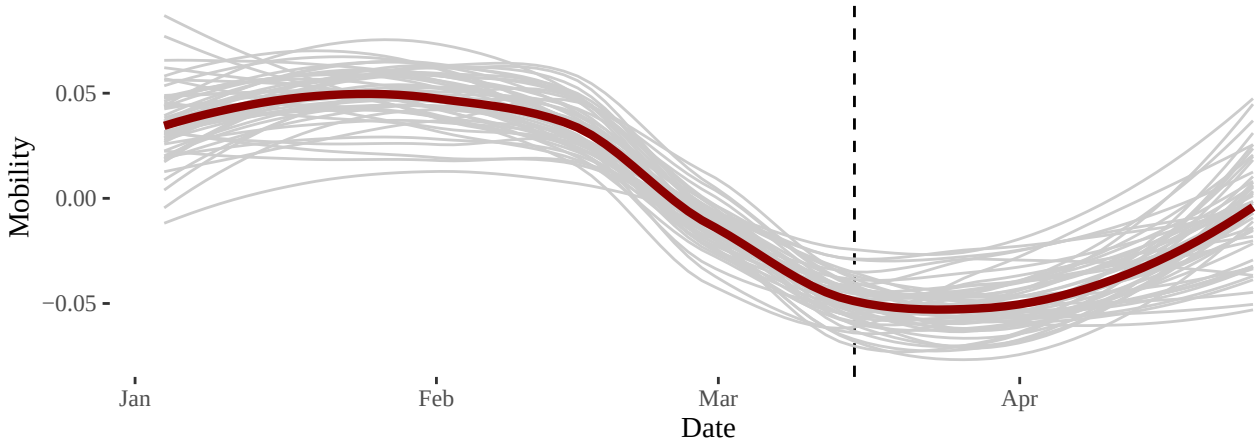

Figure 3: Aggregate mobility drops after social distancing is enforced on March 15th

weekly basis and then de-meaned the values in order the center the data points around zero and control for state fixed effects. Additionally, I decided to drop the variables that had missing observations in the majority of states, such as device exposure indices for Asian and Black populations. The unemployment claim data was downloaded from the U.S. Department of Labor as a .csv file and was already compiled on a weekly basis, so minimal data wrangling was required to tidy the variable names, date class, and de-mean the values.

For the Google Trends search terms that were included with the intention of capturing the population's shifting attitudes and perceptions on a weekly basis, I again utilized the `gtrendsR` package to draw search term frequencies of fifty terms associated with political affiliation, politcally-charged terms associated with COVID-19, and more neutral COVID-19 terms that were not intially used to re-weight the dependent variable (see Appendix 3 for the full list of terms). Similar to the data wrangling process conducted to tidy the Google queries drawn for the dependent variable, this process involved a loop in which each term had to be searched for one at a time in order to avoid being weighted against other terms in the same search. These individual keyword searches were then bound together, reshaped from long to wide format, and rescaled from the 0-to-100 range to 0-to-1. Further, missing values were imputed with zero as "NAs" in Google Trend data indicate the particular keyword do not have a significant number of queries within a given time period to record.

Once data wrangling and processing of each dataset were complete, the independent variables were joined with the dependent variable dataset. The resulting master dataset is composed of 612 observations and 70 independent variables compiled between February and April 2020. In preparation for machine learning analyses, this dataset was subsequently split into training and testing datasets; the training data was composed of state-level observations before April 11th (about 10 weeks; ~83% of the master dataset) while the testing data was composed of observations in the last two weeks of April (~17% of the master dataset). The `recipes` package was utilized to impute missing values using k-nearest neighbors imputation and normalize the scale of the continuous variables to fall between 0 and 1 (Kuhn & Wickham, 2020). Following this pre-processing, I used K-fold cross-validation with 5 folds to partition the data in order to compare model performance using the `createFolds()` function from the `caret` package (Kuhn, 2020). Finally, I used the `trainControl()` function to establish the control settings for each of the models —K-fold cross validation and establishing indices for the folds.

Model Analysis

Supervised learning regression models are ideal to predict future state-level mobility, a continuous dependent variable. For this reason, this project tested the following supervised learning models: the linear regression model (LM), k-nearest neighbors (KNN), classification and regression tree (CART), and random forest (RF). Given the non-linear nature of mobility in the training data, which includes the decline in mobility around March 15th, it is unlikely that the linear regression model, which presumes the relationships betweeen mobility and the included covariates are linear, would provide an appropriate fit. However, considering this model may potentially pick up linear relationships between mobility and certain covariates, it is worthwhile to test the model using the training data.

The k-nearest neighbors model, on the other hand, requires no training and predicts new outcomes using the k-closet neighbors in the given dataset; predicted values are given by

essentially summarizing the output variable for the specified k-nearest neighbors of a given data point. Therefore, new data points do not affect the prediction accuracy within the KNN model. Given the non-linear relationship of the mobility data, I would hypothesize the KNN model may perform better than the LM model that presumes a linear relationship to start. I hypothesize the classification and regression tree model would also perform better as this model is more robust to noise and outliers relative to the LM model. Considering the extent of noise expected from aggregate state-level estimates, I would presume CART models may fit the assumed non-linear relationship well. The final machine learning model to test is the random forest model that combines many decision trees into a single model, which ultimately builds stronger predictions by gathering information from each decision tree. These decision trees are then used to subset observations into predicted groups. RF models hold a strong assumption that mobility may be easily predicted through classification. As each of these four models possess different strengths and weaknesses, I test each on the training dataset to determine which model has the strongest predictive performance.

Results

Predictive Performance

After testing the four models on the training data, the random forest model emerges as the top performer in terms of predicting the mobility, determined by the lowest Root Mean Square Error (RMSE)[1]. The RMSE in the random forest model is 0.12, while the RMSEs for the KNN, CART, and LM models are higher at 0.14, 0.17, and 2.10, respectively. As expected, the LM model performed the worst, producing the largest RMSE of 2.10. The R-squared in the random forest model is 0.79, meaning 79% of the variation in the mobility measure can be explained by the existing independent variables in the model.

I then used the test data from the last two weeks of April to evaluate the RF model's predictive accuracy. The resulting RMSE is 0.127, roughly approximate to the RMSE produced

[1]The RMSE is the standard deviation of the residuals, which measures the prediction errors.

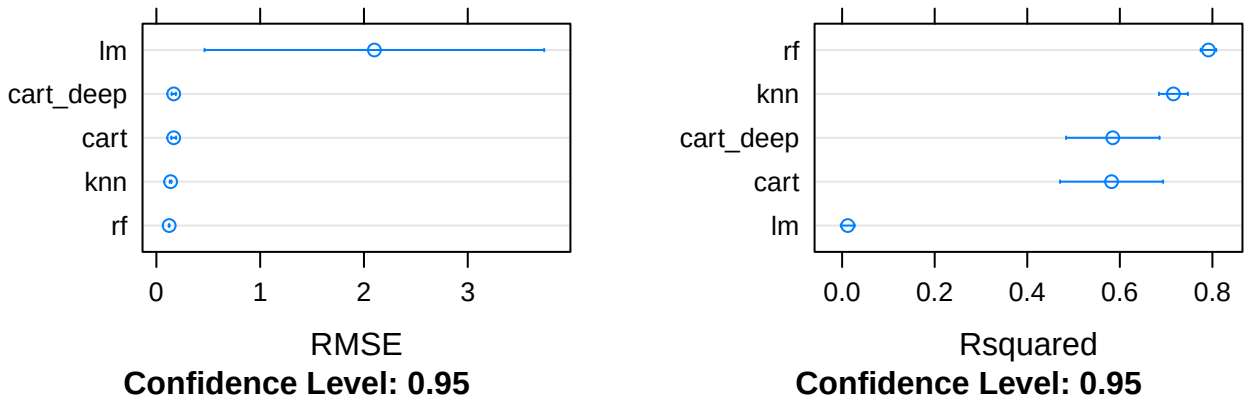

Figure 4: Machine Learning Model comparisons

from running the RF model on the training data. The consistent RMSE value suggests that the random forest model does just as well in predicting state mobility.

Variables of Importance

In the next stage of analysis, I explore the variables of importance. The five most important variables determined to predict mobility via the random forest algorithm are Google search interest in the following keywords, in order of importance ranking: 1. *china* 2. *face masks* 3. *government* 4. *coronavirus deaths* 5. *fauci*

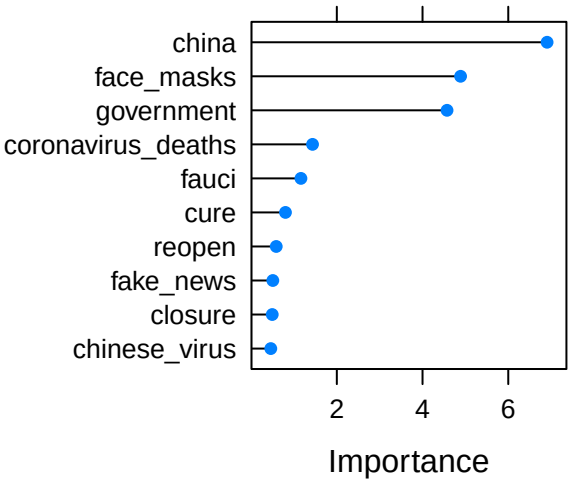

Figure 5: Random Forest: Top 10 Variables of Importance

The results indicate that the frequency of these search terms are the top ten most im-

portant variables in predicting mobility in the random forest model. In plotting the partial dependency plots for the top three most important variables — *china*, *face masks*, *government* — the general trend appears to suggest that above average search interest in these terms is associated with lower mobility. Higher search interest in these pandemic-related keywords possibly implies that state populations that are more likely to be actively concerned about pandemic issues and safety precautions are consequently less likely to be mobile.

However, this trend differs for the other two top search terms, *coronavirus deaths* and *fauci*. Increased search interest in *coronavirus deaths* reflects a slight bump in mobility when search frequency is slightly above average (~60%) and then a gradual increase in mobility as search interest ramps up from the 75th percentile. This gradual increase is similarly modelled in the partial dependency plot for the fifth most important variable, *fauci*, however the increase in mobility is comparably less drastic.

These contrasting trends present interesting potential stories. A potential story may be that state populations with higher than average search interest in *coronavirus deaths* and *fauci* may have more essential workers who must travel for work and may use these COVID-19-specific search terms to inform their own risk assessments. Higher search interest in the public health official Dr. Fauci may reflect potential interest in his advice if mobility is essential. On the contrary, above average search interest in *coronavirus deaths* may be more frequent amongst more urban states with high case rates —such as California, Michigan, or New York— where mobility is simply more frequent due to the urban lifestyle and relatively more dense urban population. Another important consideration is the fact that the original mobility data is based on Twitter activity that may be overrepresented in urban areas to begin with.

It is finally important to note that despite the ranked order of variable importance, even the top most important variable (search interest in *china*) does not appear to register very highly in the importance scale —garnering an importance rank just above 6 while the other top 5 variables rank around 4 or lower. Hence, these findings should be considered with

caution.

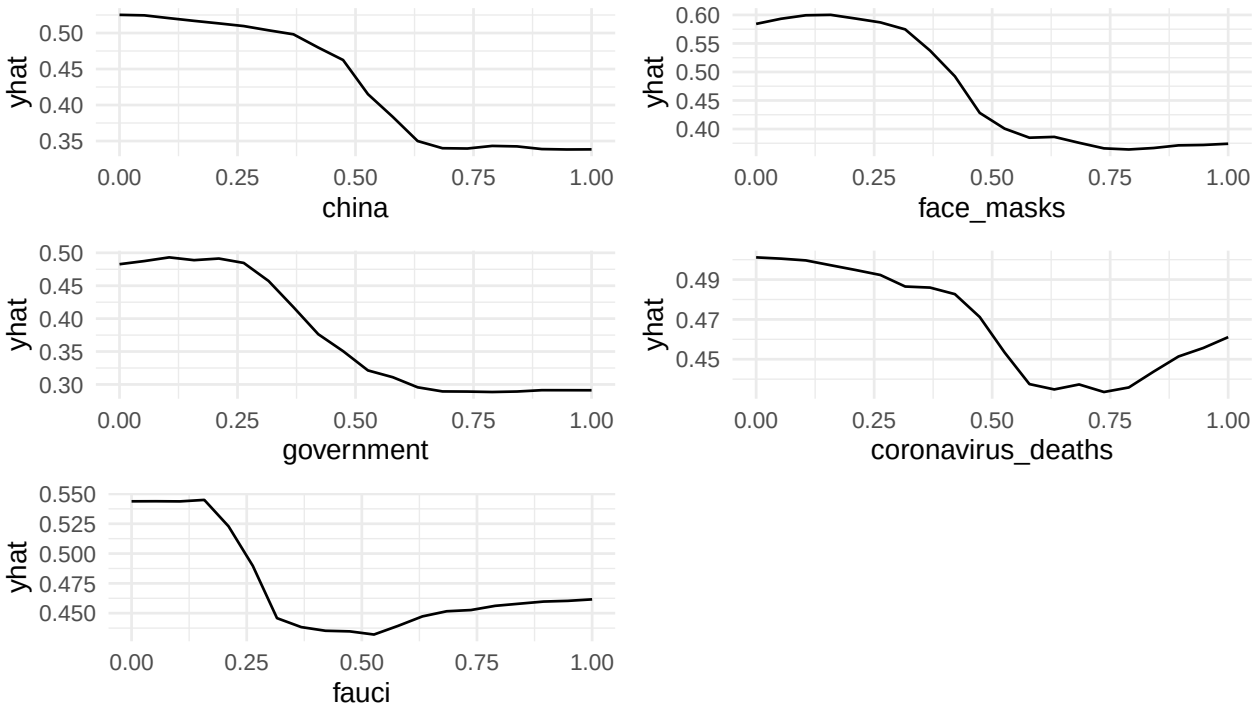

Figure 6: Partial Dependency Plots: Top 5 Most Important Variables

Discussion

As previously noted, the success of this project was to be determined by whether the best-fit machine learning model could identify the factors that best predict mobility. Accordingly, the random forest model was found to have the best fit based on data before mid-April (the training data) and explained 79% of the variation in state-level mobility when used to predict mobility in the last two weeks of April. The model then succeeded in determining the top variables of importance in the model —keyword *china*, *face masks*, *government*, *coronavirus deaths*, and *fauci* Google search interest. Thus, the project was successful in achieving its aims.

Despite having achieved the project aims, this project presents many opportunities for improvement. To start, curating a more representative, less biased mobility measure besides the Twitter Mobility Index that was used would improve the accuracy of this project of

capturing actual mobility. However, due to public data availability at the time of this project, the Twitter Mobility Index measure was the best mobility measure that was freely accessible. Expanding the unit of analysis to city or county level would also greatly increase the sample size for the machine learning models to train on, if such granular-level data were available. I had also previously compiled 200+ state-level independent variables to use as state controls, but ultimately decided to exclude these static variables in the final analysis. Finally, I had hoped to add more relevant Google search queries to capture a more accurate representation of the population's mobility and political attitudes, but was constrained by the `gtrends` 1000-queries per day limit and unexpected technical challenge when the package stopped working in the midst of the project.

As the current model only explains about 78% of the variation in mobility, there is great opportunity to expand the breadth of independent variables for the model to learn on to achieve greater explanatory power and more accurate predictions. If given more time, I would expand the volume of relevant search terms to include as independent variables. Further, I would continue to explore more time-variant independent variables that could be added to the model —such as variables that indicate when states had local shelter-in-place madates or food security measures. Further exploratory analysis would be greatly beneficial for public health official and policymakers for assessing the factors that determine American citizens' overall adherence to preventative measures such as social distancing during the age of COVID-19.

References

Broniatowski, Dredze, and Xu. (2020a). The twitter social mobility index: Measuring social distancing practices from geolocated tweets. Retrieved from https://arxiv.org/pdf/2004.02397.pdf

Broniatowski, Dredze, and Xu. (2020b). Twitter social mobility index. Retrieved from http://socialmobility.covid19dataresources.org/index

Couture, Dingel, Green, Handbury, and Williams. (2020). COVID exposure indices. Retrieved from https://github.com/COVIDExposureIndices/COVIDExposureIndices

Dredze and Paul. (2017). *Social monitoring for public health. Synthesis lectures on information concepts, retrieval, and services.*

Glanz et al. (2020). Where america didn't stay home even as the virus spread. Retrieved from https://www.nytimes.com/interactive/2020/04/02/us/coronavirus-social-distancing.html

Google. (2020). COVID-19 community mobility reports. Retrieved from https://www.google.com/covid19/mobility/

Kuhn, M. (2020). *Caret: Classification and regression training.* Retrieved from https://CRAN.R-project.org/package=caret

Kuhn, M., & Wickham, H. (2020). *Recipes: Preprocessing tools to create design matrices.* Retrieved from https://CRAN.R-project.org/package=recipes

Massicotte, P., & Eddelbuettel, D. (2020). *GtrendsR: Perform and display google trends queries.* Retrieved from https://github.com/PMassicotte/gtrendsR

The Atlantic. (2020). The covid tracking project. Retrieved from https://covidtracking.com/

Unacast. (2020). Social distancing scoreboard. Retrieved from https://www.unacast.com/covid19/social-distancing-scoreboard

U.S. Department of Labor. (2020). Unemployment insurance weekly claims data. Retrieved from oui.doleta.gov/unemploy/claims.asp

Appendix

Appendix 1 — The Twitter Social Mobility Index

- `mobility_before_distancing` The social mobility index for the entire time period of the dataset, from January 1, 2019 up to the date before social distancing, March 15, 2020.
- `mobility_after_distancing` The social mobility index since March 16, 2020. Note that social distancing started at different times for different states, but we don't include that analysis in this data.
- `reduction` The percent reduction of social mobility.

$$\text{Mobility Reduction} = 1 - \frac{\text{mobility after social distancing}}{\text{mobility before social distancing}} \tag{1}$$

Appendix 2 — Initially Proposed Methodology Details

For the methodology intially proposed in the project proposal, I compiled over 200 state-level demographic, political, healthcare system, and COVID-19-related variables[2] with the intent to use these variables as controls in the static model. However, there were many methodological concerns with static model and the Twitter Mobility Index as the outcome variable. Due to the time invariant nature of the original Twitter Mobility Index, this outcome variable limited the sample size to only 50 state observations. Further, there is a serious selection bias concern with utilizing social media data to proxy for state-wide mobility. For instance, more urbanized states with younger populations are more likely to be active on Twitter and would therefore be overrepresented in the sample. Given these misgivings, an alternative methodology was developed.

[2]In this initial model, I scraped current state-level political party affiliation data. However, this data was later unused due to the shift to the time-variant model.

Appendix 3 — Model 2 Google Search Terms

Google Search Terms Used to Re-Weight the Mobility Outcome Transportation and COVID-19 related terms

coronavirus	cdc	hand sanitizer	shortness of breath	vaccine
social distancing	fever	mask	unemployment	epidemic
hand washing	testing	covid19	stay at home	isolation
quarantine	pandemic	bus schedule	movie times	flights
travel	traffic	hotel	airport	car rental

Figure 7: Keywords included as covariates