Aubrey Brockmiller (alb3cb)
Ian Kloo (ipk8gh)
Shawn Michanco (sbm5rg)
Akeem Wells (ajw3rg)

DS 5559- Big Data Analytics
May 2021

ANALYSIS OF WEATHER EFFECTS ON SPREAD OF COVID-19 IN THE UNITED STATES

ABSTRACT

Over a year into the COVID-19 pandemic, it is still unclear if and how weather affects the spread of the SARS-COV-2 virus. Many believed the pandemic would not thrive in areas with warmer weather, while others thought there would be some seasonality to the virus like the existing respiratory viruses that spread across the country every winter. In reality, COVID-19 has spread to every corner of the earth and has waxed and waned under various weather conditions in a way that leads to no obvious conclusions about the ways the weather interacts with this virus. This study uses a data-driven machine learning approach to model reported COVID-19 infections in United States counties using weather and mobility data. Our project explored four models before conducting a grid-based hyperparameter tuning process on our best performing model, Gradient Boosted Trees. Our resulting model's R^2 value is approximately 0.35, and although this suggests limited predictive ability, this was expected because we only considered weather and mobility data when there are undoubtedly many other important variables that affect COVID-19's spread. While our models are not particularly useful for prediction, we found a strong association between temperature and COVID-19 spread, even when controlling for population mobility. This study should serve to inform future research into the ways that the weather affects respiratory viruses and presents several concrete recommendations for focus areas based on the results of this work.

INTRODUCTION

The years surrounding and encompassing the outbreak of the global COVID-19 pandemic will undoubtedly define this period in history. The COVID-19 pandemic has been characterized by tragedy and uncertainty, but has also resulted in massive data collection efforts that create an opportunity to solve complex problems in the public interest. There are opportunities to leverage these new and growing data sets to answer questions about things like controlling the spread of a virus, investigating the financial turmoil, and understanding human behavior.

Since the beginning of the COVID-19 pandemic, researchers and commenters have pointed to similarities between this virus and seasonal contagions like influenza. Several prominent research efforts found evidence that weather might play an important role in the way SARS-COV-2 spreads through the population (Gupta et al., 2020) (Tosepu et al., 2020). These papers, however, do not account for omitted variables and seek to simply prove the existence of correlations between weather and the spread of COVID-19. While there are many potential omitted variables that could affect COVID-19's spread and be associated with weather, we will focus on the hypothesis that weather could make it more likely for people to want to leave their houses and encounter others socially. Misattributing COVID-19's spread to weather when it actually is being driven by population mobility could be potentially damaging from a public health standpoint because things like holidays also drive social behavior and are uncorrelated with weather (there are holidays throughout the year). Our paper will address this issue by directly controlling for increased mobility in a population.

In addition to omitted variable bias, the existing papers studying COVID-19 and weather used data at the state and country levels. The research presented in this paper takes a much more data intensive approach, using Pyspark's machine learning capabilities to allow us to examine potential relationships between COVID-19 and weather data at the county level. We expect this increased data granularity will result in much more reliable results.

It is important to note that, much like the previous research in this space, this study will not be able to prove a direct causal relationship between specific weather features and viral behavior (e.g., cold air might lengthen viral viability on surfaces). Studies about viral behavior would be better informed by laboratory research instead of relying on population data analysis where there are myriad potential omitted variables. Instead, we will attempt to find weather features that are associated with changes in viral transmission after controlling for population mobility, regardless of the specific mechanism. The results of this study could have implications for future public health campaigns against pandemic or seasonal threats (e.g., influenza).

DATA DESCRIPTION

As previously discussed, there is a huge amount of COVID-19 data available. Data science in the early days of the pandemic was characterized by disjoint data sets at various levels of granularity and varying levels of quality. Fortunately, USA Facts now maintains an extremely reliable and publicly available COVID-19 data set at the day/county level:

1. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

When considering the most appropriate dependent variable that measures growth in COVID-19 spread, we initially considered using change in cases from the previous day. This metric, however, would not account for population size and other potentially important county-level factors. To mitigate these issues, we decided to use a measure of COVID-19 infection called infectious probability for our dependent variable. Infectious probability is an estimate of the percent of people who are currently infectious with COVID-19 in a given area that is based on reported cases but also accounts for the ways the virus interacts with county-level demographics and population size (Evangelista et al., 2021). The authors of the referenced paper use USA Facts as their data source, but make their processed data with the infectious probability metric available via Github (https://github.com/iankloo/bigmap).

For our weather features, we used National Oceanic and Atmospheric Administration (NOAA) historical data at the county level. This data is accessible in many formats from NOAA's website:

2. https://www.ncei.noaa.gov/news/noaa-offers-climate-data-counties

While NOAA's data is well-maintained, it is provided at the station level where each weather station has an associated latitude and longitude. To aggregate this data to the county level, we found the centroids of each county in the United States, found all weather stations within 50 miles, and aggregated the reported weather data. This should give a better characterization of weather in a county than taking a single weather station's readings. In some cases, there were no weather stations within 50 miles of a county centroid, so we added 50 miles to the search radius until we found valid data.

Our team accessed the NOAA data through the GSODR R package which contains functions for extracting nearby weather stations to a given latitude and longitude. (https://cran.r-project.org/web/packages/GSODR/index.html). More on this data processing effort is presented in the attached Jupyter Notebook titled "r_data_acquisition.ipynb."

3. https://www.google.com/covid19/mobility

Our final data source was Google's Mobility Reports which uses cellular phone data to characterize the percentage of normal population mobility in a given area. For example, numbers over 100% suggest more mobility than an average (non-pandemic) year. This data is provided at the county/day level.

METHODS

The full modeling process described in this section is available with additional comments and discussion in the attached Jupyter Notebook called "pyspark_cleaning_modeling_vis.ipynb."

DATA IMPORT AND PREPROCESSING

After extracting the data from the sources mentioned in the previous section, the files were merged by county and day and saved as a CSV (attached as "cov_weather_mobility.csv"). The script used to create this file is attached as "r_data_acquisition.ipynb." The final CSV file was loaded into Pyspark. Without any filtering, this data contains 1,221,090 rows with each row representing a single county on a single day.

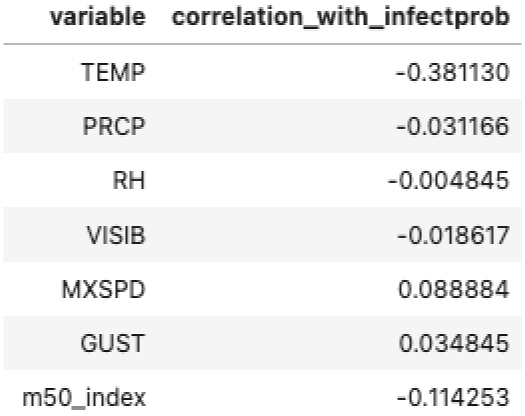
Our first preprocessing step was to limit the date range from June 01, 2020 to March 01, 2021. We expect COVID-19 case reporting was more consistent in these ranges since it excludes the early days of the pandemic when databases were not well established as well as more recent data that is often modified due to reporting corrections. Next, we filtered out some large outliers and missing data in the mobility and infectious probability metrics. Infectious probability outliers in the COVID-19 data reporting issues came from the chaotic data reporting environment that has defined the pandemic, at times. The source of the extreme values in the Google data is unknown, but we can accept that the process of aggregating cell phone GPS data is likely prone to errors. After these filtering steps, we were left with about 660,000 rows to train and test our models.

FEATURE SELECTION

Because we found it plausible that any of our weather metrics could be associated with COVID-19 transmission, we chose to use all of the weather and mobility information features from our dataset, allowing the models to determine the most influential features. A description of the weather and mobility features as they are coded in the data is as follows:

- TEMP = temperature
- PRCP = precipitation
- RH = relative humidity
- VISIB = visibility
- MXSPD = max wind speed
- GUST = wind gust speed
- m50_index = percentage of normal mobility in an area

EXPLORATORY DATA ANALYSIS



variable	correlation_with_infectprob
TEMP	-0.381130
PRCP	-0.031166
RH	-0.004845
VISIB	-0.018617
MXSPD	0.088884
GUST	0.034845
m50_index	-0.114253

Figure 1: Correlations between all independent variables and COVID-19 infectious probability

Because the aim of the project was to analyze weather correlation to COVID-19 spread, our exploratory analysis first looked at correlations between all of our independent variables and infection probability (Figure 1). It looks like temperature has the strongest (negative) correlation with COVID-19 infection, but this number does not suggest a strong linear relationship between temperature and infectious probability. Mobility also interestingly has a negative correlation, albeit with a much lower magnitude. This could be because restrictions on movement tend to correspond with increased infections.

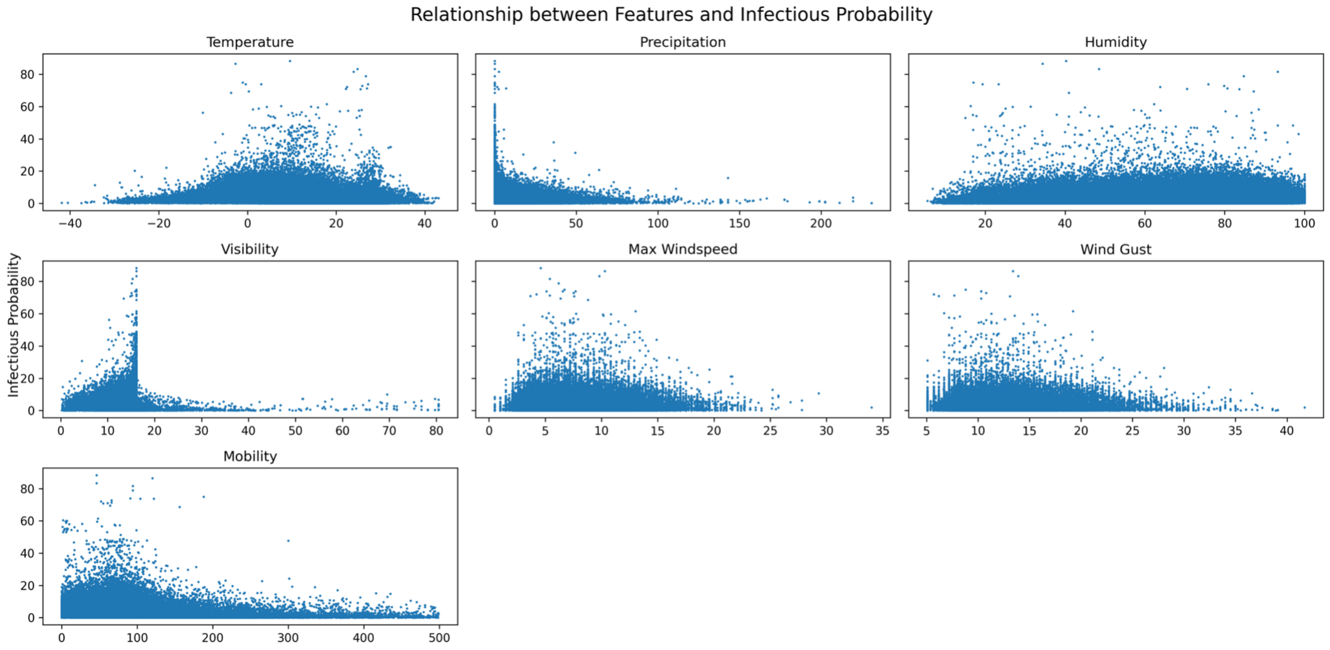


Figure 2: Scatter plots of all independent variables against infectious probability

To examine any non-linear relationships between these variables, we next created scatter plots of each variable against infectious probability (Figure 2). These plots do not reveal any obvious associations between these variables and infectious probability. There appear to be higher values in infectious probability in the middle/right part of the temperature scale, but this is likely due to the fact that these are the most common temperatures in the United States.

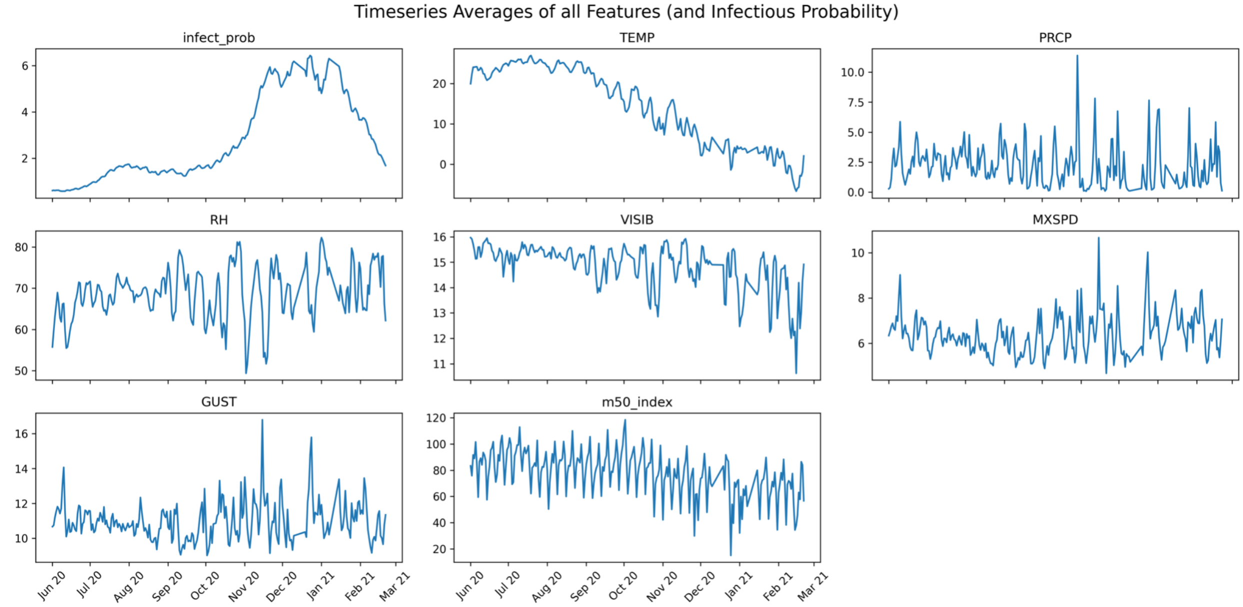


Figure 3: Time series plots of all independent variables along with infectious probability

Finally, Figure 3 examines potential similarities between time series plots of all of the variables in our model (including infectious probability). It is clear that the United States saw an increase in COVID-19 in the summer, but the biggest spike was in the winter. None of the independent variables mirror this COVID-19 data in any direct way. One could argue that temperatures fell as COVID-19 reached the highest peaks, but the trajectories of these time series are very clearly different.

All of this analysis suggests that there are no clear linear or non-linear relationships between our independent variables and infectious probability, at least in a univariate sense. This suggests that tree-based machine learning methods are good candidates for evaluation since they do not rely as heavily on the existence of linear relationships when compared to linear regression models.

FEATURE ENGINEERING

COVID-19 is an infectious disease that has an incubation time, so we would expect that any association between changes in weather (or mobility) would not appear in the COVID-19 reporting until it has had time to cause symptoms, prompting people to get tested. According to the CDC, COVID-19 has an incubation period between 3 and 14 days with a median of 5 (Center for Disease Control and Prevention, 2021). To address this issue, we created lead variables for infectious probability at intervals of 3, 7, and 10 days. We found the 7-day lead to be the most effective by a significant margin. The infectious probability metric referred to for the remainder of this paper is actually the infectious probability from 7 days beyond the described weather and mobility changes.

In addition to leading the infectious probability variable, we also log transformed it. We took this step after training some linear regression models and noticing some non-random behavior (homoscedasticity) in the residuals.

We did not find any transformations to be helpful with the independent variables, but we did scale them using Pyspark's "StandardScaler" to create features that are suitable for regression tree modeling. As part of this process, we also create a single feature vector that we used while training our models.

DATA SPLITTING/SAMPLING

In the interest of training many models and performing hyperparameter tuning in a relatively short period of time, we chose to downsample our data by half. We stratified the downsampling by county and sorted by time to make sure we did not miss or overrepresent any areas and times. The resulting sample contained ~210,000 rows. We then split the downsampled set into a 70/30 test/train split.

MODELING

Our first goal was to determine what class of models gives us the best results. The models were fit on the downsampled, scaled, training data and the resulting models were evaluated using the downsampled, scaled, testing data. To compare the results, we looked at the resulting Root Mean

Squared Error and R^2. In the interest of time, we chose to only tune hyperparameters for the model class that showed the most promise from these initial evaluations.

BASE MODEL

Our base model was a simple multiple linear regression model. We built the model with the pyspark.ml.regression.GeneralizedLinearRegression package. We did not anticipate this model to perform especially well, but it served as a good starting point and allowed for a base to work the other models off. We fit this model with all the weather and mobility features.

RANDOM FOREST

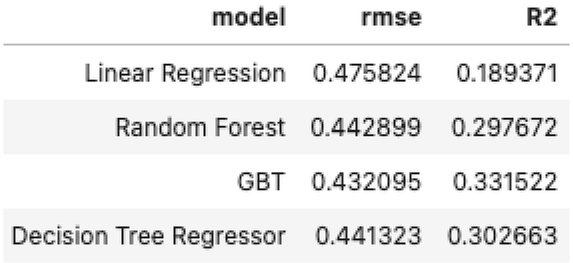
We built a random forest model utilizing the pyspark.ml.regression.RandomForestRegressor package. We fit this model with all the weather and mobility features, like the base model. We anticipated a better result than the base model provided.

DECISION TREE REGRESSION

We also built a decision tree regression model utilizing the pyspark.ml.regression. DecisionTreeRegressor package. Again, this model was fit with all the weather and mobility features. This model performed similarly to the Random Forest model.

GRADIENT BOOSTED TREE REGRESSION - The Champion

We built a gradient boosted tree model (GBT) utilizing the pyspark.ml.regression.GBTRegressor package. Again, this model was fit with all the weather and mobility features. This model was the best performer, so we chose to tune the hyperparameters and move forward with GBT.



model	rmse	R2
Linear Regression	0.475824	0.189371
Random Forest	0.442899	0.297672
GBT	0.432095	0.331522
Decision Tree Regressor	0.441323	0.302663

Figure 4: Model Evaluation Results

TUNING GBT HYPERPARAMETERS

After selecting GBT as our preferred modeling framework, we explored potential hyperparameters using a tuning grid with 4-fold cross validation. We explored different values for maximum iterations as well as maximum tree depth. In general, increasing the iterations and tree depths resulted in improved RMSE and R^2 values; however, we found diminishing returns when increasing the maximum depth beyond 5 and the maximum iterations beyond 100. While exploring these parameter grids with values larger than this, we encountered days-long runtimes and the resulting models performed similarly to those with maximum depth of 5 and maximum iterations of 100. These are the values we used for our final model training.

After tuning the hyperparameters, we improved the R^2 for our GBT model trained on the downsampled data to 0.349.

For our final model that we will use for the remainder of this paper, we trained a GBT model with maximum depth of 5 and maximum iterations of 100 on the full (not downsampled) data, again split into test and train samples. Because of our careful downsampling strategy, we did not expect significantly different results, but wanted to test this assumption. We found similar R^2 results with this final model compared to the downsampled data (~0.35).

RESULTS AND DISCUSSION

In an attempt to understand which features were driving the results of the GBT model, we created an importance plot (Figure 5).

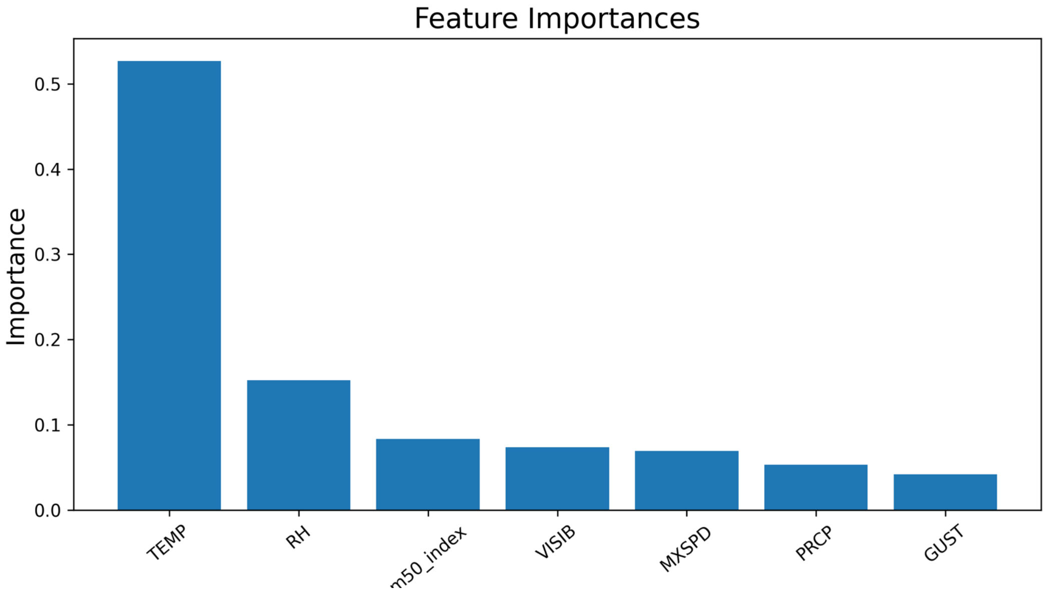


Figure 5: Feature importance in our champion model

Temperature is the most important feature in our model by a significant margin. We expected that temperature might be an important feature, but thought that it’s association with COVID-19 spread might be due to the fact that people move around more in nice weather. This plot shows that mobility (m50_index) was relatively unimportant and that temperature seems to have some association with COVID-19 spread after controlling for mobility.

We expected that the wind variables and humidity would have some associations because the relationship between a respiratory virus and wind is easy to visualize. We did not find evidence in support of this assumption.

To visualize the predictions made by our model, we mapped infectious probability along with our predictions for a single county in Figure 6. We chose Cook County which contains Chicago because it has had fluctuations in COVID-19 throughout the pandemic.

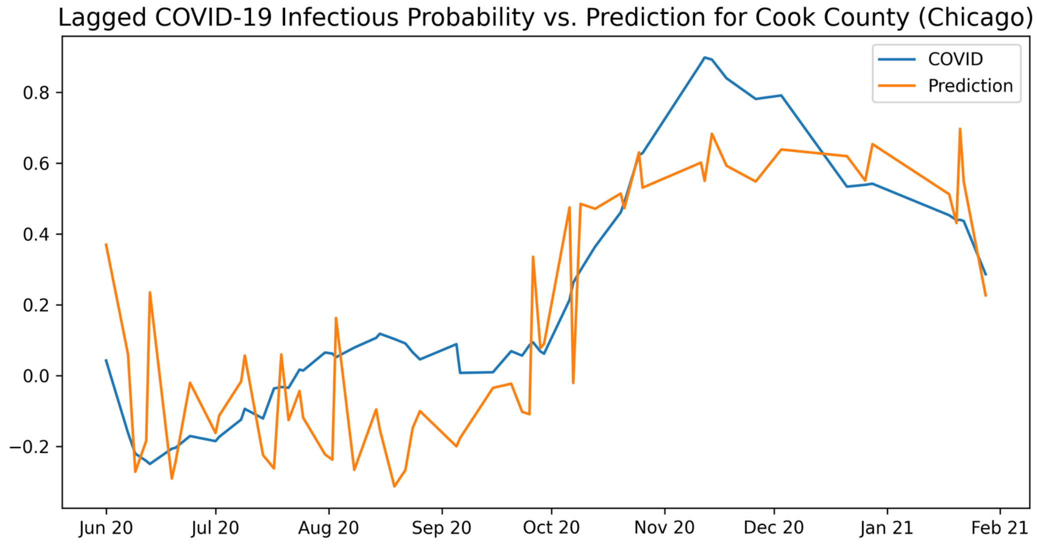


Figure 6: Actual infectious probability with our model's predictions for Cook County

Our method is causing predictions that jump around quite a bit, but we do seem to map the general trajectory of COVID-19 in the county. This suggests that we should use a smoothing factor in our predictions (more on this in the following section). We also notably lag the spike by a bit, but we do eventually catch up. This spike in COVID cases is around the time of the holidays. We hoped our mobility data would account for this, but it does not seem to help.

Our model makes predictions that are somewhat erratic, but there does appear to be at least some interesting association between temperature and COVID-19 spread. Ultimately, our group was more interested in using this as a kind of screening experiment to see if weather features were related to COVID-19 spread at all - accurate prediction was not ever really the goal. We believe this project presents more than enough evidence to suggest COVID-19 and temperature might be related in a way that goes beyond people tending to move around more in warmer temperatures.

The modeling effort presented in this paper furthered the academic knowledge about the potential relationship between weather and the spread of COVID-19 by using a finer data granularity and including mobility data (a key omitted variable in other studies).

FUTURE WORK

The findings presented in this paper suggest several interesting areas for future modeling efforts. The first, and perhaps most obvious area for future study concerns including additional omitted variables that affect COVID-19 spread. For example, data on mask wearing and local lockdown policies should be considered in future studies. Our group also discussed adding the political affiliation of each state, potentially using one hot encoding, to the dataset to control for other COVID-19 related mandates that could only be enforced on the state level. We hypothesize that political affiliation would also encode some of the local attitudes toward the virus that would be difficult to measure directly.

Beyond simply adding more variables, we could improve our existing model by incorporating time series methods that allow for things like smoothing factors. Our group did some preliminary research into existing time series machine learning models, but found it to be outside of the scope of this course project, especially since we did not find previous work on time series machine learning in Spark in our (admittedly short) search.

In addition to time series methods, this problem could also benefit from a hierarchical modeling framework like the ones taught in the Bayesian Machine Learning course in this department. The idea would be to have some commonality in the parameters that cause COVID-19 to spread but allow those to vary across the different counties to some degree to account for things like different attitudes toward the pandemic across the country. We were only familiar with hierarchical models in Bayesian frameworks, which typically do not perform well when working with large datasets.

Finally, our group ran into some computational hurdles with the timeline of this project that limited the hyperparameters that we were able to explore. Given access to more robust cluster computing, we would have explored deeper trees with more iterations. Additionally, we would have liked to tune the hyperparameters in the other model classes that we explored before selecting a champion model.

Bibliography

Evangelista, P., Clark, N., Dabkowski, M., & Kloo, I. (2021). Modeling and Analysis in Support of Organizational Decisions During the COVID-19 Pandemic. *Industrial Systems Engineering Review*, *9*(1). https://doi.org/10.37266/ISER.2021v9i1.pp2-14

Gupta, S., Raghuwanshi, G. S., & Chanda, A. (2020, April 21). Effect of weather on COVID-19 spread in the US: A prediction model for India in 2020. *Science of The Total Environment*, *728*(138860). https://doi.org/10.1016/j.scitotenv.2020.138860

Tosepu, R., Gunawan, J., Effendy, D. S., Ahmad, L. O. A. I., Lestari, H., Bahar, H., & Asfian, P. (2020, April 2). Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Science of The Total Environment*, *725*(138436). https://doi.org/10.1016/j.scitotenv.2020.138436