Team Metis 2020

Ridwan Alam, Vanessa Hu, Ryan Lewis, Ramon Martin,
Neda Saleem, Brian Tam, Nick Wilders, Andrew Zhou

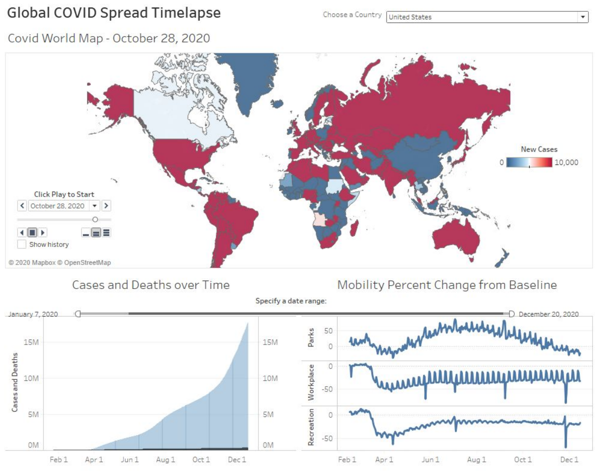
INTRODUCTION

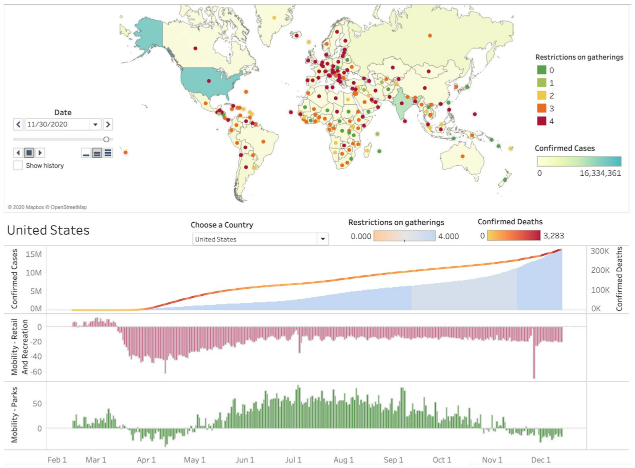
This document serves as the qualitative description and analysis for the submission to XPrize's COVID Response Challenge by Team Metis. Because COVID-19 is a global issue like no other, that affects every single individual on the planet, our team believes that the qualitative analysis and discussion of the humane aspects of this model are of utmost importance.

In this qualitative description, the model will be described, along with our strengths and areas of growth opportunities. We will also describe our diverse, qualified team along with our unique collaborative process, including our recent contributions to the larger conversation about COVID-19, the vaccine, and how data can be a vital tool in solving this problem.

EXPLANATION

Our model depends on one of the simplest and often most elegant concepts of data science: linear regression, implemented in a **generalized linear model**. Our team Tableau dashboard 1 demonstrated a quick visualization of these trends working separately. The dashboard 2 demonstrated the pandemic spread and the government response of restriction on gathering. A sample visualization below shows the disparate difference in new cases, with the bottom portion reflecting United States's mobility percent change from the baseline:





After extensive data cleaning and EDA using this tool, several models were developed; it was found that a linear relationship was found on a national and global scale, with various scaling measures dependent on the effectiveness in particular regions. We are able to predict the amount of New Cases (our target variable), and plans of developing a more telling and sophisticated metric for evaluation (outlined in the *Addressing The Challenge* section found below).

INNOVATION

Our team is an eclectic group, with backgrounds as diverse as the field of data science itself, from mathematics to finance, media to software development. We are bonded by our studies at our namesake, Metis Data Science Bootcamp, and are bonded by that shared experience despite our disparate backgrounds. We collectively bring a great deal of unique thought and perspective to this challenge. The most innovative aspect of our model and of our design process has been the space for development as our understanding of COVID progresses as more features develop. We have used our analytical minds and predictive thinking to develop robust frameworks capable of handling new parameters and producing optimized case-specific predictions.

By emphasizing interpretability and generality, we have allowed space to understand interactions between features, and how those interactions could be interpreted and highlighted with new features. Like XPrize, we value openness and clarity in how the model works, and that focus remains the heart of our innovative approach.

GENERALITY

Because our model relies on case development within each country, and the nuanced way that the country has handled COVID-19, it is optimized to ensure an accurate prediction specific to the region in question. While tuning hyperparameters to ensure optimized model performance, the team noticed that some countries (Costa Rica India, Mauritania, and the Philippines) performed better with unscaled data. This led to our conclusion that the sociopolitical ecosystem of each country is related but independent, and needs to be treated as such. We created a separate model for these particular countries, and see the continual examination of regionalized trends as an area of growth opportunity for this model.

COLLABORATIVE CONTRIBUTIONS

As mentioned above, we developed and trained our data science skills together at Metis Data Science Bootcamp. The values of collaboration and open contribution we learned at Metis were critical to our successful interactions as a team. From a back-end development standpoint, continuous code reviews ensured the capability of all members to comprehend the workflow and “big picture” of the project. While two of our team members focused heavily on the linear model development, one of our team members was heavily researching the implementation of other models and core concepts for our current and future prescriptive models alike, particularly the SIR model (detailed below) for usage in the prescriptive phase of the competition.

Our expert knowledge of collaborative tools like GitHub and Google Drive allowed us to maximize their communicative capabilities. Through the use of GitHub, the team managed parallel development and maintained strict version control, all while communicating with a robust ecosystem of documentation in our team Google Drive. We implemented formalized management tools like the Gantt chart to ensure timelines and expectations were made clear across the board, and the team adhered to a strict deadline schedule while maintaining consistent communication.

Our code is also designed to be interpreted and available in an open-source platform. Again, we highly emphasize the ability to bring on new features and team members, in order to keep this model current and effective. We also continue to participate in the larger public health and data conversation;

Vanessa Hu, our team leader, was recently selected as one of the panelists to participate in a LinkedIn Live event hosted by Harvard Business Analytics Program. She will be a student panelist in conversation with Moderna CEO, Stéphane Bancel, about their business model transformation after Moderna's COVID-19 vaccine is approved.

CONSISTENCY

The flexible and country-specific nature of our model allows us the ability to prioritize consistency in our predictions. These predictions are optimized for real-time efficiency, including established upper and lower bounds of prediction. In addition to realistic and applicable predictions, the model is as successful in the long-term as it is in the short-term, allowing the space for users to address human and dynamic problems.

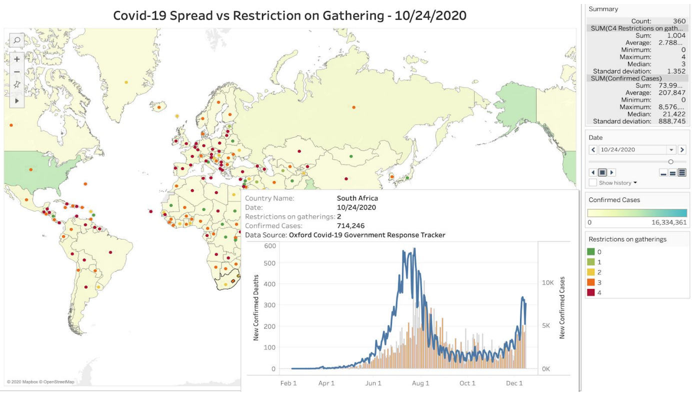
SPEED AND RESOURCE USE

A generalized linear model is computationally efficient, and minimizes mainframe use. Our model not only outputs predictions for each country as required by the competition, but also produces a separate document outlining the interpretability of each coefficient, with consistent file formatting for ease of usability. This leans into our value of creating a “glass box” model that permits and encourages full understanding of how a model works, and what features are most important in its decision-making.

This interpretability and focus on understanding of how real factors affect the model’s predictions are critical in tracking and reconfiguring this model as needed, and the ability to quickly and appropriately onboard new personnel as our team continues to grow.

ADDRESSING THE CHALLENGE

In order to implement this model as practically as possible, we have begun development of a weighted Mean Absolute Error (MAE) for the 7 day moving average of actual new cases compared to our predicted values. That MAE is then adjusted based on the relative population to a new calculated metric of MAE per 100,000 people. This allows more accurate scoring, particularly for smaller countries where the penalty is weighted against fair and accurate predictions. One of our core goals as a group is to represent this problem as realistically as possible, and that includes proper weighted evaluation metrics.



Covid-19 Spread vs Restriction on Gathering - 10/24/2020

Country Name: South Africa
Date: 10/24/2020
Restrictions on gatherings: 2
Confirmed Cases: 714,246
Data Source: Oxford Covid-19 Government Response Tracker

Summary

Count:	360
SUM(C4 Restrictions on gath...	
Sum:	1,004
Average:	2.788...
Minimum:	0
Maximum:	4
Median:	3
Standard deviation:	1.352
SUM(Confirmed Cases)	
Sum:	73,99...
Average:	207,847
Minimum:	0
Maximum:	8,576,...
Median:	21,422
Standard deviation:	888,745

Date: 10/24/2020

© 2020 Mapbox © OpenStreetMap

At this stage of the competition, we have not yet incorporated our weighted metric, and have a certain need for further feedback from judges and XPrize. We are continuing to plan and think in a forward fashion about the most effective ways to address this problem. Further aspects of our prescriptive model include a SIR (Susceptible/Infected/Recovered) model. Based on the infection and recovery rates, beta and gamma respectively, the SIR model provides insight as to the lifespan can tell us if an infection will die out or converge to an endemic equilibrium, and approximately how long. We are currently working on using the data to estimate beta and gamma for each country/region.