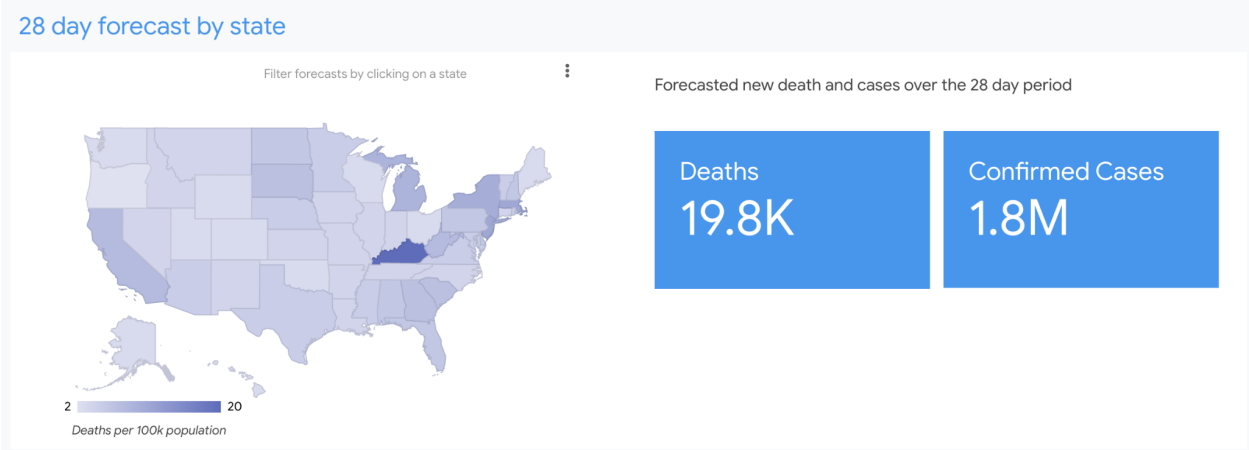
CASE STUDY: Responsible AI

COVID-19 Public Forecasts



The Tool

Google Cloud's COVID-19 Public Forecasts BigQuery Datasets Program is a tool for forecasting COVID-19 cases, hospitalizations, and deaths by county across the United States and Japan. These forecasts can help decision makers in healthcare, the public sector, and other impacted organizations prepare for the future in terms of resource allocation and other needs, and create mitigations where appropriate.

Creating a forecasting model using machine learning (ML) is in line with Google's AI Principles, especially #1: be socially beneficial. While developing this dataset and ML model it was important for the team to address Google's AI Principle #2: avoid creating or reinforcing unfair bias. By using publicly available and anonymized datasets, the team adhered to AI Principle #5: Incorporate privacy design principles. (note: the tool was designed first with U.S. data in the summer of 2020, but has since expanded to Japan in late 2020; this case study focuses on the original Responsible AI approach.)

The potential to reinforce unfair bias through unintentionally propagating the effects of structural inequalities based on historical disadvantages for certain groups of people as they are represented in existing data is a very real possibility. The Centers for Disease Control (CDC) has shared data showing that racial and ethnic minorities in the U.S. have been impacted the most by COVID-19, enduring more confirmed cases of COVID-19 and more deaths than the U.S. majority demographic group.

The Google Cloud team engaged in an ethics review and applied Responsible AI practices to check that the ML model did not ship disproportionate downstream harms to historically marginalized groups.

The Approach

In order to reduce the risk of reinforcing unfair bias, the team applied the following framework: **assess, evaluate, mitigate.**

Assess:

The team analyzed public data from sources such as Johns Hopkins University, Descartes Lab, and the United States Census Bureau, and catalogued any observed historical inequalities and implications. The team also created a proposed framework to compare the performance of the model in and across counties with both historically marginalized and at-risk sub-populations (disaggregated analysis). The team decided to analyzing the following slices, or sub-populations:

1. The 3 most common demographic groups in the US (based on 2010 Census categories): African-American, Hispanic, and White
2. Populations for whom COVID-19 spreads the most quickly: prison populations, meat-processing workers, undocumented immigrants, nursing home residents. Including at-risk groups ensures that the model is exposed to different populations during analysis so that forecasting can be fair.

Evaluate:

To evaluate, the team defined the variables and metrics they would use to evaluate the robustness of the ML model for the forecasting tool. Examples of key metrics include averages over the number of:

- Ground truth of deaths
- Predicted deaths
- Time stamps for daily forecasts

They also created a list of key assumptions that informed their decisions. Examples of key assumptions include:

- Counties with high populations of sub-group X are good proxies for that sub-group as a whole
- Reporting country-wide metrics by averaging per-county is justifiable, as individual counties test, label, and report values differently

The team then created a Colab (a collaborative data analysis using Google's Colab tool), which is generalizable. The Colab can be used in specific use cases for analyzing the model's performance for different sub-groups, by changing key assumptions and disaggregated data sources.

Mitigate:

After analyzing the results from the Colab, the team was able to determine appropriate mitigations to ensure that end users, namely decision makers for COVID-19 resource allocation, can be equipped with the knowledge to make informed decisions that did not amplify unfair historical biases.

The Outcome

The **assess, evaluate, mitigate** framework is a helpful thought exercise to go through to understand the challenges of a model and potential mitigations. To document and communicate their approach, the team created a model card, embedded within the tool's user guide, to present the model's inputs and limitations, as well as its performance across sub-groups.

The intention of creating a model card within the tool's user guide was not only for providing transparency and accountability to people (AI Principle #4), but also to mitigate the risks of deploying a model with limitations that are only known to a software developer who might be working with the dataset and the model.

Because model cards are simple, structured documents designed to share key assumptions, guide how to use the output of the model, and, most importantly, offer key limitations and tradeoffs of the model, they can be helpful in building trust among various stakeholders. As a result of using the framework, we released a forecasting tool that takes into account the socio-technical concerns raised when using historical data in real-world situations that may have structural inequalities and disadvantageous associations encoded into the data.

About Google's AI Principles

In 2018, Google published our AI Principles to help guide ethical development and use of the technology. Our objectives: 1. Be socially beneficial. 2. Avoid creating or reinforcing unfair bias. 3. Be built and tested for safety. 4. Be accountable to people. 5. Incorporate privacy design principles. 6. Uphold high standards of scientific excellence. 7.Be made available for use in accord with these principles. In addition to the above objectives, we will not design or deploy AI in the following application areas: 1. Technologies that cause or are likely to cause overall harm. Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks, and will incorporate appropriate safety constraints. 2. Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people. 3. Technologies that gather or use information for surveillance violating internationally accepted norms. 4. Technologies whose purpose contravenes widely accepted principles of international law and human rights. As our experience in this space deepens, this list may evolve.