Weekly Progress Report: Oct 30 - Nov 3

Contributors: Ryan Choe, Rakin Hussain, Sid Taneja

Progress/Accomplishments:

- Given the feedback from the midterm evaluation, we have redesigned our scope. Our initial objective was to create a model for all countries included in the data set, but we have decided to narrow our scope by choosing the most representative country per continent. We chose the country with the largest population as the metric for determining this. These countries were the United States, Brazil, Germany, Nigeria, China, and Australia. We will consider these 6 countries only, proceed to prepare the data, and model them individually.
- We are rehandling missingness and imputation. First, we pruned through all of the predictor descriptions and manually removed the ones that appear to be the least useful. These included female_smokers, male_smokers, and diabetes_prevalence. Then, given the feedback from the midterm, we realized that 20% was too low of a cutoff for missingness, as this eliminated a great majority of the original predictors. We wanted to ideally eliminate around ⅓ of predictors for each country and no more than ½. Thus, we increased our threshold up to 30%. However, 11 predictors regarding vaccine distribution displayed missingness up to 40-60%, but were kept for some countries since most of this missingness occurred within the first year of data. We anticipate that the models will be robust to this specific type of missingness, so they were not removed.
- For imputation, we decided against the KNN method according to our feedback. A few imputation methods were employed:
    - First, for reasonable variables that only had missingness in the first year of reporting, such as vaccine related predictors, we filled in those values with 0. This makes sense since countries weren't vaccinating in the initial pandemic period anyway.
    - Next, some predictors experienced a similar lack of reporting in the initial pandemic period, but were not vaccine related. For variables such as reproduction_rate and hosp_patients, though they lack initial reporting, they definitely were not 0. Thus, we backward filled this initial period with the first reported value, and forward filled the rest of the column.
    - Finally, we applied general forward filling to the rest of the predictors.

Problems/Challenges:

- Our team had to adjust its goal for a second time after initially trying to predict new_cases as the response variable for each continent. However, so much data was missing for the continents (not just from early 2020 but all over the time series). Our team then pivoted to focusing on a single representative country for each continent, determined by population.
- Deciding a new cutoff threshold for proportion missingness for variables required a lot of trial and error. The team attempted to eliminate some variables entirely based on contextual intuition and analysis. Luckily, a thoughtful process was determined to decide which variables to drop or keep for each respective country and we decided on a threshold of 50%. Our group was mindful to pick this and ensure that a thoughtful imputation process was followed to avoid any issues such as data leakage or ethical irresponsibility.
- The work for imputing missing values was quite tedious and required constant communication from all members of the team. Code had to be adjusted for each country depending on which variables were involved for a country and also at what point in time chunks of the missing data were coming from. In the end, the group trouble-shot to the best of their ability and will run it over with TA Lining and Professor Shi.

Plan/Next Steps:

- Over the weekend, our team will work on Feature Engineering, with lag features and rolling statistics and final EDA preparation steps. This will include proper Seasonal Decomposition analysis. The overall target is to have these steps done no later than by the end of Tuesday so we can begin building our models over the final week and a half before the presentation. We plan to follow-up with Lining and Professor Shi on the status of our data preparation as we near finishing it.