STAT 390 Weekly Report 2

Oct 9-13

Group 2: Cindy Ha, Willie Xie, Erica Zhang

Table of contents

1 Progress/Accomplishments

- Finalized prediction question: Predicting `new_cases` of COVID-19
- Completed additional data pre-processing (`Preprocessing Code/adv_preprocessing.R`)
  - Selected only variables with 70% completion
  - Addressed multicollinearity
  - Imputed some missingness issues
    - `new_deaths`: impute with 0
    - `total_deaths`, `total_cases`, `reproduction`: replace with the last non-zero value
    - `new_cases`: replace by change in total cases
    - `extreme_poverty`: replace by respective continent median extreme_poverty value
- Researched more about appropriate models to use for classification with R
- Brought back in some more recent 2023 data for further testing purpose; working to deal with large missingness for some significant predictor variables, especially for countries that just stopped reporting this year

2 Challenges

- Large amount of missing data for predictor variables (ie. **vaccination**, **hospital**, **icu**) in earliest time and more recent data reporting. Trying to narrow down on a time period where at least 60% of data cumulatively is not missing and then apply proper imputation; meanwhile, find if these variables are highly correlated by checking correlation matrix
- One predictor, `extreme_poverty` has missingness due to specific countries (imputed with geographic average)
- Determining the time frame for the dataset since early data lacks vaccinations while later data is not collected.

3 Next Steps

- Save finalized dataset to `data/processed_data` folder
- Split data into training and testing using `time_series_split` with older data for training and most recent data for testing
- Build recipes & models