STAT 390 Weekly Report 4

Oct 30- Nov 2

Group 2: Cindy Ha, Willie Xie, Erica Zhang

Table of contents

1 Progress/Accomplishments

- Three types of datasets are created from the initial dataset to be used for different model types: `linear`, `tree based`, and `neural nets`
- For feature engineering for all dataset types, added 2 new variables: `month` and `day_of_week`
- For the `linear` dataset, outliers are located using the Cook's distance method for regression and checking if these values exceed the threshold of 0.5. By Cook's distance criteria, there does not appear to be any influential outliers in the dataset.
- Afterwards, training and testing sets are created for each dataset type (located in `data/finalized_data` folder
  - Training sets consist of observations before 2023 and the testing sets consist of observations after 2023.
  - Most data is not missing between Feb 2021 and March 2022.
  - `linear`: Removed predictors will large missingness

    `tree based`: Keep significant predictors with large missingness, but replace missingness for observations outside Feb 2021 and March 2022 with a large value = 10^{15}.

    `neural net`: Keep significant predictors with large missingness, but for each, create an indicator predictor to indicate if the value is missing (FALSE) vs value exist (TRUE).

  - Applied K-means clustering with `life_expectancy`, `female_smokers`, `male_smokers` to impute the few random missingness between Feb 2021 and March 2022 by cluster median. The same clustering is applied to the testing sets to impute any random missingness.

- Set up tuning scripts for xgboost and arima models
  - Using a backtesting resampling method: `rolling origin forecast` to generate validation sets of 1 year length for model training + 2 month for model parameter comparisons along with 4 months skipping increments between validation sets.

2 Challenges

- There was still missingness in the testing sets because some data was completely missing (no median could be computed) so we had to impute missingness in the testing sets by the median values in the training set.
- We are removing China from list of countries we are analyzing due to different Covid reporting policies there since the data may be unreliable (irreducible error)

3 Next Steps

- Looking into lagged features over the weekend to complete feature engineering
- Each member will run the 6 models by Nov 10 to provide starting model performance