Data Science Project Proposal - Group 2

Group members: Cindy Ha, Willie Xie, Erica Zhang

Project Title:

CoronaCaster: Predicting New COVID-19 Cases

Project Overview:

For this project, we are using a complete COVID-19 dataset, which is a collection of the COVID-19 data maintained and provided by *Our World in Data* that gets updated daily. As an ongoing pandemic, COVID infection still disrupts our daily lives, especially those vulnerable people like senior citizens and immunocompromised groups. We would like to use this dataset to train and find best fit models to forecast future daily new cases in the States – whether there would be another outbreak or on a consistently declining trend. This could be meaningful as people can take predictions as a reference on what time periods they should be extra cautious and take preventative measures. Moreover, the public health system can also refer to the predictions and adjust staffing accordingly ahead of time.

Objectives:

The goal of this project is to forecast the number of confirmed daily new cases of COVID-19 in the United States based on predictors such as the total number of cases, geographic location, population, number of tests, etc. Since the project will be predicting the number of new cases, then the type of project will be regression which means the project will be maximizing metrics like RMSE or R^2. Through this project, we hope to accurately measure the degree of spread of COVID-19 and understand when COVID-19 will most likely not be a large concern.

Methodology:

From the initial data source, the dataset needs to be cleaned. Some issues and solutions that needs to be addressed at this stage are:

- Missingness
    - Predictors with large amount of data missing (>30%) needs to be excluded
    - Other types of missing data needs to be removed or imputed either through predictor averages, random forests, or k-nearest neighbors
- Data type mismatches
    - The date predictor if string needs to be converted into a data type
    - Predictors without significant value are excluded
- Multicollinearity
    - Some predictors are running averages of other predictors and needs to be excluded
    - Highly correlated predictors based on a correlation matrix need to be excluded

After these preprocessing issues are addressed, the data can be split into 80/20 training/testing using time series stratified sampling on geographic location. In order to repeatedly optimize model learning models, the training set needs to be further split into validation sets through time series resampling methods. The number of validation sets will most likely be around 5 to 10. Finally, the validation sets will be used to

train the models below to determine the model that will be used to finally predict the number of new COVID-19 cases.

- OLS regression
- Elastic net regression
- K-nearest neighbors
- Multivariate adaptive regression spline (MARS) model
- Random Forest
- Boosted Trees
- Support Vector Machines (SVMs)
- Neural Networks
- ARIMA model
- Ensemble model

To compare the accuracy of each model, the project will be primarily maximizing the RMSE while taking into account the R^2 and the model accuracy for different ranges of response values.

Data Sources:

The source is from Kaggle in which the dataset is updated daily:

Our World in Data. (2023). Our World in Data - COVID-19 [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/6559049

Data Set:

From our dataset, there are 67 variables: 1 response + 66 initial predictors.

iso_code	continent	location	date
total_cases	**new_cases**	new_cases_smoothed	total_deaths
new_deaths	new_deaths_smoothed	total_cases_per_million	new_cases_per_million
new_cases_smoothed_per_million	total_deaths_per_million	new_deaths_per_million	new_deaths_smoothed_per_million
reproduction_rate	icu_patients	icu_patients_per_million	hosp_patients
hosp_patients_per_million	weekly_icu_admissions	weekly_icu_admissions_per_million	weekly_hosp_admissions
weekly_hosp_admissions_per_million	total_tests	new_tests	total_tests_per_thousand
new_tests_per_thousand	new_tests_smoothed	new_tests_smoothed_per_thousand	positive_rate
tests_per_case	tests_units	total_vaccinations	people_vaccinated
people_fully_vaccinated	total_boosters	new_vaccinations	new_vaccinations_smoothed

total_vaccinations_per_hundred	people_vaccinated_per_hundred	people_fully_vaccinated_per_hundred	total_boosters_per_hundred
new_vaccinations_smoothed_per_million	new_people_vaccinated_smoothed	new_people_vaccinated_smoothed_per_hundred	stringency_index
population_density	median_age	aged_65_older	aged_70_older
gdp_per_capita	extreme_poverty	cardiovasc_death_rate	diabetes_prevalence
female_smokers	male_smokers	handwashing_facilities	hospital_beds_per_thousand
life_expectancy	human_development_index	population	excess_mortality_cumulative_absolute
excess_mortality_cumulative	excess_mortality	excess_mortality_cumulative_per_million	

As noted in the methodology, there are a few major issues that need to be addressed during preprocessing. Some initial data cleaning are addressed below. For more information, refer to preprocessing file at https://github.com/AzureAmber/STAT-390-Covid-Project.

- Missingness: 41 variables appear to have at least 30% missingness. It should be noted that some variables are not updated daily, but instead maybe weekly or monthly or just not recorded during some times. Thus, some of these variables might be usable for training models, but for initial impressions, these variables have missingness issues that most likely need to be addressed. Also, note that for now, only column missingness is addressed so later on, random missingness needs to be imputed with methods addressed in the methodology.

tests_units	weekly_icu_admissions	weekly_icu_admissions_per_million	excess_mortality_cumulative_absolute
excess_mortality_cumulative	excess_mortality	excess_mortality_cumulative_per_million	weekly_hosp_admissions
weekly_hosp_admissions_per_million	icu_patients	icu_patients_per_million	hosp_patients
hosp_patients_per_million	total_boosters	total_boosters_per_hundred	new_vaccinations
people_fully_vaccinated	people_fully_vaccinated_per_hundred	people_vaccinated	people_vaccinated_per_hundred
new_tests	new_tests_per_thousand	total_vaccinations	total_vaccinations_per_hundred

total_tests	total_tests_per_thousand	tests_per_case	positive_rate
new_tests_smoothed	new_tests_smoothed_per_thousand	handwashing_facilities	extreme_poverty
new_people_vaccinated_smoothed	new_people_vaccinated_smoothed_per_hundred	new_vaccinations_smoothed	new_vaccinations_smoothed_per_million
reproduction_rate	male_smokers	stringency_index	female_smokers
hospital_beds_per_thousand			

- Data type mismatches
  - Every variable appears to have the correct data type
  - Only iso_code appears to not have any significance
- Multicollinearity
  - total_cases: total_deaths (0.96) population (0.75)
  - new_cases: new_cases_smoothed (0.88)
  - new_deaths: new_deaths_smoothed (0.94)
  - median_age: age_65_older (0.91) age_70_older (0.90) life_expectancy (0.85) human_development_index (0.90)
  - age_65_older: age_70_older (0.99) life_expectancy (0.73) human_development_index (0.78)
  - age_70_older: life_expectancy (0.72) human_development_index (0.75)
  - gdp_per_capita: human_development_index (0.75)
  - life_expectancy: human_development_index (0.92)

  Most of the multicollinearity issues appear with life_expectancy, human_development_index, and population which can be removed from the dataset.

Thus, the final dataset contains 22 variables: 1 response + 21 predictors (1 date, 2 categorical, 17 numerical) still with 341,408 observations. For now, the dataset addressed some variable missingness, but later on, the dataset needs to address observations with large amounts of missingness.

References:

Alassafi M., Jarrah M, Alotaibi R. (2022). Time series predicting of COVID-19 based on deep learning. *Neurocomputing, 468*, 335-344. https://doi.org/10.1016/j.neucom.2021.10.035

Long, J. (JD), & Teetor, P. (2019). *R cookbook, 2nd edition*. 14 Time Series Analysis. https://rc2e.com/timeseriesanalysis

Painuli, D., Mishra, D., Bhardwaj, S., & Aggarwal, M. (2021). Forecast and prediction of COVID-19 using machine learning. *Data Science for COVID-19*, 381-397. https://doi.org/10.1016/B978-0-12-824536-1.00027-7

Smita Rath, Alakananda Tripathy, Alok Ranjan Tripathy. (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(5), 1467-1474. https://doi.org/10.1016/j.dsx.2020.07.045

Conclusion:

In summary, this project aims to adeptly predict the daily count of new confirmed COVID-19 cases in the US by examining different relevant variables. We will employ a regression model and seek to maximize the accuracy of predictions by minimizing RMSE and mamizing R^2. In combination with rigorous data preprocessing and exploration, we will employ multiple predictive models to work with the extensive, time-series data set. As the data source is continuously updated, the project will hopefully unravel insights into the recent patterns of COVID-19 spread up to September 2023, and provide a timeline where the disease is no longer a persistent health concern. The insights from this project will potentially enable us to develop a better understanding of the current state of COVID-19.