Global Spread of COVID-19 Prediction

Kanak Choudhury, Mingdian Liu

Introduction

- The whole world is suffering for the Covid-19 pandemic. It becomes the major concern for every country to predict number of deaths and infected persons under different conditions and measures.
- We considered two step prediction model that can incorporate different measures implemented to control this covid19 pandemic as well as country specific economic, health and demographic indicators.
- In the first set, we used popular methods (SIR, SIRD) generally used to predict deaths and infected numbers and in the second step, using these prediction as well as other country specific indicators and some factors generated based on some conditions were used to fit machine learning models to predict number of daily cumulative deaths.

Related Works

- S-I-R model (Kermack & McKendrick, 1927): does not depends on time or other factors
- Time Dependent Modified Models: (Keeling & Rohani, 2008), (Amaro, 2020), (Calafiore, Novara, & Possieri, 2020), (Chen, Lu, & Liu, 2020), (Bayes, Rosas, & Valdivieso, 2020)
- Applications on Covid-19: Keggle repository, (Dhanwant & Ramanathan, 2020).
- Covid 19 Analysis: (Fanelli & Piazza, 2020) for China, Italy and France
- (Qi, Karlsson, Sallmen, & Wyss, 2020) studied on Covid-19 using SI, SIR and SID
- The Institute for Health Metrics and Evaluation (IHME) developed mixed effects non-linear regression framework to estimate the trajectory of the cumulative and daily death rate as a function of social distancing measures, (IHME Team, 2020).

Susceptible-Infected-Recovered (SIR) Model

$$\frac{dx}{dt} = -\beta x(t)y(t) \quad (1)$$

$$\frac{dy}{dt} = \beta x(t)y(t) - \gamma y(t) \quad (2)$$

$$\frac{dz}{dt} = \gamma y(t) \quad (3)$$

where $x(t)$ is the susceptible (S), $y(t)$ is the infected (I) and $z(t)$ is the recovered (R) and at time 0, $x(0) = N1 \geq 0$, $y(0) = N2 \geq 0$ and $z(0) = N3 \geq 0$, $\beta \geq 0$ is the infection rate and $\gamma \geq 0$ is the mean recovery rate for a fixed population, N such that $N1 + N2 + N3 = N$

Susceptible-Infected-Recovered-Death (SIRD) Model

$$\frac{dS}{dt} = -N^{-1}\beta SI \quad (4)$$

$$\frac{dI}{dt} = N^{-1}\beta SI - (\gamma + \alpha)I \quad (5)$$

$$\frac{dR}{dt} = \gamma I \quad (6)$$

$$\frac{dD}{dt} = \alpha I \quad (7)$$

where S, I, R, and D represent susceptible, infected, recovered, fatal, respectively. N is the total population. In this model, α represents fatal rate γ represents recovery rate, β presents infection rate

Data set

- The covid-19 infected, recovered and deaths information for each country: Zindi competition website.
- Countries: 174
- Variables: Daily deaths, infected and recovered for each country
- Time: from 22 January 2020 to 21 April 2020.
- Train Period: until 08 April 2020
- Test Period: From 08 April 2020 to 21 April 2020

Data set

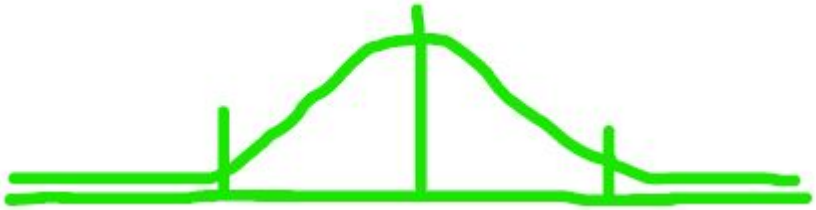
- Country specific economic, health and demographic indicator information: world bank data repository.
- 32 indicators were manually chosen from 1500 indicators

Data set

List of some important Indicators:

1. Health: Current health expenditure per capita, Domestic general government health expenditure per capita, Domestic private health expenditure per capita
2. Economy: GDP per capita, GNI per capita, Net primary income
3. Immunization: DPT, HepB3, measles , Tuberculosis case detection rate, Tuberculosis treatment success rate, Incidence of tuberculosis
4. Demographic: Life expectancy at birth, total, Population Distribution for ages 0-14, 15-64, 65+, Population density, Rural population

Features & Methods



- Change Point: Find the change point where it started to grow or fall down the curve using time serie change point trend model
- Prediction based on change point model for all 173 countries
- Cluster Countries (k-means) with similar temporal growth distribution
- Fit and predict SIR & SIRD model over time
- Cluster countries (k-means) based on similarity of infected, deaths, recovered and country indicators

Feature Generation & Methods

- Four closest neighbor countries information
- Some new features based on country specific information, e.g. how many days took to infect, deaths or recover 1000 people
- Machine Learning Models:
  - Support vector regression
  - XGBoost regression
  - Spline regression
  - Random forest regression
  - Neural network regression
- Ensemble Model: simple average

Parameter Tuning

- Random forest: `'n_estimators', 'max_depth', 'min_samples_leaf'`
- SVR: `'kernel':('poly', 'rbf', 'sigmoid'), 'C', 'epsilon'`
- Spline Regression: `'max_degree', 'penalty'`
- XGBoost Regression: `'n_estimators', 'max_depth', 'learning_rate', 'gamma', 'subsample', 'reg_alpha'`
- Neural Network: `'hidden_layer_sizes', 'alpha', 'beta_1', 'beta_2', 'epsilon', 'n_iter_no_change'`

Results and discussion

Model	Test Mean Absolute Error	
	Train	Test
SVR	383.75	314.60
XGBoost	391.97	322.14
Random Forest	732.89	722.76
Spline Regression (All Variable)	2314.61	496.89
Spline Regression (Subset 1)	1631.72	1578.33
Spline Regression (Subset 2)	6493.71	8634.68
Neural Network	2172.07	2314.88
Ensemble	1715.86	2028.80

Table 1: Test mean absolute error for different models at hyperparameter tuning

Results and discussion

Model	Mean Absolute Error (Final model)							
	USA		Italy		China		Bangladesh	
	Train	Test	Train	Test	Train	Test	Train	Test
SVR	1007.42	28995.93	2898.43	21278.93	2195.99	3804.57	1.47	56.29
XGBoost	1000.27	28953.62	2886.70	21236.64	2176.48	3762.30	18.03	24.11
RF	374.90	19250.55	1247.86	4760.96	723.58	793.61	1.23	42.37
SPR (All)	83.79	9314.05	972.75	3496.74	344.06	496.89	18.24	170.19
SPR(Set1)	87.11	11638.08	902.37	120.00	331.04	467.74	5.16	28.38
SPR(Set2)	93.20	1128.31	270.93	6170.33	146.14	2667.25	1.89	24.24
NN	143.72	10459.02	395.64	1247.01	336.57	1080.88	4.20	26.30
Ensemble	120.10	10307.36	657.08	1728.33	217.46	245.79	3.48	23.95

Table 2: Mean Absolute Error of observed and estimated cumulative deaths for selected countries.

Results and discussion

Figure 1: Train and Test MAE for different model of selected countries.

Results and discussion

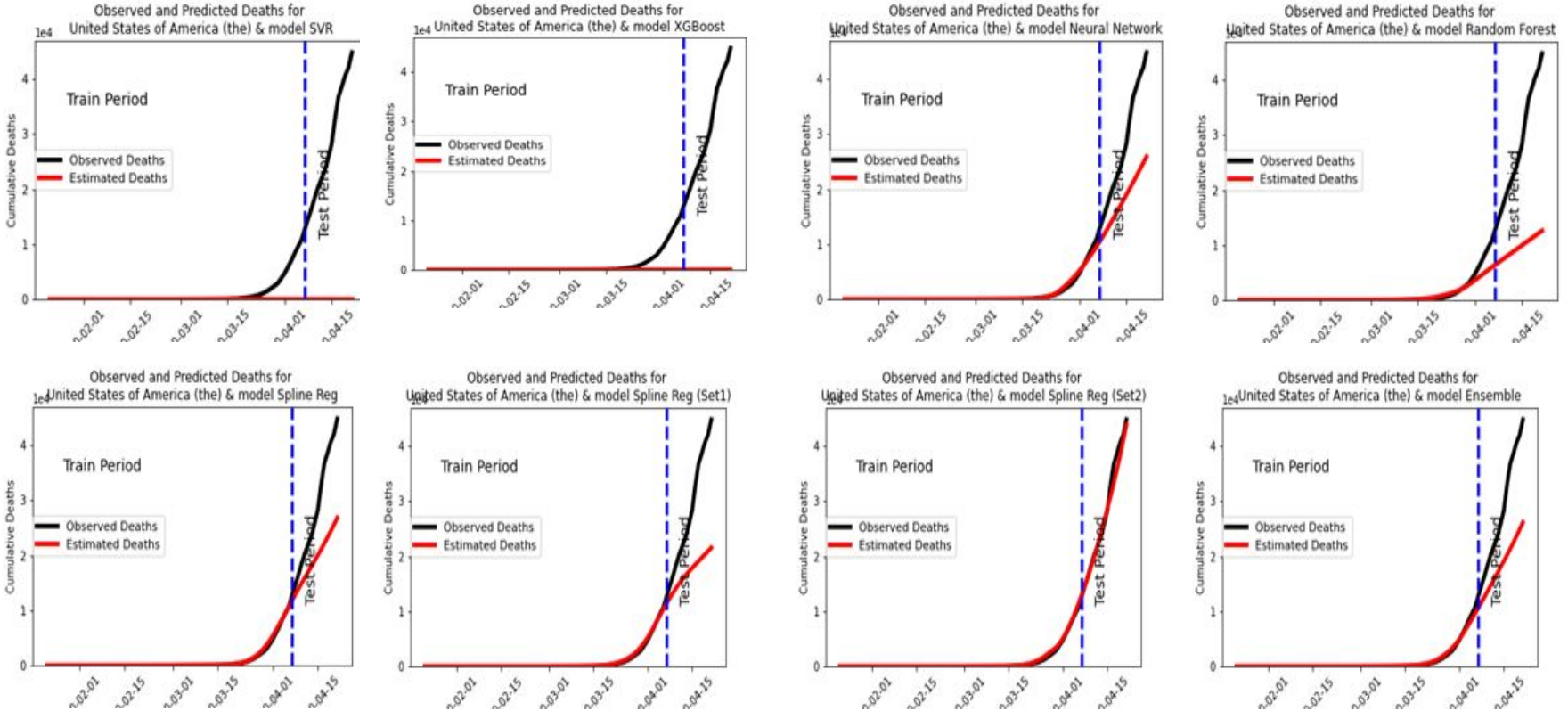
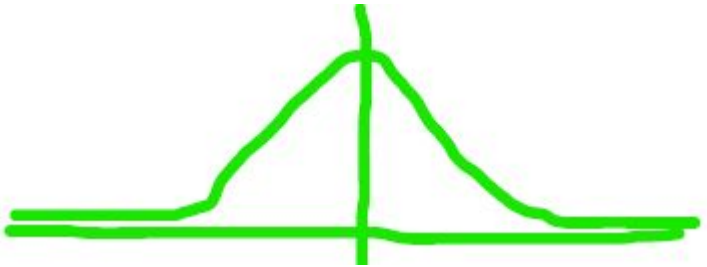


Figure 2: Observed and estimated cumulative deaths of USA for different model, (1) SVR, (2) XGBoost, (3) Ensemble, (4) Neural Network, (5) Random Forest, (6) Spline Reg. (all), (7) Spline Reg. (set1), (8) Spline Reg. (set2).

Results and discussion

- Support vector regression and XGBoost regression model perform best on test data
- Test MAE for SVR, XGBoost regression and RFR are even lower than the training MAE.
- XGboost and SVR models perform poorly for countries with highly infected and deaths rates (e.g. USA), underestimate for both training and testing data
- Spline regression (SPR) performs poorly on the overall data but performs better for countries with highly infected and deaths rates (Table 2, Figure 2).
- All model underestimates for the longer time period on test data.

Limitations and Future Work



- All machine learning model considers same distribution in train and test window. We are using half part of the curve as training and the other half as prediction.
- It would be better to predict temporal death rate instead of deaths
- Higher order polynomial and interaction terms
- Hierarchical linear and nonlinear model. For example, hierarchical model based on the change point of the curve.
- Deaths is a count data. So it would be better to consider generalized Poisson model

Acknowledgement

- This work is a solution to Zindi competition: Predict the Global Spread of COVID-19 (https://zindi.africa/competitions/predict-the-global-spread-of-covid-19)

- Thanks to my friends Sandip Sureka and Shiplu Sarker to help me to identify country indicators and to understand SIR and SIRD model

References

Amaro, J. E. (2020). The D model for deaths by COVID-19. arXiv preprint, 2003.13747. Bayes, C., Rosas, G. G., & Valdivieso, L. (2020).

Modelling death rates due to COVID-19: A Bayesian approach. arXiv. Retrieved from arXiv: https://arxiv.org/pdf/2004.02386.pdf

Calafiore, G. C., Novara, C., & Possieri, C. (2020). A Modified SIR Model for the COVID-19 Contagion in Italy. arXiv preprint, 2003.14391.

Chen, Y.-C., Lu, P.-E., & Liu, T.-H. (2020). A Time-dependent SIR model for COVID-19 with Undetectable Infected Persons. arXiv preprint, 2003.00122.

Dhanwant, J. N., & Ramanathan, V. (2020). Forecasting COVID 19 growth in India using Susceptible-Infected-Recovered. arxiv. Retrieved from https://arxiv.org/ftp/arxiv/papers/2004/2004.00696.pdf

Fanelli, D., & Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons and Fractals, 134.

Harko, T., Lobo, F. S., & Mak, M. K. (2014). Exact analytical solutions of the SusceptibleInfected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates. Applied Mathematics and Computation, 184-194. Retrieved from https://arxiv.org/pdf/1403.2160.pdf

IHME Team, C.-1. h. (2020). Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European Economic Area countries. medRxiv.

References

Johnson, T. (n.d.). Mathematical Modeling of Diseases: Susceptible-Infected-Recovered (SIR) Model. Retrieved from https://op12no2.me/stuff/tjsir.pdf

Keeling, M. J., & Rohani, P. (2008). Modeling Infectious Diseases IN HUMANS AND ANIMALS. New Jersey : Princeton University Press.

Kermack, W., & McKendrick, A. (1927). Contribution to the mathematical theory of epidemics. Proc. Roy. Soc. Lond A, 115, 700-721.

Qi, C., Karlsson, D., Sallmen, K., & Wyss, R. (2020). Model studies on the COVID-19 pandemic in Sweden. arXiv preprint, 2004.01575. Retrieved from https://arxiv.org/pdf/2004.01575.pdf

Website, 1. (2020). COVID-19 data with SIR model. Retrieved from https://www.kaggle.com/kstatucf/covid-19-data-with-sir-model/edit?rvi=1

Website, 2. (2019). Fitting the SIR model of disease to data in Python. Retrieved from https://numbersandshapes.net/post/fitting_sir_to_data_in_python/?fbclid=IwAR0uHK4iij wmT6nw_Em3YGrUukhXS-Ho-qwDrIvU4wcGgcDwt9vo6JRScdU

Thank You!