Machine-learning based prediction of COVID-19 and the effectiveness of interventions

CS229: Project Final Report

Yunjae Hwang, yunjaeh@stanford.edu
Heegwang Roh, hgroh@stanford.edu

1 Introduction

Started on November 2019, the coronavirus disease 2019 (COVID-19) has spread across the globe and has caused millions of deaths, and the pandemic resulted in significant social and economic disruption [1]. Several emergency-approved vaccines against the virus are now available thanks to the global scientific efforts, however, the emergence and the spread of variants that escape antibody neutralization are complicating the situation. It is therefore not a surprise to have another huge wave of infections anytime, as can be seen in the case of Brazil where it has been believed to have reached the herd immunity. Thus, a computational model that well predicts the future disease spread would be helpful in many aspects, including epidemiological study and policy decision.

To achieve such goal, we have developed a machine learning model that predicts the future disease spread based on the past epidemiological data. Specifically, we have established a Poisson regression model that takes the past test positive rate, percentage of fully vaccinated population, and number of holidays as the input, and predicts the future test positive rate as the output. We also investigated the impact of other features that are often believed to affect the disease spread, i.e., climate condition (ambient temperature, in particular) and variants. Ultimately, the key objectives of this project are (i) to predict the trend of disease spread in a near future, (ii) to estimate the effect of emergency-approved vaccines on the occurrence of the disease, and (iii) to suggest a possible impact of emerging variants of concern on the current and future pandemic.

2 Related work

Previous studies have reported several machine learning models for forecasting the future pandemic. For example, to build such a model, Rustam *et. al.* [2] applied several different regression models (LR, LASSO, SVM, ES), Shahid *et. al.* [3] applied deep learning models, and Pandey *et. al.* [4] used the SEIR model which takes the reproduction number (R value) into account and is frequently used in modeling disease spread. However, including these studies, most of the ML-based covid forecasting studies only incorporated the number of positive cases, recovered cases, and deaths as their input, lacking in epidemiological factors in the model [5, 6, 7].

Several studies incorporated features other than the number of cases/deaths/recoveries in their ML models and investigated their impact on the pandemic. Malki *et. al.* [8] analyzed the association between weather, demography, and COVID-19 spread in Italy, and claimed that weather-related factors (temperature, humidity, sunlight availability) are more closely related to the mortality than demographic factors (age, urbanization). However, they used national data for one single day, lacking of enough sample size for credible analysis. A recent work by Giordano *et. al* [9] took the vaccine rollout and the emergence of variants into account, but they did not use the real-word data, rather simply allowing the parameter to be flexible hoping that such change in the degree of freedom reflects the real-world situation.

3 Dataset and Features

3.1 Data selection

We chose the **test positive rate** as the random variable in the model, because it is a more relevant metric of the level of community spread. We didn't use the number of daily new cases, which in most cases have gotten more public attention, because it is directly affected by the number of tests implemented as well as the population of the area of interest thus is likely to complicate the analysis. Also, we decided to use **weekly average** values as the input/output, because it would smooth the noise in real-world data, and many data are provided in a weekly basis thus it would be more relevant to follow the same format.

The main factors for the disease spread would be (i) the level of community spread, (ii) community immunity, and (iii) public behavior. Thus, we chose **test positive rate** ($x_p \in [0, 1]$), **the number of people who are fully vaccinated** ($x_v \in [0, 1]$) and **number of holidays** ($x_h \in Z$) as the primary input features. For x_p and x_h, we included individual data for the past 4 weeks, as a previous virological assessment of the disease showed a patient swab can be tested positive up to 28 days after the onset of the symptom [10]. For x_v, we included the individual data for the current week and the past 2 weeks, as the immunity 'gradually' increases after vaccination, 2 weeks after the final dose is often considered as the 'fully-vaccinated' state, and the Phase III trial of BNT162b2 (Pfizer-BioNTech) vaccine showed efficacy plateauing after 14 days from the second dose [11]. Finally, to analyze the effect of other factors on the disease spread, the weekly average of the portion of variants circulating ($x_{var-B.1.1.7}, x_{var-B.1.351}, x_{var-P.1} \in [0, 1]$)) and weekly average temperature ($x_t \in Z$) were collected and incorporated in some models.

For simplicity, we only included the data for the 10 most populated states. Since they account for more than half of the total US population [12] and are exactly same with the 10 states with the most confirmed cases as well [13], we believe they well represent the overall behavior of the entire population.

Since the data at the very early pandemic was not as accurate and representative to be used for training, and the data for recent past were not yet available, data in between the week number 20 to 70 were collected and analyzed (corresponds to the period between early March, 2020 and mid April, 2021, as the first week of 2020 is week 1 and the first week of 2021 is week 54).

3.2 Data collection

All Covid data were collected from the repository of COVID-19 data underlying the analysis and visualizations created by the Johns Hopkins Centers for Civic Impact for the Coronavirus Resource Center(CRC) [14]. For training the model, we cleaned-up and processed time series of: (i) Cumulative number of completed PCR tests (or specimens tested); (ii) total number of confirmed plus probable cases of COVID-19; and (iii) cumulative number of second doses administered. Diving (ii) by (i) yields x_p, and (iii) directly gives x_v. For x_h, we used two (Saturday and Sunday) plus the number of federal holidays per week (e.g. New Yearś day and Thanksgiving day to the count).

All variant data were collected from the Outbreak.info created by the Scripps Research and supported by the NIH and CDC [15]. Three variants - B.1.1.7, B.1.351, and P.1 (also known as UK, South Africa, and Brazil, respectively) - were taken into account, considering their relatively high prevalence and the rich amount of available data. The daily portion of variants among all reported sequences were collected and averaged into weekly values for subsequent training.

All temperature data were collected from the database of the National Oceanic and Atmospheric Administration (NOAA) under the U.S. Department of Commerce [16], where the database archives global historical weather and climate data. Temperature history data of several weather stations in each state are averaged to train our ML model. Because temperature data in the unit of Fahrenheit have a different order of magnitude compared to the other features, they were normalized by dividing by 100 before being used in the training.

4 Methods

4.1 Initial model, training, and validation

The initial model was trained using the three most recent x_v and four most recent x_p, x_h as the input. x_p and x_v were one-dimensional in all models, and x_v was either one- or two-dimensional depending on the model. Since x_p behaves as both input and output, we recursively trained the algorithm by implementing the batch gradient for several epochs. The learning rate (10^{-2}) and number of epochs (100,000) were chosen after manually examining the average quadratic loss throughout the training, which was calculated as below:

$$J_{state}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - h_\theta(x^{(i)})\right)^2. \tag{1}$$

Training was done with the California (CA) data, because we thought it is the best representative case for several reasons. First, CA has the largest population, the number of positive cases, and the number of vaccine doses administered. Second, the overall shape of the dataset is smooth (whereas TX and GA has few noisy patterns), thus having a lower risk of overfitting. The model parameters obtained from training with CA dataset were applied to each of the 10 state dataset.

4.2 Model selection

Several different models have been tested to improve the performance of the program. Briefly, in comparison to the initial model (x_p, x_h, and x_v only), additional factors (x_t and three x_{var}) were included, or x_h was normalized differently (set 0 as the default and square the value). Below are the list of models explored in this study:

1. x_p, x_h, x_v only (Preliminary model)
2. x_p, x_h, x_v + for x_h set 0 as the default and square the value (**best**)
3. x_p, x_h, x_v + include x_t
4. x_p, x_h, x_v + include three x_{var}
5. x_p, x_h, x_v + include x_t and three x_{var} (**worst**)

The best model was chosen based on the overall performance on 10 state datasets. Explicitly, the state performance of a model, $P_{model-state}$, was defined as the ratio of the loss of a model to the loss pf the preliminary model. The overall performance of a model, $P_{model-overall}$, was defined as the product of 10 ratios and was used to gauge the overall performance of the model.

$$P_{model-state} = J_{model-state}(\theta_{model})/J_{prelim-state}(\theta_{prelim}) \tag{2}$$

$$P_{model-overall} = \prod_{state} P_{model-state} \tag{3}$$

Thus, $P_{model-overall} < 1$ means the model is better than the model in the preliminary result.

5 Results and Discussions

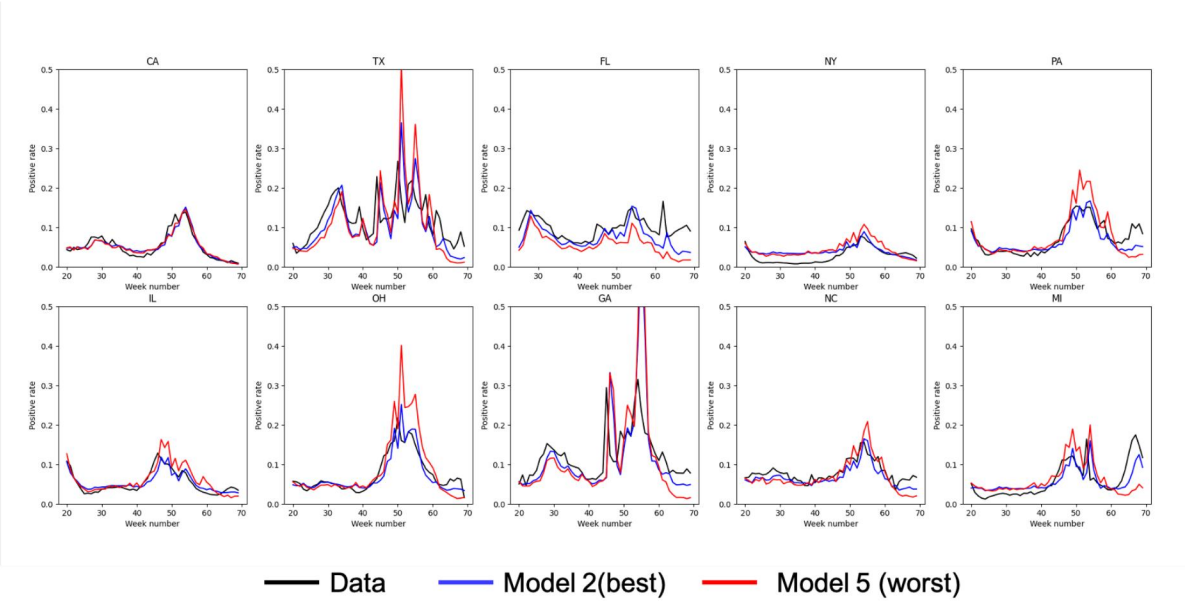

Figure 1: Measured positive rate (black line) and the predicted positive rate using the Poisson regression trained with CA data

Model No.	P_{CA}	P_{TX}	P_{FL}	P_{NY}	P_{PA}	P_{IL}	P_{OH}	P_{GA}	P_{NA}	P_{MI}	$P_{overall}$
1 (Prelim)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.00
2	1.00	0.86	1.04	1.08	1.10	1.04	0.93	0.73	1.00	1.08	0.80
3	0.85	1.33	1.49	1.54	1.77	2.49	2.89	1.16	1.29	1.35	66.5
4	1.01	1.12	1.57	0.99	1.74	1.16	1.35	1.03	1.37	2.14	14.2
5	0.84	1.51	2.23	1.48	2.78	2.55	4.07	1.19	1.82	2.58	680

Table 1: Summary of simulation results: Performance of each model

The red lines in each plot in Figure 1 shows the predicted test positive rate, which shows a moderate predictive capability for all 10 states. The values for each of the optimized parameter were in line with our expectations. For x_p, the parameter for the most recent past was the biggest, as patients infected in the near past would be mostly likely to show high viral titers and capable of infecting others. The parameters for x_v showed a large negative value, suggesting that vaccination strongly reduces disease spread.

We were able to get some improvement from the preliminary model after using different normalization for x_h. In the preliminary model, x_h has the default of 2 (Sat, Sun each week), and linearly increases with additional holidays. In model 2, we set x_h to have 0 as the default, and to increase quadratic manner. This was based on the assumption that the degree of outdoor activity would increase exponentially, rather than linearly, with the 'additional' holidays in a week. Indeed, such modification improved the model performance, and we think this enabled the model to efficiently predict the sudden surge in test positive rate after holidays.

To our surprise, including other features - average temperature (model 3) or the portion of variants (model 4) - did not improve the model. For including variant data in the input, the parameters even showed negative values in some cases (data not shown). This means the rise of variants in the US is actually a negligible feature. It is understandable since (i) although individual antibodies perform poorer in neutralizing the variant, the actual immune system works in a 'cocktail' of various antibodies, and (ii) a very recent clinical trial shows that BNT162b2 vaccine (the Pfizer-BioNTech vaccine) still has a high efficacy against variants of concern [17].

Feature No.	**Meaning**	θ
0	Intercept (x_0)	-3.459
1	Positive rate (week -4)	-2.609
2	Positive rate (week -3)	0.191
3	Positive rate (week -2)	3.877
4	Positive rate (week -1)	7.539
5	# holidays (week -4)	-0.020
6	# holidays (week -3)	0.031
7	# holidays (week -2)	0.004
8	# holidays (week -1)	0.086
9	Portion vaccinated (week -2)	-1.538
10	Portion vaccinated (week -1)	-2.226
11	Portion vaccinated (week -0)	-3.111

Table 2: Individual parameters for the best-performing model (Model 2)

6 Conclusion & Future work

In this project, we proposed a Poisson regression model for predicting the future spread of COVID-19 based on macroscopic factors of the past. Taking three main factors: (i) level of community spread, (ii) community immunity, and (iii) public behavior as the input, we have proposed a trained algorithm that moderately predicts the disease spread across different states in the US. As expected, the level of community spread at recent past affected the future disease spread more than the distant past, and the increasing portion of fully vaccinated people strongly suppressed the disease spread. Incorporating the portion of variants did not improve model accuracy, implying it does not significantly improve the disease spread in the US, and emergence of variants of concern can be still quelled by national vaccination campaigns.

Future improvements could be made on the features related to the public behavior. In this study, we only used the number of extra holidays as the yardstick. We've hypothesized that adding the weather factor will improve the accuracy of the model, but due to the large variation between average temperatures, including it only worsened the overall quality. Although we weren't able to push further due to the time constraint, we strongly suggest that the weather should have some impact on the disease spread, and trying other relevant features (e.g. rain/humidity, deviation from average temperature rather than absolute temperature, air quality) in the model should be explored in the future. Last but not least, applying different ML algorithms else than Poisson regression (Deep learning, SIR/SEIR model, etc.) would be also worth pursuing.

7 Contributions

Y.H. and H.R. equally contributed to the overall design of the project and discussions. Y.H. focused more on data clean-up and code implementation, and H.R. focused more on literature digging on coronavirus biology and writing.

References

[1] World Health Organization. WHO COVID-19 Dashboard. https://covid19.who.int.

[2] Furqan Rustam, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byung-Won On, Waqar Aslam, and Gyu Sang Choi. Covid-19 future forecasting using supervised machine learning models. *IEEE Access*, 8:101489–101499, 2020.

[3] Farah Shahid, Aneela Zameer, and Muhammad Muneeb. Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm. *Chaos, Solitons Fractals*, Aug 2020.

[4] Gaurav Pandey, Poonam Chaudhary, Rajan Gupta, and Saibal Pal. Seir and regression model based covid-19 outbreak predictions in india. *arXiv preprint*, Apr 2020.

[5] R. Sujath, Jyotir Moy Chatterjee, and Aboul Ella Hassanien. A machine learning forecasting model for covid-19 pandemic in india. *Stochastic Environmental Research and Risk Assessment*, Jul 2020.

[6] Gergo Pinter, Imre Felde, Amir Mosavi, Pedram Ghamisi, and Richard Gloaguen. Covid-19 pandemic prediction for hungary; a hybrid machine learning approach. *Mathematics*, 8(6):890, 2020.

[7] Raj Dandekar and George Barbastathis. Quantifying the effect of quarantine control in covid-19 infectious spread using machine learning. *medRxiv*, Apr 2020.

[8] Zohair Malki, El-Sayed Atlam, Aboul Ella Hassanien, Guesh Dagnew, Mostafa A. Elhosseini, and Ibrahim Gad. Association between weather data and covid-19 pandemic predicting mortality rate: Machine learning approaches. *Chaos, Solitons Fractals*, Jul 2020.

[9] Giulia Giordano, Marta Colaneri, Alessandro Di Filippo, Franco Blanchini, Paolo Bolzern, Giuseppe De Nicolao, Paolo Sacchi, Patrizio Colaneri, and Raffaele Bruno. Modeling vaccination rollouts, sars-cov-2 variants and the requirement for non-pharmaceutical interventions in italy. *Nature Medicine*, 2021.

[10] Roman Wölfel, Victor M Corman, Wolfgang Guggemos, Michael Seilmaier, Sabine Zange, Marcel A Müller, Daniela Niemeyer, Terry C Jones, Patrick Vollmar, Camilla Rothe, et al. Virological assessment of hospitalized patients with covid-2019. *Nature*, 581(7809):465–469, 2020.

[11] Fernando P. Polack, Stephen J. Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L. Perez, Gonzalo Pérez Marc, Edson D. Moreira, Cristiano Zerbini, Ruth Bailey, Kena A. Swanson, Satrajit Roychoudhury, Kenneth Koury, Ping Li, Warren V. Kalina, David Cooper, Robert W. Frenck, Laura L. Hammitt, Özlem Türeci, Haylene Nell, Axel Schaefer, Serhat Ünal, Dina B. Tresnan, Susan Mather, Philip R. Dormitzer, Uğur Şahin, Kathrin U. Jansen, and William C. Gruber. Safety and efficacy of the bnt162b2 mrna covid-19 vaccine. *New England Journal of Medicine*, 383(27):2603–2615, 2020. PMID: 33301246.

[12] United States Census Bureau. U.S. Population Clock. https://www.census.gov/popclock/.

[13] Wikepedia. COVID-19 pandemic in the United States. https://en.wikipedia.org/wiki/COVID-19_pandemic_in_the_United_States.

[14] Johns Hopkins Centers for Civic Impact for the Coronavirus Resource Center (CRC). COVID-19 data. https://github.com/govex/COVID-19.

[15] Scripps Research. SARS-CoV-2 (hCoV-19) Mutation Reports, Lineage-Mutation Tracker. https://outbreak.info/situation-reports.

[16] National Oceanic and Atmospheric Administration (NOAA). Climate of the US. https://www.ncdc.noaa.gov/climate-information/climate-us.

[17] Laith J. Abu-Raddad, Hiam Chemaitelly, and Adeel A. Butt. Effectiveness of the bnt162b2 covid-19 vaccine against the b.1.1.7 and b.1.351 variants. *New England Journal of Medicine*, 0(0):null, 0.