OneQuietNight Covid-19 Forecast

Authors	Areum Jo (areumjo1@gmail.com), Jae Cho (jaehun.cho@gmail.com)
Date	2020-11-18

Introduction

Our ability to contain the coronavirus pandemic depends on being able to forecast potential outbreaks.

In this work, we develop scientifically-driven machine learning models to accurately predict the spread of Covid-19 infections using real-time data.

C3 AI Covid-19 Data Lake[1] collects and organizes various data sets that may bear on the spread of Covid-19 -- daily case reports, movement trends, weather reports, and economic changes. Our models use this data to make predictions about future increases in Covid-19 cases at the county, state, and national levels in the United States.

Problem Description

The official CDC Covid-19 forecast[2] uses an ensemble of models to predict the number of new cases that are likely to arise in different geographic locations. The CDC Covid-19 forecast predicts the number of new Covid-19 cases per week for the next 4 weeks at the national, state, and county levels. It currently combines the forecasts from 21 modeling groups.

To aid in this effort, we develop and operationalize an accurate Covid-19 forecast based on the time series data in the C3 AI Covid-19 Data Lake. Our forecast is competitive and outperforms some well-established models in backtests. We visualize this data using an interactive web application.

Broad Approach

Coronavirus is thought to spread from person to person. A typical case starts with a person coming into contact with a patient, who may not have symptoms. The virus has a chance to

spread to the person during each contact. When the virus is successful, the person becomes infected and infectious to other people. The virus spreads exponentially in this way.

Epidemiological models use the structural knowledge of the spread of the virus to make predictions using the number of infected individuals and the number of transmittive contact. But it is difficult to measure how many infected individuals there really are and with whom they had close contact. Instead, we only have some imperfect measurements of a set of inputs that may have bearing on these components.

In order to learn the useful relations between variables with limited data, we use machine learning models with scientifically-driven features. We find that temporal and spatial features of the daily case reports and movement trends data predicts future Covid-19 cases. Our models use these to make predictions for all counties, states, and the country for the next 4 weeks.

Technical Details of the Approach

We model the number of new cases per week in the next week for region i, $y_i(t+1)$, using scientifically-driven features of the C3 AI Covid-19 data, $X(t)$.

We develop the following features:

- Sum of new cases per 100,000 people per week[1,9,10].
- Sum of new hospitalizations per licensed beds per week[1,9,10].
- Average movement trends[1,11,12].

We impute features with top-down and bottom-up hierarchical aggregations and Census Core-Based Statistical Area aggregations[13]. We stabilize features by applying winsorizations across hierarchies. We also use one-week lagged versions of each of these features to capture their dynamics. We did not include other features because they did not work consistently or intuitively or because they did not help improve the predictions when used with the features above.

We develop forecasts using different machine learning models. The models are optimized using the data set of $X(t)$ and $y_i(t+1)$ from a moving window of N_{train} training weeks, $[t-N_{train}-N_{test},\ t-N_{test})$, and evaluated on a moving window of N_{test} testing weeks, $[t-N_{test},\ t]$ in a walk-forward backtest from 2020-06-15 to 2020-09-14. We reserve 8 instances from 2020-09-14 to 2020-11-02 for validation. We find that a linear regression of the features against new cases per 100,000 people for each horizon and for each level generates the best results. These predictions are multiplied by the population in the region and combined together to produce the final forecast. The final forecast uses the optimized models trained on $t \in [t-N_{train},\ t]$ to predict $y_i(t+1)$.

Results

We compare our forecast to all the models from the CDC Covid-19 forecasts and show that our forecast is competitive and outperforms some well-established forecasts in the backtest. We backfilled our forecast by training on data up to the forecast date and making predictions with the inputs available on the forecast date in the validation period from 2020-09-14 to 2020-11-02. We used the historical forecasts from the Covid-19 Forecast Hub[3] in the same period.

Table 1. Forecasting accuracy of state-level forecasts between 2020-09-14 and 2020-11-02. We computed the mean absolute error using the daily reports containing the cases data from the JHU CSSE group as the gold standard reference for the cases in the US. The mean absolute error numbers for each of the forecast horizons are shown below (lower is better).

Forecast Horizon (Weeks)	One Quiet Night	CU-select[4]	Covid Analytics DELPHI[5]	DDS-NBDS	Google_Harvard-CPF[6]	LANL-Growth Rate[7]	LNQ-ens1	UCLA-SuEIR[8]	COVID hub-ensemble
1	1303.24	1678.55	4249.94	1650.95	3091.59	1828.46	1157.68	2016.25	1588.03
2	2608.21	3109.91	6297.43	2700.73	6785.18	3913.84	2470.56	4779.24	3271.90
3	3571.28	4380.80	7582.13	4221.48	8716.45	6389.37	3647.93	6361.37	4478.32
4	4760.87	5712.37	9060.10	7283.32	10698.24	8087.49	5129.04	7672.48	5823.79

As shown in Table 1, our OneQuietNight forecast generates accurate results across all horizons in the backtest. Our approach is different from the empirical models and dynamical models that are commonly used in the Covid-19 forecasts in that it does not make any forward-looking assumptions about the factors affecting transmission. Instead, it uses the historical dynamics between the number of cases and people's movement levels to make the forecasts. This tends to produce waves of Covid-19 peak cases rather than a continued increase over a four week time frame based on the historical patterns.

Impact

We develop scientifically-driven machine learning models to accurately predict the spread of Covid-19 infections using real-time data from the C3 AI Covid-19 Data Lake. This generates accurate forecasts that are competitive with and different from the current set of models in the CDC Covid-19 ensemble. We operationalize the forecast to retrain the model and make predictions on new data. We publish this data through a web application to help slow the pandemic and prevent future ones.

References

1. C3 AI COVID-19 Data Lake. https://c3.ai/customers/covid-19-data-lake/. Accessed: 2020-11-18.
2. Centers for Disease Control and Prevention. Covid-19 Mathematical Modeling. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/mathematical-modeling.html. Accessed: 2020-11-18.
3. Covid-19 Forecast Hub. https://covid19forecasthub.org/. Accessed: 2020-11-18.
4. Zou, D., et al. Epidemic Model Guided Machine Learning for COVID-19 Forecasts in the United States medRxiv, doi:10.1101/2020.05.24.20111989.
5. Pei, S. and J. Shaman. Simulation of SARS-CoV2 Spread and Intervention Effects in the Continental US with Variable Contact Rates, March 24, 2020.
6. Carnegie Mellon Delphi Group. https://delphi.cmu.edu/. Accessed: 2020-11-18.
7. Arık, S. Ö., et al. Interpretable sequence learning for Covid-19 forecasting. arXiv:2008.00646 (2020).
8. Castro, L., et al. COFFEE: COVID-19 Forecasts using Fast Evaluations and Estimation. https://covid-19.bsvgateway.org/. Accessed: 2020-11-18.
9. Ensheng Dong, H. D. and Gardner, L. An interactive web-based dashboard to track covid-19 in real time. The Lancet Infect. Dis. 20, 533–534 (2020).
10. The COVID Tracking Project. https://covidtracking.com/. Accessed: 2020-11-18.
11. Apple. Mobility Trend Reports. https://covid19.apple.com/mobility. Accessed: 2020-11-18.
12. Google. COVID-19 Community Mobility Reports. https://www.google.com/covid19/mobility/. Accessed: 2020-11-18.
13. Census.gov. Metropolitan and Micropolitan Statistical Area Reference Files. https://www.census.gov/geographies/reference-files/time-series/demo/metro-micro/delineation-files.html. Accessed: 2020-11-18