Using Machine Learning to Analyze COVID-19 Case Numbers

Simon Chow, Aruneshwar Venkat, Gabriel Fonseca, Jackie Vo

April 28, 2022

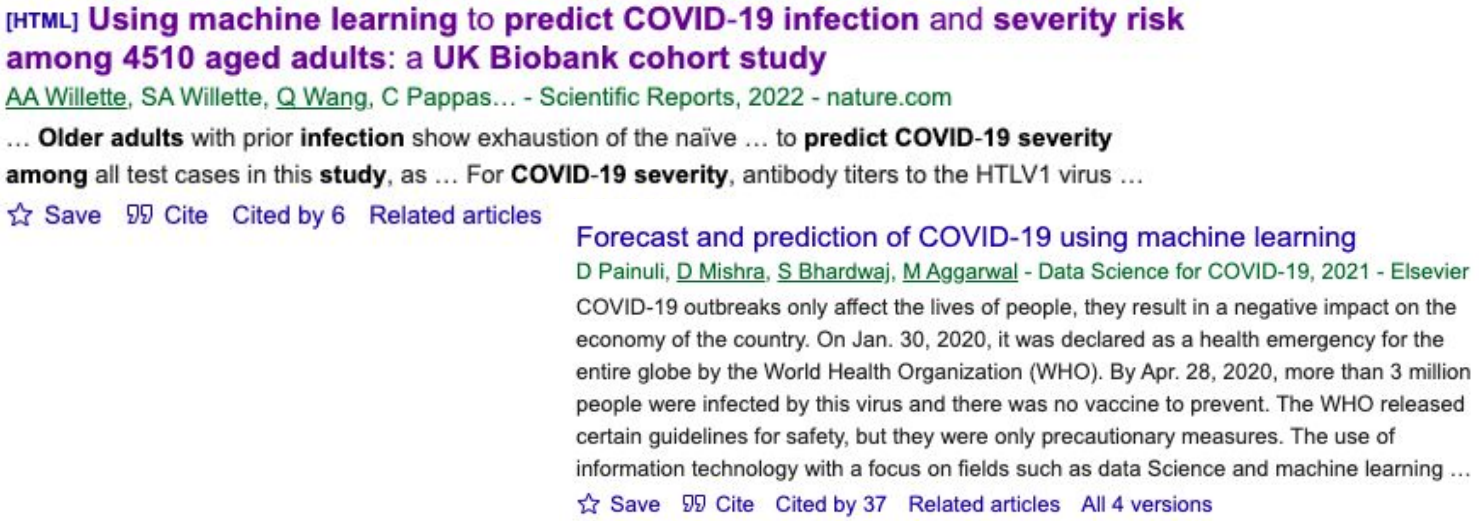
2020 …


When you are laughing at all the corona memes but the laughing suddenly turns into coughing


The Anonymous Hashtagoholic
@einfreakinstein
A Cone Like My Dog Has So I Can't Touch My Face.
#ThingsINeedInMyCOVID911Kit
GIF
11:15 AM · Mar 11, 2020 · Twitter Web App

Introduction

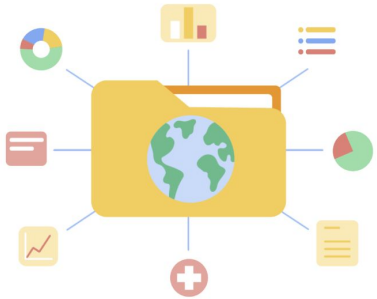
- Using COVID-19 epidemiological database to analyze disease severity in different states at different times
- Predict severity of COVID-19 at a certain time at certain location.



Data Source:

- Google COVID-19 Open Data Repository
- Provides general demographic information, other COVID-19 related information such as vaccination status, death etc.
- Focus only on United States data for analysis

COVID-19 Open Data Repository



The Google Health COVID-19 Open Data Repository is one of the most comprehensive collections of up-to-date COVID-19-related information. Comprising data from more than 20,000 locations worldwide, it contains a rich variety of data types to help public health professionals, researchers, policymakers and others in understanding and managing the virus.

Methods

1. K-means clustering
2. Time series Poisson regression
3. Time series with Random Forest
4. Time series with SVM

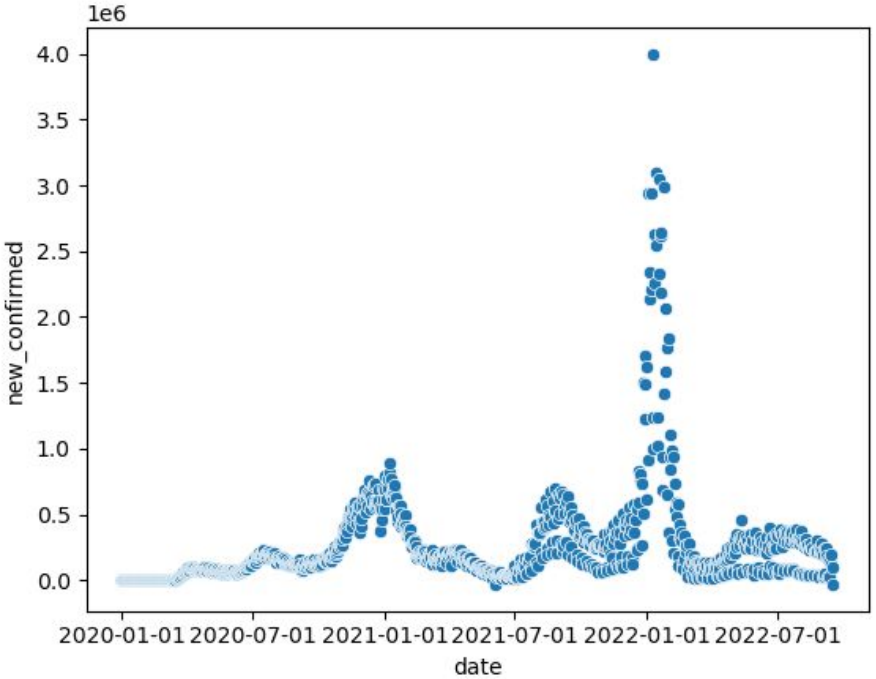
Method 1: K Means Clustering

Overview - K-means clustering

- Focused on a few features to cluster for predicting new covid cases over time
    - New deceased, new people vaccinated, new people fully vaccinated, new hospitalized patients, new intensive care patients
- N = 640 time points
- We clustered the actual new covid cases to generate true labels

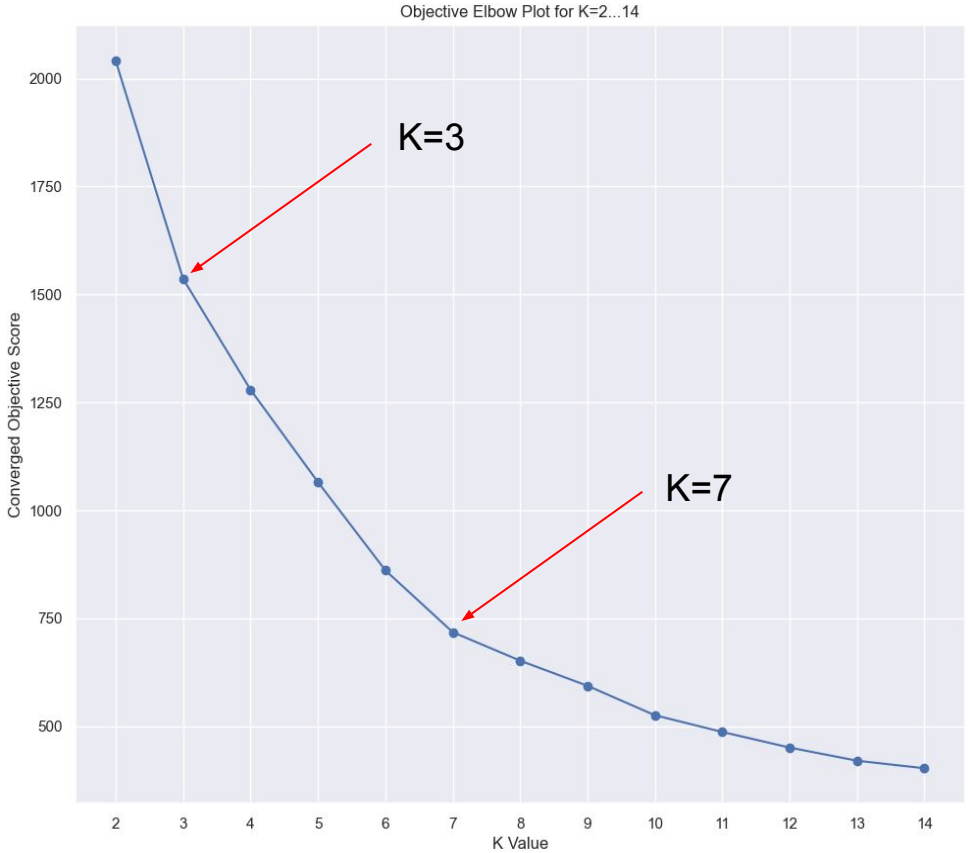
Results - K-means clustering

New cases over times plot
(no k means yet)



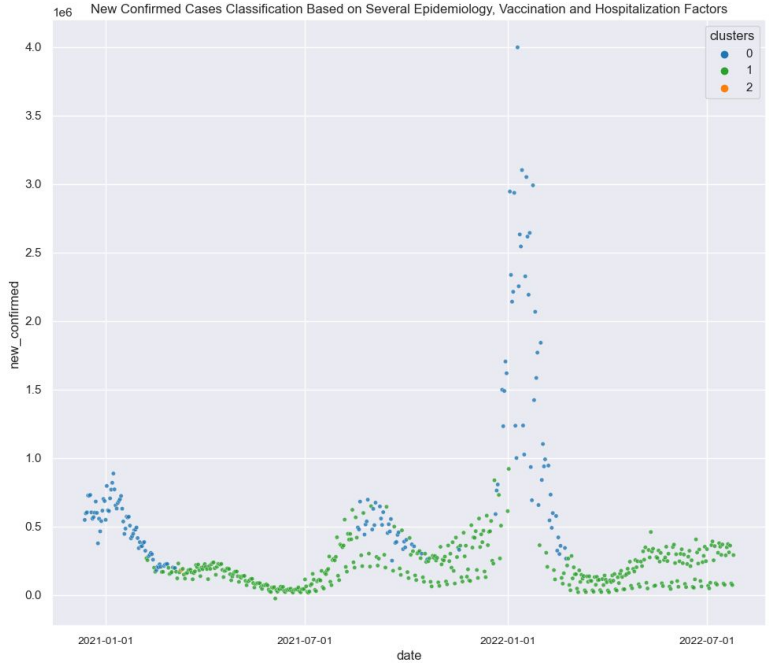
Results - K-means clustering

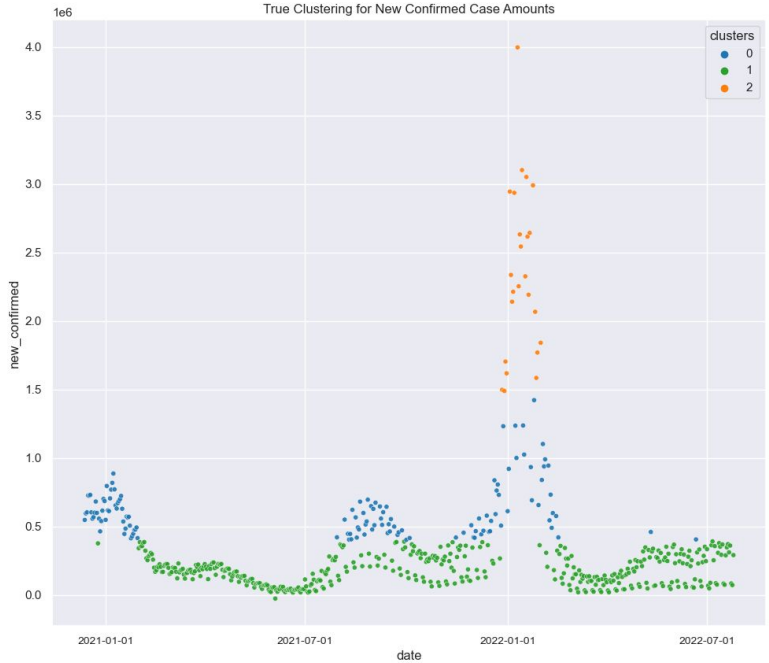
Objectives
Elbow Plot



Results - K-means clustering

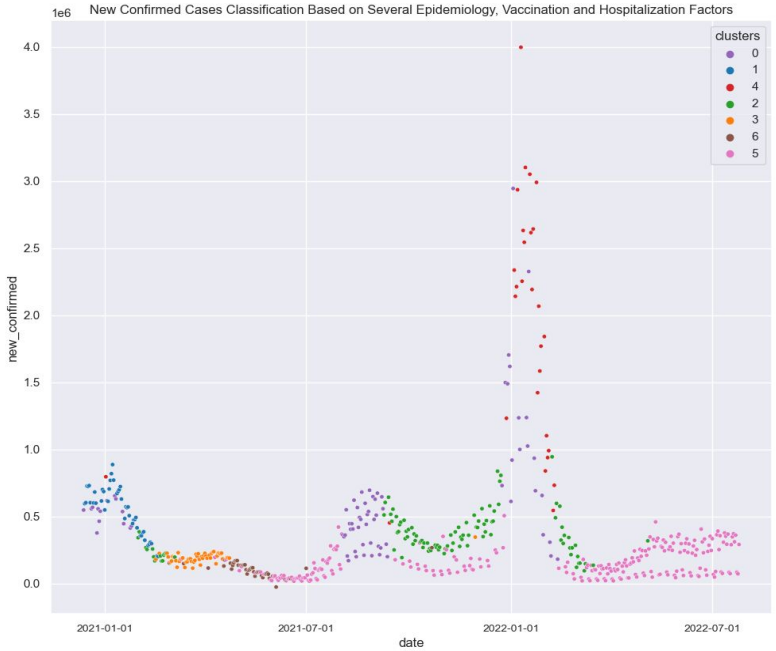
With k = 3, accuracy = ~83.1%

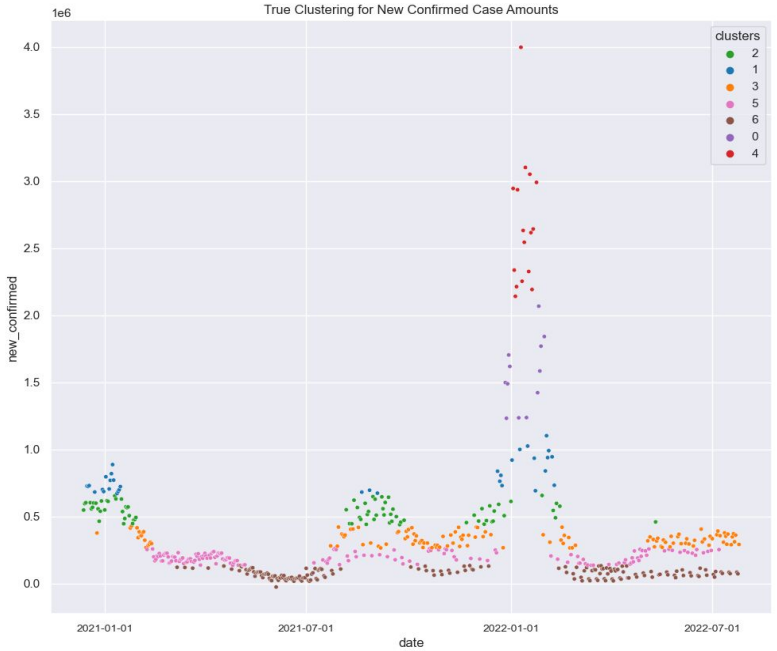




Results - K-means clustering

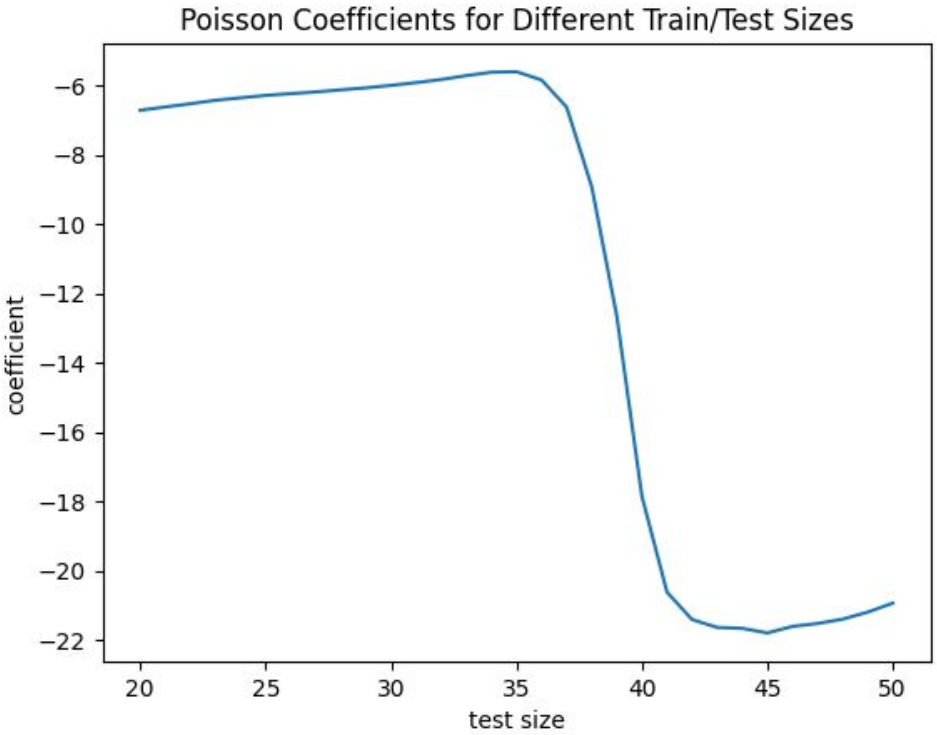
With k = 7, accuracy = ~24.9% (overfitting)

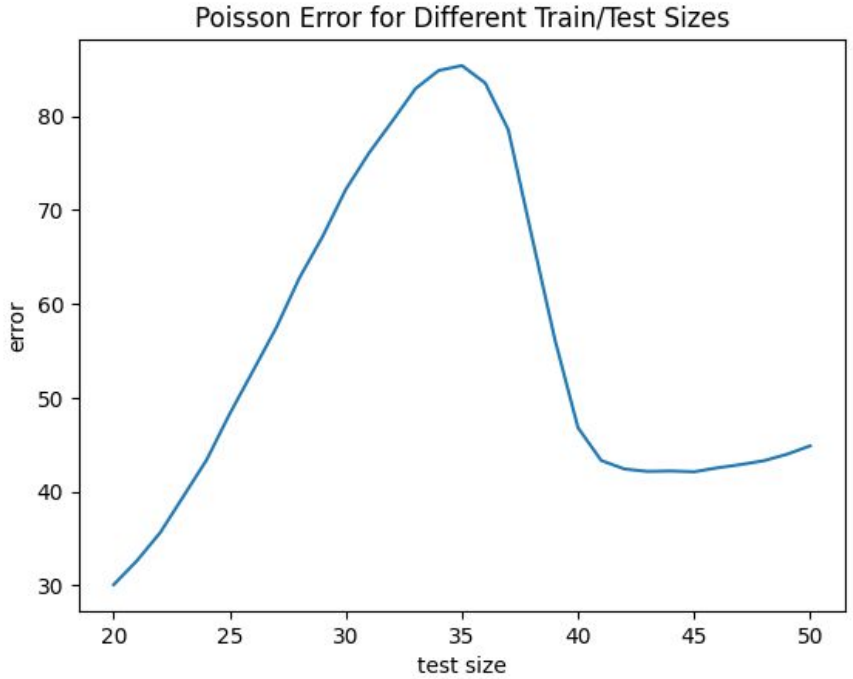




Method 2: Poisson Regression

Results - Poisson Regression



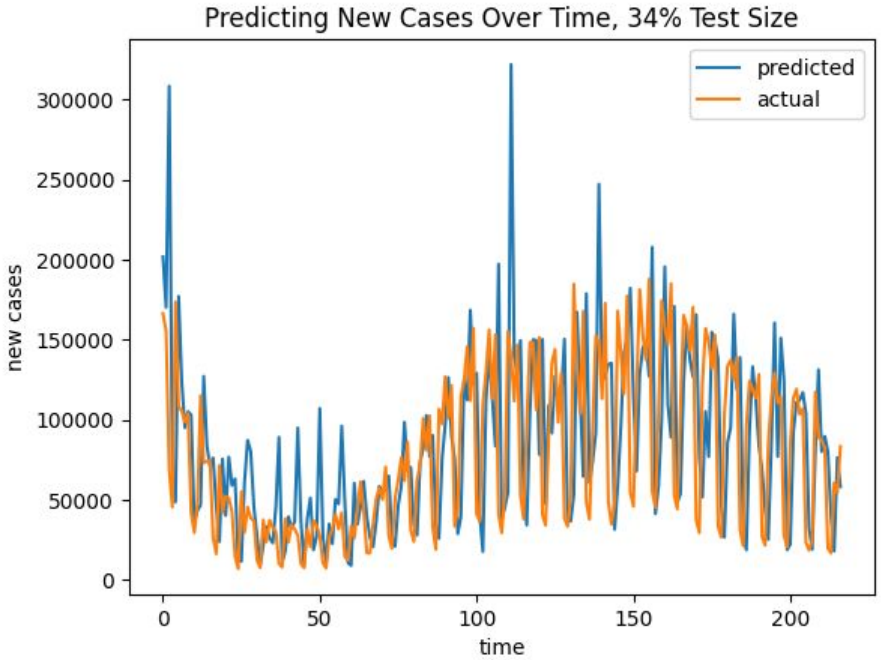


Methods 3 and 4: Time Series with Random Forest and SVM

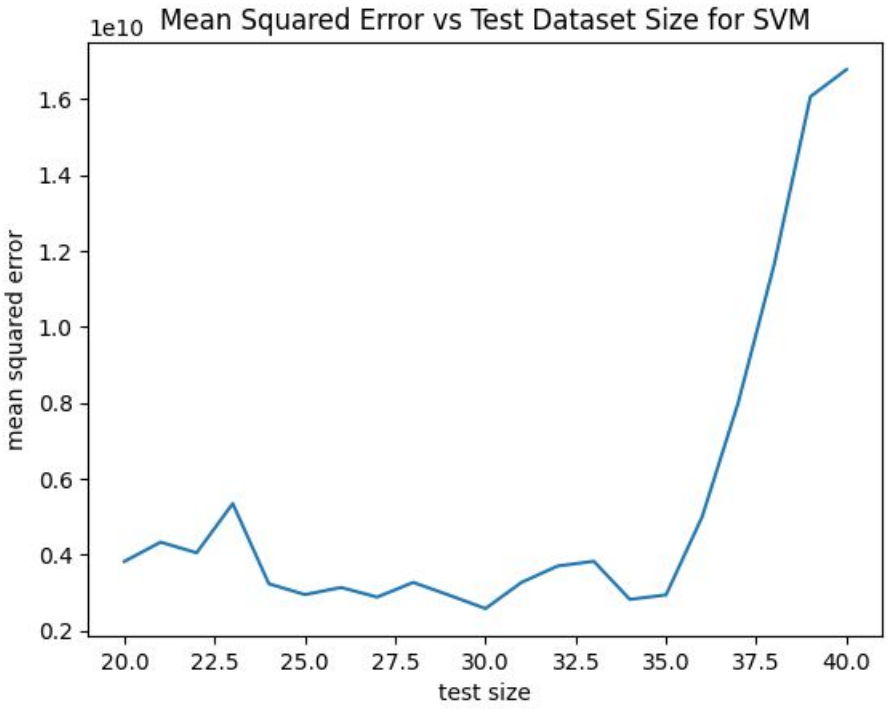
Overview - Time Series

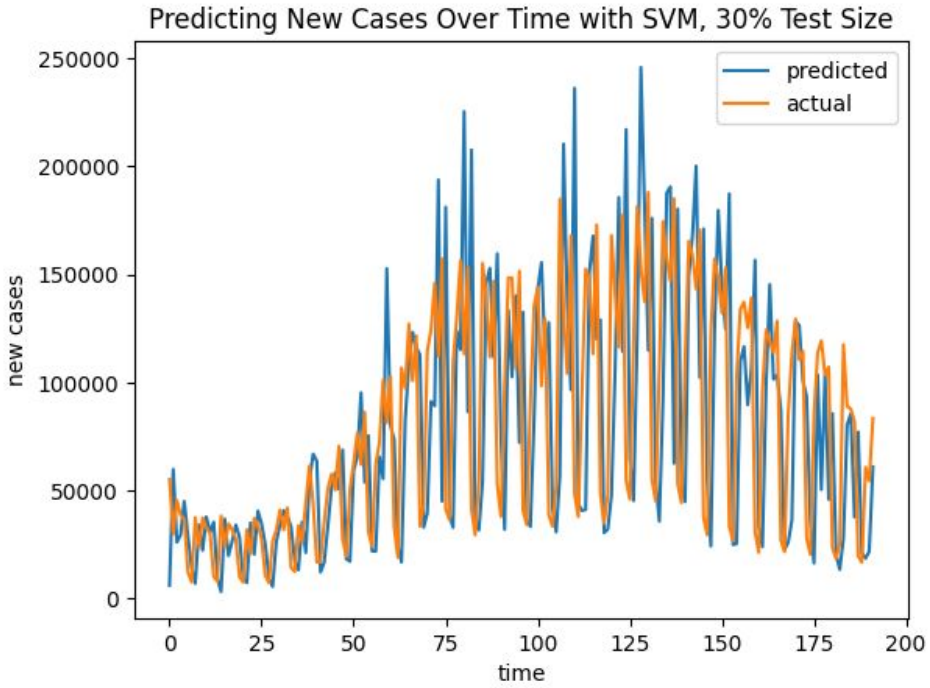
- N = 640 time points
- Converted data to sliding window
  - First time point is x, second time point y
  - Second time point is x, third is y, etc.
- Used walk-forward validation for accuracy

Results - Time Series w. Random Forest



Results - SVM


Mean Squared Error vs Test Dataset Size for SVM
1e10
mean squared error
test size
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
40.0


Predicting New Cases Over Time with SVM, 30% Test Size
predicted
actual
new cases
time
0
50000
100000
150000
200000
250000
0
25
50
75
100
125
150
175
200

Limitations

- Time series: not enough data points
- No consistent trend makes it difficult to predict
- Possible confounding variables affecting case rates
    - Hard to fit linear models

Future Directions

- Extend the analysis to worldwide
- Adjust parameters for SVM, Random Forest
- Incorporate more features into regression


I JUST WANNA
SAY...
THANK YOU.
HAHAHAHAHAHAHAH.

References

Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time [published correction appears in Lancet Infect Dis. 2020 Sep;20(9):e215]. *Lancet Infect Dis*. 2020;20(5):533-534. doi:10.1016/S1473-3099(20)30120-1

O. Wahltinez, et al. “COVID-19 Open-Data: curating a fine-grained, global-scale data repository for SARS-CoV-2”. In: (2020). URL: https://goo.gle/covid-19-open-data.

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

Willette AA, Willette SA, Wang Q, Pappas C, Klinedinst BS, Le S, Larsen B, Pollpeter A, Li T, Mochel JP, Allenspach K, Brenner N, Waterboer T. Using machine learning to predict COVID-19 infection and severity risk among 4510 aged adults: a UK Biobank cohort study. Sci Rep. 2022 May 11;12(1):7736. doi: 10.1038/s41598-022-07307-z. PMID: 35545624; PMCID: PMC9092926.