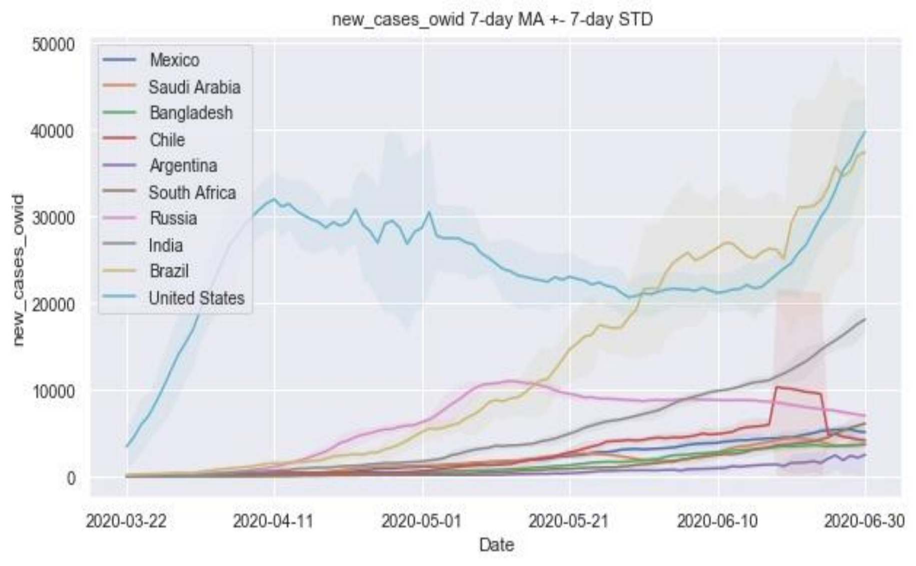
COVID-19 case number modeling

By: Matthew Gudorf

Project motivation and proposed solutions

- 7-day moving average time series for the number of new cases shows that pandemic is not yet under control
- Well informed epidemiological models exist BUT they typically require a high level of expertise to implement



My proposed methods

- Take a data-driven model approach, comparing models of varying complexity: linear regression, fully connected neural network, convolutional neural network

Data

1. John's Hopkins (JHU CSSE) : COVID-19 specific variables (cases, deaths, etc.)
2. OxCGRT : Government reponses to COVID-19 pandemic. (in the form of time series of discrete numerical variables)
3. OWID : COVID-19 specific variables (cases, tests, etc.) as well as time independent variables such as population, smoking rate, and many more
4. FIND: COVID-19 tests and cases; cases taken from JHU CSSE dataset

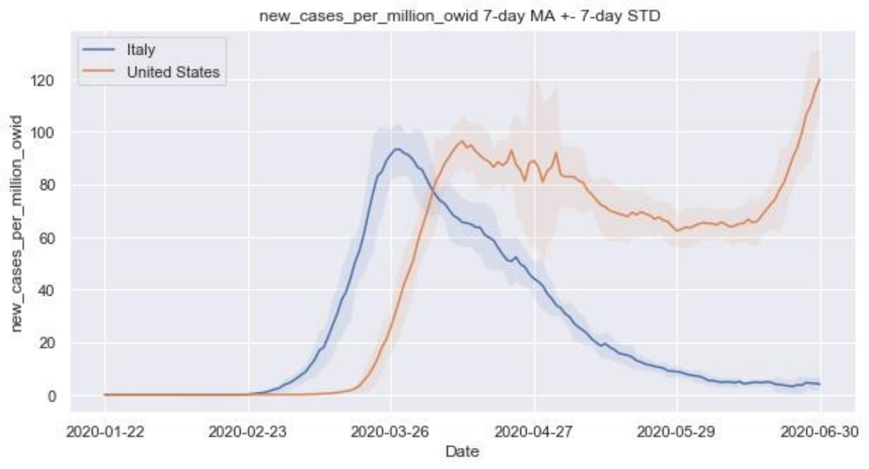
Data : cleaning and wrangling steps

1. Aggregate the multiple datasets
2. Regularize the names, dates, countries, data formats
3. Account for missing and pathological values
4. Determine which variables to use for modeling purposes (feature selection)
5. Engineer new features to better convey time dependence to the ridge regression model using moving averages

Data exploration

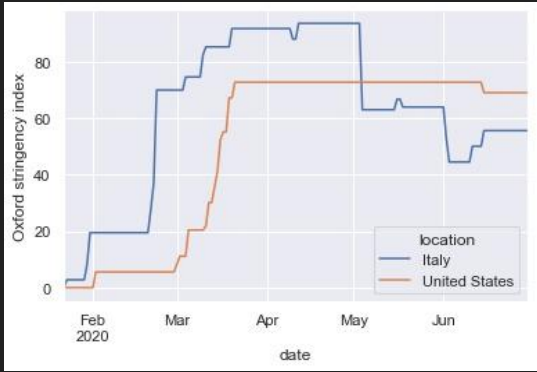
Investigate the effect of different government mandates. This information will help with the following

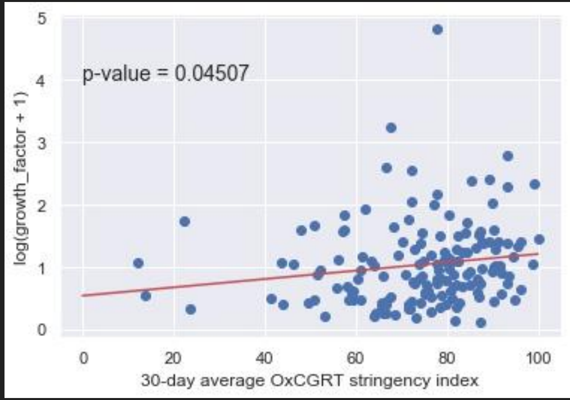
1. Feature selection for modeling
2. Determining actionable steps for governments
3. Investigate the most effective measures against future pandemics, not just COVID-19.



Growth Factor vs. Stringency

- The stringency index (defined on interval [0,100]) is a quantification/aggregation of 7 different quarantine measures (labeled S1-S7 in the OxCGRT dataset).
- Growth factor is the ratio of 30-day averages; in this instance it is the most recent 30 days and next to last 30 days (i.e. 60 days ago to 30 days ago).
- Univariate regression shows a statistically significant relationship between log(growth factors +1) and the stringency.





Stringency and its components

- Two-way ANOVA analysis, using countries and components of the stringency index
- A sample in this case is the growth factor relative to the implementation date (i.e. ratio of average new cases before and after the implementation).
- ANOVA shows that he growth factor is dependent on both the mandate type and the country

	sum_squares	Degrees of freedom	F-statistic	PR(>F)
Mandate type	0.551935	7	3.008117	0.004305
Country	2.864245	59	1.852094	0.000312

Conclusion:

The set of effective social distancing measures is unique to each country, and so an idiosyncratic approach is required.

Features used in each model.

The shape of the data (before adhering to keras or scikit-learn conventions)

(n_frames, n_countries, n_timesteps, **n_features**)

For the neural network models : (155, 146, 28, 3)

For the ridge regression model: (155, 146, 28, 14)

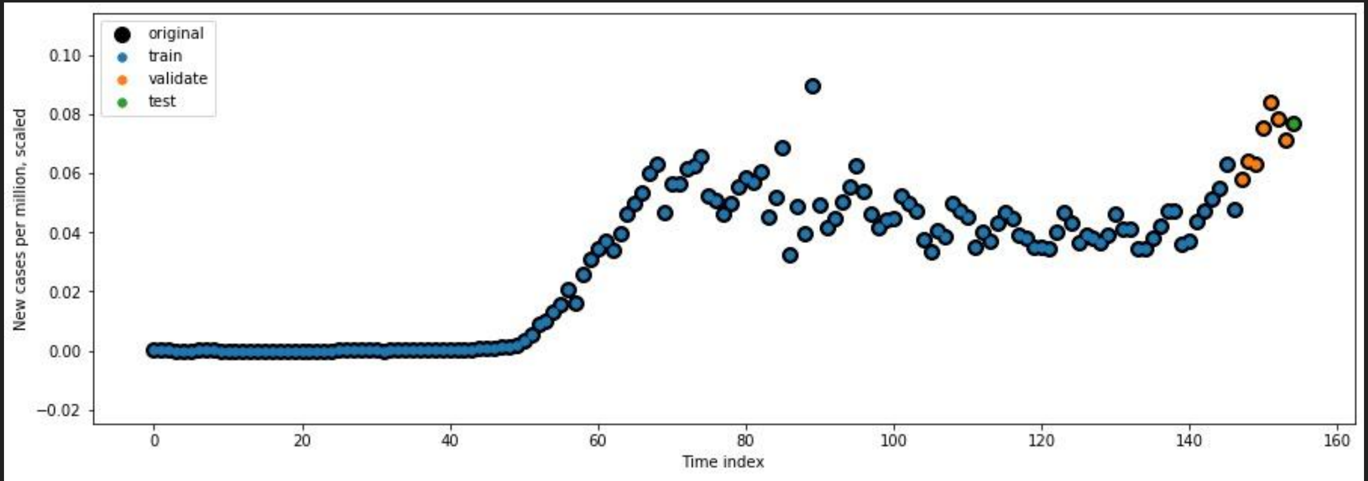
Ridge regression features = neural network features along with their moving averages

Split of the data:

1. Training set contains 147 frames
2. Validation set contains 7 frames
3. Testing (hold-out) set contains 1 frame.

Rescaling the data

Rescale the data with respect to some arbitrary maximum value such that feature data is rescaled to the interval [0, 0.5]. The data displayed here is only a partial subset, hence why the maximum is not 0.5



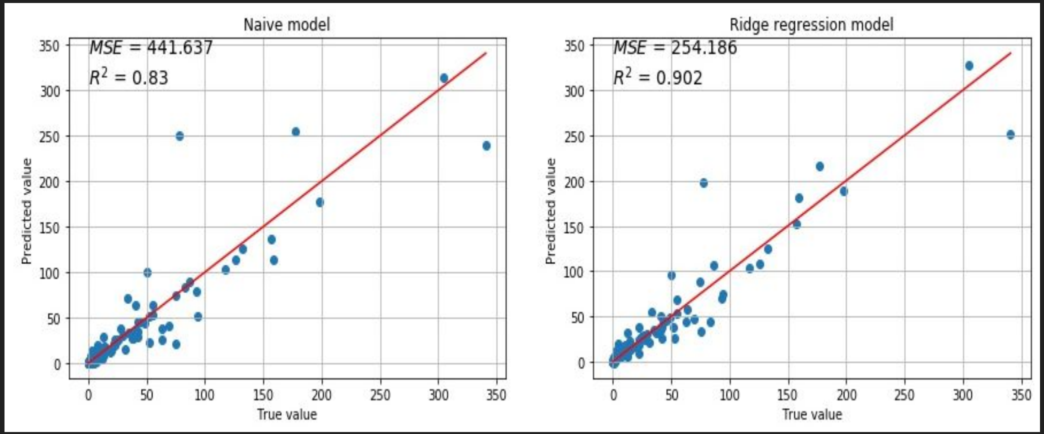
Modeling

Three different models implemented:

1. Ridge regression
2. Neural network with two fully connected layers
3. Neural network with two convolutional followed by two fully connected layers

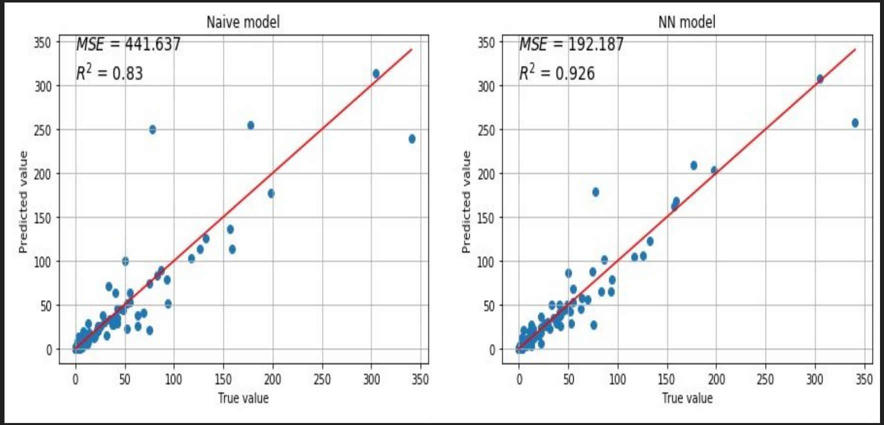
Ridge Regression

- Good performance for amount of effort required.
- Analysis of coefficients shows that government response index makes a significant contribution.
- Cross validation resulted in a stronger than default regularization constant.



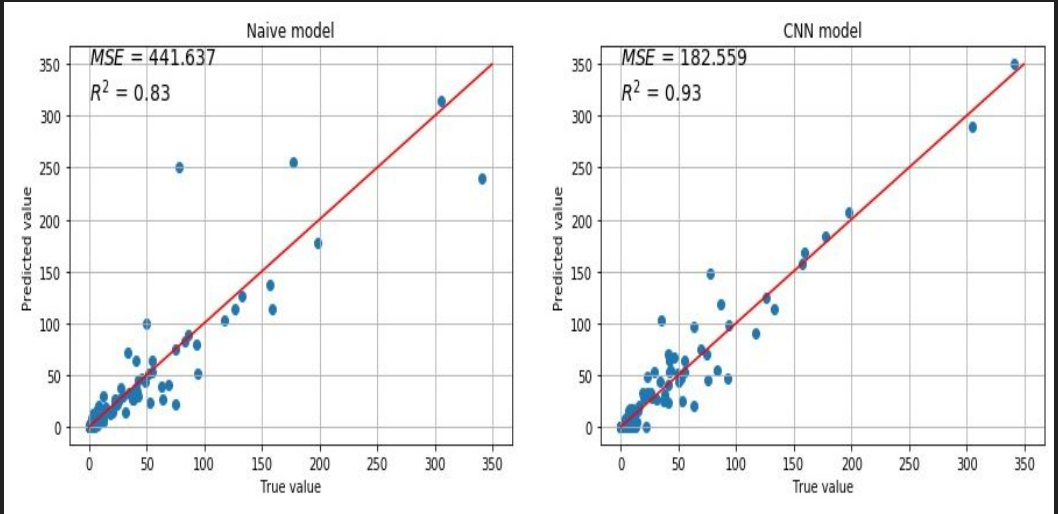
Fully connected neural network performance

- Second best performance
- Validation process chose model architecture with large number of parameters, approximately 7000
- Needs more parameter tuning



Convolutional neural network performance

- Best performance as of the creation of this presentation
- On the order of 1000 parameters
- Needs more parameter tuning



Conclusion

Neural networks performed better than Ridge regression, but took more effort to deploy.

My recommendations

1. For maximum accuracy, use a CNN model
2. For minimal effort, use a Ridge regression model
3. The growth factor is affected by on the quarantine mandate type and country; meaning an individualized approach is needed
4. As the number of training samples increases, more complex neural networks (more parameters) can be explored without fear of overtraining; this might lead to even better models in terms of accuracy, and should be explored.