COVID-19 case number modeling

By: Matthew Gudorf

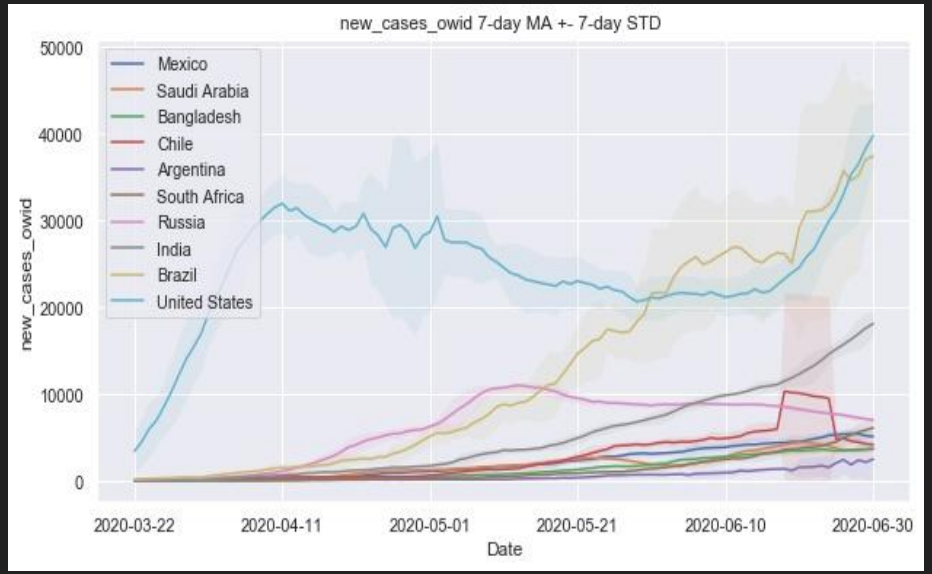
Project motivation

COVID-19 has taken a dramatic toll on the world economy.

Local fluctuations can act as nucleation events for explosive spreading. Identifying which markets are likely to remain open for business is important for designing trading / business / health strategies.

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 are harder to implement and interpret for those without expertise.
- 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 time series (cases, deaths, etc.).
2. OxCGRT : Government reponse data.
3. OWID : COVID-19 time series in addition to other national averages/categorical variables.
4. FIND: COVID-19 tests as time series.

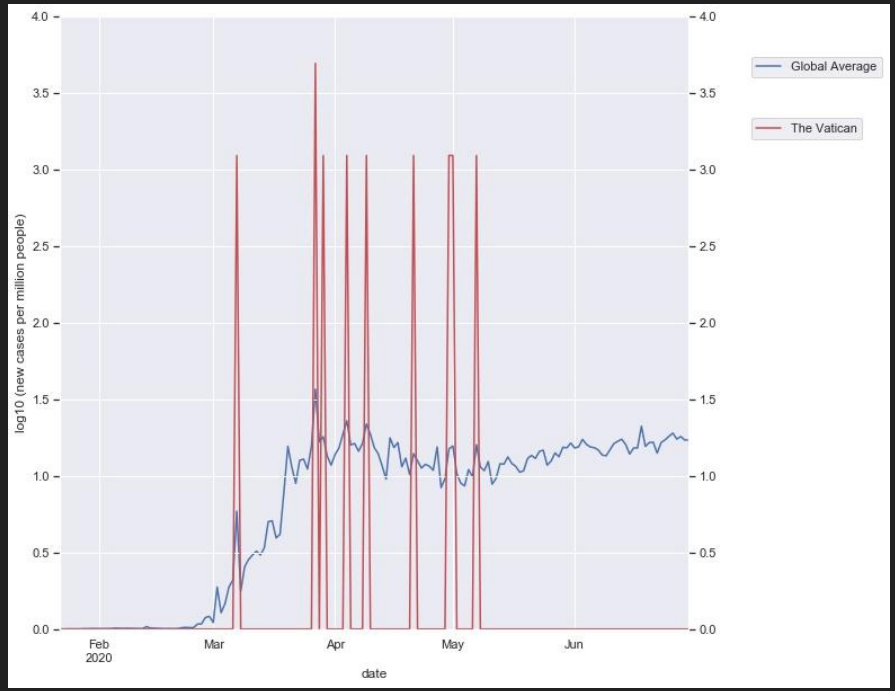
Data cleaning and wrangling

1. Regularize the names, dates, countries, data formats.

2. Account for missing and “pathological” values.

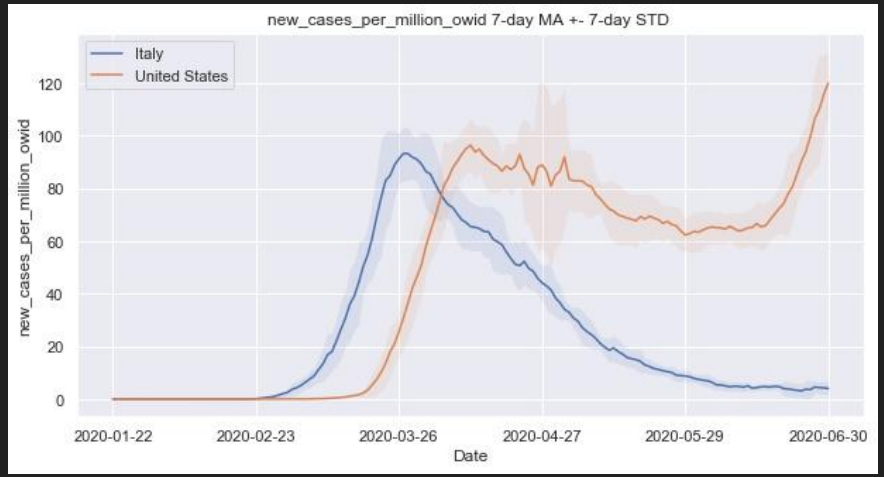
3. Determine which variables to use for modeling purposes via qualitative feature selection.

4. Engineer new features to attempt to capture time dependence explicitly.



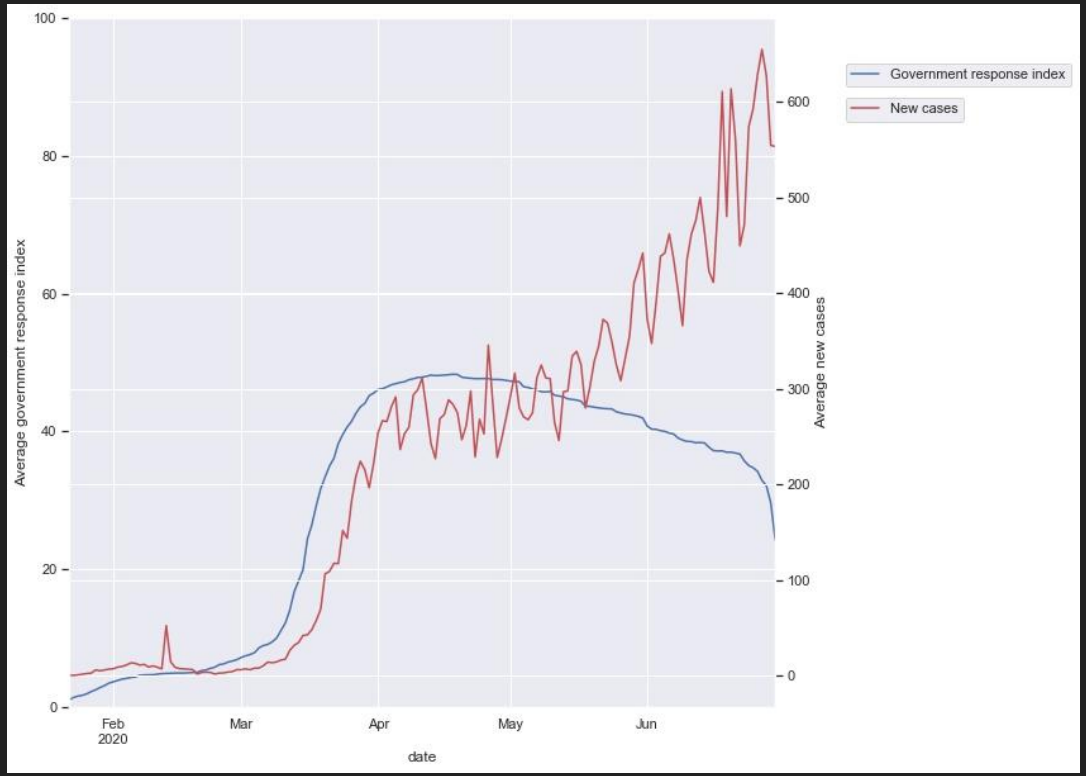
Data exploration

Investigate the effect of different government mandates, helping determine actionable steps



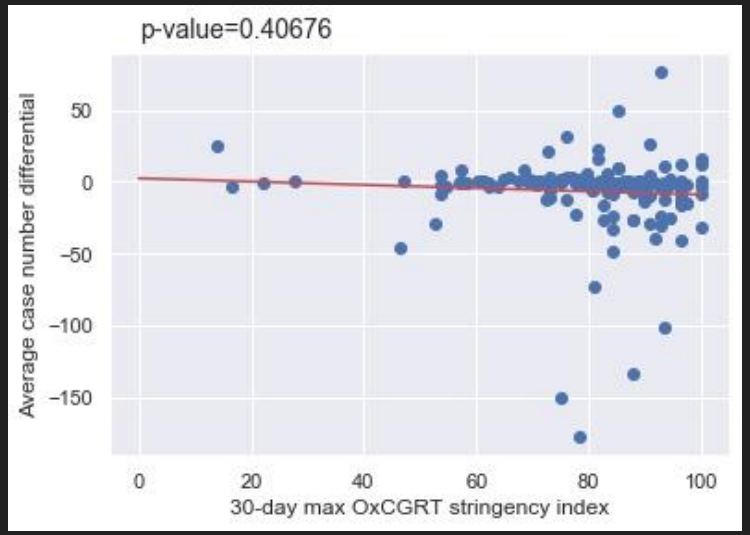
OxCGRT “Stringency index”

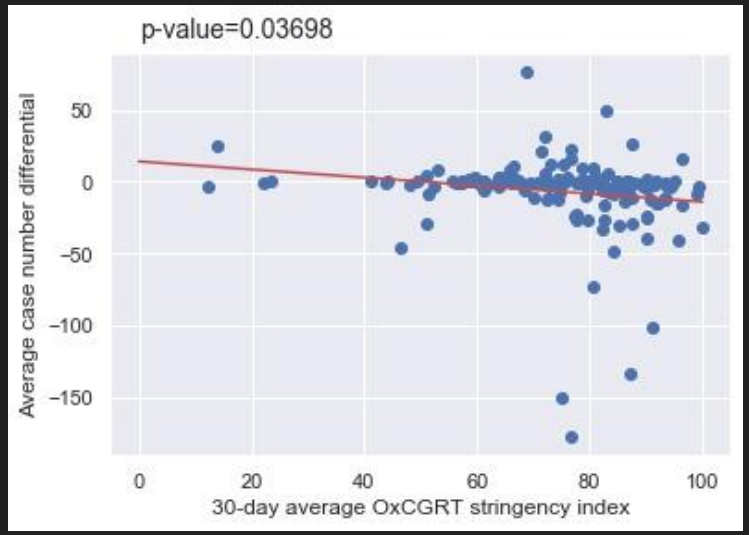
- The stringency index (defined on interval [0,100]) is a quantification/aggregation of multiple quarantine measures from the OxCGRT dataset.
- Countries have wildly different experiences. Try to account for this with a relative and not absolute valued target variable.



Statistical analysis

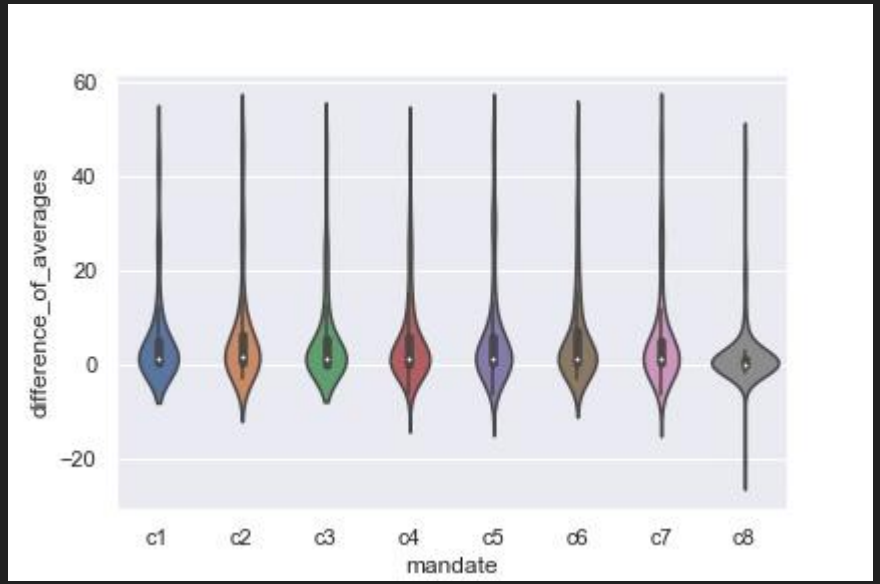
Look at univariate regressions with respect to max and mean strictness





Stringency and its components

- Can we determine which quarantine measures are effective?
- Use Kruskal-Wallis, non-parametric test on the medians between different treatments: the government mandates in this case.
- Measurements are not independent when comparing mandates from the same country, however this is just a crude approximation to compare global behavior.



P-value ~ 0, upon looking at the medians the only mandate with significant differences was the international travel controls; the magnitude of the engineered metric's median was much smaller.

Conclusion: This is likely due to time dependence and is in fact **not a significant result.**

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**)

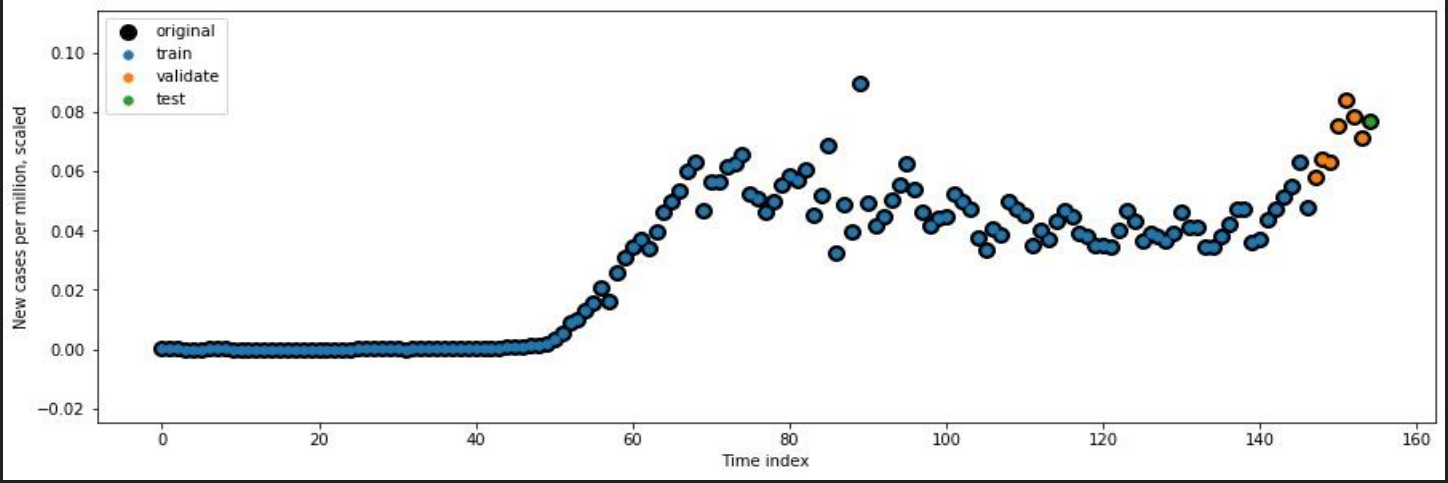
Ridge regression = 3 features from neural network, and their moving averages.

Splitting and Rescaling

MinMax rescaling, such that training values are mapped to [0, 0.5]. ‘1’ represents a fictitious maximum to account for future growth.

Each sample is a “frame” containing 28 days worth of feature data.

1. Training set contained 147 “frames”
2. Validation set contained 7 “frames”
3. Testing set contained 1 “frame”.



Modeling

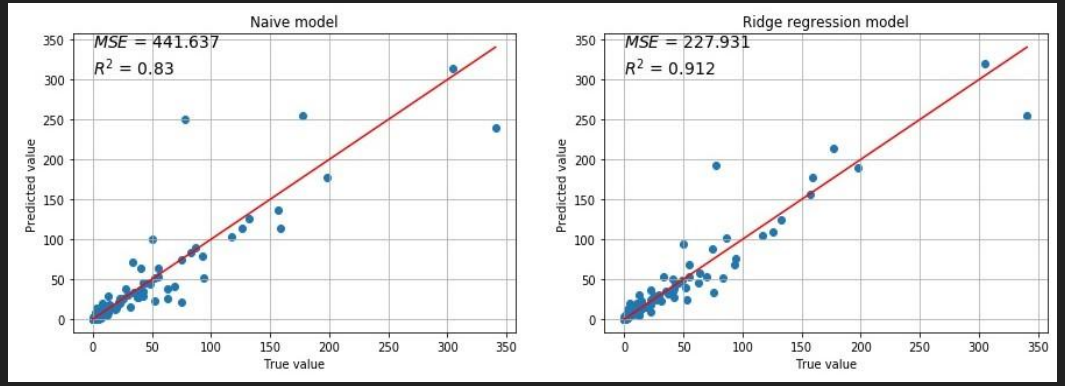
Three different models implemented:

1. Ridge regression
2. Two dense layers
3. Two convolutional layers followed by two dense layers

These are compared with a naive model. Each model uses 28 days of feature data to predict the new cases per million people on the next day.

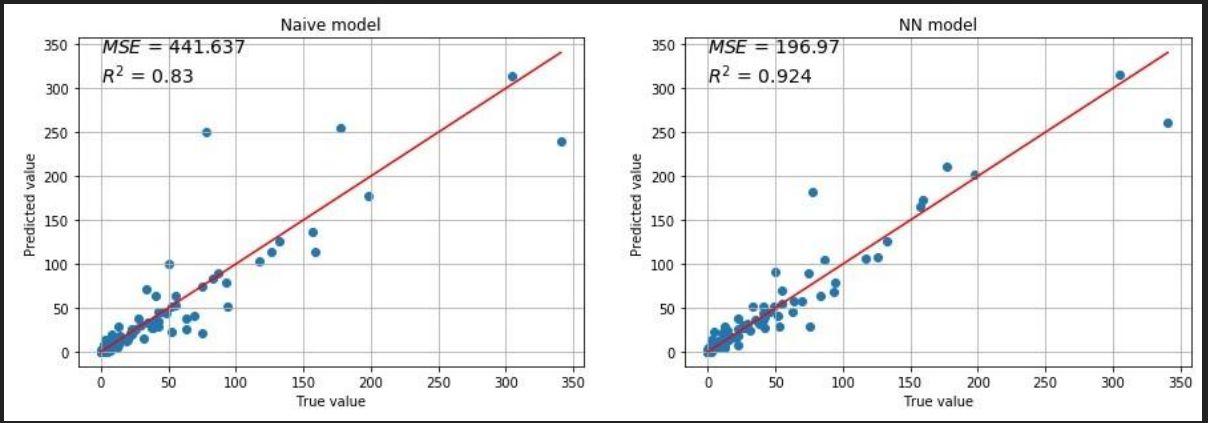
Ridge Regression

- 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.



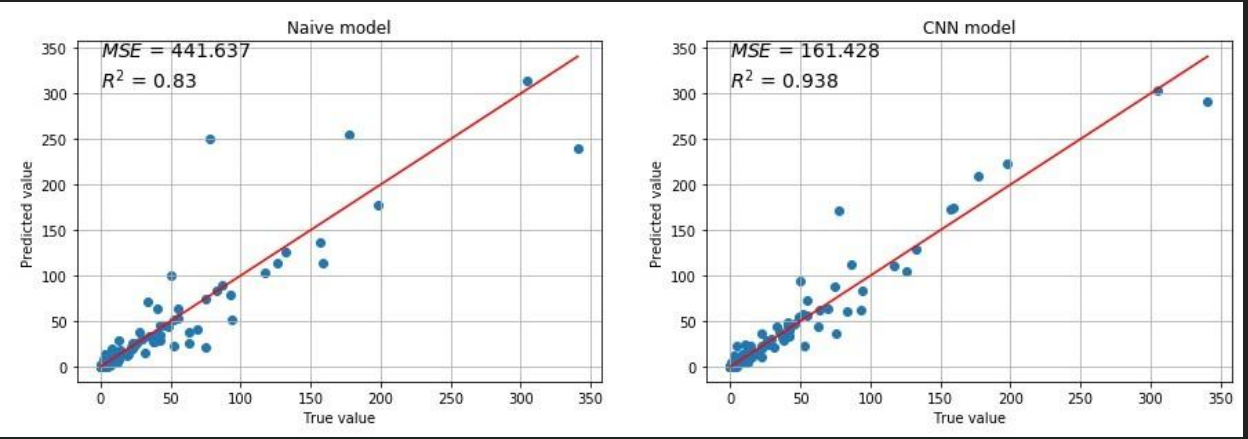
Fully connected neural network performance

- Second best performance
- Validation process chose model architecture with large number of parameters, approximately 7000, relative to the sample size.
- Was not put through rigorous parameter tuning.



Convolutional neural network performance

- Best performance as of the creation of this presentation
- On the order of 1000 parameters
- Was not put through rigorous 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. More data / investigation into quarantine measure effectiveness necessary

Future Work

1. Experiment with time frame size, feature selection, the dates included in the time series
2. Find a different loss function such as RMSLE to penalize underestimates
3. Change the architectures or the neural networks
4. Rescale the data differently
5. More parameter tuning
6. Accumulate more data, mask wearing, attitudes towards quarantines, etc.

(Extra Slide) Government mandates from Kruskal-Wallis tests

C1 = School closing

C2 = Workplace closing

C3 = Cancel public events

C4 = Restrictions on gatherings

C5 = Close public transport

C6 = Stay at home order

C7 = Internal movement restrictions

C8 = International movement restrictions