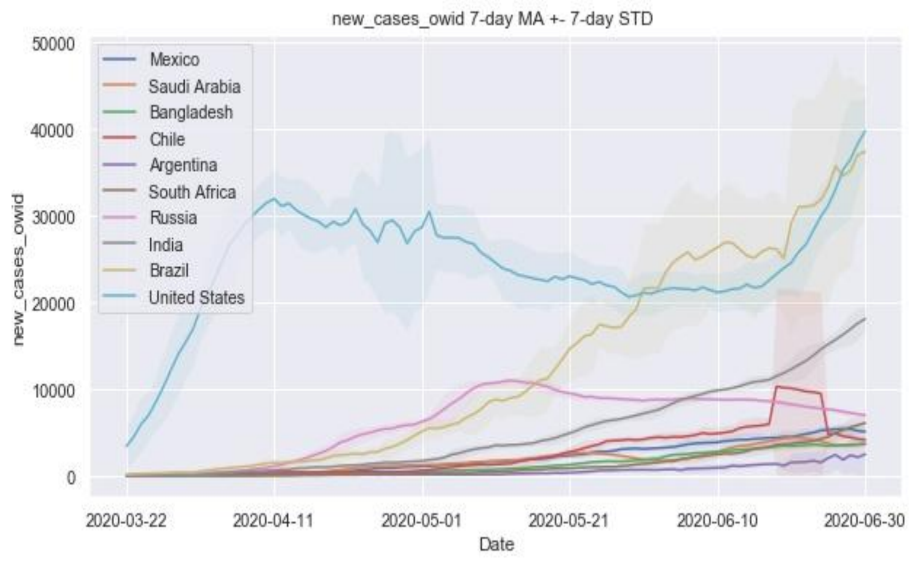
COVID-19 case number modeling

By: Matthew Gudorf

Motivations

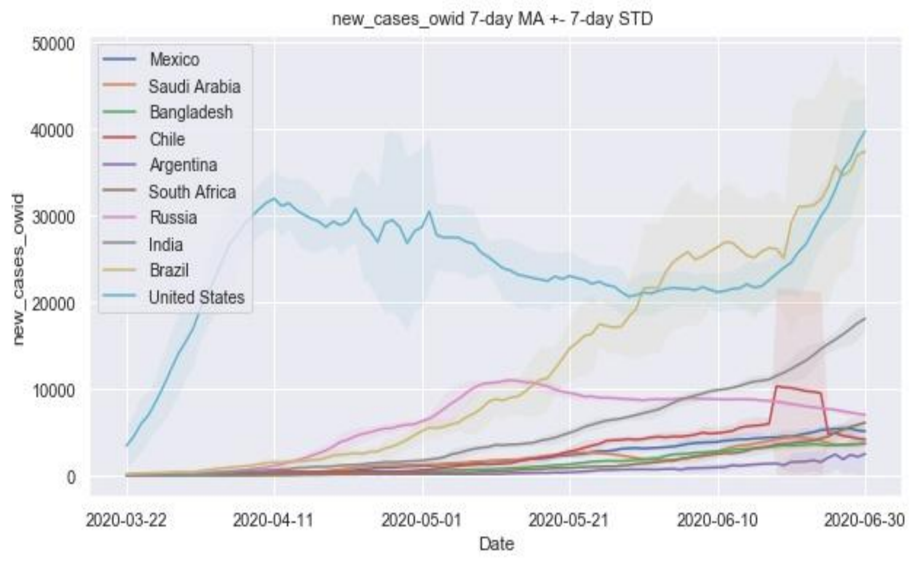
Identifying which markets are likely to remain open is important for designing trading / business / health strategies.

1. COVID-19 large effect on economy
2. Positive slope in the 7-day moving average: cases increasing.
3. Country sensitive problem.



Methodology

- Data driven approach. Investigating the use of spatiotemporal patterns via engineered features.
- Well informed epidemiological models exist, but are harder to implement and interpret for those without expertise.
- Linear regression, fully connected neural network, convolutional neural network compared to naive baseline.



Data

1. John’s Hopkins (JHU CSSE) : COVID-19 time series (cases, deaths, etc.).
2. OxCGRT : Government response data.
3. OWID : COVID-19 time series in addition to other national averages/categorical variables.
4. FIND: COVID-19 tests as time series.

What can you imagine as being possible errors in the data?

Data Quality

Why do we care?

1. Affects model accuracy

   Different units in the same variable.

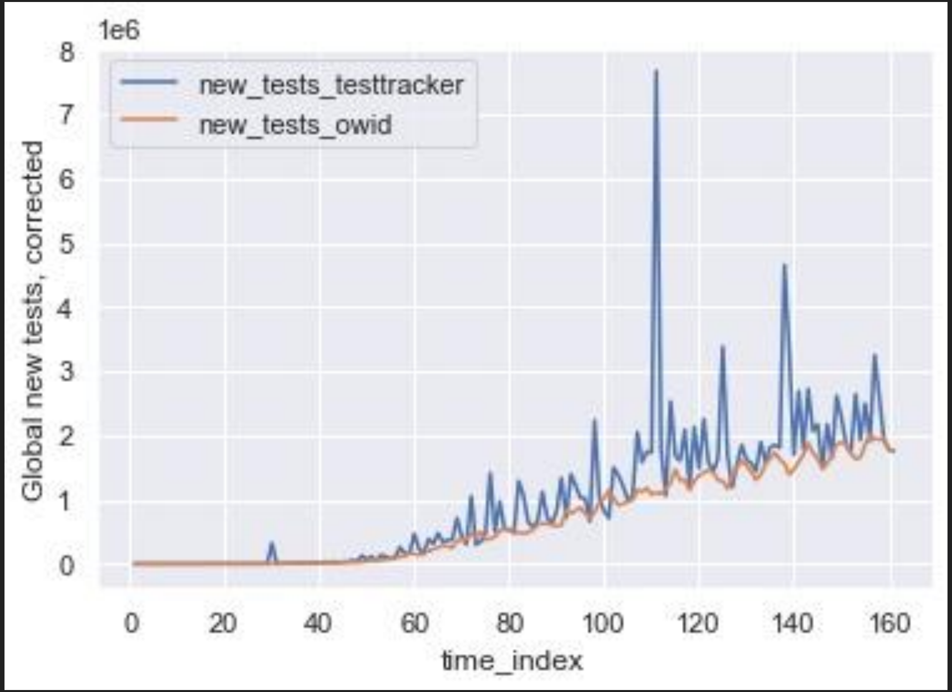
2. Model interpretability:

   How do you interpret negative new cases in a day?

3. Creating a uniform, consistent format allows for efficiency in the future.

Data quality: Part One

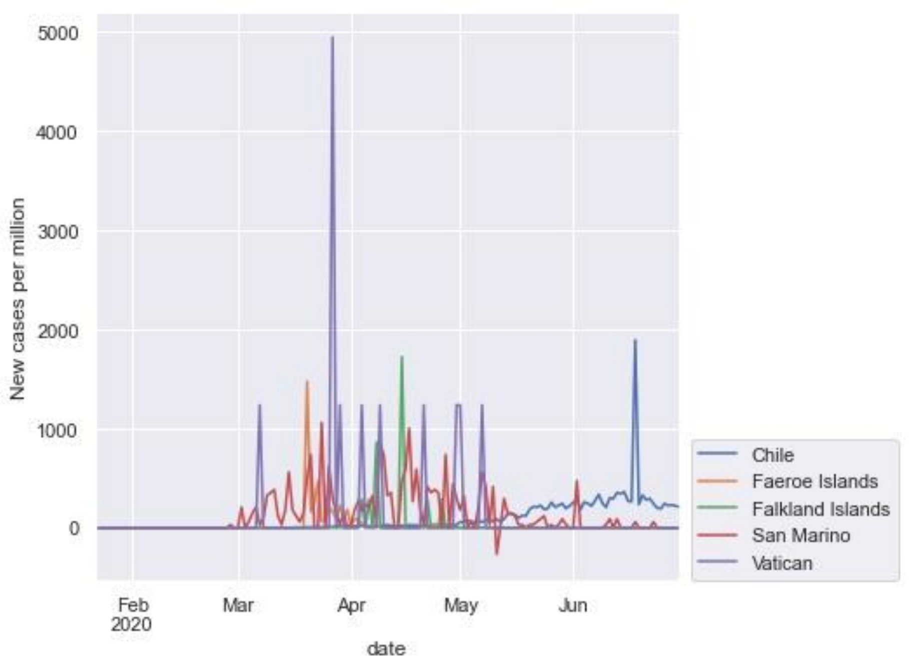
1. Units
2. Country names
3. Starting dates
4. Dataset dependent values
   who do we trust?



Data Quality: Part Two

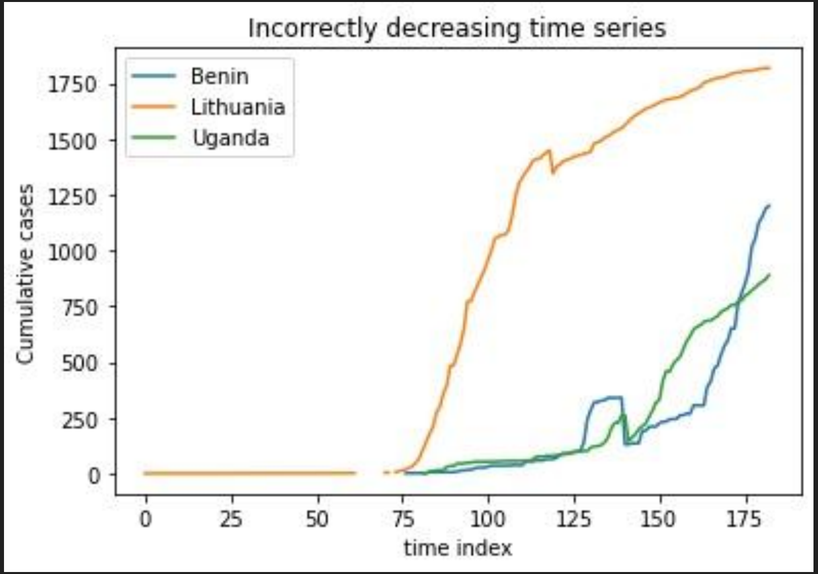
1. "Pathological" values
2. Outliers
3. Missing data

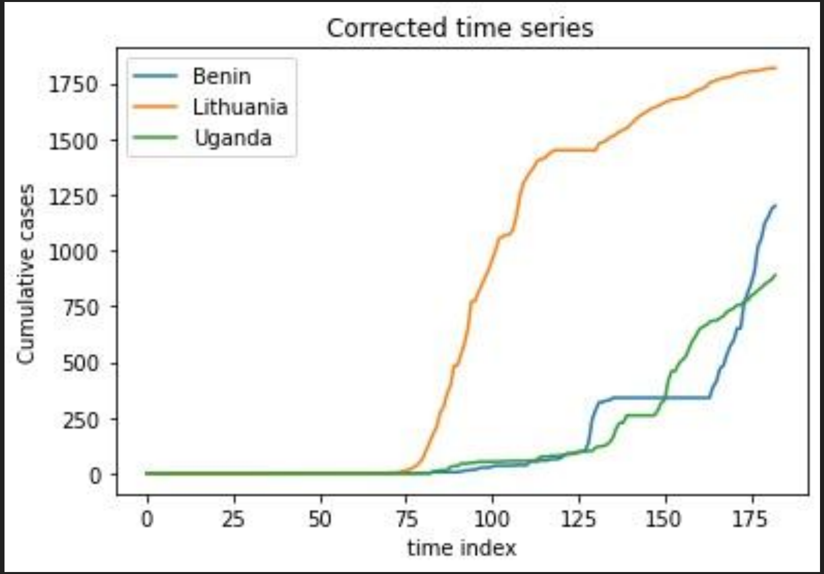
Normalizing by population weights small countries too heavily.



Data Quality: Part Three

Time series values and what to do about them. Justified for visualization, but modeling?





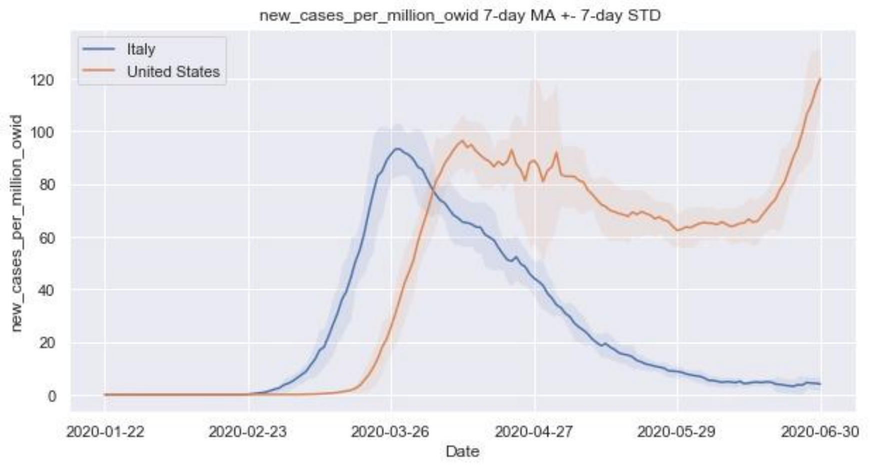
Data cleaning and wrangling

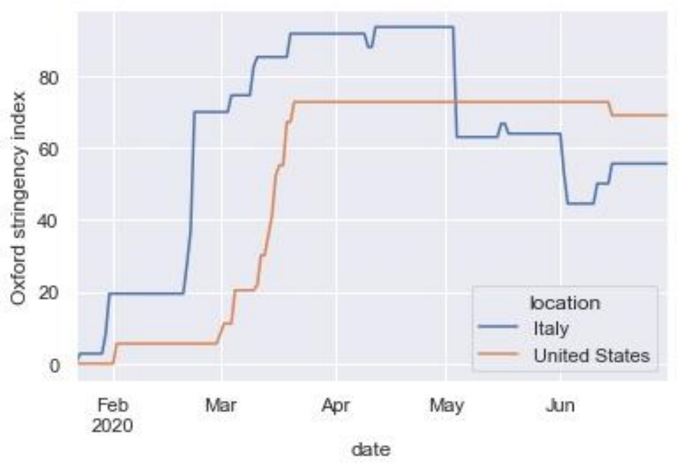
1. Regularize the names, dates, countries, data formats.

2. Account for missing and “pathological” values using imputation or custom methods.

Data exploration

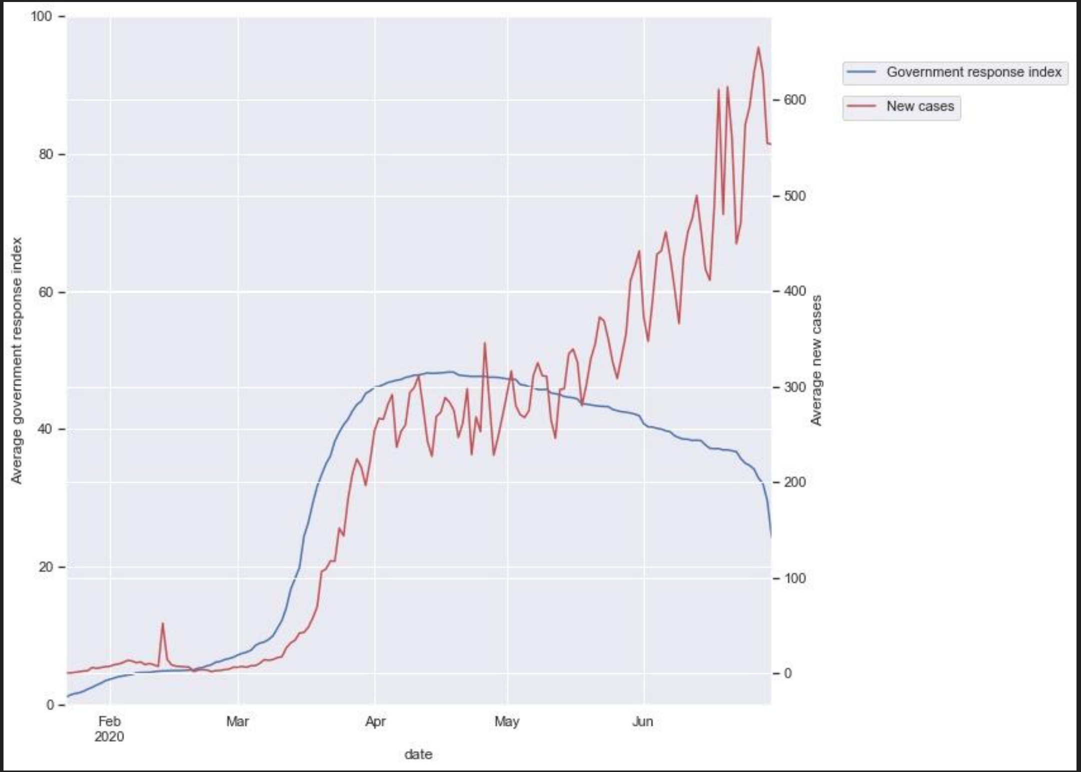
Relatively similar choices lead to drastically different scenarios; can we learn anything about this?





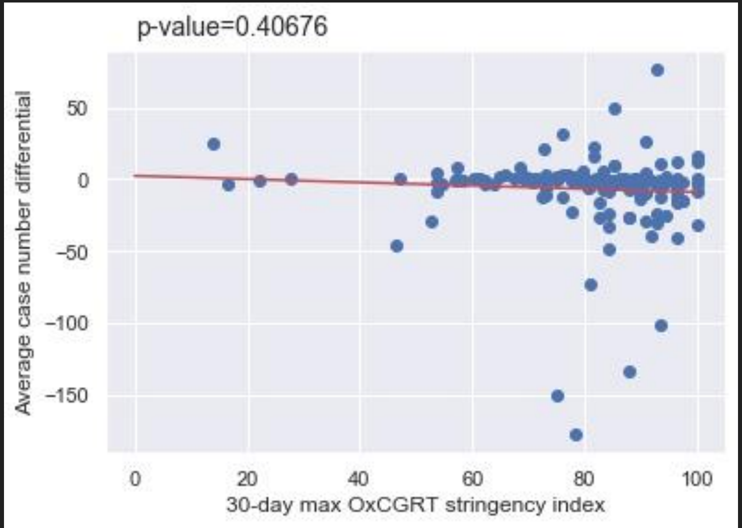
OxCGRT “Stringency index”

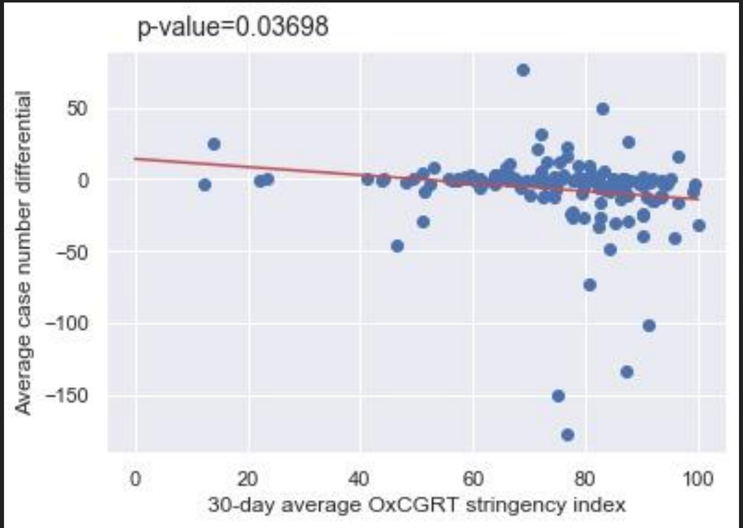
- The stringency index takes values ranging from 0 to 100. It is a quantification of various government mandates (school closings, public and international transportation, etc.)
- Aggregation hints at inverse proportionality.



Max vs Mean

Look at univariate regressions with respect to max and mean strictness.





Modeling

What:

Forecast tomorrow’s number of new covid cases.

Why:

To get a sense whether the overall situation is improving; important for allocating resources. Predicting tomorrow’s case number for simplicity.

How:

Use various models and compare against a naive baseline.

Feature Selection

Investigating the use of patterns within the time series, not all possible variables.

1. New cases per million people
2. Log(#1)
3. OxCGRT government response index.

Feature Engineering

Small dataset. Try to get more by feature engineering.

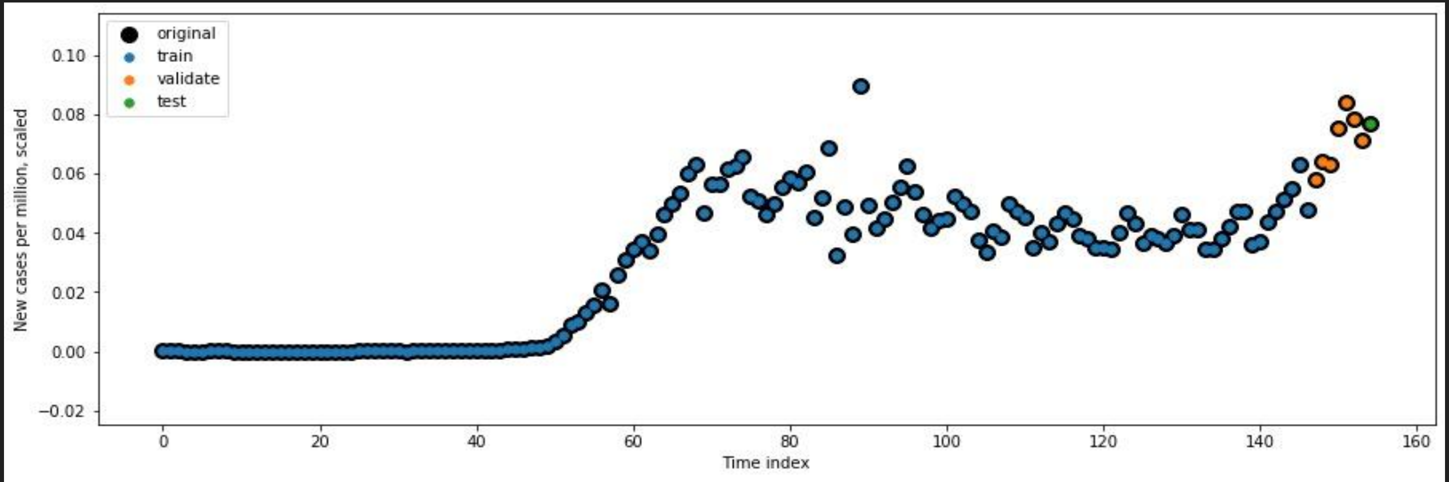
1. Use time windows to create “more” samples.
2. Normalization
3. Moving averages capture different time scales in regression.

Want our model to generalize, how do we split the data?

Splitting and Rescaling

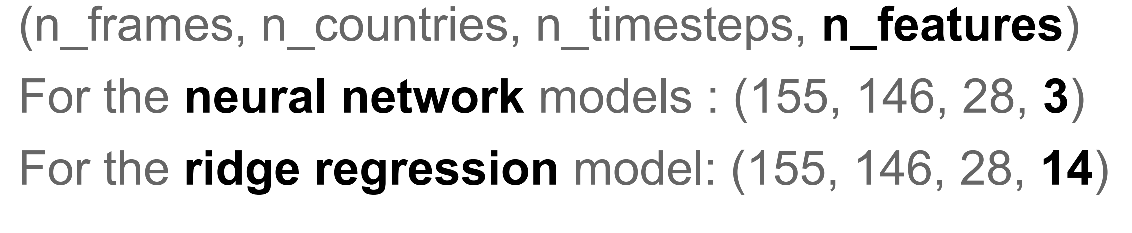
MinMax rescaling, such that training values are mapped to [0, 0.5]. ‘1’ represents a fictitious maximum to account for future growth.

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



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 + 11 moving averages.

Modeling

Model Types:

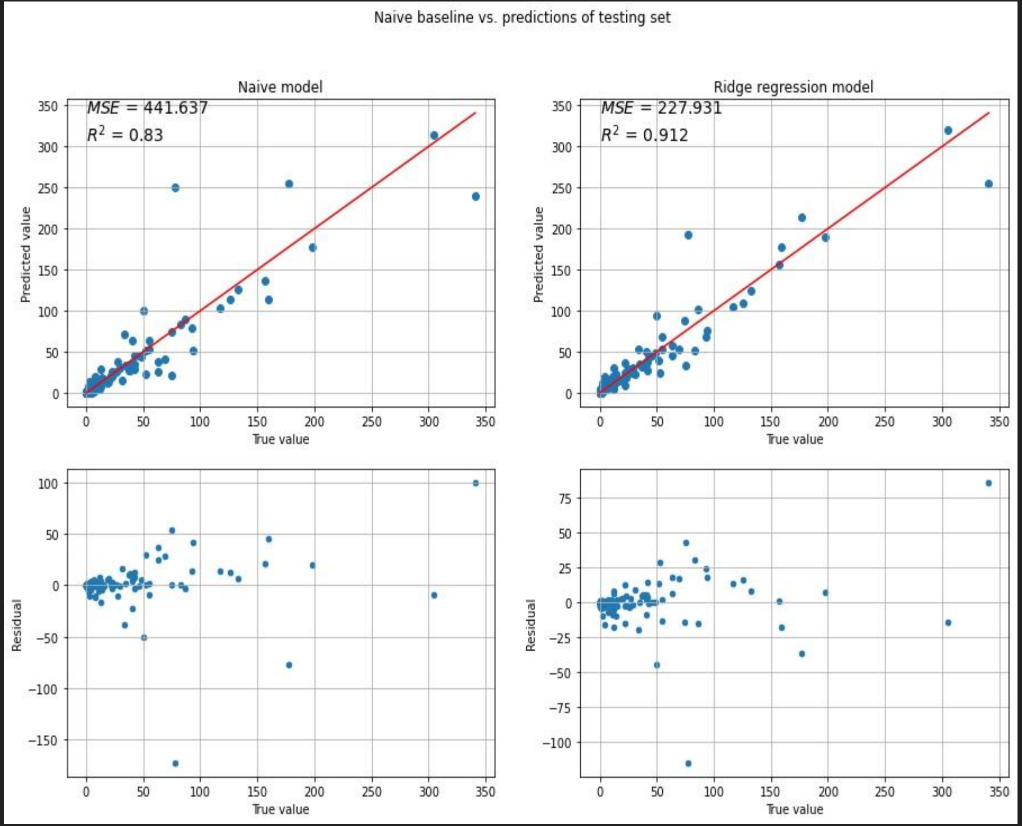
1. Ridge regression
2. Two dense layers
3. Two convolutional layers followed by two dense layers

Metrics: Used MSE for Ridge and MSLE for neural networks; MSE used for simplicity, explained variance used to compare between regression and neural models.*

*Changed NN/CNN MSLE recently, didn’t have time for Ridge due to scikit-learn telling me my target which is >= 0 has negative values. Know how to fix via variable transformation if I had time.

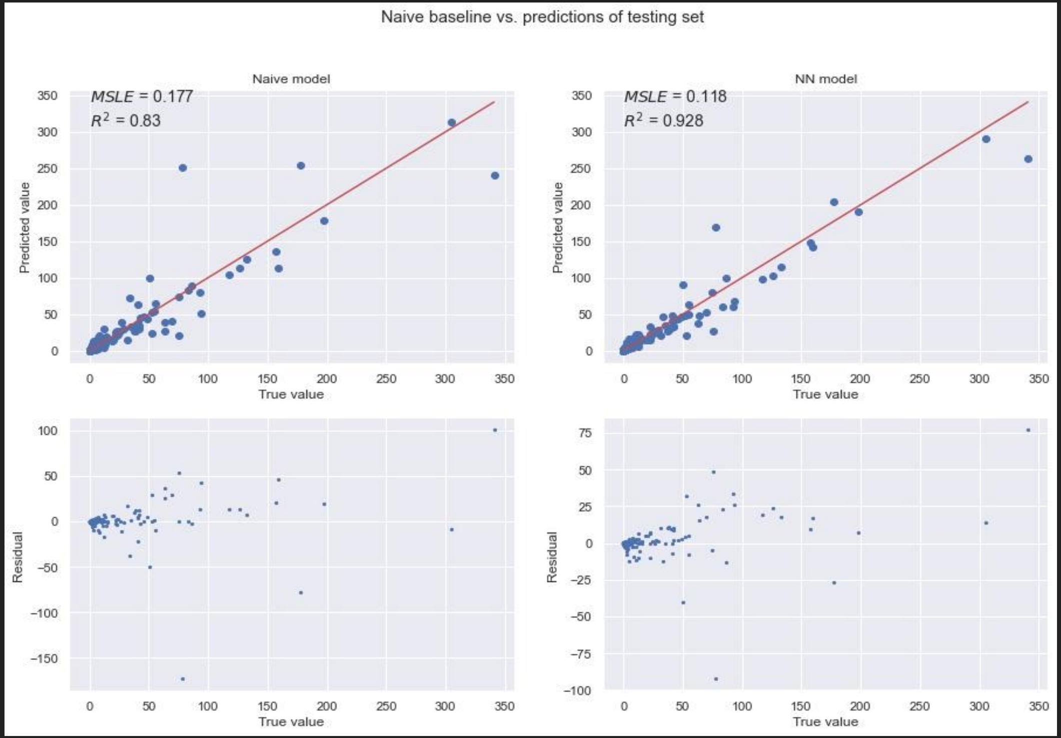
Ridge Regression

- Good performance for amount of effort required.
- Feature importance: Most significant feature was government response index.
- Cross validation resulted in a stronger than default regularization constant.
- Seemingly heteroscedasticity residuals.



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.



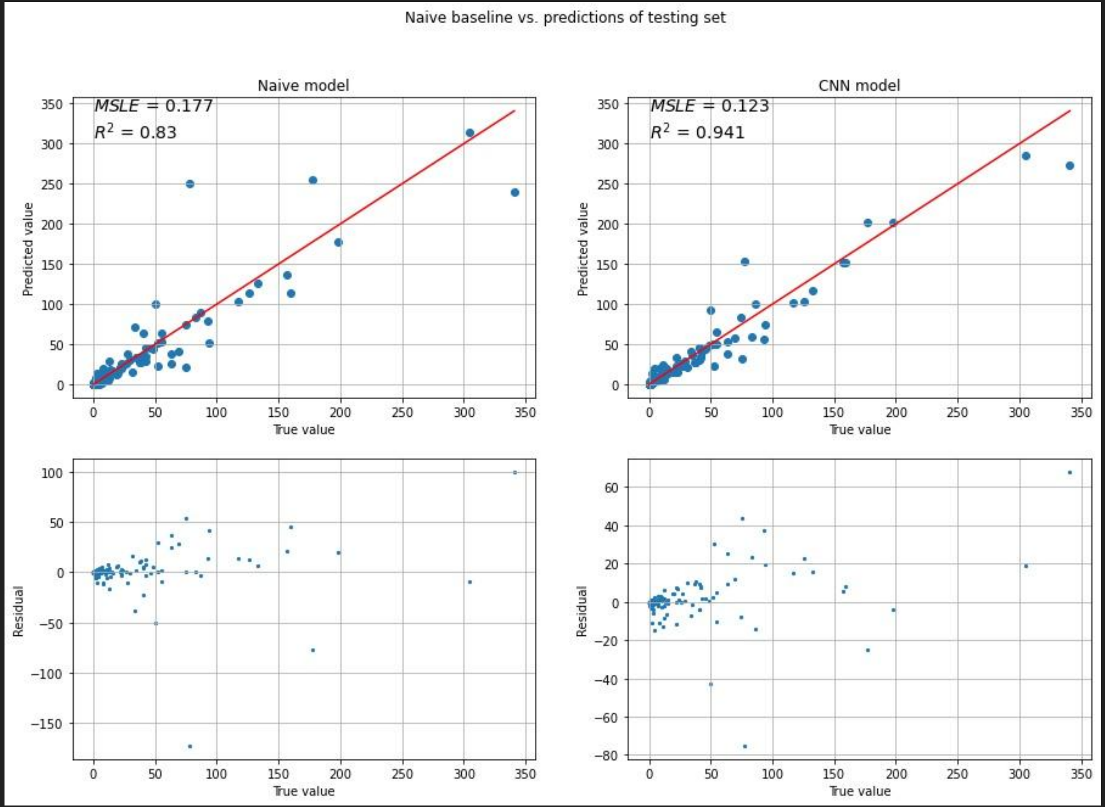
Naive baseline vs. predictions of testing set

Naive model

NN model

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.


Naive baseline vs. predictions of testing set

Naive model

CNN model

Conclusion

Small amount of data makes training cheap, but constrains the number of weights neural networks

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

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