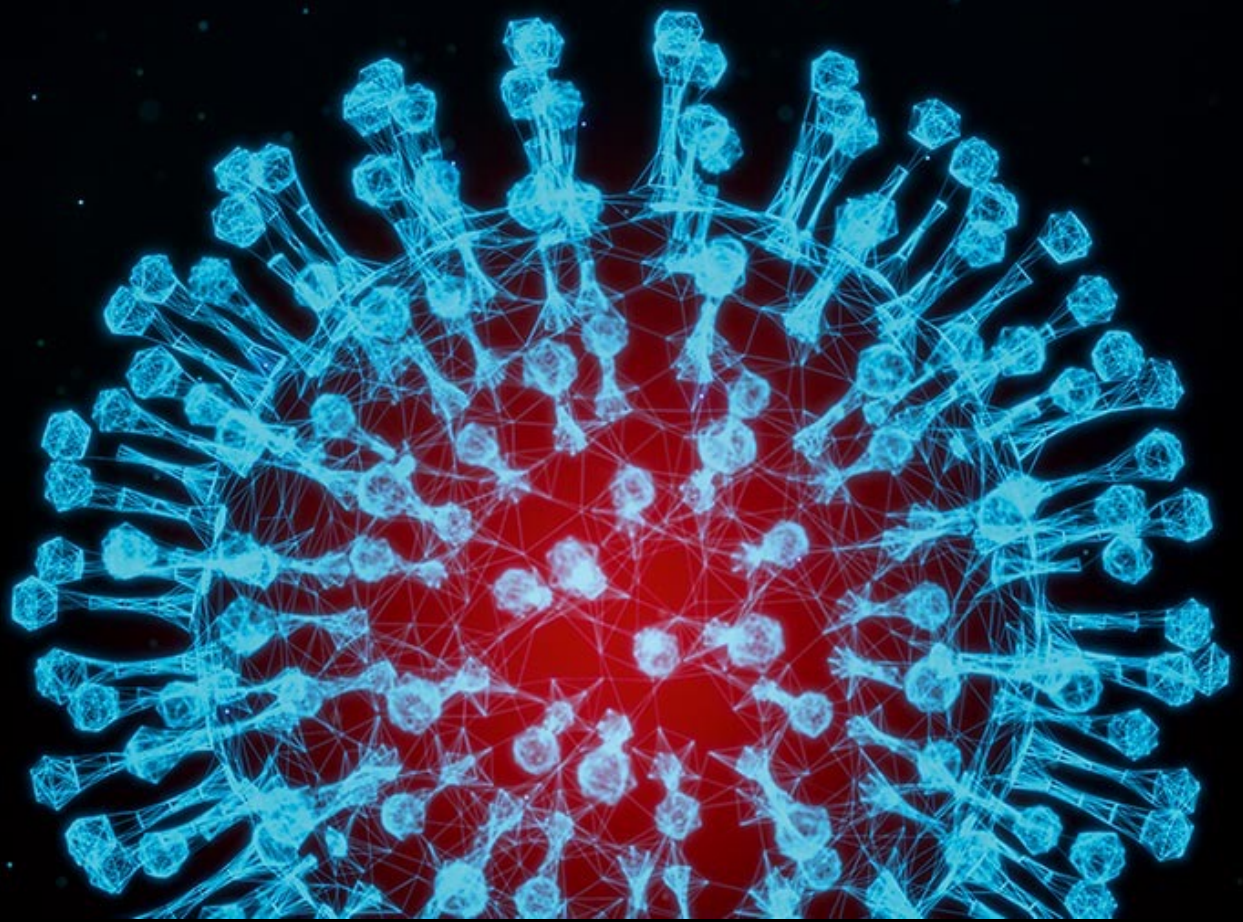
Forecasting COVID-19 cases for the next 7 days and beyond

Junchol Park

Springboard mentors:
Dipanjan Sarkar,
Ankur Verma



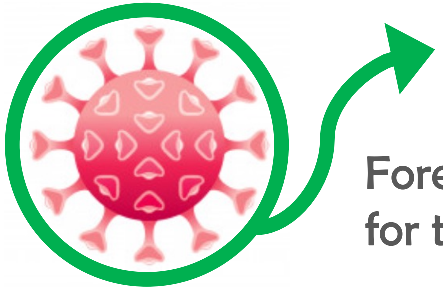
Intro.

- As of December 2021, > 261 million total COVID-19 cases, > 5 million total deaths were reported worldwide.

- Forecasting COVID-19 cases in the next several days/weeks/months is critical for coping with the pandemic.

- Many groups have been working on this problem using various methods.

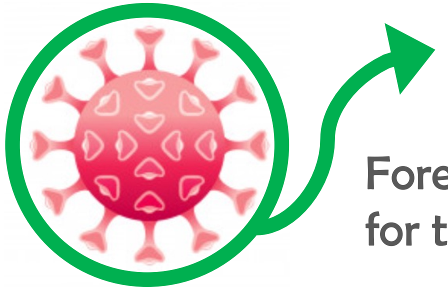
- Less effort has been made to compare performance of different model kinds, especially when they were challenged to predict cases far ahead in the future.



Forecasting COVID-19 cases
for the next 7 days and beyond

Goal

- Design and implement 4 different types of models that forecast cases in the next 7 days and further ahead into the future.
- As features, readily available data like vaccination, mobility, weather, datetime etc. should be used.
- When possible, implement the model that can generalize across diverse time-varying patterns as observed in different countries/communities.
- Compare pros/cons of the four models.

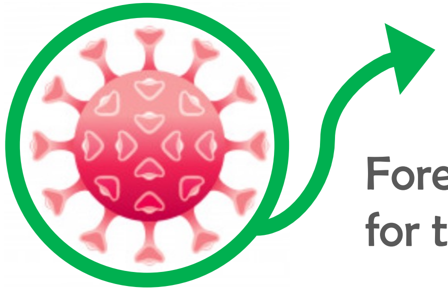


Dataset

- <u>Target</u>: COVID-19 cases in 23 countries.
- <u>Features</u>:

Static Categorical	• Country ID (country_region_code)
Temporal Categorical	• Year (year) • Month (month) • Day (day) • Week of year (week_of_year) • Day of week (day_of_week) • Holiday (holiday)

Temporal Continuous	• Google Community Mobility Data
	    • Retail & Recreation (rtrc)
	    • Grocery & Pharmacy (grph)
	    • Park (prks)
	    • Transit stations (tran)
	    • Work (work)
	    • Residential (resi)
	• Weather Data
	    • Temperature (tempC)
	    • Humidity (humidity)
	    • Cloud cover (cloudcover)
	    • Precipitation (precipMM)
	• Vaccination (vac: the number of vaccinated people per hundred)
	• Previous cases (cases during previous 14 days)



Forecasting COVID-19 cases
for the next 7 days and beyond

Models

I. SARIMAX.
(Seasonal Auto-regressive Integrated Moving Average with eXogenous factors)

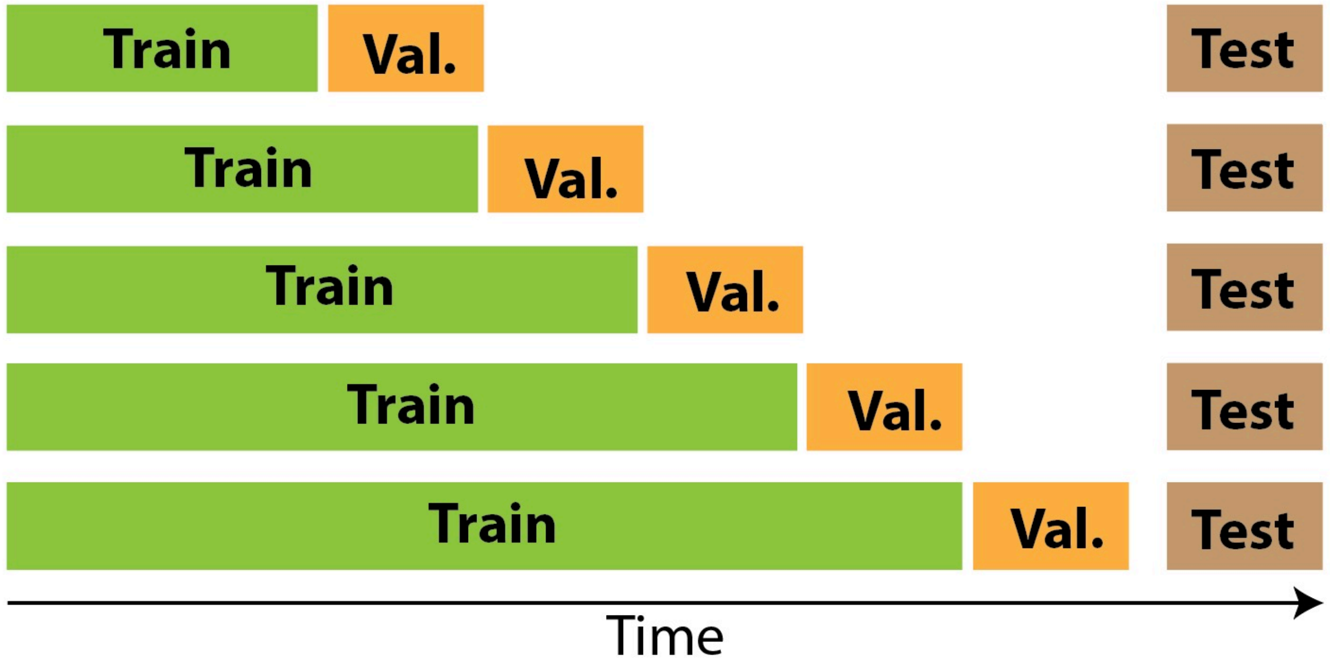
II. XGBoost regressor.

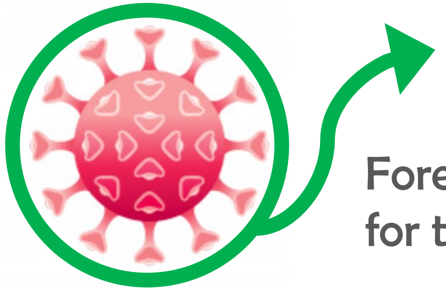
III. Multi-layer perceptron.

IV. LSTM (Long Short-Term Memory).

Forecasting COVID-19 cases
for the next 7 days and beyond

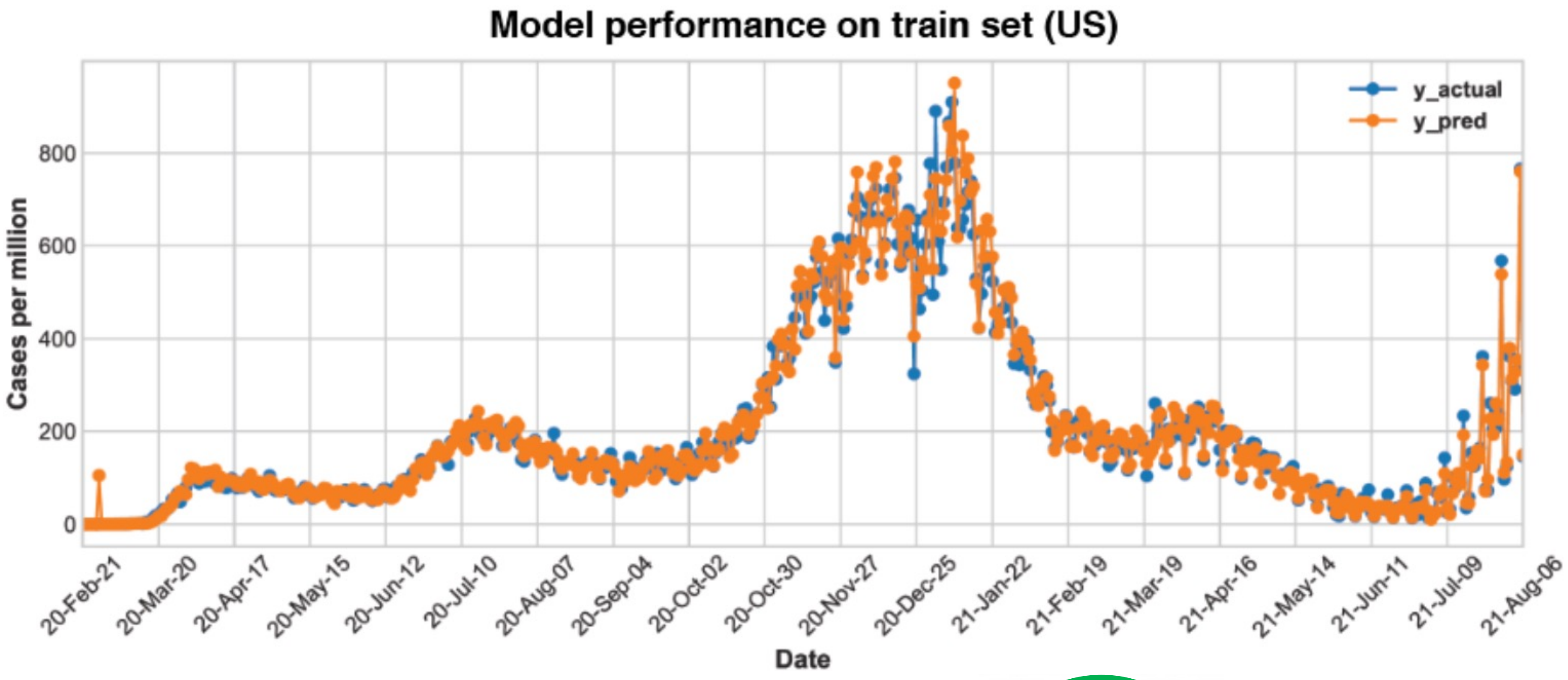
Walk forward validation

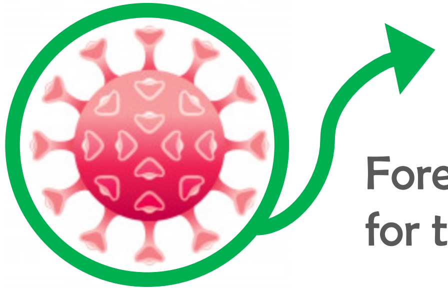




Forecasting COVID-19 cases
for the next 7 days and beyond

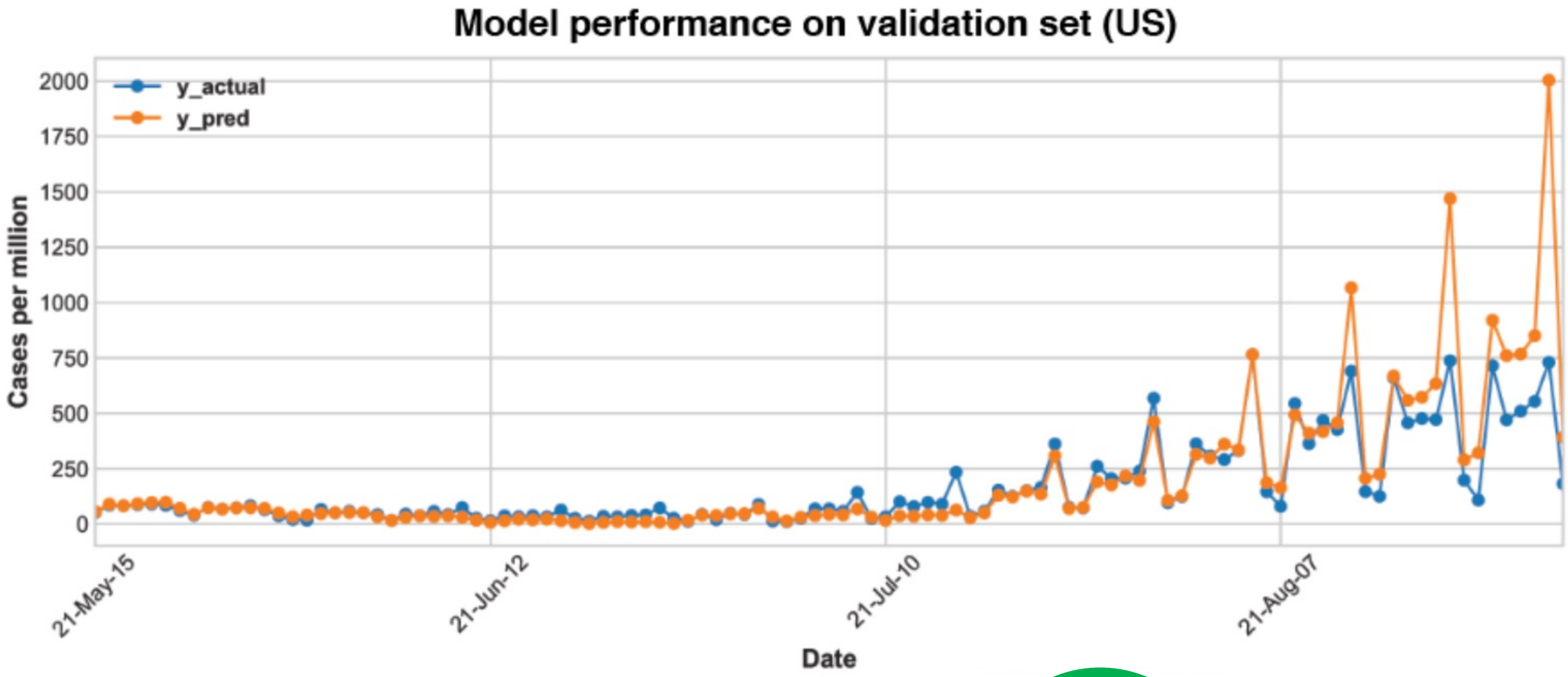
SARIMAX performance on train set

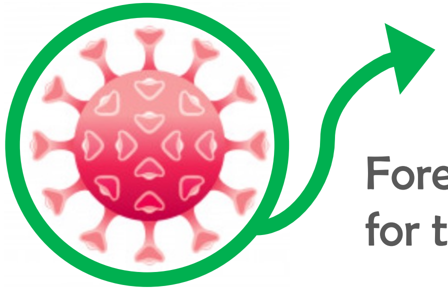




Forecasting COVID-19 cases
for the next 7 days and beyond

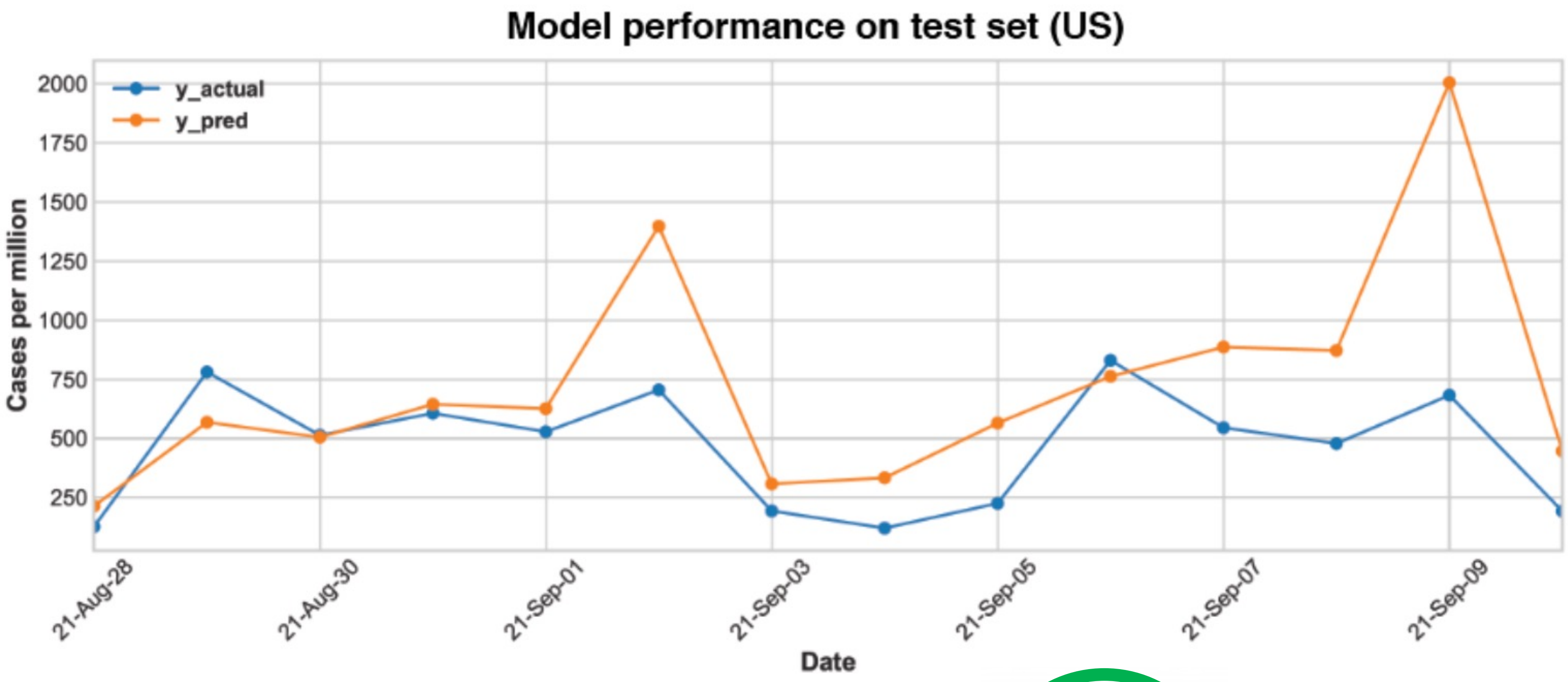
SARIMAX performance on validation set

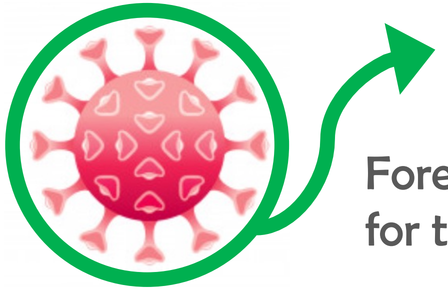




Forecasting COVID-19 cases
for the next 7 days and beyond

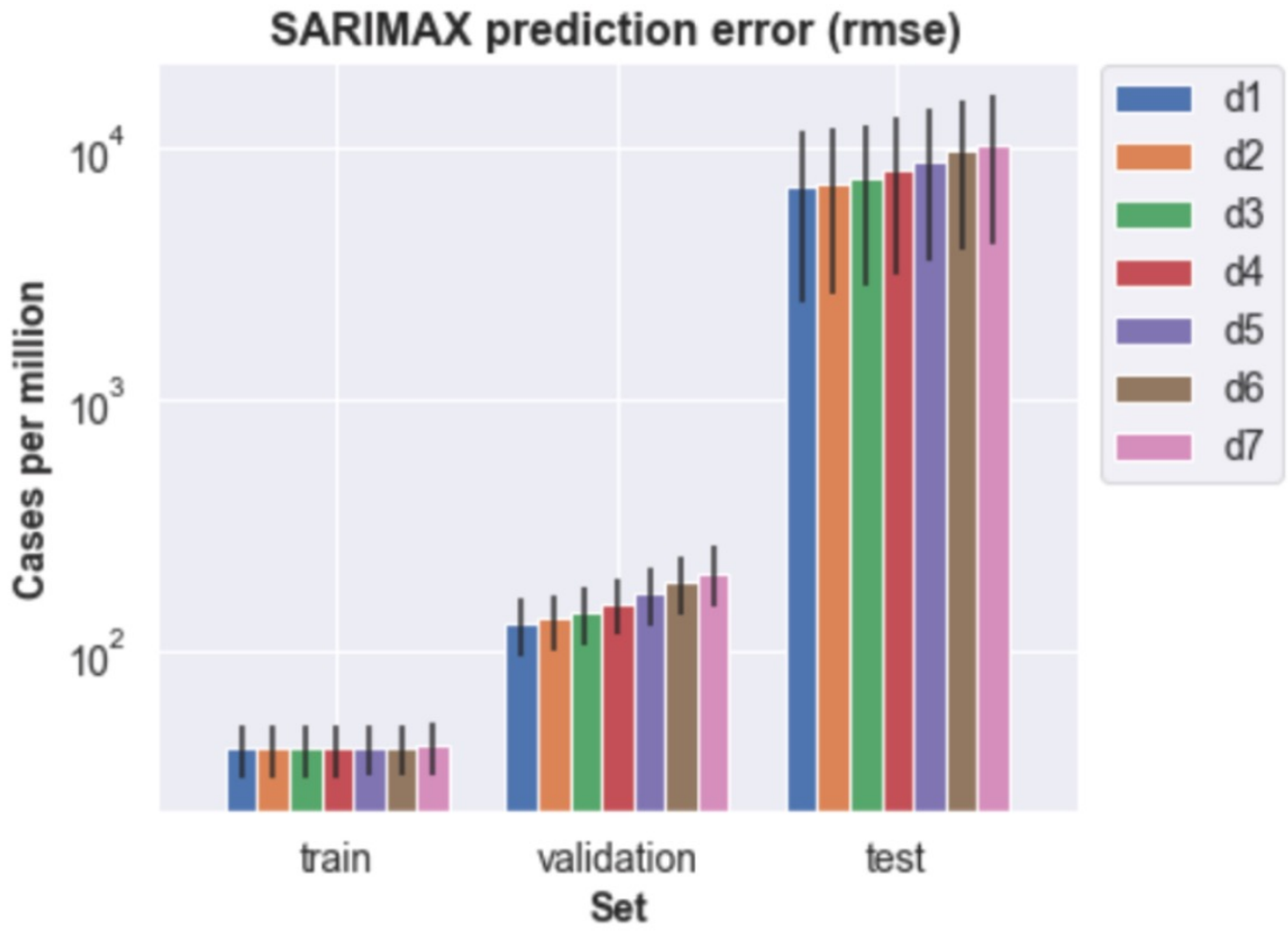
SARIMAX performance on test set



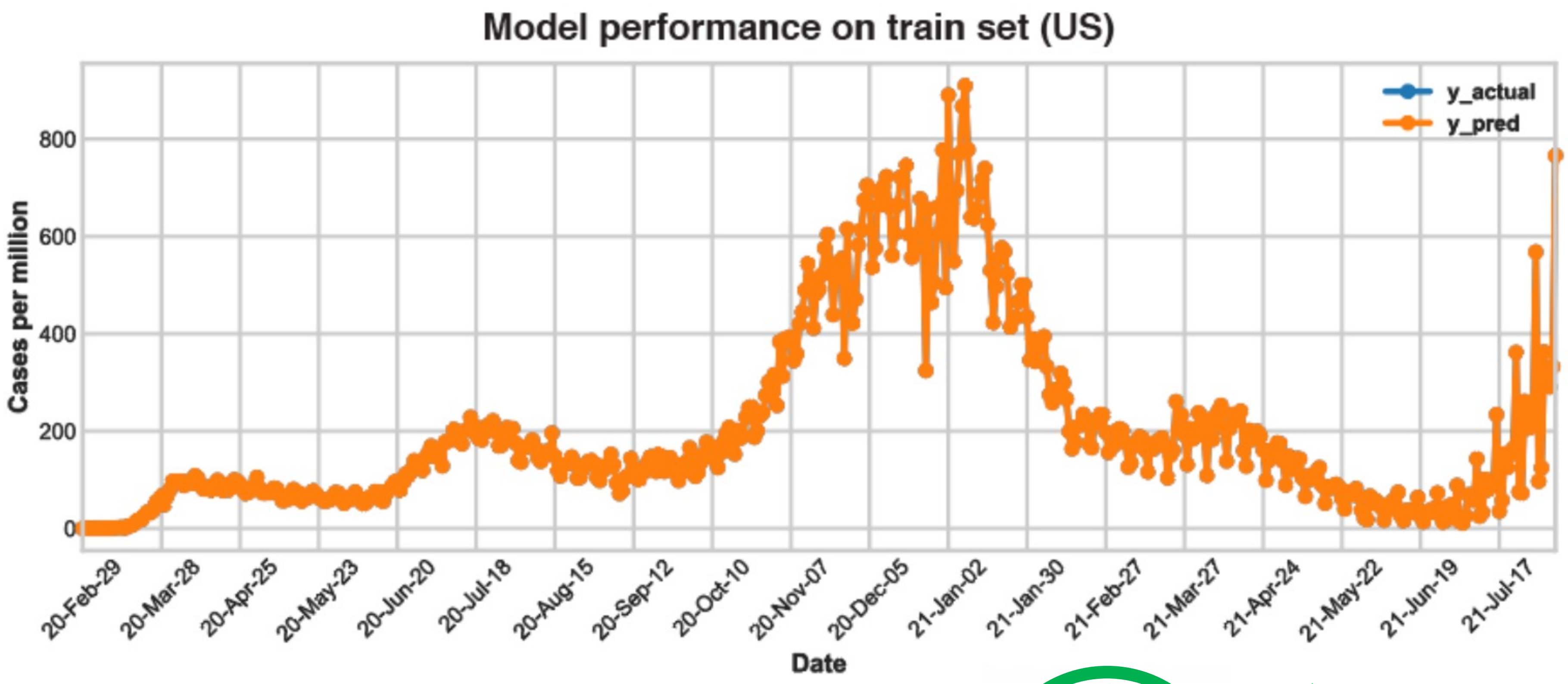


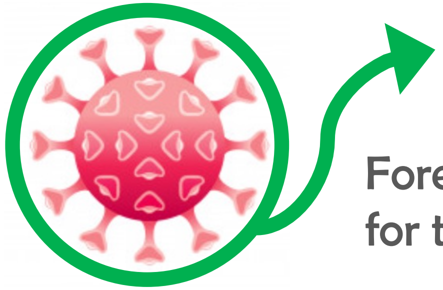
Forecasting COVID-19 cases
for the next 7 days and beyond

SARIMAX performance on train/validation/test sets



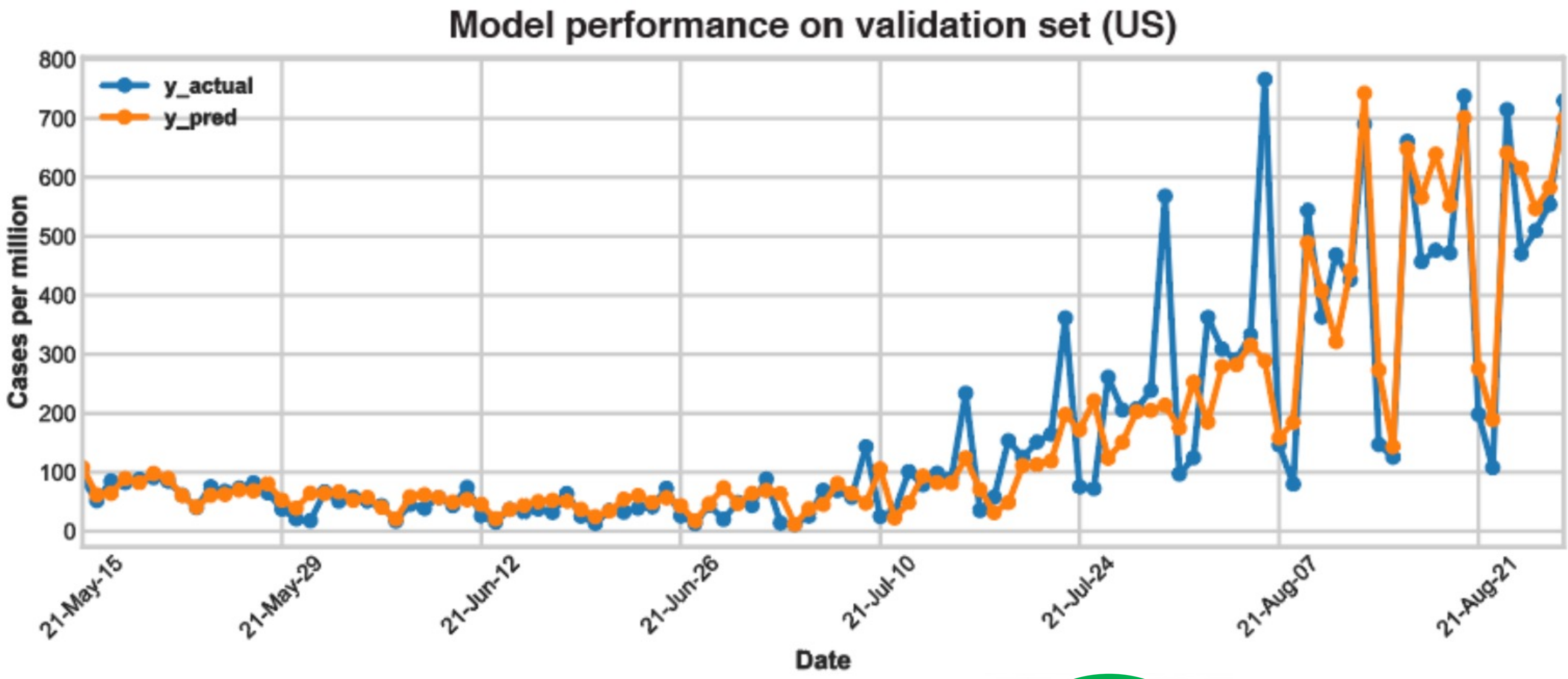
XGBoost performance on train set

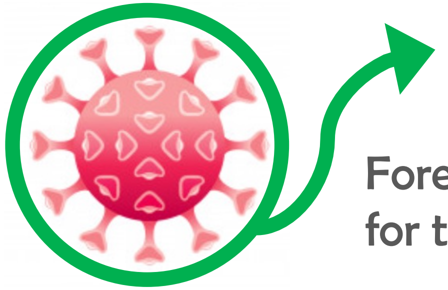




Forecasting COVID-19 cases
for the next 7 days and beyond

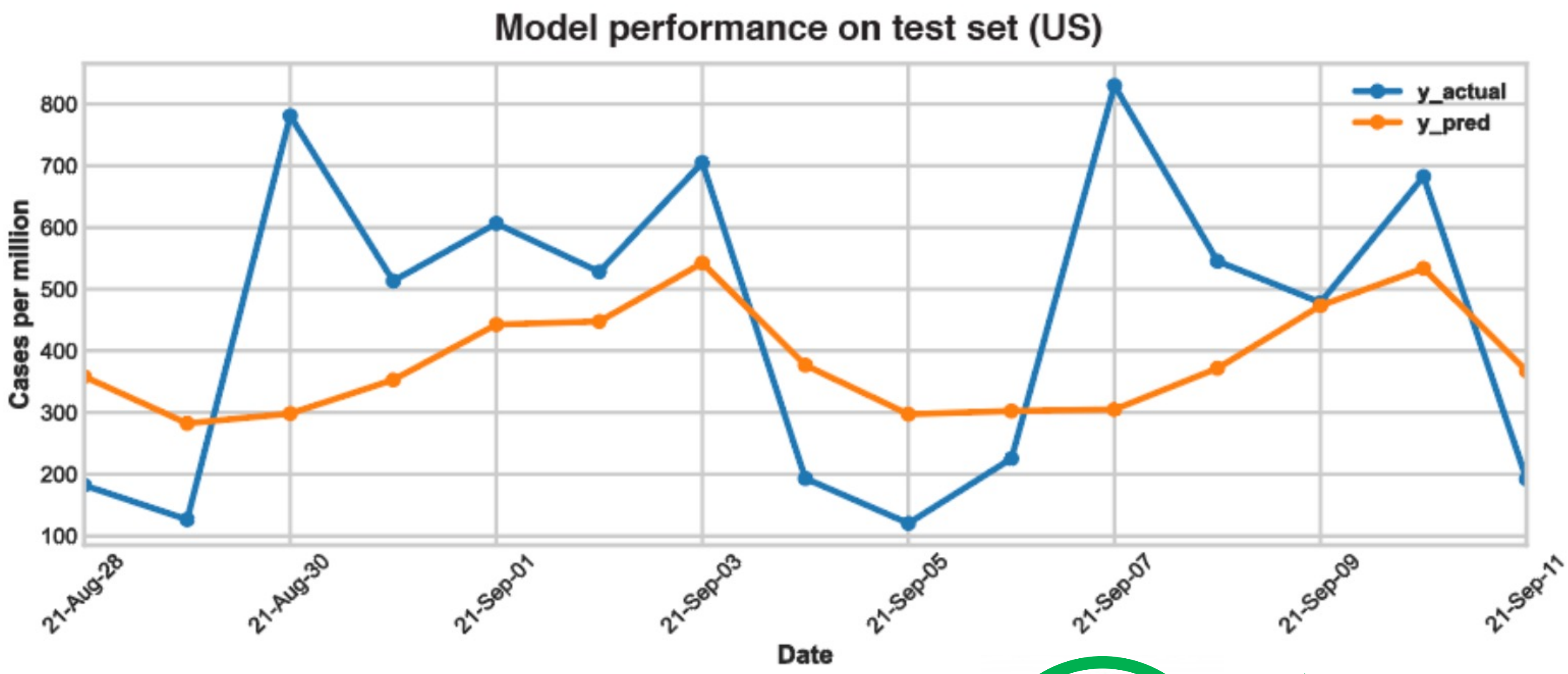
XGBoost performance on validation set

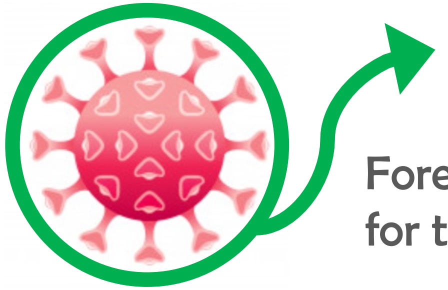




Forecasting COVID-19 cases
for the next 7 days and beyond

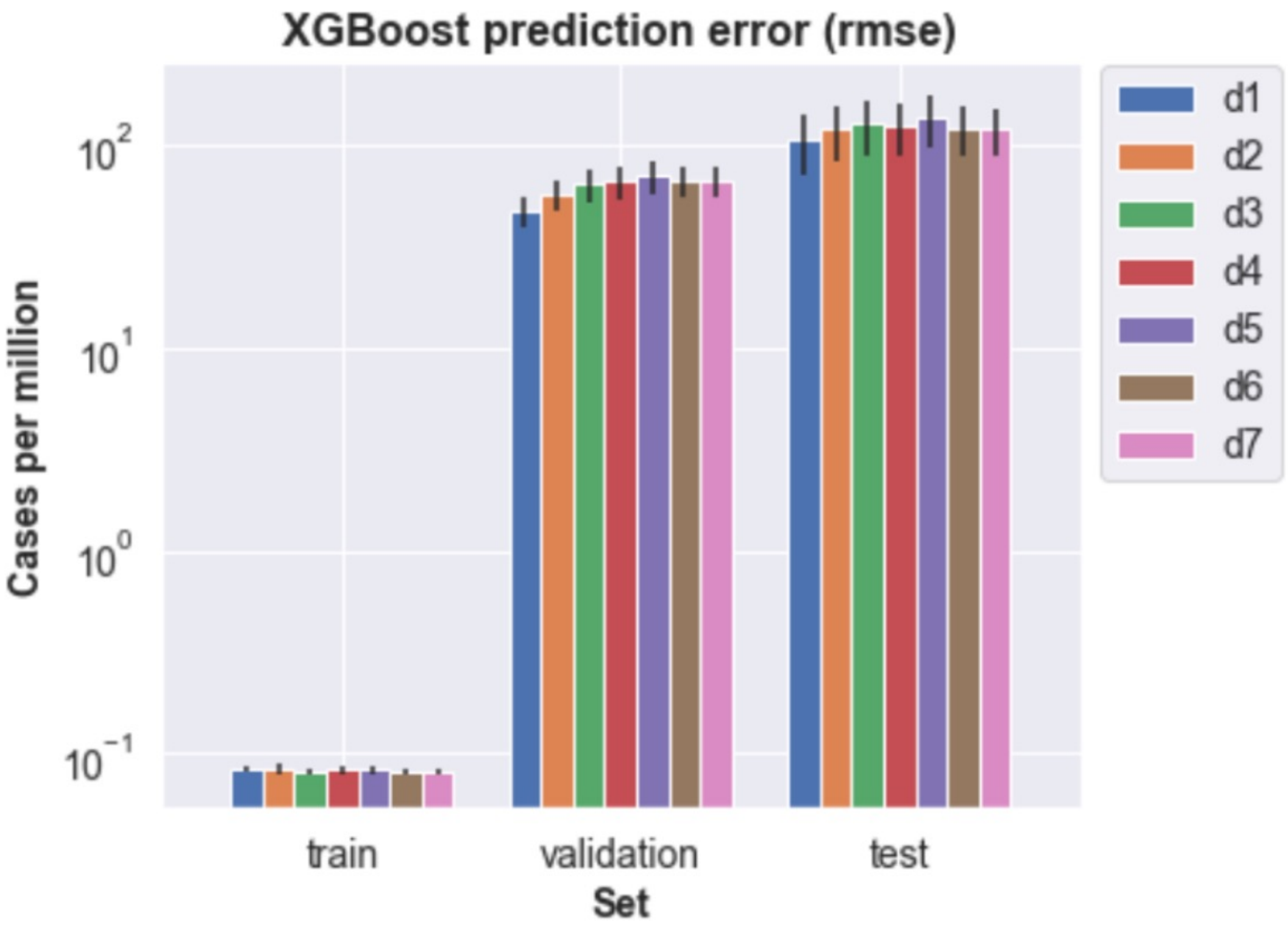
XGBoost performance on test set



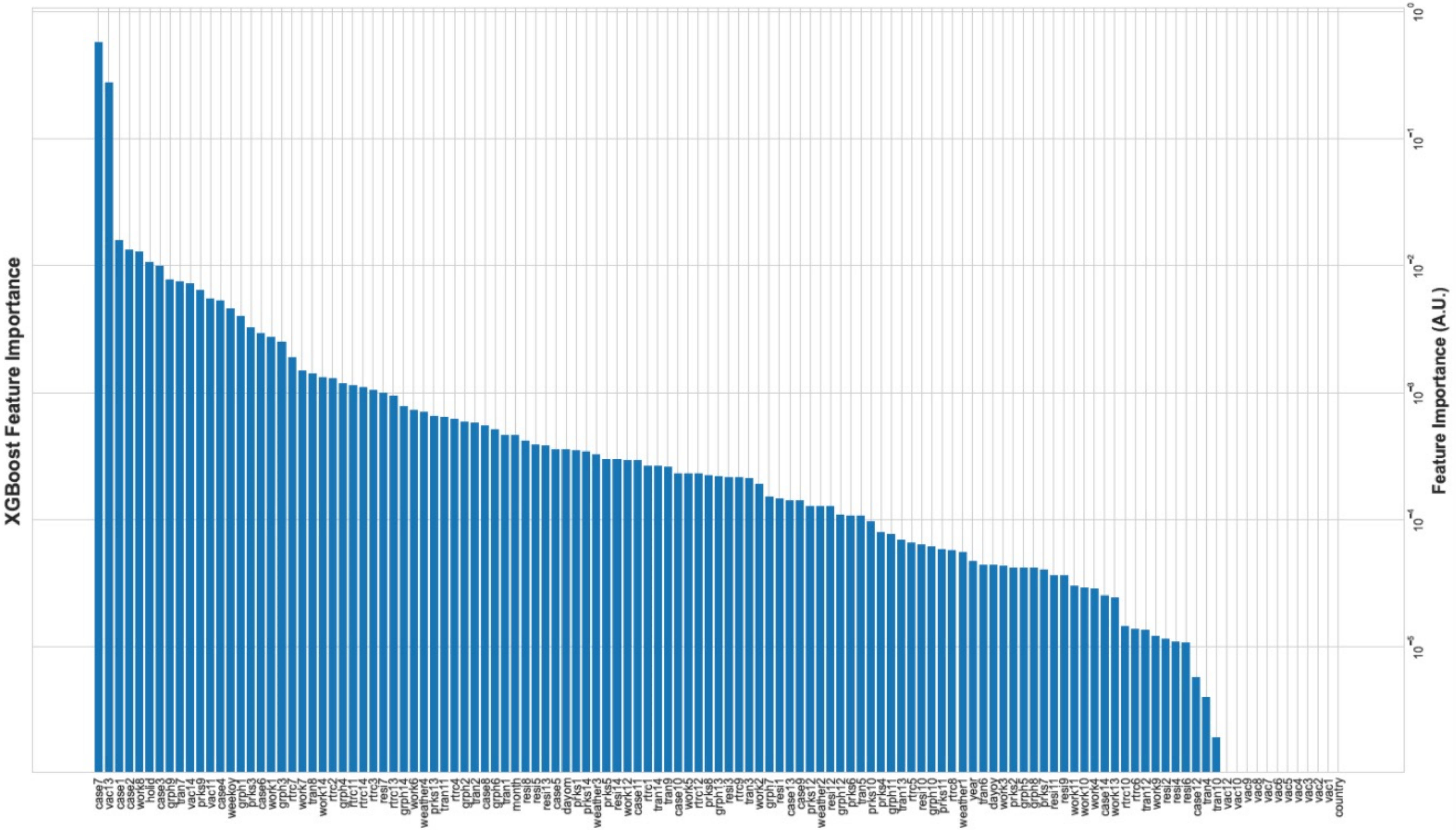


Forecasting COVID-19 cases for the next 7 days and beyond

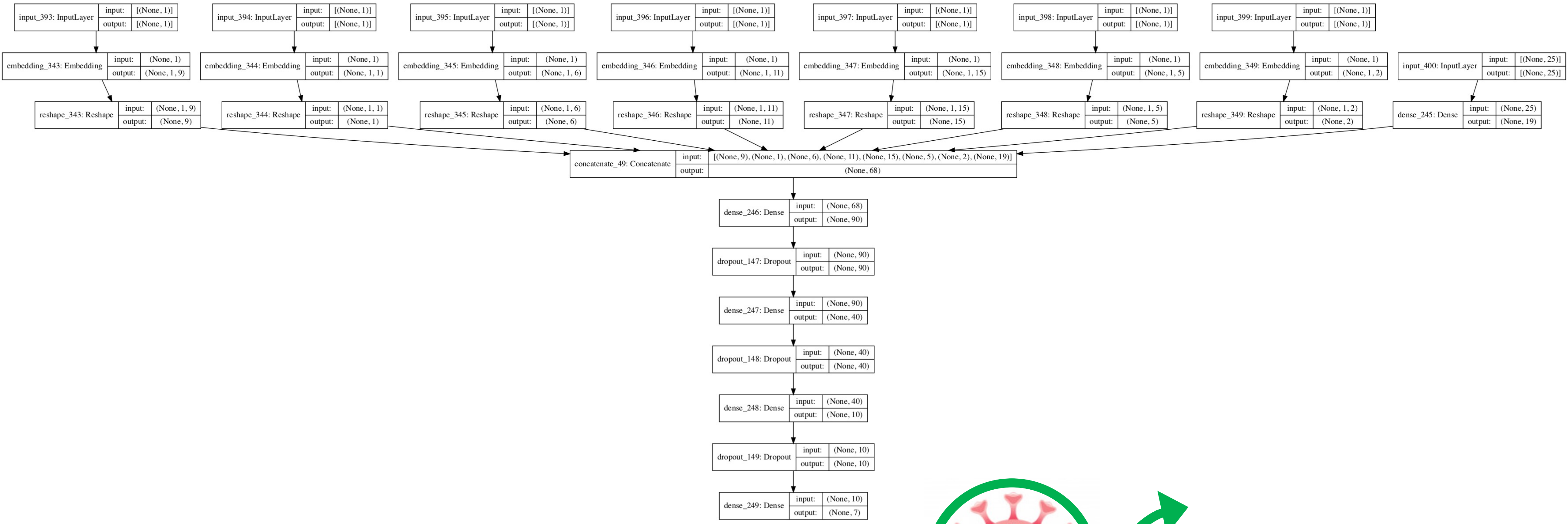
XGBoost performance on train/validation/test sets

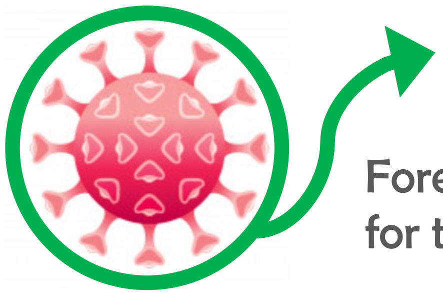


XGBoost model feature importance



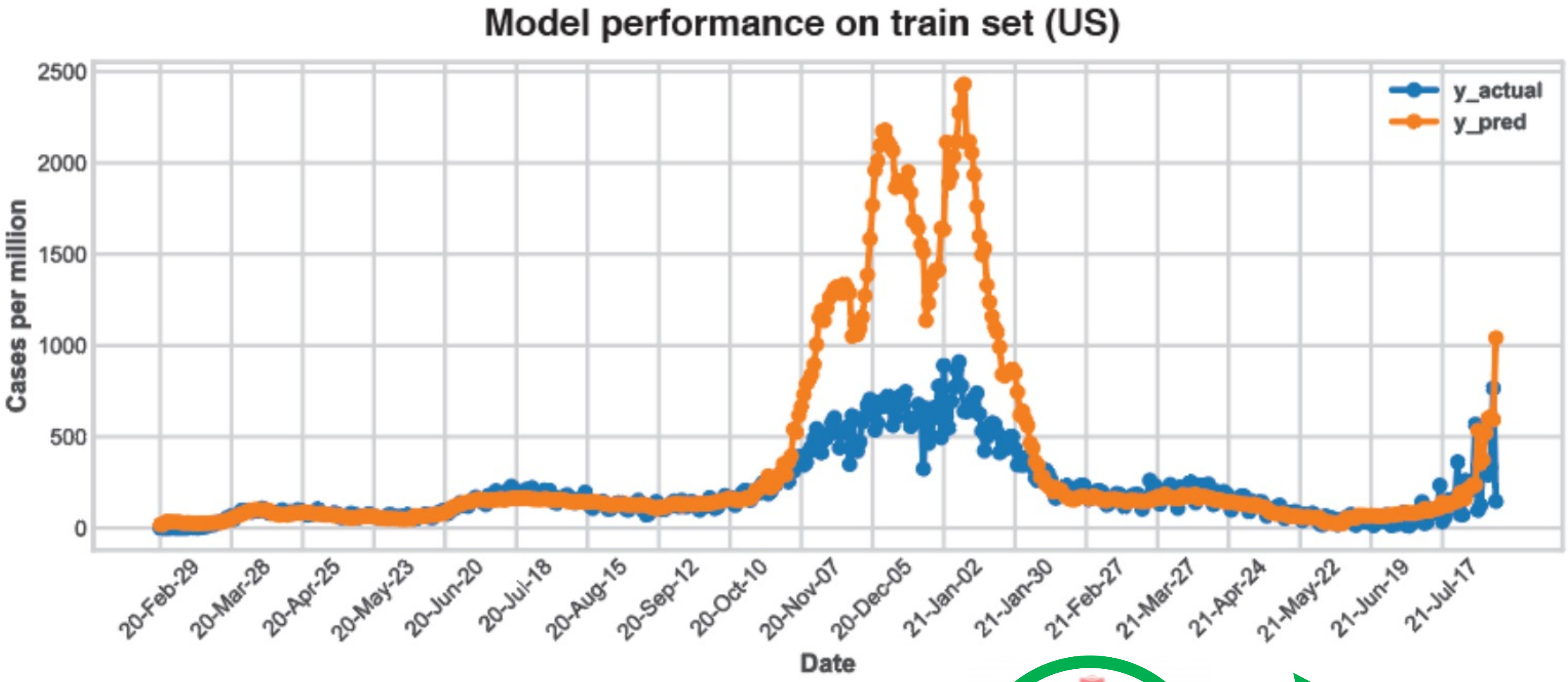
MLP model architecture with embeddings

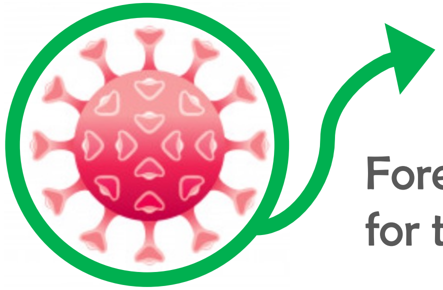




Forecasting COVID-19 cases
for the next 7 days and beyond

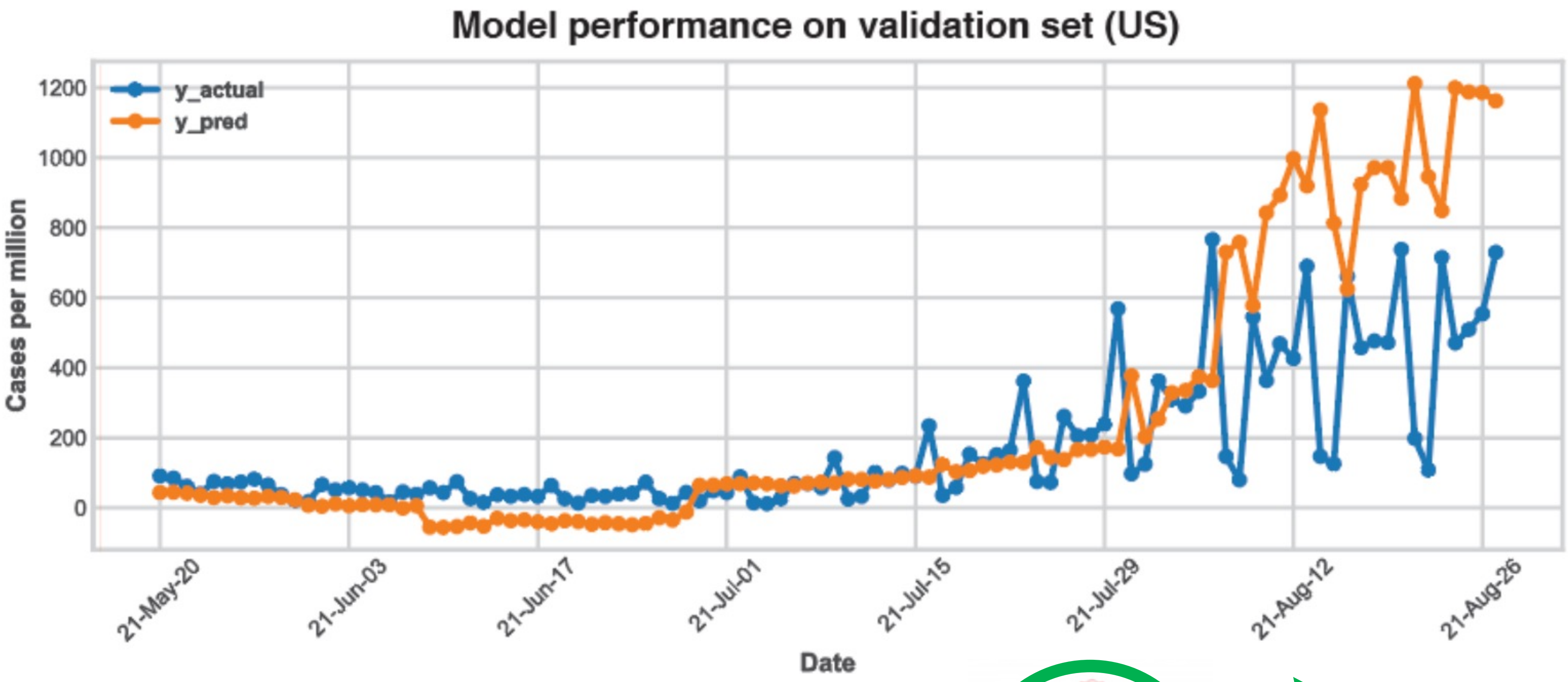
MLP performance on train set

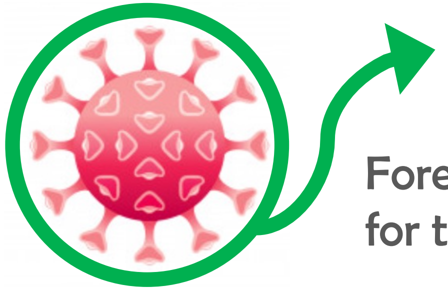




Forecasting COVID-19 cases
for the next 7 days and beyond

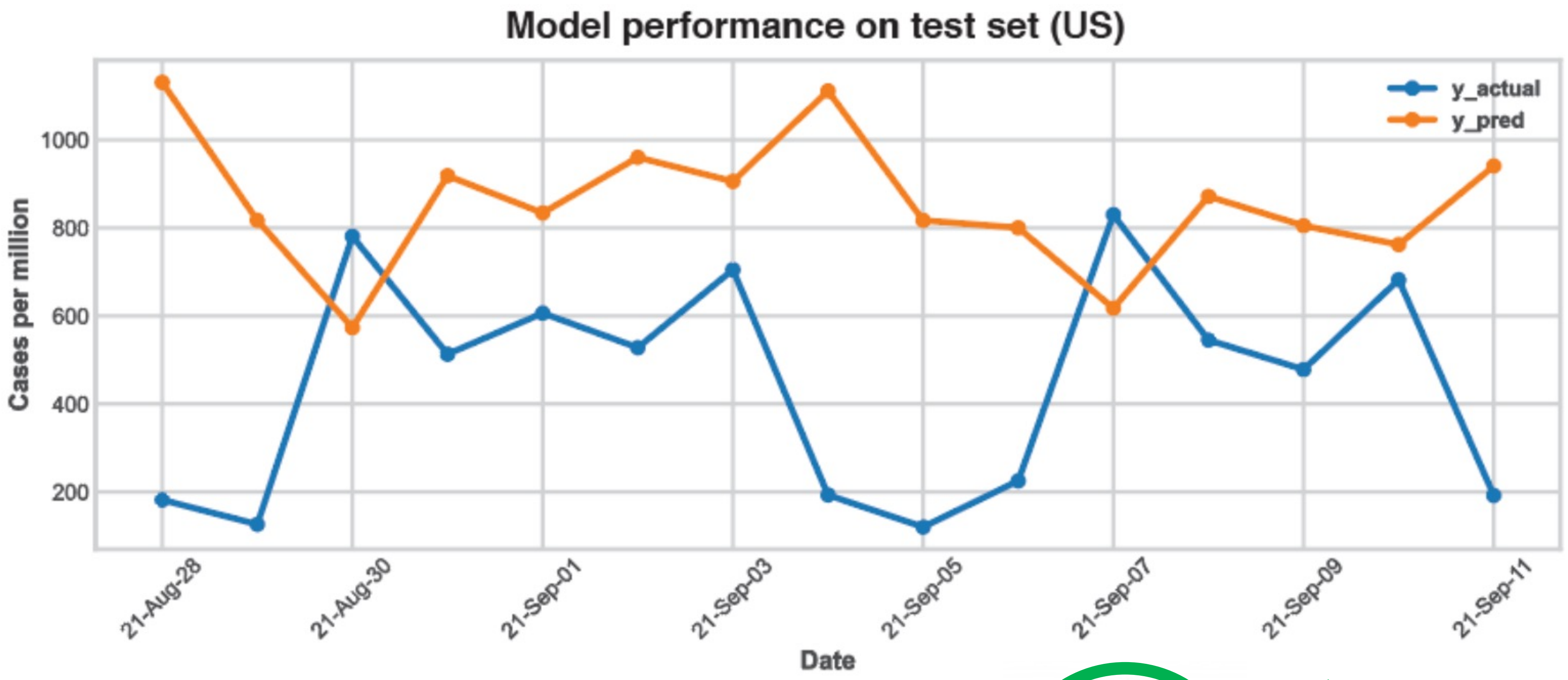
MLP performance on validation set

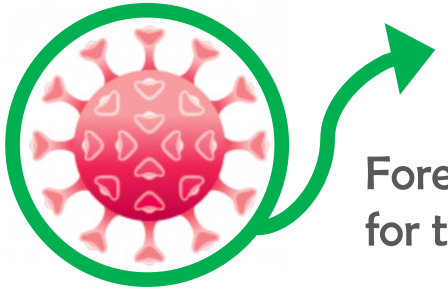




Forecasting COVID-19 cases for the next 7 days and beyond

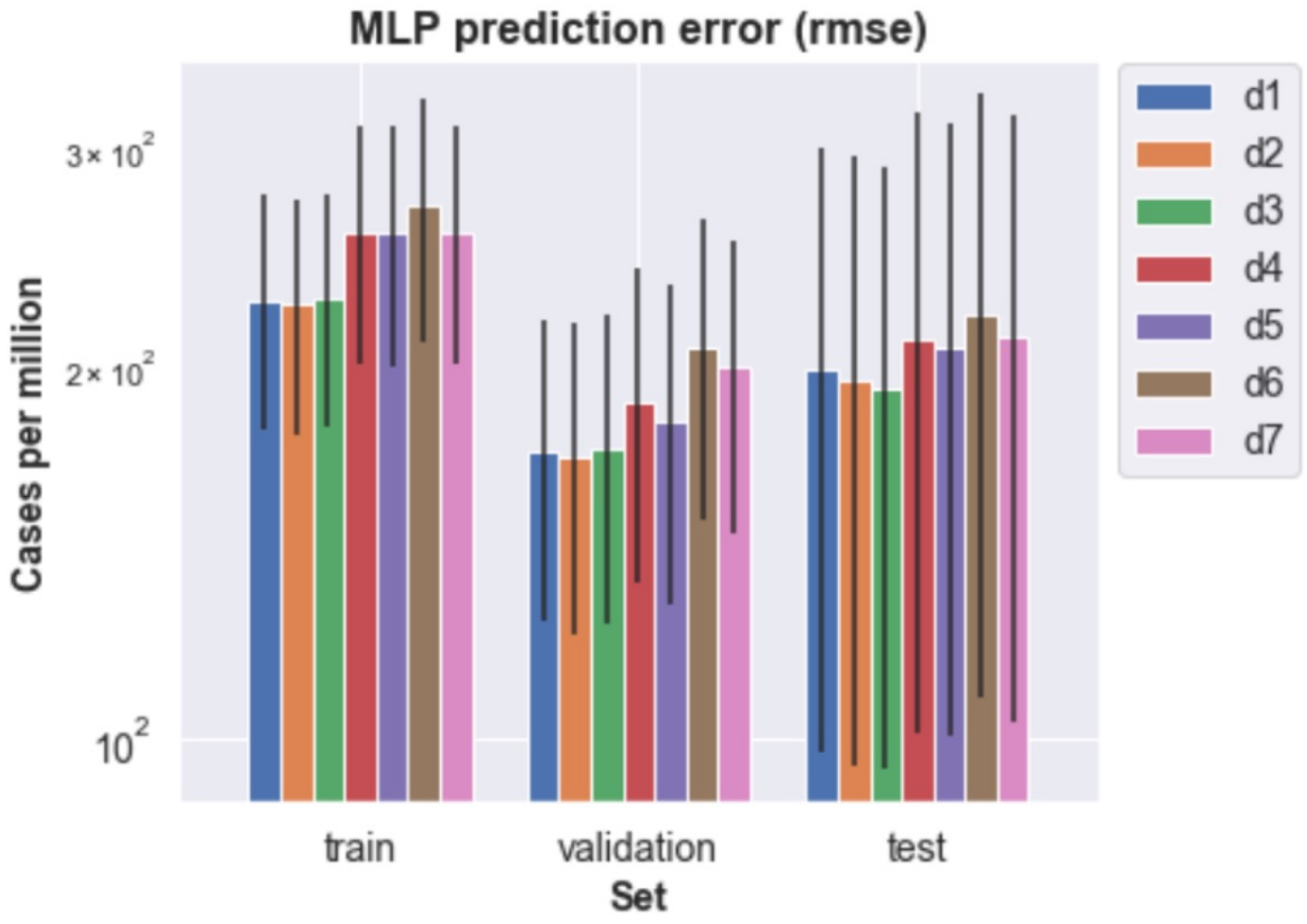
MLP performance on test set





Forecasting COVID-19 cases
for the next 7 days and beyond

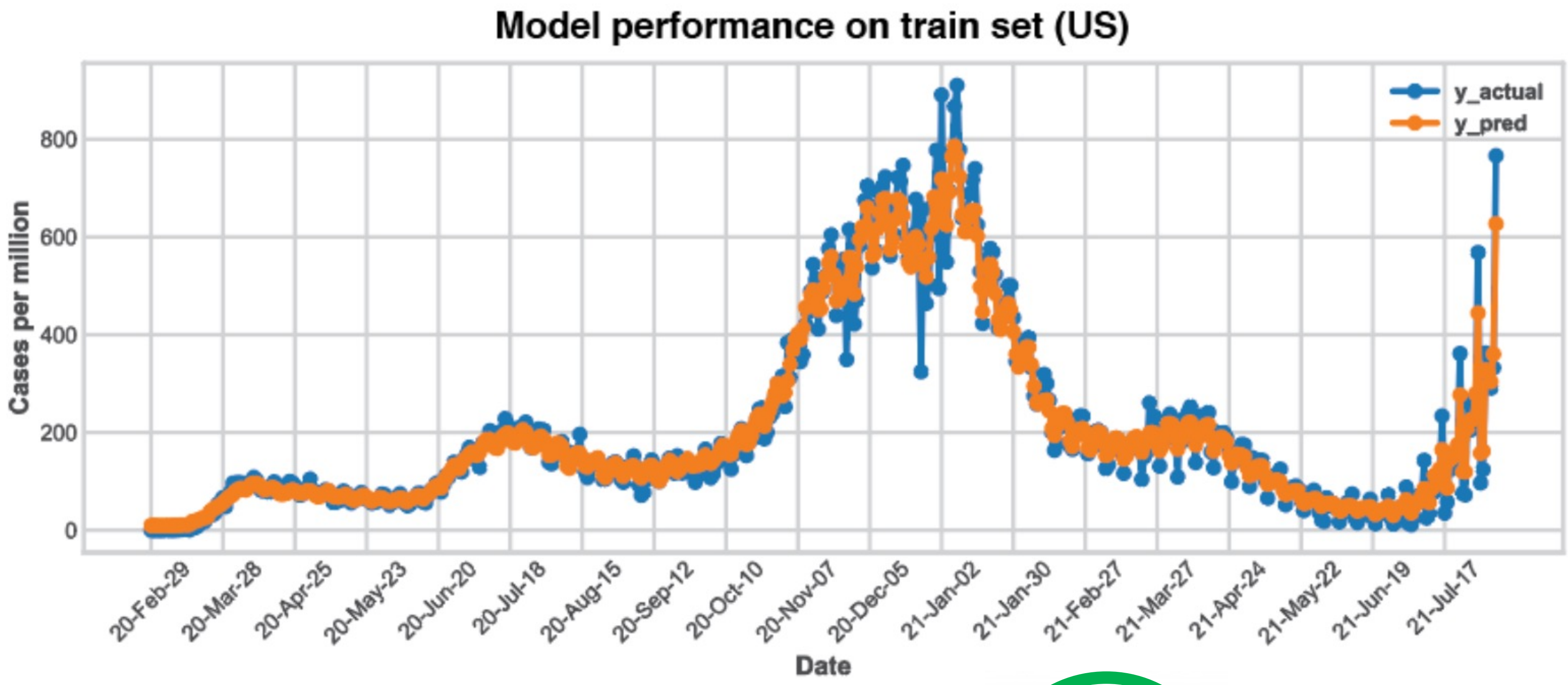
MLP performance on train/validation/test sets

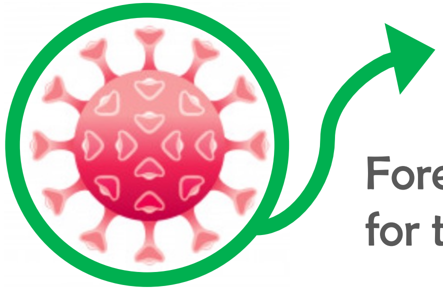


LSTM model architecture with embeddings

Forecasting COVID-19 cases
for the next 7 days and beyond

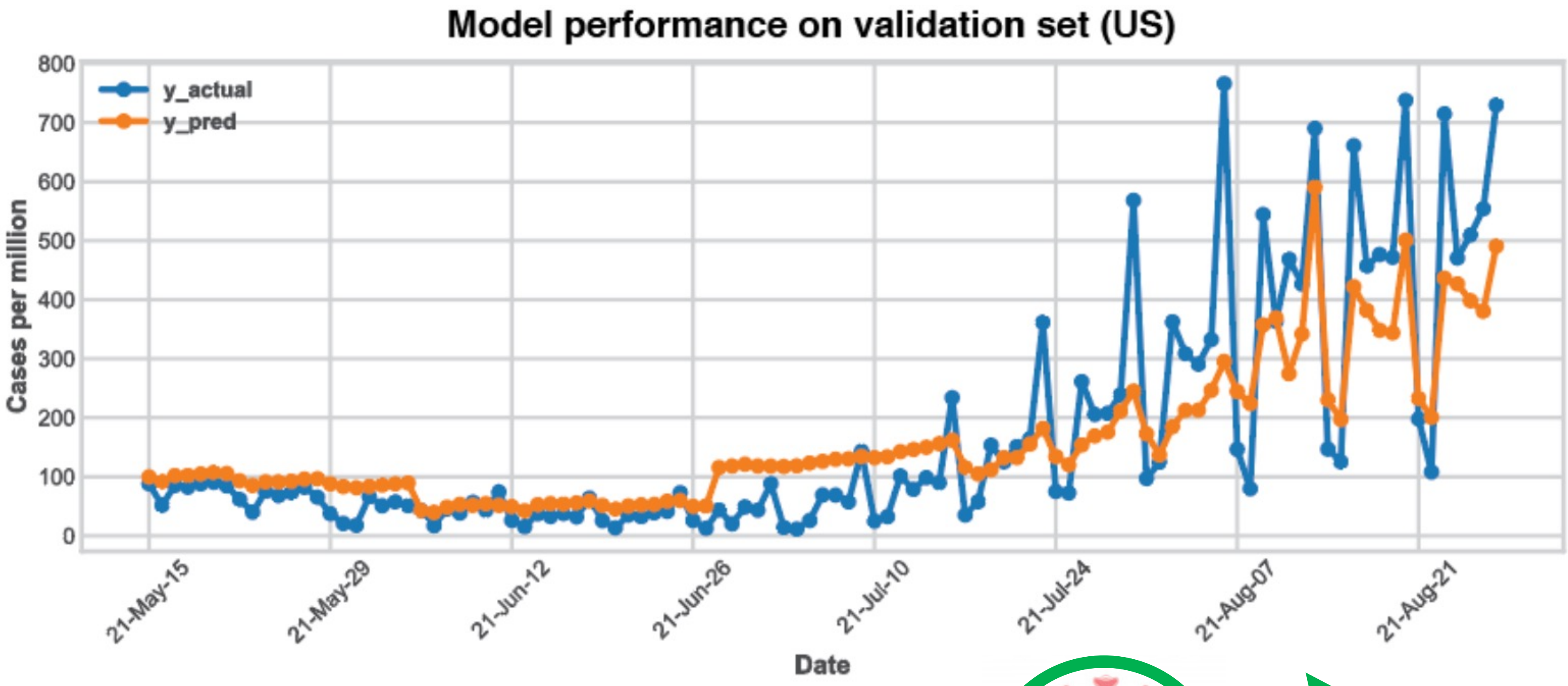
LSTM performance on train set

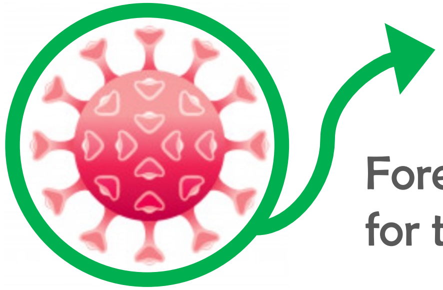




Forecasting COVID-19 cases
for the next 7 days and beyond

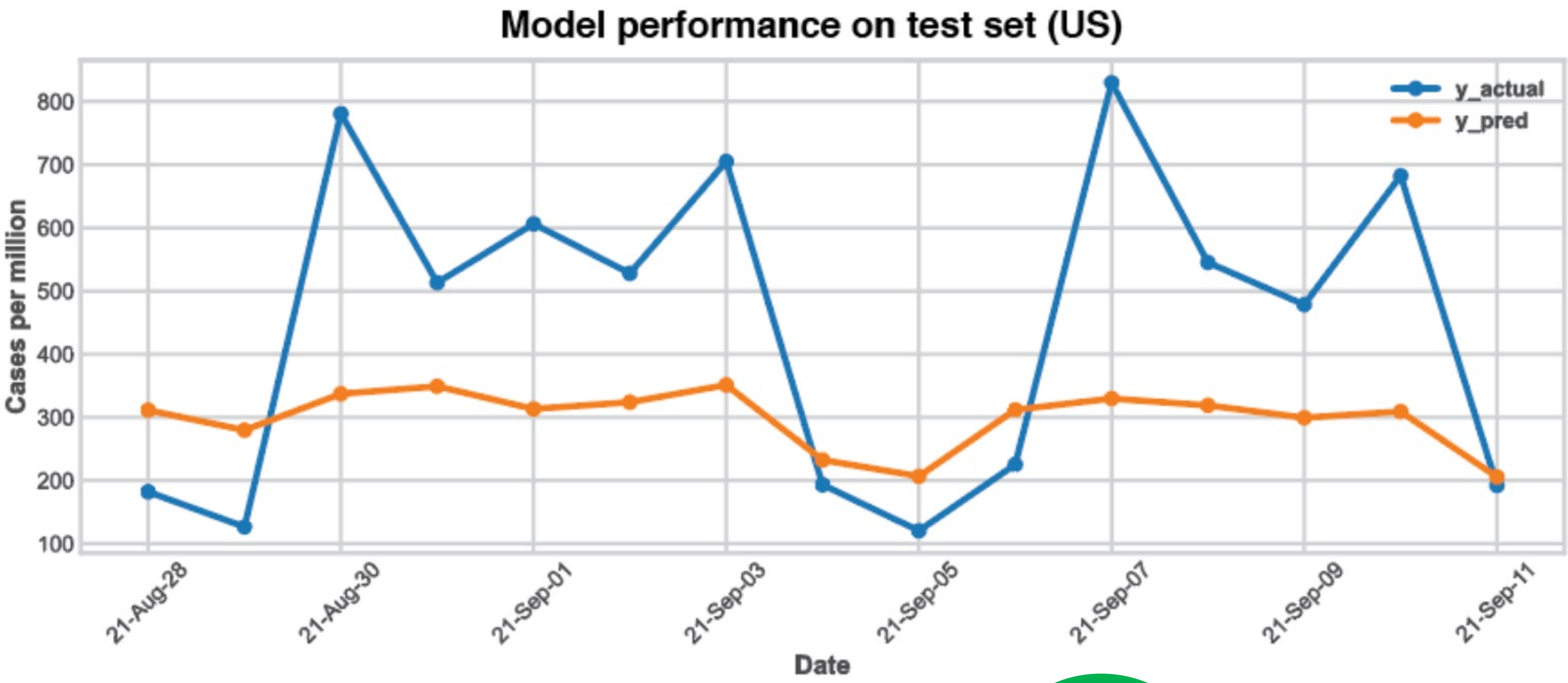
LSTM performance on validation set

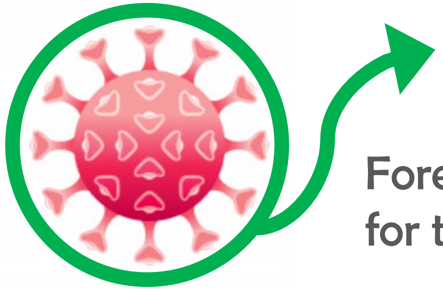




Forecasting COVID-19 cases
for the next 7 days and beyond

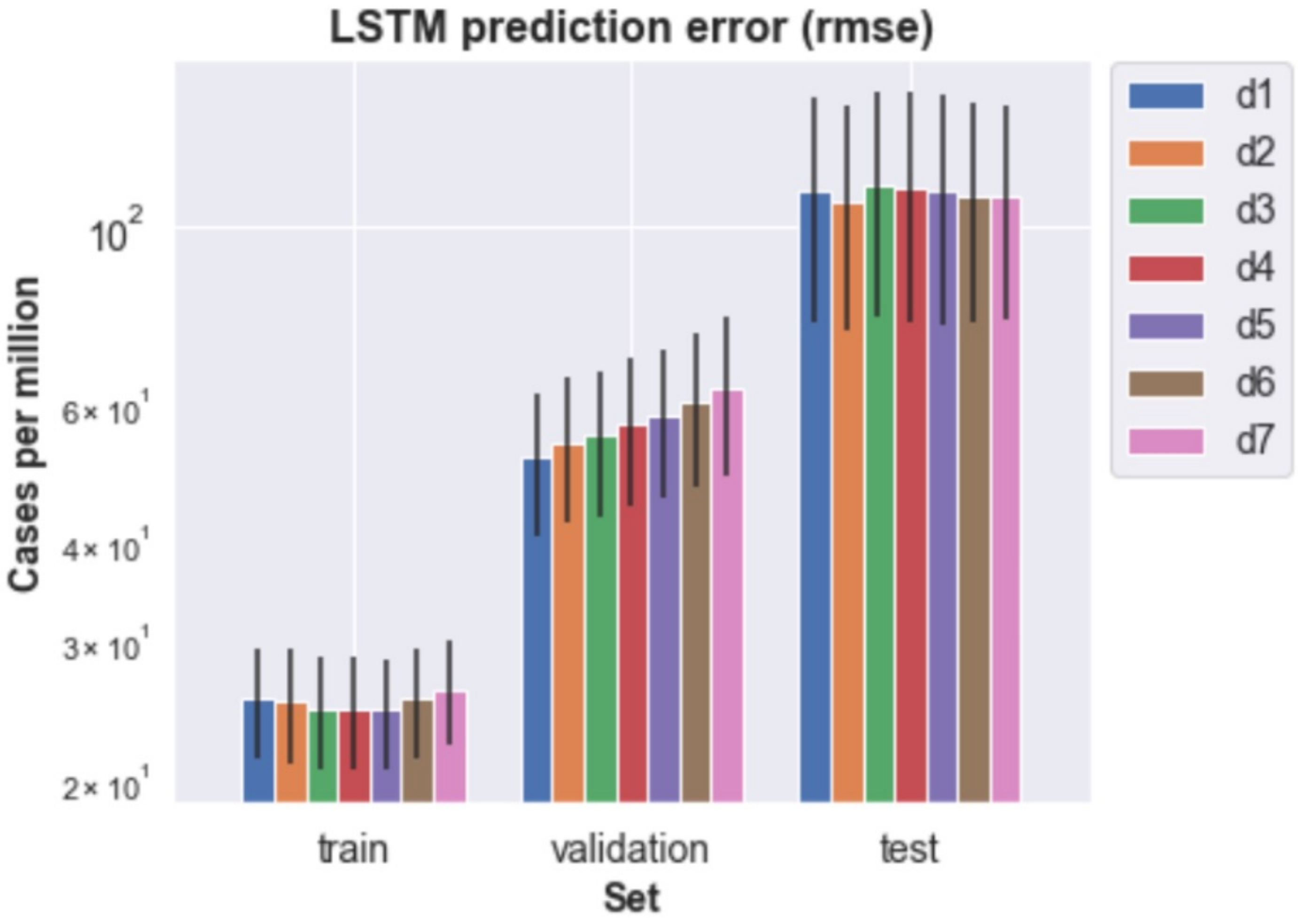
LSTM performance on test set



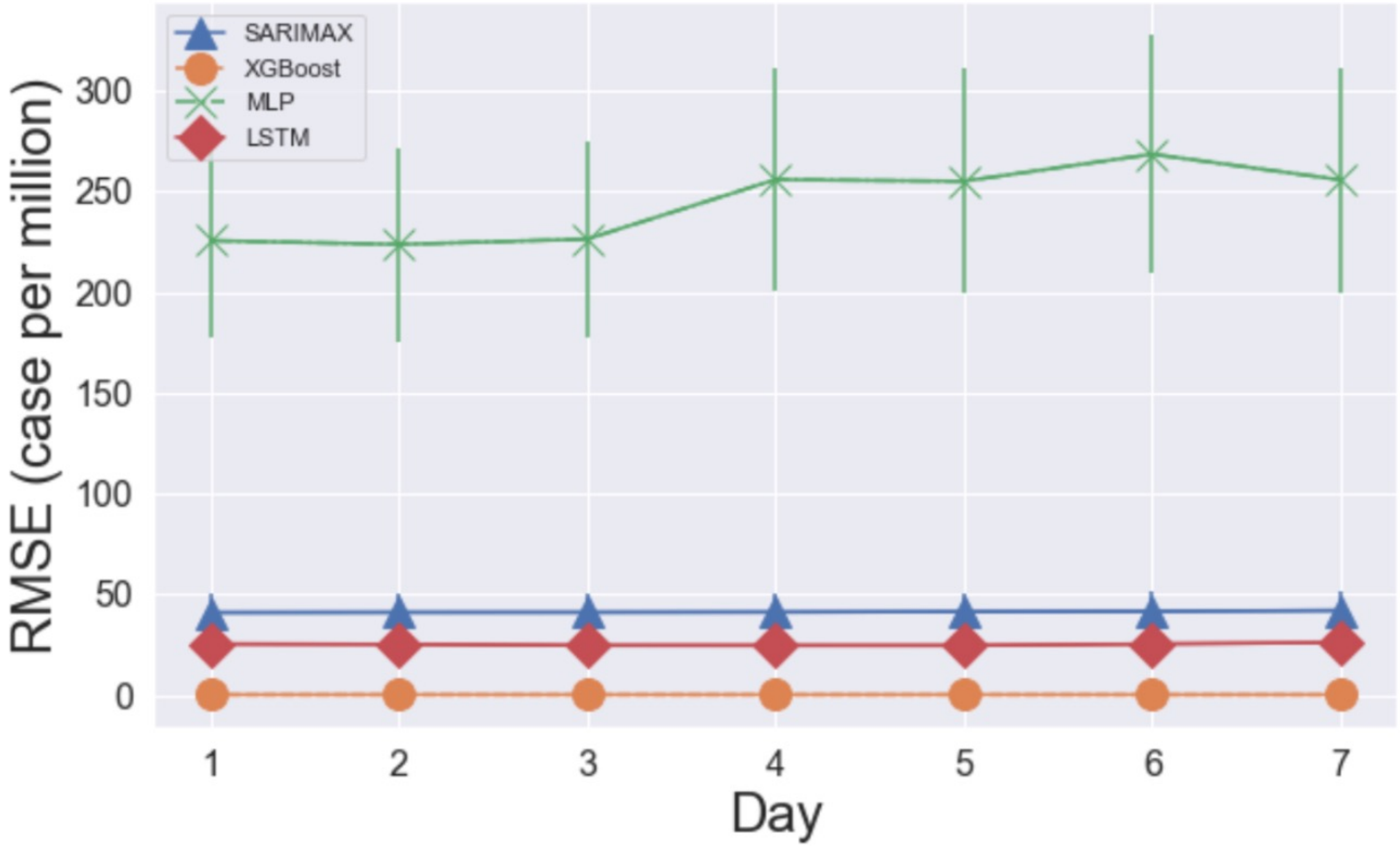


Forecasting COVID-19 cases
for the next 7 days and beyond

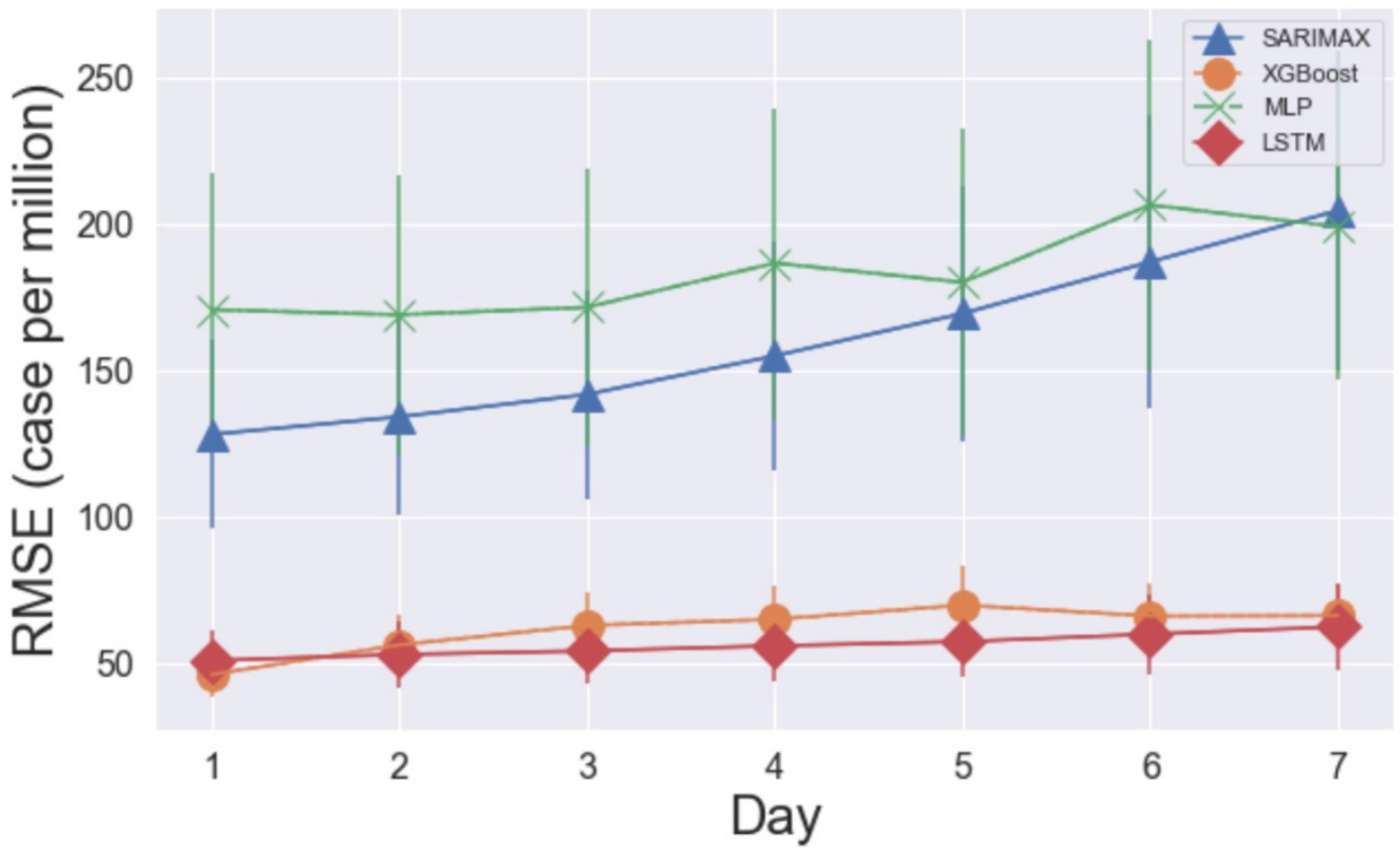
LSTM performance on train/validation/test sets



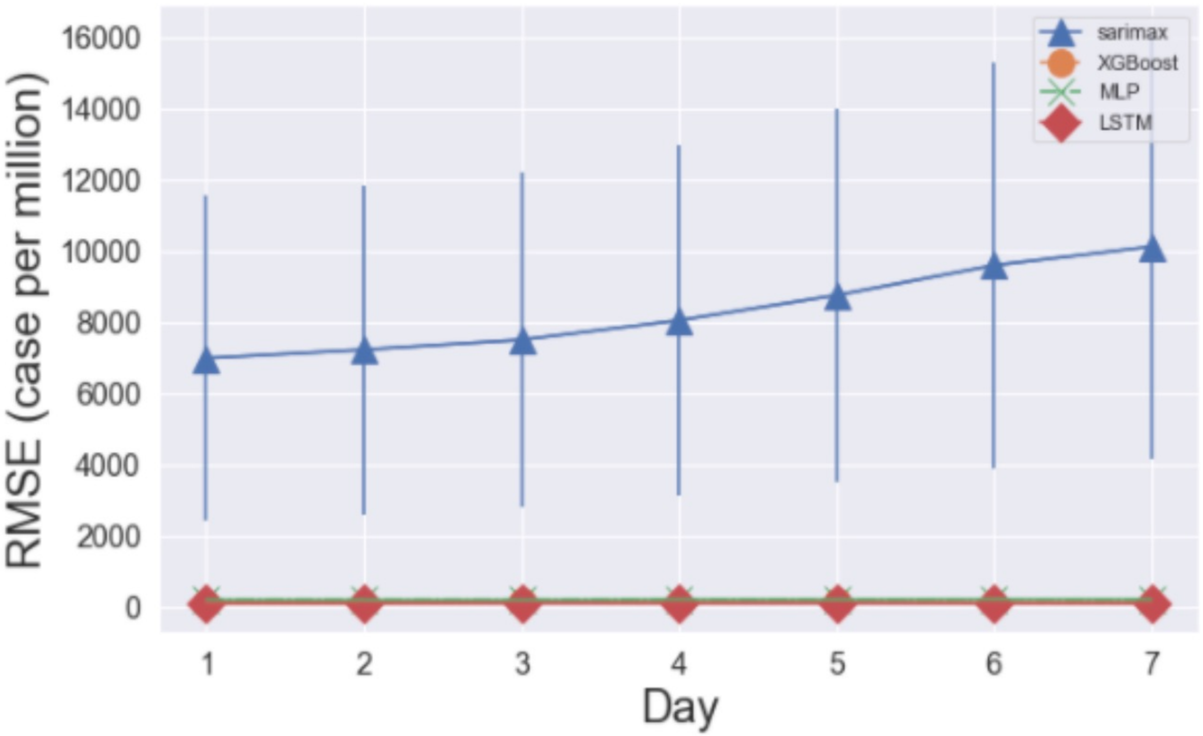
Comparing model performance on train set

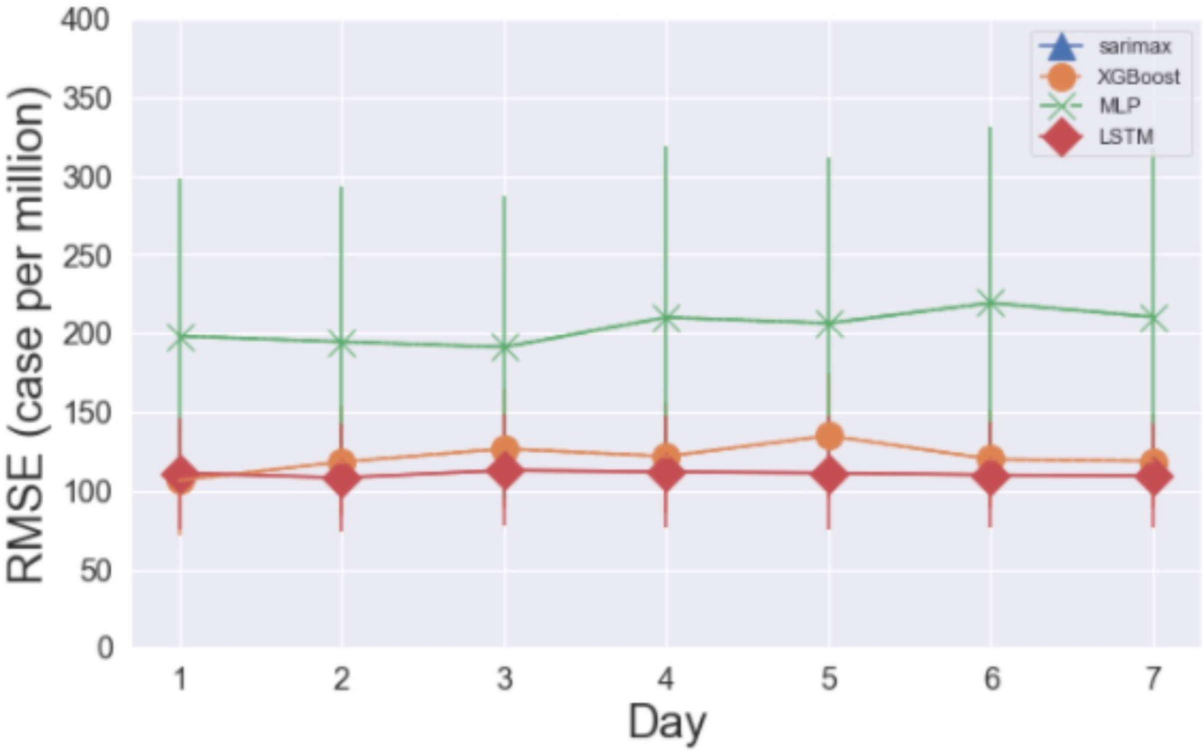


Comparing model performance on validation set



Comparing model performance on test set





Conclusion

- A traditional econometric model (SARIMAX) performed poorly when predicting far ahead (weeks or months) into the future.
- A gradient-boosting based method (XGBoost regressor) performed well even when predicting the far future.
- XGBoost was overfitted during training, meaning that with proper regularization it may perform even better.
- MLP in its current version suffered underfitting, which motivated us to try an architecture with stacked LSTM layers.
- Indeed, adding dual LSTM layers significantly improved forecasting.