Multivariate Timeseries Models

Dengue Fever Forecasting

Valérie Dier, P.Eng




Why Timeseries?

Imagine collecting a patient's health markers regularly over the decade approaching mid-life…

…or process operating conditions, already historized…

…or financial health metrics of individuals or companies…

Glimpse ahead of time, by looking back in time, to see if an event is likely to occur, how severe it could be, and how best to respond.

Modelling Timeseries

Similar, but different

Gather data

Recorded values over time

Explore

Variable change over time

Relationships across a time gap (lag)

Preprocess

Lag the data to model the impact

Model

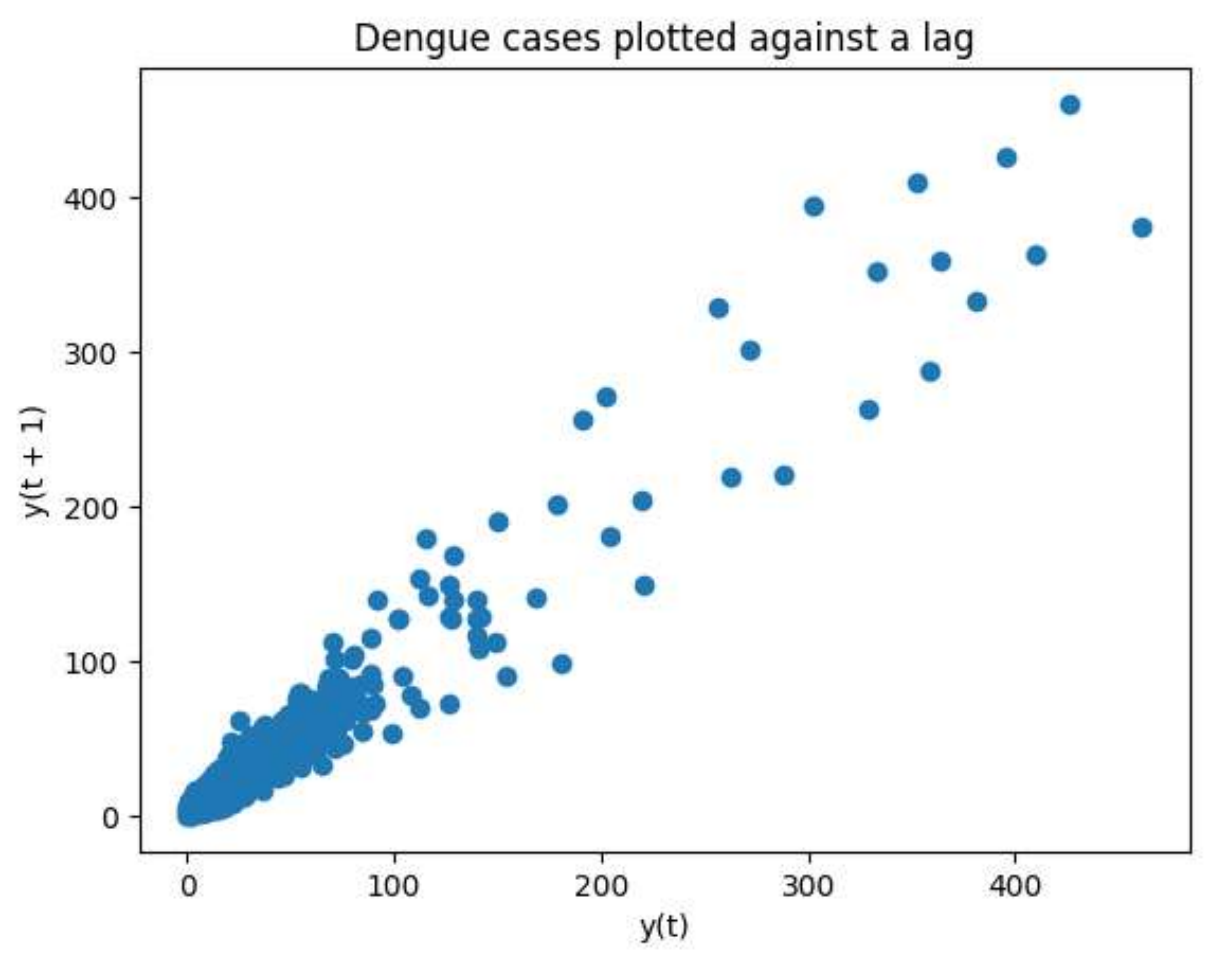
Statistical Modelling

Supervised Learning

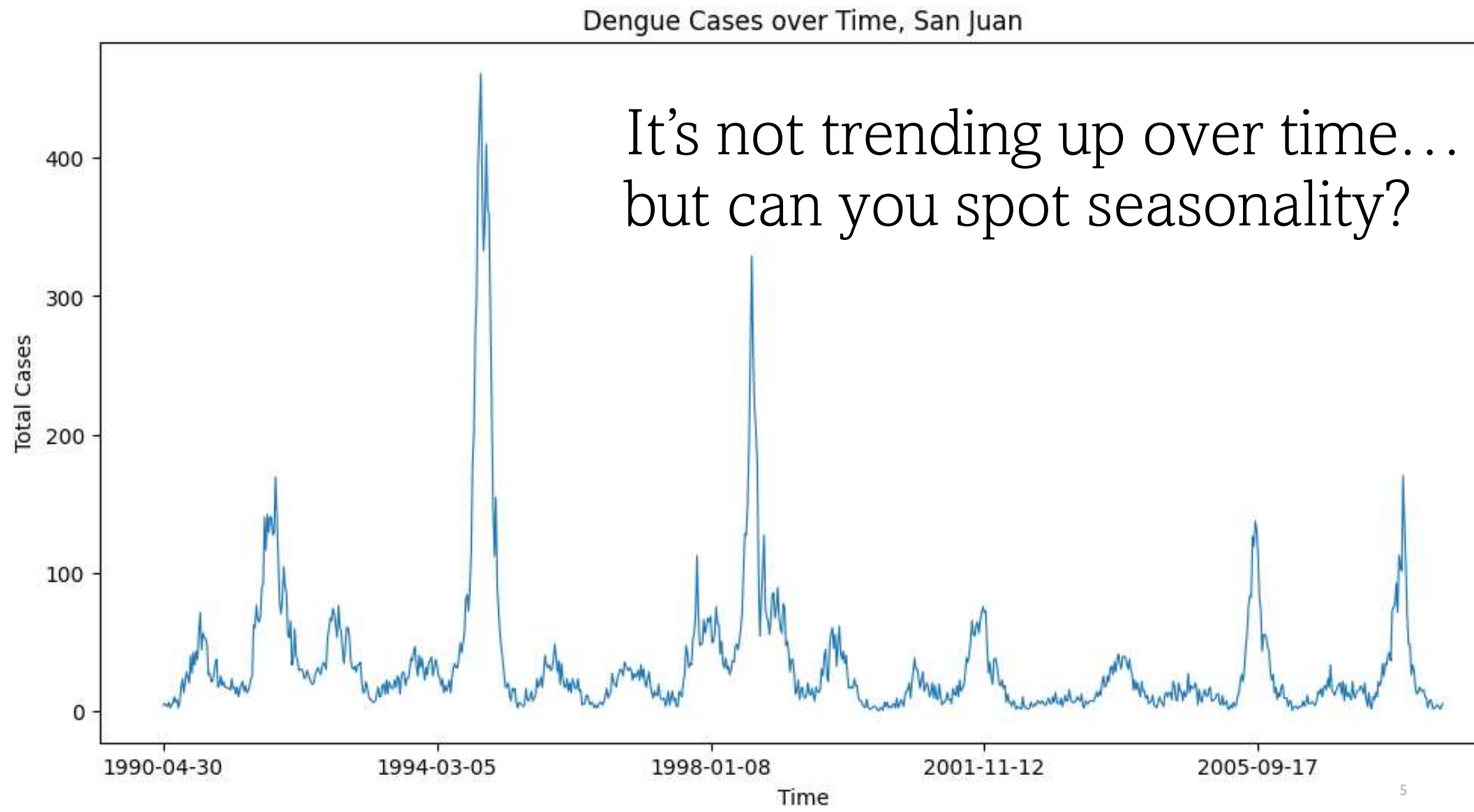
Optimize

Tune the best model

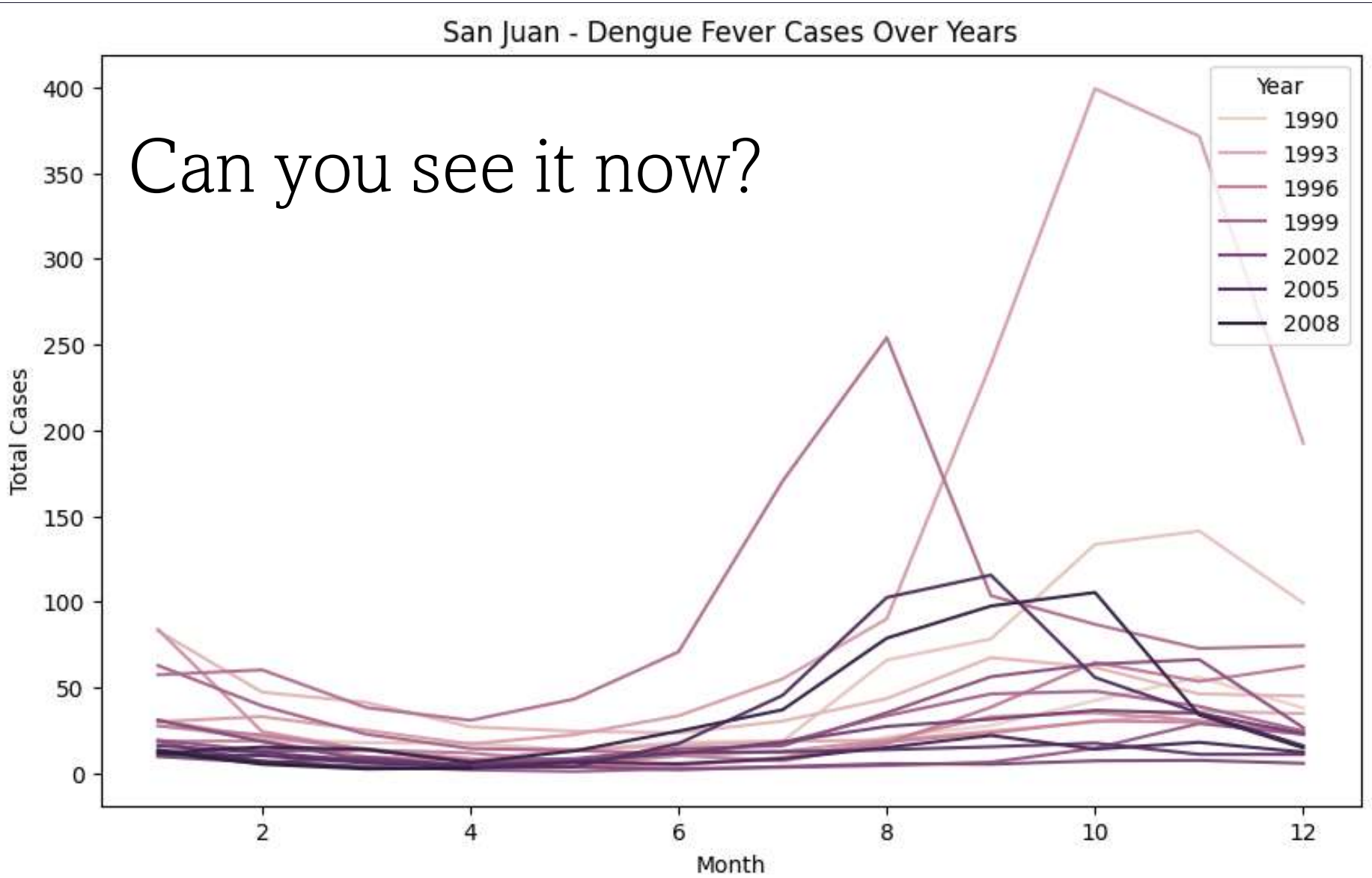
Iterate : revisit inputs?



Is there a relationship between a variable and its past?



It's not trending up over time…
but can you spot seasonality?


San Juan - Dengue Fever Cases Over Years
Can you see it now?
Year
1990
1993
1996
1999
2002
2005
2008
Total Cases
0
50
100
150
200
250
300
350
400
2
4
6
8
10
12
Month

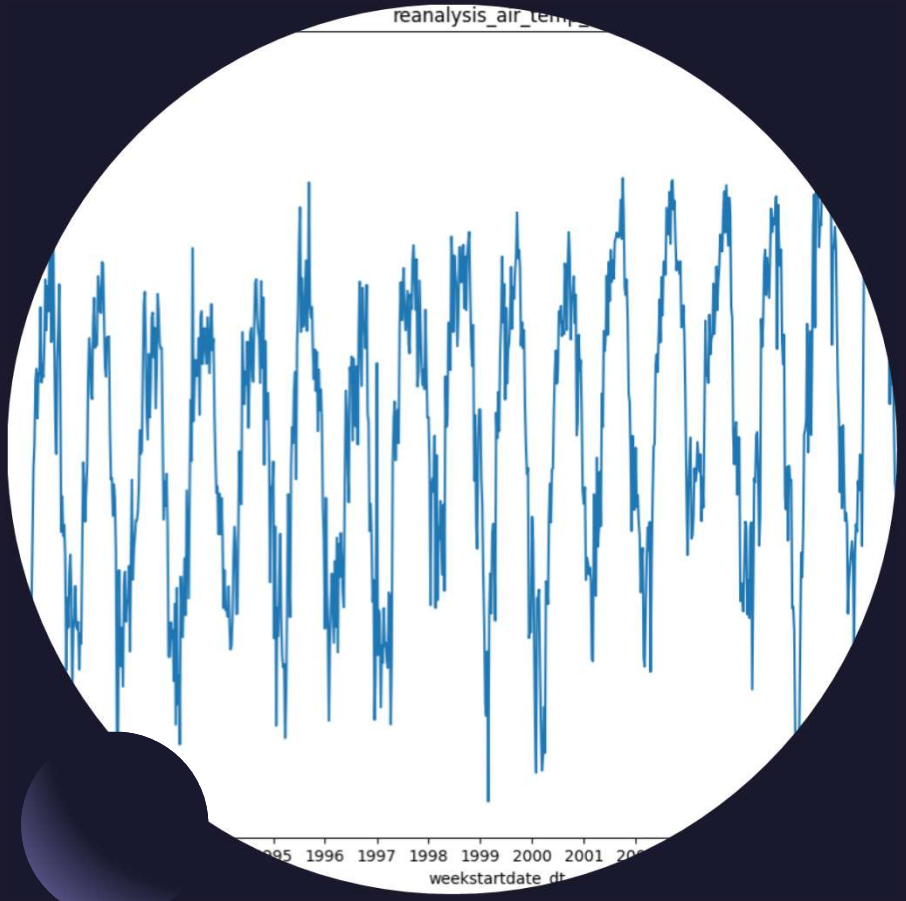
Use differences as inputs if:

- exhibiting trending (up/down), or
- seasonality is present

ADF test determines the need:

- Dengue data doesn't exhibit the undesired traits, but…
- Subjective observations may not agree. Now what?

Some algorithms aren't hampered by these requirements

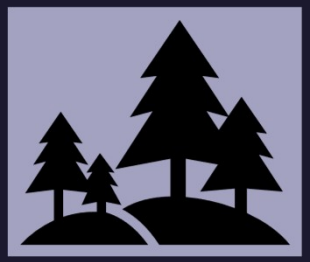


Model Frameworks Trialled

Autoregression

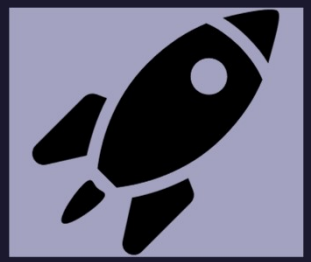
Needs stationary data

(no trending or seasonality)



RandomForest

Handles nonlinearity

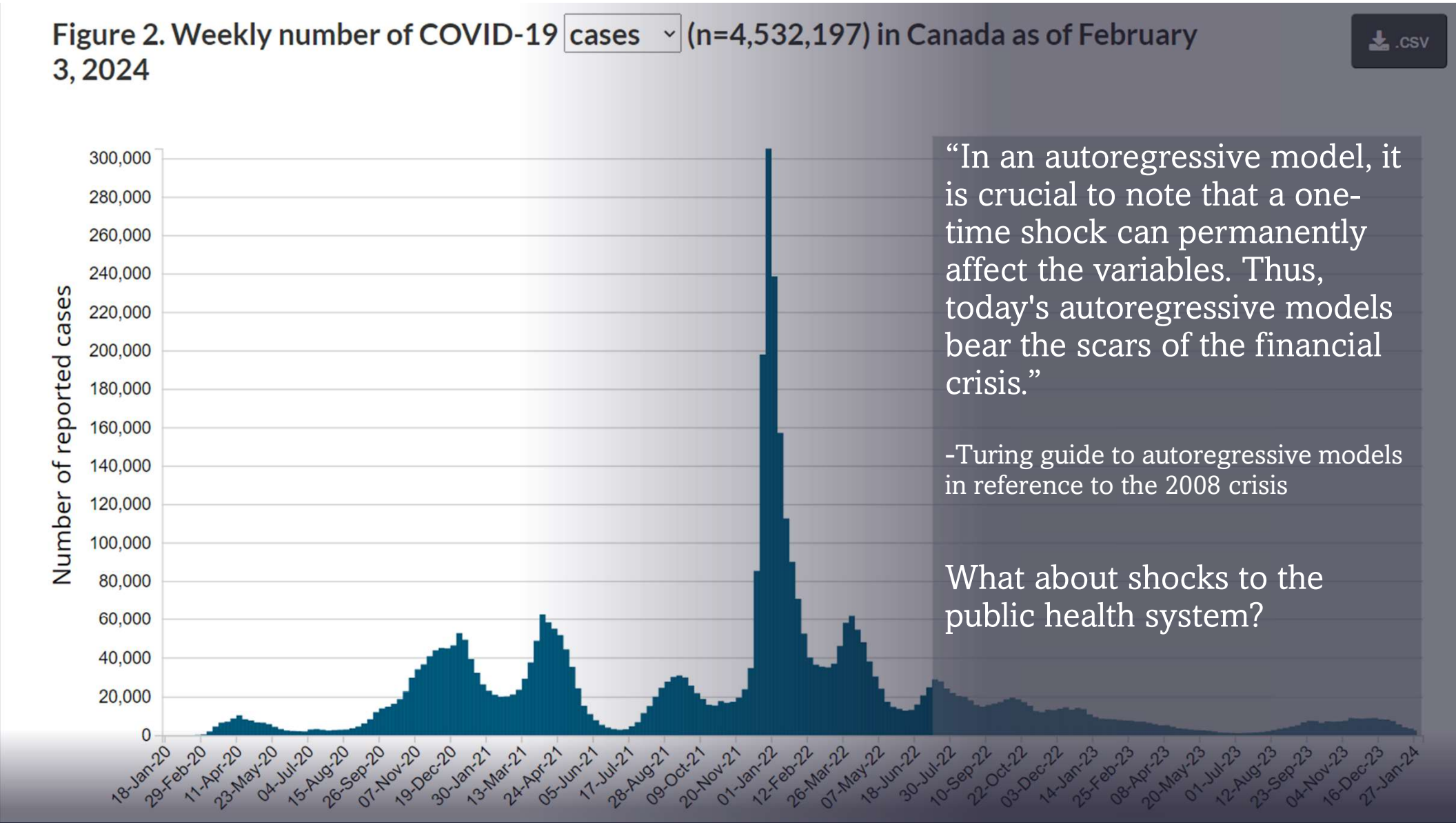


XGBoost

Fast
Accurate

LSTM

Deep Learning: Long short-term memory



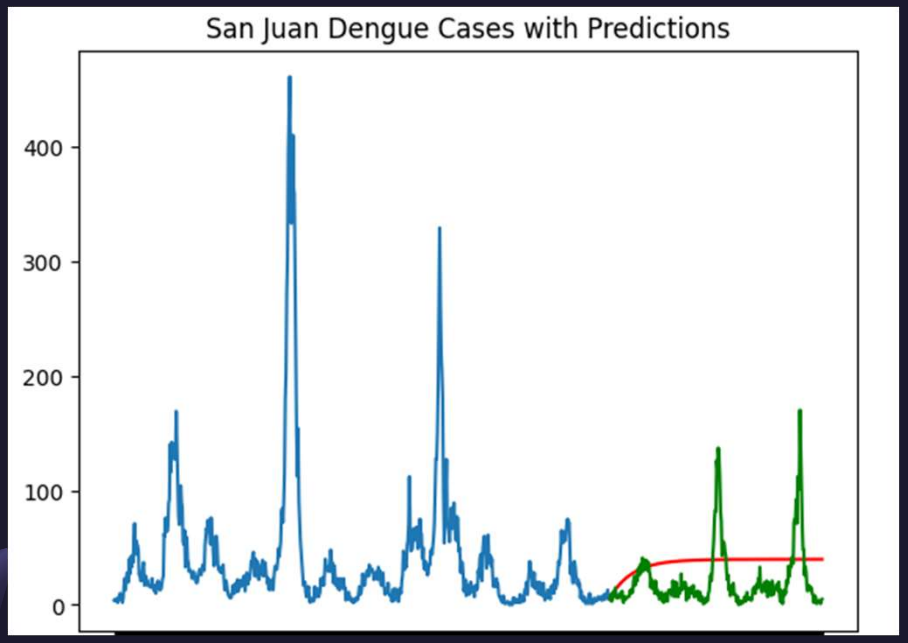
"In an autoregressive model, it is crucial to note that a one-time shock can permanently affect the variables. Thus, today's autoregressive models bear the scars of the financial crisis."

-Turing guide to autoregressive models in reference to the 2008 crisis

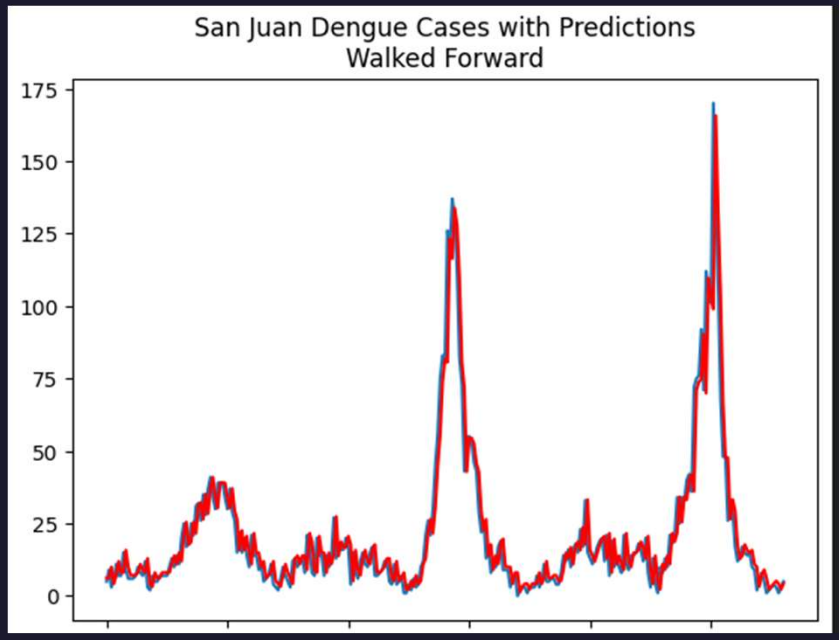
What about shocks to the public health system?

Autoregression

Forecasting without knowledge of next data point



Predictions as new data comes in



What if models were trained in real-time?

Could potentially…

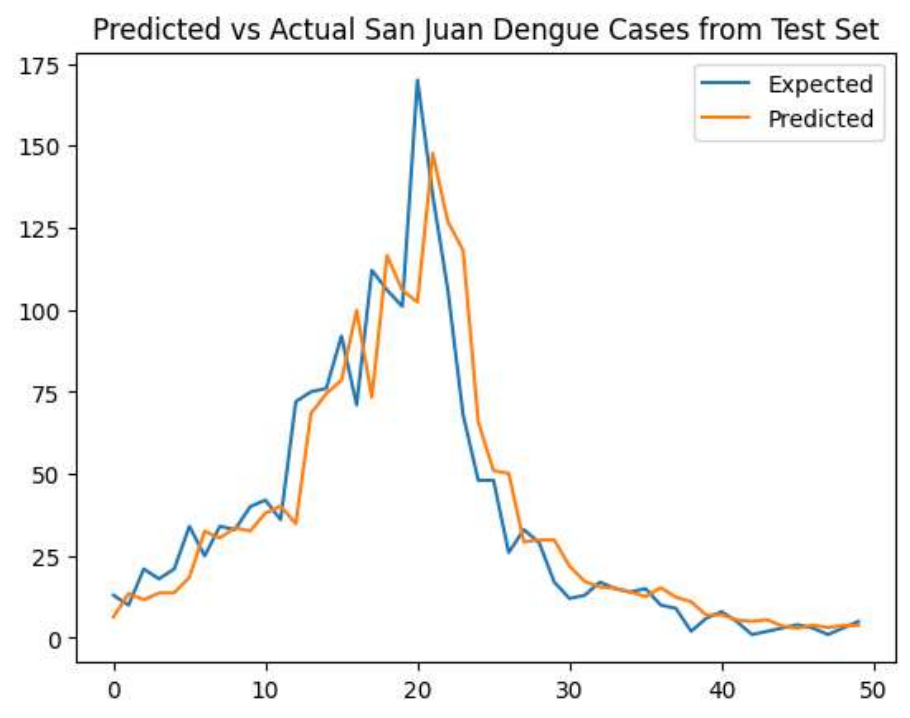
- Capture new information
- Produce representative predictions

However, we need…

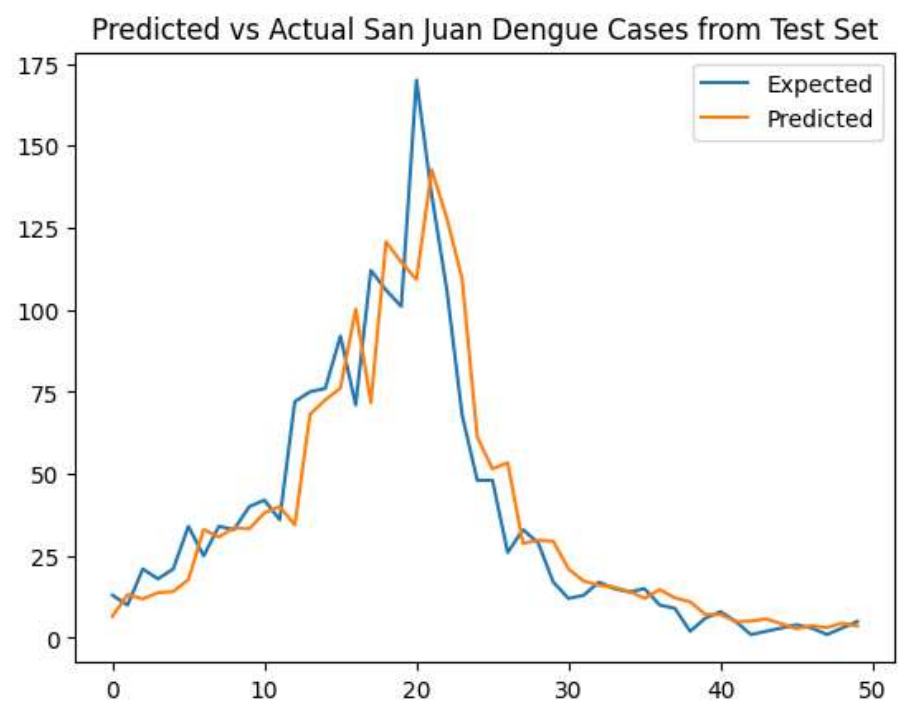
- Rigorous monitoring of data quality
- Guardrails against poor predictions

RandomForest

Dynamic model

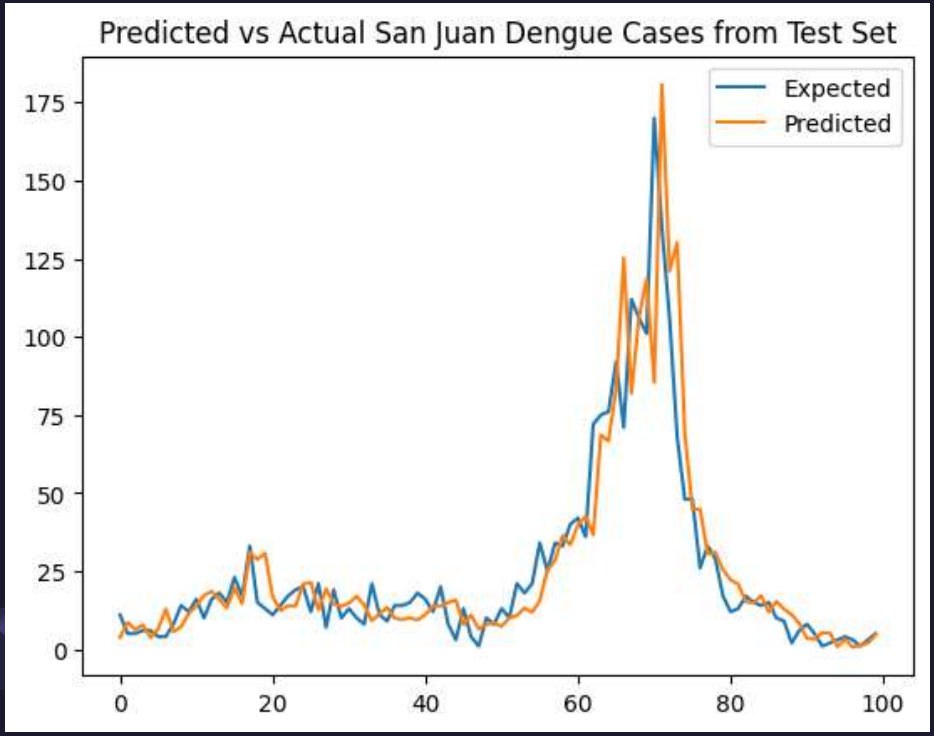


Static model

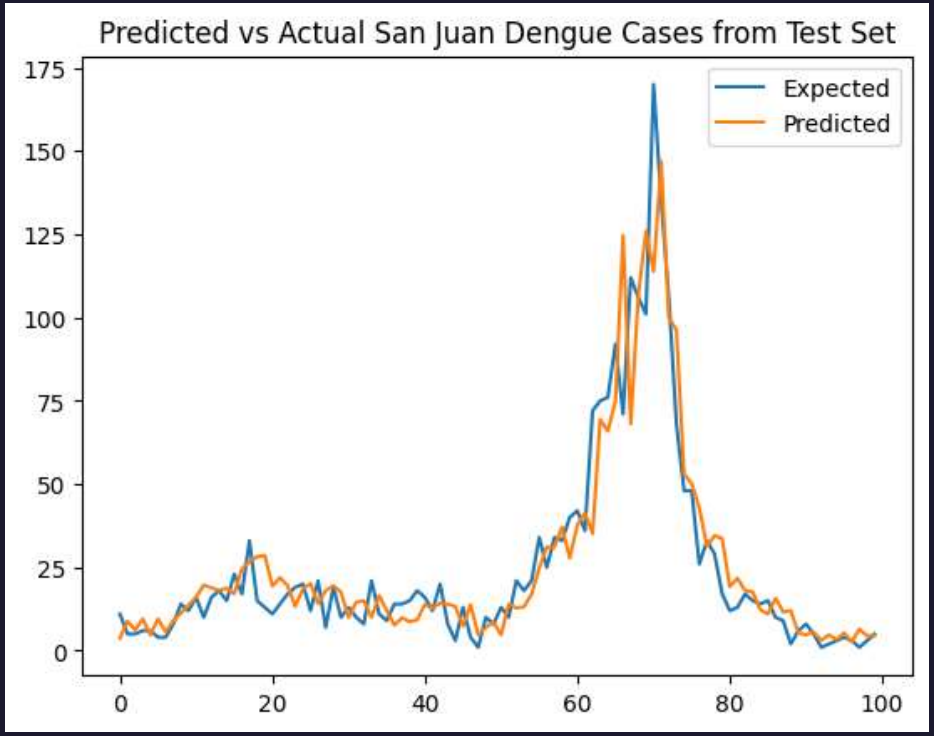


XGBoost

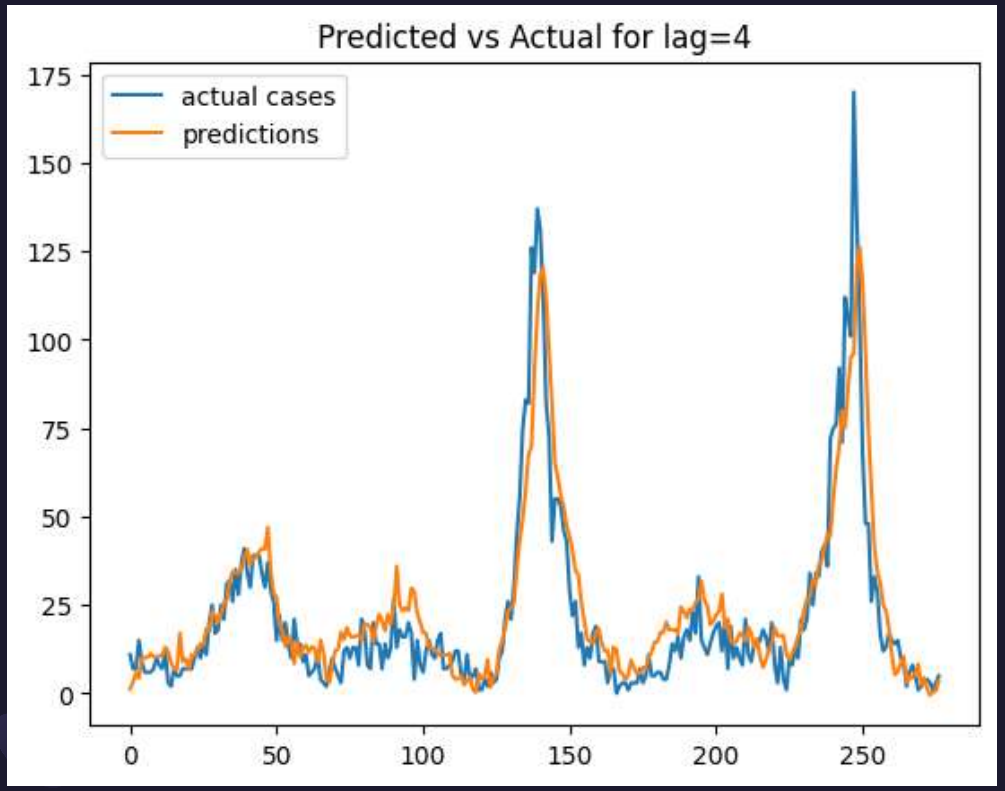
Dynamic model



Static model

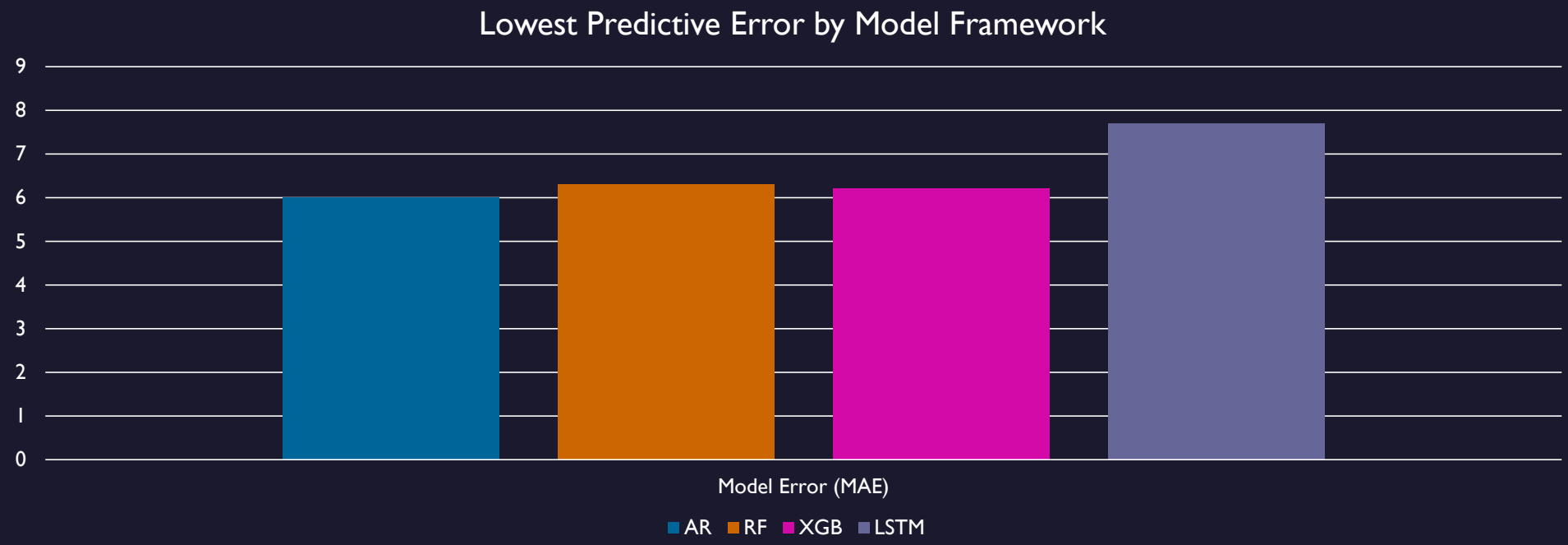


LSTM



- Long short-term memory :
  - Recurrent neural network
  - Complex to understand vs decision trees
- Inverting the scaling on the results of this model requires careful attention

Model errors : a decisional support



By far the most important feature identified by feature importance in the XGBoost model was the target variable, total dengue cases, lagged by one time step.

What's next?

Back to the features: what were the top contributors?

Explore other statistical models, deep learning algorithms beyond LSTM, and packages like Prophet

Automate via pipelining : the end goal is operationalization

Stay Connected

Valérie Dier, P.Eng

linkedin.com/in/valeriedier

github.com/ValerieDier