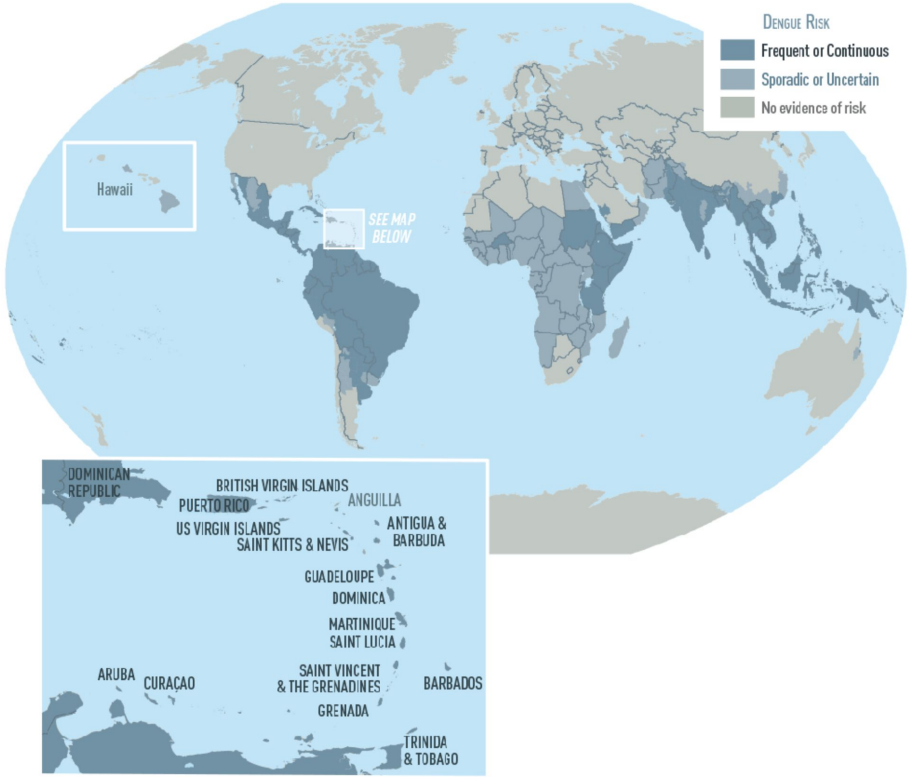
Forecasting disease from univariate time series using naive, ARIMA, exponential smoothing, additive regression, and LSTM models

Client summary

Capstone project for
Springboard Data Science Career Track

Mike Pierovich
June 22, 2020

Challenge: Create a multi-year forecast of dengue fever



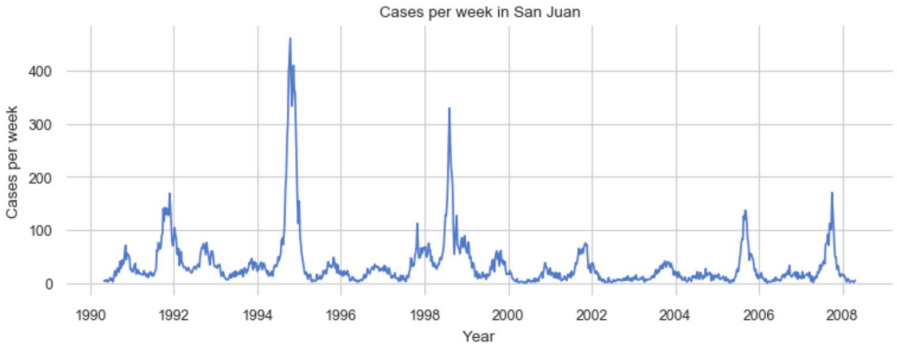
Data: Univariate time series of weekly cases

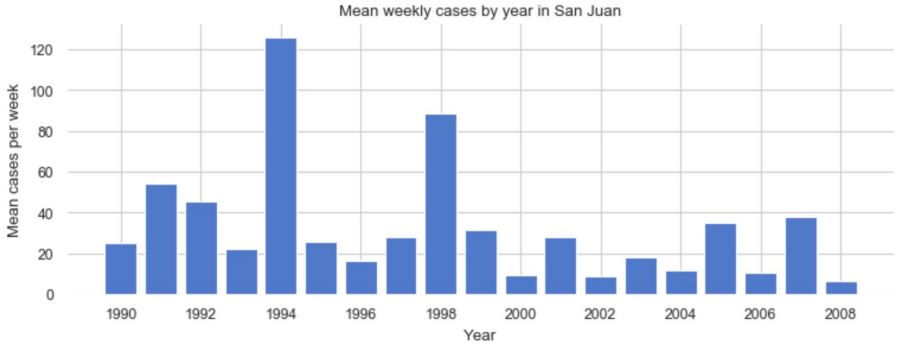
DRIVENDATA

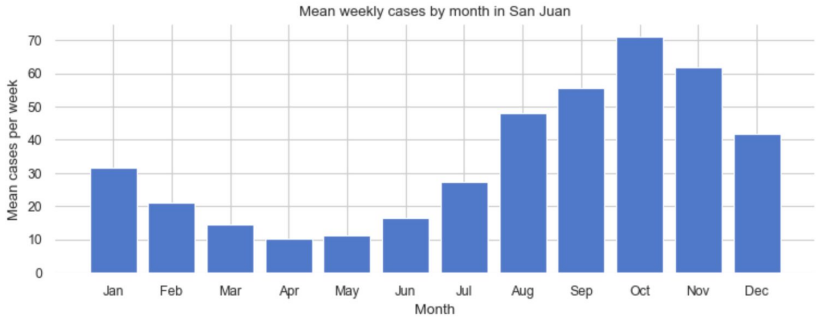
San Juan, Puerto Rico

Iquitos, Peru

Explore: Spikes, slight downward trend and strong seasonality

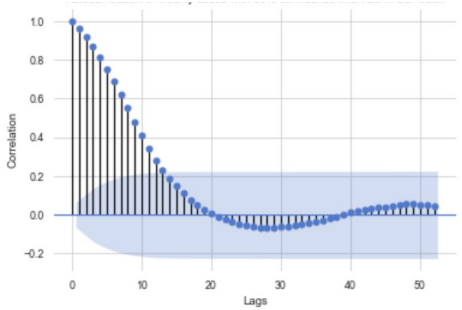




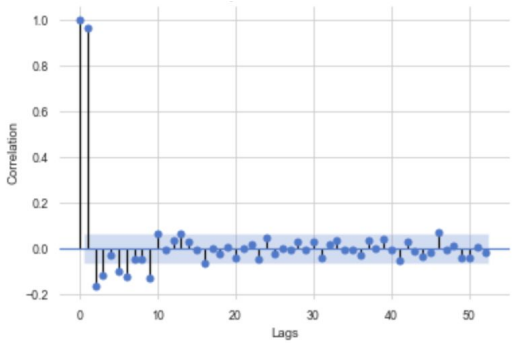


Explore: Time series statistics

- Strong autocorrelation
- Lack of stationarity



Autocorrelation of weekly cases
with 95% confidence interval in San Juan



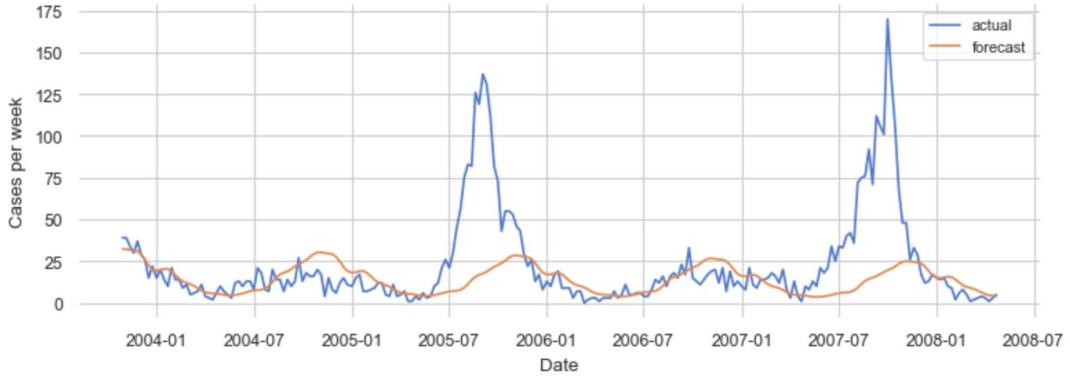
Partial autocorrelation of weekly cases
with 95% confidence interval in San Juan

Models: Additive regression on log-transformed data is best

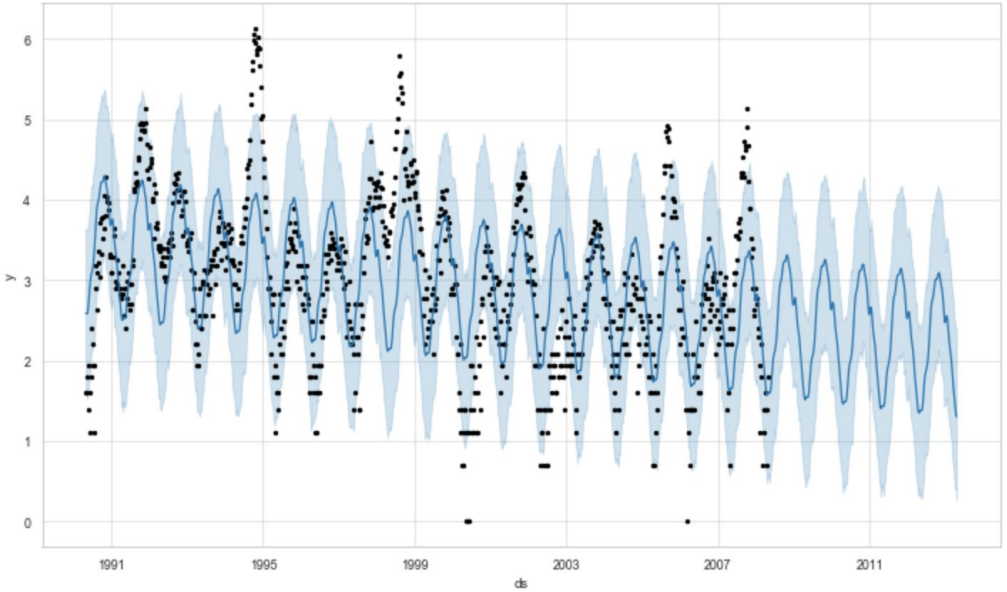
Performance of top-10 models and variations, ranked by MAE, for San Juan

Approach	Summary	Transform	RMSE	MAE	Rank
Additive Regression	Stabilized Trend	Log (x+1)	28.70	14.59	1
Additive Regression	Cap and Floor	Log (x+1)	29.94	14.64	2
Exponential Smoothing	Seasonal ES, a=.5, b=.1, g=0, optimized	Log (x+1)	26.54	15.54	3
ARIMA	SARIMAX (2, 1, 2) x (2, 0, 1, 52)	Log (x+1)	26.61	15.85	4
Exponential Smoothing	Seasonal ES, a=.9, b=.8, g=.1, optimized	None	28.60	17.14	5
ARIMA	SARMIAX (3, 1, 2) x (0, 0, 0, 52)	None	29.48	18.71	6
Exponential Smoothing	Simple ES, a=0	None	34.12	19.09	7
Additive Regression	Stabilized Trend	None	29.20	20.05	8
Naive	Seasonal Naive Method	None	35.95	21.34	9
LSTM	ConvLSTM	Log (x+1)	34.29	21.36	10

Actual and forecast of an additive model using Facebook Prophet with “stabilized trend” arguments on log-transformed test data for San Juan



Final forecast



Facebook Prophet's visualization of the final forecast on a log(x+1) scale for San Juan

Example forecast value:

- Week of August 19, 2008, which is 6 months into forecast period
- Forecast value is 26 cases
- An 80% constant confidence interval
  - Lower bound is 9 cases
  - Upper bound is 72 cases

Lessons and improvements

Lessons from modeling:

- Importance of well-transformed data
- Connection between data and model
- Limits of univariate data
- LSTM’s lack of success

Improvements:

- Better understand business context and domain
- Refine forecast horizon
- Explore additional time-series models
- Leverage multivariate data
- Explicitly predict spikes