A Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks

- Daniel Hans Munk

UNIVERSITY OF COPENHAGEN

Background: Avian influenza H5N1 (Bird Flu)

- Disease Outbreaks
  - Show Patterns
- Covid-19
- Time Series Analysis
  - ARIMA
  - Random Forest

Trial Data:

- H5N1 outbreaks, 2005-12-08 to 2012-10-28.
  - Temperature
  - Relative humidity

ARIMA (AutoRegressive Integrated Moving Average)

AutoRegressive

$$AR(n):\ Y_{\{future\}} = B_0 + B_1 Y + \sum_{i}^{\{n-1\}} B_{Y_{\{lag:i\}}}$$

The target: ”Y ” is based on its past ”lagged” values: $Y_{\{lag:1\}}, Y_{\{lag:2\}}, \ldots$

Integrated

$$\mathrm{I(n)}:\ Y_{\{future\}} - Y = B_0 + B_1\left(Y - Y_{\{lag:1\}}\right) + \sum_{i=1}^{n-1} B_{i+1}\left(Y_{\{lag:i\}} - Y_{\{lag:i+1\}}\right)$$

Differencing removes: *the changes in the level of a time series *eliminates trends and seasonality -> consequently stabilizing the mean of the time series. Features is not dependent on time.

Moving Average

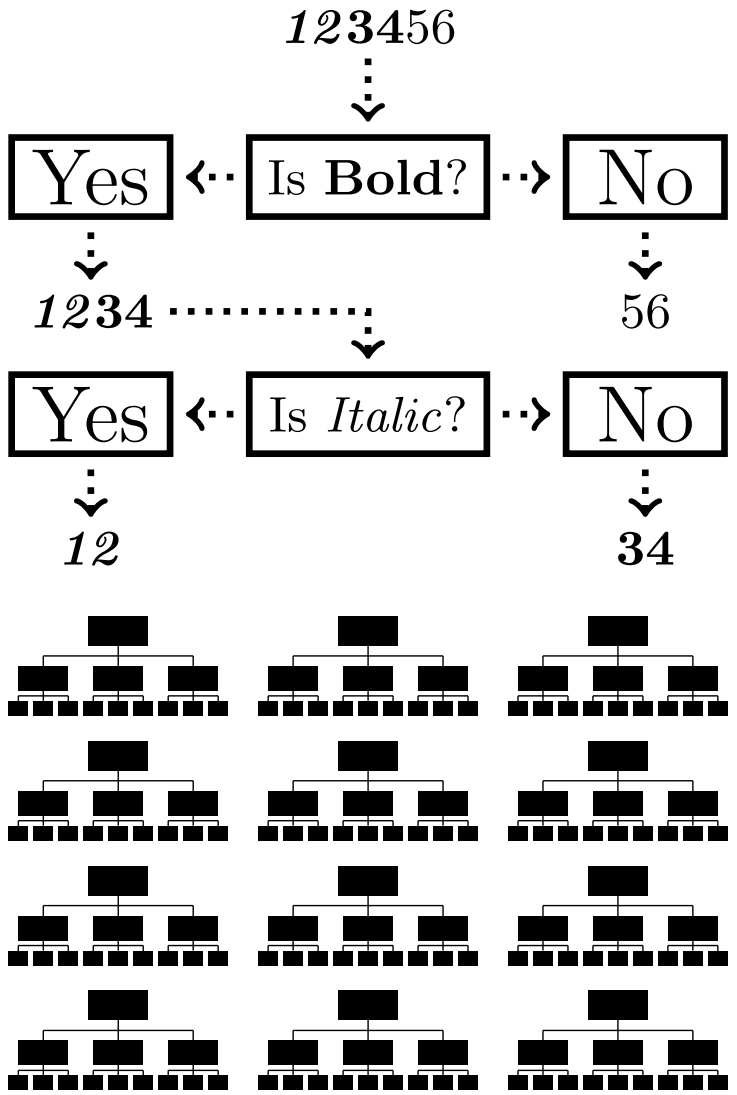
$$MA(n):\ Y_{\{future\}} = B_0 + B_1\varepsilon + \sum_{i}^{n-1} B_i \varepsilon_{\{lag:i\}}$$

The residual error: ε for each lagged versus model value to predict "Y"

Random Forest

- Decision Tree → Random Forest

- Uncorrelation?
  - Bagging/Bootstrap Aggregation
  - Feature Randomness



Retrospective versus Prospective

- Retrospective
  - All prior data → predict all of future
- Prospective
  - 30 week prior → 1 week at a time
  - Append 1 week, train → predict 1 week.
- Create Confusion Matrix
  - Test $\chi 2$ → 0

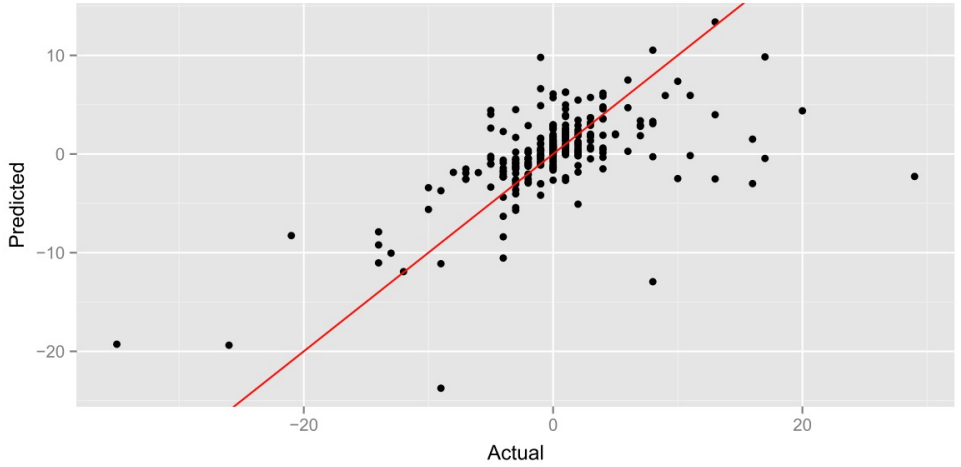
(a) Comparing the MSE of the models. (Table 3 in [1])

	Retro[b]	Pro[c]
ARIMA	26.96	28.74
R.F.[a]	6.32	24.81

(b) Pro[c] R.F.[a] confusion matrix under null. (Table 4 in [1])

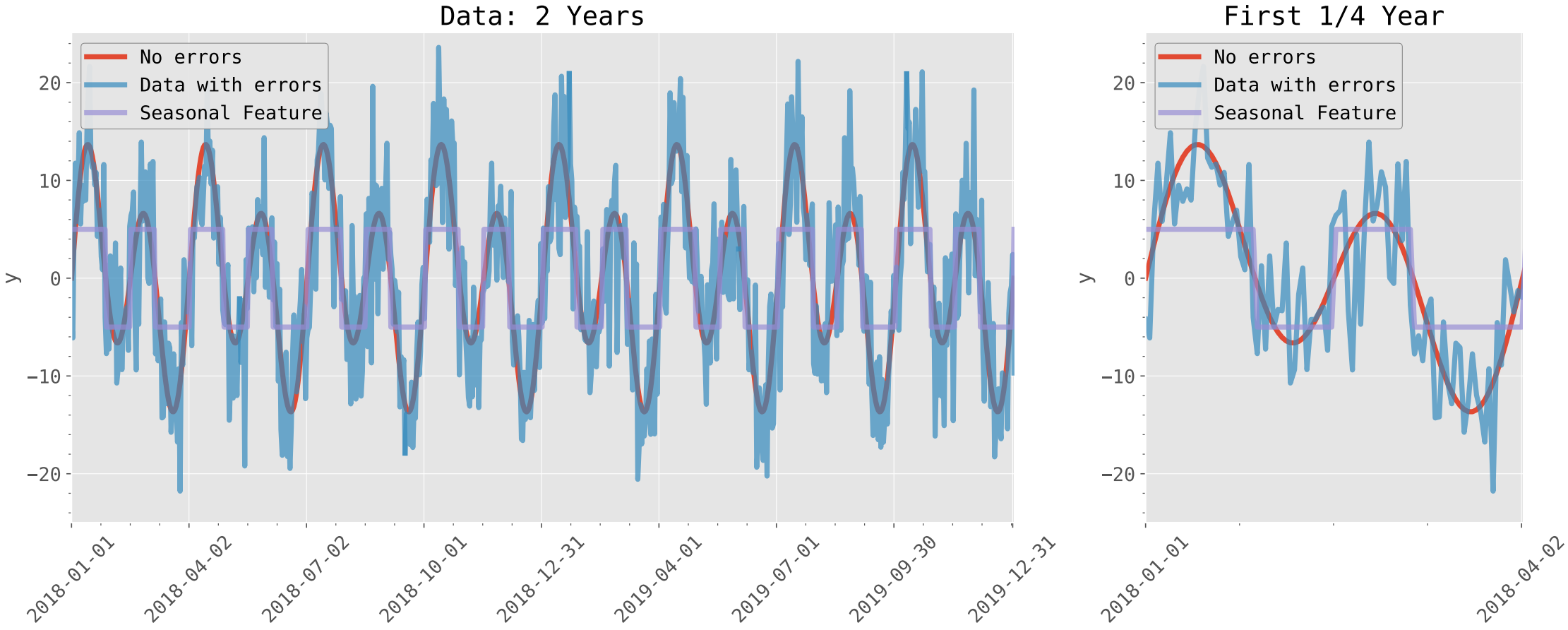
		Predicted	
		Up	Down
Actual	Up	0.3685	0.2222
	Down	0.2553	0.154

[a] Random Forest, [b] Retrospective, [c] Prospective

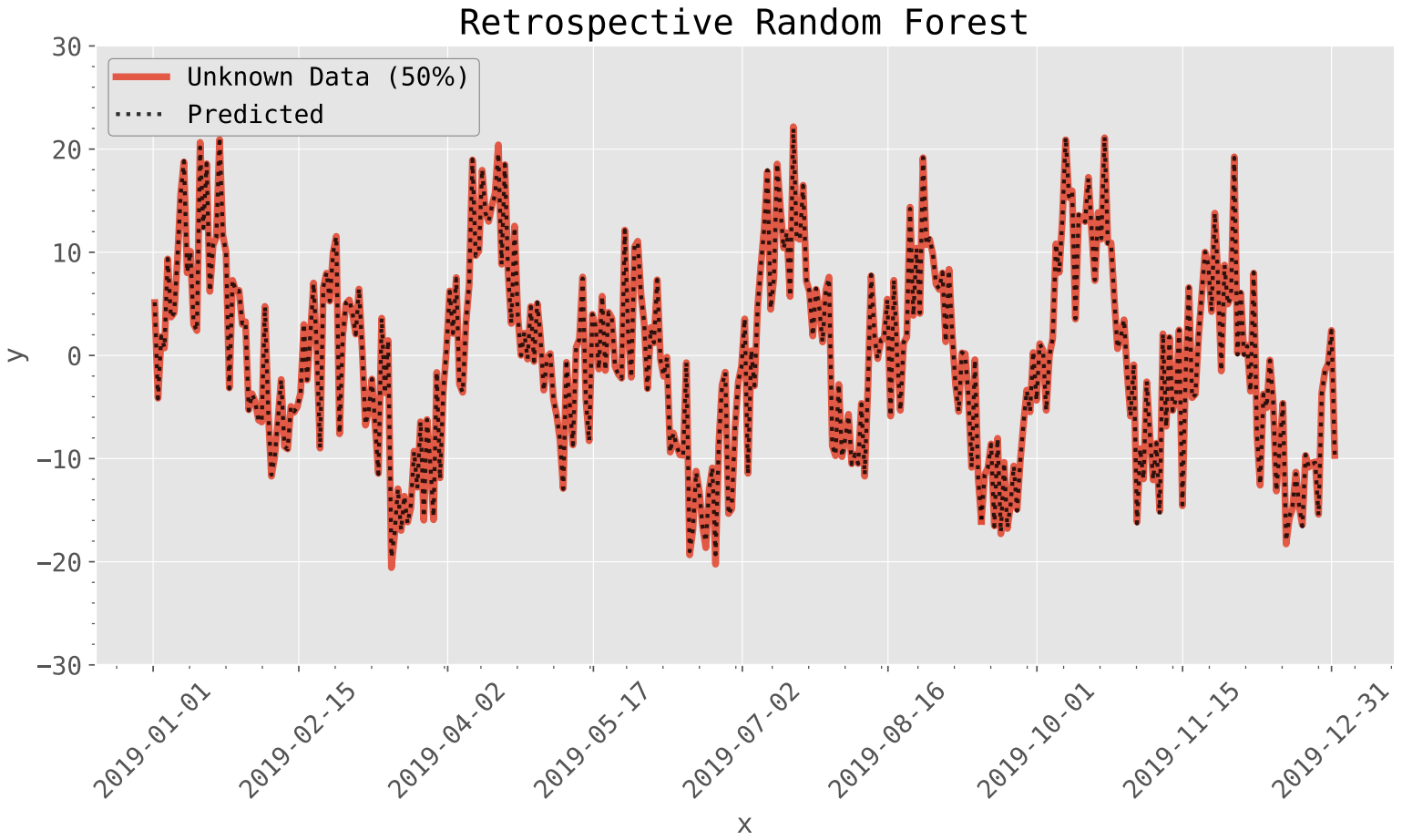


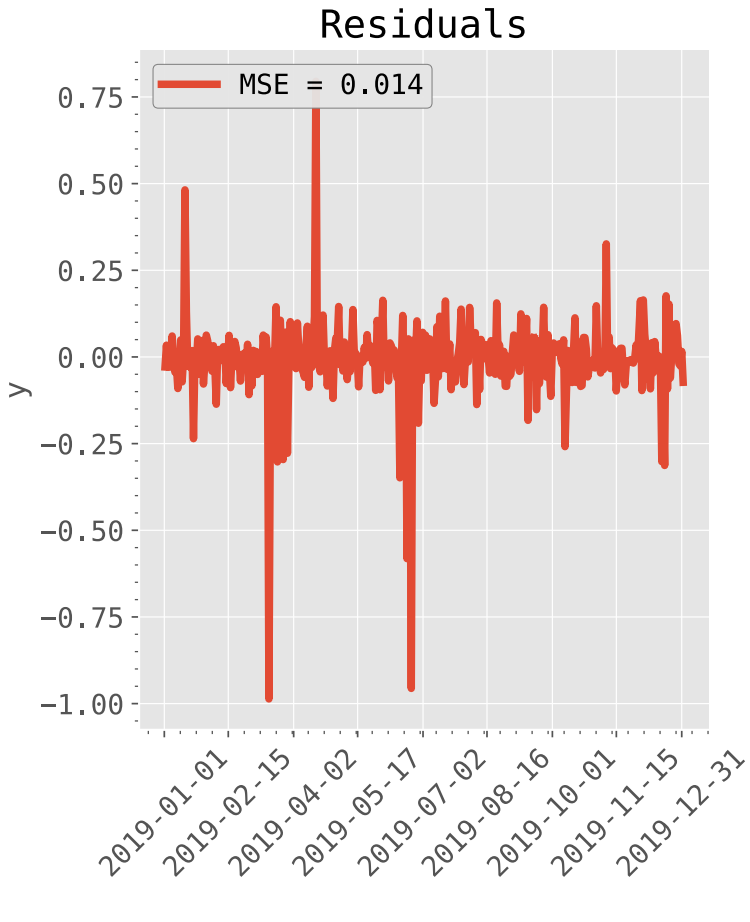
Data Creation

$$Error: \ \varepsilon = \mathcal{N}(\mu = 0, \sigma = 5)y$$
$$= 5sin\left(x\frac{8}{365}\pi\right) + 10sin\left(x\frac{16}{365}\pi\right) + \varepsilon$$

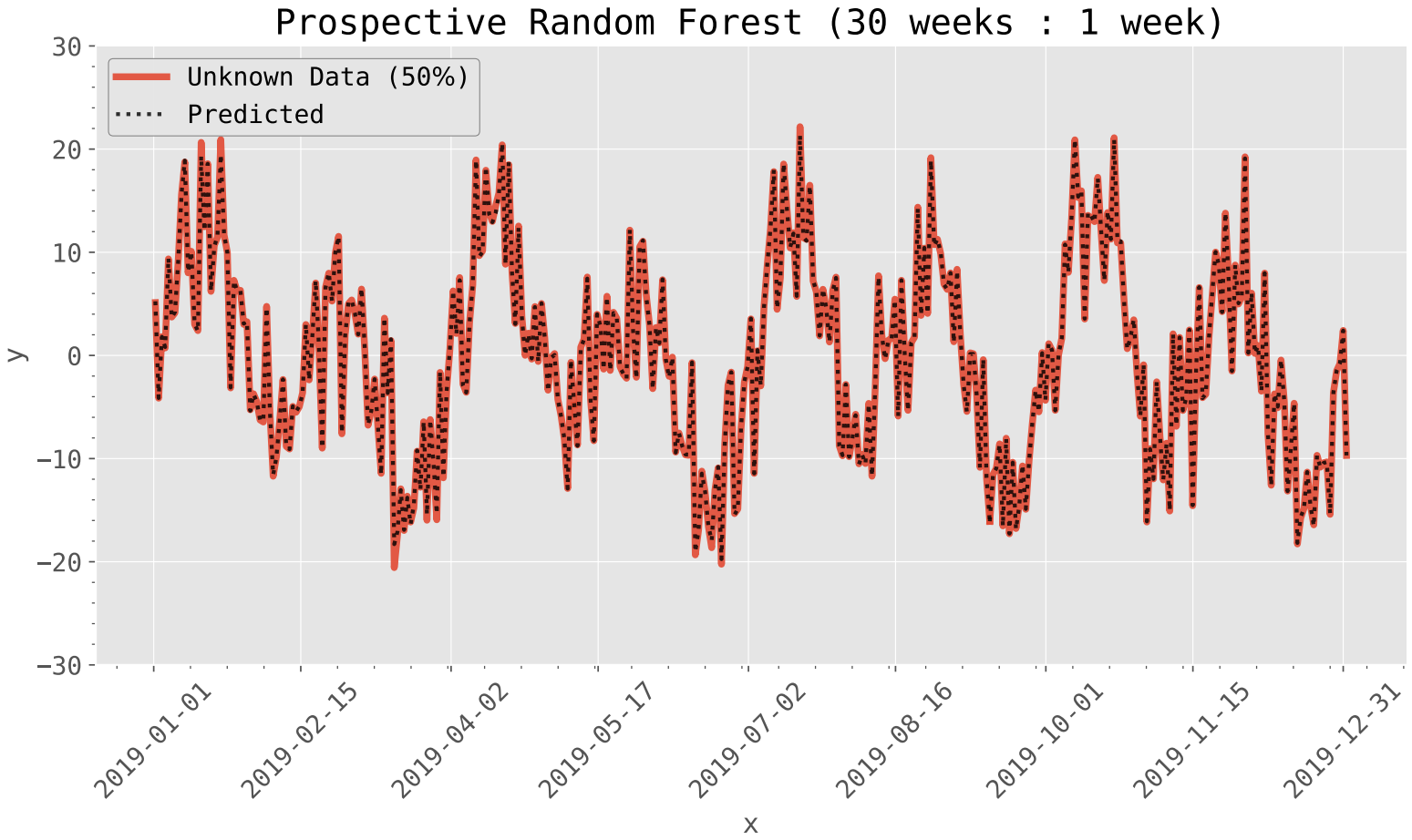


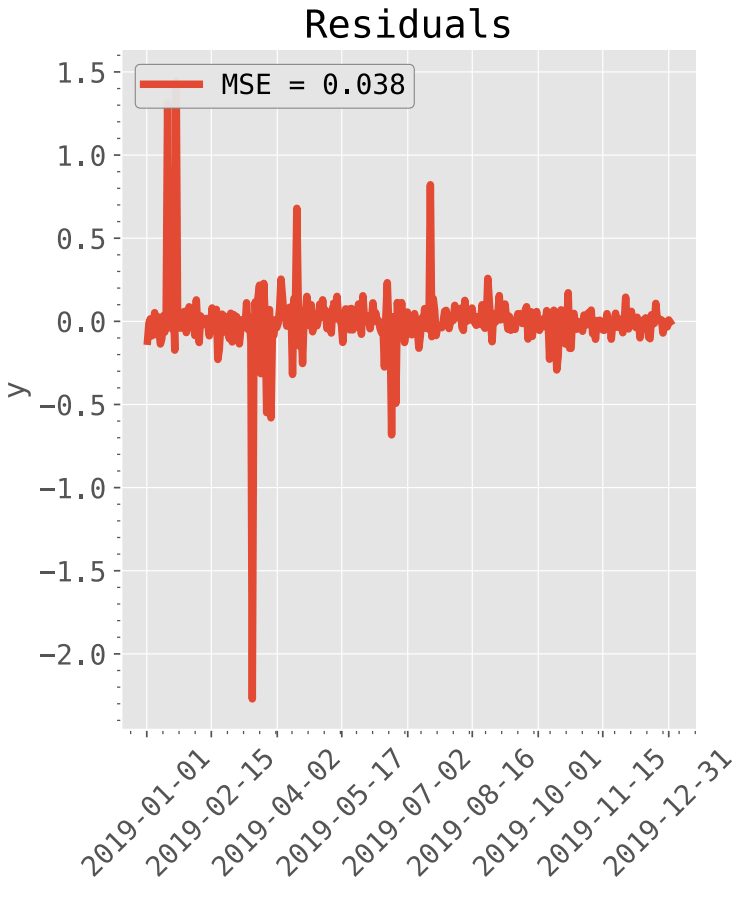
Retrospective



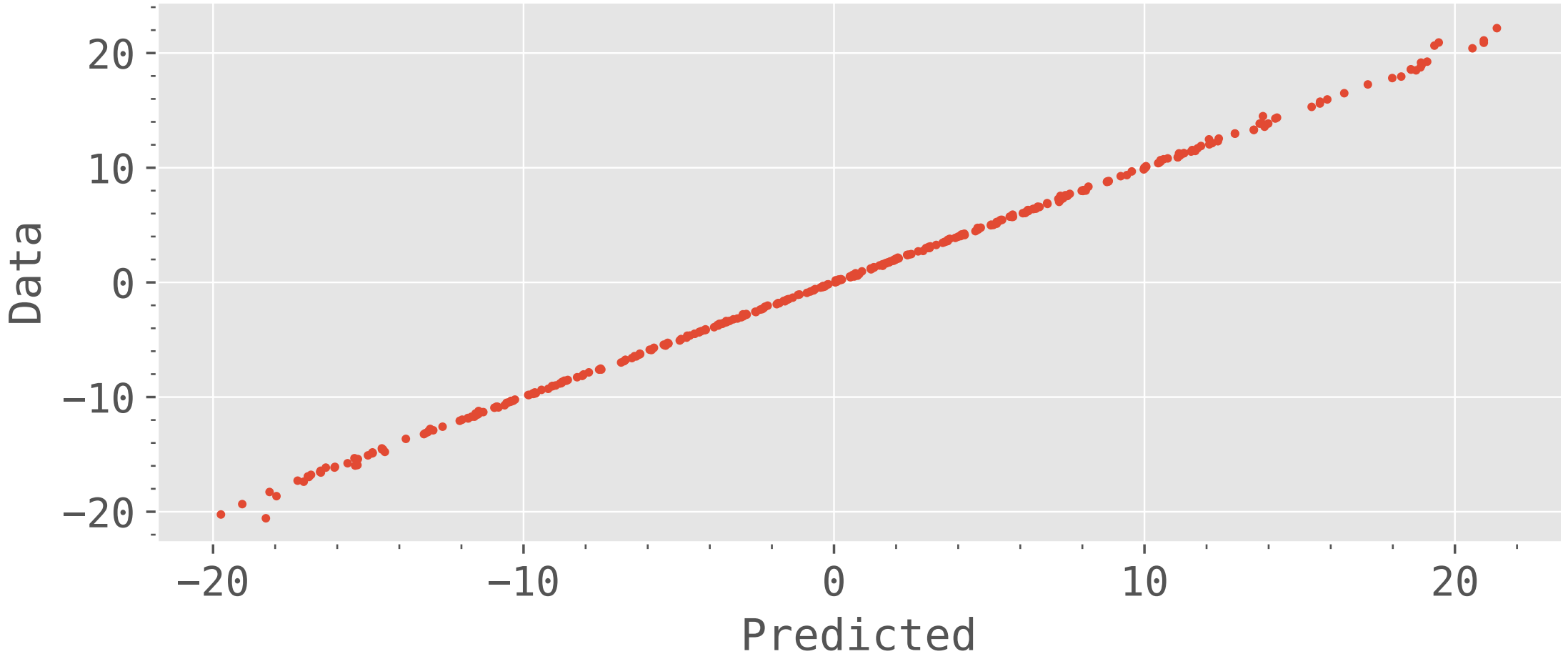


Prospective

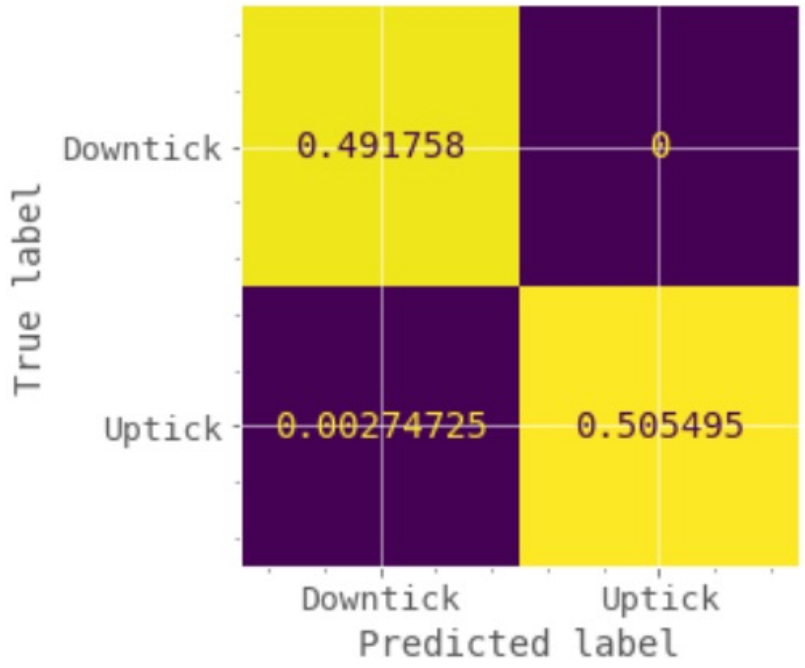




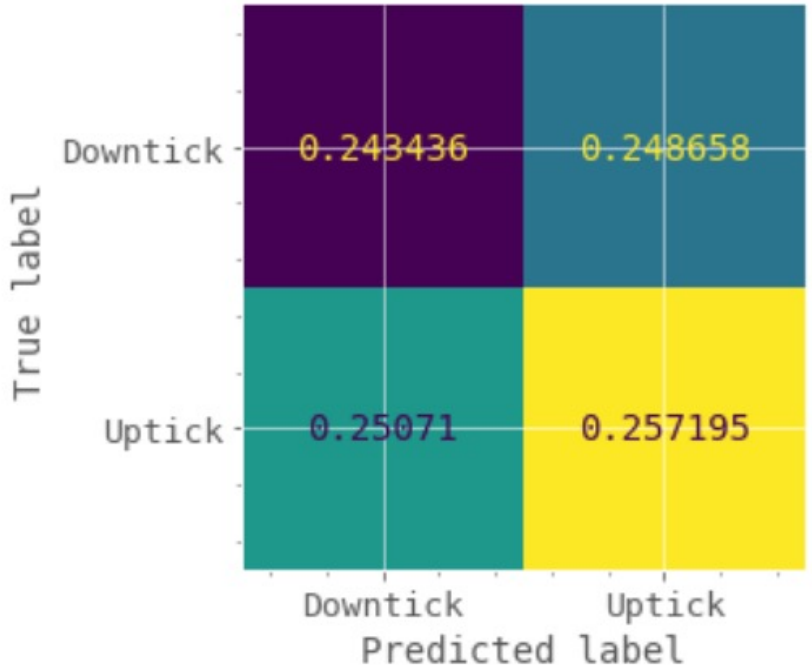
Data versus Predicted



Confusion Matrices



Proportion of TRUE downticks = 0.49176
Proportion of PREDICTED downticks = 0.49451



Proportion of TRUE downticks = 0.49209
Proportion of PREDICTED downticks = 0.49415