

Methods section of final report

Nathen Byford

The goal of this project is to identify anomalous data points in the twitter tag data utilizing data driven methods, specifically isolation trees.

1 Methods

In this report we will compare the performance of four classes of anomaly detection algorithms. These classes are: linear regression, seasonal decomposition of time series by Loess, neural network, and isolation forest. Some of these models are post hoc and intended to be performed after the data is collected, others can be trained, and then the new data can be input into the model for decision-making. All models will be trained on the first half of the time series if necessary and all test measures are from the second half of the time series.

1.1 Linear regression model

Anomalies are similar to outliers, points that are not expected and further from the other data points. It's possible to use a common method of outlier detection with the leverage calculations for a simple linear regression. For the time series in this study the x variable is date/time and the y variable is the number of twitter tags. Then calculating the leverage measures of cooks distance, covariance ratio, and DF beta if a point has high leverage for any of these measures it's considered an anomaly.

In figure 1 the model is shown, the blue line is the simple linear regression, the green points are identified outliers using this method, and the red x is a true anomaly from the data set.

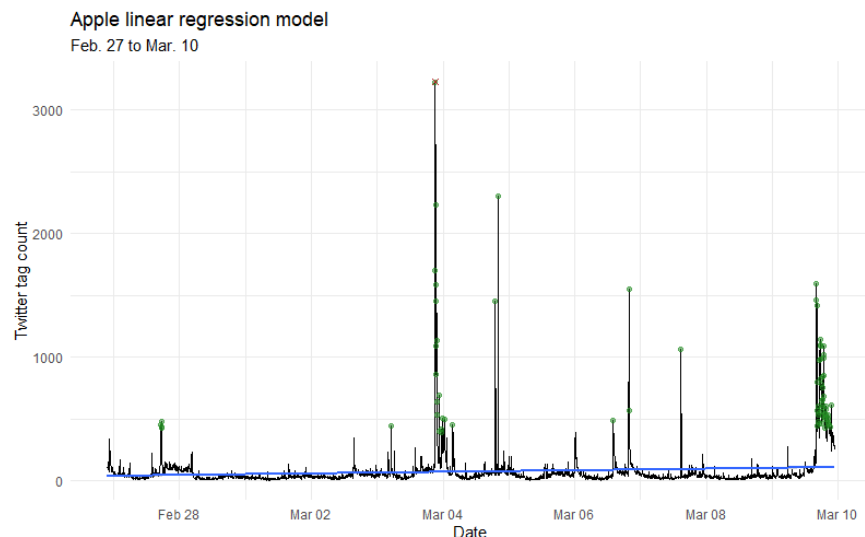


Figure 1: Example linear model

Thinking about this model, the linear regression slope and intercept are not of interest. This is counterintuitive from the typical construction of a linear regression model, here the primary interest is what

data points are leverage points. So there will not be any mention of the coefficients of the model here, only mention of the anomalies identified and the AUC value.

1.2 Seasonal decomposition of time series by Loess (STL)

A common method of anomaly detection with time series data is to utilize the seasonal decomposition of time series by Loess or STL. The STL is as its name suggests a decomposition of the time series by its trend component, seasonal component, and some residual remainder. This can be seen in figure 2. For anomaly detection we can place a bound on the remainder portion of the STL and determine that any remainder outside these bounds is an anomaly. This would be an indication that the values observed are further than expected from the seasonal plus trend components of the time series.

Similar to the linear regression model, the STL model is post hoc and needs the full data, as the classification comes as a byproduct of the model fit to the complete data. The coefficients of the seasonal and trend portions are not of interest. The package that contains the function in R for STL anomaly detection is `timetk`.

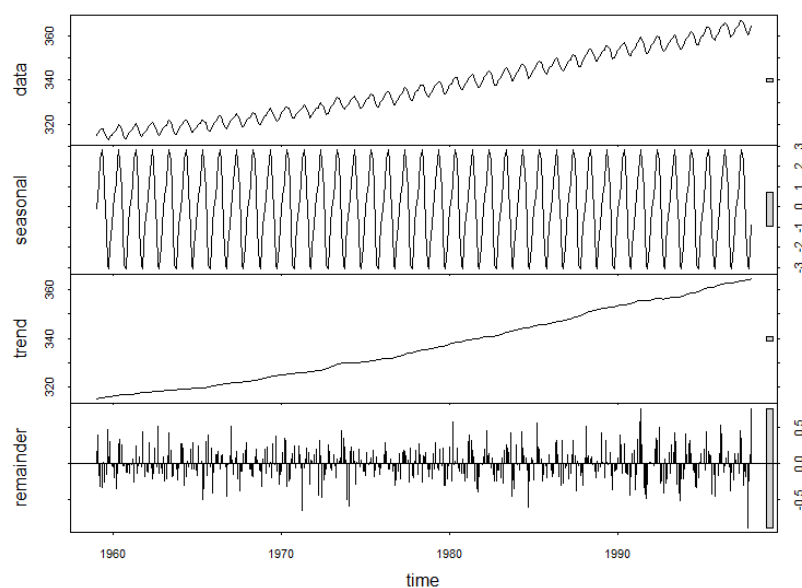


Figure 2: Example plot of STL

1.3 Stochastic gradient descent

1.4 Isolation forest