# Anomaly Detection of Twitter Tags Using Isolation Trees

Nathen Byford

## Contents

**1 Introduction** **2**
1.1 Data . . . . . 2

**2 Methods** **2**
2.1 Linear regression model . . . . . 3
2.2 Seasonal decomposition of time series by Loess (STL) . . . . . 3
2.3 Artificial neural network . . . . . 3
2.4 Isolation forest . . . . . 4

**3 Results** **4**

**4 Conclusion** **4**

# 1 Introduction

In data analysis one of the most prevalent issues among all areas of data collection is anomaly detection. Anomalies can cause numerous problems in a statistical analysis and effect the results in unwanted ways. Many modern data pipelines rely on data streamed at high rates to make near instantaneous decisions for business needs.

There are many possible causes for anomalies in data. Possible causes are: faulty sensors, poor data quality, external actors, and others. Identifying anomalies is the first step to identifying the cause of the anomaly. Knowing the cause of the anomaly can help determine if the anomalies should be included in the analysis of the data and the decision-making process.

## 1.1 Data

The data comes from the Numenta Anomaly Benchmark data set, specifically the twitter tag part of the data set. This part of the data has 158,631 observations of twitter tag counts every 5 minutes from Feb. 26, 2015 to Apr. 23, 2015 at 5 minute increments. These observations are split between 10 companies, Apple, Amazon, Salesforce, CVS, Facebook/Meta, Google, IBM, Coca-cola, Phizer, and UPS. With each observation there is a logical assigned, true or false if it is an anomaly.

In figure 1 we can see the time series of each company with the identified anomalies overlaid as Xs. Looking at the time series' in figure 1 it is clear that there is an imbalance problem in the data, there are far more non-anomalous points compared to the number of anomalous points. This leads to the possibility of the models performing extremely well by determining that all points are normal. In anomaly detection there is often a high false positive rate due to the nature of anomaly detection being related to outlier detection. Because of the imbalance of the data the measure of model performance will be the AUC and the true positive rate, because it's important that the model does identify the anomalies since there are so few.

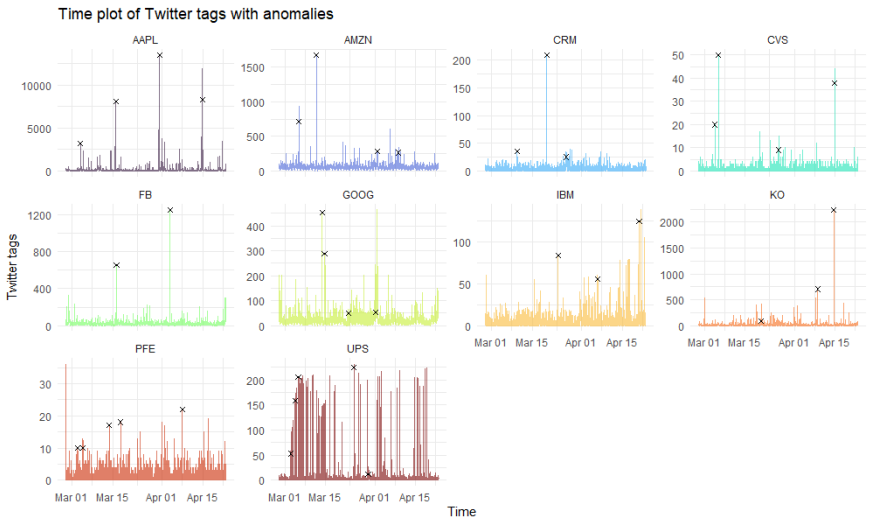


Figure 1: Plot of each company time series

# 2 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.

## 2.1 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 2 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.

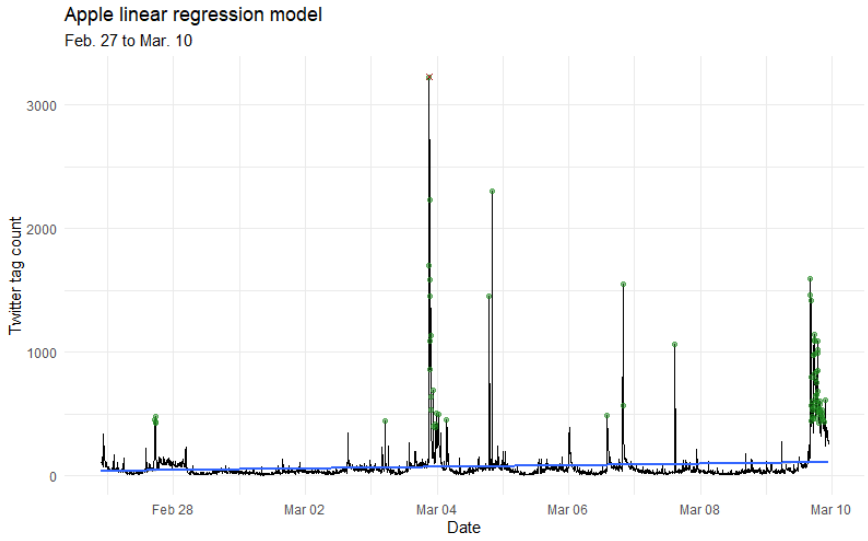


Figure 2: Example linear model

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.

## 2.2 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 3. For anomaly detection we can place a bond on the remainder portion of the STL and determine that any remainder outside the 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 if interested. The package that contains the function in R for STL anomaly detection is `timetk`.

## 2.3 Artificial neural network

Neural networks try to replicate how the human brain thinks when classifying data. Neural networks can find connections that other methods don't see, but they are also more of a black box method where we don't learn much about what is important. Using the package `ANN2` in R for artificial neural networks for anomaly detection we can make a classification network based on stochastic gradient descent using a log loss function.

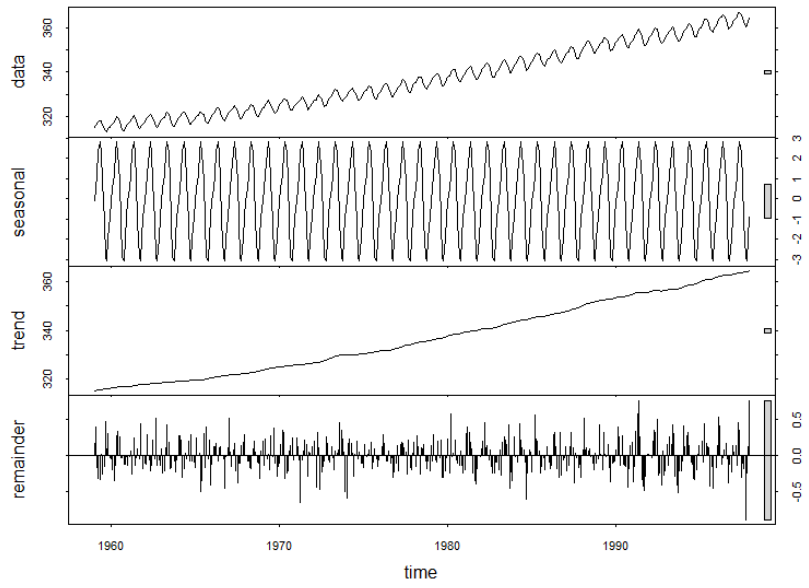


Figure 3: Example plot of STL

Using a neural network is good for prediction and classification, so that is why it can be useful for anomaly detection. Here the model will be trained in the `ANN2` package and utilizing a single hidden layer. One thing to note is that the time stamp is needed to be converted to a numeric as the function doesn't know how to use a date time for classification. This could take away the time structure of the data and cause some negative consequences. There are new methods being introduced for time series neural networks, but I was unable to get any to run.

### 2.4 Isolation forest

The last method is isolation forests, this method is similar to random forest and regression trees with a different goal and measure. Similar to trees we take the data and split it into partitions as seen in figure 4, the data will continue to be partitioned until all points are the only member of their own partition. The data are split randomly by variable and value, this is why an ensemble of these isolation trees is utilized. As each point gets its own partition we count the number of splits necessary to isolate the point, this is the isolation depth of the point for the tree. Using the isolation depth of each point for each tree in the ensemble, we take the average depth of each point and that is the "anomaly score". Placing a threshold on the anomaly score we are able to classify the anomaly status of a point.

There are other metrics to determine the anomaly score of the partition, there is the average of the isolation depth and another option is a ratio of the variables on each side of the split. This second measure is called the density anomaly score. The density isolation forest tests to preform better for categorical variables. Both methods will be tested for comparison with the twitter tags data.

## 3 Results

## 4 Conclusion

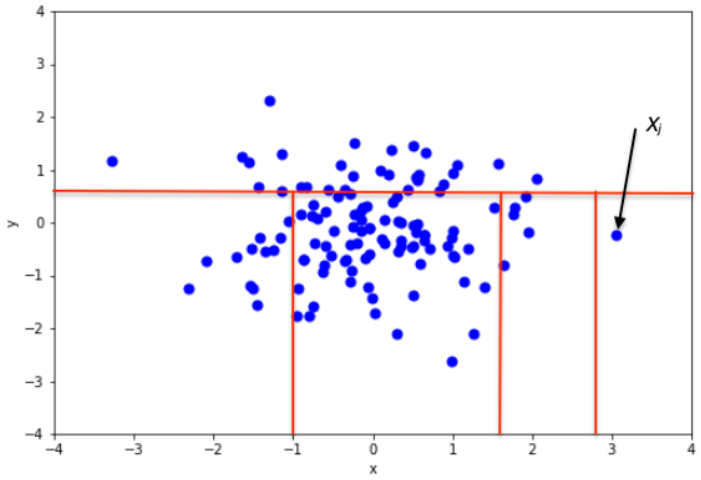


Figure 4: Example of data partition for isolation forest

Table 1: AUC results of each model

| Model | AAPL | AMZN | CRM | CVS | FB | GOOG | IBM | KO | PFE | UPS |
|---|---|---|---|---|---|---|---|---|---|---|
| lm | 0.9962 | 0.9888 | 0.9814 | **0.9899** | 0.9945 | 0.4851 | **0.9899** | 0.9968 | 0.9790 | 0.4939 |
| stl | 0.9554 | **0.9920** | **0.9892** | 0.8667 | 0.9868 | 0.4839 | 0.9869 | 0.9843 | **0.9869** | 0.4817 |
| SGD | **0.9994** | 0.5000 | 0.5000 | 0.5000 | **0.9999** | 0.5000 | 0.5000 | 0.5000 | 0.5000 | **0.5000** |
| IF | 0.9967 | 0.5000 | 0.5000 | 0.7497 | 0.5000 | 0.4999 | 0.5000 | **0.9991** | 0.5000 | 0.4981 |
| Density IF | 0.9665 | 0.9859 | 0.9713 | 0.9611 | 0.9905 | **0.9736** | 0.9746 | 0.9912 | 0.9758 | 0.4888 |

Table 2: True positive rates

| name | AAPL | AMZN | CRM | CVS | FB | GOOG | IBM | KO | PFE | UPS |
|---|---|---|---|---|---|---|---|---|---|---|
| lm | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |
| stl | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |
| SGD | 1 | 0 | 0 | 0.5 | 1 | 0 | 0 | 0.5 | 0 | 0 |
| Isolation Forest | 1 | 0 | 0 | 0.5 | 0 | 0 | 0 | 1 | 0 | 0 |
| Density Isolation Forest | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |