Anomaly Detection in Time Series Data with Data-Driven Approaches

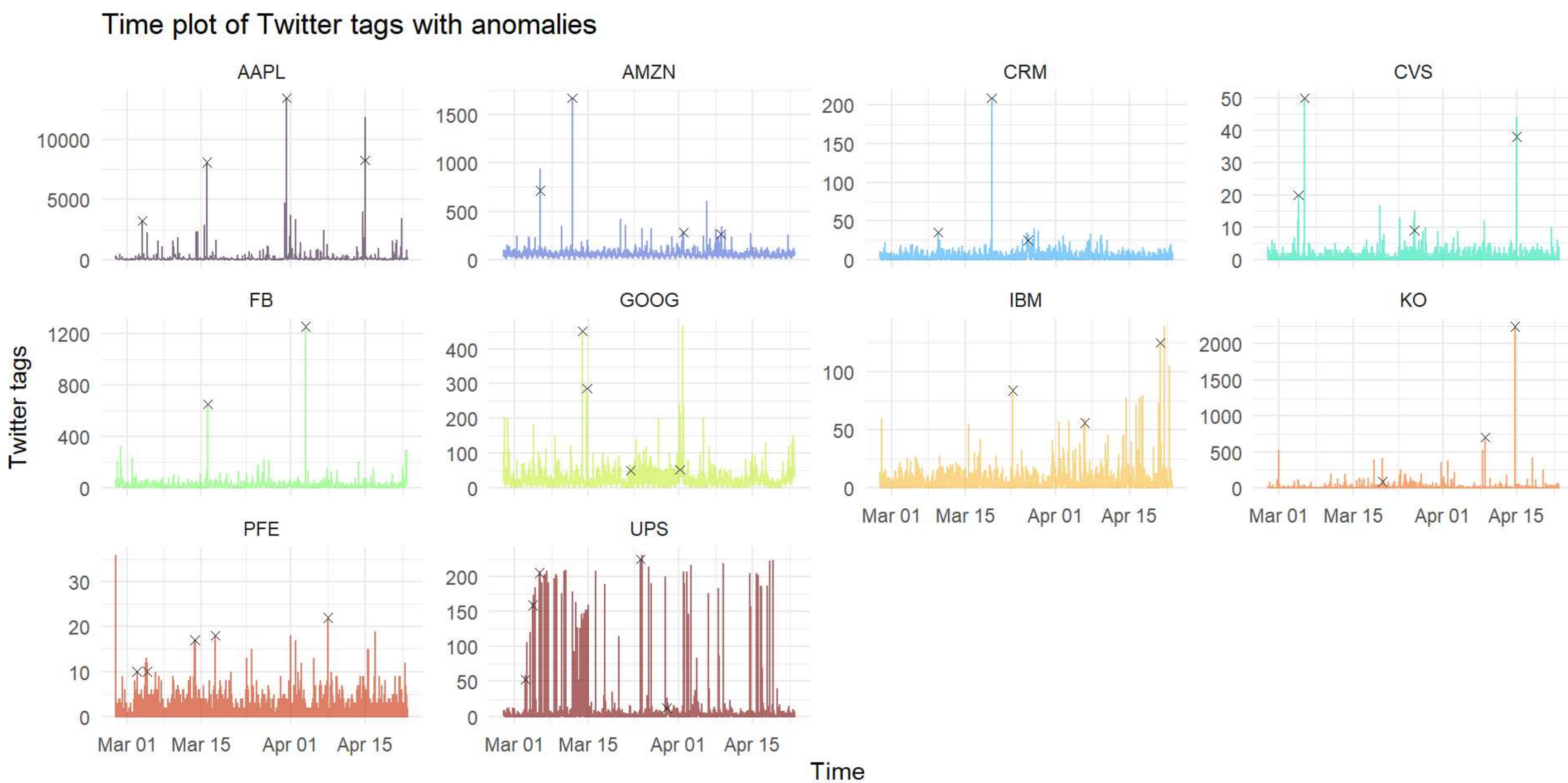
Nathen Byford

Project Motivation

Project Review

- Anomalies are a common problem in statistics analysis
- Anomalies can be caused by different factors
  - Faulty sensors
  - Bad data
  - Some outside actor
- **Goal:** Use data-driven approaches to determine what data points are anomalous in the Twitter anomaly data set.

Data



Time plot of Twitter tags with anomalies
AAPL
AMZN
CRM
CVS
FB
GOOG
IBM
KO
PFE
UPS
Twitter tags
Time
Mar 01
Mar 15
Apr 01
Apr 15

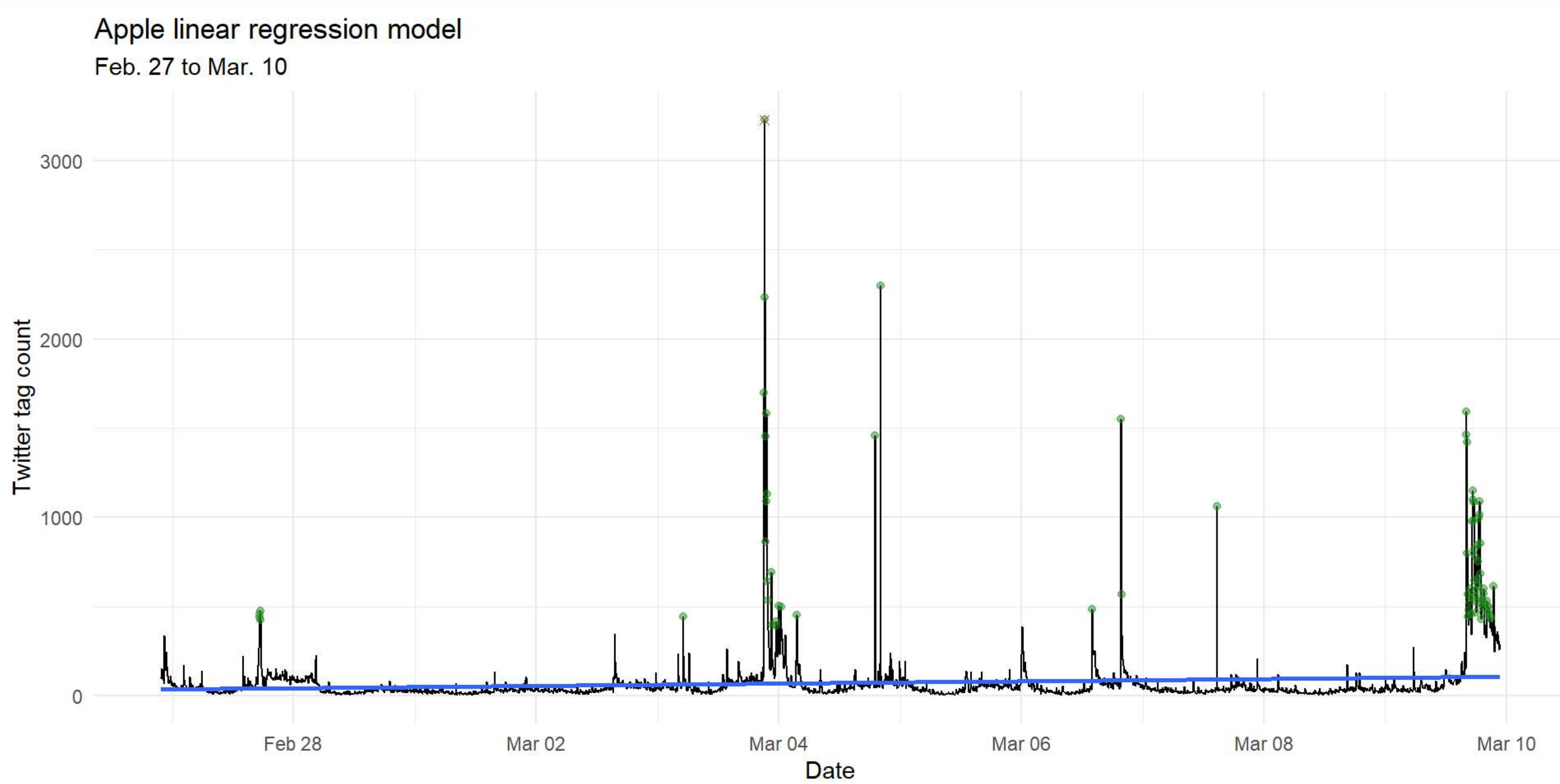
Methods

Methods

- Comparing 4 methods of detection:
  - Linear regression (leverage points)
  - Seasonal decomposition of time series by Loess (STL)
  - Stochastic gradient descent (Neural Network)
  - Isolation Forests
- For cross validation the data is split in half with the first half as training.

Linear regression

- Post hoc method
- Utilize linear regression leverage measures to identify anomalies/levereage points.
  - Cook’s Distance
  - Covariance ratios
  - DF beta
- Declare point as anomalous if it has high leverage.



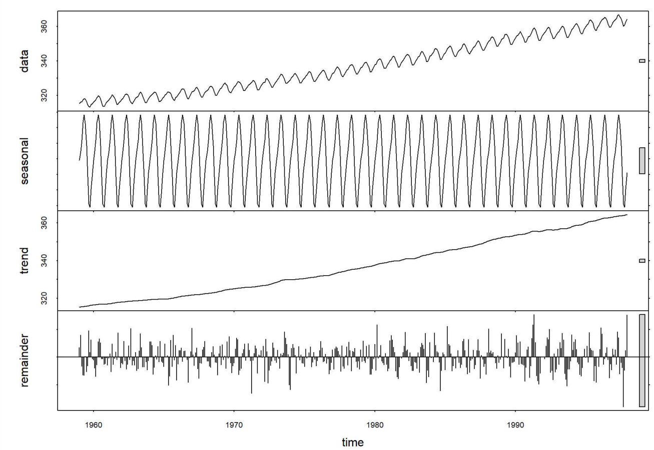
Blue line is linear regression line fit to the data.

Green dots are points that are identified as anomalies/high-leverage points.

Red ‘X’ is the true anomaly value.

Seasonal decomposition of time series (STL)

- Post hoc method.
- Common technique in time series analysis to identify trends in the data.
- Utilize white noise/residuals to determine how far observations is from trend.



- Compare realization to trend + seasonal model using residuals.
- Using the residuals we can set a threshold and determine if values are anomalies.

Neural Network

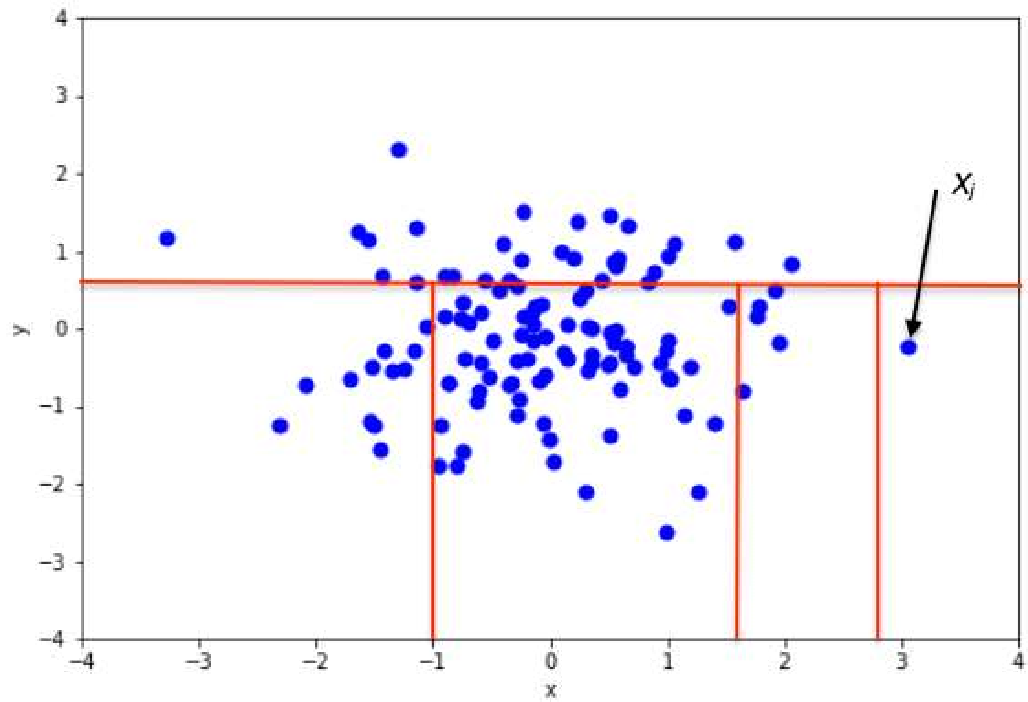
- Training and testing cross validation
- Using a RNN for classification of the training data
  - One hidden layer (found to have best results from tuning)
- Uses nodes to determine outcome based on other nodes
- Not optimized for time series data
- Would be better to use an autoencoder[1] method

1. Like a GAN model, specifically TimeGAN or anoGAN.

Isolation Forest

- Split training and testing
- Similar idea to random forest and classification trees
  - No longer interested in classifying at each split
  - Split until all points are their own partition
  - Interested in the amount of splits it took to isolate each point
  - anomalies will/should require less splits
  - Anomaly score is the average number of splits to isolate a point across all trees

Isolation Forest visual Aid



- Point x_j has isolation depth of 4 in this case
- Repeat process for ensemble of trees and average depth

Isolation Forest Types

- Isolation Forest
  - Basic isolation forest described on previous slide.
- Density Isolation Forest
  - Changes the scoring method to density based on the ratio of the characteristics on each side of split.
  - Tends to be better for categorical variables.

Results

Results

- We are not concerned with the accuracy or misclassification rate in this case
  - Due to the imbalanced nature of the data
- Using area under the curve (AUC) gives a better perspective of what models are better
  - AUC of 0.5000 is the same as if the model is guessing every value is false.

AUC for models

name	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	0.7499	**0.9999**	0.5000	0.5000	0.7500	0.5000	**0.5000**
Isolation Forest	0.9967	0.5000	0.5000	0.7497	0.5000	0.4999	0.5000	**0.9991**	0.5000	0.4981
Density Isolation Forest	0.9665	0.9859	0.9713	0.9611	0.9905	**0.9736**	0.9746	0.9912	0.9758	0.4888

No model is overall the "best", but overall the density isolation forest performs very well and is only worse than guessing on UPS while all other models

Results True Positives

- Another critical measure in this case is true positive rate
- In each testing set there is roughly 2 anomalies

True positive rate

name	AAPL	AMZN	CRM	CVS	FB	GOOG	IBM	KO	PFE	UPS
lm	1	1	1	1.0	1	0	1	1.0	1	0
stl	1	1	1	1.0	1	0	1	1.0	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	0
Density Isolation Forest	1	1	1	1.0	1	1	1	1.0	1	0

- No model does any well on UPS

Density isolation forest is the only model that captures all anomalies aside from the UPS data set. Both the linear model method and STL also perform

Conclusion

- Detecting anomalies in time series can be difficult
  - Here there is only one variable total
  - Possible improvements could come from incorporating other time series as well
- The STL model provides acceptable results
- The linear regression model does surprisingly good even if it is basic
- The best model with the data is the density isolation forest

Thank you!