Report

Task 1

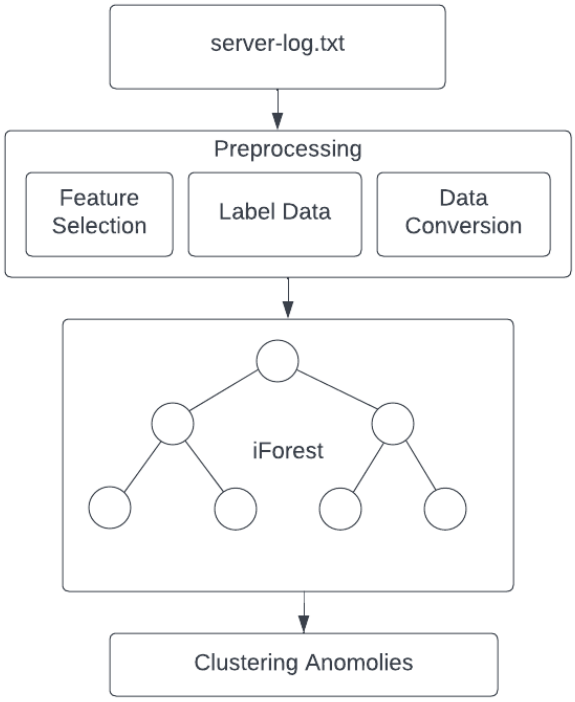
Preprocessing

Feature	Action	Justification
No	Removed	Provided no value
Start-Date	Removed	Redundant - used to create timestamp
Start-Time	Removed	Redundant - used to create timestamp
Duration	Removed	Redundant - used to create Duration Seconds
Service	encoded	Used label encoding to create numeric features while retaining label value
Source-IP	encoded	Used label encoding to create numeric features while retaining label value
Destination-IP	encoded	Used label encoding to create numeric features while retaining label value
Source-Port	Modified	Replaced “-” with -1 value as some protocols do not require a port and therefore an impossible value was assigned
Destination-Port	Modified	Replaced “-” with -1 value as some protocols do not require a port and therefore an impossible value was assigned
Timestamp	Created	Converted date and time to seconds since epoch as the retain ordinal value as an integer for training
Duration Seconds	Created	Converted duration to seconds since epoch as the retain ordinal value as an integer for training

Note: I decided not to use one-hot encoding for any of the features as I was concerned it would introduce too many redundant features, diminishing results.

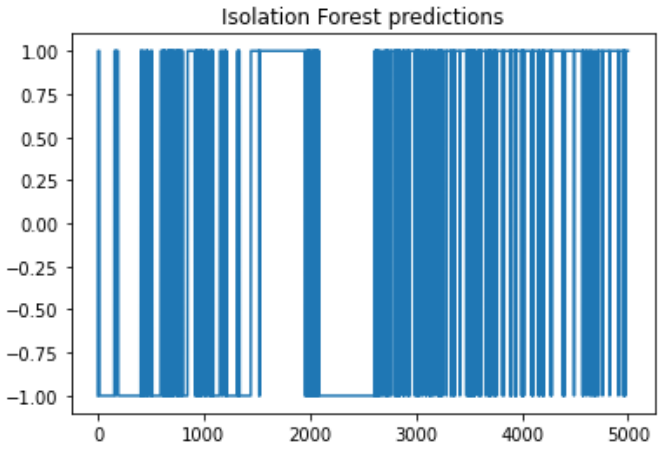
iForest

iForest was used as it is commonly used in anomaly detection and there is a significant amount of literature that exists regarding this application.



Part 1

The first attack occurred on 06/04/2014 at 00:26:1. iForest was used to determine outliers. The outliers were then displayed in a graph to understand what trends exist.



The graph highlights that there are clear periods of significant anomaly rates. Using this information a function was written to capture these clusters of log events. Each cluster of events must have a minimum of 200 events to count.

Cluster 1

Cluster 1 starts on 06/04/2014 at 00:26:12 and ends at 02:27:26. More than 200 requests were sent in this period with http being the most common service.

Cluster 2

Cluster 2 starts on 06/03/2014 at 08:12:16 and ends at 08:12:52. 525 requests were sent during this period, across a range of services.

Attack Type

The attack is likely to have been a DDoS attack given the significant number of requests in a short period of time.

Task 2

Explanation

Preprocessing

The decision was made to validate utf-8 encoding of the files as some files contained issues that prevented TF-IDF Vectorising. Any not utf-8 encoded files are skipped in the process. This affected 2 files:
pintersts-facebook-cdomaincom3-useless.js
pintersts-facebook-cdomaincom2-useless.js

There was no other preprocessing required.

Task 2.1

The TF-IDF Vectorizer was used with the default values. This could be a potential area for refinement. There is a significant similarity between the Tracking and Functional javascript files with the present configuration.

Task 2.2

The OCSVM was tuned as described in the following task. The features extracted from the TF-IDF were used to train the OCSVM classifier. The TF-IDF features were fitted to the tracking JS file and all JS files had features extracted for classification.

Task 2.3

The OCSVM parameters were: Gamma: 0.2, Nu: 0.8.
These values were chosen after running an algorithm that used every combination of gamma 0.2 to 1.0 at an increment of 0.2 and Nu 0.2 to 1.0 at an increment of 0.2.

The result of this was an accuracy of 63.48% accuracy (true positives).

The graph below shows there was no change in the accuracy when running the best params function in initial testing. However, since producing the graph there has been a change in the best parameters. The python file will automatically output the best parameters calculated for the selected dataset.