Un-Supervised Case-Study

Adithya Vikram Raj Garoju

5.18.2022

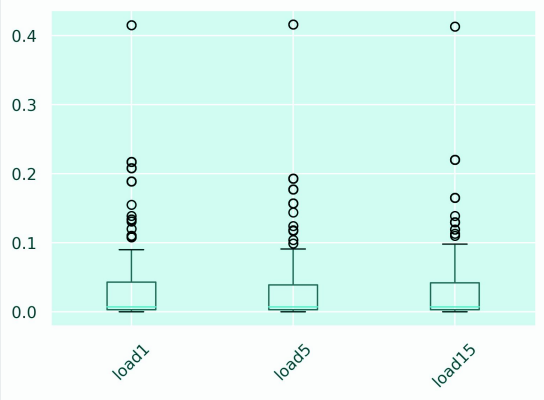
Objective

To be able to detect when a host is misbehaving in comparison to the rest of the hosts at any point in time, regardless of the time of day/day of week/week of month and so on.

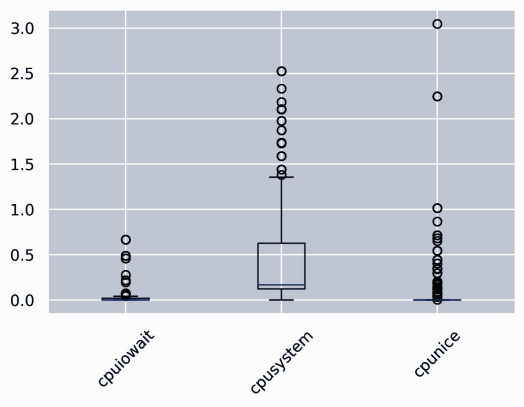
Pre-Processing Steps

1. ELK logs are transformed and reorganized accordingly to accomodate more data comparatively in lesser rows and columns.
    * Segregated data Metric wise
    * Merged them based on hostname in spite of timestamp
    * The actual metric component names have been fetched from description csv and re-mapped to corresponding columns.
2. Selected independent features to contribute to the model/
3. Excluded Highly Correlated Values from data.
4. Created Compressed/ Reduced Training set for Visualization purposes.

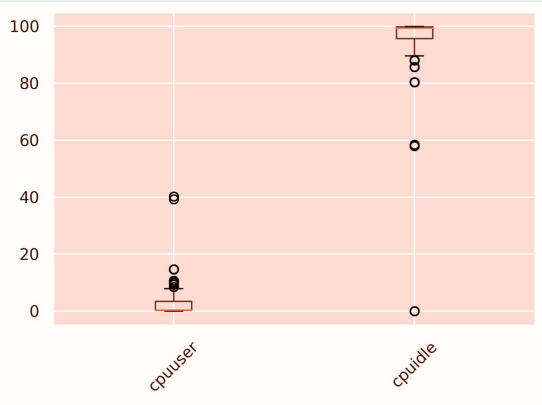
Data Ranges



The load components correlate really well and the ranges are equal for all the three.



When cpu is waiting, the cpu nice time is allocating resources to other priority processes.

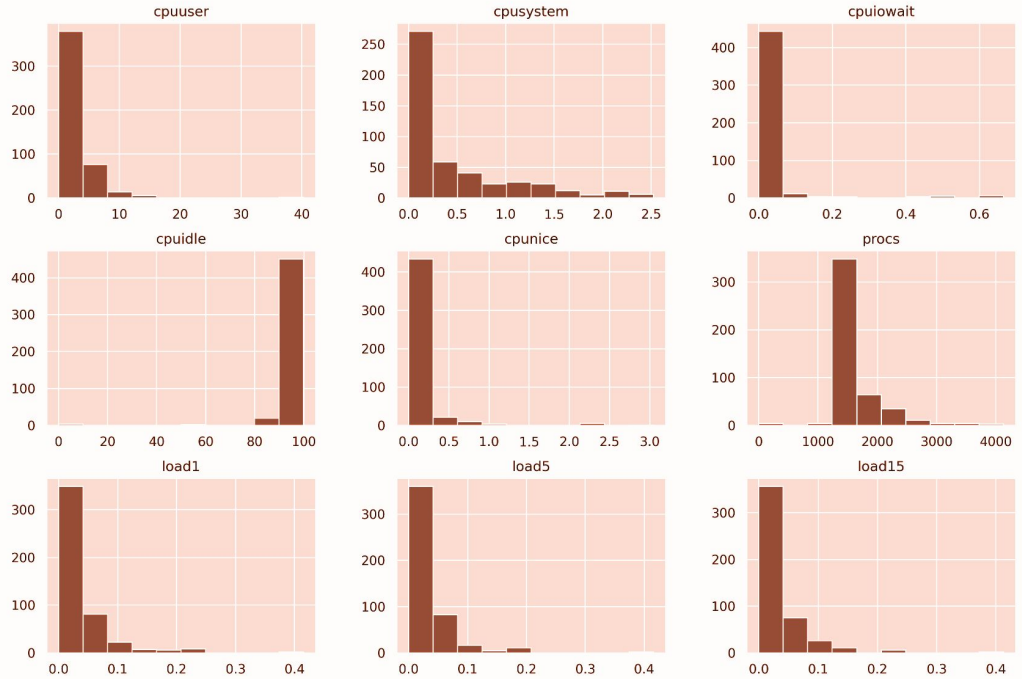


No of CPU-users are inversely proportional to CPU Idle time

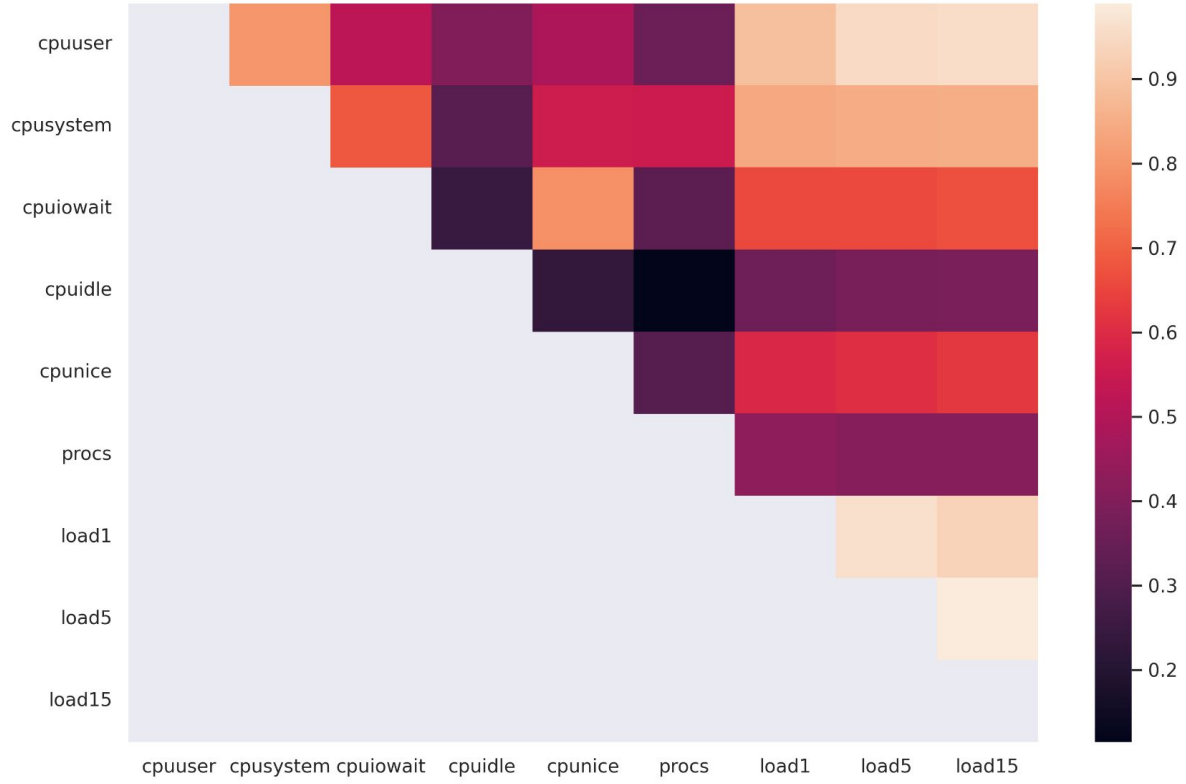
The peak number of users observed is of 40

Can imply that cpuidle is 0 when its loaded with users

Histograms


cpuuser
cpusystem
cpuiowait
cpuidle
cpunice
procs
load1
load5
load15

Correlation

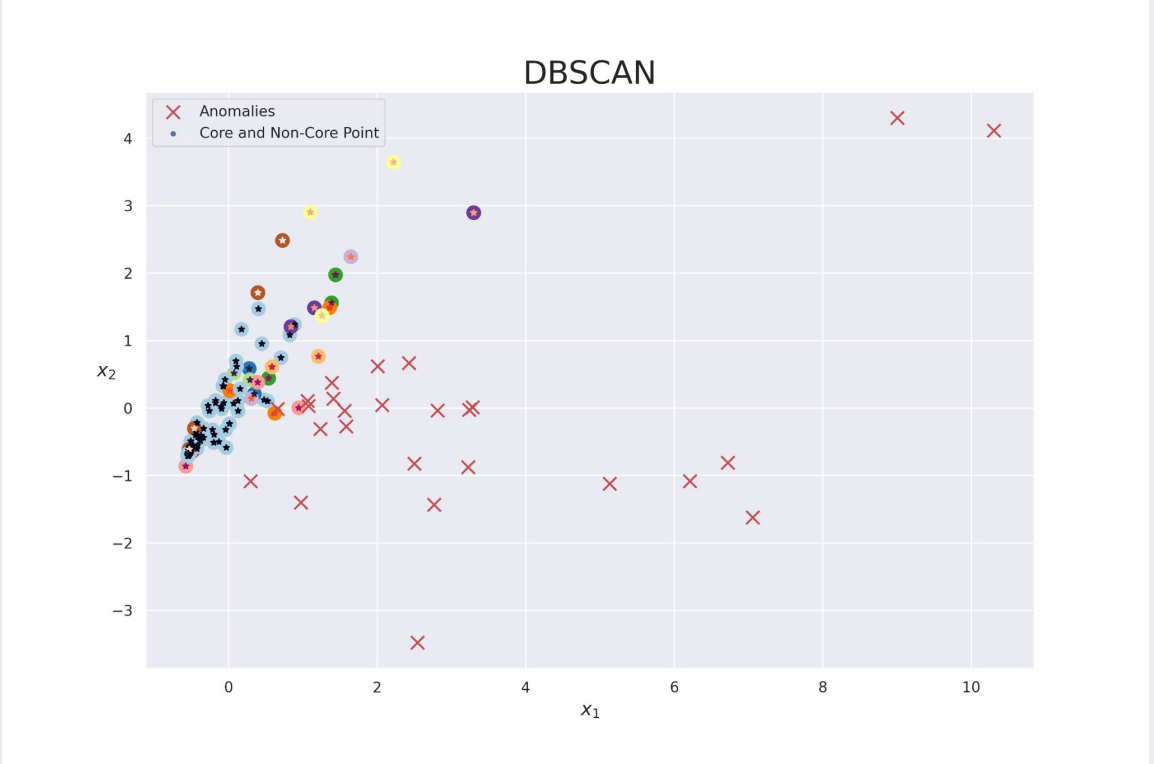


DBSCAN

- Started out with random eps and min_samples.

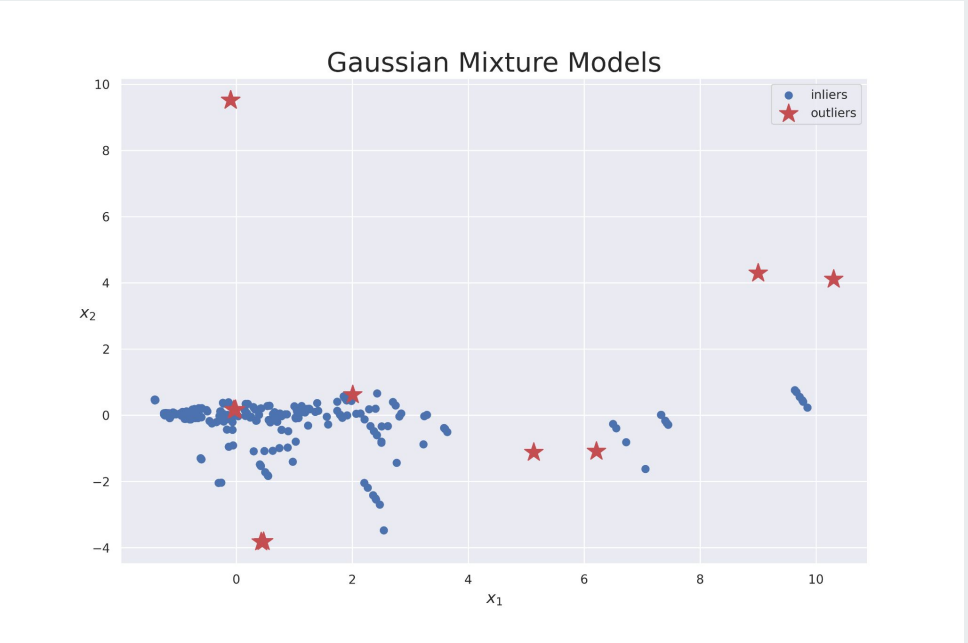
- The Figure on right is of setting {eps : 0.5 , min_samples=2}

- Due to variance in continuous distribution of the data, majority of the instances are left out of cluster



Gaussian Mixture Models

- Chosen n_components based on AIC(Akaike Information Criterion) and BIC(Bayesian Information Criterion) trend. The elbow point has been recorded at 2
- The No of Iterations took to converge are 4
- The covariance type used is ‘tied’



Isolation Forest

- Chosen 5% contamination factor and picked the default 100 n_estimators (n_jobs=-1 to utilize all cpu_scores)

- Would fit our use-case properly since we can control the outlier fractions.


Tnk!
you