Enhanced Anomaly Detection

Scott Garcia and Iman Makaremi

Vijay Veggalam

Enhanced Anomaly Detection

Scott Garcia
MTS | T-Mobile

Iman Makaremi
Principal PM – ML & AI | Splunk

splunk> .conf19

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or plans of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results may differ materially. The forward-looking statements made in the this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, it may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements made herein.

In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only, and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionalities described or to include any such feature or functionality in a future release.

Machine Learning Advisory Program

splunk> .conf19

Splunk Machine Learning Advisory Program

1. Get help from the Splunk Data Scientists to solve your business use case with Machine Learning Toolkit
2. Complimentary support with your Enterprise or Cloud license
3. Early access to new Machine Learning features
4. Results in opportunity to tell your success story with Splunk
5. Contact mlprogram@splunk.com for more information

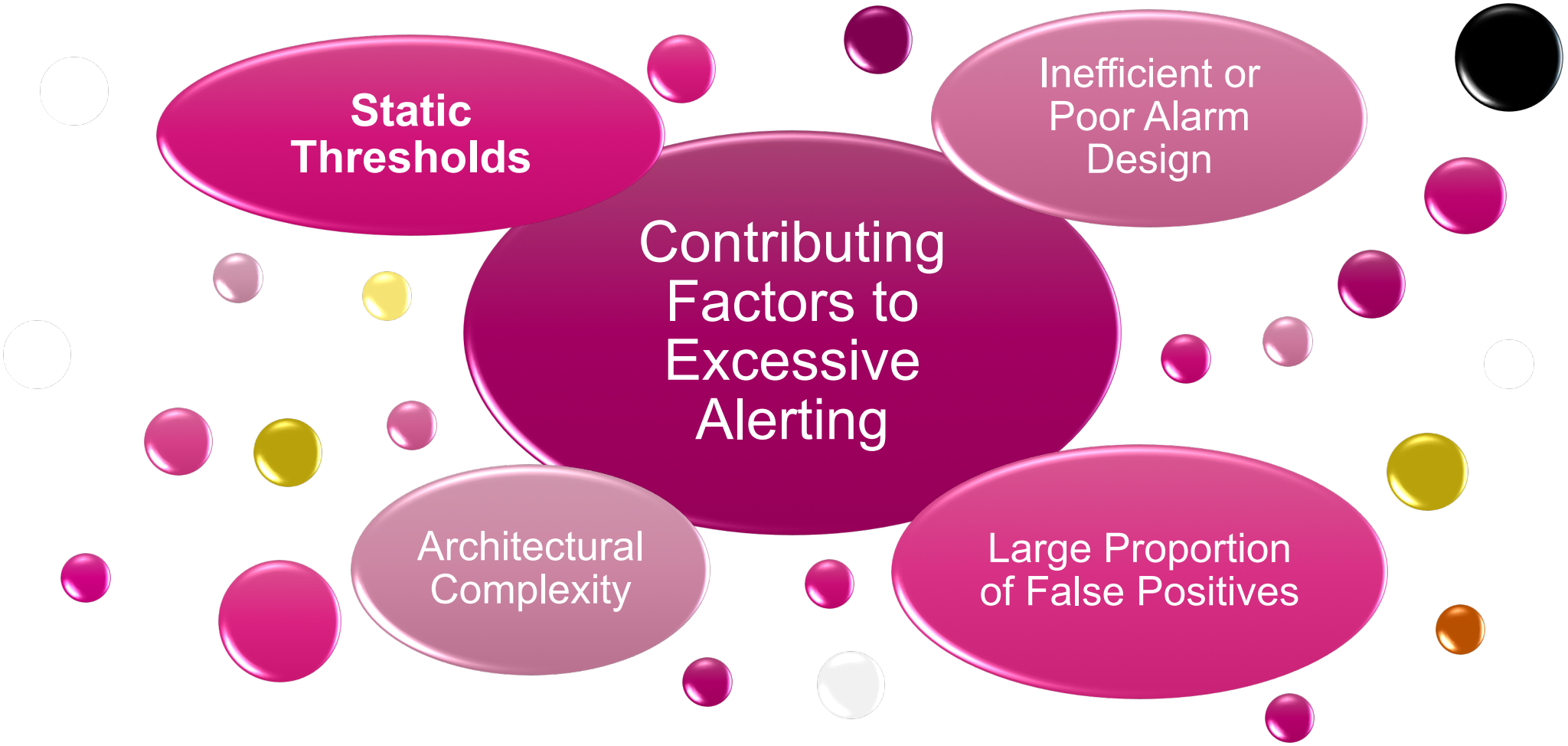
Machine Learning Use Case

Ineffective & Inefficient Alerting

splunk> .conf19

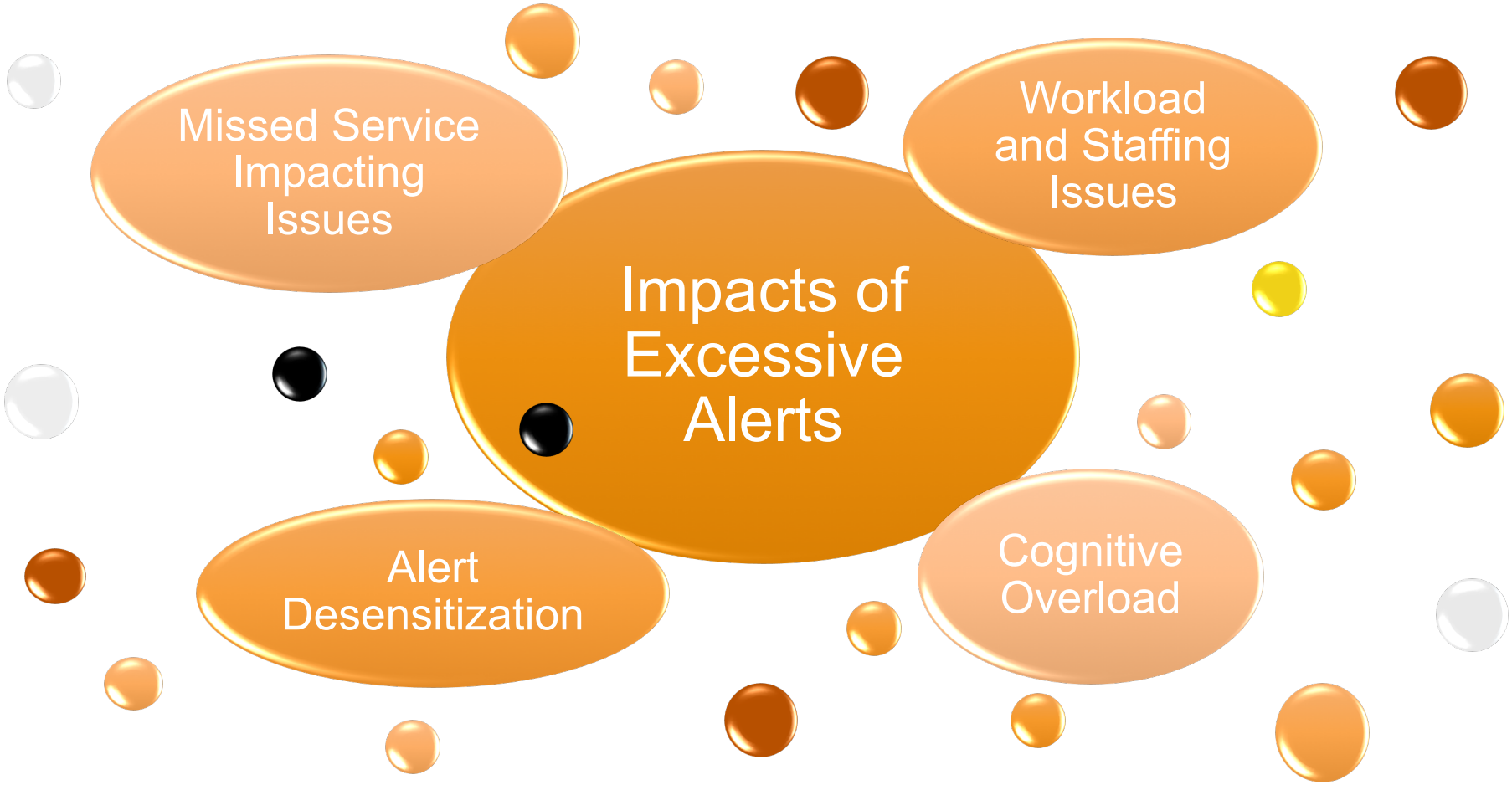
Alert Fatigue

Contributing Factors



Alert Fatigue

Impacts



Solution Strategy and Workflow

splunk> .conf19

Splunk ML Advisory Team Partners with T-Mobile

T-Mobile and Splunk collaborate on an ML use case

1. Preprocess the Data

- Centralize data in Splunk
- Preprocess and extract features

2. Generate 1st level Models

- Create Anomaly Detection models and send results to a summary index

3. Fit the 2nd Level Models

- Anomalies of Anomalies

4. Tune the Models

- Iterate over models to validate and tune

5. Splunk .conf2019

- T-Mobile and Splunk co-present results in October

Splunk ML Advisory Team Partners with T-Mobile

T-Mobile and Splunk collaborate on an ML use case

1. Preprocess the Data

- Centralize data in Splunk
- Preprocess and extract features

2. Generate 1st level Models

- Create Anomaly Detection models and send results to a summary index

3. Fit the 2nd Level Models

- Anomalies of Anomalies

4. Tune the Models

- Iterate over models to validate and tune

5. Splunk .conf2019

- T-Mobile and Splunk co-present results in October

Splunk ML Advisory Team Partners with T-Mobile

T-Mobile and Splunk collaborate on an ML use case

1. Preprocess the Data

- Centralize data in Splunk
- Preprocess and extract features

2. Generate 1st level Models

- Create Anomaly Detection models and send results to a summary index

3. Fit the 2nd Level Models

- Anomalies of Anomalies

4. Tune the Models

- Iterate over models to validate and tune

5. Splunk .conf2019

- T-Mobile and Splunk co-present results in October

Splu
T-Mo

T-Mobile

For Example,

```
| fit DensityFunction ChubErrorRate by
        "DayHourDef,weekDayDef,operation"
        into ChubErrorRate_OPER_model
```

New Search

Save As ▾ Close

```
| summary ChubErrorRate_OPER_model
| table DayHourDef weekDayDef operation type
```

Last 60 minutes ▾

✓ 306 results (9/7/19 11:52:00.000 PM to 9/8/19 12:52:14.000 AM) No Event Sampling ▾ Job ▾ Verbose Mode ▾

Events (0) Patterns **Statistics (306)** Visualization

100 Per Page ▾ Format Preview ▾ < Prev 1 2 3 4 Next >

DayHourDef	weekDayDef	operation	type
Day	Weekday	Individual_spcCustomer	Auto: Exponential
Day	Weekend	UserNotificationLookup	Auto: Gaussian KDE
Night	Weekday	upsertIndCustomer	Auto: Exponential
Night	Weekday	SearchCustomerSummary	Auto: Gaussian KDE
Day	Weekend	lookupCustomer	Auto: Gaussian KDE
Day	Weekday	HNEUpsert	Auto: Gaussian KDE
Day	Weekend	UpsertEnrollment	Auto: Exponential
Night	Weekday	CreateBAN	Auto: Gaussian KDE
Night	Weekday	IndCustomerCreditUpdate	Auto: Gaussian KDE

splunk> .conf19

Splunk ML Advisory Team Partners with T-Mobile

T-Mobile and Splunk collaborate on an ML use case

1. Preprocess the Data

- Centralize data in Splunk
- Preprocess and extract features

2. Generate 1st level Models

- Create Anomaly Detection models and send results to a summary index

3. Fit the 2nd Level Models

- Anomalies of Anomalies

4. Tune the Models

- Iterate over models to validate and tune

5. Splunk .conf2019

- T-Mobile and Splunk co-present results in October

Splunk ML Advisory Team Partners with T-Mobile

T-Mobile and Splunk collaborate on an ML use case

1. Preprocess the Data

- Centralize data in Splunk
- Preprocess and extract features

2. Generate 1st level Models

- Create Anomaly Detection models and send results to a summary index

3. Fit the 2nd Level Models

- Anomalies of Anomalies

4. Tune the Models

- Iterate over models to validate and tune

5. Splunk .conf2019

- T-Mobile and Splunk co-present results in October

Splunk ML Advisory Team Partners with T-Mobile

T-Mobile and Splunk collaborate on an ML use case

1. Preprocess the Data

- Centralize data in Splunk
- Preprocess and extract features

2. Generate 1st level Models

- Create Anomaly Detection models and send results to a summary index

3. Fit the 2nd Level Models

- Anomalies of Anomalies

4. Tune the Models

- Iterate over models to validate and tune

5. Splunk .conf2019

- T-Mobile and Splunk co-present results in October

Anomaly Detection in Machine Learning Toolkit

splunk> .conf19

Detecting Numeric Outliers in Splunk MLTK

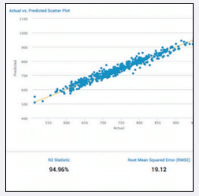
Showcase | Experiments | Search | Models | Classic | Settings | Docs | Video Tutorials | Splunk Machine Learning Toolkit

Showcase

Welcome to the Showcase, which exhibits some of the analytics enabled by this app.
Click on one of the examples to see that Assistant applied to a real dataset.
Please see the video tutorials for more information.

Select which examples to show

All Examples

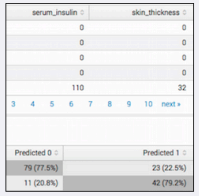


Predict Numeric Fields

Predict the value of a numeric field using a weighted combination of the values of other fields in that event. A common use of these predictions is to identify anomalies: predictions that differ significantly from the actual value may be considered anomalous.

Examples

- Predict Server Power Consumption
- Predict VPN Usage
- Predict Median House Value
- Predict Power Plant Energy Output
- Predict Future Logins
- Predict Future VPN Usage (sinusoidal time)
- Predict Future VPN Usage (categorical time)

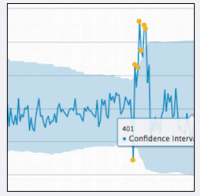


Predict Categorical Fields

Predict the value of a categorical field using the values of other fields in that event. A common use of these predictions is to identify anomalies: predictions that differ significantly from the actual value may be considered anomalous.

Examples

- Predict Hard Drive Failure
- Predict the Presence of Malware
- Predict Telecom Customer Churn
- Predict the Presence of Diabetes
- Predict Vehicle Make and Model
- Predict External Anomalies



Detect Numeric Outliers

Find values that differ significantly from previous values.

Examples

- Detect Outliers in Server Response Time
- Detect Outliers in Number of Logins (vs. Predicted Value)
- Detect Outliers in Supermarket Purchases
- Detect Outliers in Power Plant Humidity
- Detect Cyclical Outliers in Call Center Data
- Detect Cyclical Outliers in Logins

splunk> .conf19

Calculate the outliers

```
| inputlookup hostperf.csv | eval _time=strptime(_time, "%Y-%m-%dT%H:%M:%S.%3Q%z") | timechart span=10m max(rtmax) as responsetime | head 1000
| streamstats window=200 current=true median("responsetime") as median                    // calculate the median value using a sliding window
| eval absDev=(abs('responsetime'-median))                                                // calculate the absolute deviation of each value from the median
| streamstats window=200 current=true median(absDev) as medianAbsDev                      // use the same sliding window to compute the median absolute deviation
| eval lowerBound=(median-medianAbsDev*exact(20)), upperBound=(median+medianAbsDev*exact(20))   // calculate the bounds as a multiple of the median absolute deviation
| eval isOutlier=if('responsetime' < lowerBound OR 'responsetime' > upperBound, 1, 0)    // mark values outside the bounds as outliers
```

Detect Num
Find values that diffe
Assistant Setting
Enter a search
| inputlookup h
✓ 1,000 results (12
Field to analyze
responsetime
Detect Outliers
All time
Smart Mode
Data and Outliers
125
100
75
50
25
0
-25
Series 2
4
outliers

fit DensityFunction X

fit DensityFunction X by “host”

fit DensityFunction X
by “app,host”

```
fit DensityFunction X
by “date_hour,app,host”
```

fit DensityFunction X into model

apply model

Solution Architecture

splunk> .conf19

Incident Detection (Phase 1 Architecture)

Data Sources

Data Types

Splunk | ExtraHop | AppD

Log Events | Wire Data | Metrics

FeatureExtraction

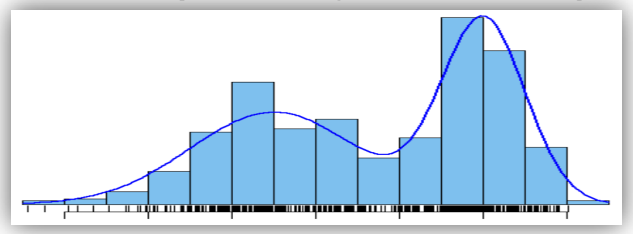
Parse | Transform

Normalize | Scale

Splunk Anomaly Detection (Density Function)

Per Metric Distributions
(Automatic Identification)

- Normal
- Gaussian KDE
- Exponential



Anomaly Detection (1st Stage)

Volume | Memory | Error Rate | Performance | Error Messages | Error Codes | Additional Metrics

Applications / Resources

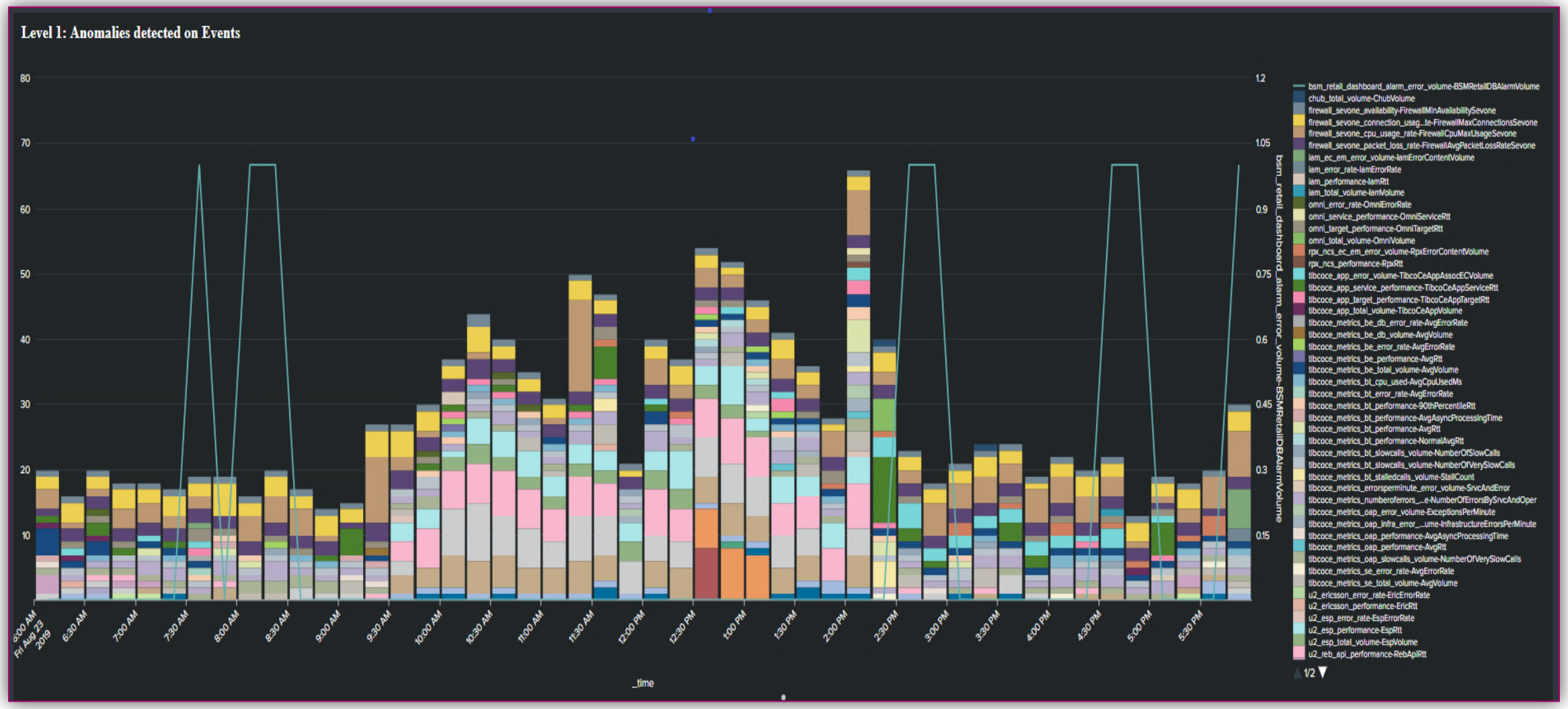
App 1 | App 2 | App 3 | App 4 | App 5 | App 6 | App (n)…

Anomaly Detection (2nd Stage)

Anomalies to Summary Index → Application of Anomalies Anomalies model → Alert

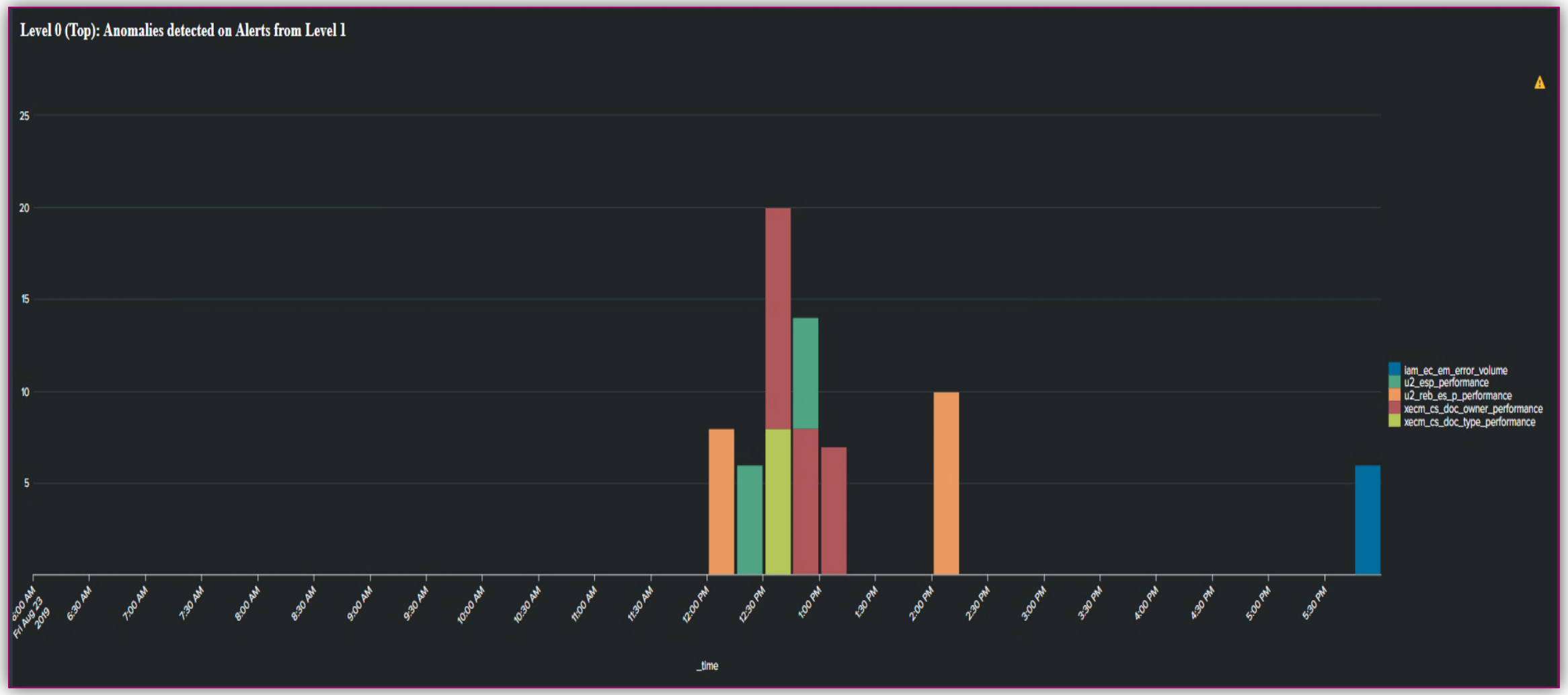
splunk> .conf19

Detected Anomalies – Level 1

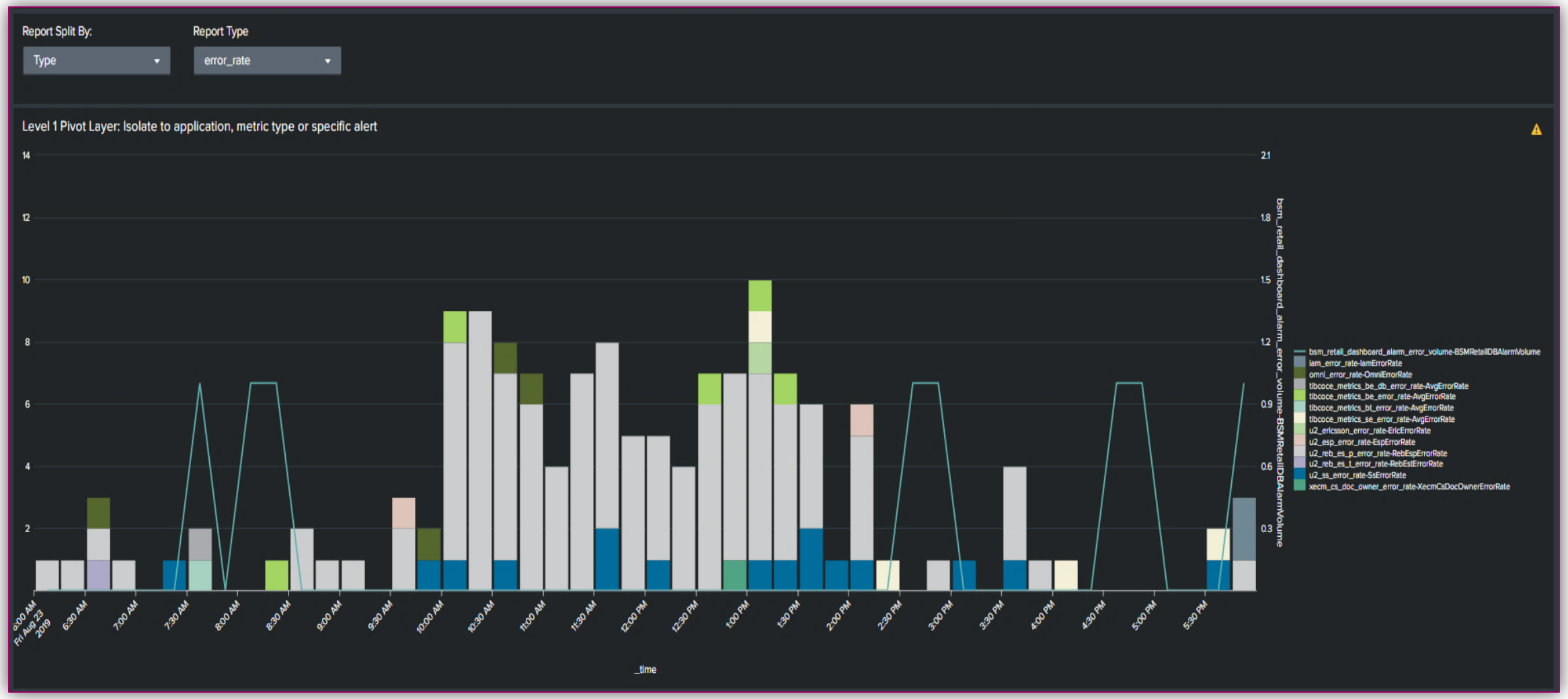


splunk> .conf19

Anomalies of Anomalies – Level 2

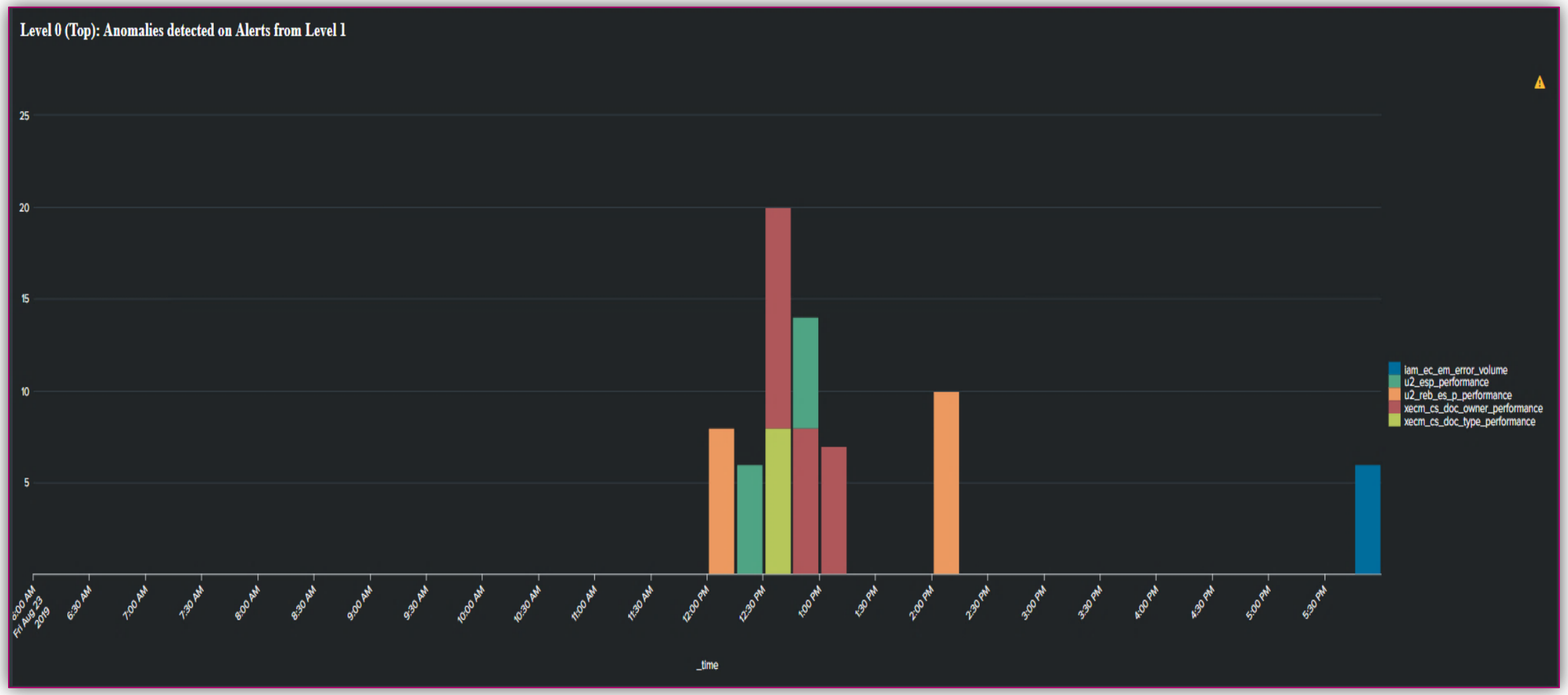


Detected Anomalies (Pivot on Error Rate) - Level 1



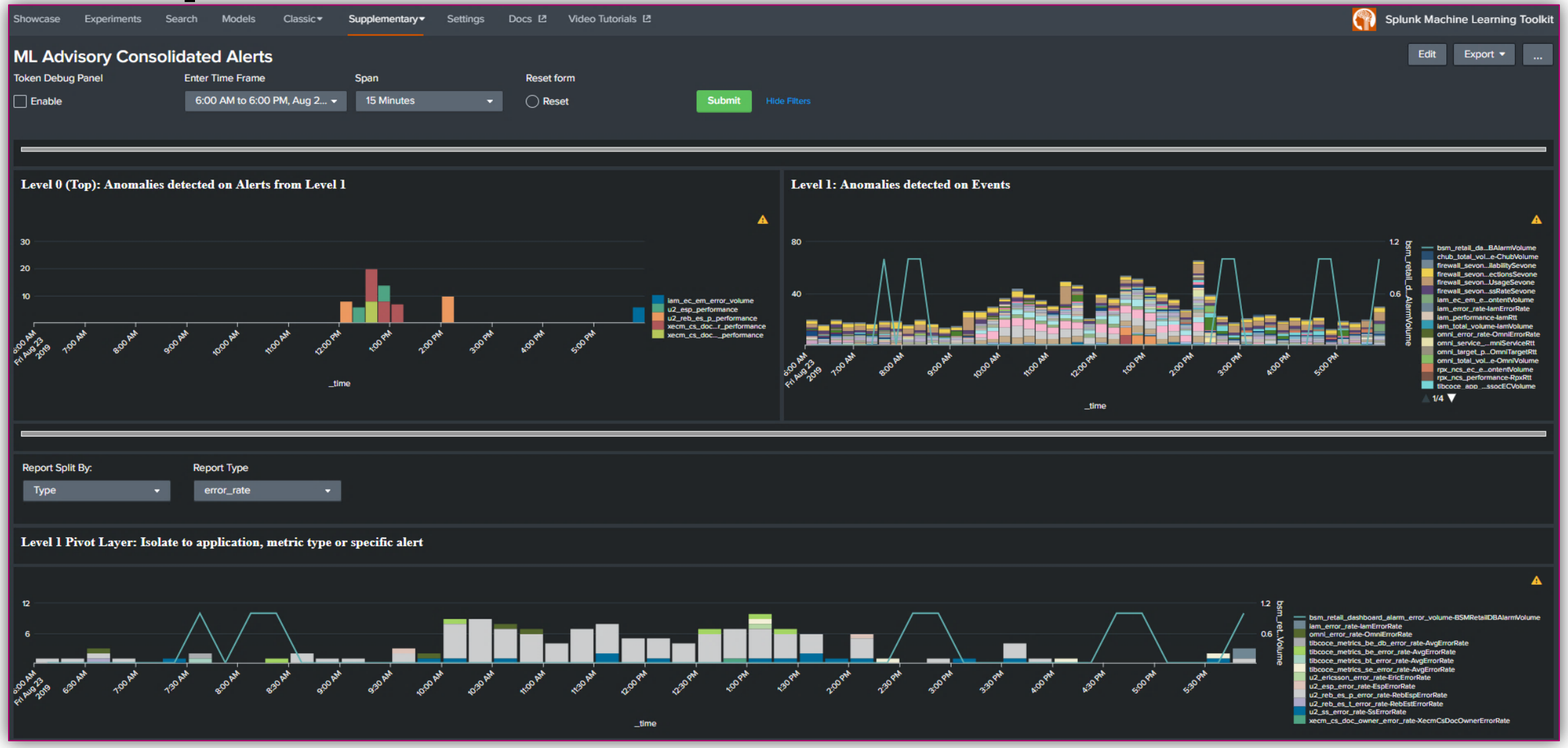
splunk> .conf19

Anomalies of Anomalies – Level 2



splunk> .conf19

Composite Dashboard



Phase 2
Incident Detection and RCA

splunk> .conf19

Strategy – Phase 2

Score and prioritize anomalous events to identify related anomalies

Score events

Enrich Data with Topology

Correlate and Classify related events

Ensure alerts contain adequate contextual content

Identify root cause analysis

Automate remediation

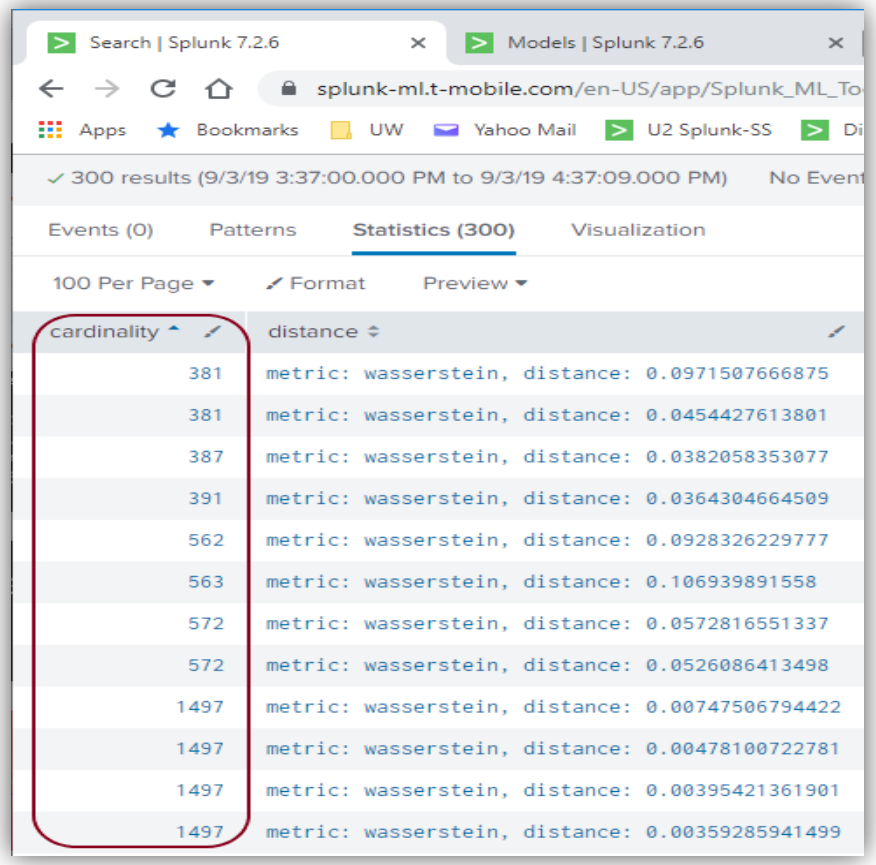
Lessons Learned

splunk> .conf19

Model Accuracy

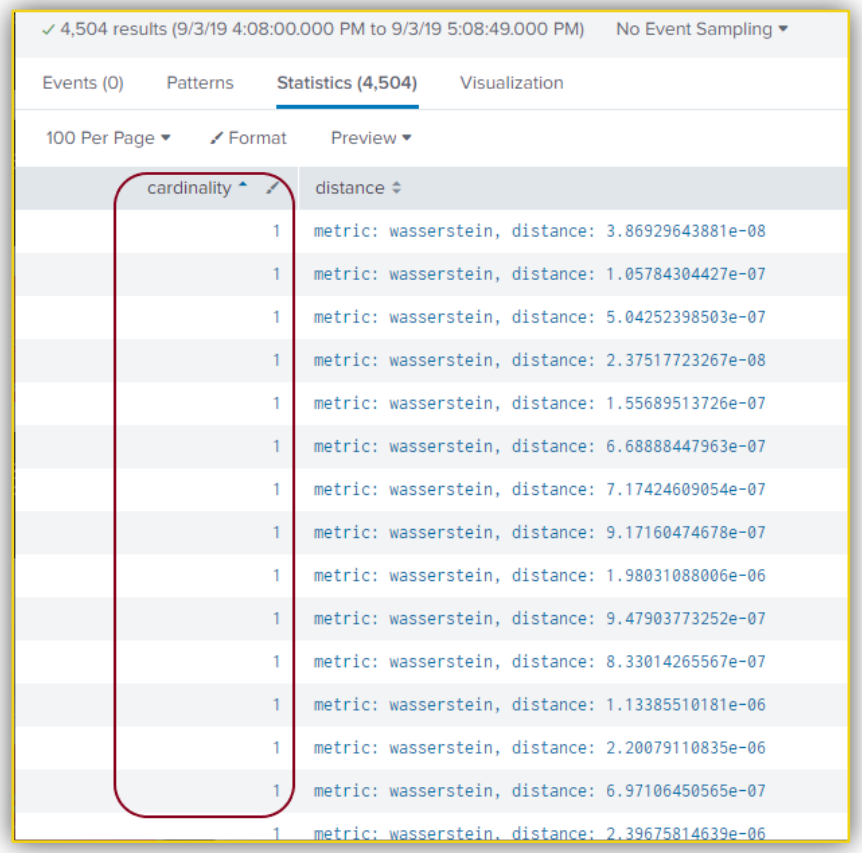
Unable to ascertain the model accuracy other then displaying cardinality

Superior Accuracy



cardinality	distance
381	metric: wasserstein, distance: 0.0971507666875
381	metric: wasserstein, distance: 0.0454427613801
387	metric: wasserstein, distance: 0.0382058353077
391	metric: wasserstein, distance: 0.0364304664509
562	metric: wasserstein, distance: 0.0928326229777
563	metric: wasserstein, distance: 0.106939891558
572	metric: wasserstein, distance: 0.0572816551337
572	metric: wasserstein, distance: 0.0526086413498
1497	metric: wasserstein, distance: 0.00747506794422
1497	metric: wasserstein, distance: 0.00478100722781
1497	metric: wasserstein, distance: 0.00395421361901
1497	metric: wasserstein, distance: 0.00359285941499

Inferior Accuracy



cardinality	distance
1	metric: wasserstein, distance: 3.86929643881e-08
1	metric: wasserstein, distance: 1.05784304427e-07
1	metric: wasserstein, distance: 5.04252398503e-07
1	metric: wasserstein, distance: 2.37517723267e-08
1	metric: wasserstein, distance: 1.55689513726e-07
1	metric: wasserstein, distance: 6.68888447963e-07
1	metric: wasserstein, distance: 7.17424609054e-07
1	metric: wasserstein, distance: 9.17160474678e-07
1	metric: wasserstein, distance: 1.98031088006e-06
1	metric: wasserstein, distance: 9.47903773252e-07
1	metric: wasserstein, distance: 8.33014265567e-07
1	metric: wasserstein, distance: 1.13385510181e-06
1	metric: wasserstein, distance: 2.20079110835e-06
1	metric: wasserstein, distance: 6.97106450565e-07
1	metric: wasserstein, distance: 2.39675814639e-06

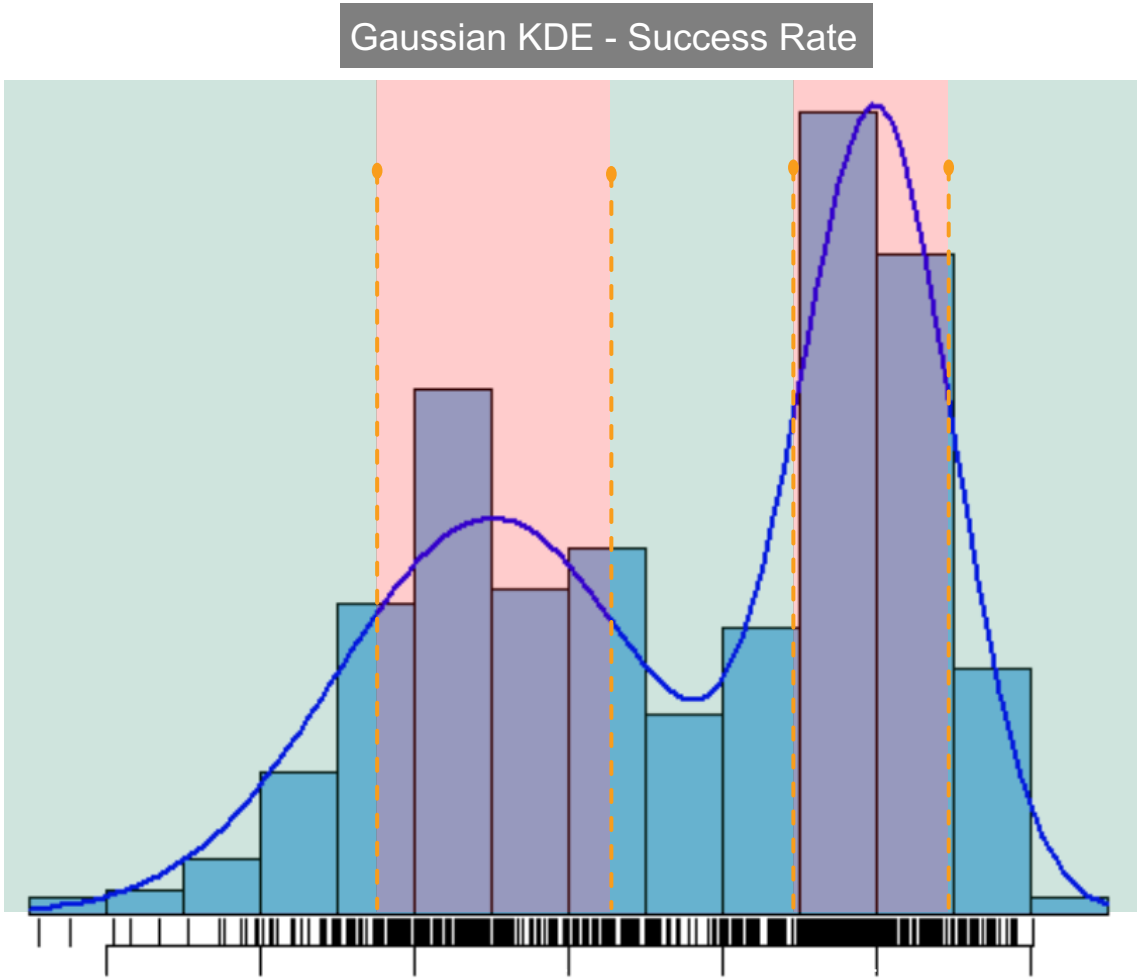
splunk> .conf19

Selective Boundary Limitations

With the exception of the normal distribution, selecting only upper or lower bounds was not possible

Example

- *For success rate, the upper boundary is the only boundary which matters*
- *Given a KDE distribution, segments of the population below upper minimum will be classified as Non-Anomalous*
- *Solution: Create a macro to parse the boundary_ranges field and manipulate modeled outliers which are set beneath the upper minimum*



Fit: Limit dimensions when appropriate

How and when to select fields to split by in order to improve model performance

- Example, CPU Resource may not require seasonality
- By removing seasonality, the number of data points increase which results in a significantly better model.

```
| eval hourOfDay=strftime(_time, "%H")
| eval dayOfWeek=strftime(_time, "%A")
| eval dayHourDef = case(
    hourOfDay >= 00 AND hourOfDay < 06, "Night",
    hourOfDay >= 06 AND hourOfDay < 22, "Day",
    hourOfDay >= 22 AND HourOfDay <= 23, "Night“)

| eval weekDayDef = if(
    like(dayOfWeek, "Saturday") OR like(dayOfWeek, "Sunday"),
        "Weekend",
        "Weekday“)

| fit DensityFunction cpu_usage_rate
  by “dayHourDef,weekDayDef, cpu_usage_rate"
  into cpu_usage_rate_model
```

Versus

```
| fit DensityFunction cpu_usage_rate
    by “cpu_usage_rate"
    into cpu_usage_rate_model
```

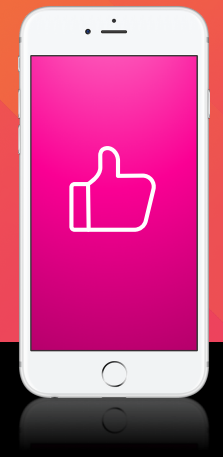
.conf19

splunk>

Thank You!

Go to the .conf19 mobile app to

RATE THIS SESSION



Q&A

splunk> .conf19