Multi Layered Unsupervised Machine Learning for Detection of Real Time Network Service Impairments

An Operational Practice prepared for SCTE by

Eric Frishman
Senior Principal Engineer
Comcast
1800 Bishops Gate, Mount Laurel, NJ 08054
215-510-6347
Eric_frishman@cable.comcast.com

Devangna Kaushish
E4 Product Development Engineering
Comcast
1800 Bishops Gate, Mount Laurel, NJ 08054
901-605-7629
Devangna_kaushish@cable.comcast.com

Russell Harlin
Senior Engineering Manager
Comcast
1050 Enterprise Way, #500, Sunnyvale, CA 94089
650-766-7256
Russell_harlin@comcast.com

Aditya Vallabhajosyula
E4 Software Development and Engineering
Comcast
1050 Enterprise Way, #500, Sunnyvale, CA 94089
aditya_vallabhajosyula@comcast.com

Eswaramoorthy Subramaniam
Principal Engineer Product Development
Comcast
11951 Freedom Dr, Ste 900, Reston, VA
Eswaramoorthy_Subramaniam@cable.comcast.com

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

1. Introduction

Finding and categorizing network impairments is a significant challenge. Using Machine Learning (ML) to study network traffic patterns, we can quickly discover places where behaviors depart from statistical norms and highlight them. As the divergences may stem from common underlying characteristics themselves, ML modeling can further separate and classify situations to speed in diagnoses and resolution, leading to improved customer experience and satisfaction as well as more efficient use of staff resources.

In this paper and presentation, we will walk through project and process flow, and ML considerations.

2. Background

Unsupervised ML techniques offer great methodologies of discovery to distinguish underlying behavior through data organization, unlabeled classification, density analysis, and outlier detection. Though these methods do not specifically identify or label outcomes, these techniques utilize sophisticated hunt and search capabilities to first highlight areas of interest and thereafter parse and filter them into more organized and simplified data structures. From these updated structures, user feedback through semi-supervised ML learning additionally refines the outcomes into appropriate anomaly labeled classifications. Thus, we begin with unsupervised techniques and through our workflow we finish with supervised and labeled solutions as outlined in Figure 1.

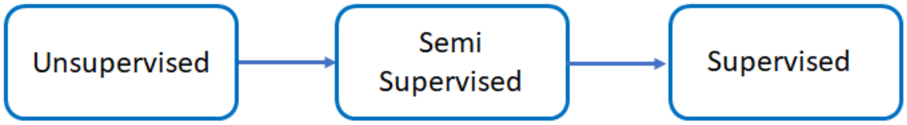


Figure 1 – ML Workflow

With the above broadly structured approach to both organize and thereafter classify the information for outlier detection, the generalized solution is highly flexible and adaptable to a wide variety of situations and data sets. In addition, an unsupervised ML leading stage can simplify extremely large data sets of hundreds of millions of data points into more reasonably organized and explorable subsets.

In the following diagram, Figure 2, the flow from left to right highlights the fundamental functional components and ML structures we used to detect outlier traffic behavior within a Cable Model Termination System (CMTS). The benefit of this approach is layering suitable ML techniques to refine, adapt, and classify broad and small-scale outliers.

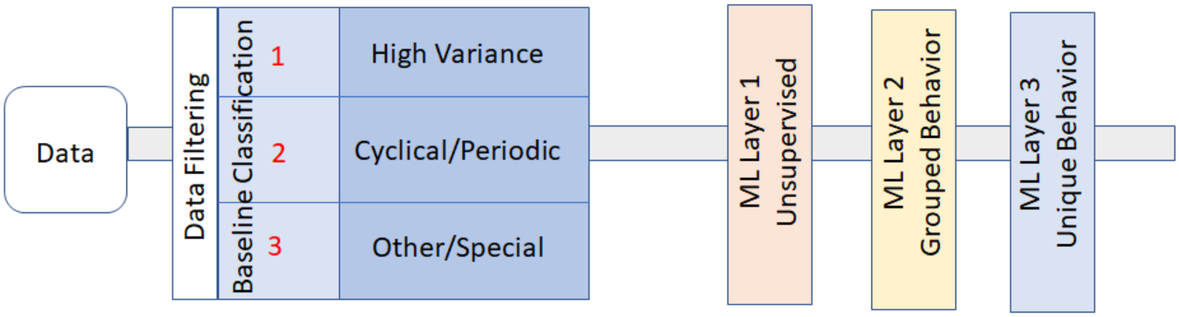


Figure 2 – High Level ML Structure

3. Exploratory Data Analysis

Data exploration is a key critical first step in developing robust solutions. The understanding and consideration of signal data changes, variations, and patterns with respect to time and frequency, as well as how outliers manifest themselves, permits a stronger selection of appropriate ML tools, techniques, and solutions. By choosing to create baselines against the lowest granular element in the network, the cable modem, higher level aggregates and analysis become possible, corresponding to ML Layer 2 within Figure 2. We used anonymized cable modem traffic data to better understand usage patterns and impacting outliers. Data usage, with strong outlier detection methods, can be turned into high value data streams when other informational sources and data are integrated Quality of Service (QOS), latency, bandwidth utilization, software releases, hardware versioning, etc.).

The data set consists of millions of elements which are evaluated for every time interval. This results inhundreds of millions to a billion analyses and status determinations in a day. With large data output volumes, a critical consideration is reducing both the False Positive and False Negative rates in alerting/alarming. This is a primary concern since appropriately focusing precious operational resources reduces mean time to resolve (MTTR). False positives, even in small numbers, may become prohibitive unless efforts are well aligned. Therefore, any aspect which improves signal-to-noise ratios through data cleaning and robust model development is critical. For simple signals, ML modeling does not pose a significant challenge. For very complex signals, variable design and simplifications though classification assist in running models effectively at scale.

Figure 3 shows a few examples of cable modem traffic variations:

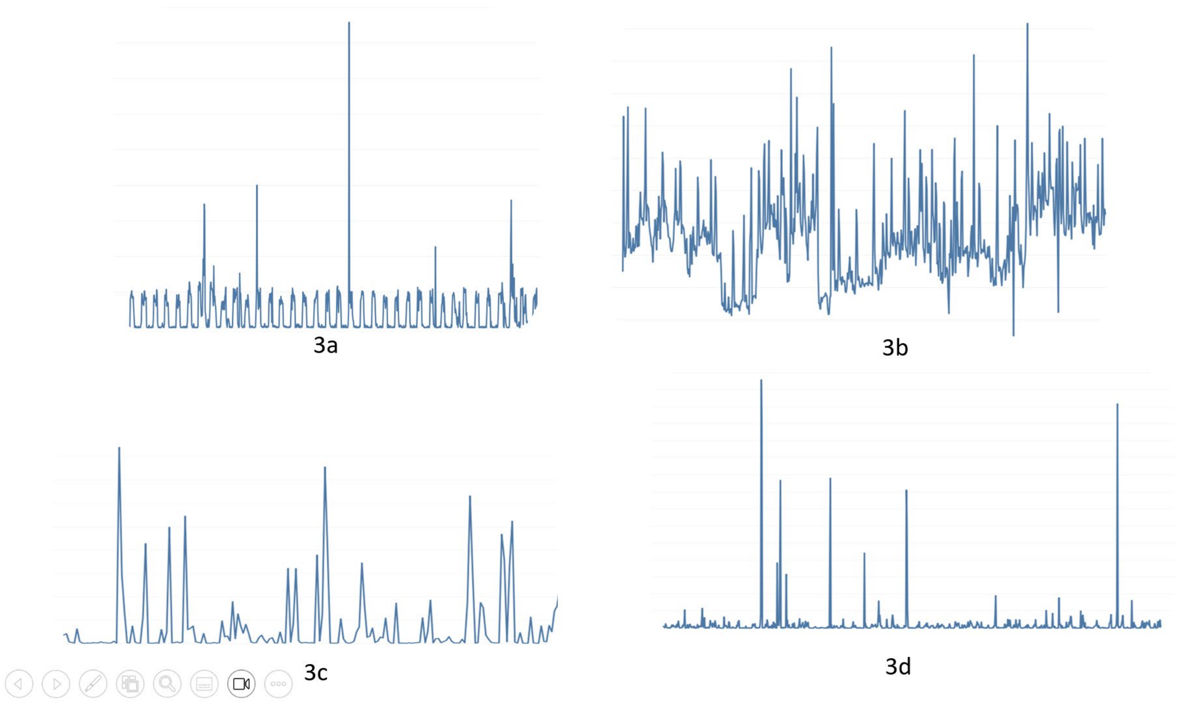


Figure 3 – Traffic Samples

The above examples have a variety of individual features. Note that all of the behavior within these graphs is expected without any outliers or points of interest. Within Figures 3a and 3b, there are components of cyclical traffic behavior as well as a variety of signal changes and variations. Figure 3a has high periodicity with some higher usage spikes. Figure 3b shows persistent higher usage with large fluctuations. Day of the week, as well as hour of day usage patterns, may be evident in some examples, however, usage itself is rarely a constant and is subject to instantaneous change.

Figures 3c and 3d emphasize very low traffic usage preceding much heavier volumes with significant amplitude differences. Traffic usage is often quite bursty with large volumes in short periods of time. Traffic peaks may be periodic, though in many cases sporadic with large magnitude differences and long and short time pauses in between peak usage.

In many cases historical signal usage is only loosely correlated with potential current expectation forecasts. Therefore, when modeling individual elements, to account for these potentially significant changes and shifts, individual element modeling must incorporate appropriate biases when possible. Adaptable models which characterize signal components with respect to intensity, symmetry, time, variance, and signal randomness may create robust signal expectations. Please note that long-time interludes of low usage may not be representative of significant change, and simply reflective of normal usage.

As noted, there are a variety of attributable signal differences at the cable modem level. Due to high variance elements having much lower predictability, we decided to model high variance elements as their own class separately. Categorically defining baseline characteristics enables specific signal modeling refinements for complex behavior patterns.

Figure 4 depicts snapshot views of signal characteristics based on signal intensity and time distributions.

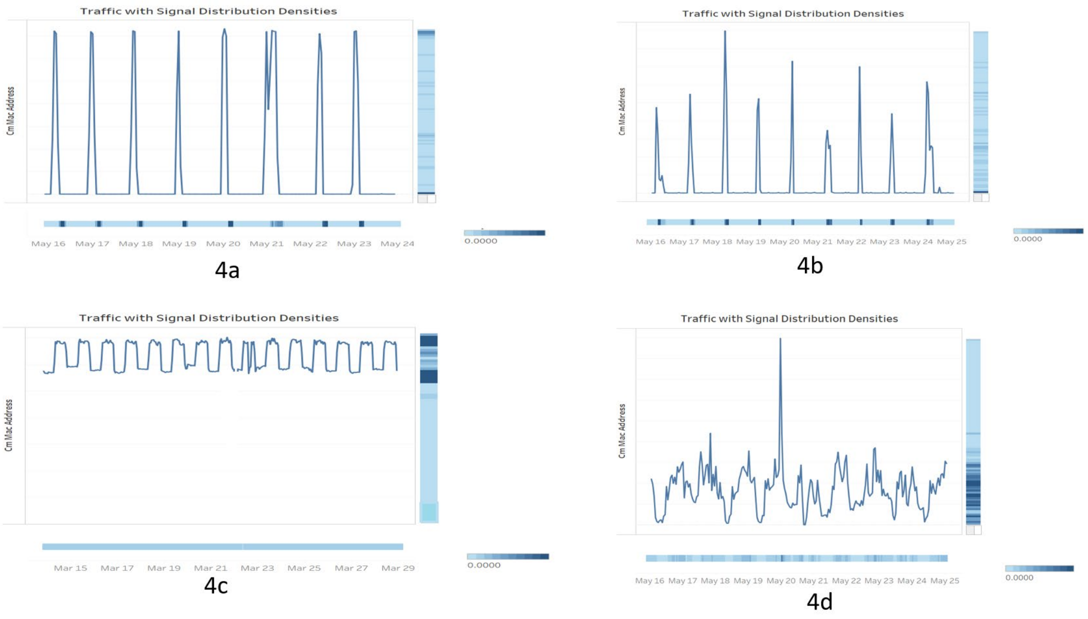
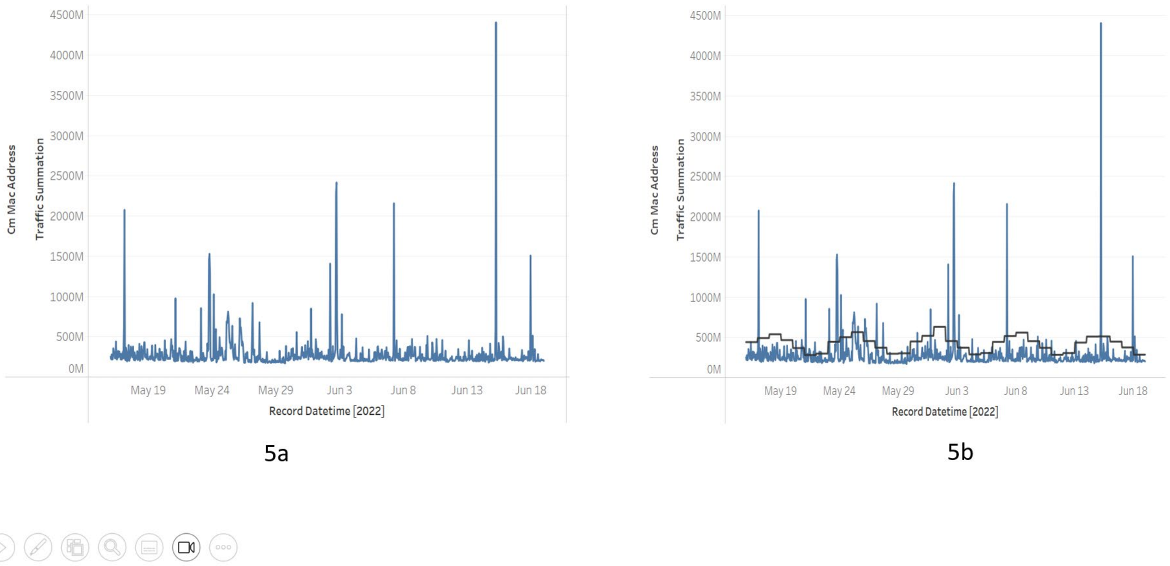


Figure 4 – Traffic and Density Views

Through variable design, relative descriptive features are developed to generalize, differentiate, and classify the raw signal data into categories. In the four examples in Figure 4, there are distribution density bar analyses included on the horizontal time axis and the vertical traffic volume axis. Darker colors indicate higher data density presence as well as positioning within the respective axis. We quantify and qualify important signal aspects using the density qualities and signal position references to create assignable model classifications. In comparing Figures 4a and 4b against Figures 4c and 4d, the time signal density distributions along the horizontal time axis appear very different with the highly variant signals showing great concentration in very narrow bands. Traffic densities on the vertical traffic axis may often be largely unpredictable but are recorded characteristics. While sometimes highly variant signals show periodicity, this was not shown to be reliably predictive of outlier detection in many cases. In contrast, signals of lower variability had greater predictability in periodicity and usage but were also subject to data shifts simply because of usage changes. Through separation, each class may be distinctly modeled.

The vertical traffic density mappings show signal concentrations and divergences, with high variability. Using both traffic intensity and time density mappings, we create model classifications with respect to groups as well as individual element models. The time density characteristics are a key mapping for differentiating highly variant signals and the combination of both the horizontal and vertical densities contribute to an overall understanding of cyclical behavior and persistence.

The first stage within the Figure 2 solution is focused on signal classification and reliable baseline model development using historical patterns and density signal mappings. Signals which are highly variant may be so in both intensity and/or time. While highly variant signals might be periodic, a noteworthy element is that the length of time between usage may often be sporadically distributed with long periods of time between large amplitude peaks. Because of this, low periods of utilization may have little predictive value and modeling these areas may hinder model convergence. Instead, low traffic periods may be compressed to quantify the peak amplitudes better.



5a

5b

Figure 5 – Baseline Modeling

The graphics in Figure 5 depict the raw usage signal in Figure 5a as well as the interpolated baseline in black, Figure 5b. The baseline includes aspects of time-based periodicity including day of the week and hourly components. Note that there are instances of higher intensity beyond the baseline model. These aspects and their interpretations are managed within Layer 1 of the ML model for appropriate outlier detection.

4. ML Modeling: Primary and Secondary Variables

After the creation of individual baselines there will be instances where signals may exceed modeled expectations as shown in Figure 5. These areas are not necessarily indicative of outliers but high variability and lower predictability. Our objective is to characterize both signal characteristics and signal changes. We do this by creating primary and secondary evaluation variables analyzed through sequential ML layers. This progression moves from unsupervised/unknown descriptive features to classified subsets. The primary variables relate to signal changes with respect to the developed modeled baselines including changes of intensity, time, density, and position. These variables relate to characteristics of unsupervised modeling in ML Layer 1 in Figure 2. The secondary variables consist of grouped common classification sets evaluated by ML Layer 2 in Figure 2. These secondary characteristics include software type, hardware type, physical network layers [CMTS, service group, virtual group, etc.], Ip Version, or any other grouped classification of interest.

Variation metrics are derived to reflect how much outlier traffic signals deviate from the expectations using ML Layer 1, for each baseline. There are both parametric and non-parametric methods for these derivations, and the metrics create and define the appropriate boundaries in Figure 2, ML Layer 1, as described in Section 5.

5. 1st ML Layer: Outlier Distribution Analysis

Using the modeled baselines for each traffic subgroup, we then created intensity boundaries to define appropriate outliers. A baseline sets the key characteristics of expectation for outcome modeling. However, outlier sensitivity may be specific to project or end user perspectives. Because element traffic behavior may exceed developed baseline models simply because of instantaneous changes in pattern usage, ML's non-linear solution space helps to clarify outcomes. With the integration of density-based ML methods to define intensity levels, and through semi-supervised feedback through sampling, zones or areas of true outlier interest may be quickly derived, explored, and refined, at a cable modem level.

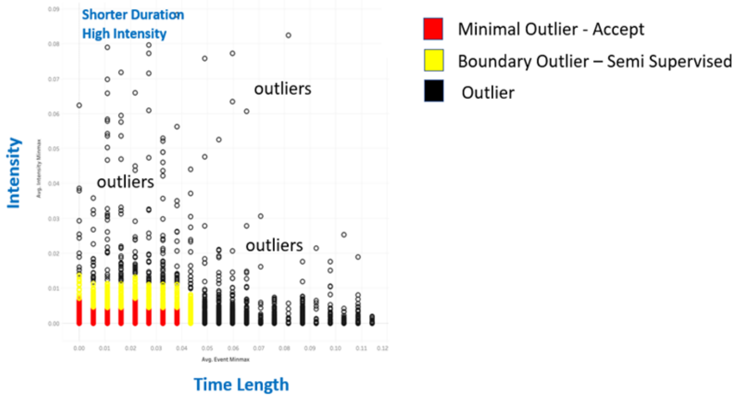


Figure 6 – ML Layer 1 Outlier View

Figure 6 has simplified the primary variables, which may include relative signal position, velocity and acceleration components, or other characteristics, into a two-dimension perspective showing the intensity of variation (y-axis) and time dimensions (x-axis). Please note that the axes have been normalized. Since each sample point falls outside of baseline expectations, all are technically outliers. However, because of the variability of signals, and the complexity of modeling, the simple presence and position of a sample outside of a modeled boundary may not be sufficient in intensity to classify it as an outlier. The great benefit of unsupervised modeling is that characteristics of these key variables and density changes throughout the distribution organize and identify specific areas of interest. With a statistical sampling of the distribution at various intensity points based on time, intensity, or both variables, semi-supervised client feedback may be propagated into the distribution so that millions of data points are classified into an appropriate data stratum, using only a few representative points.

In Figure 6, the red points reside closely to the baseline boundaries with respect to intensity and/or time. For outliers that occur in shorter time durations, oftentimes greater magnitudes of intensity might be necessary for alerting, as opposed to lower intensity outliers which may only be of interest if they exist for longer durations. Therefore, areas that are highlighted in red may not be of interest. Yellow demarked boundaries might be of sufficient intensity and are dependent on semi-supervised client feedback. The black marked areas are reflective of outliers which should be alarmed. The boundary definitions between red, yellow, and black areas are quickly discernible through sampled feedback.

6. 2nd ML Layer: Grouped Outlier Distribution Analysis

The first ML layer defined potential issues at a cable modem level against baselined attributes and key performance indicators (KPI), in this case, intensity and time. ML layer 2 in Figure 2 organizes and evaluates additional shared descriptive features among device groups. For these secondary descriptive variables, grouped baselines and correlative analysis may provide insight into common behavior shifts. These characteristics may include groupings of different network layers [CMTS, Service Group, Virtual Group, etc.], software releases, hardware associations or other similar collections. The first ML layer defines individual performance, while the second layer organizes outlier groupings for correlated behavior shifts.

For network layer associations, the highest level of incident impact will most often characterize the affected network level. An impacted lower-level service group usually will not result in a higher-level CMTS aggregate data alarm. However, CMTS level issues may propagate across many lower network levels; this differs from descriptive grouped collections of software or hardware that would show effects within their specific classifications.

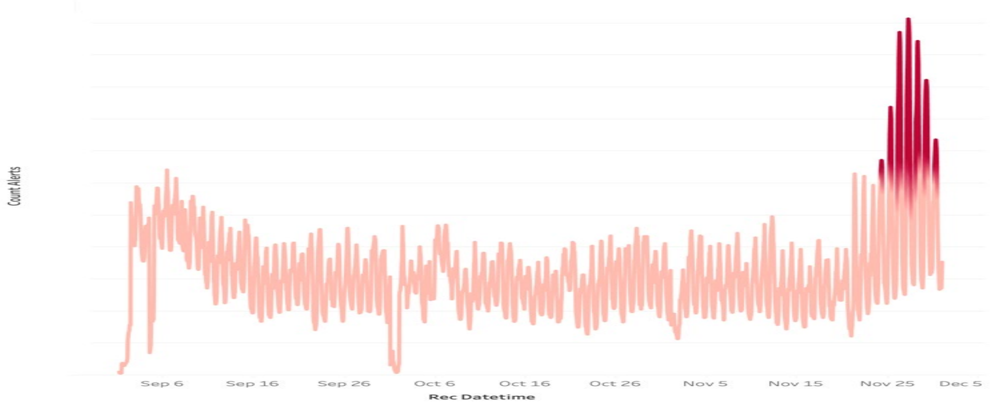


Figure 7 – ML Layer 2 Grouped Outliers

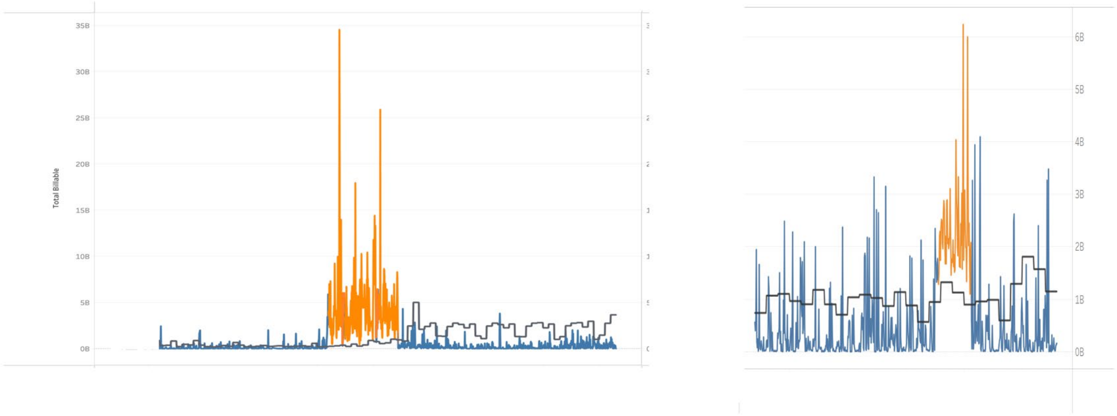
Figure 7 shows historical outlier counts from ML Layer 1 organized and modeled into a network CMTS ML Layer 2. The dark red highlights show baseline outlier events quickly rising and being detected. Large scale element counts are easily monitored and tracked using ML strategies.

7. 3rd ML Layer: Unique Element Distribution Analysis

The first layer of ML provides single-element interpretation of outlier or normal expectation status for each data point. Because the model baselines provide a firm foundation, the characterization of group behavior is then easily interpreted by ML layer 2 for shifts in anomaly counts and type qualifications. With strong individual model baselines, outlier evaluations at the cable model level, the lowest common denominator, become possible despite this resolution’s high variance and variability.

The statistics and evaluations from the proposed multi layered model indicated a 0.05% outlier issue rate at ML Layer 2 during a normal operating cycle. At this issue rate, outlier presence is sufficiently low such that individual cable modem investigations become operationally possible through an additional layer of ML modeling and filtering. Figure 2 ML Layer 3 creates individual element density models and expectations for alerting confirmation at a cable modem granularity.

Figures 8 shows areas of interest based on intensity and/or time where baselines differed from expectations.



8a

8b

Figure 8 – Outlier Detection Examples

Figure 8 shows a divergence from baseline expectations in both magnitude and time whereas, Figure 8B reflects a time-based length event. While visual outliers may be easy to discern, these examples exist within hundreds of millions of examples, and the machine learning process is quickly able to highlight them.

8. Large Scale Implementation

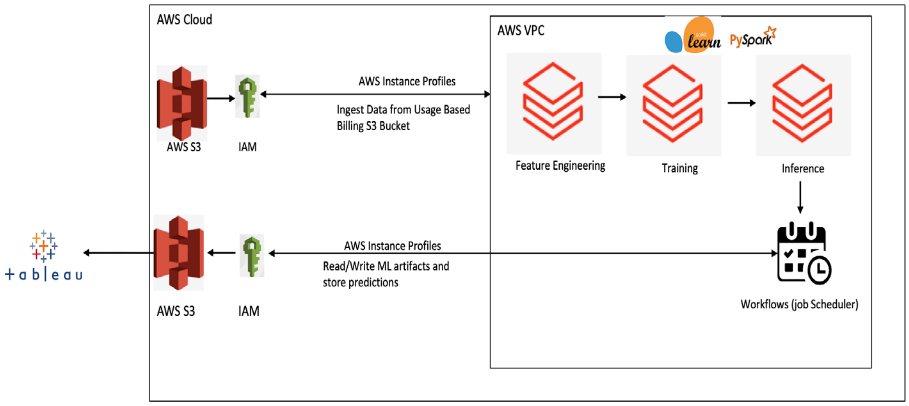


Figure 9 – Large Scale Design

Figure 9 shows the implementation of the model through Databricks on AWS as it easily processes large datasets and multiple models quickly and efficiently. The design is comprised of feature engineering, model training, and model inference. The predictions are stored in AWS S3, and Tableau is used for visualizations.

The model is in a preproduction phase at present. The addition of centralized logging and alerting will complete the deployment.

9. Conclusion

Through this paper, we have explored how it is possible to model highly variable and dynamic signals by creating a construct that initially develops modeled baselines using interpreted signal characteristics, creates data organization through unsupervised ML models, and then parses the information through outlier detection ML Layers 2 and 3 for additional refinement and characterization. With these methods, it is possible to provide reasonable coverage for very large systems and optimize outputs such that operational investigation of outliers is both reasonable and possible.

We have shown that the described methodology has captured potential network anomalies of a software issue because of the ability to model at a very granular resolution, and capture grouped behavior and outliers well. Finally, with the ability to model at granular levels, data sets and models become value added data streams when joined with latency, dropped packet segments, and other Quality of Experience (QOE) metrics.

Abbreviations

CMTS	cable model termination system
MTTR	mean time to resolve
QOE	quality of experience
QOS	quality of service