Predicting Service Impairments from Set-top Box Errors in Near Real-Time and What to Do About It

How Machine Learning Can Preempt Calls and Tickets: Results from a Trial

A Technical Paper prepared for SCTE•ISBE by

Justin Watson
Senior Manager, Product Management
Comcast
Philadelphia, PA
Justin_Watson@comcast.com

Roger Brooks
Chief Scientist
Guavus, Inc.
San Jose, CA
roger.brooks@guavus.com

Andrew Colby**,** Office of the CTO, Innovation Lead, Guavus, Inc.

Pankaj Kumar**,** Senior Manager Analytics, Guavus, Inc.

Anant Malhotra**,** Principal Analytics Engineer, Guavus, Inc.

Mudit Jain, Principal Analytics Engineer, Guavus, Inc.

Table of Contents

List of Figures

List of Tables

Introduction

Getting ahead of subscriber problems is a difficult but powerful way to reduce costs and increase customer satisfaction. This paper describes a proof of concept (POC) trial that harnessed machine learning and Comcast X1 service impairment data to identify at-risk subscribers and risk drivers, and to further indicate next best-actions to take in response to the predicted issues.

Machine learning is well-positioned to address the blind spots of customer support teams. It can be architected to scale to tens of millions of simultaneous event streams and handle real-time, complex predictive analytics. The highly accurate analytics used in this trial enabled us to identify subscribers potentially affected by impairments responsible for generating 36 percent of calls and 46 percent of tickets. Only 5 percent of the device population drove these care events. To enable action, our analytics also identified the associated risk drivers. In another exercise, we predicted a large quantity of true positive tickets per year related to 13 newly clustered ticket classes, with known resolution paths, and associated 57 percent of those tickets with single problem codes. (Note: All tickets referenced in this paper are technical tickets.)

Shared among internal stakeholders, these kinds of insights can drive numerous benefits. They can enable service providers to proactively address technical problems of subscribers; reduce the number of calls, tickets and truck rolls as a result; and more quickly resolve impairment events that do arise. The net result is reduced costs and improved customer experience.

1. Problem Statement and Challenges

1.1. Customer Care Constraints

Facilities-based service providers know that providing excellent customer care takes tremendous effort. From NOCs to support personnel to maintenance technicians to software, equipment, and fleets of vehicles, it requires an extensive combination of resources to take care of customers. These costs add up. Scaled to millions of subscribers, customer care becomes a big number, one that even gets the attention of financial analysts.

Yet despite the investment and efforts, visibility remains limited. “We’re still reactive” - that’s what one operator admitted to the author of a paper on customer experience delivered at this conference last year [Cunha.] The assessment still applies widely across the industry. Past patterns help us schedule resources, network monitoring and telemetry provide device-level insight, and integrated ticketing systems gather what we know onto one screen, but it remains difficult to get ahead of the customer.

How to address the technical problems of subscribers who are likely to call, before they call, was the guiding question of a two-part data analytics project, undertaken in late 2017 and early 2018, whose results we share in this paper. Before getting to the study and results, let us first point to some challenges posed by the data and discuss how machine learning is well suited to meet them.

1.2. Challenges: Data, Calls and Tickets

The problem today is not too little data. In addition to information drawn from customer database and network telemetry common to most MSOs, Comcast can leverage the X1 set-top box. This next-generation IP video platform, which was launched five years ago, not only proved popular among subscribers, it has also generated tremendous amounts of data. Those include the large numbers of error streams that provide the foundation of this analytic exercise. (See Figure 1.)

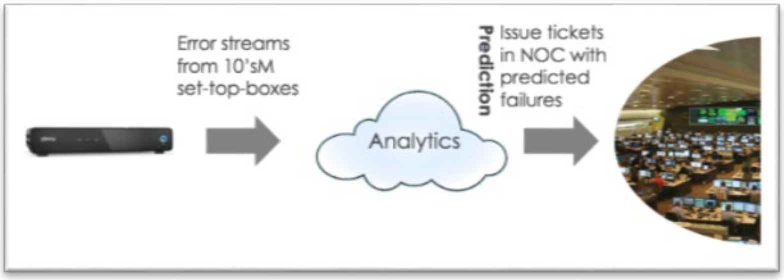
Error streams from 10'sM set-top-boxes
Analytics
Prediction
Issue tickets in NOC with predicted failures

Figure 1 – Predictive Maintenance Use Case

To anticipate customer calls, you need to not only process massive quantities of data, but also do so at near real-time speed. Legacy approaches associated with manual correlation, centralized processing and storage bottlenecks are poorly equipped to deliver the desired results.

Handling the data in this case also means prepping them for analysis. The challenge is common to other industries: raw data rarely arrive in neat taxonomies. Streaming off the X1 set-tops are 1-2 thousand types of errors codes that must be grouped, hierarchically ordered and contextualized within meta information linked to particular boxes.

The calls and ticket have other limits. Several tickets may arise from a single customer, each of which may have an incorrect or correct assessment. Because tickets bear subscriber account numbers, not MAC-level addresses, there is the further problem of not knowing which box in a multi-box home is impacted, forcing the machine to learn only from single-box homes. Given customer delays, the time stamp on tickets also may not convey when the impairments actually occurred.

Then there are the tickets that never arrive. Between 10 and 40 percent of subscribers who have some reason to call in to complain, do not do so. Sometimes called “silent sufferers,” they may just be apathetic or distracted or waiting for things to improve. In this study, however, they are also conservatively tagged as ‘incorrect care event predictions’, even though the errors codes did correlate to incidents. Consequently, this has the adverse effect of negatively impacting our accuracy score and should be kept in mind when reviewing the final accuracy scores.

A final issue with tickets is that some percentage are non-technical. As noted at the outset, for present purposes, drawing largely from X1 error codes, along with outage and reconnect data, this analysis is concerned with technical tickets, not those associated with billing or other types of problems.

2. Machine Learning

Machine learning is a good fit for these kinds of data. Rightly applied, this field of AI has numerous use cases. In a paper last year, several Comcast colleagues discussed using machine learning to simplify field operations, in particular, to detect spectral impairment [Dorairaj, et al.] In another paper, a Guavus colleague pointed to how a combination of machine intelligence and operational analytics is an effective way to assure virtualized networks and services [Sundelin].

Machine learning uses various tools across the entire process, from exploration and feature engineering to training, evaluation, tuning and deployment. The solution used in this exercise is a modular data ingest and analytics platform, containerized for the cloud, and highly scalable over a distributed architecture. Among the algorithms employed are the following:

- Quantile Transformer - enables features to follow a more uniform distribution
- Linear Support Vector Classification - allows selection of features based on weights
- Spectral Clustering - reduces spectrum of the similarity matrix before clustering data
- Probability Chunking - provides probability-based segregated chunks of predictions
- Hierarchy of Models - combines multiple models, all trained at each level of Tree-based taxonomy

One aspect worth underscoring is machine-learning's orientation toward probabilities. Unlike more deterministic techniques that provide yes/no answers, machine learning generates probabilistic results. The tradeoff between precision (correctness of the prediction) and recall (breadth of coverage of the predictions) is a common way to both measure the accuracy of the predictions and align the models to the business value sought. A lot depends on how averse one is to false positives.

3. Customers at-Risk and Risk Drivers

3.1. Trial Set Up

To preemptively address technical problems of individual subscribers, the first step is to identify who is at risk of calling or ticketing. Doing so also involves looking at the risks that are driving those incidents.

Who are the at-risk customers? Given that X1 set-top errors reflect actual impairments in service, we can assume those data are associated with some percentage of customers who do call. The initial data selected for this project followed from that premise. Our population of data came from all subscribers who had any X1 set-top errors over a 7-day period.

In this same exploratory phase, we assessed the problem portions of the ticket data, or problem codes. Using several machine-learning algorithms, including Spectral Clustering, we organized these codes into a taxonomy with the leaves of the taxonomy tree representing "classes" containing similar error data. (See Figure 2.)

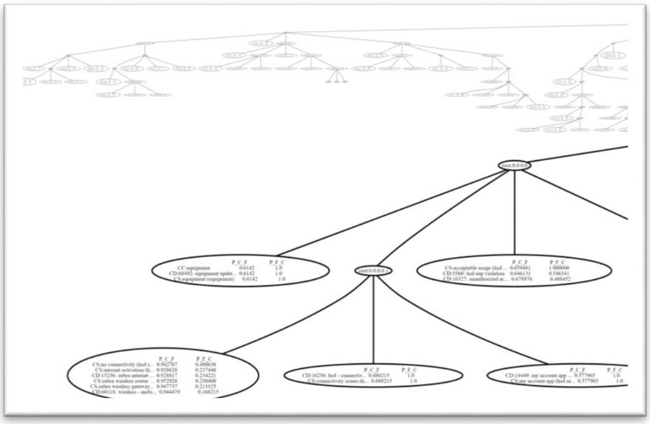
root.0.0.0.0
P_C_F P_F_C
CC:equipment 0.6142 1.0
CD:60492: equipment updat... 0.6142 1.0
CS:equipment (equipment) 0.6142 1.0
root.0.0.0.0.1
CS:acceptable usage (hsd ... 0.659481 1.000000
CD:5560: hsd aup violation 0.646131 0.546341
CD:16327: unauthorized ac... 0.678976 0.488452
CS:no connectivity (hsd s... 0.942767 0.480636
CS:internet activation (h... 0.928628 0.237446
CD:15236: zebra internet ... 0.928817 0.234221
CS:zebra wireless router ... 0.952928 0.230400
CS:zebra wireless gateway... 0.947737 0.213325
CD:60118: wireless - mobi... 0.944479 0.168215
CD:16256: hsd - connectiv... 0.680215 1.0
CS:connectivity issues (h... 0.680215 1.0
CD:14449: my account app ... 0.577965 1.0
CS:my account app (hsd se... 0.577965 1.0

Figure 2 – Problem Taxonomy Generated from Ticket Data

The next phase was identifying those in our subscriber pool most at risk of calling or ticketing. We combined the stream of X1 errors with reconnect, historical outage and other contextual information about individual subscribers and boxes. Feature engineering was then done to reformulate the properties of the data to best align with the use case and the machine's ability to interpret the data. We seeded the model with all calls and tickets and those aggregated features from three rolling windows of 1hr, 24hrs and 168 hrs. The output was an hourly prediction, aggregated for 7 days.

To assess greater or lesser risk, we used risk bucketing, a machine-learning method for correlating variables, in which similarity is calculated by rank score comparison and then displayed, largest group to smallest, with the smallest bearing the greatest probabilities.

3.2. Outsized Risk Factor Impact

The outcome of these exercises reveals striking results. The multi-bucket model, indeed, shows increasing number of predicted calls and tickets as group-sizes diminish. Bucket 4 warrants the most attention. It represents only 5 percent of total subs with errors, but drives 36 percent of actual calls and 46 percent of actual tickets. (See Tables 1 and 2.) This set offers the best target for preemptive action and cost reduction. Knowing that a certain percentage of issues can be handled proactively, we can envision reducing a number of these calls, tickets and other support on a recurring basis.

Table 1 – Risk Buckets: Predicted and Actual Calls vs. Total Subs with Errors (percentages)

Risk Bucket	Predicted	Actual	Subs w/Errors
0	8%	6%	35%
1	15%	14%	25%
2	19%	18%	20%
3	24%	26%	15%
4	34%	36%	5%
total	100%	100%	100%

Subscribers in Risk Bucket 4, while accounting for only 5% of the total subscribers with set-top errors, drive 36% of actual technical calls.

Table 2 – Risk Buckets: Predicted and Actual Tickets vs. Total Subs with Errors (percentages)

Risk Bucket	Predicted	Actual	Subs w/Errors
0	6%	5%	35%
1	15%	13%	25%
2	18%	16%	20%
3	19%	20%	15%
4	42%	46%	5%
total	100%	100%	100%

Subscribers in Risk Bucket 4, while accounting for only 5% of the total subscribers with set-top errors, drive 46% of actual technical tickets.

3.3. Drivers of Risk

To dive deeper into high-risk Bucket 4, we built another ranking that correlated risk drivers, i.e. our newly organized problem codes, with tickets and calls.

In descending predictive power are X1 errors, previous outages, previous calls and device model. At the top for calls are X1 errors and previous outages, which drive a roughly equal number of actual calls, and together account for three-fourths of the total. (See Table 3.) For predicted Bucket 4 tickets, X1 errors drive an even larger number of actual calls, and twice the events as previous outages. (See Table 4.) In both cases, previous calls outrank device model as explanatory features.

Table 3 – Predicted and Actual Calls Associated with Risk Drivers (percentages)

Risk Drivers	Predicted	Actual
X1 Errors	41%	37%
Previous outage	34%	37%
Previous calls	19%	20%
Device model	6%	6%
Total	100%	100%

Table 4 – Predicted and Actual Tickets Associated with Risk Drivers (percentages)

Risk Drivers	Predicted	Actual
X1 Errors	51%	48%
Previous outage	23%	21%
Previous tickets	18%	22%
Device model	8%	9%
Total	100%	100%

3.4. Model Accuracy

As one can see from Tables 1–4, predicted calls and tickets track closely with the actual ones. Figures 3 and 4 provide additional evidence for the high accuracy of this model. They also indicate the cyclical, time-series nature of these data and the 7-day length of both training and evaluations periods.

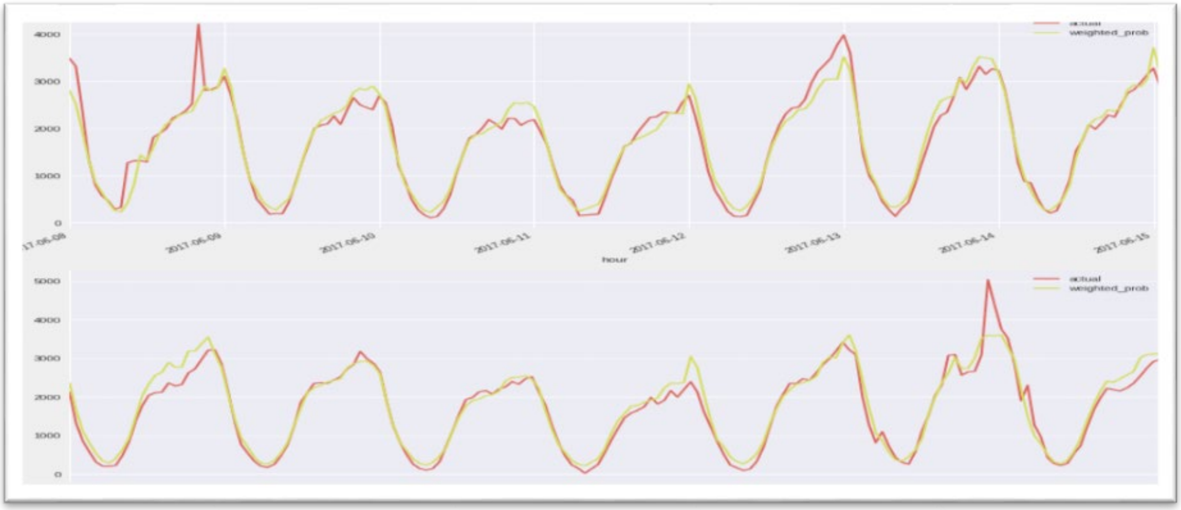

Figure 3 – Aggregated Actual (red) and Predicted (green) Calls

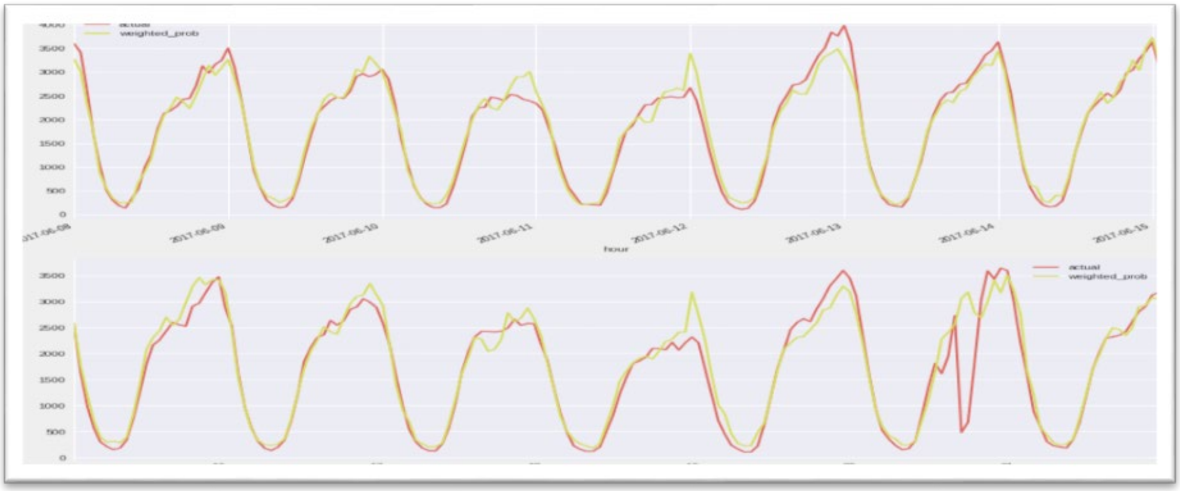

Figure 4 – Aggregated Actual (red) and Predicted (green) Tickets

3.5. Risk Drivers and Incidents

Turning from prediction accuracy to causal correlation, we look again at the drivers identified in Tables 3 and 4, here from a slightly different angle. The X1 Errors remain worthy of attention. For tickets, 20 percent of high-risk Bucket 4 subs correlate strongly with X1 Errors, higher than the other three drivers combined. For calls, 14 percent correlated strongly with X1 Errors, still higher than the next highest driver. (See Table 5.)

Table 5 – Percentage of Incidents Correlating Strongly with Risk Drivers

Risk Drivers	Tickets	Calls
X1 Errors	20%	14%
Previous outage	9%	11%
Previous calls	7%	9%
Device model	3%	2%

3.6. Treating Risk Proactively

The end game of this exercise is proactive measures. The first area we examined was what percentage of at-risk calls and tickets might be saved if any given risk drivers were removed. With the X1 placing high in the previous risk analyses, drivers related to X1's Cross-Runtime Environment (XRE) or possibly the Reference Design Kit (RDK) software stack were logical candidates. Moreover, the focus of this study was not to address outages that occur beyond the control of the operator, rather issues that could be resolved remotely. There will always be previous calls; and device failure is, in part, simply related to device lifecycle.

For both tickets and calls, this learning exercise revealed some top candidates for feature removal. We have anonymized the actual error codes, but if risk drivers A and E were successfully addressed, then 42 percent of calls and 53 percent of tickets would be saved, respectively. (See Tables 6 and 7.)

Table 6 – Calls Saved by Addressing Risk Drivers (percentage)

Risk Drivers	Call saved if feature is removed	Fraction of subs impacted	Call propensity impacted subs
A	42%	91%	0.092
B	31%	95%	0.063
C	14%	100%	0.038
D	12%	100%	0.048

Table 7 – Tickets Saved by Addressing Risk Drivers (percentage)

Risk Drivers	Ticket saved if feature is removed	Fraction of subs impacted	Ticket propensity impacted subs
E	53%	92%	0.128
F	17%	100%	0.039
G	14%	100%	0.043
H	13%	100%	0.039

3.7. X1 Errors and Incidents

Another look at X1 data disclosed more relationships. By extracting the at-risk subscribers for given risk drivers, we discovered that about 20 percent of tickets are highly correlated with XRE errors, and 8 percent highly associated with particular X1 errors, impacting thousands of predicted tickets. We also found that 12 percent of calls were highly correlated with XRE errors, and 5 percent highly associated with particular X1 errors, also impacting thousands of predicted calls.

Removing a high-risk feature affects both those who call or register tickets and those who do not. In the first case, it eliminates the time and effort of contacting customer care; but for everyone it removes a service impediment, arguably improving service. The benefit redounds to the service provider, as well, reducing the number of calls received and a percentage of truck rolls sent to subs who might have simply needed a software patch.

4. Ticket Problem Code Prediction

4.1. One Suspect Set, Many True Positives

A separate exercise in addressing technical problems of subscribers who are likely to call, before they call, involved ticket problem code prediction. Extracting distinct and homogenous ticket problem code classes, we used a clustering algorithm to improve classification. Then over a four-day period, we trained the model, with an evaluation period on the fifth day. Hourly predicted results, aggregated for 1 day, were based on a correlation of X1 errors, outages, and reconnects with ticket problem codes.

The result, once again, reveals that certain risk factors have outsized influence. The overall prediction was 13 ticket classes associated with a large population of true positive tickets per year. From that large sample of tickets emerged three class-related sets (See Table 8):

- Set 1: Eight classes with a single problem code; total tickets covered, eight; true positive tickets predicted per year, 57 percent of total.
- Set 2: Four classes with two problem codes; total ticket problem codes covered, eight; true positive ticket predicted per year, 33 percent of total.
- Set 3: One class with three problem codes; total ticket problem codes covered, 3; true positive tickets predicted per year, 10 percent of total.

In Table 8, we report those classes for which our prediction model achieved the highest precision, irrespective of the associated recall. The tradeoff between precision and recall is adjustable and can be calibrated separately for each of the problem classes. The key takeaway, however, is that a majority (57.4 percent) of the total true-positive tickets are associated with eight ticket classes that have a single code each. Conveyed to the right stakeholders, that kind of insight can help drive both quicker and deeper resolution of issues, reducing average care handling time and costs.

Table 8 – Problem Code Prediction: 13 Classes, 3 Sets of True-Positive Tickets

Predicted ticket classes, problem codes	Precision	Recall	True positives/year
Class 1, code A	98%	21%	Set 1: 57.4%
Class 2, code B	91%	38%	
Class 3, code C	100%	74%	
Class 4, code D	100%	28%	
Class 5, code E	98%	50%	
Class 6, code F	90%	27%	
Class 7, code G	81%	32%	
Class 8, code H	86%	5%	
Class 9, code J, K	99%	20%	Set 2: 33.0%
Class 10, code L, M	83%	37%	
Class 11, code N, P	92%	37%	
Class 12, code Q, R	73%	4%	
Class 13, codes S, T, U	96%	35%	Set 3: 9.6%

4.2. High-Risk Sub and Highly Probable Resolution

This true-positive exercise yielded further insight into proactive problem-solving. The method is to identify those ticket classes for which accurate predictions can be made regarding an appropriate resolution, and then link the fix to an individual subscriber. One example is a provisioned modem for an incorrect boot file, for which our classification analytics found the most likely solution among several for this problem. (See Table 9.)

Table 9 – Proactive Steps: Accurate Predictions on Resolutions

Sub Id	Risk score	Potential ticket problem code	Possible resolution codes
XXX	0.44	Incorrect Boot file	Provisioned modem: 80.96% Customer equipment: 9.21% SIK to customer: 2.32% Excluding Voicemail: 1.71% Reconfigured: 0.12%

Conclusion

The cable industry has invested heavily in customer care and frontline technical support. But so far, we have yet to get very far ahead of the subscriber – and know who is going to call before they call, and why. One development capable of changing the game is the combination of machine learning and rich operational data, such as error data from the IP-enabled X1 set-top box.

As our results from the POC exercise show, there is potential for considerable insight and follow-up actions. This falls into three domains:

A. We found that 5 percent of subscribers, in an identifiable high-risk category, are driving 36 percent of all technical calls and 46 percent of all technical tickets. We were also able to associate 57 percent of total true-positive, predicted tickets with one large set of ticket classes notable for having a single problem code.
B. For a significant number of the devices exhibiting errors, we were able to identify the likely diagnosis that would have been made by the agent, as reflected in the ticket if that person sought support. This information can be passed to agents to reduce their call times and better inform the actions taken, thus improving customer experience and making more efficient use of internal resources.
C. It is not a stretch to say that these findings – if shared among leaders in customer care, finance, quality engineering and other teams – could lead to coordinated strikes against the leading risk drivers and preemptive actions. The outcome, again, could be notable reductions in operating costs (calls, truck rolls, customer care, etc.) and increased customer satisfaction.

Bibliography & References

Cunha, G. Approaches for Proactively Managing Customer Experience and Reducing OPEX in a Cable Operations Environment. SCTE-ISBE 2017.

Dorairaj, S., and Chris Bastian, Bernard Burg, and Nicholas Pinkernell. Simplifying Field Operations using Machine Learning. SCTE-ISBE 2017.

Sundaresan, K., and J. Zhu. Access Network Data Analytics. SCTE-ISBE, 2017.

Sundelin, A. Leveraging Machine Intelligence and Operational Analytics to Assure Virtualized Networks and Services, SCTE-ISBE, 2017.