Extracting information for failure prediction from intermittent data

Balakrishna S Kesavan[1] and Amol B Mahamuni[2]
[1] Principal Data Scientist, IBM India
[2] Program Director, IBM India

e-mail: balakrishna.kesavan@ibm.com

Abstract. The edge devices used for edge analytics usually record data intermittently. So, it is hard to extract patterns of device failures, using common temporal approaches, like Long Short Term Memory, or multivariate models, like boosted trees. Survival analysis can produce actionable insights in such scenarios.

Key words. Edge analytics, survival analysis, failure prediction, predictive maintenance

1. Introduction:

With Industry 4.0 gaining acceptance, there will be an increase in the number of intelligent devices, like Industrial IoT (IIoT) devices, that will be used at the edge locations of businesses. Many use cases for edge analytics have been identified at factories, retail stores, warehouses, hotels, distribution centers, and in vehicles. In all of these, the edge device itself will need predictive maintenance, so that the intended use case is reliably supported.
One challenge we face is that edge devices record observations intermittently. This could be because these devices are powered down during non-working hours or they may have limited storage space etc.

2. The challenge – intermittent data:

To explain the problem and proposed solution better, we use S.M.A.R.T. (Self-Monitoring, Analysis and Reporting Technology) data about Hard Disk Drive (HDD) failures, hosted for public use by Black Blaze[1]. The SMART dataset has 3+Mn observations, covering 69 HDD models, of which only 215 observations show HDD device failures. The observations were recorded once a day starting from 1st Jan 2016 to 29th Apr 2016. There are 94 features (sensor readings), in addition to the 'failure' flag. The data contributors, Black Blaze, have found the following features, in Table 1, to be useful for failure prediction[2]. This is an important point of collaboration between device engineering and data science.

SMART 5		Reallocated Sectors Count
SMART 187		Reported Uncorrectable Errors
SMART 188		Command Timeout
SMART 197		Current Pending Sector Count
SMART 198		Uncorrectable Sector Count

Table 1. Features with engineering relevance to failure prediction

To simplify this discussion, we subset the data by picking only 4 models, 'Hitachi HDS5C4040ALE630', 'WDC WD800BB', 'WDC WD800AAJS' and 'WDC WD3200BEKT'.

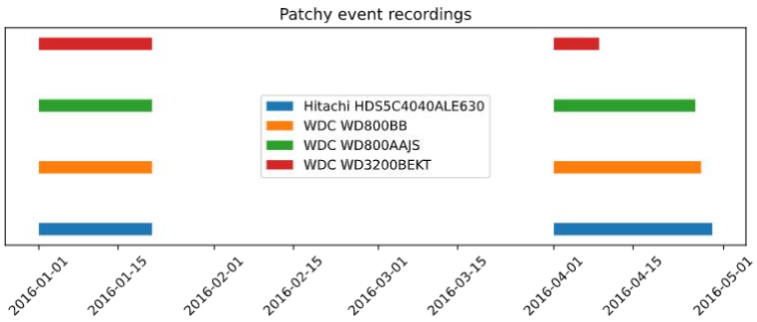

Fig 1. Intermittent observations

As seen, in Fig 1, data has been recorded in an irregular manner that varies by device. All devices' data have been recorded for a couple of weeks in January. Then, nothing is recorded till April when, data is recorded for some devices. But not for the same days.

3. Temporal analysis:

Temporal models, like Long Short Term Memory (LSTM), try to detect patterns of sensor readings that indicate device degradation and impending failure. For example, as the value of SMART_5_Reallocated Sectors Count goes up, the probability of device failure goes up[3].

But with the data being recorded intermittently for varying durations, as seen in Fig 1, it is impossible to even set a reasonable time 'window'/ timesteps (variable N in Fig 2) for an LSTM. For example, if we picked N = 14 days (refer Fig 1), it might work in the first couple of weeks of the dataset. But after that, data is missing for nearly 3 months. It would be wrong to impute missing data when the duration of missing data is longer than available data. This obstacle applies to other time series analysis techniques as well.

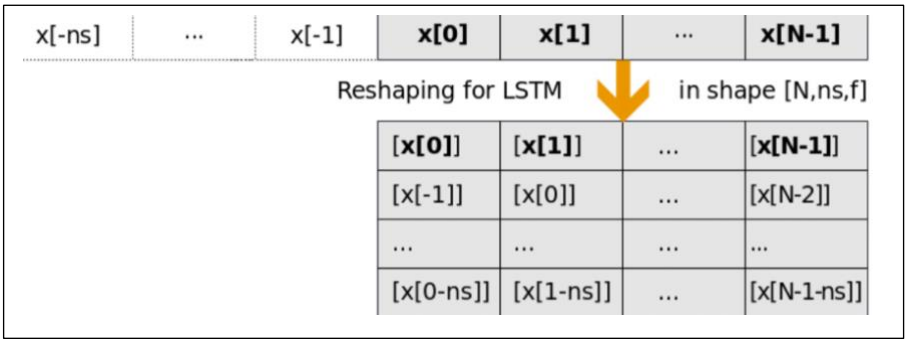

Fig 2. Reshaping timeseries data for LSTM

4. Multivariate analysis:

Multivariate models, like boosted trees and Cox Proportional Hazard were built. The full dataset was used, and the device model names were included as dummy features. The boosted tree showed poor fit with cross validated f1 score mean = 0.229, standard deviation = 0.043. (The target f1 score is 1.0). Imbalanced classes (very few failures) is the likely cause. Imbalanced class treatment techniques, like Synthetic Minority Oversampling Technique (SMOTE), could potentially inject fake patterns and were eschewed.

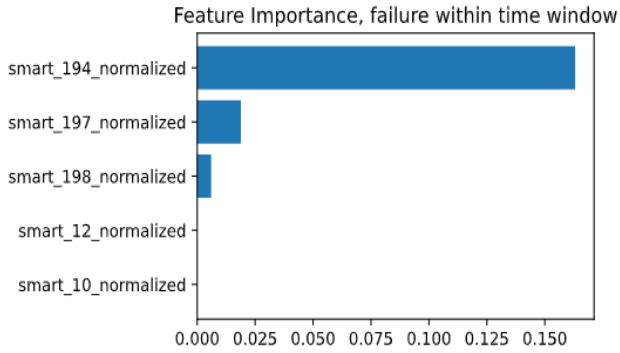

Fig 3. Near zero feature importance in boosted trees

The feature importance plot in Fig 3, from the boosted tree model, shows that even the top 5 features have very poor contribution, explaining the low f1 score. Only 2 of the features that Black Blaze suggested, for hard disk engineering reasons, appear here (smart 197 and smart 198). Smart 194 has 77 distinct values compared to 6 & 5 for smart 197 and smart 198 respectively. This is the likely reason for the inclusion of this feature, which was already flagged as non-informative by Black Blaze.

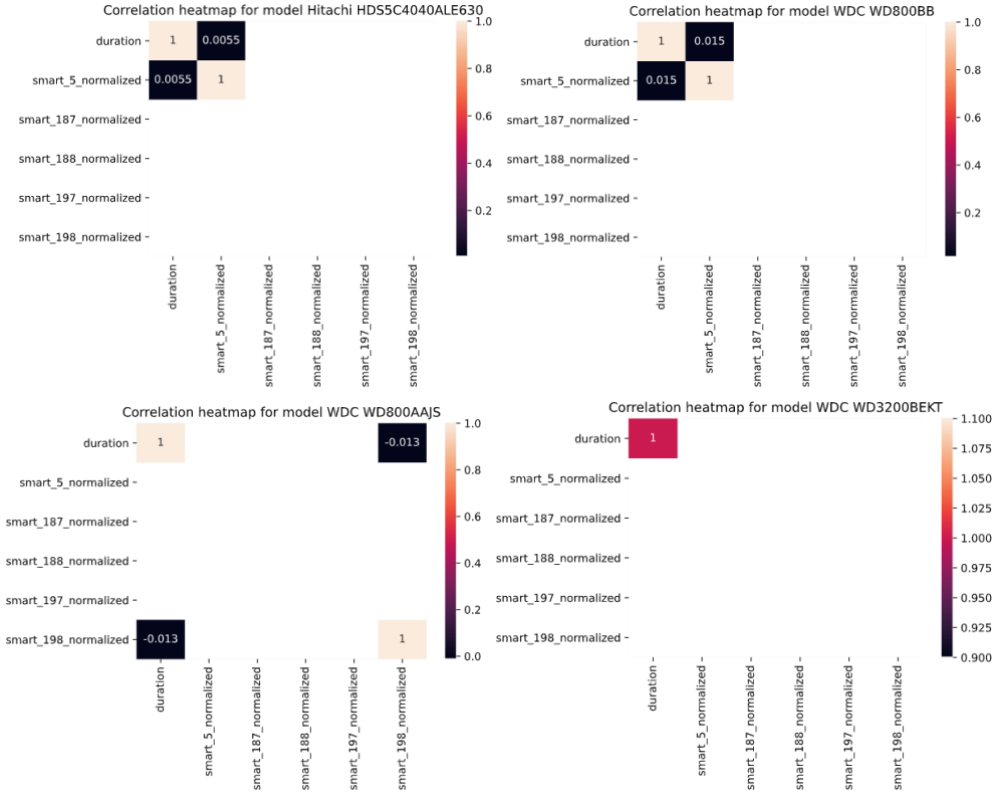

Fig 4. Correlation heatmap between suggested features to the time to failure (duration)

Survival regression using Cox Proportional Hazard was also tried but produced the error “ConvergenceError: Convergence halted due to matrix inversion problems. Suspicion is high collinearity”. But as seen in Fig 4, there isn’t any such collinearity. So, the more likely reason is that the features were not informative, as already seen in Fig 3.

Some information can be inferred from just the device failure observations in the data, as seen in Fig 5. For example, for the model "Hitachi HDS5C4040ALE630", there were 6 failures out 15,814 observations. Since 4 of these failures occurred on day zero (refer Table 7 for failure events), the median is 0 days and the mean value of age at failure was 3.5 days. However, these values cannot be generalized to the devices that are yet to fail. We cannot expect more than half of all "Hitachi HDS5C4040ALE630" devices to fail on either day 0 or 4 (the median and mean values of failure duration).

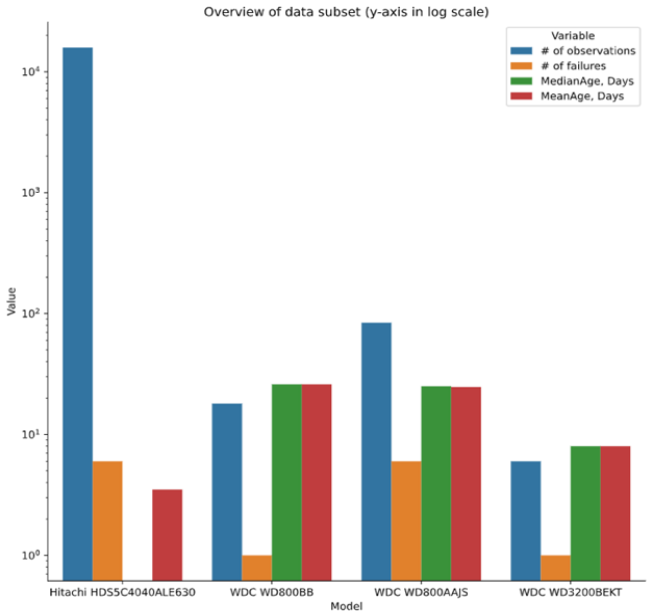

Fig 5. Overview of the data for four selected models (y-axis in log scale)

5. Squeezing out insights using survival analysis:

Using survival analysis, a univariate approach that uses the Kaplan Meier algorithm, we can still obtain usable insights out of this dataset.

serial_number	start	end	model	failure	duration
WD-WX71A9290300	2016-01-01	2016-01-01	WDC WD3200BEKT	0	0
WD-WX71A9290300	2016-01-01	2016-01-01	WDC WD3200BEKT	0	0
WD-WX71A9290300	2016-01-01	2016-01-21	WDC WD3200BEKT	0	20
WD-WX71A9290300	2016-04-01	2016-04-01	WDC WD3200BEKT	0	0
WD-WX71A9290300	2016-04-01	2016-04-01	WDC WD3200BEKT	0	0
WD-WX71A9290300	2016-04-01	2016-04-09	WDC WD3200BEKT	1	8

Table 2. Partially processed observations for the model "WDC WD3200BEKT"

Consider the partially processed observations for the model "WDC WD3200BEKT" in Table 2. There are only 6 of them in the dataset, of which only the last observation shows a failure.

WDC WD3200BEKT:

event_at	removed	observed	censored	entrance	at_risk
0	4	0	4	6	6
8	1	1	0	0	2
20	1	0	1	0	1

Table 3. Survival for the model "WDC WD3200BEKT"

Using the observations from Table 2, a survival table, in Table 3, extracts information about how many devices of this model were being tracked ("at risk") at different points in time ("event_at", in days), how many of these recorded no failures during or at the end of the study ("censored") and how many failed ("observed"). Survival analysis uses information even from devices that never failed and hence, were "censored". Instead of discarding observations relating to devices that didn't fail, censoring allows using all available information. That is, the fact that all 6 devices survived day 0 and 1 device has survived ***past*** day 20 is used in survival probability computation.

The survival probability[4, 5] is calculated using this formula where 'n' is the number of devices at risk ('at_risk') just before time 't'('event_at') and 'd' is the number of deaths ('observed') at 't'.

$$\hat{S} = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$$

Day	Number of devices at risk	Number of devices that failed	Survival probability (product of terms)
0	6	0	(6-0)/6 = 1
8	2	1	{(6-0)/6}x{(2-1)/2} = 0.5
20	1	0	{(6-0)/6}x{(2-1)/2}x{(1-0)/1} = 0.5

Table 4: Survival probability calculations (Refer Table 3 for the variable values)

It is important to note that censoring is assumed to be non-informative and that censored devices are not excluded because of increased risk of failure.

It is also possible to extrapolate past our observed durations by using parametric models, like Weibull. But this is not covered here.

Insight 1 – survival probabilities for a device model: The survival function plot, in Fig 6, provides an easy to interpret view of survival probabilities. For example, for model "WDC WD3200BEKT" the probability of survival is 50% on the 10th day. Using this information, predictive maintenance schedules can be set. For example, inspect/ maintain/ replace components before probability of survival drops below 50%.

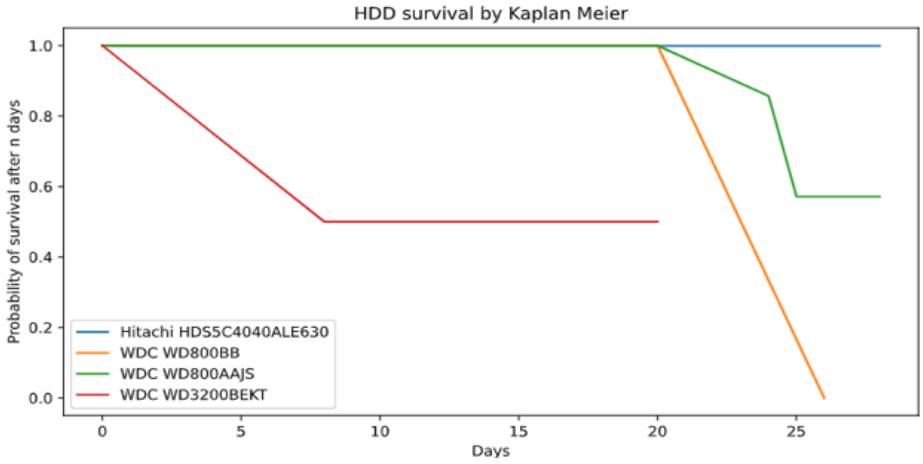

Fig 6. Survival function plot showing survival probability over time

Insight 2 – comparing device models at a future point in time: The Restricted Mean Survival Times (RMST), shown in Fig 7, another measure from survival analysis, is the area under the survival curve up to a prespecified time horizon.

All else being equal, the model with the larger RMST has the higher probability of surviving at a given point in time (day 28 in this case). While the RMST for "WDC WD800BB" and "WDC WD800AAJS" are deceptively close, the plot shows that some "WDC WD800AAJS" devices are still running whereas no more "WDC WD800BB" devices are. Cross check with details in Table 7. If these device models were equivalents, the expected duration of operation before testing/ replacement can help choose the model with lower maintenance costs. (Use a log rank test to test if two survival curves are statistically different).

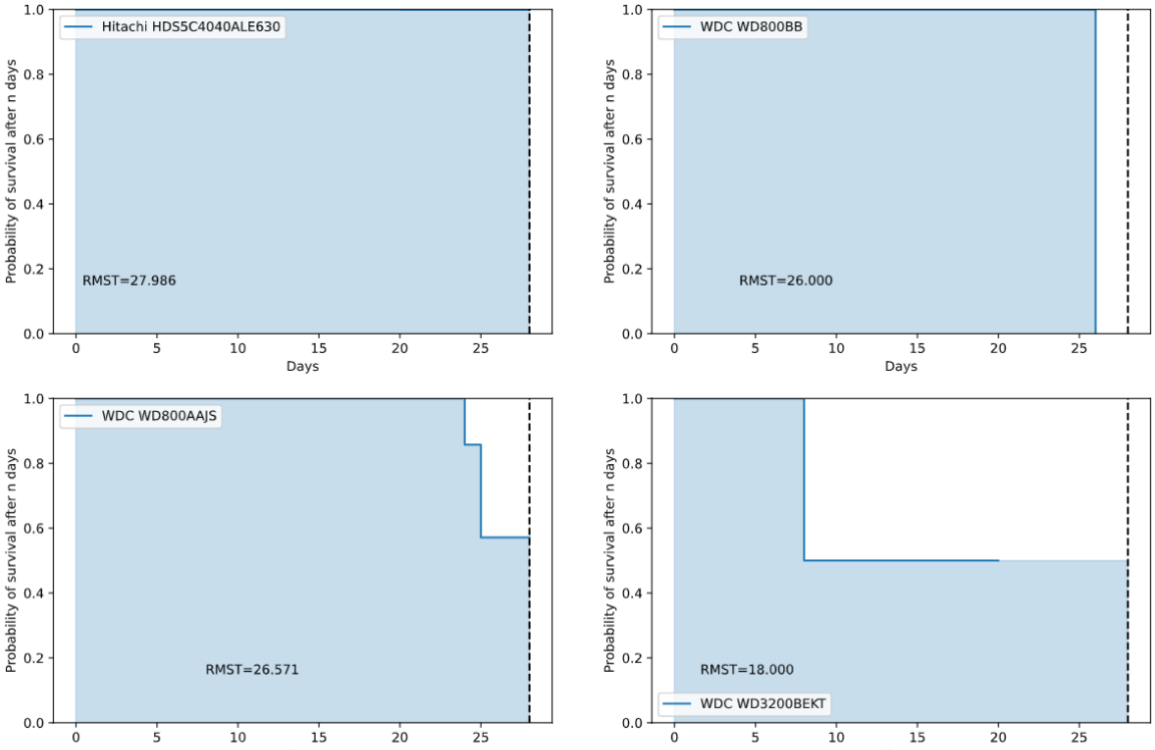

Fig 7. Comparing survival of models at a future point in time

Insight 3 – Confidence in estimates. Continuing the above comparison between "WDC WD800BB" and "WDC WD800AAJS", the 95% confidence interval around the median survival time is instructive. We can tell for sure that more than half the "WDC WD800AAJS" devices were running on day 26. But we can't say the same for "WDC WD800BB". See Table 5.

0	Lower CI	Median Survivial Time, Days	Upper CI
Hitachi HDS5C4040ALE630	inf	inf	inf
WDC WD800BB	26.0	26.0	26.0
WDC WD800AAJS	25.0	inf	inf
WDC WD3200BEKT	8.0	8.0	inf

Table 5. Median survival time from survival analysis

But by considering only the failure observations, in Table 6, we underestimate the median & mean ages. As stated earlier, we cannot expect more than half of all "Hitachi HDS5C4040ALE630" devices to fail on either day 0 or 4 just because 6 out of 15,814 devices failed by then.

model	# of observations	# of failures	MedianAge, Days	MeanAge, Days
Hitachi HDS5C4040ALE630	15814	6	0.0	3.500000
WDC WD800BB	18	1	26.0	26.000000
WDC WD800AAJS	84	6	25.0	24.666667
WDC WD3200BEKT	6	1	8.0	8.000000

Table 6: Median survival time from failure observations

In the case of model "WDC WD800BB" of 18 devices, 17 were censored and at the end, 1 failed on the 26th day. Hence the single median survival time estimate. In the case of "WDC WD800AAJS", since less than half the number of devices at risk at any point in time failed, the median cannot estimated. Refer Table 5 and Table 7.
For "WDC WD320BEKT", since half the devices at risk have failed, a median can be estimated but since one more is still running, upper bound cannot be estimated.

6. Conclusion:

We saw that due to the intermittent data, temporal analysis approaches could not be used to extract patterns of failure. We also saw that multivariate approaches, like boosted trees survival regression are not helpful when the classes are imbalanced, and the features are not informative. However, survival analysis is able to provide actionable insights, when other methods cannot.

These are the events,by model, by time (column "event_at"):

****

Hitachi HDS5C4040ALE630:

event_at	removed	observed	censored	entrance	at_risk
0	10544	4	10540	15814	15814
1	1	1	0	0	5270
20	2638	1	2637	0	5269
26	1	0	1	0	2631
28	2630	0	2630	0	2630

WDC WD800BB:

event_at	removed	observed	censored	entrance	at_risk
0	12	0	12	18	18
20	5	0	5	0	6
26	1	1	0	0	1

WDC WD800AAJS:

event_at	removed	observed	censored	entrance	at_risk
0	56	0	56	84	84
20	14	0	14	0	28
24	2	2	0	0	14
25	4	4	0	0	12
28	8	0	8	0	8

WDC WD3200BEKT:

event_at	removed	observed	censored	entrance	at_risk
0	4	0	4	6	6
8	1	1	0	0	2

Table 7. Full survival table for the four selected models

7. References:

[1] Data source: https://www.kaggle.com/backblaze/hard-drive-test-data

[2] Black Blaze feature recommendation: https://www.backblaze.com/blog/what-smart-stats-indicate-hard-drive-failures/

[3] Source: https://en.wikipedia.org/wiki/S.M.A.R.T.

[4] "Survival Analysis A Self-Learning Text" by David G. Kleinbaum and Mitchel Klein, published by Springer.

[5] Python Lifelines: https://lifelines.readthedocs.io/en/lat est/index.html