Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Interpretable predictive maintenance for hard drives

Maxime Amram, Jack Dunn [*], Jeremy J. Toledano, Ying Daisy Zhuo

Interpretable AI, Cambridge, MA 02142, USA

ARTICLE INFO

Keywords:
Predictive maintenance
Decision trees
Interpretable machine learning

ABSTRACT

Existing machine learning approaches for data-driven predictive maintenance are usually black boxes that claim high predictive power yet cannot be understood by humans. This limits the ability of humans to use these models to derive insights and understanding of the underlying failure mechanisms, and also limits the degree of confidence that can be placed in such a system to perform well on future data. We consider the task of predicting hard drive failure in a data center using recent algorithms for interpretable machine learning. We demonstrate that these methods provide meaningful insights about short- and long-term drive health, while also maintaining high predictive performance. We also show that these analyses still deliver useful insights even when limited historical data is available, enabling their use in situations where data collection has only recently begun.

1. Introduction

Backblaze is a large data storage service provider with over 130,000 hard drives in its data centers. Hard drives have a number of sensors that are continuously monitoring and reporting on the health of the drive, known as SMART (Self-Monitoring, Analysis and Reporting Technology). Backblaze regularly publishes the historical daily snapshots of the SMART statistics for each of its hard drives (Backblaze, 2020a). Additionally, they record whenever a hard drive fails or is removed from service.

The ability to understand and predict when a hard drive is going to fail is extremely valuable to the operations of such a data center, as it offers the ability to avoid downtime through better timing for replacement and repair of equipment (preventive maintenance), as well as providing more input into strategic planning and forecasts of operational needs. Based on univariate analysis and domain knowledge, Backblaze (2014) identified five SMART metrics (numbers 5, 187, 188, 197, and 198) that it uses as indicators of impending failure (these findings will be presented in more detail in Section 3.)

The use of machine learning methods for such predictive maintenance tasks is growing in popularity, particularly with the rapid increase in deployment of IoT (internet of things) devices and sensors. While humans have trouble considering more than a few variables at a time, machine learning and artificial intelligence methods are capable of handling thousands of variables at a time when building predictive models. These approaches are more powerful than simple univariate analyses as they can model complex interaction effects between features rather than simply considering each variable in isolation.

Despite their increased power over simpler approaches, these machine learning methods are often ‘‘black-box’’ approaches whose prediction logic is difficult or impossible for humans to comprehend. This is undesirable, as although we may be able to build a predictive model that can predict drive failure on historical data with high accuracy, it is difficult to extract insights about how the features interact to indicate impending failure, and also hard to have full confidence that such a system will continue to perform well if deployed in production.

An active area of research is interpretable machine learning methods, which unlike black-box methods are completely understandable by humans. Recent work has demonstrated that modern optimization techniques can be leveraged to construct interpretable methods with performance rivaling the black-box methods, enabling us to leverage the insights and confidence that interpretability brings without sacrificing performance (Bertsimas & Dunn, 2019).

The main goal of our analysis is to develop a data-driven model for understanding how the SMART metrics relate to impending failure. Given the desire to learn and understand how these metrics can indicate failure, it is natural that we consider approaching this task with interpretable models. In particular, in this paper we concern ourselves with understanding the following two dimensions of hard drive health:

- the long-term health of a hard drive over its expected lifespan (roughly 3 years)
- the short-term health of a hard drive and whether it is likely to fail in the immediate future (30 to 90 days)

We also consider these problems from the perspective of limited data. In the case of Backblaze, there are many years of rich data

[*] Corresponding author.
E-mail addresses: maxime@interpretable.ai (M. Amram), jack@interpretable.ai (J. Dunn), jeremy@interpretable.ai (J.J. Toledano), daisy@interpretable.ai (Y.D. Zhuo).

https://doi.org/10.1016/j.mlwa.2021.100042
Received 6 November 2020; Received in revised form 25 March 2021; Accepted 14 May 2021
Available online 21 May 2021

available that we can use for modeling. However, we would like to understand whether we can use data from a limited timespan to conduct the same analysis and arrive at similar conclusions. Such a result would prove useful in situations where sensors have only recently been installed, permitting this kind of analysis without waiting years to collect sufficient data.

We summarize our contributions in this paper:

- Using data across a three-year time horizon, we build an Optimal Survival Tree (OST) to model the long-term health of drives. We observe that the resulting model places importance on metrics known to be correlated with drive failure, as well as additional features not previously highlighted. Moreover, the model demonstrates that these features interact in specific ways when indicating the outlook for drives, where some metrics are only useful for healthy drives, while others are relevant when examining drives with high risk of imminent failure.
- Using data from a shorter time horizon, we build both OST and Optimal Classification Tree (OCT) models to predict the risk of short-term failure (30–90 days). These models again surface feature interactions that are highly predictive of drive failure, and both have strong predictive performance (sensitivity approximately 50% with a false alarm rate around 10%).
- Finally, we demonstrate that these analyses can be applied to deliver similar insights even in scenarios where a wealth of historical data is not available. This is particularly significant as it is often the case that sensors have only recently been installed, limiting how much data is available. We show that even in this data-poor case, our approach can still be leveraged to deliver useful insights.

The structure of the paper is as follows. In Section 2, we review the classical approaches used in the literature for these problems. In Section 3, we give an overview of the data published by Backblaze, as well as their existing findings. In Section 4, we consider the problem of predicting long-term health using OST. In Section 5, we use OCT and OST to predict short-term failure. In Section 6, we repeat our analysis using data from a limited time frame to simulate the case where limited historical data is available. Finally, in Section 7 we present a summary of our findings.

2. Literature review

2.1. Overview

Machine learning has become a popular approach for solving predictive maintenance problems in the literature (see Carvalho et al., 2019 for a comprehensive review). When data on the occurrence of failures is available, *supervised* machine learning is typically the preferred approach, as they tend to be more powerful than *unsupervised* approaches, which learn purely from the sensor data and do not require failure labels (Kanawaday & Sane, 2017). In some cases, the occurrence of failure may not be recorded, or failures may not be observed (e.g. if machines are rotated out of service before failure or preemptively maintained), necessitating the use of unsupervised methods. In our case, failure data is available so we focus on supervised approaches.

Supervised machine learning uses data comprised of pairs of the form $(\mathbf{x}_i, y_i), i = 1, \ldots, n$. Each $\mathbf{x}_i$ is a p-dimensional vector containing the measurements for each sensor (in our case, the daily SMART values), and y_i represents the outcome variable. Depending on the choice of outcome variable, we can generate different classes of problem:

- y_i representing a binary or categorical outcome (e.g. whether the drive failed within 30 days) gives a *classification* problem
- y_i representing a numerical outcome (e.g. the time until the drive failed) gives a *regression* or *survival* problem

Many different algorithms can be used to solve a classification problem, classical approaches include both interpretable methods like logistic regression and decision trees (CART), as well as black-box methods like random forests and gradient boosting (Friedman et al., 2001). Traditionally, black-box methods have outperformed interpretable methods in terms of predictive performance, resulting in a significant price of interpretability and a difficult tradeoff for practitioners. Recent works combining machine learning with modern mathematical optimization have developed Optimal Classification Trees (Bertsimas & Dunn, 2017, 2019) which permits constructing a single decision tree that has similar performance to the black-box methods, thus delivering interpretability without sacrificing performance.

When it comes to predicting continuous outcomes such as the remaining useful life, a standard regression analysis is often not well-suited to solving the problem. This is because it is common that the majority of the records in the data have not experienced a failure, and thus we have no measure of their remaining useful life. However, we can still use these data for learning by exploiting the fact that we have observed these machines without failure for some time, and using this to form lower bounds for remaining lifespan. For example, if we have observed a machine without failure for 90 days, then we know that as of the first day, it has a remaining useful life of *at least* 90 days. Survival analysis is a specialized class of models designed to deal with such so-called *censored* data. A classical survival analysis tool is the Cox proportional hazards model, which models impact of features on the lifespan using an approach analogous to linear regression (Cox, 1972). Another traditional approach is the Kaplan–Meier estimator, which is a non-parametric model of survival over time (Kaplan & Meier, 1958). Decision tree models can also be applied to survival problems for enhanced power and interpretability.

2.2. Optimal survival trees

Just as Optimal Classification Trees improve significantly over traditional tree methods for classification problems, Optimal Survival Trees (Bertsimas et al., 2020) use modern optimization approaches to improve significantly over existing greedy survival tree methods (LeBlanc & Crowley, 1992) while maintaining interpretability.

Optimal Survival Trees, instead of making a single prediction on the probability of failure over a pre-defined period, estimate the entire survival probability over any time period. In addition, another major benefit is that this method accounts for censored outcomes instead of removing them altogether, presenting a more efficient use of data when we do not have enough time to observe failures in all hard drives.

The detailed problem formulation can be found in Bertsimas et al. (2020), to which we refer the reader for full details. At a high level, the problem aims to find best splits and hazard functions in each leaf node such that the within-leaf sample likelihood is maximized. The Optimal Trees framework presented in Bertsimas and Dunn (2019) is then used to solve this optimization problem holistically with a view towards global optimality.

The Optimal Trees framework automatically tunes the tree to optimize the tradeoff between training error and tree complexity. Thus, when training Optimal Survival Trees the only key parameter that need to be selected is the maximum depth of the tree.

3. Overview of the Backblaze data and insights

In this section, we present an overview of the SMART data collected and published by Backblaze. We also review their existing research and efforts to link SMART metrics to hard drive failure.

Table 1
Descriptions of relevant SMART metrics.

SMART	Attribute name	Description
3	Spin-up time	Average time (in milliseconds) of spindle spin up from zero RPM to fully operational.
5	Reallocated sectors count	Count of bad sectors that have been found and reallocated. A hard drive which has had a reallocation is very likely to fail in the immediate months.
7	Seek error rate	Rate of seek errors of the magnetic heads, due to partial failure in the mechanical positioning system.
187	Reported uncorrectable errors	Count of errors that could not be recovered using hardware ECC (Error-Correcting Code), a type of memory used to correct data corruption errors.
188	Command timeout	Count of aborted operations due to hard drive timeout.
190	Temperature difference	Difference between current hard drive temperature and optimal temperature of 100°C.
197	Current pending sectors count	Count of bad sectors that have been found and waiting to be reallocated, because of unrecoverable read errors.
198	Offline uncorrectable sectors count	Total count of uncorrectable errors when reading/writing a sector, indicating defects of the disk surface or problems in the mechanical subsystem.

3.1. Backblaze data

Each of Backblaze's 130,000 hard drives is monitored daily and its activity is recorded as a single row in a daily data file. Every day, for each hard drive, Backblaze records several daily aggregated SMART metrics as well as whether it failed on this day. Each attribute measures different characteristics about the drive's operation, and we present a summary of the most relevant quantities in Table 1.

Each SMART attribute has a raw value whose meaning is determined entirely by the drive manufacturer (but often corresponds to the raw physical unit that is being measured), as well as a normalized value between 1 and 253, where 1 is worst, and 100 is usually the starting value (Wikipedia, 2020). Fig. 1 shows the mapping between the normalized and the raw measurements for SMART metrics 197 and 7. We see that for metric 197 the raw values map directly to the normalized values, whereas for metric 7 there is little correlation between the raw and normalized metrics. Although most metrics in the dataset show a similar pattern to 197 with a clear mapping between the values, there are many metrics that, like metric 7, show little correlation between the values. As a result, in our analysis we will opt to retain both raw and normalized values in case having these different signals results in more predictive power.

At the end of Q1 2020, Backblaze had recorded a total of 7033 hard drive failures since April 2013 for the 130,000 hard drives still in service in 2020, for an annualized failure rate of 1.71% (Backblaze, 2020b).

Backblaze notes that there is significant variation in failure rates across different models of hard drive. Coupled with the potential for inconsistent meanings across models and manufacturers, we chose to limit our analysis to a single model of hard drive. Fig. 2 shows the number of failures per model for Q1 2020. We see that model ST12000NM0007 accounts for roughly 30% of the overall failures. Throughout the paper, we will thus focus our analysis on this specific hard drive model to maximize the amount of available data.

Hard drives are organized into "storage pods", which are distinct servers each with 60 hard drives. Backblaze labels a hard drive as failing whenever it is removed from its storage pod (Backblaze, 2014). There are two possible reasons for such a removal:

1. The hard drive has stopped working: it is impossible to turn it on, it does not respond to console commands, or it cannot be read or written anymore.
2. The hard drive has been deemed "about to fail" (indicated by a positive value for SMART metric 187, which means that the hard drive has started experiencing uncorrectable errors).

It is necessary for Backblaze to anticipate such failures as a single failing hard drive could impede the whole storage pod's operation.

3.2. Existing analysis of SMART data

Backblaze

In an effort to understand how SMART statistics relate to drive failure, Backblaze (2014) ran a simple univariate correlation analysis between each SMART metric and their failure records. Based on this analysis, coupled with their own domain knowledge, they concluded that the following SMART metrics are good predictors of impending failure:

- 5 (Reallocated Sectors Count)
- 187 (Reported Uncorrectable Errors)
- 188 (Command Timeout)
- 197 (Current Pending Sectors Count)
- 198 (Offline Uncorrectable Sectors Count)

Additionally, they claim these metrics have the benefit of generally being consistent across hard drive manufacturers.

Google

Pinheiro et al. (2007) conducted a detailed correlation analysis of SMART metrics for a large number of hard drives in operation at Google. We summarize their findings as follows:

- Four SMART metrics are consistently highly correlated with failure: 5, 187, 197 and 198.
- Six additional SMART metrics were found to be relevant in predicting failures, but are not always consistent across models and manufacturers, including 7 (Seek Errors).
- They find that any change in metric 187 is highly predictive of failure: "after their first scan error (i.e. when a positive value for 187 is observed for the first time), drives are 39 times more likely to fail within 60 days than drives with no such errors".
- Temperature, which is often considered the most important factor in predicting failures by many in the industry, was not found to be positively correlated with failure in relatively young hard drives.

Additional machine learning analyses on SMART data

More broadly speaking, there is ample research on using SMART data to predict hard drive failures. In Succi et al. (2018), the authors reviewed a broad range of techniques used on SMART data, including recursive feature elimination to rank and select most important features, and support vector machines, k-nearest neighbors, random forests, and classification and decision trees for predictions. Studies such as Pitakrat et al. (2013), Rincón et al. (2017), Xiao et al. (2018) further investigate the relative performance among multiple machine learning methods.

In many of these studies, Decision Trees were the best performing method (Rincón et al., 2017). In particular, in the area of decision-trees, studies such as Li et al. (2014), Li et al. (2017) used CART on SMART data, which was shown to outperform models such as neural networks.

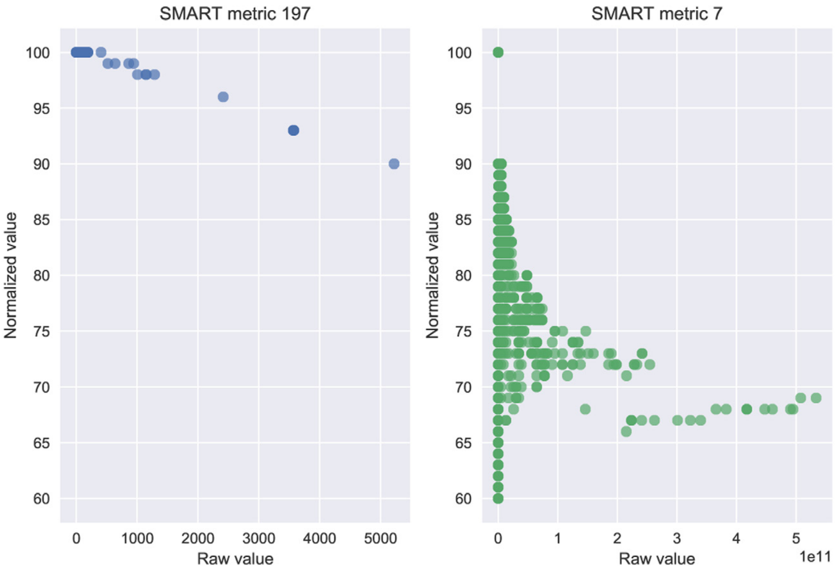

Fig. 1. Normalized values as function of the raw values for metrics 197 and 7.

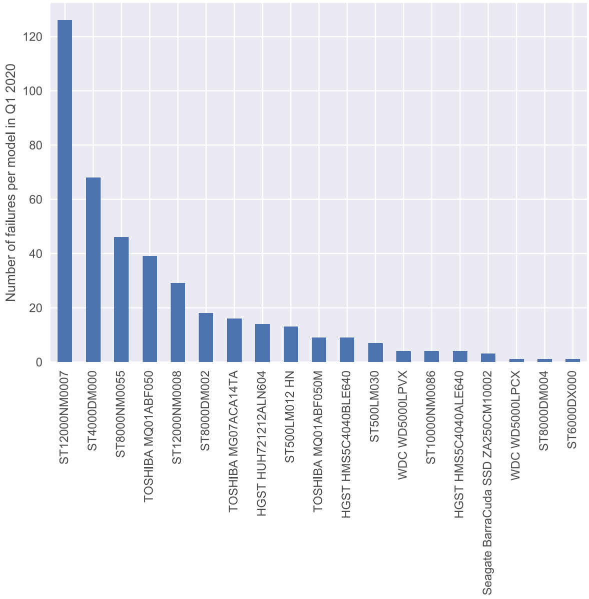

Fig. 2. Number of failures per model in Q1 2020.

Limitations

The analyses from Backblaze and Google were univariate and only considered correlation between failures and a single metric at a time. As such, they would not be able to detect any nonlinear interactions between metrics that affected the chance of failure. Another limitation of this analysis is that it leaves humans to choose the cutoff values that will raise alerts if exceeded.

While some of the methods in the additional machine learning-based analyses using SMART data were interpretable such as CART, the majority such as random forest and neural networks were not. This hinders the practical usefulness of the predictions as they cannot be readily understood by humans, but the interpretability of CART often comes at a significant cost in terms of performance.

However, Bertsimas and Dunn (2017) established that Optimal Trees maintain the interpretability of CART while achieving performance competitive with black box methods, due to its global optimization perspective. This motivates revisiting the previous analyses to construct an approach that is both interpretable and highly-performant.

Equally importantly, the existing studies answer the question of predicting failures over a pre-defined time period. There is very little research in characterizing the short-term and long-term health of hard drives over their entire lifespan, which is useful for making operational decisions, such as when to swap out equipment and when to conduct maintenance. Optimal Survival Trees, as introduced earlier, address this question and have the ability to provide additional insights that a traditional classification approach to this problem might not.

4. Predicting long-term health

In this section, we address the problem of trying to predict the overall health of a hard drive over a long time horizon using the SMART metrics.

4.1. Data processing

As stated earlier, our analysis is restricted to model ST12000NM0007. We isolated hard drives that failed between Q1 2019 and Q1 2020 (inclusive), and gathered their daily SMART metrics records for the past three years (back to Q1 2017). This data captured the entire lifespan for the majority of these drives, indicating we have data on their full trajectory since being put into service. There were a total of 172,104 observations for this model in the dataset during the time period of interest. We used a random subset of 50,000 observations for training and the remainder as testing.

To keep our analysis interpretable and intuitive, we used only the raw and normalized SMART metrics captured in the data, and did not engineer any more complex features. Additionally, to allow our predictive models to focus on characterizing the health of the drive, we removed SMART metrics which represent cumulative counts over the lifespan of the drive and thus are highly correlated with time (4, 9, 12, 192, 193, 240, 241, and 242). This allowed us to remove from the model the obvious result that older drives are more likely to fail, and as a result we can uncover more interesting predictive insights based on metrics that are more related to the overall health status of the drives, regardless of age.

For each hard drive and for each daily snapshot, we compute the *remaining useful life*, being the number of days until failure. Note that because we captured the complete lifespan for most of these drives, we have very little censored data.

4.2. Models

We used the daily SMART feature records coupled with remaining useful life values to train an Optimal Survival Tree that models how the remaining life is affected by SMART values. The trees were cross validated with maximum depth between 2 and 4. The resulting model is a single tree that uses the SMART metrics to segment the data into subpopulations with similar survival prognosis.

The analysis in this section and following sections were conducted on a Macbook Pro 3.3 GHz Intel Core i7 with 16 GB of memory. Model training takes less than 1 h on this single machine.

We show the trained survival tree in Fig. 3. Starting from the top of the tree, the split nodes use one SMART feature at a time to recursively divide the hard drives. Each leaf node displays the expected survival time among hard drives in that leaf, with a darker color shading indicating longer survival. Each leaf node is also characterized by a Kaplan–Meier curve, showing the survival outlook over time for hard drives that fall into that leaf. Fig. 4 shows examples of these curves for two of the leaves in the tree (the full tree showing the Kaplan–Meier curves in each node is available in the supplementary materials).

We see that three out of the five SMART metrics identified by Backblaze as indicative of impending failure (namely 5, 187, and 197) are used by our model. In particular, the tree begins splitting on 5, indicating this is the most impactful metric to consider before following up with subsequent questions.

If metric 5 is below 1.5, we go to the left side of the tree, where survival tends to be higher, and subsequent splits are based on metrics 3 and 7. If on the other hand metric 5 is above 1.5, we go to the right side of the tree, with lower survival times that are refined by examining metrics 3, 5 (again), 187 and 197. This dichotomy shows the power of our advanced modeling over simple univariate analysis: we are able to identify cases where the features interact to affect survival differently, and some metrics only become relevant based on the values of others.

From the model we can also derive paths to failure that showcase how, under certain conditions (represented by SMART metrics values), failures are historically more likely to occur. In particular, we see that Node 18 is a good example of an extreme failure mode. This leaf node aggregates a population of machines, which, based on certain SMART metric conditions, have the lowest expected survival time of all of the leaf nodes displayed: after 50 days, most of the machines had failed, and all had failed within a year. The conditions chosen by the tree to assign a machine to Node 18 are symptomatic of an extreme failure mode, as such machines are experiencing:

- High count of reallocations (bad sectors that have been found and reallocated), as shown by the positive value for raw SMART metric 5.
- High spin-up time (these machines are thus slower than usual, potentially indicating accumulated wear on the drive), as shown by a lower than 100 value for normalized SMART metric 3.
- Higher than usual count of uncorrectable errors, as shown by a lower than 100 value for normalized SMART metric 187.

This provides a very interpretable and understandable characterization of a set of hard drives that are in very poor health and should be expected to fail imminently.

The survival tree also uncovers subpopulations with characteristically healthy behaviors in hard drives. In particular, we consider Node 11, which has the highest expected survival time of all leaf nodes. The SMART metric conditions leading a hard drive to Node 11 also make intuitive sense: healthy values for raw SMART metrics 3, 5 and 7. Around half of the drives in this subpopulation end up lasting another two years, indicating they are indeed particularly healthy.

Finally, to quantify the relative importance of the metrics in modeling long-term health, we can analyzing the variable importance plot for the Optimal Classification Tree model that was presented in Fig. 3, shown in Fig. 5. The variable importance metric is calculated as the sum of the relative improvements in model objective made by the split variable at each split. This is a common metric used in all decision-tree based approaches such as CART, random forests, and gradient boosted trees. We see that metrics 3 and 5 are the most important determinants of the long-term health, followed by 7. The importance of 5 is particularly evident in the tree, as the life expectancy drops dramatically as soon as more than one bad sector has been found and reallocated.

5. Predicting short-term health

In this section, we focus on the problem of predicting short-term drive failure from the SMART metrics. First, we formulate the problem as a short-term survival analysis and use Optimal Survival Trees to model drive health over this horizon, and then we pose the problem as a classification task and predict the occurrence of failure within this time window.

5.1. Short-term health as a survival problem

For this analysis, we will consider only a single quarter's worth of data (Q1 2020), and use Optimal Survival Trees to model the drive health over this 90-day window. This gives us 16,511 number of observations for the short-term analysis, with 13,208 used for training and the remainder for testing.

Unlike Section 4 where we limited the analysis to failing hard drives, in this analysis we will consider both hard drives which failed and hard drives which did not fail within the time window. This results in a large number of *censored* observations (as discussed in Section 2), as most of the drives observed do not fail within the quarter. However, as discussed previously, the absence of failure of such drives can be utilized by survival models, because they still provide a lower bound on the remaining useful life of each drive.

We used this dataset to train an Optimal Survival Tree, which is shown in Fig. 6 (the full interactive tree with Kaplan–Meier curves for each leaf is again available in the supplementary materials). The trees were crossed validated with maximum depth between 2 and 6. As in Section 4, each leaf displays the expected survival time for drives falling into the leaf, with a darker color shading indicating longer survival.

We see that many of the features used by this survival tree are very similar to the ones used by the long-term health tree, and overlap heavily with the metrics identified by Backblaze and Google as highly correlated with failure.

The variable importance scores for this tree are shown in Fig. 7. We see that metric 5 is again the most important, however metric 3 is significantly less important here compared to its importance in modeling long-term drive health. Instead, metric 187 becomes very important, mirroring the conclusion from Google that any change in this metric from baseline is a very significant predictor of failure.

By examining the Kaplan–Meier curve in each leaf, we see that the survival tree is successfully discriminating between failing and healthy leaf nodes on this short time scale. Fig. 8 shows the curves for selected leaves. We see that certain leaf nodes like Nodes 21 and 24 are comprised of extremely unhealthy drives, with near-certain failure within 90 days, while Nodes 12 and 15 contain seemingly healthy hard drives with almost no risk of failure within the 90-day window.

The Kaplan–Meier curves in the leaves are also a powerful tool for telling us more about the drive behavior than we can get from the expected survival time alone. Nodes 6 and 27 have very similar expected survival times (39.4 and 41.7 days, respectively), yet the curves show they exhibit distinctly different behaviors over time. It appears that drives in Node 6 tend to either fail within 40 days or not fail at all during the 90-day window. However, Node 27's drives are failing at a rate that is roughly constant over time. We can try to understand the differences in behavior by analyzing the conditions that lead to these leaves:

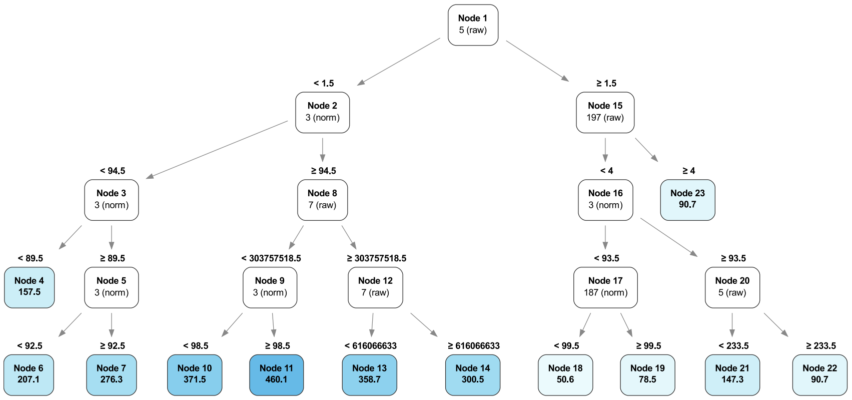

Fig. 3. Optimal Survival Tree for predicting long-term health. The variable in the split nodes refer to the splitting SMART variable (either as raw or normalized value). The numbers in the leaf nodes refer to expected survival times in days, where darker shading in color corresponds to longer survival.

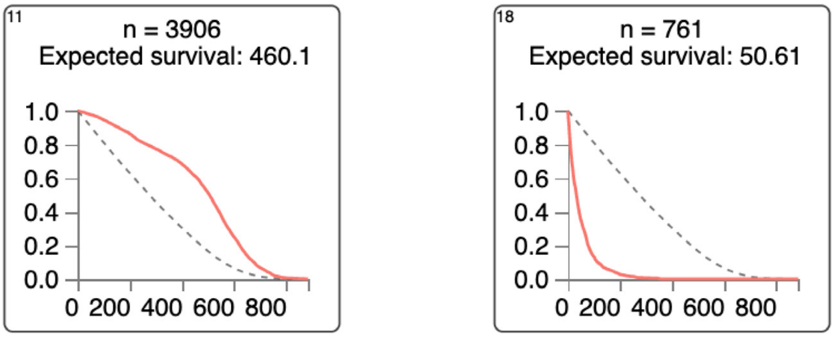

Fig. 4. Survival curves for Nodes 11 and 18 in the Optimal Survival Tree for predicting long-term health. The x-axis is the time in days and the y-axis is the probability that a given drive will not fail before this time. The black dotted line in each curve is the overall survival curve for all hard drives, whereas the red line refers to the survival curve for hard drives in the specific leaf nodes.

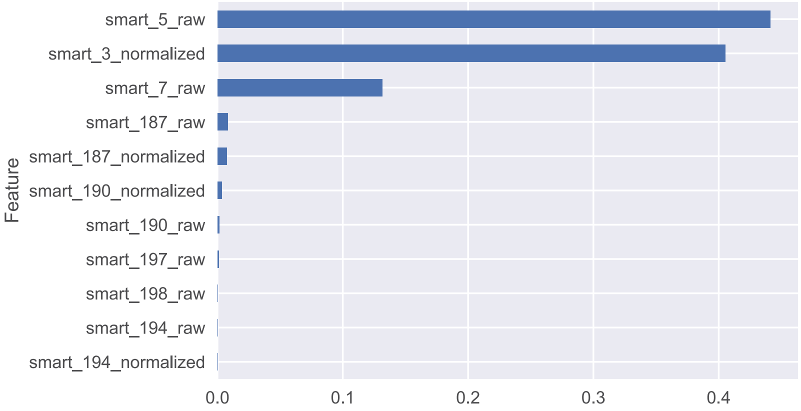

Fig. 5. Variable importance for long-term Optimal Survival Tree.

- Node 6 is characterized by a high seek error rate (low values for normalized and high values for raw SMART metric 7), which is symptomatic of a partial failure in the mechanical positioning system: we might imagine that such a partial failure could be lethal in certain cases and surmountable in others, leading to the two classes of behavior seen in this node.
- Node 27 contains drives with relatively high values for raw SMART metric 5, which would normally be indicative of impending failure. However, these drives also have high values for normalized SMART metric 187, which might explain the steady pace of failure among these drives. This is in contrast to the extreme failure modes we observed in Nodes 21 and 24, which have anomalous values for both metrics 5 and 187.

In particular, comparing the results for Node 27 against those for Nodes 21 and 24 demonstrate the power of a non-linear model like Optimal Survival Trees. Individually, metrics 5 and 187 seem to be correlated with impending short-term failure, but the non-linearity of the tree uncovers that drives with abnormal values for metric 5 but normal values for 187 are in fact significantly healthier than those with abnormal values for both metrics.

5.2. Short-term health as a classification problem

Approaching the question of short-term health with survival analysis allows us to model the health of the drive over the entire 90-day

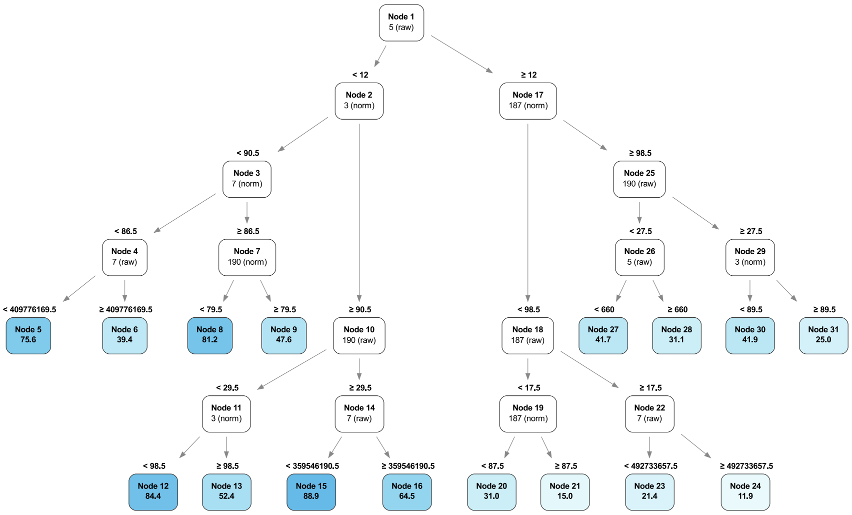

Fig. 6. Optimal Survival Tree for predicting short-term health.

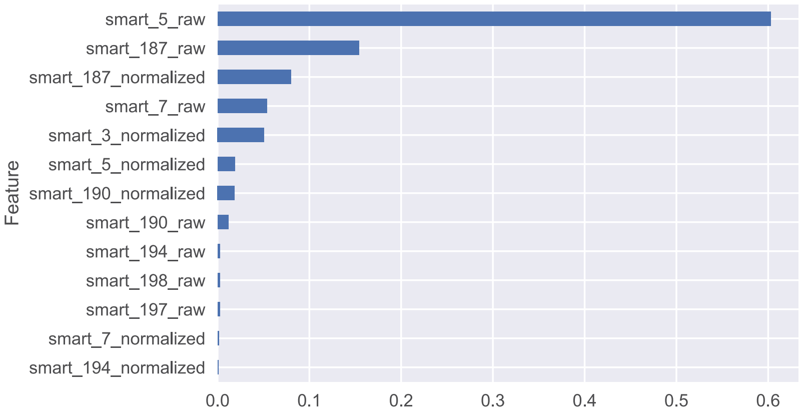

Fig. 7. Variable importance for short-term Optimal Survival Tree.

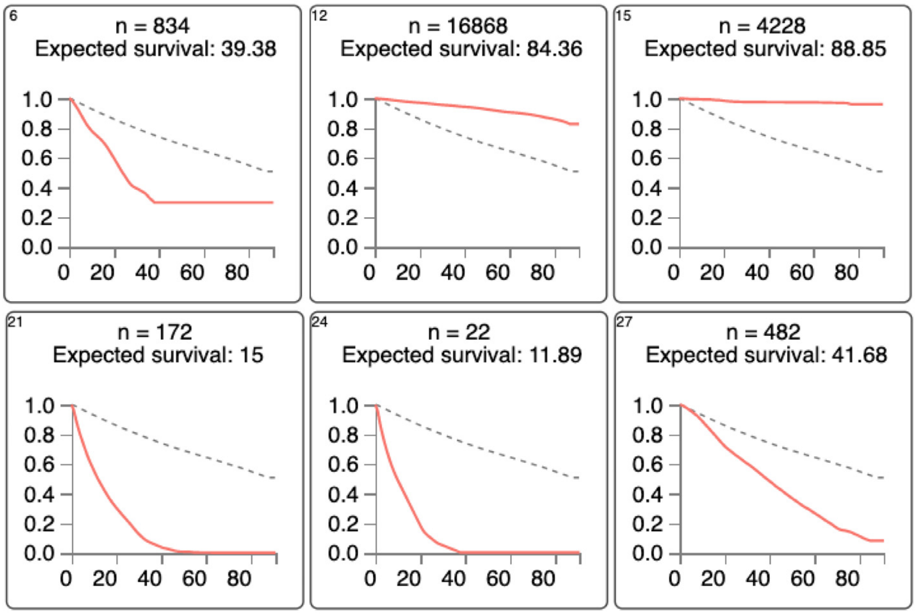

Fig. 8. Survival curves from short-term Optimal Survival Tree for Nodes 6, 12, 15, 21, 24 and 27.

time window, but sometimes we might want to focus on a specific operational concern, such as predicting whether drives are likely to fail within the next 30 days, to plan enough time for replacements. The survival analysis can be used to answer these questions, but because they model the health of the drive over the entire time window they may be less precise than a classification model trained specifically with this single focus.

To investigate this, we will pose the task of predicting drive failures within 30 days as a classification problem, and use Optimal Classification Trees to generate an interpretable predictive model for the probability of failure. Our target variable for the classification problem is whether or not the hard drive failed in the 30 days following the date the SMART metrics were recorded. We used one full year worth of data (Q2 2019 to Q1 2020, inclusive), and separated the data into training and testing based on the hard drives' serial numbers so that no drive had observations appearing in both the training and testing sets. The full data has 524,845 observations, with 419,027 used for training and 105,818 for testing. The raw data was unbalanced with very few hard drives failed during the 30-day period. As a result, we used gini impurity index as the training criterion to handle the unbalanced data.

As comparison against state-of-the-art methods, we also trained CART and Random Forest on the same dataset. The model for CART was cross validated with maximum depth between 2 and 6. The random forest model was cross-validated with maximum depth between 2 and 6, as well as number of trees between 50 and 500 with increments of 50.

We show the trained Optimal Classification Tree in Fig. 9 (again, an interactive visualization with additional details is available in the supplementary materials). The trees were crossed validated with maximum depth between 2 and 6. The probability of failure within 30 days as predicted by the model is displayed in each leaf node of the tree, with a darker color shading indicating higher probability of failure.

Inspecting the tree, we see that the structure appears similar to the survival tree presented in Section 5.1, and similar variables are used in the splits. Fig. 10 shows the variable importance scores for this model, and indeed SMART metric 5 is again the most important, followed by 187. The probabilities of failure in each leaf range from 3.0% to 62.2%, indicating that the tree can indeed distinguish between the healthy drives and those with impending failure.

We compare the model performance between Optimal Classification Trees, CART, and Random Forest using AUC, the area under the ROC (Receiver Operating Characteristic) curve. It is a holistic measure of the ability of the model to discriminate between failing and healthy drives, where 0.5 corresponds to random guessing and 1 indicates perfect predictions. Fig. 11 shows the ROC curves for each of the three methods. We can see that Random Forest has the highest AUC and CART the lowest, with Optimal Classification Trees coming inbetween. This demonstrates the edge of global optimization in finding the best single-tree solution.

5.3. Comparison of survival and classification approaches

We have considered the short-term health from both survival and classification perspectives. The Optimal Classification Tree makes predictions of failure probability within 30 days, while the Optimal Survival Tree allows us to make probabilistic predictions at any given time in 90-day window. Because the curve is monotone, the predicted probability of survival is always lower with longer evaluation time, consistent with intuition.

When assessing the quality of the predictions, we will report the following measures:

- AUC as described above
- Accuracy is the proportion of drives that are labeled correctly.
- Sensitivity, or true positive rate, is the percentage of failing machines which are correctly identified as failing: we want it as high as possible, as we would like to isolate failing machines for maintenance.
- The false alarm rate is the percentage of machines that are predicted to fail but do not actually fail: this should be as low as possible to minimize the number of drives falsely identified as being at risk of failure.
- The precision is the percentage of machines that actually failed among those we predict failure. Also known as positive predictive value, a higher number is desired.

Table 2
Out-of-sample performance results for OCT and OST at various evaluation times. Accuracy, sensitivity, false alarm rate, and precision are evaluated under a threshold of 0.05.

Metric	OCT	OST		
	30 day	30 day	60 day	90 day
AUC	0.7238	0.6915	0.6637	0.7011
Accuracy	0.8614	0.8526	0.8386	0.8328
Sensitivity	0.5468	0.5200	0.4634	0.4500
False alarm rate	0.1185	0.1259	0.1157	0.0887
Precision	0.4432	0.2096	0.3275	0.5094

Accuracy, sensitivity and the false alarm rate all depend on the selection of a threshold for the probability of failure: if the predicted probability is above this threshold then the drive is predicted to fail. If the threshold is too low, then we will have high sensitivity but also a high false alarm rate, and conversely, if the threshold is too high, we will have a low number of false positives but also very low specificity. Fig. 11 shows the ROC curve which visualizes the trade-off between sensitivity and false alarm rate as the threshold is varied. To maximize the sensitivity while minimizing the false alarm rate, we elected to have the model operate at the inflection point of the curve, corresponding to a threshold of 5%. In simple terms, if a machine's predicted probability of failing in the next 30 days is higher than 5%, it is predicted as a failure.

Table 2 shows the performance of the OCT (predicting failure within 30 days), and the OST when predicting failure within 30, 60, and 90 days. For each, we evaluate the AUC, accuracy, sensitivity, and false alarm rates in testing data, using the threshold of 5% where required.

We see that the performance of OST across all three time windows is similar, with sensitivity of roughly 45%–50% and false alarm rate around 10%, indicating that we identify roughly half of the failing drives while only having a 10% false positive rate. Comparing to the results of the OCT with the OST at 30 days, we see that the OCT has a small advantage in all metrics. This confirms our intuition that by directly formulating the problem as a classification problem for a specific time horizon of interest we can achieve better performance. In contrast, the survival tree is able to make predictions for any time horizon and guarantees monotonicity of these predictions over time, with a slight impact on performance compared to the classification approach.

6. Conducting analysis with limited data

In the last two sections, we have seen that extensive granular data representing the behavior of machines allows us to both assess the overall long-term health of such machines and predict which machines will fail over a shorter timeframe. However, it is not always the case that such a wealth of historical data is available for this modeling. In particular, collection of this kind of sensor data has often only begun recently, and in this situation a common question is *how much data is required for an insightful analysis?* In this section, we will see that both analyses can still be conducted with data from a shorter time frame.

6.1. Long-term health

In Section 4, we analyzed the behavior of hard drives using a three-year time window, which covered the entire lifespan of nearly all the

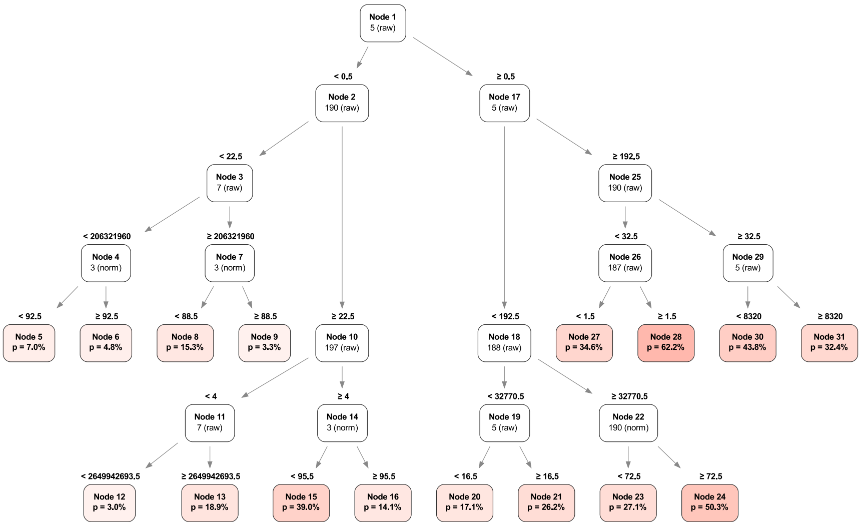

Fig. 9. Optimal Classification Tree for predicting failure within 30 days. The variable in the split nodes refer to the splitting SMART variable (either as raw or normalized value). The probability in each leaf node is the predicted probability of failure, where darker shading indicates higher probability.

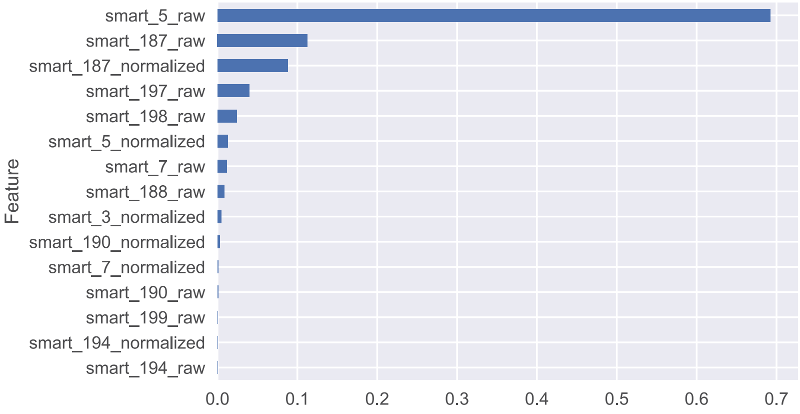

Fig. 10. Variable importance for Optimal Classification Tree predicting failure within 30 days.

drives considered. To examine what is possible under a more data-limited scenario, here repeat our analysis using only a single year of data (Q1 2019 through Q1 2020, inclusive).

In the three-year case, we had sufficient failure data to restrict our analysis to hard drives that failed as we had sufficient data to observe their entire lifecycle. However, with only one year of data, the time window is too short to observe the entire lifespan for most failing drives. For this reason, we utilize all hard drives in the year of data for our analysis, and treat the drives that did not fail as censored data, as in Section 5.1. This gives us 557,936 number of observations. We randomly resampled 50,000 observations for model training and used the rest for testing.

We show the trained survival tree in Fig. 12 (with the full interactive tree available in the supplementary materials). We see many similarities in structure between this tree and the one trained with full data from Section 4. For instance, the leaf with the lowest expected survival time in this tree is Node 5, characterized by a high value for raw SMART metric 5, and lower-than-baseline value for normalized SMART metrics 3 and 187. This is similar to the leaf with lowest expected survival in the original tree (Fig. 3), which is Node 18 and is characterized by a high raw value for metric 5 and 197, and deviations from baseline in metrics 3 and 187. Despite the shorter timeframe for the data, the tree is still able to discover similar modes of accelerated failure.

We also see that the tree includes a split on metric 190 (Temperature Difference). Recall that the univariate analysis conducted by Google concluded that this temperature was not predictive of impending hard drive failure. However, the tree uses it to fine-tune its prediction after analyzing metrics 3, 5, and 187, demonstrating the power and flexibility of such a non-linear model over simpler approaches.

Fig. 13 shows the variable importance for the tree trained on limited data. We see that metric 5 is by far the most important, whereas metric 3 is much less important than it was for the original tree. This makes sense as metric 3 was used heavily on the left side of the tree (Fig. 3) to refine the survival estimates for the health drives with expected survival times between over 6 months. Given that the new tree is only exposed to a single year's worth of data, it makes sense that it cannot learn much about the longer-term survival in this way.

Overall, with only one year worth of data we see that it is still possible to recover useful insights about how the various SMART metrics indicate the overall health of a drive.

6.2. Short-term health

Analyzing the short-term health of drives requires no changes from before, provided we have data longer than the window of interest. The survival analysis in Section 5.1 was already conducted using just a single quarter of data. The classification approach of 5.2 used one year of data to run a thorough training and testing comparison analysis, but we could just as easily have obtained similar results with less data, provided that we observe a sufficient number of failures in line with our

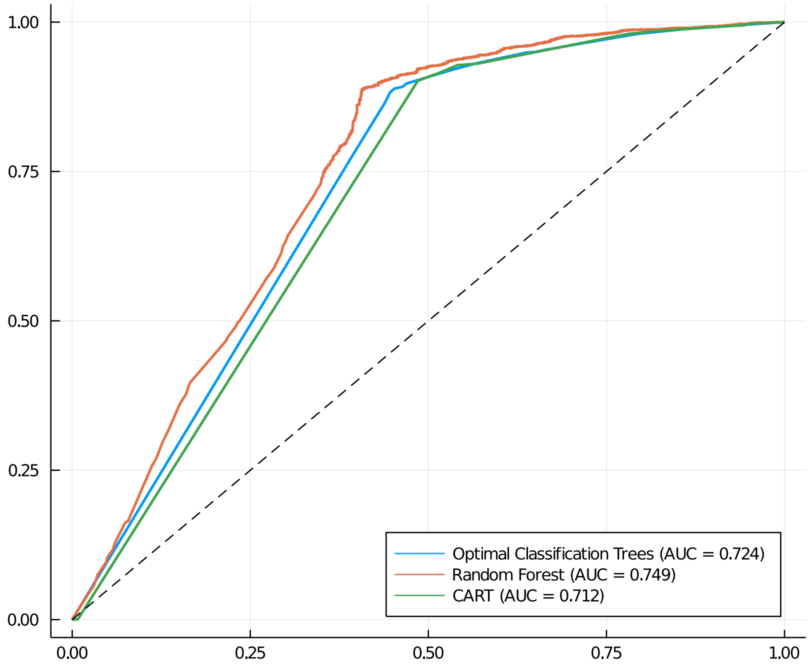

Fig. 11. ROC curves for Optimal Classification Trees, Random Forest, and CART predicting failure within 30 days.

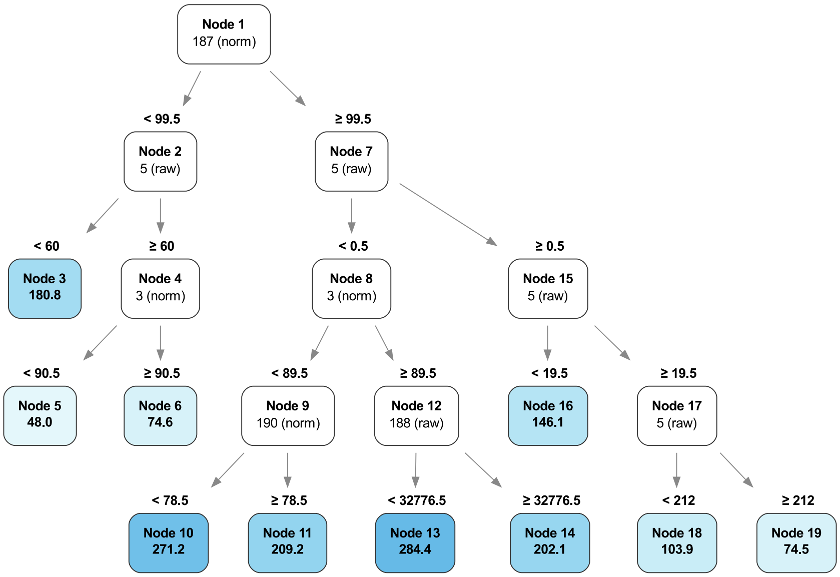

Fig. 12. Optimal Survival Tree trained with limited data.

time window. Our intuition is that in this case as little as 1–2 quarters of data would be sufficient to predict 30-day failures.

7. Conclusion

In this paper, we showed that interpretable machine learning algorithms can be used to both assess the long-term health of hard drives, and to predict short-term failure. In addition, the approach can be adapted to limited data scenarios without significantly affecting the quality of the models.

The Optimal Classification Tree and Optimal Survival Tree models that we use bring a number of advantages to predictive maintenance over correlation analyses and classical predictive maintenance machine learning techniques:

- **Detecting interpretable paths to failure**
  Compared to Random Forest and other black-box models, Optimal Trees can automatically identify both paths leading to accelerated failure as well as paths indicating healthy behaviors. Each leaf of the tree defines a cohort of hard drives with similar survival outcomes based on a series of conditions on their current SMART metric values, providing a discrete, easy-to-understand description that can be validated against expert knowledge. In this case, they uncovered interactions between several SMART metrics simultaneously in addition to confirming findings from existing univariate correlation analysis.
- **Tailored model to address each question**
  We presented many model variants each targeted at different questions. In Section 4, we used OST to understand long-term

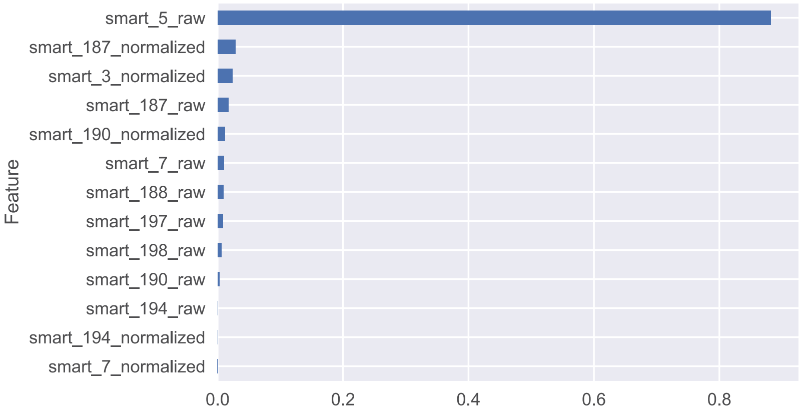

Fig. 13. Variable importance for Optimal Survival Tree with limited data.

health. In Section 5.1, we again used OST but for the purposes of characterizing short-term health. Finally, in Section 5.2, we used OCT to predict occurrence of failure within a specific time window. In contrast to the one-size-fits-all style of previous analysis, this flexibility allows us to construct models that are targeted towards answering specific questions. These interpretable models can uncover simple-yet-powerful insights allowing data-oriented people and domain experts to have a common communication ground.

- **Applicable with limited data**

  If measurement data is scarce or is just starting to be collected, the approaches we utilized can still be applied to whatever limited data is present, and still generate useful and predictive models. In particular, the ability of Optimal Survival Trees to deal with censored data permits it to exploit any data that is available.

The analysis we conducted only considered a single model of drive, and most likely could be strengthened through approaches like engineering new derivative features based on changes in the SMART metrics over time. Nevertheless, the approach we present is a powerful way to derive useful insights and strong predictive performance by applying interpretable and non-linear methods to the problem of predictive maintenance.

CRediT authorship contribution statement

Maxime Amram: Writing - review & editing, Visualization. **Jack Dunn:** Conceptualization, Software, Writing - review & editing, Supervision. **Jeremy J. Toledano:** Methodology, Software, Investigation, Data curation, Writing - original draft, Visualization. **Ying Daisy Zhuo:** Validation, Writing - review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mlwa.2021.100042.

References

Backblaze (2014). Hard drive SMART stats. URL: https://www.backblaze.com/blog/hard-drive-smart-stats/. (Accessed 14 September 2020).

Backblaze (2020). Hard drive data and stats. URL: https://www.backblaze.com/b2/hard-drive-test-data.html. (Accessed 14 September 2020).

Backblaze (2020). Hard drive failure rates: A look at drive reliability. URL: https://www.backblaze.com/blog/backblaze-hard-drive-stats-q1-2020/. (Accessed 14 September 2020).

Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine Learning*, *106*(7), 1039–1082.

Bertsimas, D., & Dunn, J. (2019). *Machine learning under a modern optimization lens*. Dynamic Ideas LLC.

Bertsimas, D., Dunn, J., Gibson, E., & Orfanoudaki, A. (2020). Optimal survival trees. *Mach. Learn.*, (under review).

Carvalho, T. P., Soares, F. A., Vita, R., Francisco, R. d. P., Basto, J. P., & Alcalá, S. G. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, *137*, Article 106024.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *34*(2), 187–202.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning (vol. 1), no. 10*. Springer Series in Statistics New York.

Kanawaday, A., & Sane, A. (2017). Machine learning for predictive maintenance of industrial machines using IoT sensor data. In *2017 8th IEEE international conference on software engineering and service science* (pp. 87–90). IEEE.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457–481.

LeBlanc, M., & Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics*, 411–425.

Li, J., Ji, X., Jia, Y., Zhu, B., Wang, G., Li, Z., & Liu, X. (2014). Hard drive failure prediction using classification and regression trees. In *2014 44th annual IEEE/IFIP international conference on dependable systems and networks* (pp. 383–394). http://dx.doi.org/10.1109/DSN.2014.44.

Li, J., Stones, R. J., Wang, G., Liu, X., Li, Z., & Xu, M. (2017). Hard drive failure prediction using decision trees. *Reliability Engineering & System Safety*, *164*, 55–65. http://dx.doi.org/10.1016/j.ress.2017.03.004, URL: https://www.sciencedirect.com/science/article/pii/S0951832016301569.

Pinheiro, E., Weber, W.-D., & Barroso, L. A. (2007). Failure trends in a large disk drive population. In *5th USENIX conference on file and storage technologies* (pp. 17–29).

Pitakrat, T., van Hoorn, A., & Grunske, L. (2013). A comparison of machine learning algorithms for proactive hard disk drive failure detection. In *Proceedings of the 4th international ACM sigsoft symposium on architecting critical systems* (pp. 1–10). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/2465470.2465473.

Rincón, C. A. C., Pâris, J., Vilalta, R., Cheng, A. M. K., & Long, D. D. E. (2017). Disk failure prediction in heterogeneous environments. In *2017 international symposium on performance evaluation of computer and telecommunication systems* (pp. 1–7). http://dx.doi.org/10.23919/SPECTS.2017.8046776.

Succi, G., Romanov, V., Reznik, A., Litvinov, S., Kozar, A., Ivanov, V., & Garcia, M. (2018). Review of techniques for predicting hard drive failure with SMART attributes. *International Journal of Machine Intelligence and Sensory Signal Processing*, *2*, 151. http://dx.doi.org/10.1504/IJMISSP.2018.10014099.

Wikipedia (2020). S.M.A.R.T. URL: https://en.wikipedia.org/wiki/S.M.A.R.T. (Accessed 14 September 2020).

Xiao, J., Xiong, Z., Wu, S., Yi, Y., Jin, H., & Hu, K. (2018). Disk failure prediction in data centers via online learning. In *Proceedings of the 47th international conference on parallel processing*. New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3225058.3225106.