A practical approach to predicting remaining useful life of hard disk drives

Cristian Seceleanu
Prof. Richard Mitchell (sup)
08 September 2023

Outline

- Why this project?
- Literature
- RF, BiLSTM, RF & BiLSTM
- Dataset
- Some findings
- pRUL (Agent, Web API, Oracle)
- Benefits of this approach
- Demo

Illustrations by Pixeltrue on icons8

Why this project?

Simply put because I like storage and related problems

> Actually, because I've been a systems engineers for too many years, my main reason is that "proactive" is always better than "reactive". Knowing when a drive will fail we can schedule maintenance windows or at the very least make sure we have enough spare parts in stock.

Literature

In the field of predicting remaining useful life of HDDs

Most of the literature in the field tackles the problem from a classification pov (some of it from a regression pov).
Most works use the Backblaze dataset and selecte one particular HDD model then perform various operations on the dataset in the preprocessing step before training a ML model on the data.
Some excellent results (state-of-the-art) have been obtained thanks to RF and BiLSTM together with data preprocessing techniques.

Random Forest

RF is probably the best ML algorithm to use for classifying the state of HDDs where by looking at a set of measurements a model can predict its status (class)

LSTM

LSTM is excellent at predicting RUL over long windows of time based on measurements captured periodically

BiLSTM

BiLSTM is better than LSTM as it runs in both directions (forward – towards failure and backwards – from failure) and is able to better learn the relationship between measurements and RUL.

RF + BiLSTM

Research has shown that in order to predict RUL one cannot simply create sets of lookback windows of different sizes over the entire lifetime of HDDs in order to train a BiLSTM model to predict RUL so a combination of the two would produce the best results/.

A classification model, using the RF algorithm, is trained to identify prefailure at various windows before failure (15/30/60/90/etc days).

After that a regression model using the BiLSTM neural network is trained with 30 x 15/30/60/90/etc lookback windows to generate RUL predictions when a HDD is identified as being in a prefailure state.

Dataset

In the field of predicting remaining useful life of HDDs

Most of the literature in the field tackles the problem from a classification pov (some of it from a regression pov).
Most works use the Backblaze dataset and select one particular HDD model then perform various operations on the dataset in the preprocessing step before training a ML model on the data.
Some excellent results (state-of-the-art) have been obtained thanks to RF and BiLSTM together with certain data preprocessing techniques.

Findings

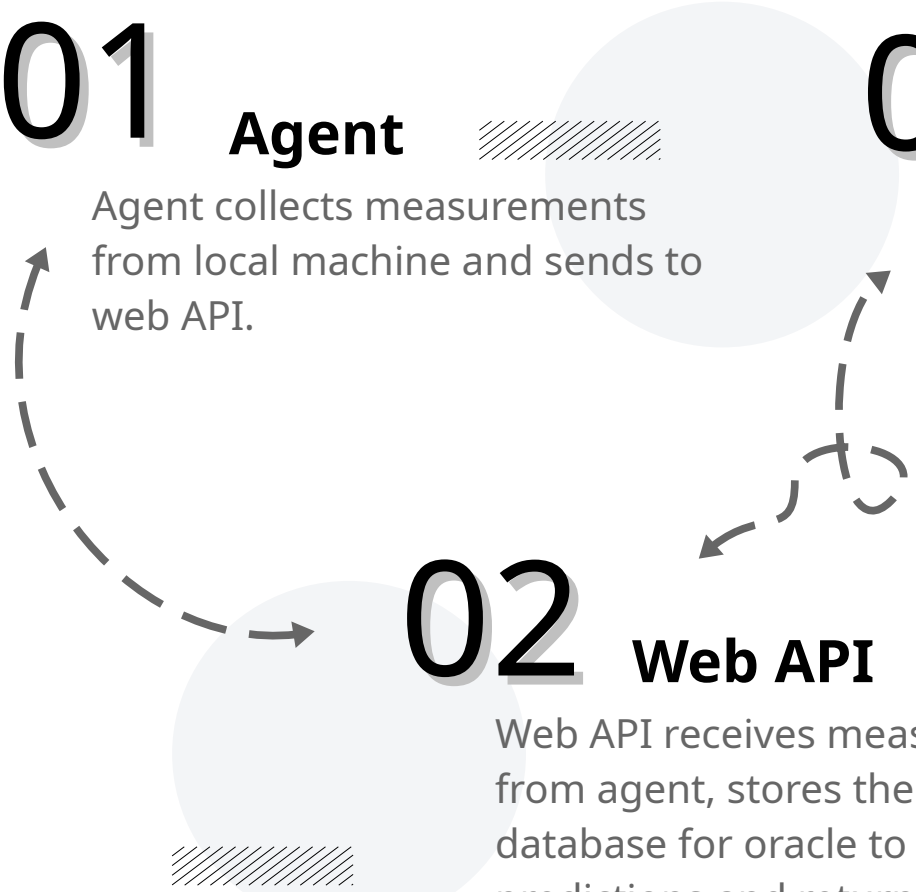
Through experimentation we found that:

> The curent health state of a HDD can be predicted with incredible accuracy and precision by using RF but only if the ML model was trained on the same HDD model (or a model that uses the same SMART reporting) as the one being assessed.
> RUL can be predicted with better confidence levels the further away we are from failure (prediction at 90 day lookback is better than at 60,30...).
> A model can be built and trained on a diverse dataset (multiple HDD models) to generate predictions with good accuracy, precision and confidence.
> All of these can be achieved in a practical way.

pRUL

Its components and how they work

Components: agent, web api and oracle (update and prediction) service
The role of the pRUL agent is to collect S.M.A.R.T. measurements from the local drives installed on the local machine and report health state and RUL. The Web API is the central service which collects measurements from the agents and returns predictions generated by the oracle service. As for the Oracle service, it is responsible for identifying the health state of HDDs and predicting RUL as well as for updating the ML models. (next slide shows how the components work together)



01

Agent

Agent collects measurements from local machine and sends to web API.

02

Web API

Web API receives measurements from agent, stores them in the database for oracle to issue predictions and returns current state and predictions to agent.

03

Oracle

Once enough measurements have been collected , the oracle service generates predictions and stores them in the db for API to return to agents.

04

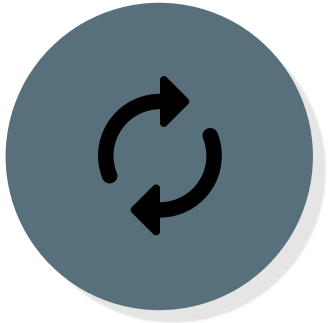
ML model update

When enough measurements have been collected the oracle service fits the existing BiLSTM models with the new data and starts using the new obtained models for generating predictions

Benefits of using pRUL

Easy to use

Simple deployment and configuration, basic software requirements, no need for a Data Science degree to use it



Easy to maintain

Once in production it just works, it updates itself, uses standard technologies

Scalable

It works at any scale from a personal laptop to the datacenter of a cloud service provider.

Demo

Let’s see it in action

> What follows is a live demo of the pRUL technical solution (what the agent is and how to install and configure it, how the web api service works and how easy it is to deploy, how the database behind it looks like and how the oracle service works to generate health assessments and RUL predictions and keep the models up-to-date.)

Thank you!

Remember: It is free! ... as in distributed under a FREE license

> To download a copy for yourself or if you are interested in more details about the project or how it came to be or perhaps if you are interested in the research behind it please visit its website:
>
> https://prul.seceleanu.co.uk