To Fail or not to Fail

-

How long will it last?

Guardians of the Memory

Stakeholder - Cloudwave

- A new startup offering cloud storage as a service
- Lack of backup options due to limited budget
- Options to use cheaper 2nd-hand hard drives
- Wants to maximize hard disk usage before failing

Our Goal - Failure Prediction

Predict if a hard drive will fail within the coming 30 days

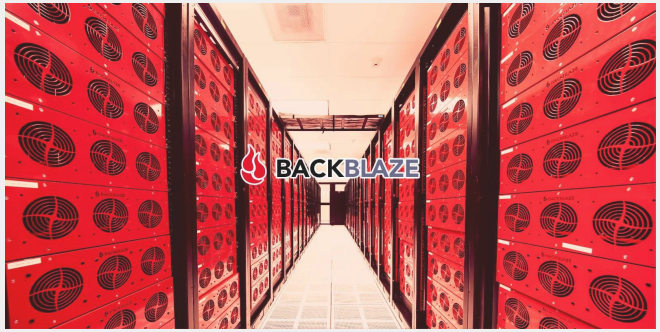
Predict long-term health of hard drives

Business Value:

- ✓ Money can be saved by fully utilizing hard drives to the very end
- ✓ Valuable user data (like photos or memories) can be saved

Hard Drive Data

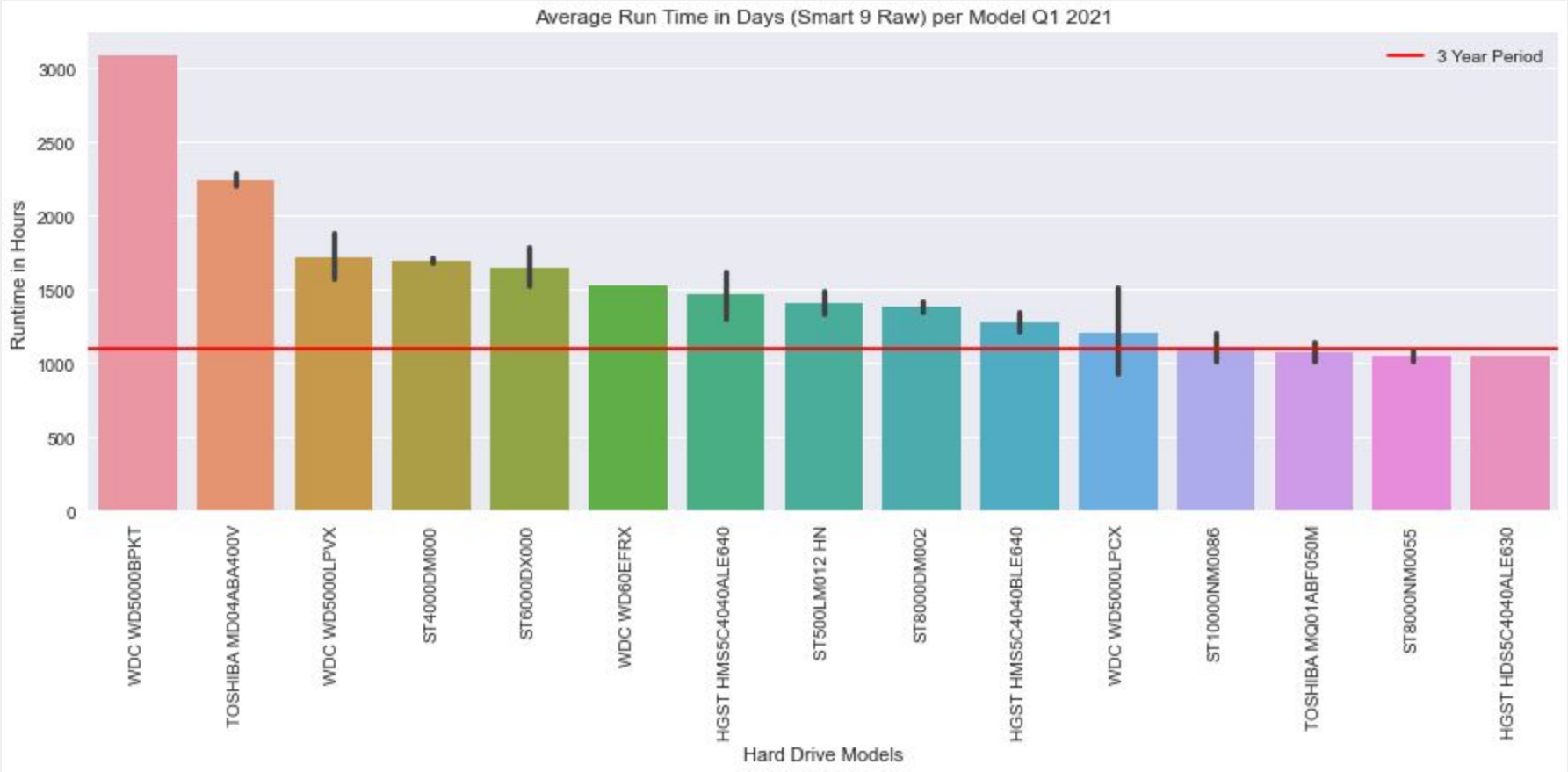
Data is offered by Backblaze:



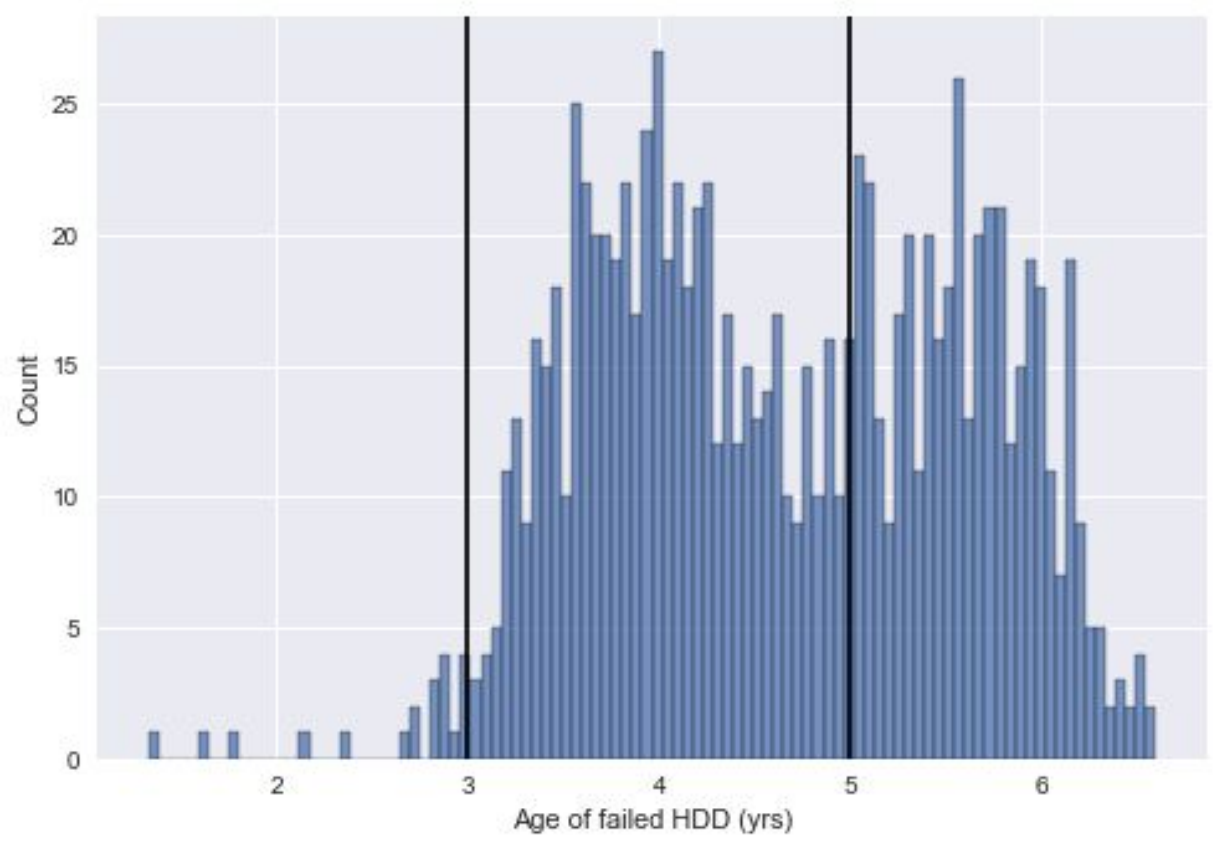
- Roughly *205k hard disks* among *65 models* (on 31/12/2021)
- *174 SMART* parameters recorded everyday (2019 to 2021)
- **Highly unbalanced data**
- Focus on the model of interest for Cloudwave

Data: https://www.backblaze.com/b2/hard-drive-test-data.html

EDA - Life Expectancy



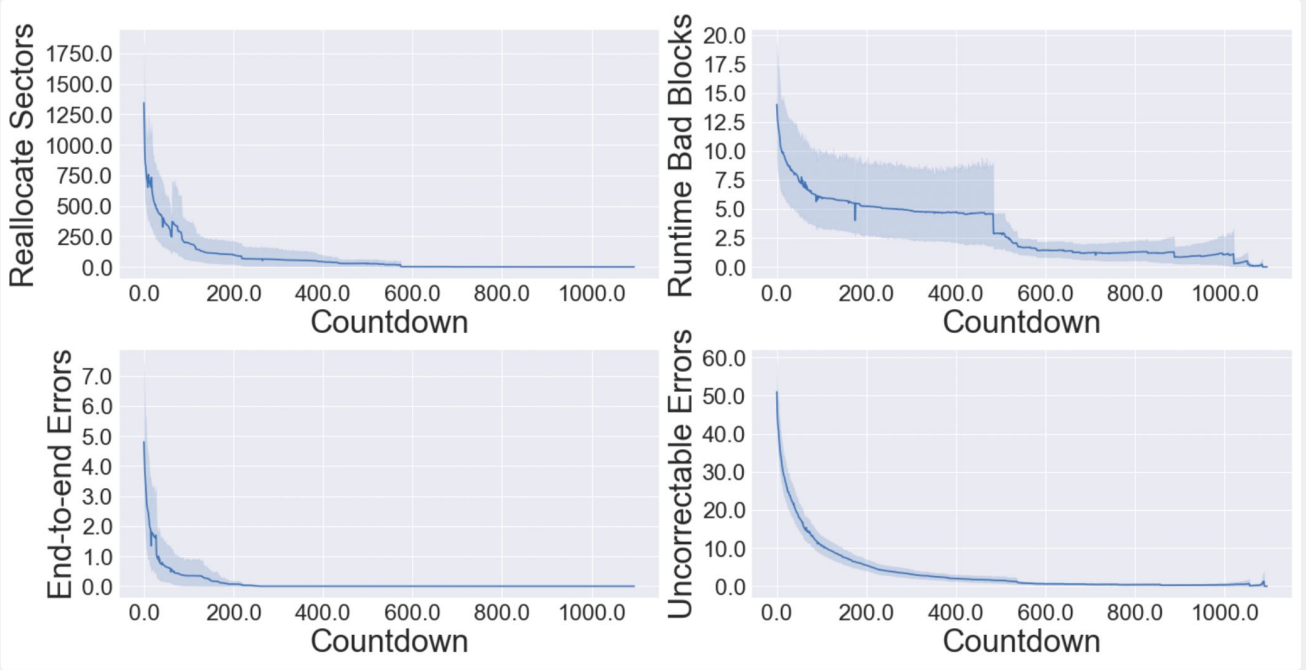
EDA - Life Expectancy for our Model



EDA - Important SMART Features

Examples of SMART features:

- error rates
- **reallocated sectors**
- power on time
- **runtime bad blocks**
- **end-to-end errors**
- **uncorrectable errors**
- temperature
- bad sectors



Model Evaluation and Data Splitting

For evaluation of prediction model, ***f2-score*** is chosen

- it accounts for both ***recall*** and ***precision***, with emphasis on recall

Data is split into train and test datasets according to serial numbers

- ensuring that certain disks in test dataset will not be seen during training

Baseline Model

If any 2 of the 4 important features are above their median, the hard drive is predicted to fail in 30 days

Test Data	
Metric	**Result**
f2-Score	29 %
(Recall)	45%
(Precision)	12%

- Baseline model is able to make an educated guess about the classification problem
- With achieved f2-score there is some room for improvement for modelling

Next Steps

- Apply machine learning models
- Do fancy feature engineering
- Detect anomaly
- Deploy the model and create Dashboard (Heroku or streamlit.io)