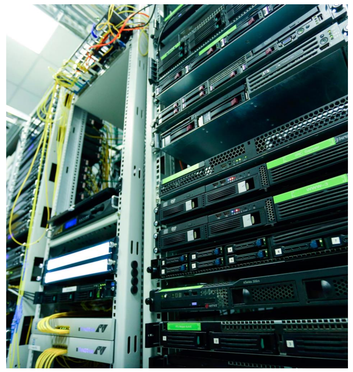
Data Driven Mission Readiness

Divyanshu Bhadoria - Brian Gan - Daivik Ghosh - Eric Lin - Gabriel Muzio - Jaxson Pahukula - Dhiya Pereira - Jiaxuan Wang - Kapil Manicka - Delaney Elder

Project Introduction

RUL Prediction for Improved Operational Capability

Our goal for this year was to develop a model to predict remaining useful life (RUL) and anomaly detection for hard drives. Doing so, will improve the operation capabilities by reducing downtime of equipment and increasing overall efficiency and productivity. Accurate RUL predictions enable proactive maintenance and replacements, minimizing unexpected failures, optimizing resource allocation, and enhancing system reliability and performance. The architecture of our solution is shown in the diagram. We aim to enhance maintainers' and operators' situational awareness regarding the state of various hard drives in the data farm, thereby aiding in the improvement of operational efficiency and effectiveness.

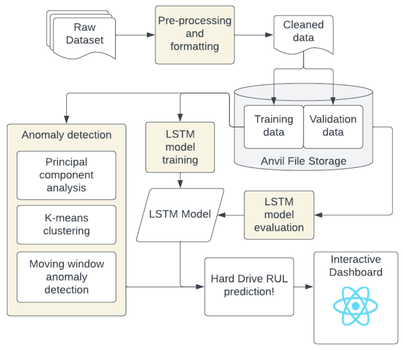


Dataset

The Backblaze hard drive dataset describes the lifecycle of a hard drive, offering statistics such as drive count, failures, drive days, and an annualized failure rate since 2013. With over 269,000 drives and a 1.46% failure rate in 2023, this dataset makes drive performance analysis more comprehensive and repeatable

Future Enterprise Solution

Pipeline Overview



Prediction Methodologies

We used many different approaches to help us achieve accurate failure prediction. Techniques we used include:

- Principal component analysis
- Various anomaly detection strategies
- Autoregressive integrated moving average (ARIMA)
- Long short-term memory (LSTM) recurrent neural networks
- K-means clustering
- And more

You can read more in-depth about some of these techniques in the panel to the right

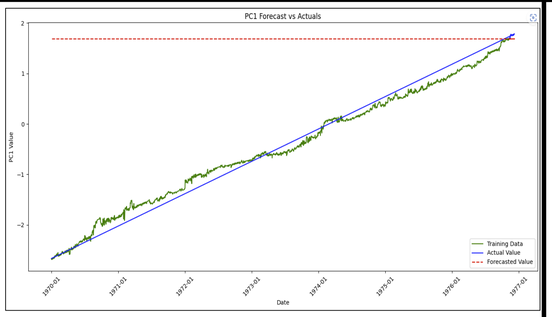
Data Storage

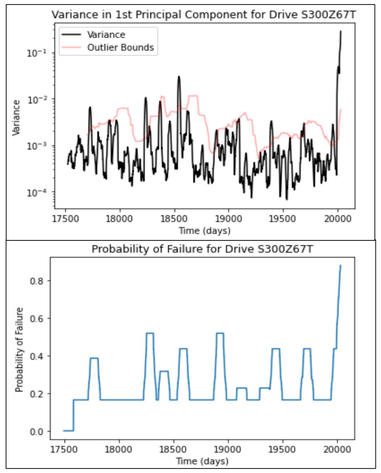
We explored many options for storing our data, including SQL and noSQL databases, however after evaluating our use case and requirements, and because file I/O wasn't a bottleneck, we settled on simply storing our data in plaintext tables, in CSV format.

RUL Prediction

ARIMA Model

We utilized an ARIMA model to project the trajectory of our primary principal component (PC1). The model's parameters were optimized using an automated selection process, enhancing prediction accuracy. By plotting actual against predicted PC1 values clearly shows when the drive will fail. This approach allows us to preemptively predict drive failures by scrutinizing PC1's future behavior.



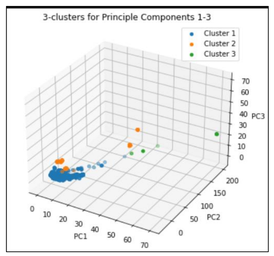


Anomaly Detection

We conducted rolling window analyses of our principal components (PCs) using primary statistical moments such as mean, variance, and skew. When comparing anomalies derived from these moments across a rolling window, variance proved to be most useful in warning of a potential failure. By observing the variance of the first PC over a rolling window for 150+ drive failures, we calculated a probability of failure for each number of anomalies observed. This allowed us to predict the probability of failure at any given time during a drive's operation.

Clustering

We applied k-means clustering to group our observations into clusters based on the values of the PCs. However, this approach was unsuccessful for forecasting drive failure due to the lack of a strong relationship between specific constant PC values and drive failure.



	index	date	serial_number	model	capacity_bytes	failure	smart_1_ra
1	2171	2021-12-16	S300WCSA	ST4000DM000	4000787030016.0	0	242329848
2	2172	2021-12-17	S300WCSA	ST4000DM000	4000787030016.0	0	83991320
3	2173	2021-12-18	S300WCSA	ST4000DM000	4000787030016.0	0	219565232
4	2174	2021-12-19	S300WCSA	ST4000DM000	4000787030016.0	0	84518200
5	2175	2021-12-20	S300WCSA	ST4000DM000	4000787030016.0	0	31332104
6	2176	2021-12-21	S300WCSA	ST4000DM000	4000787030016.0	0	199469288
7	2177	2021-12-22	S300WCSA	ST4000DM000	4000787030016.0	0	86422160
8	2178	2021-12-23	S300WCSA	ST4000DM000	4000787030016.0	0	65233128
9	2179	2021-12-24	S300WCSA	ST4000DM000	4000787030016.0	0	130375256
10	2180	2021-12-25	S300WCSA	ST4000DM000	4000787030016.0	0	230472544
11	2181	2021-12-26	S300WCSA	ST4000DM000	4000787030016.0	0	85618296
12	2182	2021-12-27	S300WCSA	ST4000DM000	4000787030016.0	0	61504392
13	2183	2021-12-28	S300WCSA	ST4000DM000	4000787030016.0	0	226500416

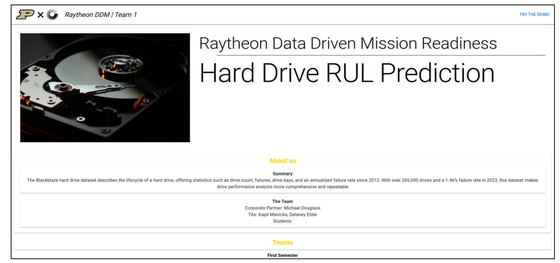
Data Engineering

The raw drive data were available as daily snapshots, which needed to be transformed into time series format. We also calculated the actual RUL for every data entry and assigned each one a "health status" (red/yellow/green).

UX

RUL Prediction Application

Throughout the entire process, we have been dedicating additional time and effort to conducting user experience (UX) research, trying to find out more about the problem we are trying to solve and how our data could solve it. The resultant final product is an interactive application that our target users can utilize to get an accurate RUL prediction of their hard drives. While our product is not polished quite yet, below is work-in-progress image of our prototype web-app.



Future Work And Lessons Learned

Future Work

- Aggregate data for a Naïve Bayes Classifier
- Shannon's "signal entropy"
- Polish interactive dashboard
- Explore needed level of interaction with the interface
- Research infrastructure to support such an interface at scale

Conclusion/Lessons Learned

Overall, we learned how to not only perform predictive maintenance on HDDs, but also how to work on a semester-long project on a team and work together to meet stakeholder expectations. Along the way, we learned about the different aspects of working in a professional collaborative environment under Agile/Scrum methodology.

Tools/Libraries

The primary tools used were Python and Jupyter. Our data analysis was performed using libraries such as Numpy, Pandas, Tensorflow, and scikit-learn. We use Matplotlib to generate graphcis, and the UI of our dashboard was built using React as a frontend framework.

Acknowledgements

We'd like to acknowledge and give our warmest thanks to Raytheon, specifically our Mentor Michael Douglass, and the Data Mine Staff, Dr. Ward, and Emily Hoeing for their constant support.