Predicting Storage Failures

by Ahmed El-Shimi
aelshimi@alum.bu.edu

VAULT - Linux Storage and File Systems Conference
Cambridge MA. March 22, 2017

This Talk

- Part I: Motivation
  - Drive Failure & Common Mitigations
  - Seeking a Better Recover/Rebuild
  - Use Cases & Goals
- Part II: Examining the Data
  - Dataset
  - Features, Trends, Challenges
- Part III: Predicting Drive Failure
  - How to Evaluate
  - Baseline
  - Approach & Models
  - Evaluation & Results

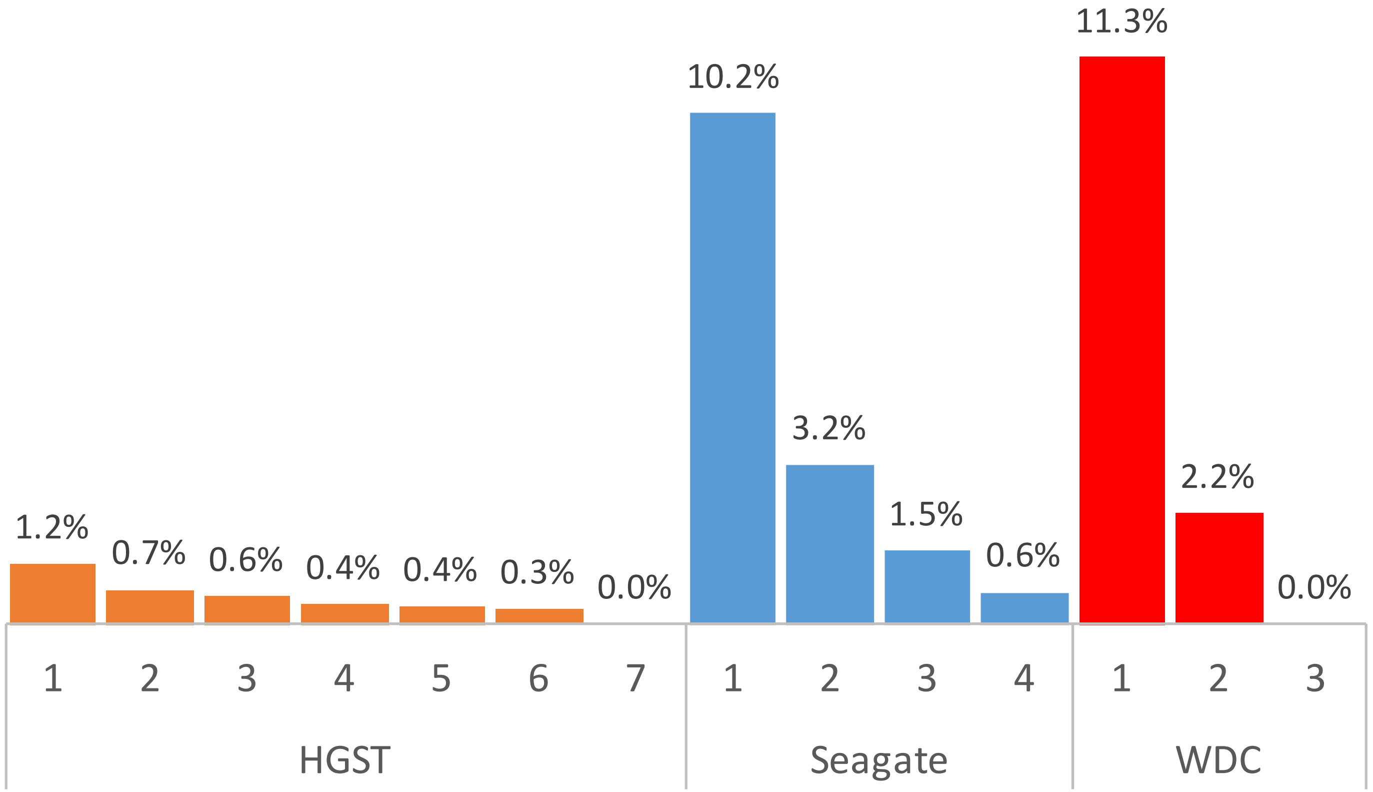
Part I

Disks fail. And even with redundancy failure has costs.

The Problem

2%

Drive Failure Rate
Averaged, Annualized.



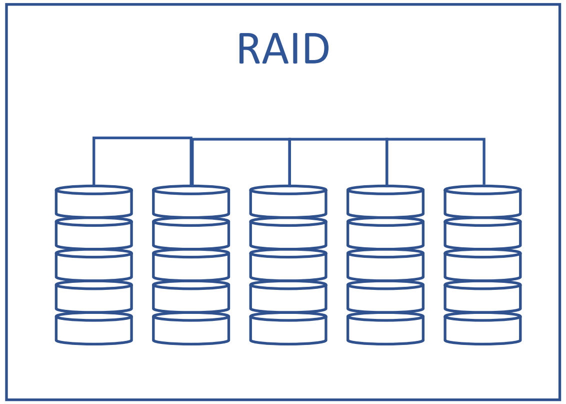
Drive Annual Failure Rate by Manufacturer and Model

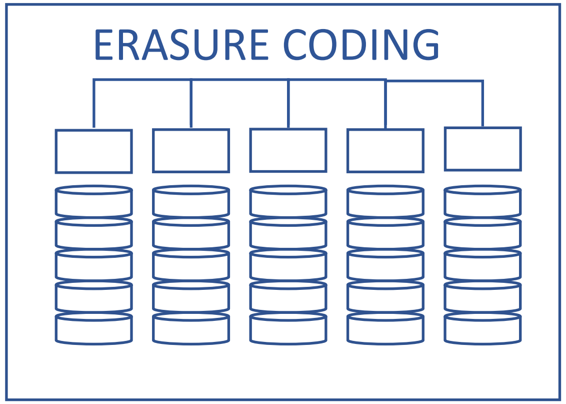
Source: Hard Drive Failure Rate Report Q3 2016, Backblaze.

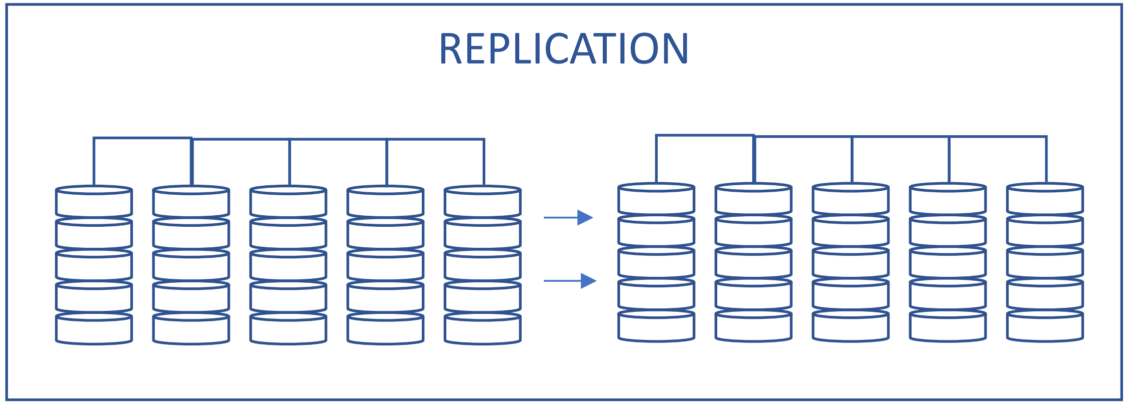
Today’s Mitigation

<u>Assume:</u> “Everything that can fail will fail”

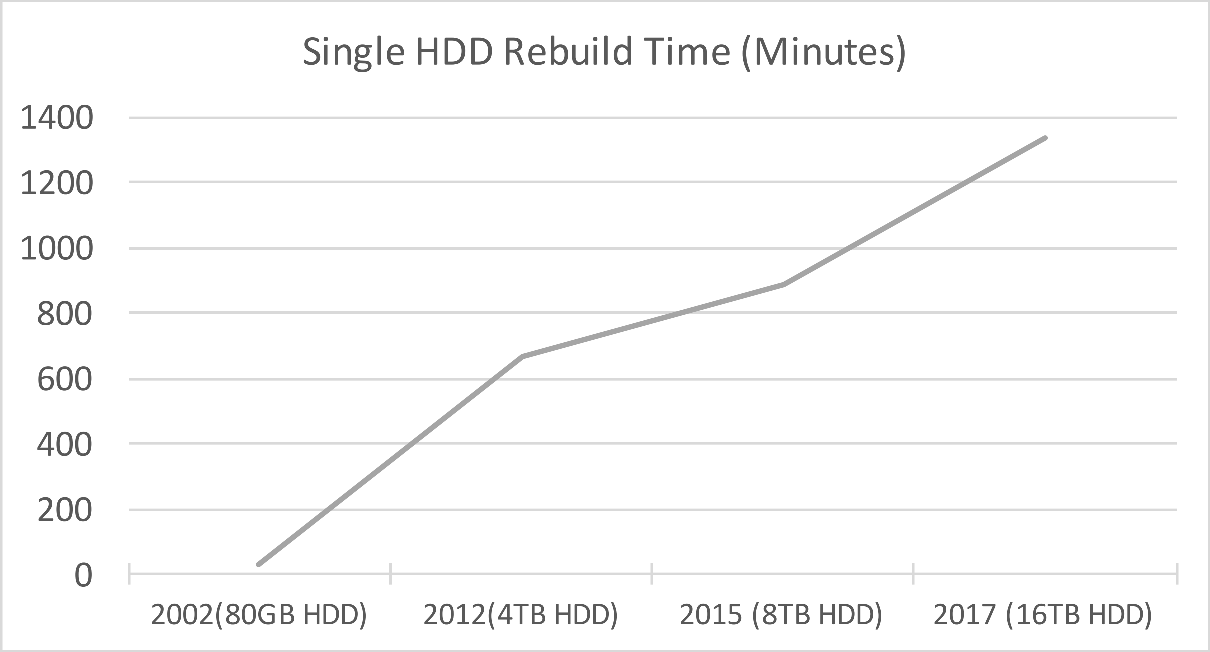
<u>Design:</u> Redundancy at every point of failure







But Rebuild is not free



	2012	2015	2017
Capacity	4TB HDD	8TB HDD	16TB HDD
Throughput	100 MB / Sec	150 MB / Sec	200 MB / Sec
1-Disk Rebuild Time	11 hours	15 hours	22 hours

Rebuild Inflation has consequences:
- Availability and Durability 9s
- Rebuild is a workload!
- Resilience, Reliability
- Disk Capacity & Network Management
- Failure Modes
- Lots of complexity to address edge cases

Can we do better if we had an early warning?

Use Cases & Goals

Cloud	Enterprise/Field	End-User PC
• Proactive Rebuild • Smarter Ops Scheduling	• Proactive Rebuild • Better FRU SLA	• Backup Now • Contingency planning

Part II

Examining the Data.

The Backblaze Dataset

- Backblaze.com: Online Backup and Cloud Storage provider.
- 83+K drives
- 2013-2016
- Seagate, Hitachi, HGST, Western Digital, Toshiba, Samsung

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

- Hats off to them for sharing their data openly.

Understanding the Data

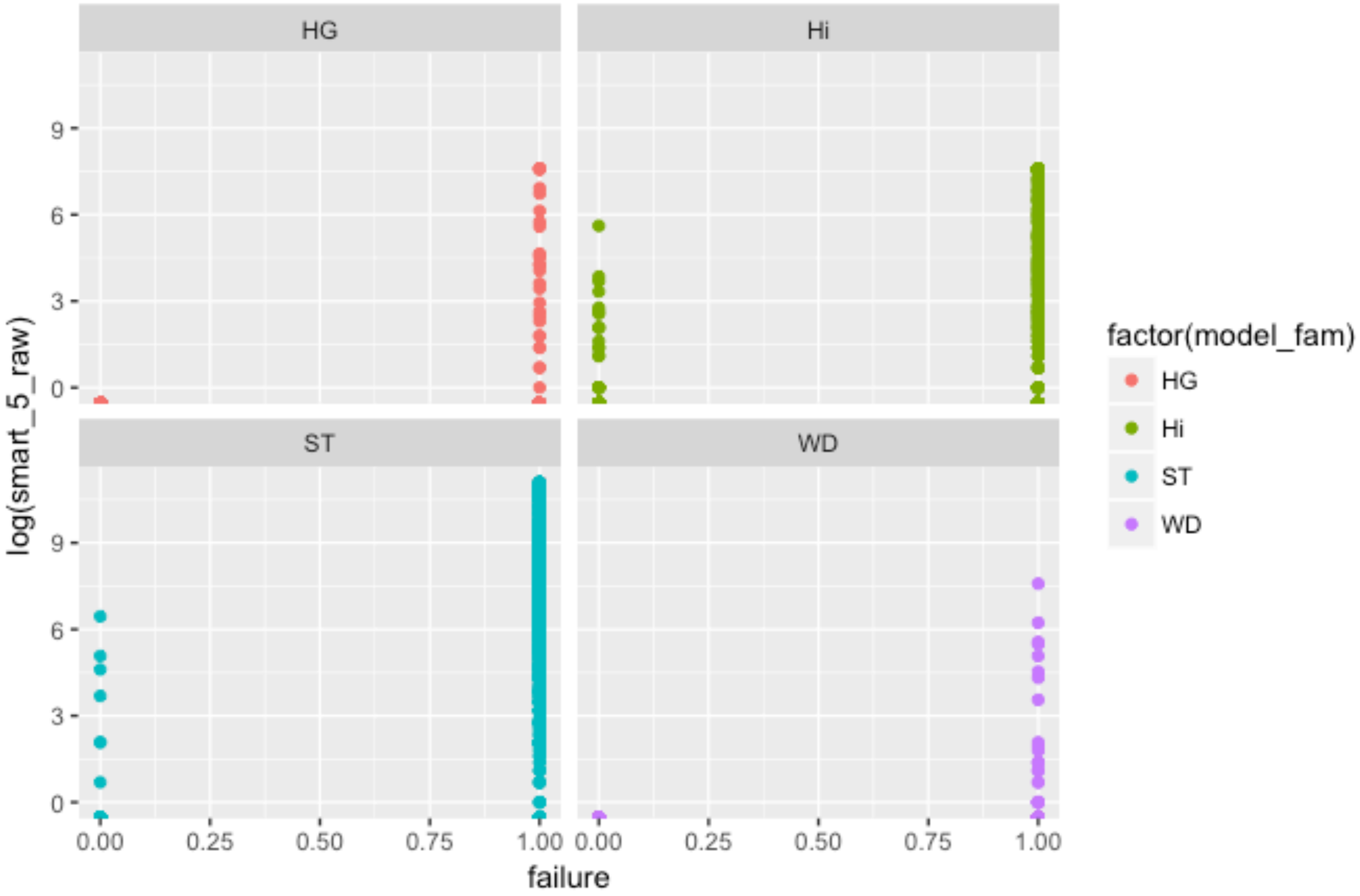
	date	serial_number	model	capacity_bytes	failure	smart_1_normalized	smart_1_raw	smart_2_normalized	smart_2_raw	smart_3_n
2	2016-04-01	Z305B2QN	ST4000DM000	4000787030016	0	117	140875840			
3	2016-04-01	MJ0351YNG9Z7LA	Hitachi HDS5C3030ALA630	3000592982016	0	100	0	136	104	
4	2016-04-01	MJ0351YNGABYAA	Hitachi HDS5C3030ALA630	3000592982016	0	100	0	136	104	
5	2016-04-01	WD-WMC4N2899475	WDC WD30EFRX	3000592982016	0	200	0			
6	2016-04-01	Z305DTP7	ST4000DM000	4000787030016	0	117	118868640			

Dataset:
- 83K drives
- 46M Drive Days (2013-2016)
- >5000 failures
- Daily snapshot for each drive's health state + SMART metrics
- SMART (Self-Monitoring, Analysis and Reporting Technology) a standard monitoring system included in HDDs and SSDs

Example SMART Attributes:
- SMART_1: Read Error Rate
- SMART_5: Reallocated Sectors Count
- SMART_9: Power On Hours
- SMART_7: Seek Error Rate
- SMART_197: Current Pending Sector Count

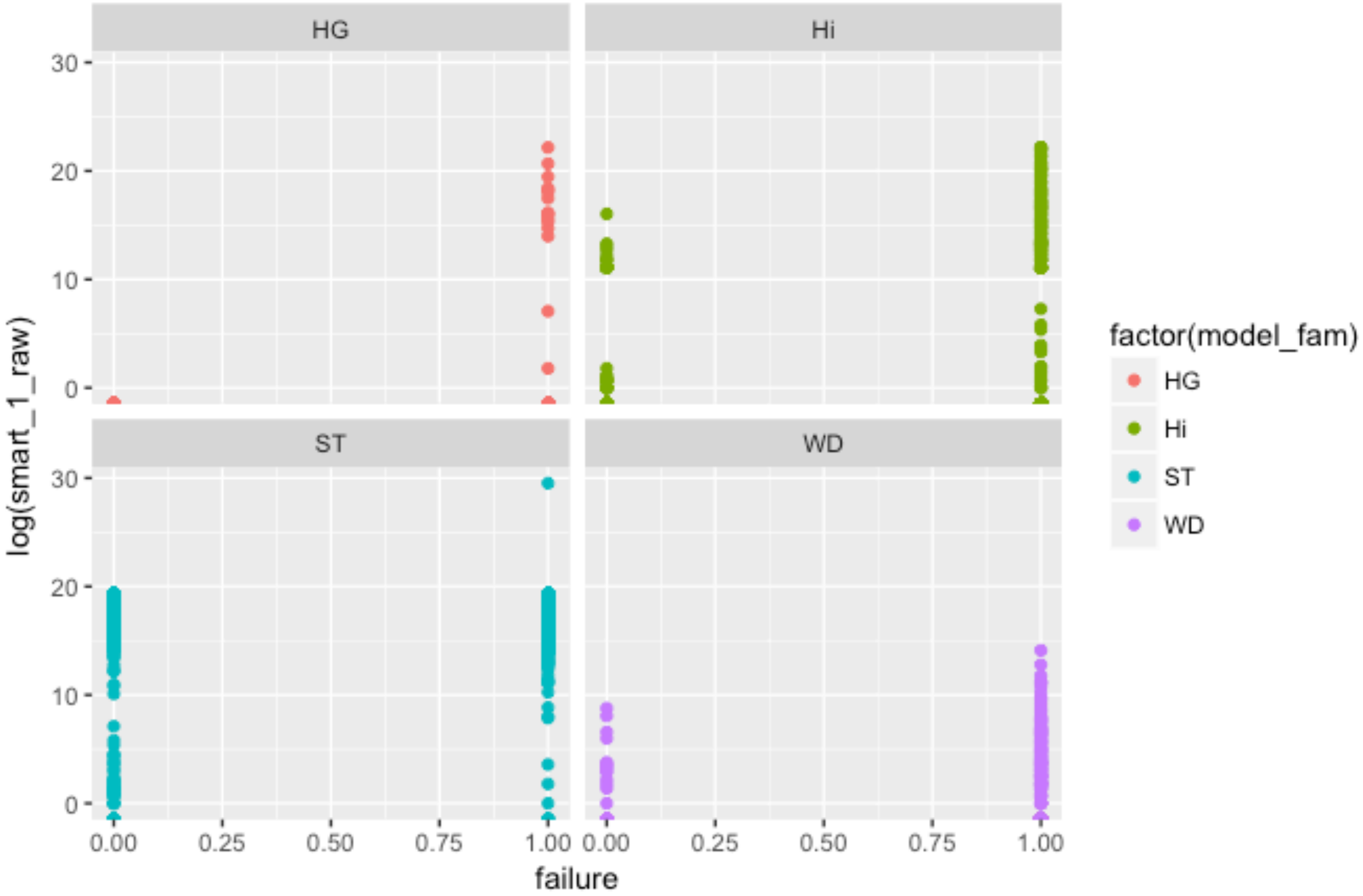
https://en.wikipedia.org/wiki/S.M.A.R.T.

SMART5: Reallocated Sector Count



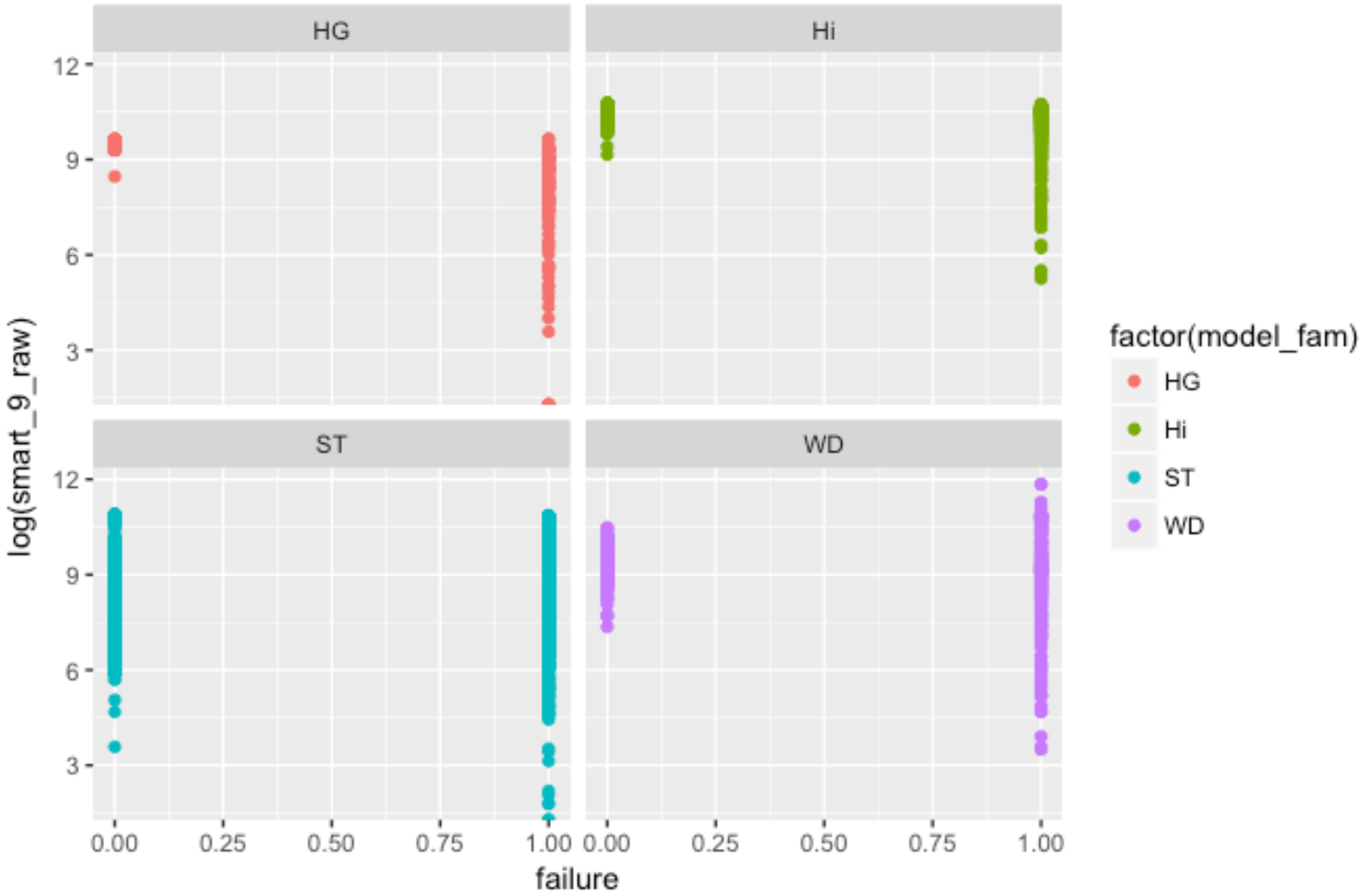
Sample of 7678 drives
(50% failed/50% healthy)

SMART1: Read Error Rate



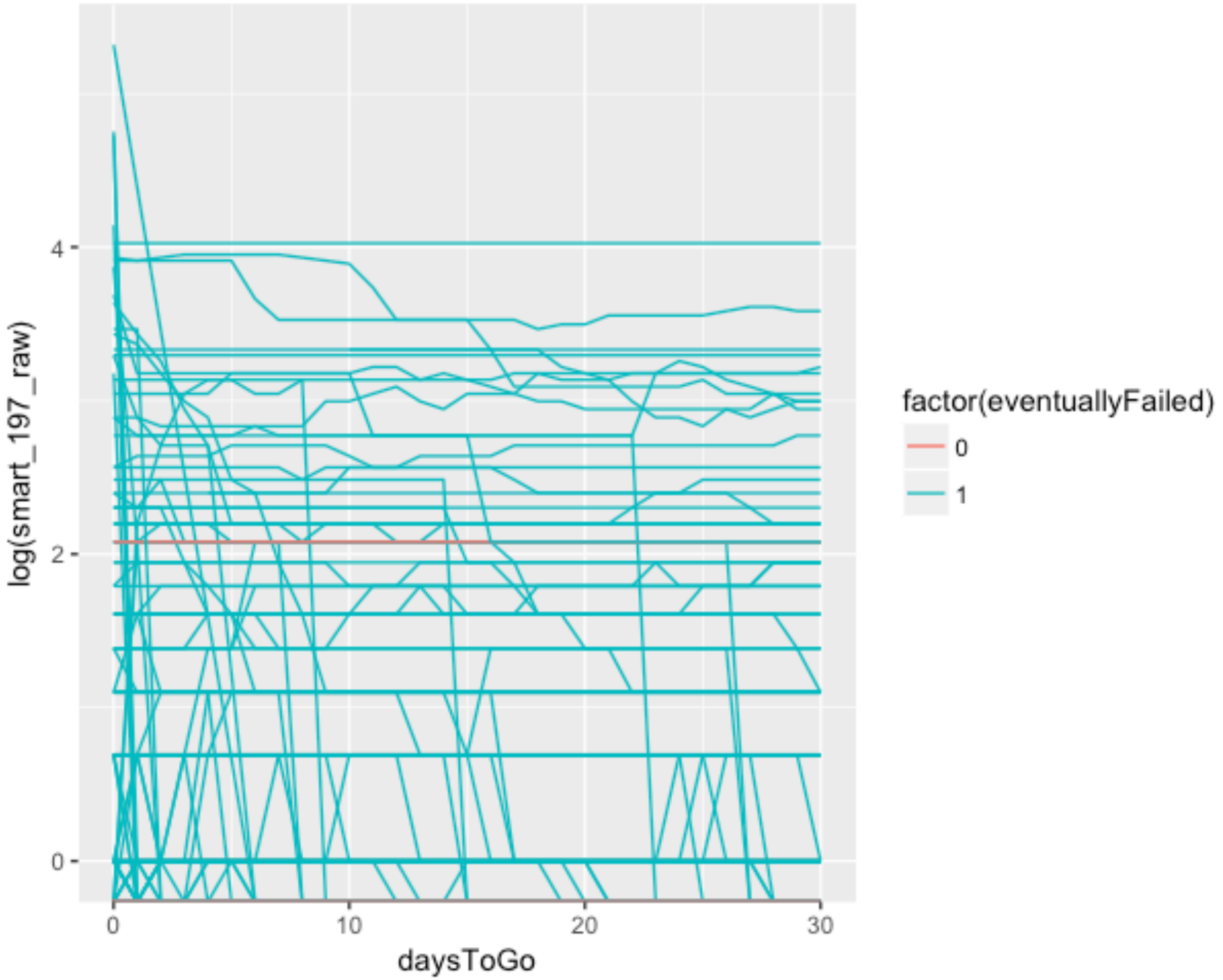
Sample of 7678 drives
(50% failed/50% healthy)

SMART9: Power On Hours



Sample of 7678 drives
(50% failed/50% healthy)

SMART197: Current Pending Sector Count over 30 days



Sample of 7678 drives
(50% failed/50% healthy)

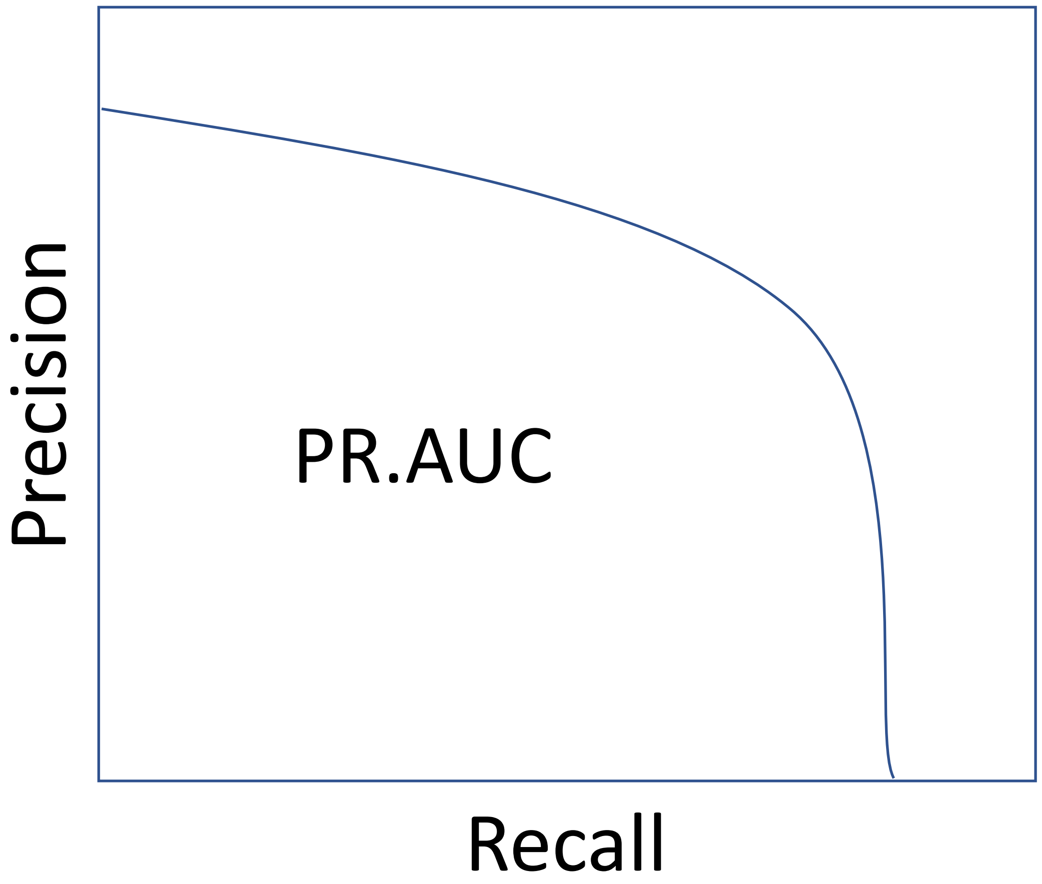
Part III

Disk failures can be predicted. Not perfectly. But better than simple heuristics.

Performance Metric

- Goals:
  - Detect rare event (1/20 or less)
  - Tune depending on Use Case
    - Tolerance for False Positives
    - vs. Tolerance for False Negatives
  - We want to maximize our ability to make better tradeoffs

- Performance Metric:
  - PR.AUC: Area Under Precision-Recall curve



Baseline Heuristic

- If any of the critical SMART attributes > 0 then the drive is likely to fail

- SMART_5: Reallocated Sectors Count
- SMART_187: Reported Uncorrectable
- SMART_188: Command Timeout
- SMART_197: Current Pending Sector Count
- SMART_198: Offline Uncorrectable

Baseline Performance

- Evaluation dataset:
  - 13980 drives
  - 699 failed
  - 13281 healthy
- Precision: 42%
  - (i.e. 58% false positives)
- Recall: 68%
  - (i.e. 32% false negatives)

Confusion Matrix

		Baseline Prediction	
		Healthy	Failed
Truth	Healthy	12625	656
	Failed	223	476

Approach

- Split the data into train/test and Eval
  - 2013-2015: Train/Test
  - 2016: Eval
- Sample from the train data at 50/50
  - (Learn equally from failure/health)
- Sample from the Eval data at 95/05
  - (Evaluate at a fixed real-life failure/healthy mix)

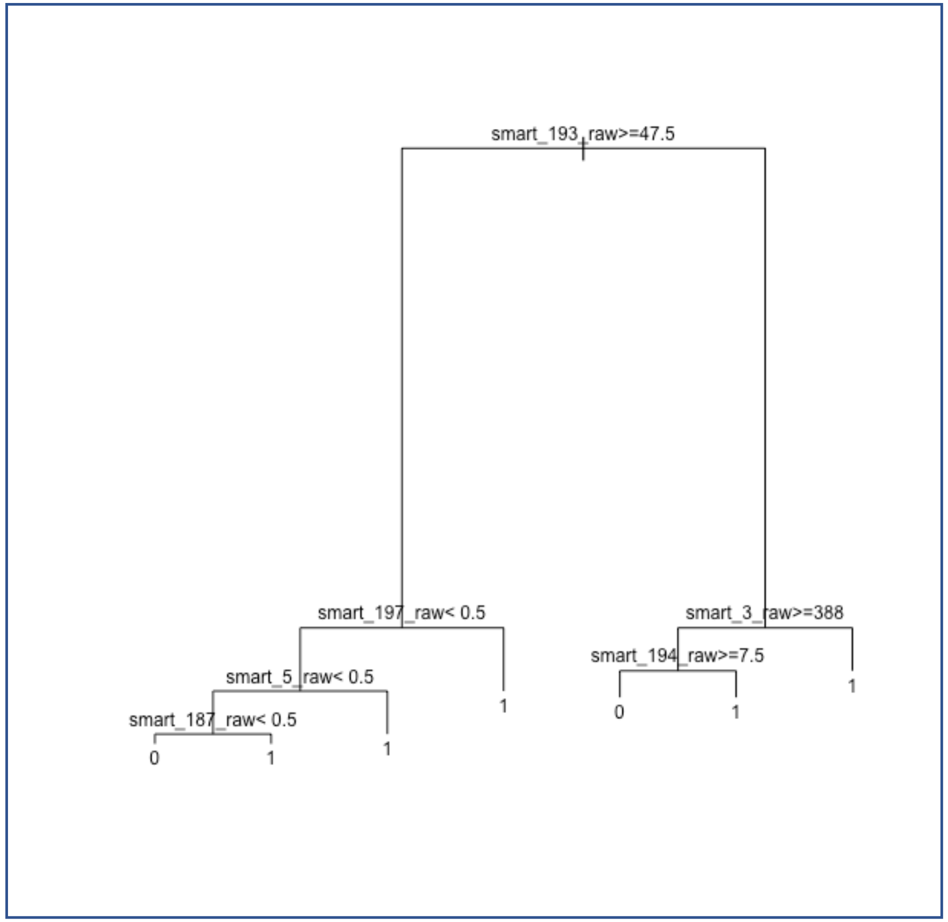
Feature management challenges

- Inconsistency of SMART data support across vendors and drive models
- Data Sparseness
- Opacity of most SMART metrics
- Further opacity of Normalized SMART values
- Wide Range for most SMART values
- Dataset skewness by vendor and model
- Some gaps and inconsistencies in the telemetry data

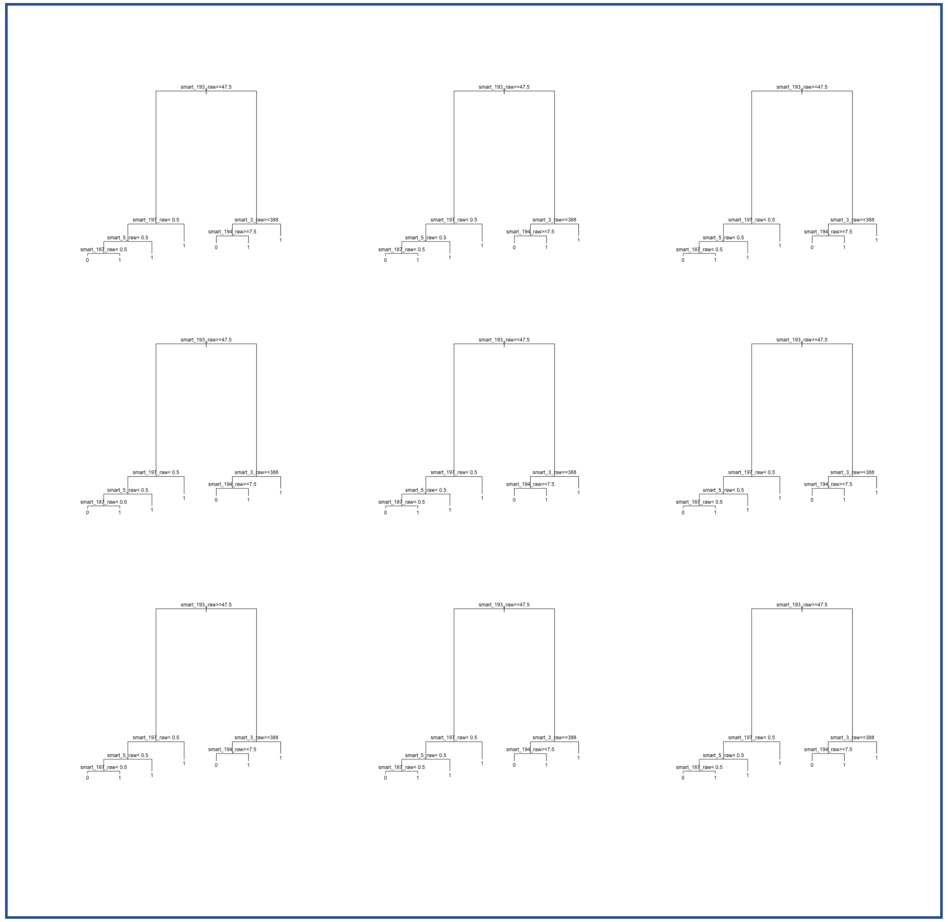
Feature Selection & Models

- Feature Selection:
  - Raw over Normalized SMART Data
  - Created model_fam feature to collapse vendor/model
  - Z-Score Normalization of RAW values
  - 3-day, 5-day rolling Variance for all SMART RAW values
  - and a few things which didn’t work as well as initially hoped ☺
- Models:
  - Random Forests
  - Logistic Regression
  - Support Vector Machines
- Goal:
  - Improve on the baseline
  - Expand options to tune Decision Threshold for various Precision vs. Recall tradeoffs (based on Use Case)

Model 1: Random Forests



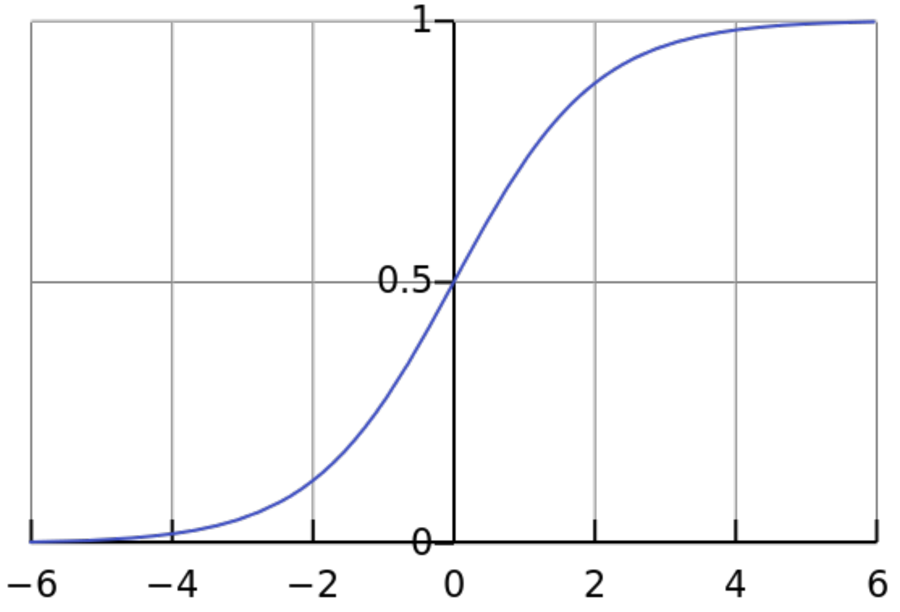
Single Decision Tree



Random Forest

Model 2: Logistic Regression

- Sigmoid (Step) function to model a binomial output (0/1)
- Works well for linear continuous numerical inputs
  - Corollary: Horribly for categorical non-linear variables appearing to be continuous numerical
- In our case the key is to:
  - Isolate right SMART metrics
  - Normalize where needed

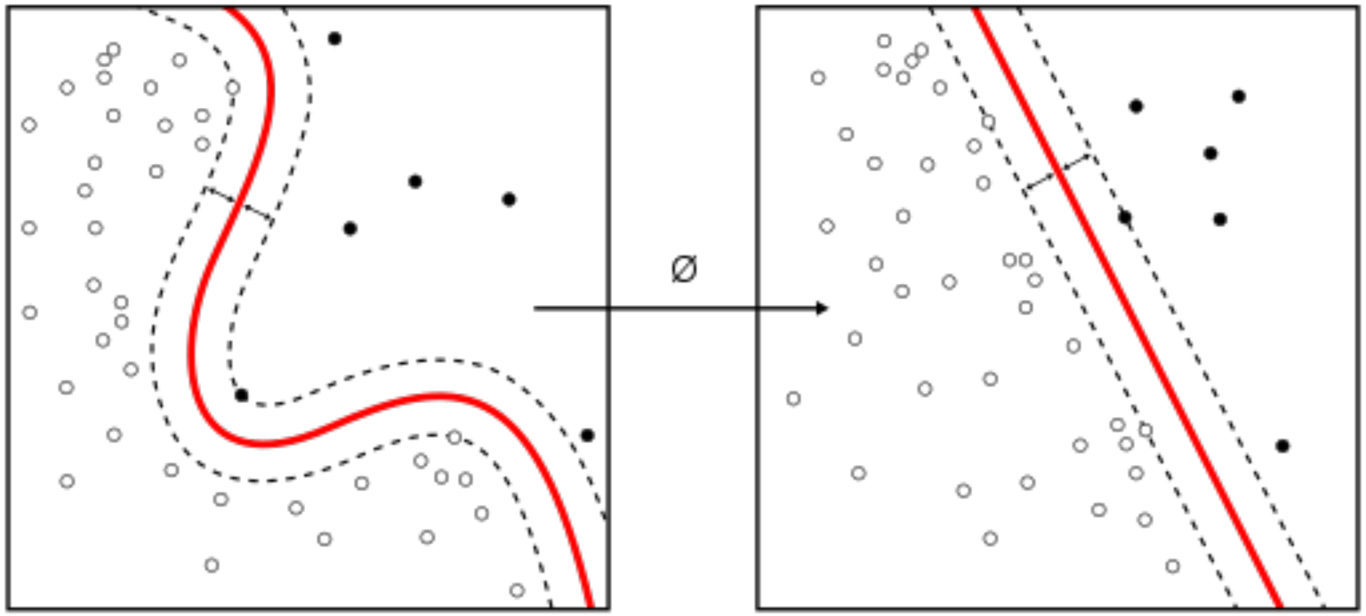


Model 3: Support Vector Machines

- Binary Classifier based on mapping non-linear data into a higher dimension to make linear separation possible

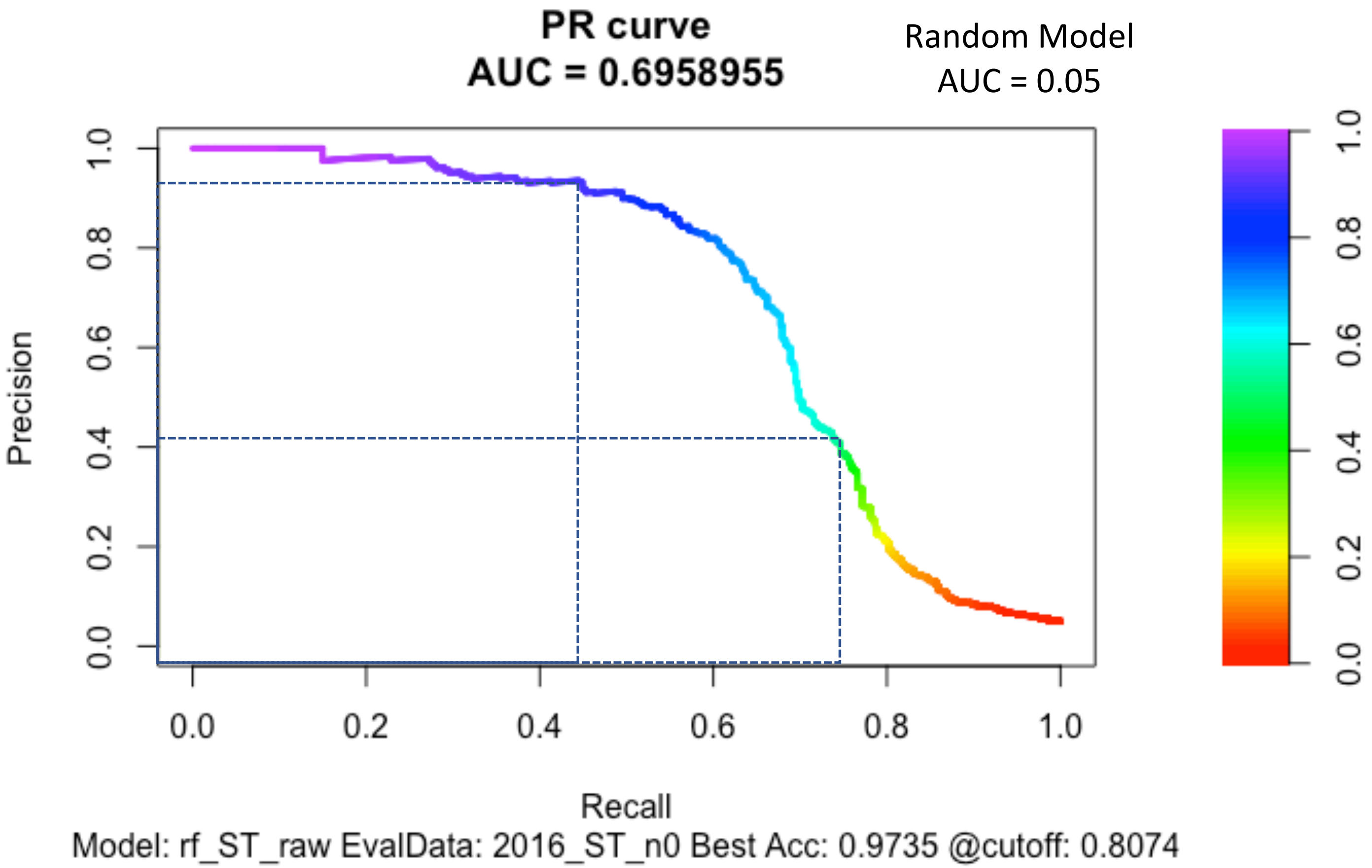
- Maximizes margin of separation between +/- categories

- In our case the key was still to isolate right SMART metrics before training

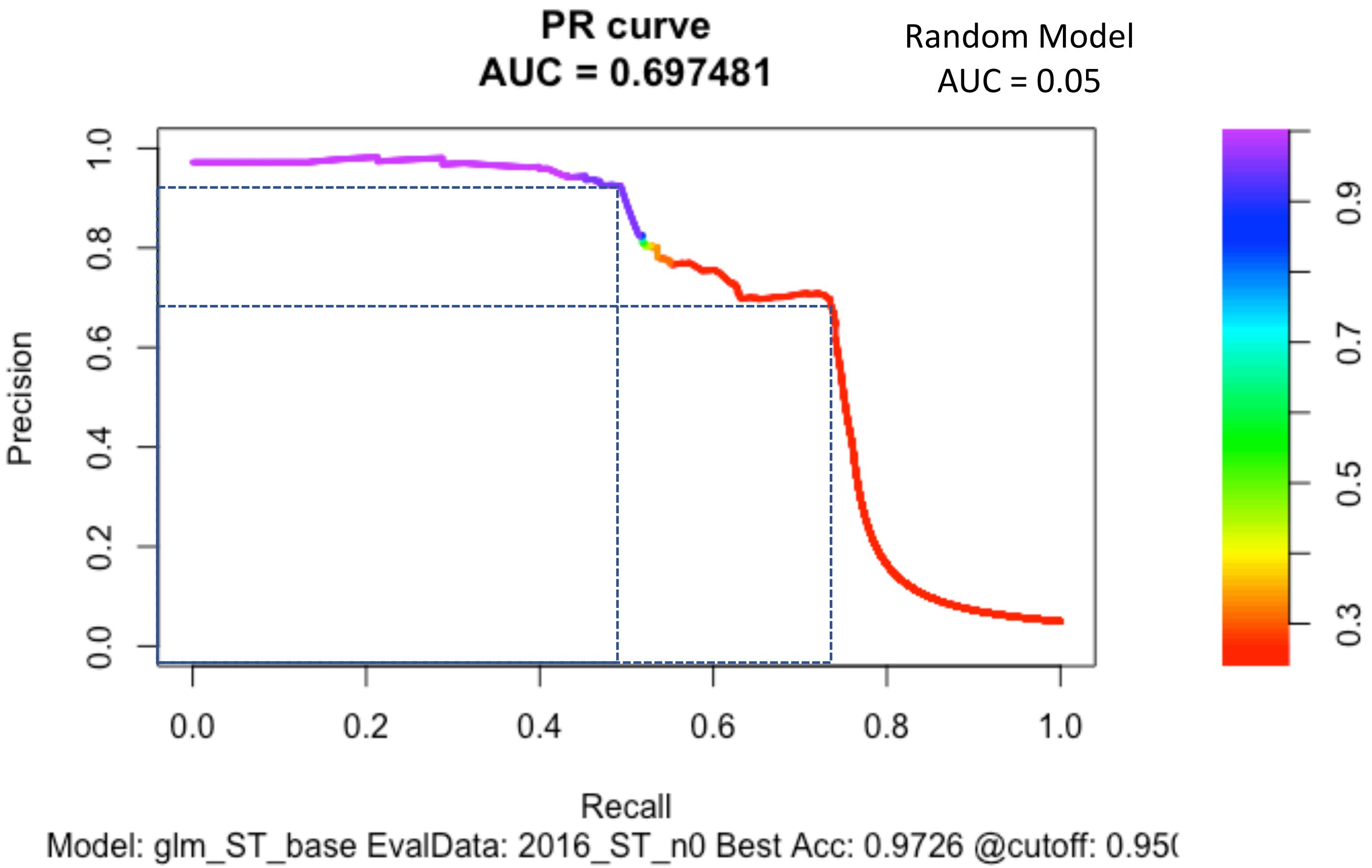


By Alisneaky, svg version by User:Zirguezi - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=47868867

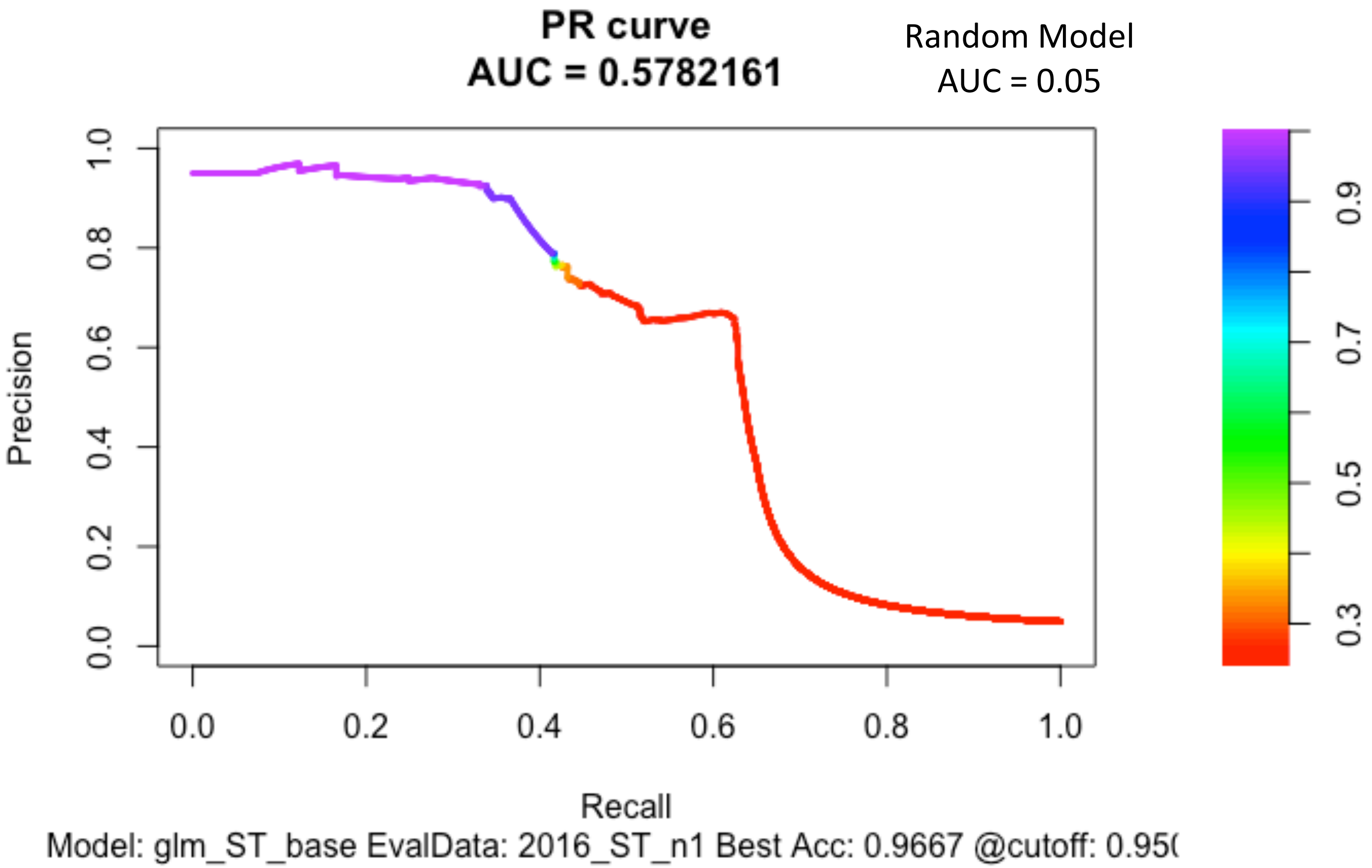
Results: Seagate drives. Day Zero



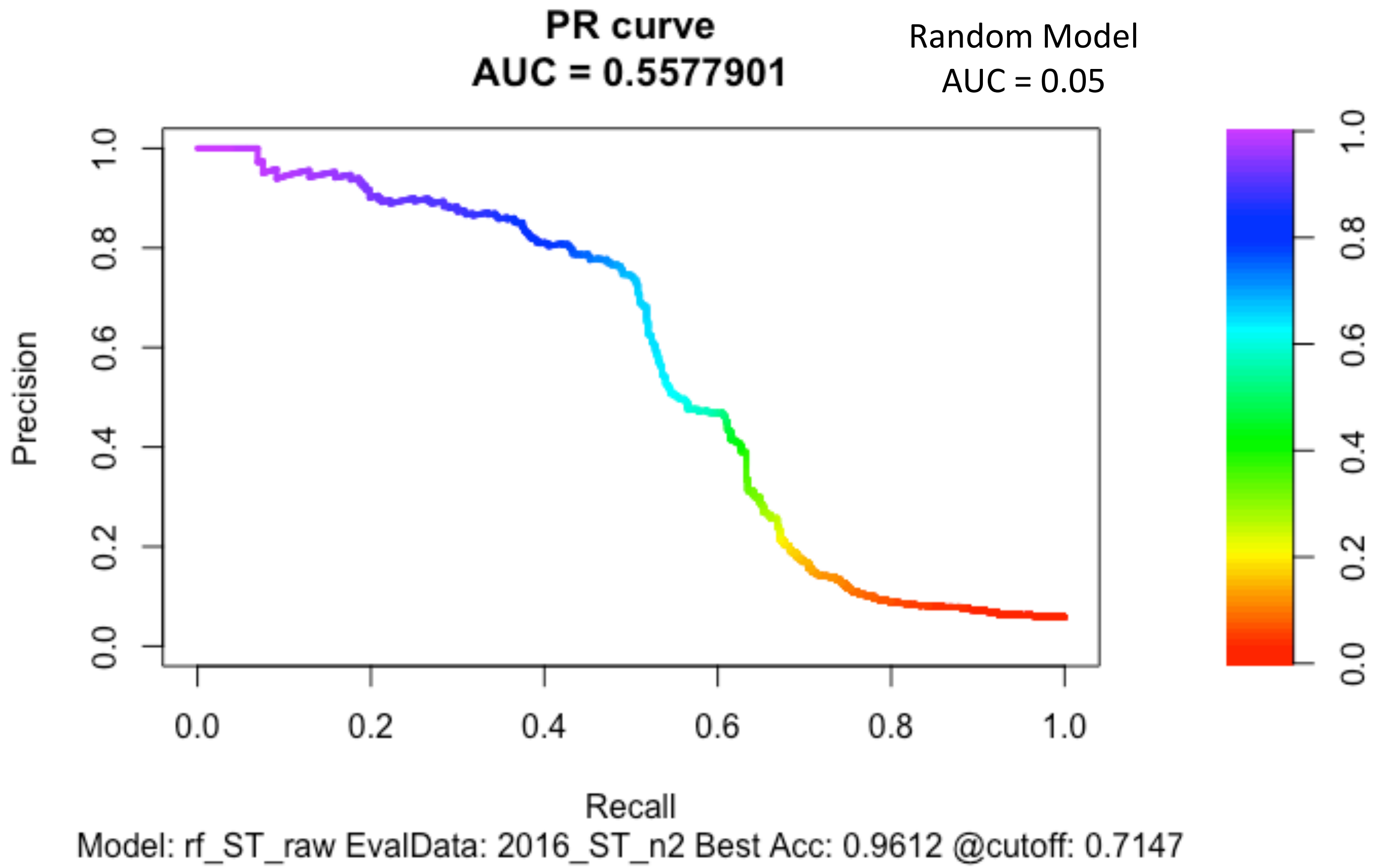
Results: Seagate drives. Day Zero



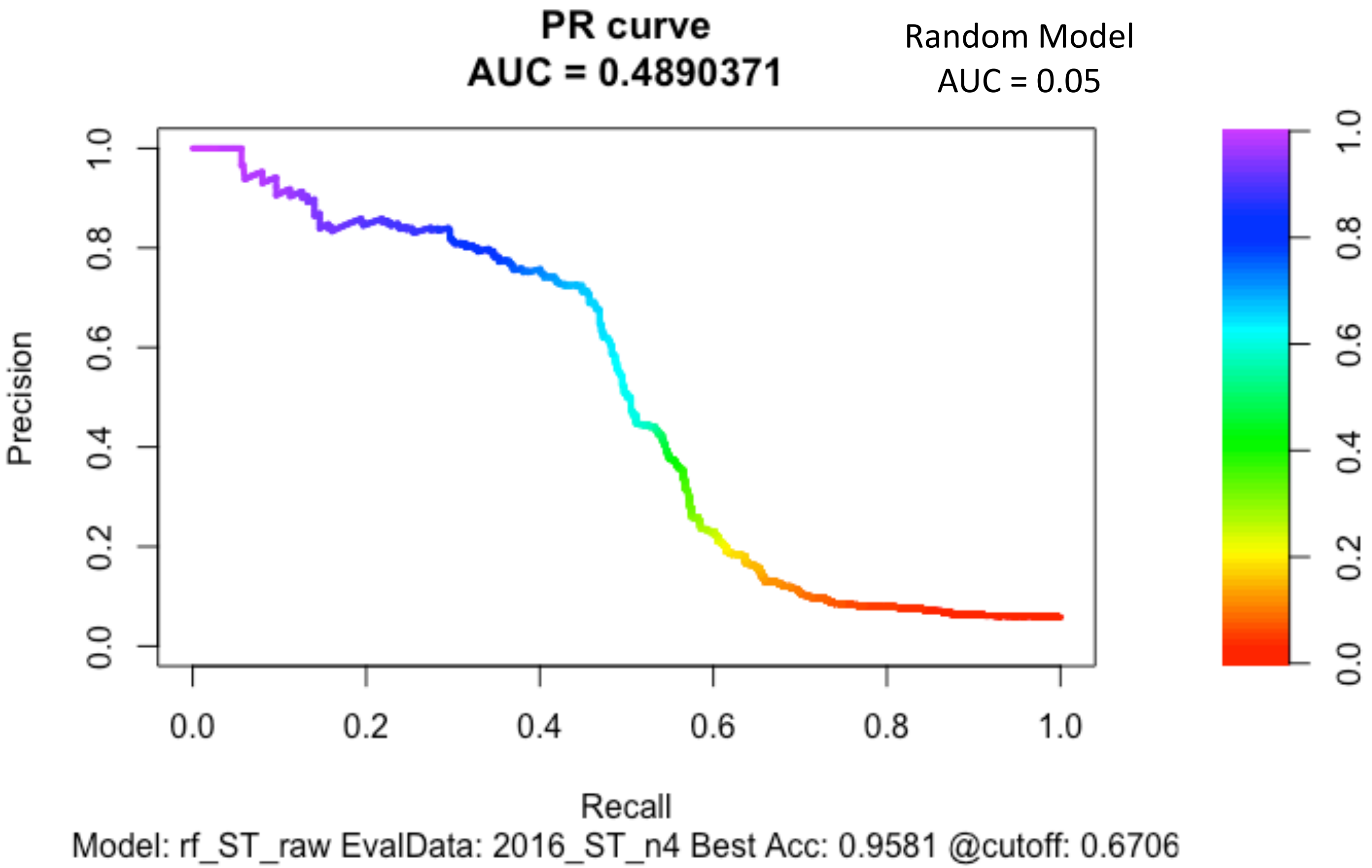
Results: Seagate drives. Day -1



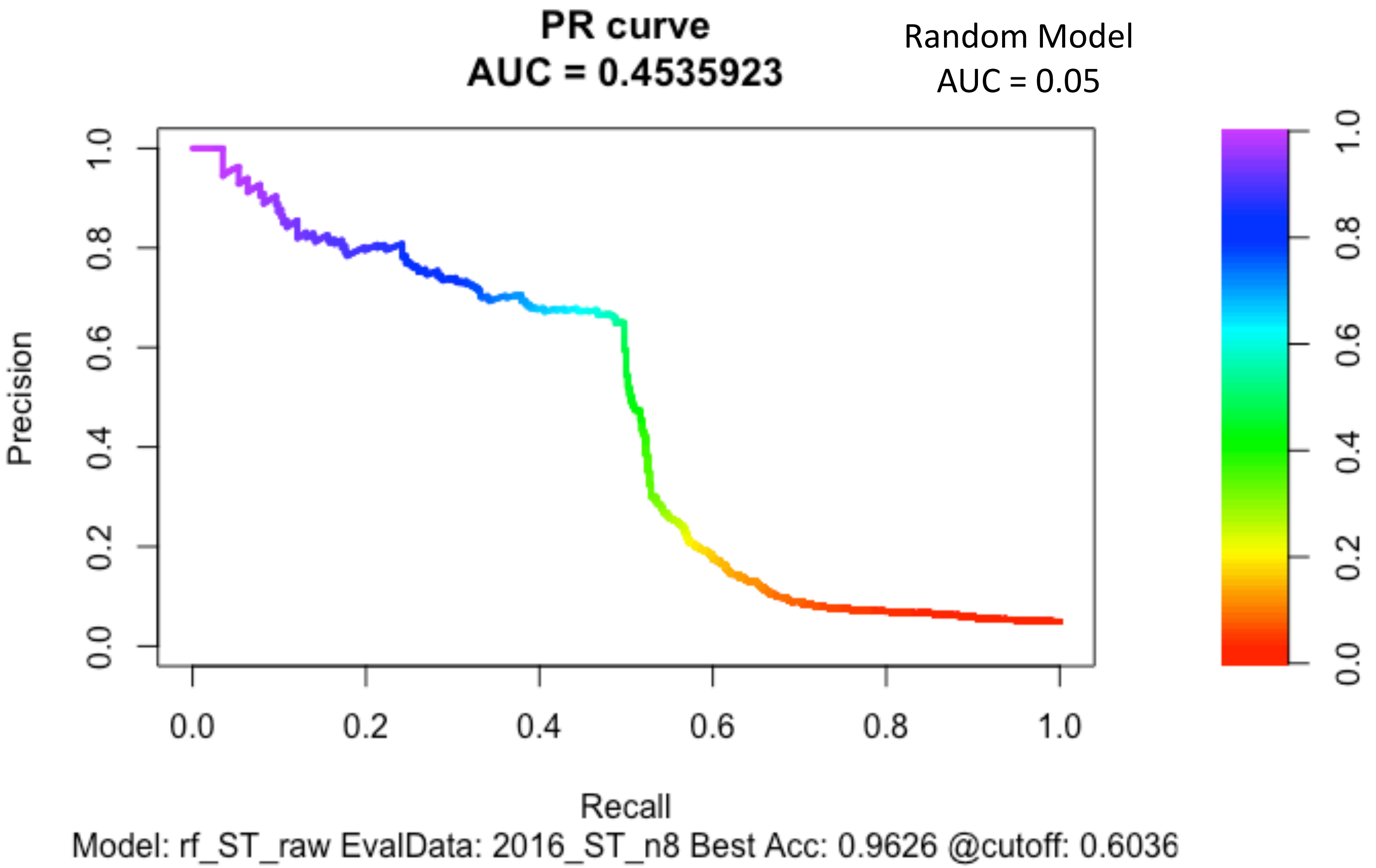
Results: Seagate drives. Day -2



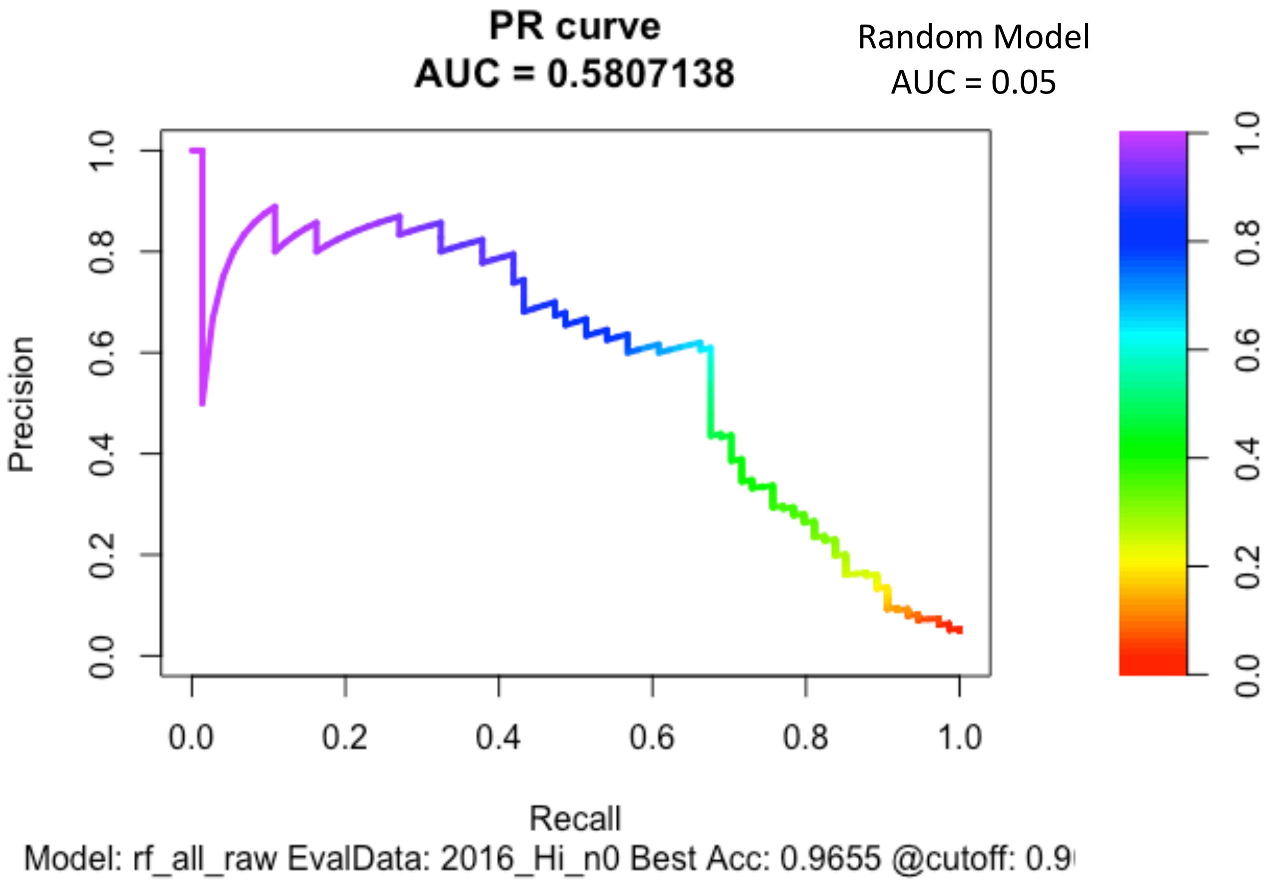
Results: Seagate drives. Day -4



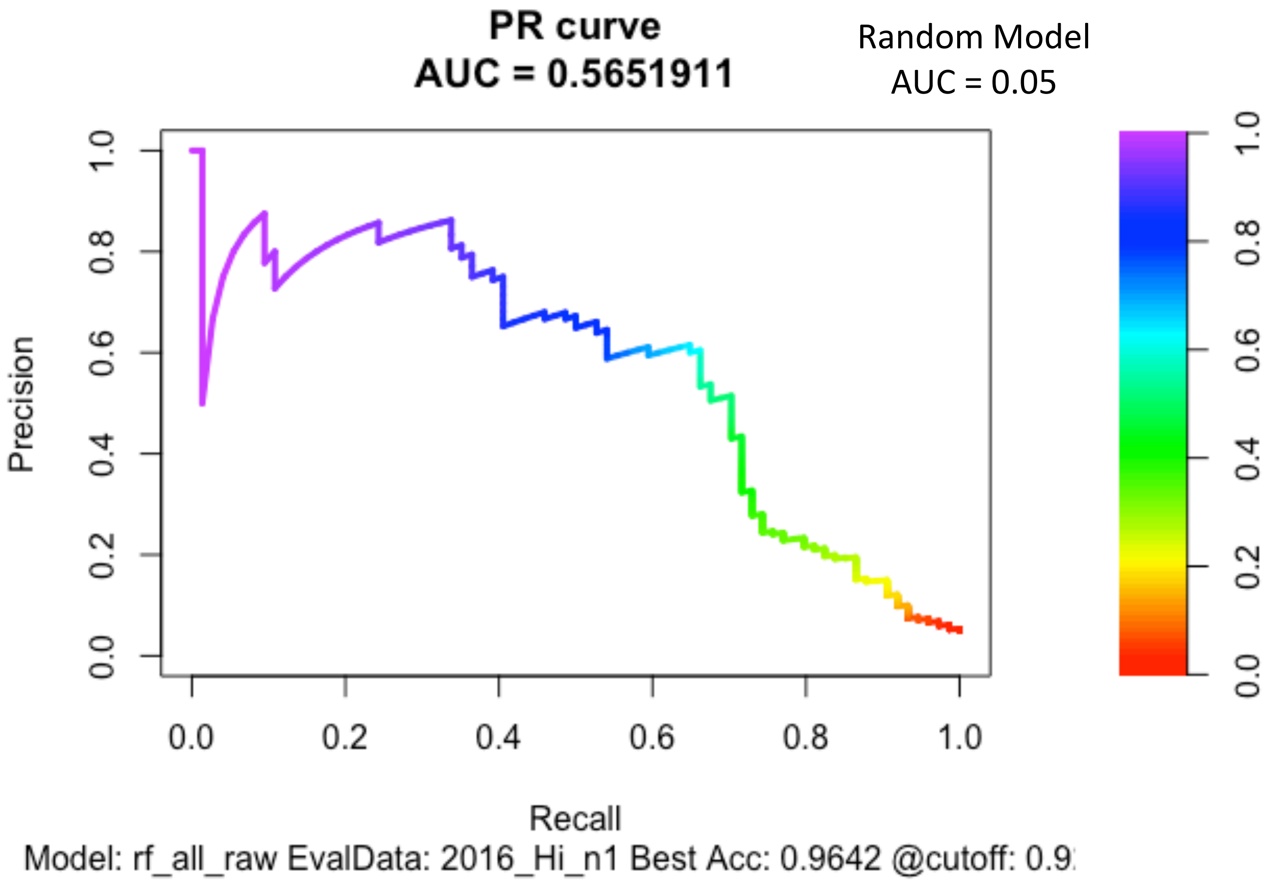
Results: Seagate drives. Day -8



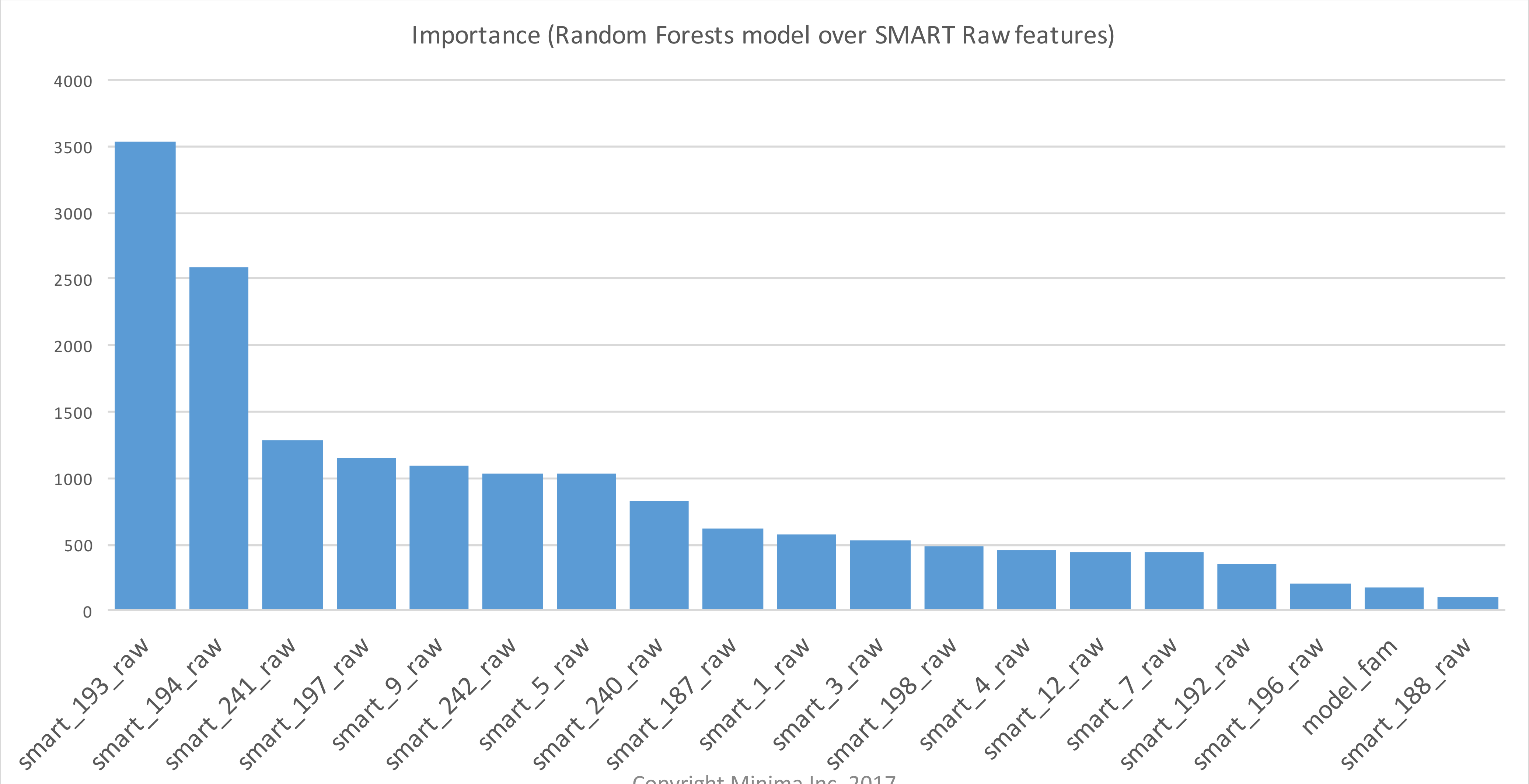
Results: Hitachi drives. Day Zero



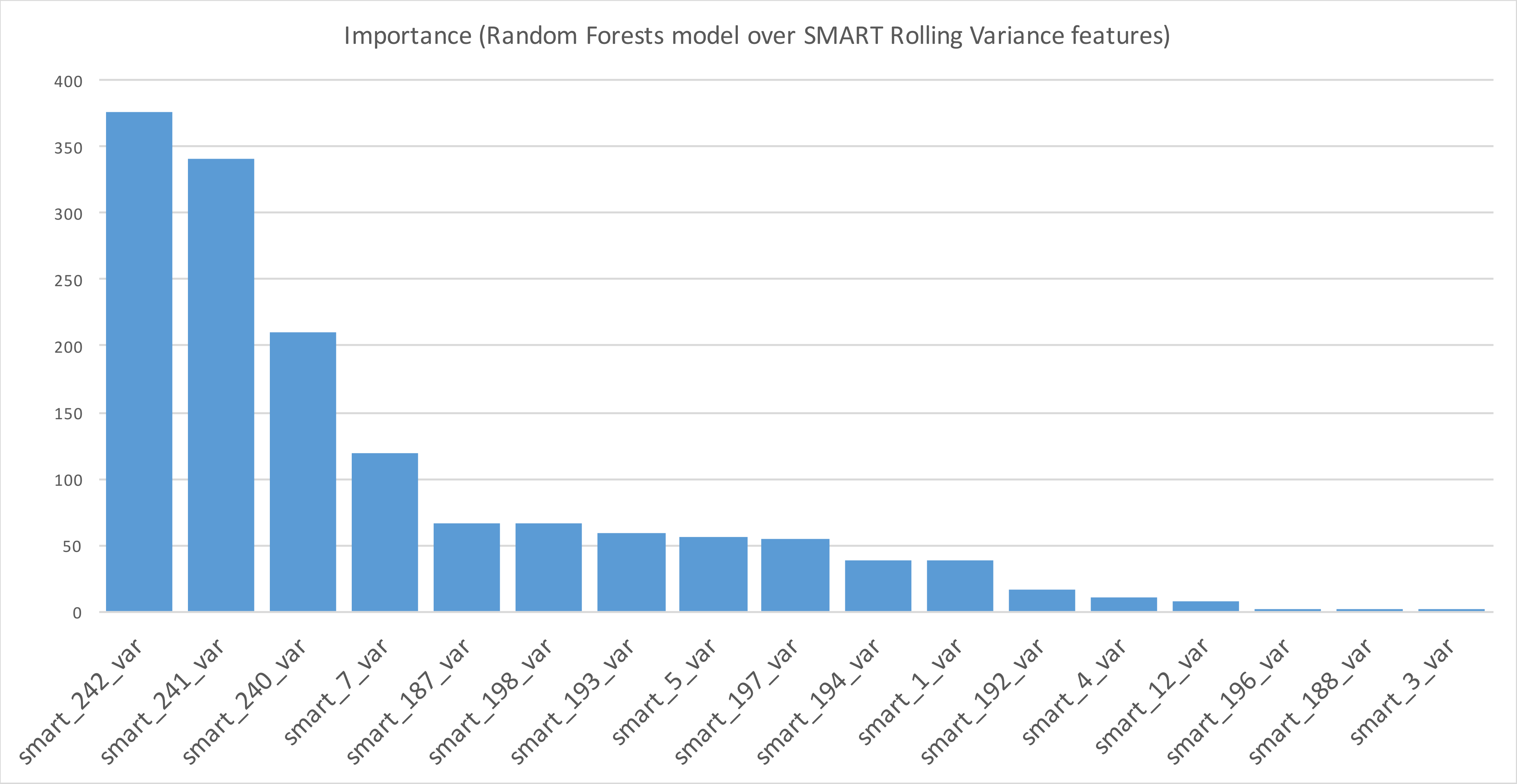
Results: Hitachi drives. Day -1



Feature Importance – SMART Raw – All Models



Feature Importance – SMART Variance - Seagate



Summary

- 2% of Disks fail annually, on average. Mileage varies by model.
- SMART metrics can clearly signal failures, sometimes days before it happens
- We can train/predict across some of the drive models overcoming training data sparsity
- We can reasonably train models to predict drive failure using SMART data and improve upon existing heuristics
- Picking and tuning the right model depends on Use Case and goals
  - Precision vs. Recall tradeoff
  - Different models give you more options
- More Data is better. Duh!