# Unit 2 Build Project Outline & Primer

These are informal notes

**Topic :**

Using BackBlaze’s data repository, I will create a machine learning classification model that will predict and classify a hard drive’s status in terms of functional or failure.

**Data Scope:**

BackBlaze has an astonishingly low failure rate, with less than 10k failures over the past 6 years. Because of this huge class imbalance, I feel it necessary to take at least half a year or a year’s data in order to have enough presence of both classes such that the majority class baseline does not prevent the model from choosing the minority class ‘failure’. My starting point is 2019 Q1 - 2019 Q2.

My data can be found at the bottom of the page here: [Backblaze Hard Drive Stats](https://www.backblaze.com/b2/hard-drive-test-data.html)

**Target:**

Failure Status / Functionality

**Features:**

* Date
* Model
* Serial No
* Capacity
* SMART Stats

As this is supposed to be a general model, I will not use the date and serial number features so that the model will not ‘learn’ on a drive by drive basis.

**What constitutes a failed drive:**

In BackBlaze’s data, a hard drive is marked functional until the day before it fails. Drives that fail are marked as failed on the previous day’s log because if the drive is failed then it can’t report it’s status at the end of the daily reporting period. From BackBlaze’s Website @ [What SMART Hard Disk Errors Actually Tell Us](https://www.backblaze.com/blog/what-smart-stats-indicate-hard-drive-failures/) Regarding failure markers:

“Drives which have failed are marked as such and their data is no longer logged. Sometimes a drive will be removed from service even though it has not failed, like when we upgrade a Storage Pod by replacing 1TB drives with 4TB drives. In this case, the 1TB drive is not marked as a failure, but the SMART data will no longer be logged.”

This reporting methodology means that a lack of entry does not mean failure. Only those flagged failure and lack of future entries in the log are failed.

---

**Features of Interest:**

[What SMART Hard Disk Errors Actually Tell Us](https://www.backblaze.com/blog/what-smart-stats-indicate-hard-drive-failures/)

In the article linked above, BackBlaze references 5 specific SMART statistics they commonly use to determine drive failure likelihood.

|  |  |
| --- | --- |
| Attribute | Description |
| SMART 5 | Reallocated Sectors Count |
| SMART 187 | Reported Uncorrectable Errors |
| SMART 188 | Command Timeout |
| SMART 197 | Current Pending Sector Count |
| SMART 198 | Uncorrectable Sector Count |

Following this, BackBlaze gives a statistical breakdown of these five stats in operation drives and failed drives; however, to avoid bias at this early stage of the process, I will not take that insight into consideration.

Other SMART Stats:

Outside of these 5 SMART Stats, BackBlaze has started reviewing SMART 189 - High Fly Writes and SMART 12 - Power Cycles in their role to determine failure candidacy. These are worth investigating. In its breakdown of the SMART System, Wikipedia lists the following as critical and potentially indicative of drive failure: SMART 5, SMART 10, SMART 184, SMART 187, SMART 188, SMART 196, SMART 197, SMART 198, SMART 201

**Managing and Engineering Features:**

An important note from BackBlaze is that it is relevant to consider that SMART Stat error reports should be considered in regards to drive usage time. 5 reported errors within a specific SMART Stat over a year is contextually relevant in comparison to the same error count within a time frame of only a couple days.

Although I mentioned above that I do not think the “Log Date” should be used as a modeling feature, engineering a feature that measures drive install/usage time is very important.

\*\* Issue \*\*

* Drive age isn’t a logged statistic, and it is not likely possible to figure out the first date of logging for every drive

EDIT: SMART 9 reports number of hours a drive has been in usage.

---

**Managing the Data:**

The data is structured as follows:

* Each .csv file is a daily (1-day) snapshot of the drive’s basic information and SMART Stats
* Date is in format YYYY-MM-DD
  + Files are named according to this, e.g. 2013-04-10.csv
* A folder of a Quarterly data is zipped together.

The files are too numerous to load in individually. Overall, the files are too large to be processed and held in RAM locally. This means file merges and preliminary data cleaning will need to take place before all of the data is able to be condensed.

Data Processing Plan:

* Load a single .csv file and do preliminary analysis and important feature identification
* Create a ‘template’ from this first .csv that I want all files to be reduced to before merging together.
* Create a script that will reduce these files and merge them into fewer files

OR:

* Create a database?
  + Would require learning how to create, manage, and access a database.

In either scenario I will have to learn methods to make my data manageable.

**Train / Validation / Test:**

Splitting my dataset from the selected range is another reason my data management plan is so important. Because of the nature of the data and needing to frame it in the context of changes over time, Date Splits will likely result in mismatched proportions of class presence. My current plan is to use train\_test\_split and stratification based on failure so that each set has a proper proportion of operational/failure.

**Evaluation Metrics:**

Due to the large class imbalance, accuracy scores would likely be misleading. Choosing an appropriate metric will require further exploration and insight into the data than what is currently available right now. Researching various metrics will be required.

---

**Data Structuring:**

Creating a plan for how to structure my data

* Each drive should have a unique Serial Number. This can be used to track and ‘index’ the drive day by day.
* Data needs to track a Serial Number’s SMART stats over time.
  + But this must also function in a way that it can process thousands of drives’ data over time

What does a single day's worth of logging convey?

* The state of 110K drives
* The statistics of a drive up until it fails

Is tracking the drives over time necessary?

* Tracking drives over time would allow the model to predict trends in drive behavior based on model / manufacturer / capacity

What can the model solve by only training on a single day’s worth of data

* Potentially identify stats that are critical on the near-failure drives (because last log for failed drives is the day prior)

Does data need to be arranged such that each observation/drive is in-line with daily data?

Example:

**Wide**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Drive Serial | Info1\_Day1 | Info2\_Day1 | …. | InfoX\_Day1 | Info1\_Day2 | Info2\_Day2 |
| 1ZXLM000001 |  |  |  |  |  |  |
| 1ZXLM000002 |  |  |  |  |  |  |

Or

**Wide Transposed**

|  |  |  |
| --- | --- | --- |
|  | 1ZXLM000001 | 1ZXLM000002 |
| Info1\_Day1 |  |  |
| Info2\_Day1 |  |  |
| …. |  |  |
| InfoX\_Day1 |  |  |
| Info1\_Day2 |  |  |

Or

**Tidy**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Info1 | Info2 | ….. | InfoX |
| 1ZXLM000001 | Day 1 | Day 1 | Day 1 | Day 1 |
| 1ZXLM000002 | Day 1 | Day 1 | Day 1 | Day 1 |
| 1ZXLM000001 | Day 2 | Day 2 | Day 2 | Day 2 |
| 1ZXLM000002 | Day 2 | Day 2 | Day 2 | Day 2 |
| …. |  |  |  |  |
| 1ZXLM000001 | Day X | Day X | Day X | Day X |
| 1ZXLM000002 | Day X | Day X | Day X | Day X |

**Data Structuring Pros & Cons**

Before discussing data orientation, basic shape should be considered

Number of drives: 106,918

Number of “Info” per day (before cleaning): 129

Wide Structure:

Pros:

* Basic structure data already exists in
* Merges will mostly be concatenation of tables on serial number

Cons:

* Each new day of data adds 129 columns, which can quickly increase the complexity significantly
  + 365 \* 129 = 47085 columns
    - Final shape roughly 110k x 47k

Wide Transposed Structure:

Pros:

* Relatively fixed number of columns

Cons:

* Organization will be more difficult since some drives are replaced and new ones added

Tidy Structure:

Pros:

* Merging multiple days’ datasets will not increase number of columns
* Effectively multiply sample rate and data if model is not time-sensitive

Cons:

* Lose the identification and in-line nature of using serial number as index for the row
* It will be up to the model to draw that each repeated serial number is the same unit
  + Can be remedied with clever encoding
  + No different than players’ names on a sports team
  + Is tracking drives over time and each day relevant?

---

# General Notes

Here I will write general information gleaned from my process

**Raw and Normalized Values**

* Raw values are generally set by the drive manufacturer, and each attribute is relatively up to that party.
  + Can be either a decimal or hexadecimal
* Normalized values attempt to solve this problem of units/values/scale by expressing the values on a range of 1 to 253
  + Normalized 1: Worst Case
  + Normalized 253: Best Case
  + Initial Default: 100
    - Can vary by manufacturer

**The SMART System**

* The SMART System is a statistical data reporting system hard drive manufacturers build into their products
* Due to it not being standardized across brands, some manufacturers measure the statistics in different units / attributes
  + Would this be a reason to group drives by manufacturer?

**SMART 9**

* Power-On Hours
* A count of hours (or minutes or seconds) in power-on state.
  + Exception to Rules?

**SMART 22**

* ????????????
* <https://www.backblaze.com/blog/smart-22-is-a-gas-gas-gas/>

**Why no raw values?**

* <https://www.smartmontools.org/wiki/TocDoc#SMARTAttributes>

**????**

* <https://www.micromat.com/product_manuals/drive_scope_manual_01.pdf>

**SMART 168**

* Attribute 168 - Maximum Erase Count by Spec This attribute is the maximum number of programerase counts by specification.

---

# Resources:

**BackBlaze Articles & Blog Posts:**

* Data Home Page: <https://www.backblaze.com/b2/hard-drive-test-data.html>
* Official 2019 Q1 Report: <https://www.backblaze.com/blog/backblaze-hard-drive-stats-q1-2019/>
* Official 2019 Q2 Report: <https://www.backblaze.com/blog/hard-drive-stats-q2-2019/>
* Official 2019 Q3 Report: <https://www.backblaze.com/blog/backblaze-hard-drive-stats-q3-2019/>
* What SMART Stats Tell Us About Hard Drives: <https://www.backblaze.com/blog/what-smart-stats-indicate-hard-drive-failures/>
* Hard Drive SMART Stats: <https://www.backblaze.com/blog/hard-drive-smart-stats/>

**Wikipedia:**

* SMART: <https://en.wikipedia.org/wiki/S.M.A.R.T.>