Hard Disk Drive Reliability Statistics and Analysis

Group – 12

Submitted by: Saurabh Bhagwat (17211349)
Lavleen Bhat (17210637)

Contents:

1. Introduction

With increase in popularity and usage of Cloud systems, there is an ever-increased demand of reliable storage centres. When companies setup their data centres, it’s important to know which manufacturer can provide the best hard disk drive to provide the uninterrupted storage services to clients or for their own use. To make an informed decision, companies can either first need to go through and use all the hard drives available, use their market knowledge along with monitoring the hard drives performance to reach a decision but technologically we’ve become so advanced that we don’t need to go through that cumbersome and time consuming task when instead we can just utilize the already available data to get insights to make a good decision. And that is the main motivation behind this project – to help companies make an informed decision about the manufacturer they want to purchase hard drives from, for reliable and durable storage of data.

Now, to make a good decision, we need to dig into the huge amount of historical data to gain fruitful insights.

2. Data Understanding

Backblaze has been recording data about the hard drives they use since 2013. Recently, they published this data publicly, so anyone can utilize the data to generate statistical reports from that data.

Here is the link to the data source:

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

The dataset consists of snapshots of each operational hard drive on each day of the year. The data is available for last 5 years i.e., 2013-2017. The daily snapshot files record the data in below format:

- Date
- Model
- Capacity
- Serial Number
- Failure (0 &1)
- SMART Stats (Temp, Power-on-hours, Command-timeout, etc.)

Hard drives record a total of 45 SMART statistics about their current state containing both, normalized and raw values of the state.

Based on our research, we figured out that out of all the 45 SMART statistics, major contribution in determining the health condition of hard drive can be attributed to below five:

SMART 5	- Reallocated Sector Count.
SMART 187	- Reported Uncorrectable Errors.

- SMART 188 - Command Timeout.
- SMART 197 - Current Pending Sector Count.
- SMART 198 - Offline Uncorrectable

Not to miss any important SMART statistics, we went through all the SMART statistics one by one to understand what impact other SMART statistics can have on the health of hard drive. We finally shortlisted additional below SMART statistics to study their effect on the failure rate/health of hard drives.

- SMART 2 - Throughput performance
- SMART 194 - Temperature (Celsius)
- SMART 9 - Power on hours
- SMART 192 - Power-Off Retract Count

3. Data Quality

Now, that we have data, we need to find out how suitable the data is to start analysing it.
Data quality is as below:

- Missing Data
    - Data is missing in SMART statistics.
    - There are no missing values in first 5 columns.

- Data is normalized
    - As the SMART statistics already record current state in raw as well as normalized form, we picked only normalized columns for further analysis.

- No out-of-bound values
    - Since, SMART statistics are already normalized and recorded in the range 1 to 253 (1 – worst and 253 – best), there are no outliers in the data.

4. Data Cleaning and Preparation

4.1 Data Transformation

- Capacity of the hard drives are recorded in Bytes. So, converted the capacity from Bytes to Tera-Bytes.

4.2 Feature Extraction

- Find out the name of the manufacturer from Model
- Extract month from Date field

4.3 Imputation

To handle missing data, we divided imputation into two parts.

Imputation Part 1	**Imputation Part 2**
• Some of the SMART statistics are not recorded by some manufacturers. There is no information regarding the same on the web. • So, in part 1 of imputation, we're marking such SMART statistics for specific manufacturer as 'NOT RECORDED'. **Logic:** • To find out which manufacturer does not record which SMART statistics, we first counted the total number of hard drives available for that manufacturer. • Then, took out the total number of null values for each SMART statistic for each manufacturer. • If both the numbers (total null and total count of hard drives) match, cells are marked as 'NOT RECORDED'.	• For imputing remaining null values, we first decided to find out the percentage of null values for each manufacturer for each SMART statistic. **Logic:** • Null values percentage > 50%. If the percentage of null values present in any SMART stat for a manufacturer is greater than 50%, cells are again marked as 'NOT RECORDED'. • To impute rest of the null values, we took out mean of each SMART statistic grouped by manufacturer and imputed the null values of that SMART for a particular manufacturer with mean value of that specific manufacturer only.

5. Modelling

5.1 Statistical Analysis

5.1.1 Main challenge

The main problem that we faced initially was the size of the dataset. On each day, there are around 60k-70k operational hard drives, which increases the size of the data. We faced memory issues while loading the data into python notebook for analysis with multiple failed attempts. We tried the following methods to handle the huge amount of data:

- Dask (Python library for parallel computation) – Didn't work
- SqlLite – Didn't work
- Pyspark – Didn't work
- Sampling (Aggregation) – Worked for us

We faced similar issues while implementing the first three methods. Finally, we used sampling and aggregated the results of each sample. We divided the yearly data into 6 samples of 2 months each and aggregated the results of each sample and stored it in a separate csv file. We then combined all the csv files of all the aggregated results and calculated the final results.

5.1.2 Analysis

To do the statistical analysis on the hard drives data, we explored the data in mainly two dimensions:

1. Failure rate of hard drives
2. Average life expectancy of hard drives

Both the analysis dimensions are divided into analysing the data for each manufacturer and for each model as well.

5.1.3 Results of statistical Analysis

1. The results of the failure rate are as follows:

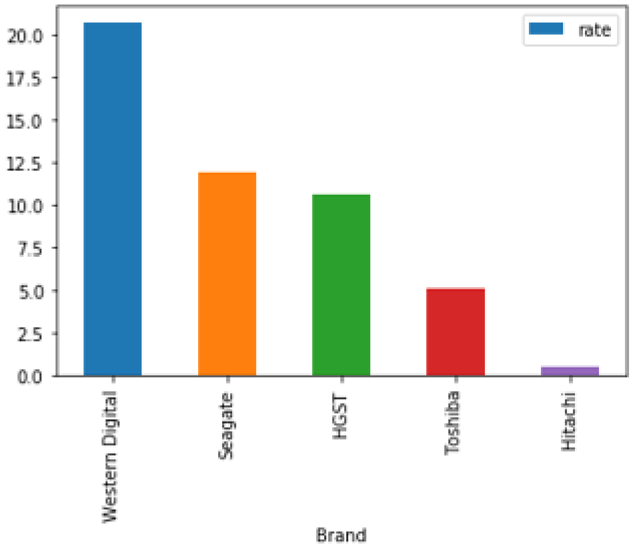


Fig 1: Failure Rate of top 5 hard drives manufacturer

Western Digital has the highest failure rate whereas Hitachi has the lowest failure rate. We also found out the failure rate model wise. Following is a list of top 5 models having the least failure rate.

Model	**Failure Rate**
HGST HMS5C4040ALE640	0.19%
HGST HMS5C4040BLE640	0.4%
Hitachi HDS5C4040ALE630	0.5%
ST8000NM0055	1.26%
ST8000DM002	1.28%

Table 1: Failure Rate for top 5 hard drive models

2. The results for the average life of the hard drives based on Brands and Models are as follows:

Brand/Manufacturer	Average Life (years)
Western Digital	2.9
HGST	1.82
Seagate	1.72
Toshiba	1.17

Table 2: Average Life expectancy for top 4 drive manufacturers

Model	Average Life (years)
ST3160318AS	6.59
C WD3200AAKS	6.51
Hitachi HDS722020ALA330	6.26
ST3160316AS	5.87
WDC WD5002ABYS	5.07

Table 3: Average Life expectancy for top 5 hard drive models

5.2 Prediction of failure of hard drives

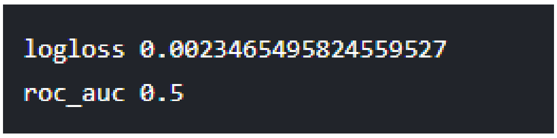
In this section, we’re studying the impact of shortlisted SMART statistics on the failure of hard drives. We tried to predict whether a drive will fail in future or not based on current SMART statistics for each manufacturer present in the data.

Input features: shortlisted SMART statistics

Target: Failure (0-Good, 1-Failure)

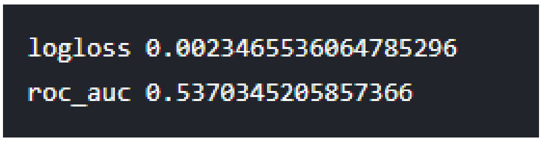
Models used:

1. Logistic Regression:
   For predicting the failure of hard drives based on current SMART statistics, we trained the Logistic Regression Classifier. Here are the results:



```
logloss 0.0023465495824559527
roc_auc 0.5
```

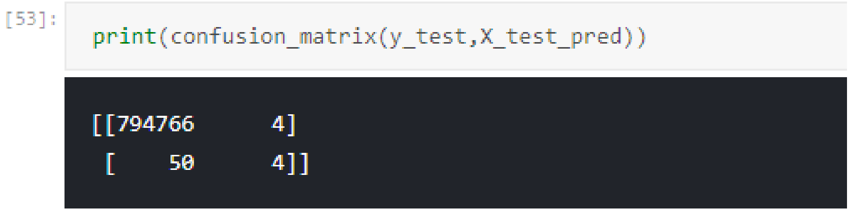
2. Random Forest Classifier:
   Since, we did not get a good accuracy, we decided to implement a Random Forest Classifier to compare the accuracy. Below are the results:



```
logloss 0.0023465536064785296
roc_auc 0.5370345205857366
```

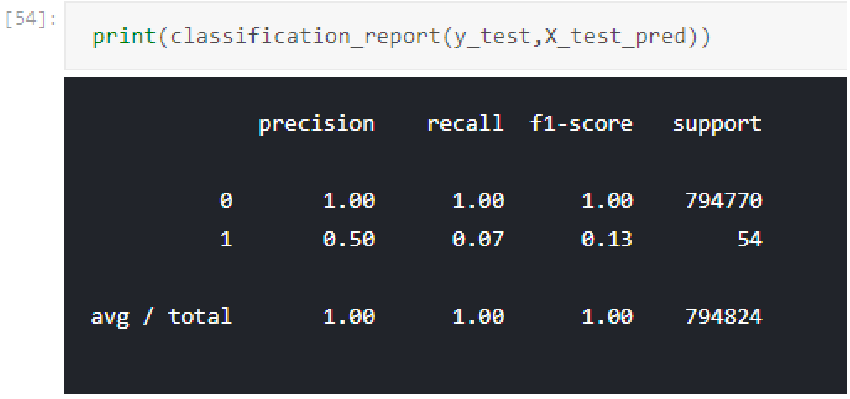
The results are not at all good because of the class imbalance in the sample of dataset we picked. Due to limitation in the size of the data we could upload in the notebook, models were trained on sample dataset and thus because of class imbalance model could not learn to predict failure (value 1) correctly.

We used other metrics such as Confusion Matrix and Classification Report just to be sure that we are evaluating the models correctly. Here are the results:



```
[53]: print(confusion_matrix(y_test,X_test_pred))

[[794766      4]
 [    50      4]]
```



```
[54]: print(classification_report(y_test,X_test_pred))

             precision    recall  f1-score   support

          0       1.00      1.00      1.00    794770
          1       0.50      0.07      0.13        54

avg / total       1.00      1.00      1.00    794824
```

6. Future work

To extend the project, the data can be used to do survival analysis of the hard drives.

7. Conclusion

There are a lot of hard disk drive manufacturers out there in the market. To understand which manufacturer provides the best hard drives, a statistical analysis is done on the Hard Disk Drive historical data. Results show that Western Digital has the highest rate of failure whereas Hitachi comes out to be best manufacturer in building robust hard drives. Insights with respect to average life of hard drives show that Western Digital lasts longer with an average of around 3 years and Toshiba is not that durable with average life just above 1 year. Prediction of failure of hard drive using SMART statistics did not give good results with accuracy just above 50% which is like the toss of a coin.

8. GitHub Link:

Please find the code on below GitHub repository link:

https://github.com/lavleenbhat/Data-Mining.git