Hard Drive Fault Prediction

Members:

Abu Mathew Thoppan A0178303H (e0267614@u.nus.edu)

Balagopal Unnikrishnan A0178398E (e0267709@u.nus.edu)

Gopalakrishnan Saisubramaniam A0178249N (e0267560@u.nus.edu)

Nivedita Valluru Lakshmi A0178253Y (e0267564@u.nus.edu)

Yesupatham Kenneth Rithvik A0178448M (e0267759@u.nus.edu)

Description

- Backblaze takes a snapshot of each operational hard drive that includes basic hard drive information (e.g., capacity, failure) and S.M.A.R.T. statistics reported by each drive.
- Data spanning four quarters in 2018 and contains basic hard drive information and 50 different S.M.A.R.T. statistics.
- Each row represents a daily snapshot of one hard drive.

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

Kaggle link: https://www.kaggle.com/backblaze/hard-drive-test-data/home

Why S.M.A.R.T. Metrics?

- SMART monitors the performance of a hard drive in real time. Analysis of collected data and evaluation of each characteristic in two groups takes place inside the system every second:
  - Signs of storage device normal wearing (the number of cycles, heads movements, spindle hub rotations) device current status (the number of errors and the time of searching for a track, the elevation of heads above the drive, the total number of active sectors)
  - Performance assessments typically are in the range from 0 to 100. The higher is the number, the better is the performance of a data storage device in this particular characteristic. A low number indicates a high probability of future failure. https://howtorecover.me/best-programs-read-smart-attributes-hdd

- If the S.M.A.R.T. status indicates that you have an error, it does not necessarily mean that your hard drive is going to fail immediately. However, if there's a S.M.A.R.T. error, it would be wise to assume that your hard drive is in the process of failing. A complete failure could come in a few minutes, a few months, or—in some cases—even a few years. https://www.howtogeek.com/134735/how-to-see-if-your-hard-drive-is-dying/

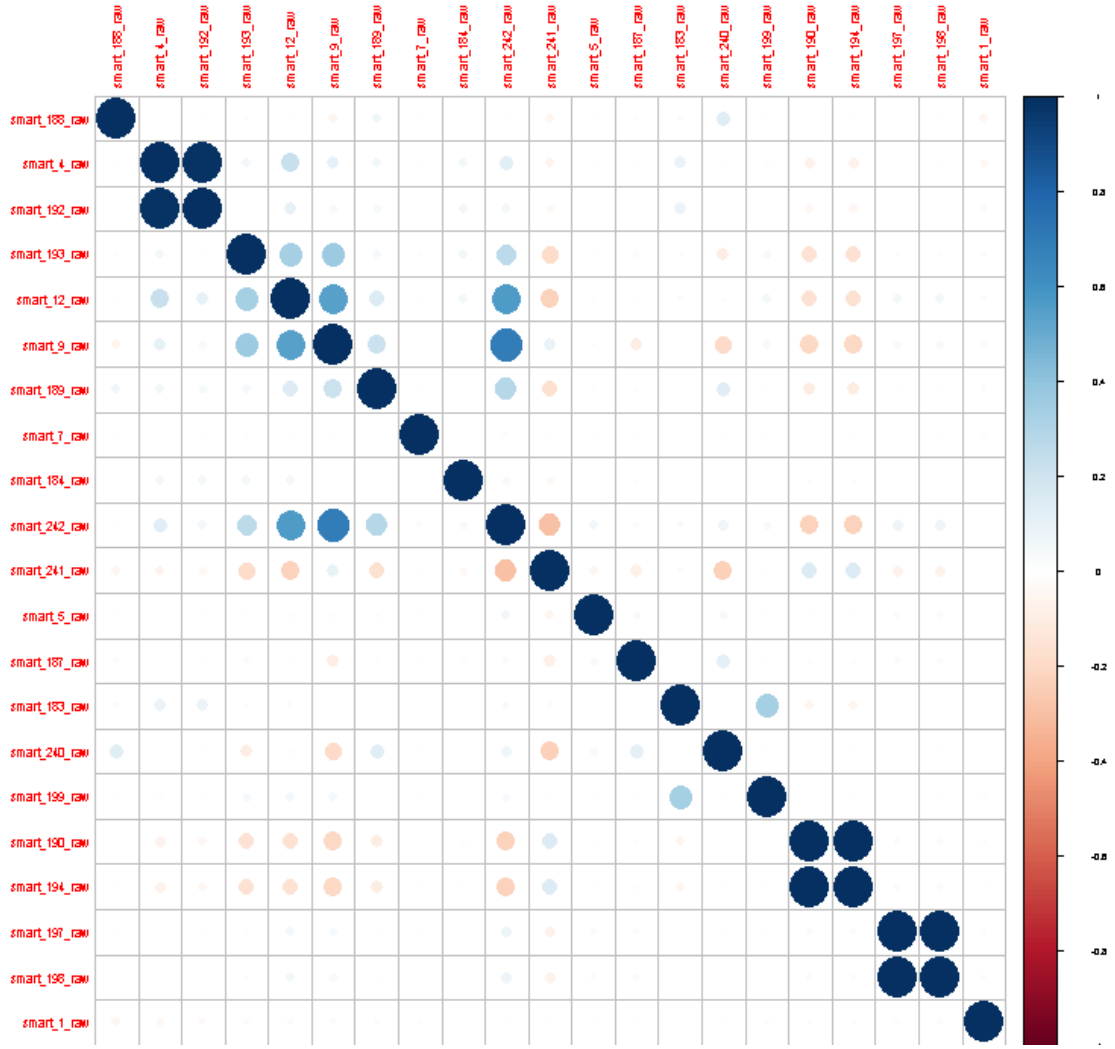
Metadata

- **date**: Date in yyyy-mm-dd format
- **serial_number**: Manufacturer-assigned serial number of the drive
- **model**: Manufacturer-assigned model number of the drive
- **capacity_bytes**: Drive capacity in bytes
- **failure**: Contains a “0” if the drive is OK. Contains a “1” if this is the last day the drive was operational before failing
- **variables that begin with 'smart'**: Raw and Normalized values for 50 different SMART stats as reported by the given drive
- The chosen objective is to **classify whether a hard drive will fail or not**

Dataset Pre-processing

- 9,992,362 unique rows from 32,164 unique hard drives (2018 Q1,Q2,Q3,Q4)
- After removing columns and rows with NA (thresh=999999), we end up with (9992205, 21) features and 1 target
- For phase -1 we **remove duplicates and retain one record** per serial number
- Per Class distribution:
  - 0 class – 31583
  - 1 class – 581
- Split train and test – 50-50 with **stratified sampling** (16082 samples per file)
- Apply Z-score (**Standard scaling**) to fit the training set and transform the test set

Data Understanding – Correlation Analysis



- SMART 4 and 192 exhibit high correlation as they relate to the number of cycles on start after shutdown. 192 captures power off cycles and is complemented by 4 which increments the value on startup.
- SMART 190 and 194 deal with temperature, hence highly correlated.
- SMART 197 and 198 exhibit high correlation because 197 defines unstable sectors due to read errors and 198 gives count of uncorrectable errors while read/write to a sector.
- SMART 9,12 and 242 are correlated to an extent as they cover related features - number of hours the drive is up, count of full power on/off cycles, and the Logical Block Addresses read during the time it was up.

Data Understanding – based on Wikipedia

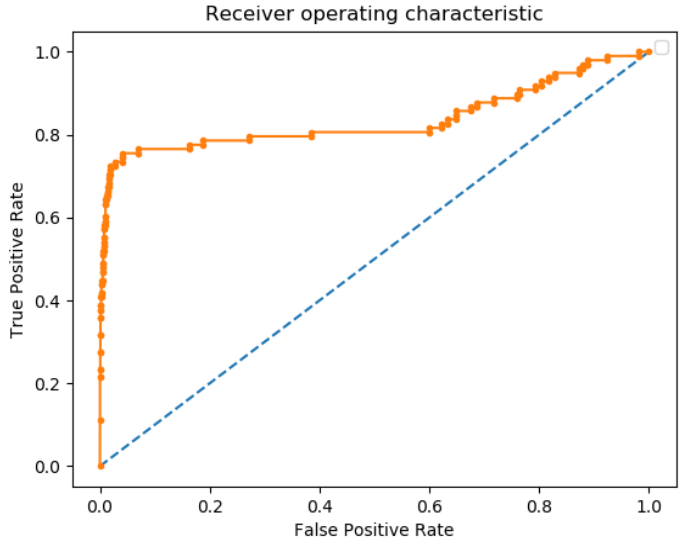
SMART ID	Attribute Name	Description	Comments
5	Count of re-allocated sectors	The raw value of the no. of bad sectors that were found and remapped.	This metric has been used to indicate the life expectancy of the drive.
7	Seek Error Rate	The raw value gives the drive's magnetic head seek error rate.	Different value measurements reported by different vendors. For the same vendor and model, the values should be consistent.
183	Runtime Bad Block	The total no. of data blocks with detected and un-correctable errors occurred during regular operations.	An indicator of drive aging and/or potential electromechanical problems
184	End-to-End error / IOEDC	Contains the parity error count that exists in the data path to the media through the drive's cache RAM.	Ideal value should be low as parity errors occur when data gets corrupted during transmission.
187	Un-correctable errors	No. of errors that could not be corrected using hardware ECC	High error count is a sign of failing drive.
188	Aborted operations	The no. of operations aborted due to hard disk drive timeout.	This value is close to 0 for healthy drives.
197	Pending sector count	The no. of unstable sectors that are to be remapped due to un-recoverable read errors.	The sector is remapped and this value is decreased over time on subsequent successful reads.
198	Count of un-correctable errors	The total count of uncorrectable errors when reading/writing a sector.	A rise in the value of this attribute indicates defects of the disk surface and/or problems in the mechanical subsystem.

Phase 1

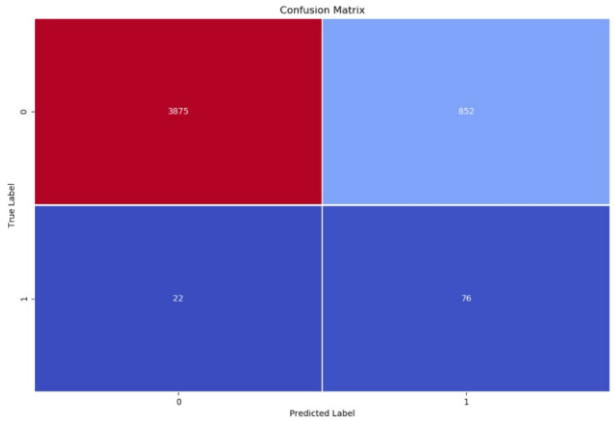
Treating each serial number as unique record without consideration of time (date)

ROC - All Features

- Chosen classifier: SVM
- AUROC: 0. 836
- Selected threshold: 0.013339
  FPR:0.187856991749524
  TPR: 0.775510204081632



	precision	recall	f1-score	support
0	0.99	0.79	0.88	4727
1	0.07	0.79	0.13	98
micro avg	0.79	0.79	0.79	4825
macro avg	0.53	0.79	0.51	4825
weighted avg	0.98	0.79	0.87	4825



Feature Selection Strategies

- **Variance threshold** - Feature selector that removes all low-variance features.
  - Unsupervised technique
  - Threshold: 0.025
  - Removed features: None

- **RFE** - Feature ranking with recursive feature elimination.
  - Estimator: Logistic Regression (default 100 iterations)
  - # features to select: 15
  - # features to discard per iteration: 1
  - Criterion: Weight coefficients
  - Removed features: 6

- **Sequential Forward Selection (SFS)** - SFS is a greedy search technique which returns a subset of features; the number of selected features k, where k<d, has to be specified a priori.
  - # features to select: 15
  - Criterion: AUROC (each step will add a feature k*i* that that maximizes the AUROC at that step)
  - Removed features: 13

- **Sequential Forward Selection (SBS) -** Similar to SFS described above
  - Removed features: 6

Choosing Feature Selector

Results on svm with rbf kernel on all 4 feature selectors:

rfe

[[7884 1]
[137 19]]

	precision	recall	f1-score	support
0	0.98	1.00	0.99	7885
1	0.95	0.12	0.22	156
micro avg	0.98	0.98	0.98	8041
macro avg	0.97	0.56	0.60	8041
weighted avg	0.98	0.98	0.98	8041

seq_fwd

[[7883 2]
[143 13]]

	precision	recall	f1-score	support
0	0.98	1.00	0.99	7885
1	0.87	0.08	0.15	156
micro avg	0.98	0.98	0.98	8041
macro avg	0.92	0.54	0.57	8041
weighted avg	0.98	0.98	0.97	8041

forward

[[7883 2]
[122 34]]

	precision	recall	f1-score	support
0	0.98	1.00	0.99	7885
1	0.94	**0.22**	0.35	156
micro avg	0.98	0.98	0.98	8041
macro avg	0.96	0.61	0.67	8041
weighted avg	0.98	0.98	0.98	8041

seq_bwd

[[7884 1]
[142 14]]

	precision	recall	f1-score	support
0	0.98	1.00	0.99	7885
1	0.93	0.09	0.16	156
micro avg	0.98	0.98	0.98	8041
macro avg	0.96	0.54	0.58	8041
weighted avg	0.98	0.98	0.97	8041

Handling Class Imbalance

- SMOTE
  - Minority class upsampled (1:0.6 for 0 to 1 classes)
  - (array([0, 1], dtype=int64), array([15792, 9475], dtype=int64))
  - New sample distribution:
    - Class 0 – 15792
    - Class 1 – 9475

Model Training

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine

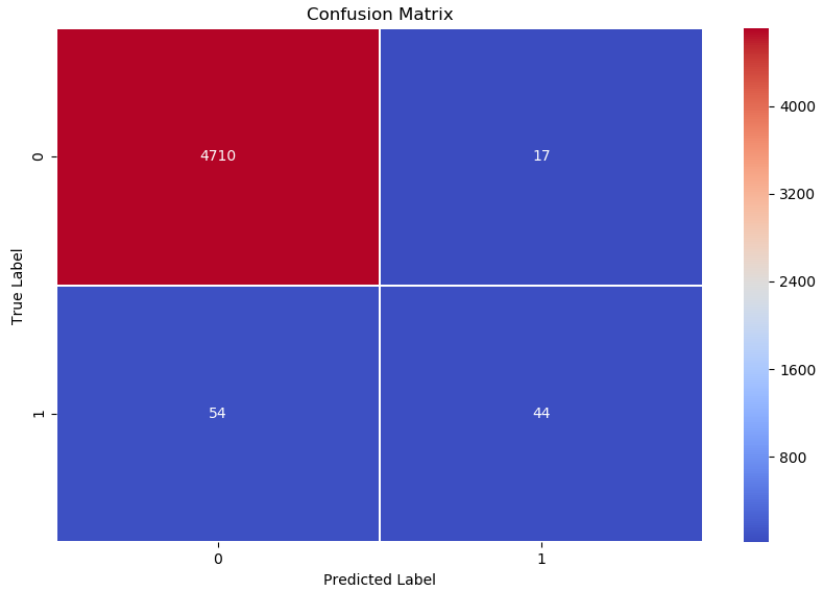
- The best parameters for each model is selected using Grid Search and using k-fold cv (k=7)

Decision Tree Classifier

Grid Search tuning_parameters = {'min_samples_split': range(10, 500, 10),
'max_depth': range(1, 20, 2),
'max_features':range(1,X_train.shape[1])}

Best score for data: 0.9880074620236298
Best value for min samples to split: 10
Best max depth value: 7
Best value to decide how many features to consider for splitting: 7

	precision	recall	f1-score	support
0	0.99	1.00	0.99	4727
1	0.81	0.30	0.43	98
micro avg	0.98	0.98	0.98	4825
macro avg	0.90	0.65	0.71	4825
weighted avg	0.98	0.98	0.98	4825

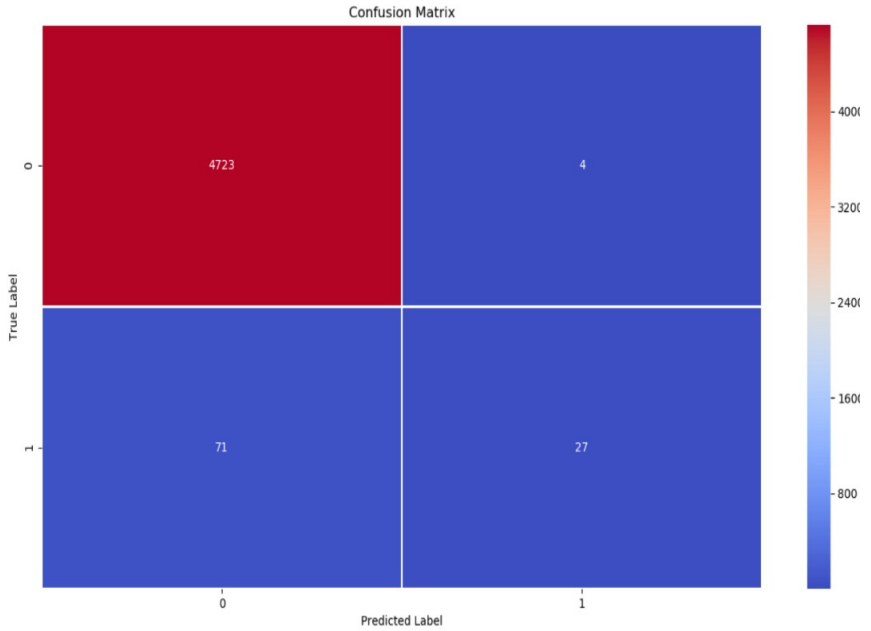


Support Vector Machine

Grid Search tuned_parameters = [{'kernel': ['rbf'], 'gamma': [1e-3, 1e-4], 'C': [1, 10, 100, 1000]},
{'kernel': ['linear'], 'C': [1, 10, 100, 1000]}]

best score for data1: 0.9880074620236298
Best C: 10
Best Kernel: linear
Best Gamma: auto_deprecated

	precision	recall	f1-score	support
0	0.99	1.00	0.99	4727
1	0.87	0.28	0.42	98
micro avg	0.98	0.98	0.98	4825
macro avg	0.93	0.64	0.71	4825
weighted avg	0.98	0.98	0.98	4825

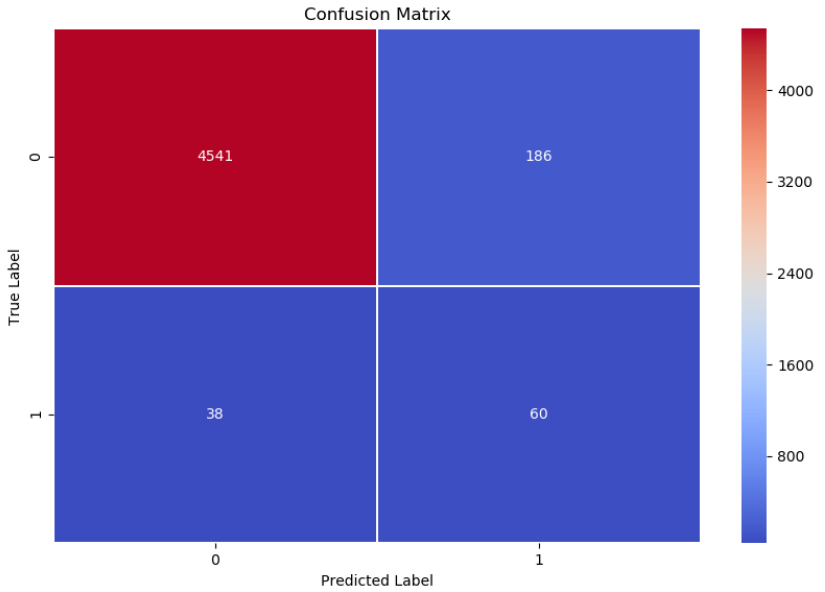


Logistic Regression

Grid Search tuning_parameters = {'penalty': ["l2"],
'class_weight': ['balanced'],
'random_state': [42],
'tol': [1e-3, 1e-4],
'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga’],
'C': [0.1, 0.5, 1, 2, 10, 100, 1000]}

Best score for data: 0.9563826952118681
Best tolerance value: 0.001
Best solver: newton-cg
Best value for C 0.5

	precision	recall	f1-score	support
0	0.99	0.96	0.98	4727
1	0.24	0.61	0.35	98
micro avg	0.95	0.95	0.95	4825
macro avg	0.62	0.79	0.66	4825
weighted avg	0.98	0.95	0.96	4825

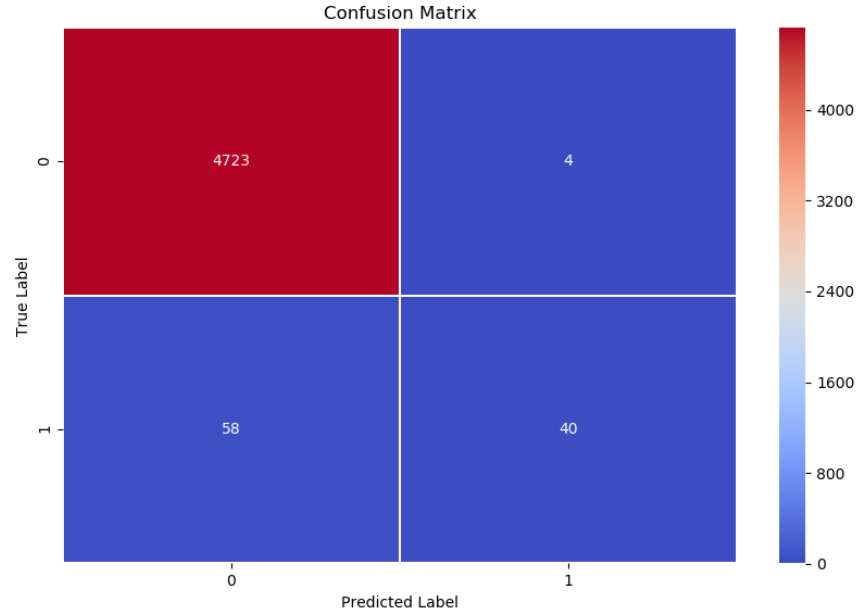


Random Forest

Grid Search tuning_parameters = {'criterion':['gini','entropy'],
'n_estimators': range(10, 20, 5),
'min_samples_split': range(10, 100, 10),
'max_depth': range(4, 10, 2)}

Best score for data: 0.9888957981700275
Best no. of estimators: 10
Best criterion for splitting: entropy
Best value for min samples to split: 30
Best max depth value: 8

	precision	recall	f1-score	support
0	0.99	1.00	0.99	4727
1	0.91	0.41	0.56	98
micro avg	0.99	0.99	0.99	4825
macro avg	0.95	0.70	0.78	4825
weighted avg	0.99	0.99	0.98	4825



Test Set

- Best among models: Random Forest

Best score for data: 0.9882739628675491
Best no. of estimators: 10
Best criterion for splitting: gini
Best value for min samples to split: 10
Best max depth value: 6

	precision	recall	f1-score	support
0	0.99	1.00	0.99	15791
1	0.88	0.36	0.51	291
micro avg	0.99	0.99	0.99	16082
macro avg	0.93	0.68	0.75	16082
weighted avg	0.99	0.99	0.98	16082