International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

SUPERVISED LEARNING BASED HARD DISK FAILURE PREDICTION

Mr. B. B. K. Prasad[1],A. Lakshmi Priyanka[2], M. Anjali[3], K. Devi Supriya[4]

- 1.Associate Professor Dept. of IT, NRI Institute of Technology, A.P, India-521212
- 2,3,4 UG Scholar, Dept. of IT, NRI Institute of Technology, A.P, India - 521212.

Abstract

Hard disk failures can be catastrophic in large scale data centres. It can lead to potential loss of all important and sensitive data stored in these data centres. To alleviate the impact of such failures, companies are actively looking at ways to predict disk failures and take pre-emptive measures. If companies are able to predict the failure of their hard-drives, it would reduce the economic impact incurred by the company due to these failures greatly, and protect data thereby maintaining customer trust. Admittedly, there are situations such as electricity failure in the server, natural hazard, etc. where the failure of disks cannot be predicted. However, most of the hardware failures don't happen overnight and hard disks starts to show significant reduced performance over the last few days of their lifetime before failing. Uncovering these patterns, recognizing features that may be attributed to the failure of a hard disk, and predicting the event of hard disk crash through machine learning, is the main goal of our project. Our project explores unsupervised and supervised learning techniques to predict and analyse hard drive crashes. The objective of using both supervised and unsupervised algorithms is to make a comparison between them.

Keywords: Random Forest Algorithm, Decision Tree Algorithm

Introduction

The task of hard disk failure prediction has been the primary focus of many researches over the recent few decades. Traditional approaches used a threshold-based algorithm. These however, were successful in predicting drive failures only 3-10% of the time [1]. Thus, we saw a shift to more proactive, learning-based algorithms that use S.M.A.R.T attributes to make predictions. These attributes are different hard drive reliability indicators of imminent failure.

In "Predictive models of hard drive failures based on operational data" [4], Nicolas and Samuel proposed using Random Forest and its variants for hard disk failure prediction. They achieved a very high accuracy of 99.98% and reported precision of 95% and recall of 67% when using Random Forest on the 2014 Backblaze dataset. The gradient boosted trees also performed similarly well, reaching a precision of 94% and recall of 67%. They used a subset of the S.M.A.R.T parameters (5, 12, 187, 188, 189, 190, 198, 199 and 200). [3] explores classification trees, recurrent neural networks, part voting random forests and random forests. They trained their algorithms for one hard disk model from the Backblaze data set. Part voting random forests were able to attain a failure detection rate of 100% and a false alarm rate of 1.76% for model ST3000DM001. Select features of this

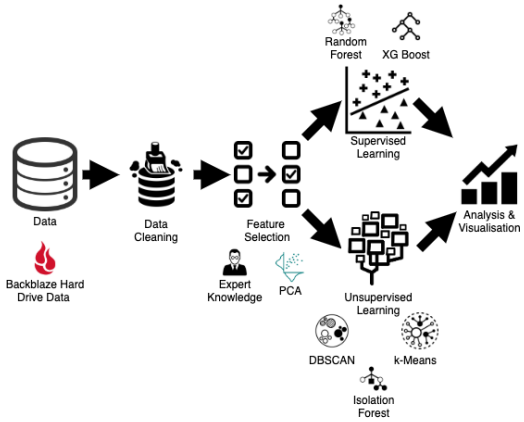
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

model were used in the training. S.M.A.R.T attributes are hard drive model-specific i.e the meaning of these attributes might differ across manufacturers. To accommodate these nuances, previous works [2] [3] explored ways to train algorithms on specific models instead of one generic model for predication. The most recent studies leverage Transfer learning techniques [2] where classifiers trained on one model are used for predicting failures of other models. This however did not perform as well as they had hoped. Since the failure of hard disks is a rare event, the dataset is highly unbalanced. hence, in order to overcome this imbalance problem, work has also been done in exploring the efficiency of SMOTE (Synthetic Minority Oversampling Technique) [4] and resampling [5] techniques. All the implemented supervised learning techniques try to address this problem.

In contrast to all the aforementioned works, we decided to focus on supervised as well as unsupervised learning techniques in this project. Instead of just looking at accuracy, we used F1 score as the primary metric to evaluate our algorithms. We have employed anomaly detection and clustering based techniques and contrasted their performance against supervised learning techniques that use tree-based classifiers. We have used only a subset of data (last 10 days of a hard disk lifetime) to train our models. This along with resampling helps us tackle the class imbalance problem. We have also ensured to maintain the time-sequence in the dataset in order to train better models.

Every disk drive includes Self-Monitoring, Analysis, and Reporting Technology (S.M.A.R.T) statistics, which reports internal information about the drive and its primary function is to detect as well as report multiple indicators of hard disk drive reliability with the intent of anticipating imminent hardware failures. Backblaze takes a snapshot of each operational hard drive daily and the data includes drive's serial number, model number, disk capacity, a label indicating disk failure, and S.M.A.R.T stats. Data for the project was collected from January 1st, 2019 to December 31st, 2019 and data was in 365 CSV files with each representing one day of the year. Each file has 129 columns. 62 distinct S.M.A.R.T attributes are measured and represented both as raw values as well as normalized values totalling to 124 columns. The other columns provide information about the hard disk and the date of the record. The data is temporal in nature and is more than 10 GB in size. We have 40.7 million data points or records in the dataset in total.

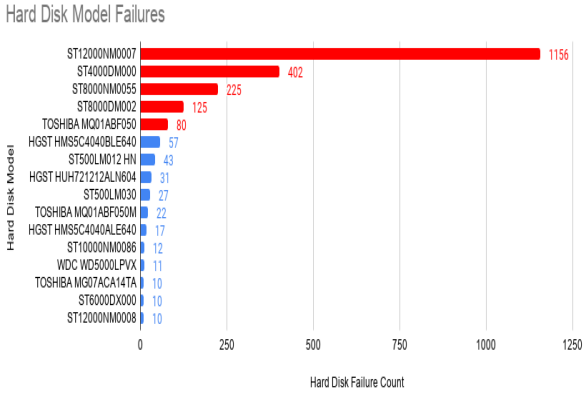
Architecture:



We started by observing the raw Backblaze data to get a better understanding of what all preprocessingtechiniques were needed to be employed. We observed that different hard disks showed significantly different

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

behavior in terms of S.M.A.R.T statistics at the time of failure. Since the failure of a hard disk is a very rare event given its life span, we also observed a heavy bias in the dataset. Just a few rows were labeled 1 indicating a failure of hard disk on the given day while on all other days of its lifetime, the label remained 0.



The records corresponding to this hard disk are no longer available in the dataset. To reduce this bias, we only worked with dataset in which the hard-disk was failing frequently and only sampled a few of the data-points where the hard-disk was operational.

In this project we created two programs called 'HardDiskFailure1.py' and 'HardDiskFailure2.py' by using different machine learning algorithms

HardDiskFailure1: In this program I have used Random Forest, Decision Tree and Logistic Regression and the calculate accuracy, precision, recall and FSCORE

HardDiskFailure2: In this program I have used Random Forest, SVM and Gradient Boosting algorithms

To implement this project I have used same dataset given by you.

To implement this project I have designed following modules

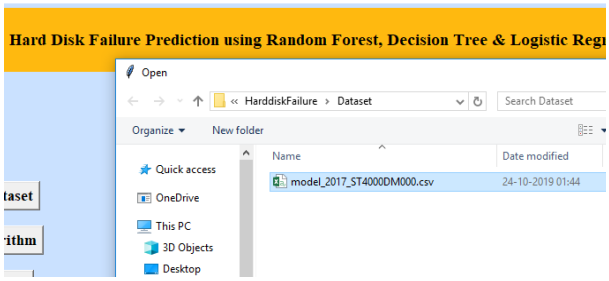
1) Upload Harddisk Dataset: using this module we will upload dataset to application and then read all records from dataset
2) Preprocess Dataset: using this module we will remove empty and missing values from dataset and then replace with 0
3) Generate Train & Test Dataset: using this module we will divide or split dataset into train and test where application used 80% dataset for training machine learning algorithm models and used 20% dataset to test trained model. Trained model will be applied on test data to predict class labels and then this class label will be compared with original dataset to calculate accuracy.
4) Run Random Forest Algorithm: using this module we will trained random forest algorithm with above dataset and then calculate accuracy.
5) Run Decision Tree Algorithm: using this module we will trained decision tree algorithm with above dataset and then calculate accuracy.
6) Run Logistic Regression Algorithm: using this module we will trained Logistic Regression algorithm with above dataset and then calculate accuracy.
7) Comparison Graph: using this module we will plot accuracy, precision, recall and FSCORE comparison graph of all algorithms

To run project double click on 'run_HardDiskFailure1.bat' file to get below screen

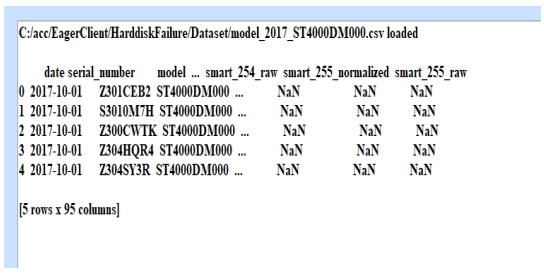
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Results:

In above screen click on 'Upload Harddisk Dataset' button to upload dataset



In above screen selecting and uploading 'model_2017' dataset file and then click on 'Open' button to load dataset and to get below screen

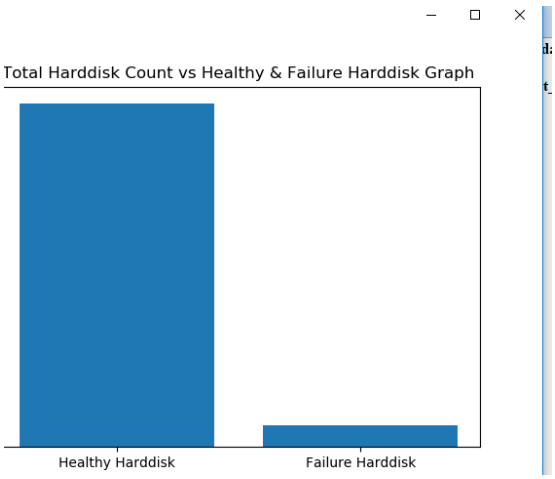


C:/acc/EagerClient/HarddiskFailure/Dataset/model_2017_ST4000DM000.csv loaded

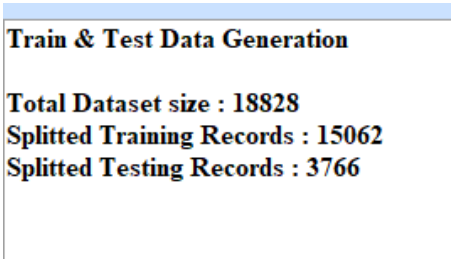
	date	serial_number	model	...	smart_254_raw	smart_255_normalized	smart_255_raw
0	2017-10-01	Z301CEB2	ST4000DM000	...	NaN	NaN	NaN
1	2017-10-01	S3010M7H	ST4000DM000	...	NaN	NaN	NaN
2	2017-10-01	Z300CWTK	ST4000DM000	...	NaN	NaN	NaN
3	2017-10-01	Z304HQR4	ST4000DM000	...	NaN	NaN	NaN
4	2017-10-01	Z304SY3R	ST4000DM000	...	NaN	NaN	NaN

[5 rows x 95 columns]

In above screen in text area we can see dataset loaded and this dataset contains lots of 'NaN' missing values and machine learning algorithms will not accept NaN values so we need to remove those NaN values by clicking on 'Preprocess Dataset' button

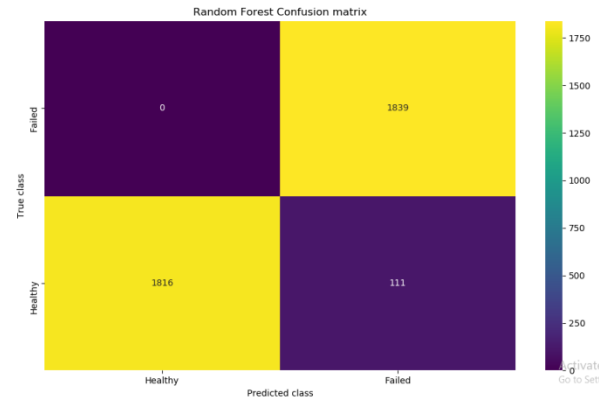


In above screen text area all NaN values are replaced with '0' and in above graph we can see number of healthy and failure records. Now close above graph and then click on 'Generate Train & Test Dataset' button to divide dataset into train and test parts

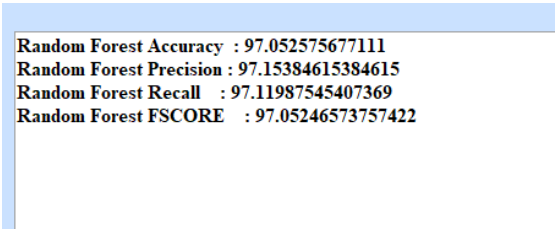


In above screen dataset contains total 18828 records and application using 15062 (80%) records to train Machine learning algorithms and 3766 (20%) records for testing machine learning. Now dataset is ready and now click on 'Run Random Forest Algorithm' to trained above dataset with Random Forest and then will get below confusion matrix graph

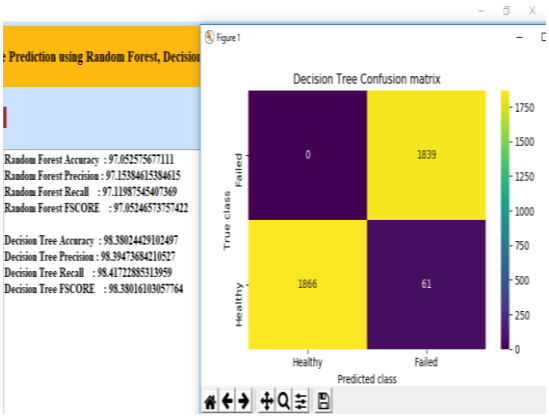
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362



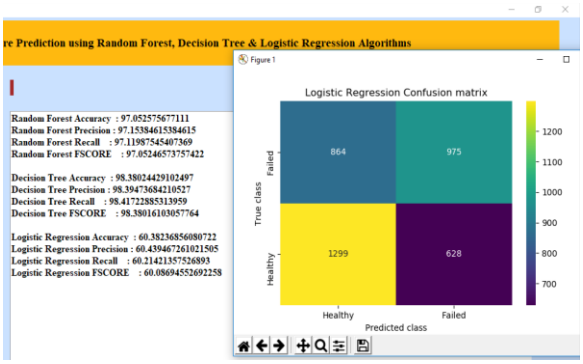
In above graph we can see confusion matrix graph for random forest and now close above graph to get below screen



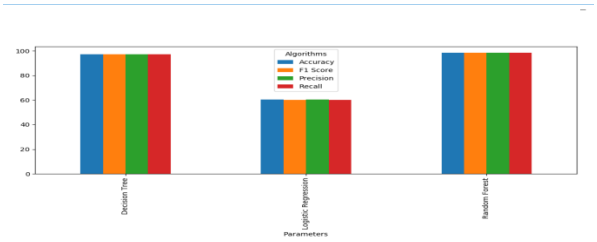
In above screen random forest accuracy, precision, recall and FSCORE is calculated and now click on 'Run Decision Tree Algorithm' button to train decision tree with above dataset



In above screen decision tree confusion matrix and accuracy is calculated and now closed above graph and then click on 'Run Logistic Regression Algorithm' button to train this algorithm with above dataset



In above screen Logistic Regression confusion matrix and accuracy is calculated and in all 3 algorithms decision tree has got high accuracy and now close above graph and then click on 'Comparison Graph' button to get below graph



In above graph 4 different colour bars represents accuracy, precision, recall and FSCORE values and in above graph x-axis represents algorithm names and y-axis represents values.

Model	Accuracy	Label	Precision	Recall	F1	Support
ST12000NM0007	0.9997123	0	1	1	1	74210
		1	0.99	1	1	2252
ST4000DM000	0.9999358	0	1	1	1	45926
		1	1	1	1	794
ST8000NM0055	1	0	1	1	1	28906
		1	1	1	1	436
ST8000DM002	1	0	1	1	1	19627
		1	1	1	1	239
TOSHIBA MQ01ABF050	0.9990637	0	1	1	1	912
		1	0.99	1	1	156

Conclusion

We predicted hard disk failure based on its S.M.A.R.T attributes. We used data

augmentation techniques like SMOTE and data resampling to handle the class imbalance problem. We were not able to implement a generic model to determine failure, since S.M.A.R.T attributes are model and manufacturer specific. We applied supervised learning techniques like Random Forest and XGBoost on individual hard disk models to predict hard disk failure. We were able to obtain a very high F1 score for all the hard disk models trained using tree-based classifiers. We further extended the study to predict hard disk failure using unsupervised learning techniques like DBSCAN and K-Means to cluster them into groups of failing and non-failing hard drives. We also explored a novel approach of applying anamoly detection techniques for the hard disk prediction problem. Unsupervised learning techniques however performed poorly due to the the nature of the dataset when compared to supervised learning.

References:

1. C. Xu, G. Wang, X. Liu, D. Guo, and T. Liu. Health status assessment and failure prediction for hard drives with recurrent neural networks. IEEE Transactions on Computers, 65(11):3502–3508, Nov 2016.
2. Mirela Madalina Botezatu, Ioana Giurgiu, JasminaBogojeska, and Dorothea Wiesmann. Predicting disk re- placement towards reliable data centers. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
3. Jing Shen, Jian Wan, Se-Jung Lim, and Lifeng Yu. Random-forest-based failure prediction for hard disk drives. International Journal of Distributed Sensor Networks, 14(11):1550147718806480, 2018.
4. Nicolas Aussel, Samuel Jaulin, Guillaume Gandon, YohanPetetin, ErizaFazli, et al.. Predictive models of hard drive failures based on operational data. ICMLA 2017 : 16th IEEE International Conference On Machine Learning And Applications, Dec 2017, Cancun, Mexico.
5. Wendy Li, Ivan Suarez, Juan Camacho,Proactive Prediction of Hard Disk Drive Failure-Project
6. Backblaze. Backblaze hard drive state, 2020.
7. J. Li et al. Hard drive failure prediction using classification and regression trees. In 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, Atlanta, GA, 2014, 2014.
8. Blagus, R., Lusa, L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14, 106 (2013)
9. Andy Klein, “What SMART Stats Tell Us About Hard Drives”, October 6, 2016, Available : https://www.backblaze.com/blog/what-smart-stats-indicate-hard-drive-failures/. [Accessed: April 11, 2020]
10. https://www.backblaze.com/blog/hard-drive-stats-for-q2-2018/. [Accessed: April 13, 2020]