Report 5

LI, Ruizhe | 1155076990

Github: https://github.com/rzli6/ML-Storage.git (private)

Target Paper: Improving Storage System Reliability with Proactive Error Prediction

Goal: Realize the method descripted in the paper.

Data Set: Blackblaze 2015, 12 months in total.

In the 2015's dataset, the overview is like this:

	Models	smart_5_raw	smart_187_raw	smart_196_raw	smart_197_raw	Capacity (TB)	# Drives
0	ST4000DM000	0.007381	0.015470	0.000000	0.015066	4.0	29670.0
1	ST3000DM001	0.120719	0.177226	0.000000	0.070205	3.0	1168.0
2	Hitachi HDS5C3030ALA630	0.032349	0.000000	0.032132	0.012375	3.0	4606.0
3	Hitachi HDS722020ALA330	0.133675	0.000000	0.133675	0.036515	2.0	4683.0
4	Hitachi HDS5C4040ALE630	0.015038	0.000000	0.015414	0.007519	4.0	2660.0
5	HGST HMS5C4040ALE640	0.006032	0.000000	0.006032	0.002665	4.0	7129.0
6	HGST HMS5C4040BLE640	0.000967	0.000000	0.000967	0.002578	4.0	3103.0

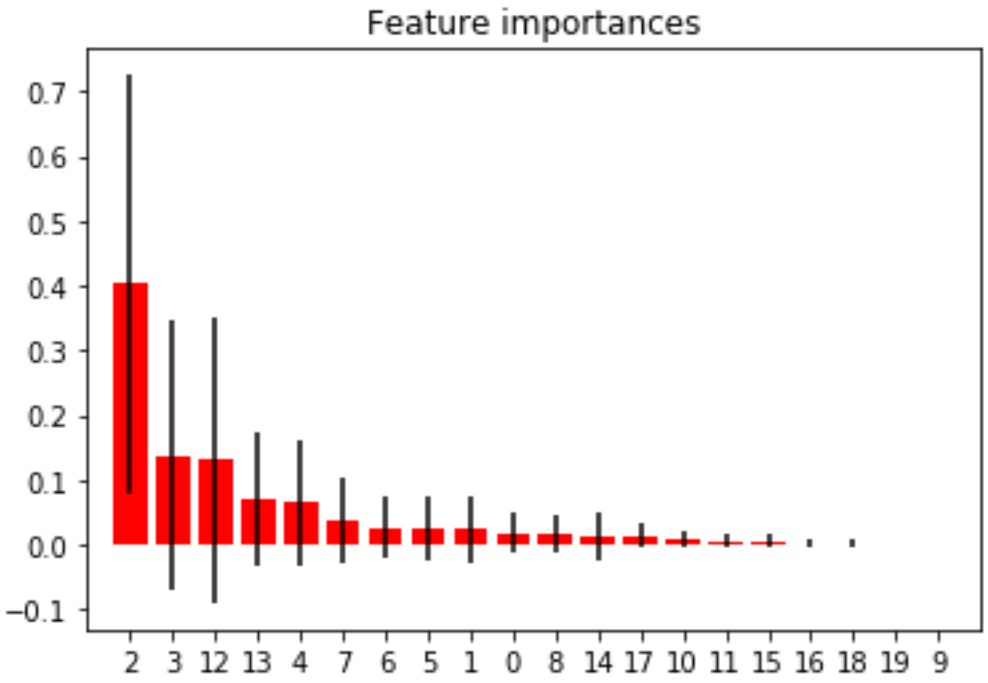
And when I trained and tuned each machine learning method on disk model ST3000DM001, which is the same choice as paper, I got figures like this:

	CART	SVM	NN	LR	RF
P	0.968137	0.991304	0.992000	0.900943	0.984615
R	0.865217	0.906522	0.930797	0.939493	0.974638
F	0.911131	0.946601	0.943269	0.915293	0.979045
Sd	0.089723	0.018153	0.052120	0.046580	0.012678

Note that in the model ST3000DM001, there are only **1168** disks in total, and **118** of them were failed. To make the data set more balanced, I down sampled the number of healthy disks to **3** times the number of the failed disks. In other words, there are only less than 500 samples in the dataset. And I used a 5 folds cross validation to train my models and got the figures above.

From the above figure we can see that, although the dataset is rather small, we can still use RF machine learning model to get a F1 score as high as 97.9%. And similar to the paper, following the forests model are the SVM and NN models, both of their f1 scores were around 95%. And the LR and CART models can also reach 91% accuracy.

Smart_5_raw: reallocated sectors count



This picture portrayed the Feature Importance when the RF model doing the prediction.

```
(0, 'smart_1_raw'),
 (1, 'smart_4_raw'),
 (2, 'smart_5_raw'),
 (3, 'smart_7_raw'),
 (4, 'smart_9_raw'),
 (5, 'smart_12_raw'),
 (6, 'smart_187_raw'),
 (7, 'smart_193_raw'),
 (8, 'smart_194_raw'),
 (9, 'smart_197_raw'),
 (10, 'smart_199_raw'),
 (11, 'smart_4_raw_increase'),
 (12, 'smart_5_raw_increase'),
 (13, 'smart_7_raw_increase'),
 (14, 'smart_9_raw_increase'),
 (15, 'smart_12_raw_increase'),
 (16, 'smart_187_raw_increase'),
 (17, 'smart_193_raw_increase'),
 (18, 'smart_197_raw_increase'),
 (19, 'smart_199_raw_increase')
```

And the above statistics is the sequence of the features. We can see that, the ‘smart_5_raw’ contributes the most to the prediction, which means the current smart_5 value could indicate an increasement in next week. And smart_7 is the second

Smart_5_raw: reallocated sectors count

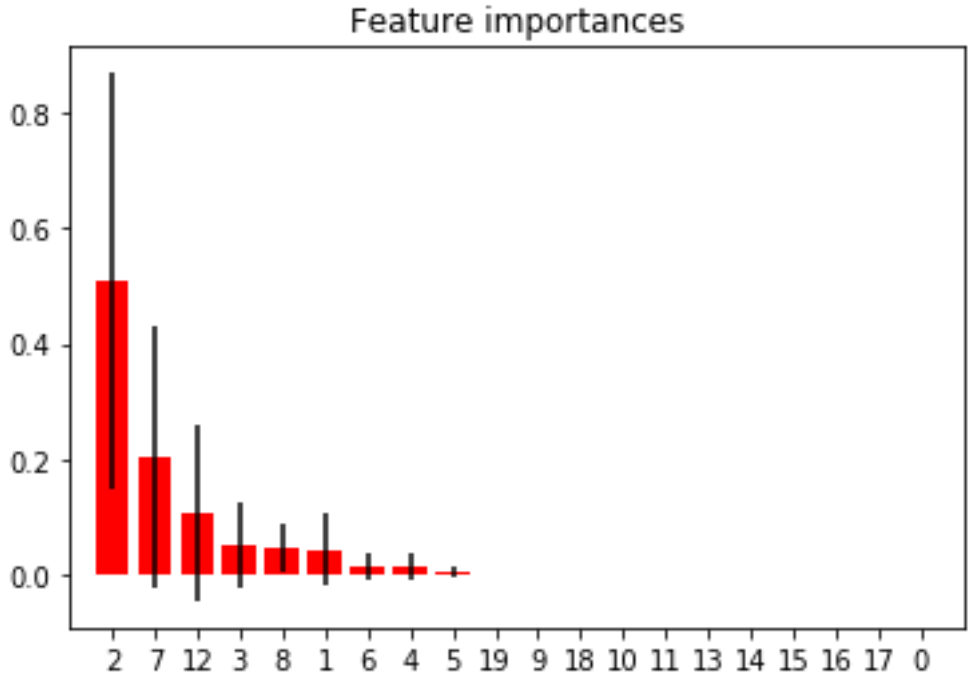
most important attribute. And their increasements also contribute a lot. And other values are not that important in prediction.

And as following indicated, all other HDD models can achieve a high accuracy to predict the smart_5_raw's change. What's more, the ST4000DM000 model can achieve a 100% accuracy, which is so weird to me…

model	Metrics	CART	SVM	NN	LR	RF
ST4000DM000	P	1.00	1.00	1.00	1.00	1.00
	R	1.00	1.00	1.00	1.00	1.00
	F	1.00	1.00	1.00	1.00	1.00
	Sd	0.00	0.00	0.00	0.00	0.00
ST3000DM001	P	0.97	0.99	0.99	0.90	0.98
	R	0.87	0.91	0.93	0.94	0.97
	F	0.91	0.95	0.94	0.92	0.98
	Sd	0.09	0.02	0.05	0.05	0.01
Hitachi HDS5C3030ALA630	P	1.00	1.00	1.00	1.00	1.00
	R	0.97	0.97	0.94	0.94	1.00
	F	0.99	0.99	0.96	0.97	1.00
	Sd	0.01	0.02	0.05	0.04	0.00
Hitachi HDS722020ALA330	P	1.00	1.00	1.00	1.00	1.00
	R	0.97	0.98	0.97	0.97	0.97
	F	0.98	0.99	0.98	0.98	0.98
	Sd	0.04	0.02	0.04	0.04	0.04
Hitachi HDS5C4040ALE630	P	0.91	0.94	0.95	0.79	1.00
	R	0.89	0.88	0.92	0.84	0.89
	F	0.89	0.88	0.93	0.76	0.93
	Sd	0.11	0.16	0.10	0.20	0.10
HGST HMS5C4040ALE640	P	1.00	1.00	1.00	1.00	1.00
	R	0.99	0.98	0.94	0.98	1.00
	F	1.00	0.99	0.97	0.99	1.00
	Sd	0.01	0.02	0.03	0.03	0.00

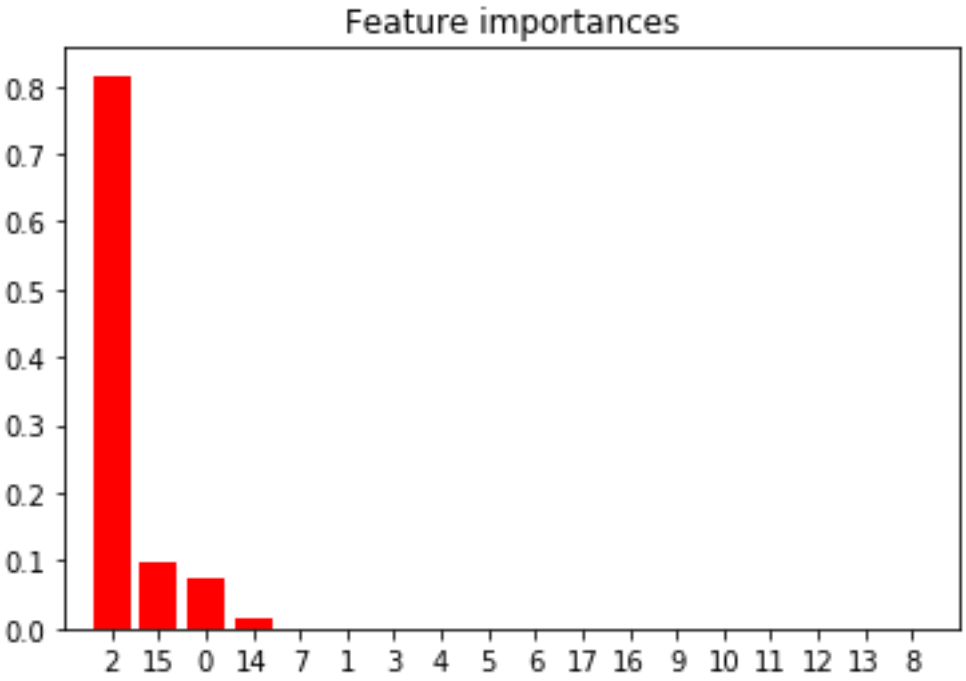
Smart_5_raw: reallocated sectors count

The following pictures are the feature importance for **RF model** in each different disk models. We can see that, the smart_5_raw, smart_7_raw, and their corresponding increasement are always the dominant features.



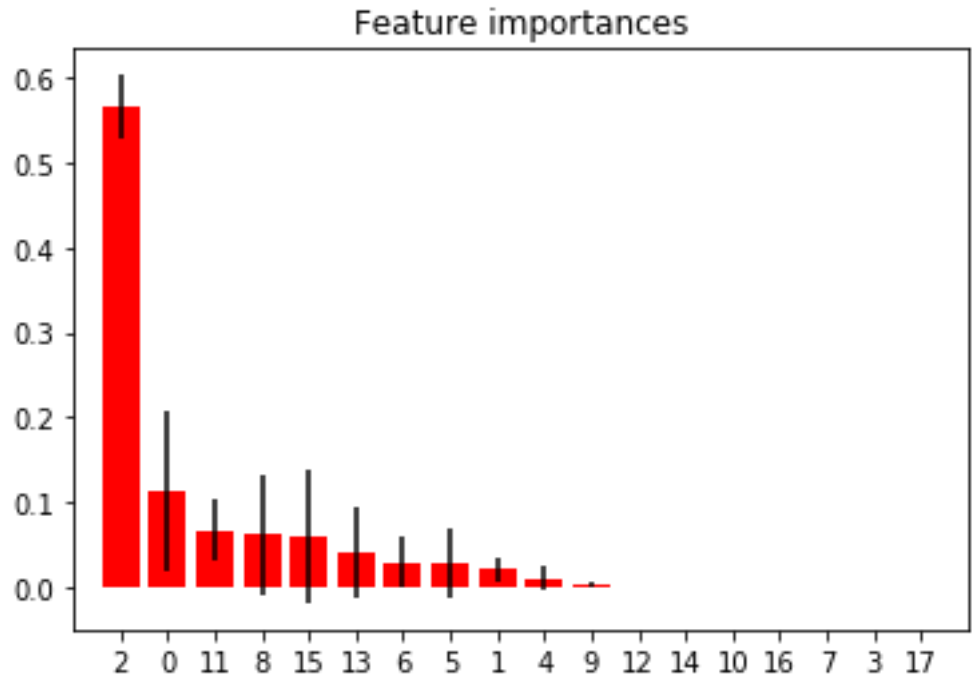
7_

ST4000DM000

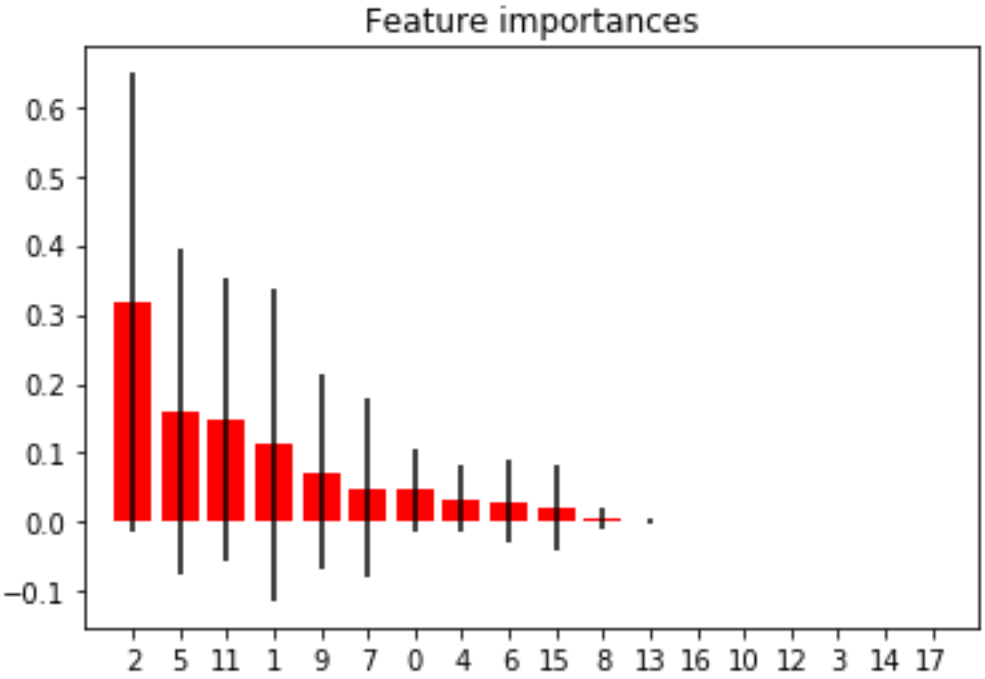


Hitachi HDS722020ALA330

Smart_5_raw: reallocated sectors count



Hitachi HDS5C3030ALA630



Hitachi HDS5C4040ALE630

Smart_5_raw: reallocated sectors count