**Report 5**

**LI, Ruizhe | 1155076990**

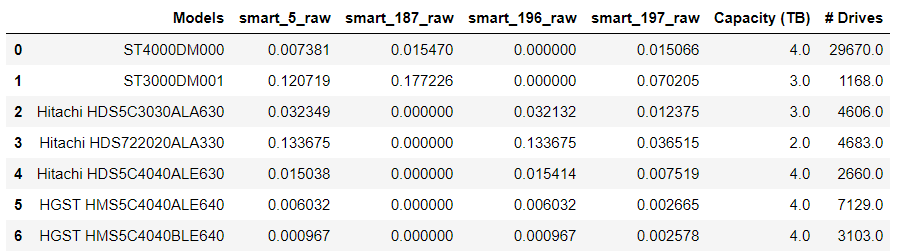
**Github:** <https://github.com/rzli6/ML-Storage.git> (private)

**Target Paper:** Improving Storage System Reliability with Proactive Error Prediction

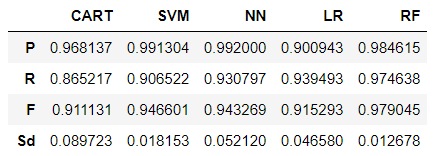
**Goal:** Realize the method descripted in the paper.

**Data Set:** Blackblaze 2015, 12 months in total.

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

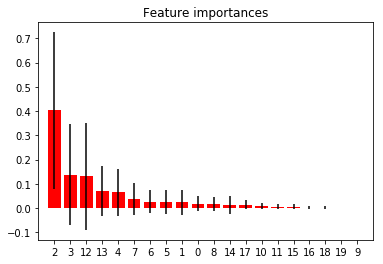


And when I trained and tuned each machine learning method on disk model ST3000DM001, which is the same choice as paper, I got figures like this:



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.



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'),

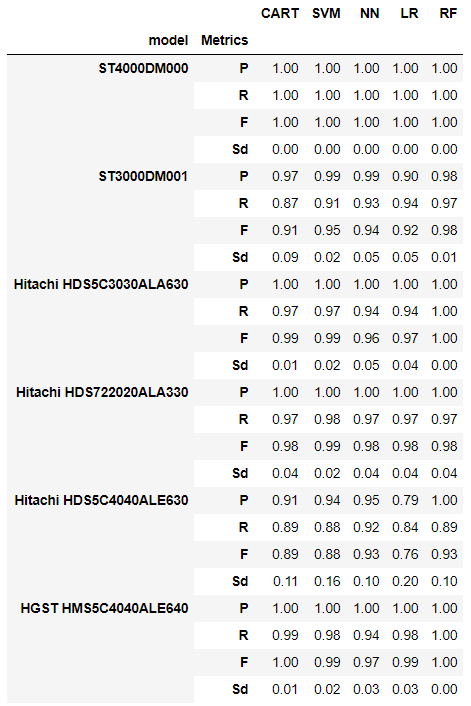
(17, 'smart\_193\_raw\_increase'),

(18, 'smart\_197\_raw\_increase'),

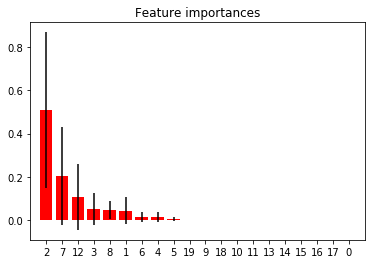
(19, 'smart\_199\_raw\_increase')

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

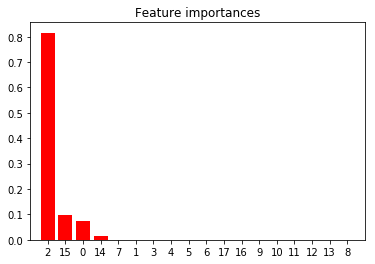
And as following indicated, all other HDD models can achieve a high accuracy to predict the smart\_5\_raw’s change. What’s more, the ST4000DM000 model can achieve a 100% accuracy, which is so weird to me…



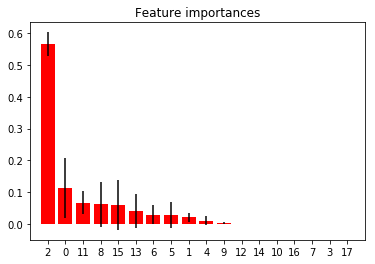
The following pictures are the feature importance for **RF model** in each different disk models. We can see that, the smart\_5\_raw, smart\_7\_raw, and their corresponding increasement are always the dominant features.

7\_

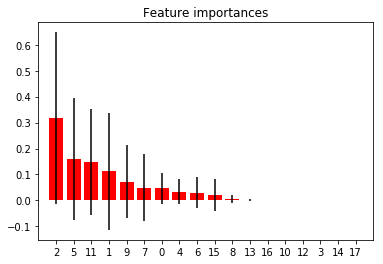
ST4000DM000



Hitachi HDS722020ALA330



Hitachi HDS5C3030ALA630



Hitachi HDS5C4040ALE630