Department of Computer Science, IIT Madras

Cloud Computing (CS-6847)
Course Project Report

TITLE

Feature Selection and Prediction of Remaining Useful life of a HDD using LSTM

Submitted By -
Gaurav Singhal
(CS18S032)

Submitted To-
Prof. D. Janakiram
(Computer Science Dept)

Problem Abstract -

Hardware components of any system have a specific period for which they perform their tasks sleekly. After a particular time, their performance starts degrading, and eventually, these components crash. HardDisks are essential hardware components for any computer system. The failure of this component can severely impact an organization in terms of the loss of their data. One of the solutions to this problem is to clone the data in multiple replicas. For large organizations, the replication of data is not a cost-effective solution.

The cost-effective solution is to predict the remaining useful life(RUL) of HDD and clone the data before failure. We use Machine Learning algorithms to predict the remaining useful life of HDD on the Backblaze HDD dataset.

This project aims to perform feature selection on the Backblaze HDD dataset and reiterate the research paper "A Data-driven Prognostic Architecture for Online Monitoring of HardDisks Using Deep LSTM Networks."

Introduction -

Backblaze HDD Dataset-

The Backblaze HDD dataset contains the daily snapshot of operational hard disks at the Backblaze data center. This snapshot of a hard disk comprises a serial number, Date, capacity, failure, and the SMART (Self-Monitoring, Analysis, and Reporting Technology) attributes. SMART is a monitoring system included in HDDs to detect and report various indicators of drive reliability to anticipate imminent hardware failures.

Feature Selection-

Feature/Attribute selection is a process of finding the relevant feature set for a specific target variable. As mentioned in [cite], Feature selection techniques are used for four reasons, which are, the simplification of the model (which makes it easier to interpret by the users), reducing training time, avoiding the curse of dimensionality and reducing Over-fitting.

Backblaze Dataset contains SMART features for each HDD device, but all SMART features are not essential for our target variable Failure of Disk. So we

need to find the features which are more relevant for the prediction of target variable. To attain this, we are using a tool feature-selector authored by Will Koehrsen. This Feature-selector tool has below five methods to identify the features to remove.

- Missing Values - It removes the features which do not contain values in some rows. We have filled these missing values with adjacent rows as most of the features/attributes are showing an increasing trend.

- Single Unique Values - It removes the features which contain the same values across all the rows, as these will not be contributing anything in the prediction of RUL.

- Collinear Features - It removes one of the features from the two if both of them are highly correlated with one another.

- Zero Importance Features - It removes the features that have zero importance according to a gradient boosting machine (GBM) learning model.

- Low Importance Features - It removes the features that have less importance according to a gradient boosting machine (GBM) learning model, and the threshold is decided to retain a certain percentage of the variance.

LSTM-

LSTM(Long Short term memorry networks) are an special kind of RNN (Recurrent neural networks) which is capable of learning long term relationships. LSTM performs selective delete, add and update of information at each layer. First LSTM decides what information it is going to throw away from the cell state (forget gate layer). The next step is to decide what new information it is going to store in the cell state (input gate layer). Finally, it need to decide what it is going to output, based on its cell state. All these decisions are made by various layers present in LSTM network.

The mentioned research paper trains the LSTM networks to predict the remaining useful life of a given Harddisk. It proposes two models for predicting RUL. Summary of these model is shown in figure[1].

SUMMARY OF LSTM MODEL 1 USED

Layer (type)	Output Shape	Parameters
LSTM layer 1	(None,25,100)	42400
Dropout layer 1	(None,25,100)	0
LSTM layer 2	(None,50)	30200
Dropout layer 2	(None,50)	0
Dense	(None,1)	51

SUMMARY OF LSTM MODEL 2 USED

Layer (type)	Output Shape	Parameters
LSTM layer 1	(None,25,100)	42400
Dropout layer 1	(None,25,100)	0
LSTM layer 2	(None,100)	80400
Dropout layer 2	(None,100)	0
Dense	(None,1)	101

Figure 1: Summary of LSTM Model in RUL prediction

Analysis and Graphs-

I have performed the feature engineering over BackBlaze HDD dataset. This dataset contains the timestamp based data as it is captured as daily snapshot of harddisks maintained in Backblaze data center. This prediction required the data that indicates the SMART attributes based on each harddisk. So we reformed the data to required format.

After reforming the data we needed to fill the missing values to have a clean data. As these attributes are continuous in nature, We can fill these missing values with mean value of top and bottom cells.

Feature Selector tool is used to identify the SMART attributes that have more importance to predict the remaining useful life of hard disk. In Backblaze data set first it removes attributes with missing values greater then a threshold percentage. Featue selector finds 42 features with greater than 0.60 missing values. These attributes will not play any role for predicting RUL. Figure[2] shows the Histogram representing the number of features with their fraction of missing values.

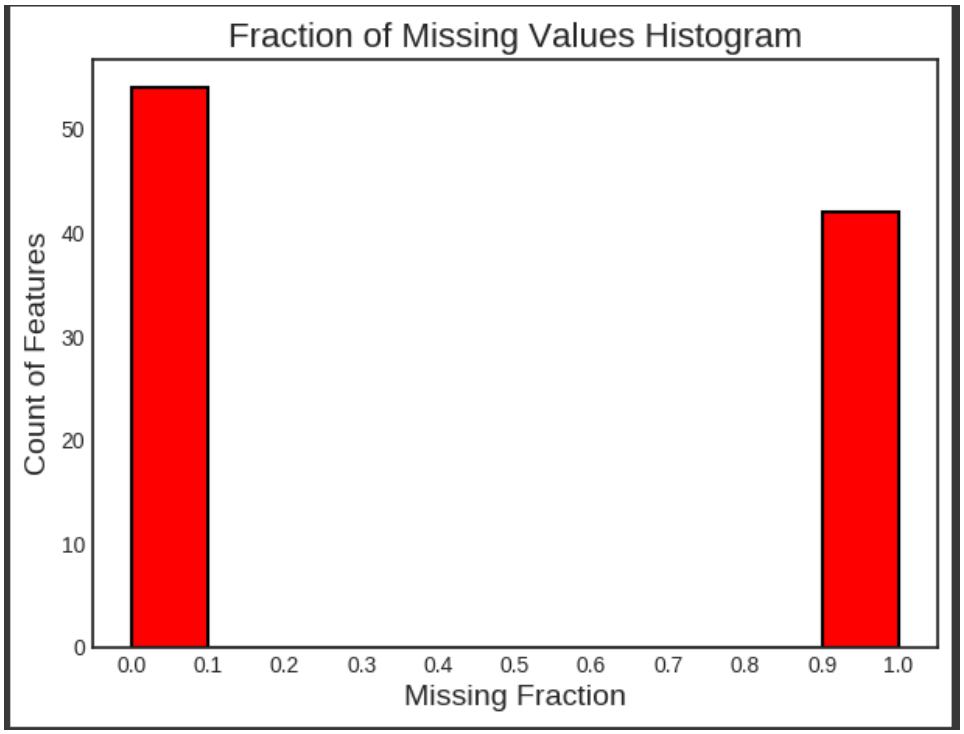

Figure 2 : Histogram representing number of features with fraction of missing value

After removing the attributes with missing values, we remove the attributes that contains a unique value for all the rows. Figure[3] shows the histogram of all the values in the dataset and finds the frequency of unique values. In this datase feature selector identifies 14 features with a single unique value.

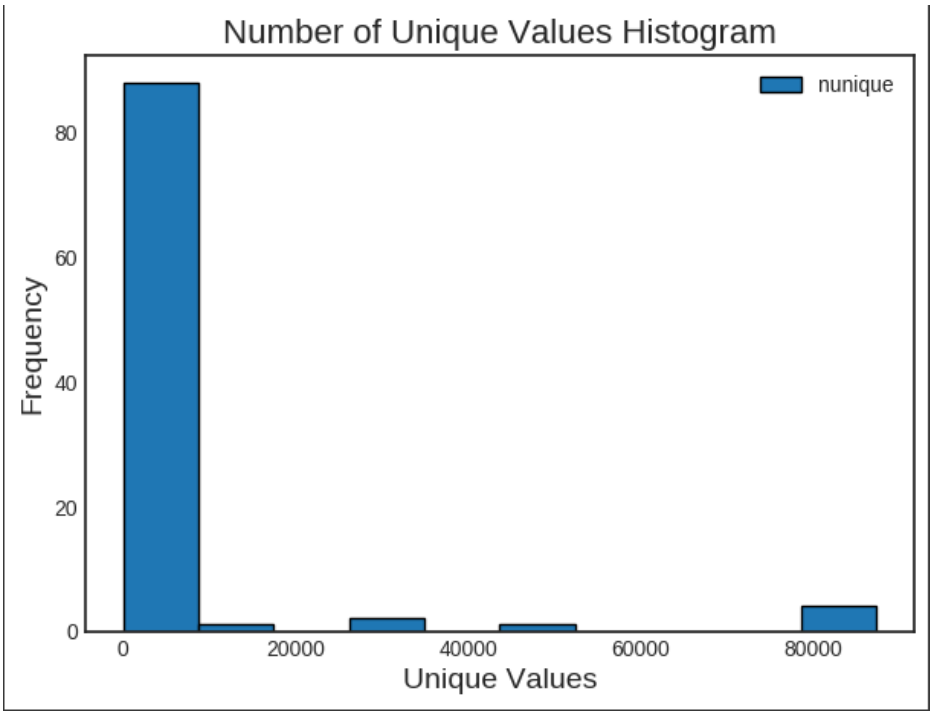

Figure 3 : Histogram representing unqiue values and their frequencies

We need to remove these attributes with single unique values as they are not going to contribute for predrcition of RUL. Next we need to remove the co-related features. Figure[4] shows the attributes which have corelation factor more then threshold value. Feature selector mentions 11 features with a correlation magnitude greater than 0.98.

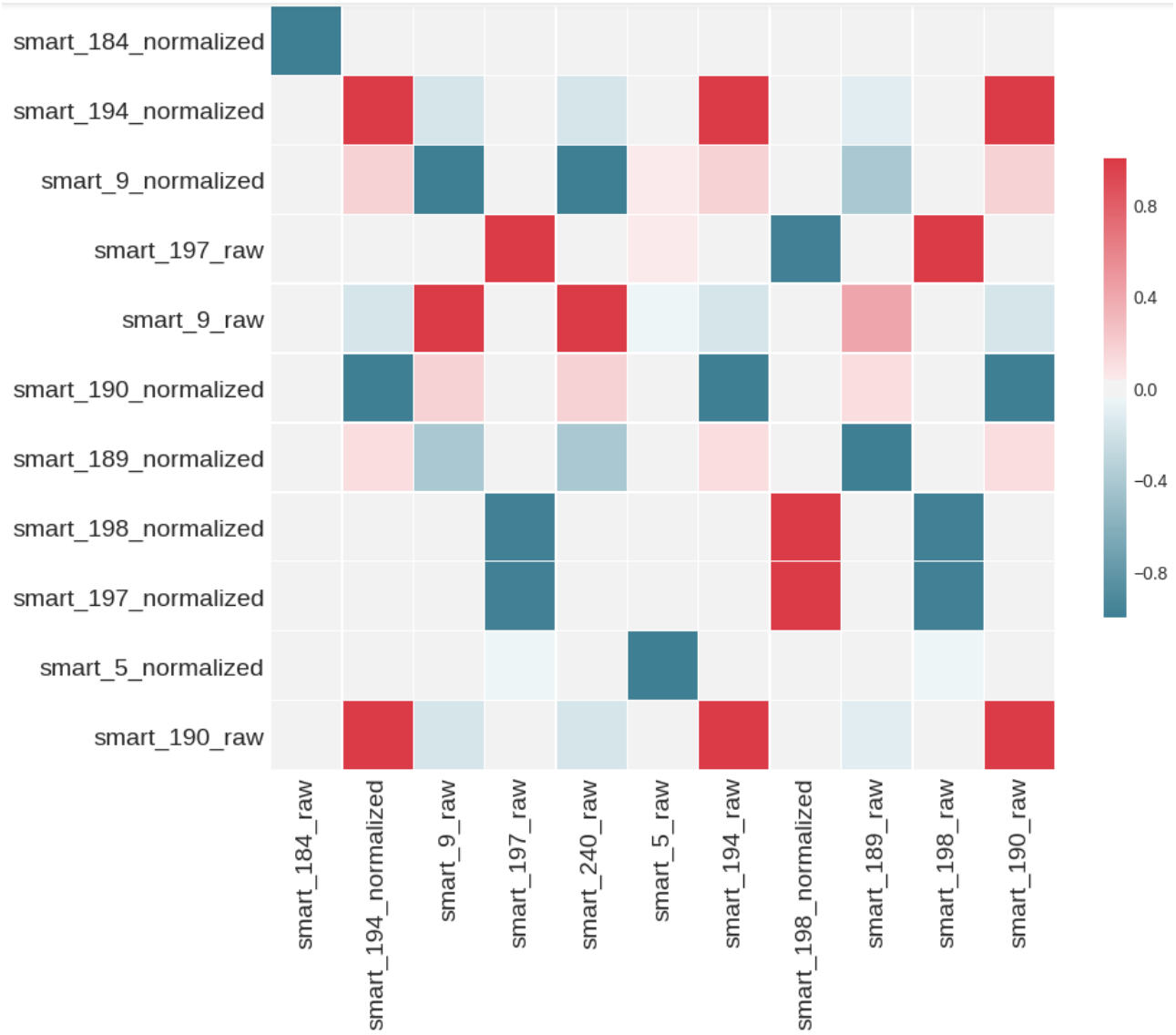

Figure 4: Corelation above Threshold

Last part is to calculate the importance of attributes to predict the RUL of HDD. Then remove the attributes with zero importance and attributes with importance less then threshold. To calculate normalized importance(Figure[5]) feature selector is using Gradient boosting model.

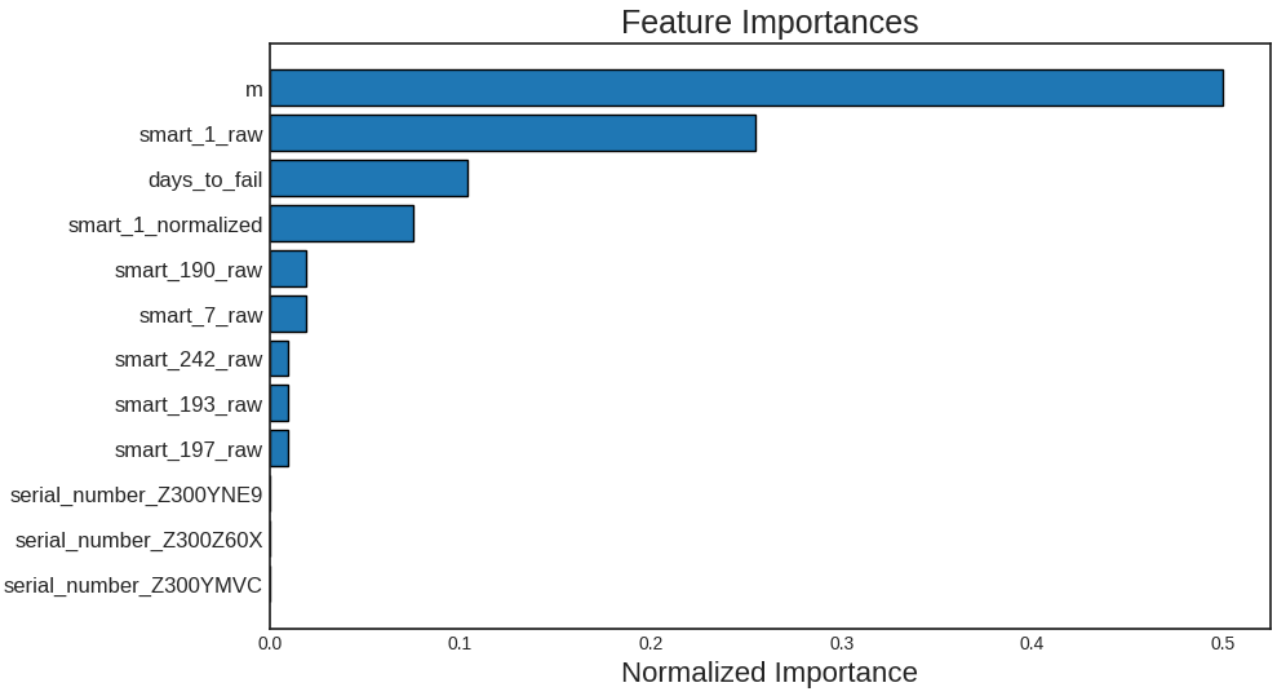

Figure 5: Attribute vs Normalized Importaance

Feature Selector returns the following top 5 attributes -

	feature	importance	normalized_importance	cumulative_importance
0	smart_1_raw	153.0	0.106918	0.106918
1	smart_242_raw	135.6	0.094759	0.201677
2	smart_241_raw	130.6	0.091265	0.292942
3	smart_7_raw	127.1	0.088819	0.381761
4	smart_9_raw	108.7	0.075961	0.457722

In the paper mentioned, they uses following attributes to train LSTM model.

SUMMARY OF SMART FEATURES USED

ID	Attribute Name	Description
SMART 7	Seek error rate	Frequency of the errors during disk head positioning and rises with approaching failure.
SMART 9	Power-on-hours count	Estimated remaining lifetime, depending on the time a device was powered on. Raw value indicates the actual powered-on time, usually in hours.
SMART 240	Head flying hours or transfer error rate	Time spent during the positioning of the drive heads
SMART 241	Total LBAs written	Related to the use and hence indicating the aging process of hard drives
SMART 242	Total LBAs read	Related to the use and hence indicating the aging process of hard drives

Second part of the project was to reiterate the results of the mentioned research paper. For this part, I have used Keras liberary to train LSTM Model. I have created all the layers mentioned in Figure[1] with associated value. But this model was not able to get the results which were mentioned in paper. After the analysis and information recieved from author, in their model they are no shuffling the training and testing data. Because of this, their LSTM model learning directly the decreasing value of remaining useful days. Hence getting the minimum loss.

PROJECT CODE LINK-

1. Feature Selector-
https://colab.research.google.com/drive/1gfj013NCKq2Q9aL76jv_zUmd86kGiL3K?usp=sharing

2. RUL Implementation using LSTM-
https://colab.research.google.com/drive/1BqbS24m9XRp0D2c9KjnEQbFxFcC_P_ru?usp=sharing