OPA Towobola

Remaining Useful Life Characterization and Prediction for Components

16 November

Provided to meet the requirements of the Galvanize Data Science Immersive Capstone 2

**Data requirement:**

* Contains >1000 observations/rows.
* Stick to data sets smaller than 1 GB if you want to load all the data into pandas. If you choose a larger data set, you will need to spend additional time learning how to work with it out-of-core, or reduce the number of observations.
* Now that you have experience cleaning data, the focus should be on modeling. Avoid data that needs 2+ days of cleaning.
* There are countless data sets online without information about when they were collected, and without good descriptions of features, or anonymized features. Please be mindful that you should be trying to answer an interesting question or solve a practical problem, so spend time investigating a potential data set to find out when it was collected and whether it provides data that leads to practical applications.

**Proposal:**

Three proposals with each having:

1. A description of each columns contents within the data set
2. State 3 problems you are trying to solve or questions you are trying to answer
3. Screenshot of the data loaded in python showing rows and columns

**Opportunity to practice:**

* Feature engineering
* Predictive modeling
* Inferential modeling
* Unsupervised learning

**Required deliverables:**

* At least 2 non-neural network models.
* You may pursue neural network or time series forecasting models after producing 2 non-neural network models. The purpose of this is for you to gain practical knowledge of popular classical machine learning algorithms.
* [**LA Campus DSI Capstone Rubrics**](https://docs.google.com/spreadsheets/d/1Xlh3hySKEZi7EGVHLtFGg8Hymoln1yKDmci3RzIoNAk/edit?usp=sharing)

**RESOURCES**

Capstone Brainstorm Sources:

* [Past Galvanize Student Capstones repository](https://github.com/GalvanizeDataScience/project-proposals/blob/master/past_student_projects.md)
* [Past Los Angeles Cohort Capstones](https://docs.google.com/spreadsheets/d/1gHGXiTkexGmwOnd66-vI2XECOQLccucNxn9DK_5U6Uc/edit?usp=sharing)

List of Open-Source Datasets:

* [Google public datasets](https://www.google.com/publicdata/directory)
* [Data.world](https://data.world/)
* [FiveThirtyEight](https://data.fivethirtyeight.com/) -
* [Awesome public datasets](https://github.com/awesomedata/awesome-public-datasets)
* [Data.gov](https://www.data.gov/)
* [Google Dataset Search Engine](https://datasetsearch.research.google.com/)
* [Kaggle](https://www.kaggle.com/)
* [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/index.php)
* [OpenML](https://www.openml.org/search?type=data)

Project Ideas

**Project Idea #1:**

* **Topic (Title): Machine Remainin Useful Life**
* **State 3 problems that can be solved or questions that can be answered with your data:**

Useful life When a company or an individual buys an engine, on the forefront of their minds is the question of how much useful life can be expected from said engine.

a) how much remaining life is there on an engine

b) when should the engine be replaced

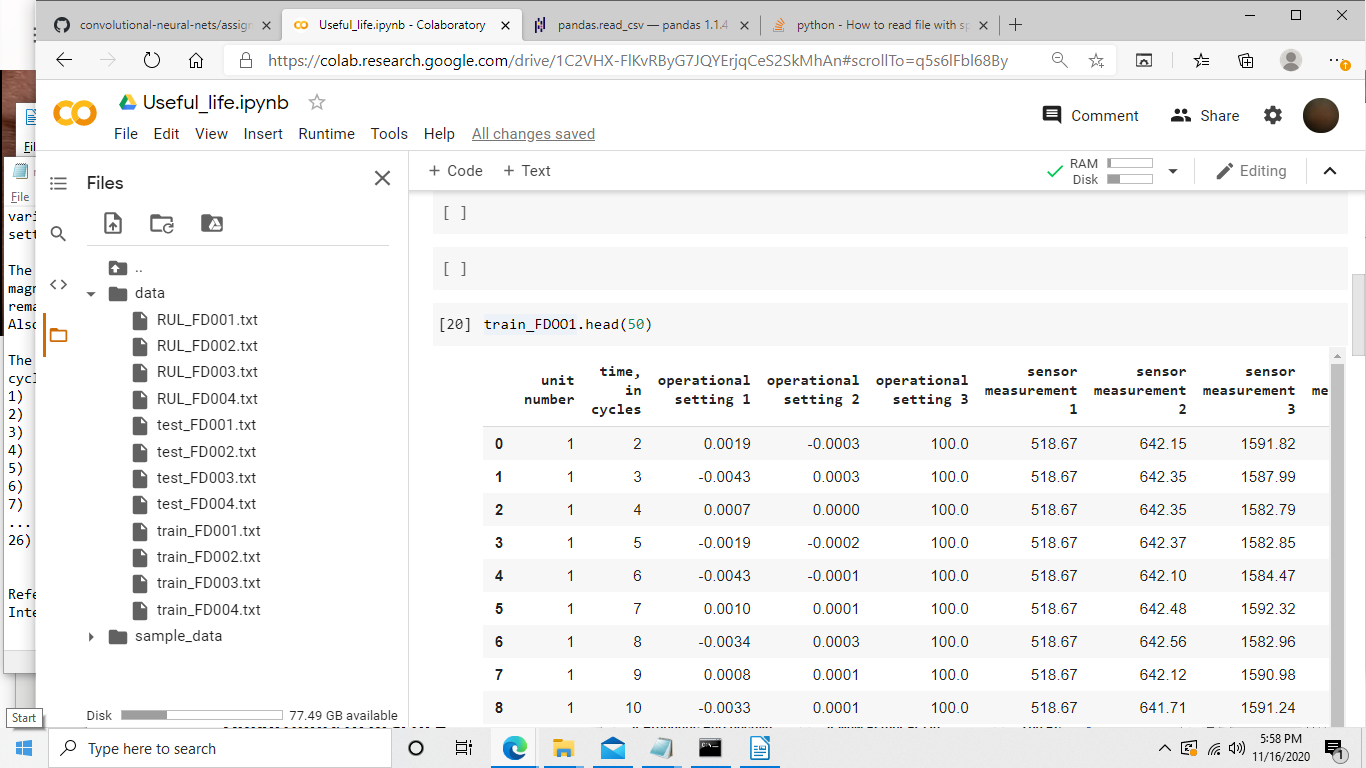
c) what factors indicate remaining useful life

* **Data Source(s):** 
  + **Websites and/or databases: https://catalog.data.gov/dataset/c-mapss-aircraft-engine-simulator-data**
  + **Number of data points (rows):** 20630 entries
  + **A description of each columns contents in the data set**

**Multiple time series are provided. As described by the publisher**

‘Each time series is from a different engine – i.e., the data can be considered to be from a fleet of engines of the same type. Each engine starts with different degrees of initial wear and manufacturing variation which is unknown to the user. This wear and variation is considered normal, i.e., it is not considered a fault condition. There are three operational settings that have a substantial effect on engine performance. These settings are also included in the data. The data is contaminated with sensor noise.  
  
The engine is operating normally at the start of each time series, and develops a fault at some point during the series. In the training set, the fault grows in magnitude until system failure. In the test set, the time series ends some time prior to system failure. The objective of the competition is to predict the number of remaining operational cycles before failure in the test set, i.e., the number of operational cycles after the last cycle that the engine will continue to operate. Also provided a vector of true Remaining Useful Life (RUL) values for the test data.  
  
The data are provided as a zip-compressed text file with 26 columns of numbers, separated by spaces. Each row is a snapshot of data taken during a single operational cycle, each column is a different variable.” The columns correspond to:  
1) unit number  
2) time, in cycles  
3) operational setting 1  
4) operational setting 2  
5) operational setting 3  
6) sensor measurement 1  
7) sensor measurement 2  
...  
28) sensor measurement 3

* + **Screenshot of first 5 rows of data loaded into python:**



* *(Optional)* **Potential Future Employer: Boeing, SpaceX**

**Project Idea #2:**

* **Topic (Title): Remaining hard drive cycles**
* **State 3 problems that can be solved or questions that can be answered with your data:**

One of the most frustrating incidents a computer user can experience is the failure of their hard drive.

a) how much remaining life is there on a hard drive

b) when should the hard drive be replaced

c) what factors indicate remaining useful life

* **Data Source(s):** 
  + **Websites and/or databases:** https://www.backblaze.com/blog/hard-drive-data-feb2015/
  + **Number of data points (rows):**

21688

* + **A description of each columns contents in the data set**

**As described by the publisher,** Backblaze has released the raw data collected from the more than 41,000 disk drives in our data center. To the best of our knowledge, this is the largest data set on disk drive performance ever to be made available publicly. What Does It Look Like?

Each daily stats file is in CSV (comma-separated value) format. The first line lists the names of the columns, and then each following line has all of the values for those columns. Here are the columns:

Date: The date of the file in yyyy-mm-dd format.

Serial Number: The manufacturer-assigned serial number of the drive.

Model: The manufacturer-assigned model number of the drive.

Capacity: The 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.

SMART Stats: 80 columns of data that are the raw and normalized values for 40 different SMART stats as reported by the given drive. Each value is the number reported by the drive.

The Wikipedia page on SMART (https://en.wikipedia.org/wiki/S.M.A.R.T.) has a good description of all of the data, and what the raw and scaled values are. The short version is that the raw value is the data directly from the drive. For example, the “Power On Hours” attribute reports the number of hours in the raw value. The normalized value is designed to tell you when the drive is OK. It starts at 100 and goes down to zero as the drive gets sick. (Some drives count down from 200.)

How to Compute Failure Rates

One of my statistics professors once said, “It’s all about counting.” And that’s certainly true in this case.

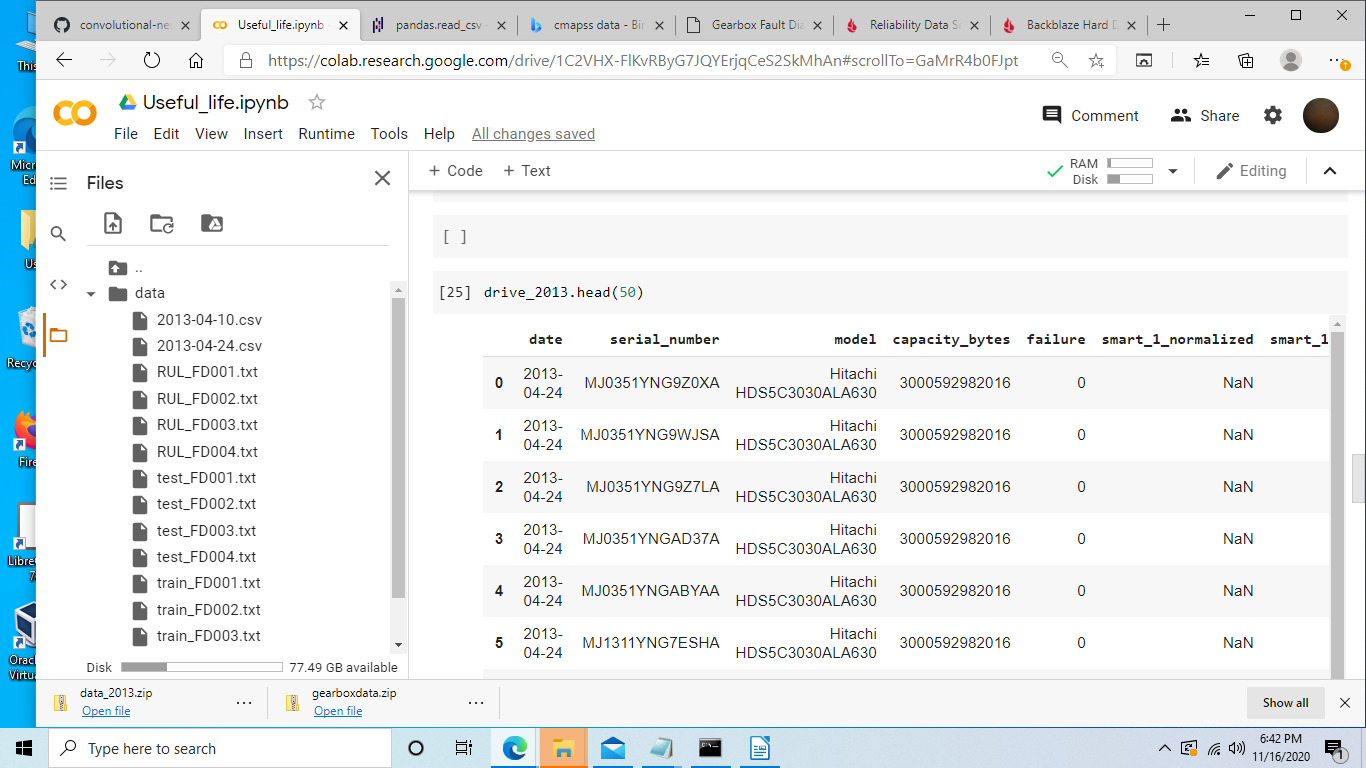
A failure rate says what fraction of drives have failed over a given time span. Let’s start by calculating a daily failure rate, which will tell us what fraction of drives fail each day. We’ll start by counting “drive days” and “failures.”

To count drive days, we’ll take a look every day and see how many drives are running. Here’s a week in the life of a (small) data center:

blog\_datacenter\_dots\_1

Each of the blue dots represents a drive running on a given day. On Sunday and Monday, there are 15 drives running. Then one goes away, and from Tuesday through Saturday there are 14 drives each day. Adding them up we get 15 + 15 + 14 + 14 + 14 + 14 + 14 = 100. That’s 100 drive days.

* + **Screenshot of first 5 rows of data loaded into python:**



* *(Optional)* **Potential Future Employer: HP**

**Project Idea #3:**

* **Topic (Title):**

**Remaining gear box life**

* **State 3 problems that can be solved or questions that can be answered with your data:**

Design engineers and users want to know the remaining useful life on a gear box

a) how much remaining life is there on a gear box

b) when should the gear box be replaced

c) what factors indicate remaining useful life

* **Data Source(s):** 
  + **Websites and/or databases:**

https://openei.org/datasets/dataset/gearbox-fault-diagnosis-data/resource/affa53da-cae6-42f2-b898-ad018ff91641

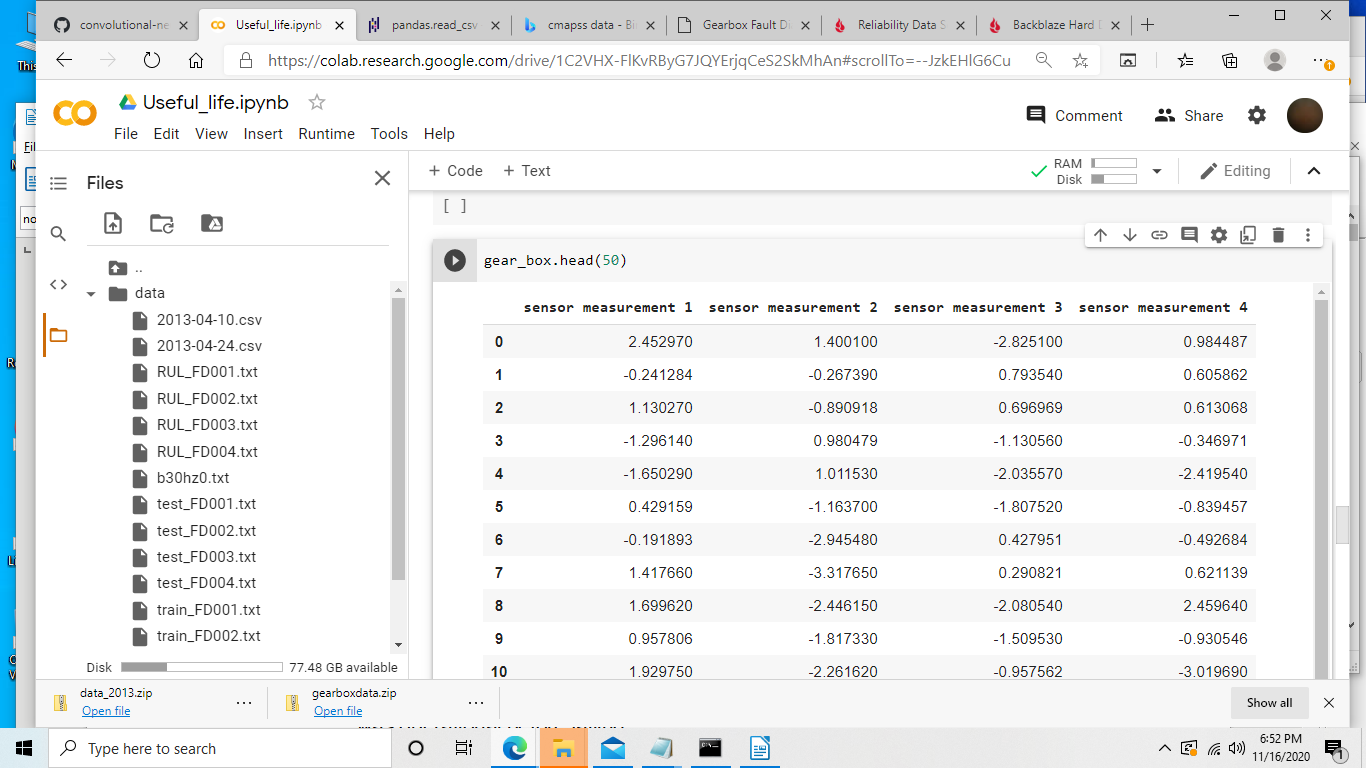
* + **Number of data points (rows):**

88319

* + **A description of each columns contents in the data set**

Multiple time series are provided. As described by the publisher, data set includes Healthy and Broken Tooth data set. Data set has been recorded under variation of load from '0' to '90' percent load with 4 different sensors in four directions. Ten different text files are available for each case

* + **Screenshot of first 5 rows of data loaded into python:**



* *(Optional)* **Potential Future Employer: Pacific Gas & Electric, Southern California Edison, HP**