**PSYENCE 산학프로젝트**

**Nasa Turbofan RUL prediction(잔여수명) 예측 위주로**



학번: 201820196

이름: 김진찬

<목차>

1. **연구 문제 및 주제**
2. **데이터 셋 가공 과정**

2.1 EDA

1. **방법론**
   1. 데이터 전처리
   2. Fast Fourier Transform
   3. Discrete Wavelet Transform
   4. 유사성 모델
   5. 건전 지표 모델
   6. 생존 모델
2. **실험**
   1. 유사성 모델 실험
   2. 건전 지표 모델 실험
   3. 생존 모델 실험
3. **결과**
4. **결과 해석**
5. **출처**
6. **연구 문제 및 주제**

예측 및 상태 관리는 다운타임 및 장애를 방지하기 위한 자산 상태 예측을 위해 업계에서 중요한 주제이다. 특히 고가의 장비들이 즐비한 공장에서 장비 이상으로 인해 공정이 중단하게 되면 장비 수리비와 함께 공정 중단으로 인한 막대한 손실이 발생하게 된다. 따라서 사전에 유지 보수를 수행해야 하는 시기를 추정하는 잔여 수명 예측은 공장 운영의 성패를 가를 수 있는 중요한 문제라고 할 수 있다. 본 보고서는 NASA에서 제공하는 NASA Turbofan Engine 데이터를 사용하여 잔여수명에 사용되는 대표적인 3가지 방법들을 활용하여 잔여수명 예측하고 예측한 결과를 바탕으로 더 좋은 방법을 제시하고자 한다. 더 나아가 산학 협력 관계를 맺고 같이 프로젝트를 진행한 Pysence 기업에 다변량 시계열 데이터에서 잔여수명 예측을 할 때 사용될 수 있는 머신러닝 모델을 제시하여 인사이트를 제공하고자 한다.

Pysence에서 제시한 과제는 주조 공정에 설치된 다양한 센서들을 통해 장비의 고장을 예측하고 이를 사전에 정비하는 것이다. 슬라브 주조 공정에서 사용되는 다양한 장비들에 센서를 설치한 후 측정된 센서 데이터 값을 통해 장비를 교체할 시기를 구하고자 한다. 해당 문제는 Nasa Turbofan에서 제공하는 데이터와 유사하다고 판단하여 Nasa Turbofan 데이터를 활용하여 다변량 시계열 데이터에서 잔여수명을 예측하는 방법론을 제시하겠다.

공장의 상태를 관리하기 위해 사용되는 대표적인 방법에 이상 감지가 있다. 이상 감지는 공장에서 상태를 점검하고 싶은 장비에 센서를 부착하여 비정상적인 동작이 감지되면 즉각적으로 조치를 취하는 방법이다. 이를 통해 생산하는 제품에 비정상적인 동작이 감지되면 제품 불량률을 줄이고 더 나은 제품을 생산하여 제품의 품질을 향상시킬 수 있고 미리 이상을 감지하여 조치를 취함으로써 장비 고장 및 생산 중단을 방지하여 비용을 절감 할 수 있다. 하지만 이는 Pysence에서 제시한 문제를 푸는 데에는 다소 어려움이 있다.

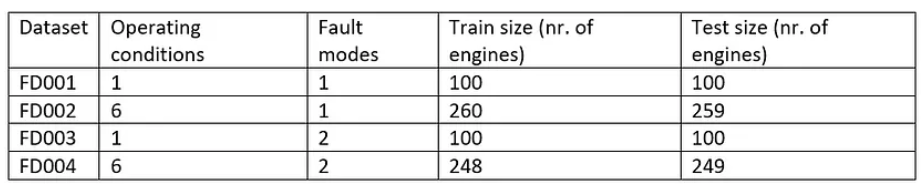
가장 대표적인 이유는 주어진 데이터가 실시간 데이터가 아니라는 점이다. 이상감지는 실시간으로 데이터를 감지하고 실시간으로 조치를 취할 때 유의미한 결과를 얻어낼 수 있다. 예를 들어, 반도체 공정에서 웨이퍼를 이송 시키는 oht에 이상이 생겨 작동을 멈췄다고 하자. 이때 반도체 공정은 스마트 팩토리 시스템이 적용되어 있기 때문에 즉각적으로 다른 oht의 운송 루트를 변경하거나, 해당 생산라인의 생산량을 줄이는 등의 조치를 취해 bottleneck이 발생하는 것을 줄일 수 있다. 하지만 주조 공정에서는 반도체 공정과 달리 스마트 팩토리 시스템이 구비되어 있지 않아 즉각적으로 이상 상태에 대처하는 것이 쉽지 않다. 따라서 센서 데이터를 활용하여 이상치를 감지하여 조치를 하는 것보다 각 센서들을 통해 장비의 수명을 예측하여 초기 조치를 취하는 것이 더 적합하다고 판단하였다.

잔여 수명 예측은 통상적으로 3가지 방법론을 사용한다. 첫 번째 방법은 유사성 모델을 활용하는 것이다. 유사성 모델은 비슷한 동작을 보이는, 유사하거나 다른 구성 요소의 Run-To-Failure 데이터를 활용해 RUL을 추정하는 방식이다. 두 번째 방법은 생존 모델을 활용하는 것이다. 생존 모델은 수명 데이터가 주어졌을 때 사용된다. 마지막 세 번째 방법은 건전 지표를 활용하는 것이다. 이때는 도메인에 따라 규정된 임계 값이 주어졌을 때 사용될 수 있다. Nasa Turbofan 데이터 셋에 위 3가지 방법들을 모두 적용시키고 가장 성능이 좋은 방식을 제안하겠다.

한 공정 안에 있는 다변량 시계열 데이터를 다룰 때 서로 다른 변수들이 서로 영향을 주는 지 확인하는 것이 중요하다. 한 장비에서 이상 상태가 발생하면 다른 장비에 영향을 주게 되어 센서에 측정되는 데이터가 해당 장비의 고유 데이터가 아닌 다른 장비의 결함으로 인해 노이즈가 추가된 데이터가 측정되기 때문이다. 측정된 데이터에서 각 장비의 고유 센서 값을 구하기 위해서는 별도로 데이터를 필터링 하는 과정이 필요하다. 본 보고서에는 Nasa Turbofan 엔진들의 고유 주파수를 구하기 위해 DWT(discrete wavelet transform)을 하는 방식을 적용하였다. 기존에 대부분의 연구에서 진행된 필터링 방식은 이동평균법인데 해당 방법과 dwt를 진행한 방법과 서로 비교를 해보려고 한다.

1. **데이터 셋 가공**

분석에 활용된 데이터 셋은 Nasa에서 제공하는 Nasa TurboFan 데이터 셋 중에서 FD001 데이터 셋이다.



각 데이터 셋은 훈련, 테스트, RUL 세트로 나뉜다. 각 엔진은 초기 마모 및 제조 변동 정도가 다른 상태에서 시작한다. 엔진은 정상적으로 작동하다 어느 시점에서 고장이 발생한다. 훈련 셋에서는 시스템 고장이 발생할 때까지 선형적으로 고장 정도가 심해진다. 테스트 셋에서는 시계열이 시스템 고장보다 어느 정도 먼저 종료된다. 해당 분석 목적은 테스트 세트에서 고장이 발생하기 전 사이클 수, 즉 엔진이 계속 작동할 마지막 사이클 이후의 작동 사이클 수를 예측하는 것이다.

FD001 데이터 셋은 100개의 엔진을 작동시켜 21개의 센서 값을 통해 데이터를 측정하였다. 데이터 셋은 26개의 속성이 있고 엔진 고유 번호, 사이클 수(시간), 세가지 세팅 상태, 21개의 센서 값으로 이루어져 있다. 인스턴스는 훈련 데이터 셋 기준으로 20631개 있다.

[FD001 훈련/테스트 데이터 셋 Attribute]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| UnitNumber | Cycle | OpSet1 | OpSet2 | OpSet3 | Sensor 값 | RUL |

UnitNumber: 엔진 고유 번호

Cycle: 사이클 수

OptSet1~3: 3가지 설정(해당 분석에는 활용하지 않을 예정이다)

Sensor1~21: 센서 값

RUL: 잔여 수명

[FD001 RUL 데이터 셋 Attribute]

|  |  |
| --- | --- |
| RUL | UnitNumber |

RUL: 잔여 수명

UnitNumber: 고유 기계 번호

* 1. **EDA**

**[고유 기계 번호와 사이클 수 기술 통계량]**

각 고유 기계 번호와 사이클 수의 통계 값을 통해 데이터 분포를 확인하였다.

.

* Count 값을 확인하면 데이터 집합의 행 수는 총 20631개라는 것을 알 수 있다.
* Min-Max 값을 통해 기계는 1에서 100개까지 있다는 것을 알 수 있다.
* 평균과 분위수가 기술 통계량과 깔끔하게 일치하지 않는 것을 확인 할 수 있다. 이는 각 기계마다 Max Cycle이 다르기 때문이다.

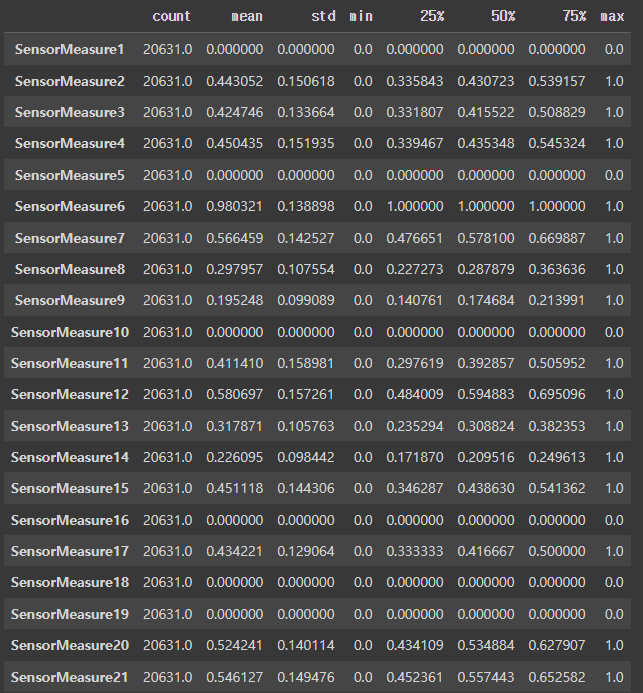
**[고유 기계 별 사이클 수 기술 통계량]**

고유 기계 별 사이클 수를 확인하는 과정을 거쳤다.



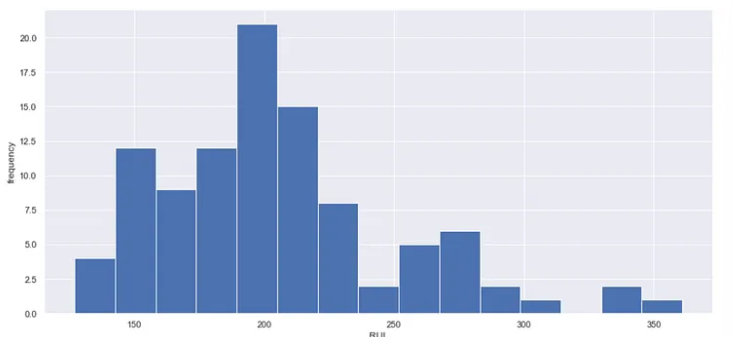
* Min 값을 보면 가장 일찍 고장난 엔진은 128, 가장 오래 작동한 엔진은 362 사이클 후 고장 나는 것을 알 수 있다.
* 평균 엔진은 199 사이클과 206 사이클 사이에서 고장나지만 표준편차를 보면 46 사이클이 나는 것을 알 수 있다. 이는 비교적 표준편차가 크다고 볼 수 있다.

**[각 센서 값의 기술 통계량]**



해당 통계 값을 보면, 센서 1, 10, 18, 19가 변동하지 않는 것을 알 수 있다. 이는 유용한 정보를 전달하지 않는다고 판단할 수 있다. 또한 분위수를 통해 센서 5, 6, 16도 변동이 거의 없음을 알 수 있다. 이 또한 데이터를 한번 알아봐야 한다는 것을 알 수 있다.

**[RUL 값의 히스토그램]**



RUL 데이터 셋을 확인하면 좌측으로 skewed 되어 있다는 것을 알 수 있다.

1. **방법론**

본 보고서에는 유사성 모델, 건전 지표 활용 모델, 생존 모델 총 3가지 방법을 통해 Nasa Turbofan 데이터 셋의 잔여 수명 예측을 진행하고 그 중 가장 적합한 모델을 찾아내는 것을 목표로 한다. 해당 장에서는 각 방법론들의 이론에 대해 설명하려고 한다.

**Data Preprocessing(데이터 전처리)**

대부분의 경우 한 공정 안에 있는 장비들의 주파수를 구할 때 각 장비의 고유 주파수가 측정되는 것이 아닌 다른 장비의 결함 요소로 인한 측대파로 나타난다. 따라서 받은 측대파를 filtering 방법을 거쳐 고유 주파수를 구해주는 과정을 거쳐야 한다. 이때 filtering 하는 방식으로는 급속 푸리에 변환(Fast Fourier Transform), 이동 평균 방법(Moving Average), 이산 웨이블릿 변환(Discrete wavelet Transform) 등이 있다. 기존 연구에서는 대부분 이동평균법을 활용하여 데이터를 전처리 하는 과정을 거쳤으나 본 보고서에는 Discrete wavelet Transform을 활용하여 데이터를 전처리한 방식으로 진행하였다.

**Fast Fourier Transform**

신호처리를 할 수 있는 가장 대표적인 방법 중 하나가 바로 Fast Fourier Transform이다. 푸리에 변환은 신호의 주기성을 공부할 때 사용되는 방법이다. 푸리에 변환은 신호들을 주파수 성분으로 분해하는 방법이다.

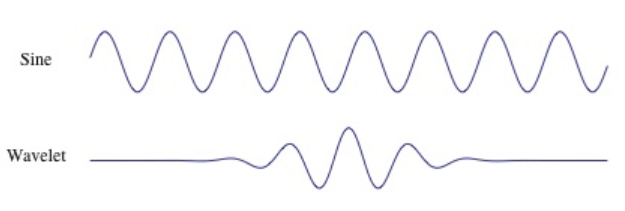
모든 신호는 더 간단한 형태의 신호 싸인 혹은 코싸인 형태의 신호의 합 형태로 분해가 가능하다. 시간 영역에서 주파수 영역으로 변환하는 것을 푸리에 변환이라고 부른다. 반대 과정을 하는 것을 역푸리에 변환이라고 한다.

**Discrete Wavelet Transform**

푸리에 변환은 신호를 시간 차원에서 주파수 차원으로 변환하는 데 획기적인 기법이지만 시간을 반영하지 못한다는 치명적인 단점이 존재한다. 이를 해결하기 위해 사용되는 방법이 STFT(Short Term Fourier Transform)이다. 해당 방법은 원본 신호를 동일한 길이의 window를 가지고 나눠서 푸리에 변환을 하는 것이다.

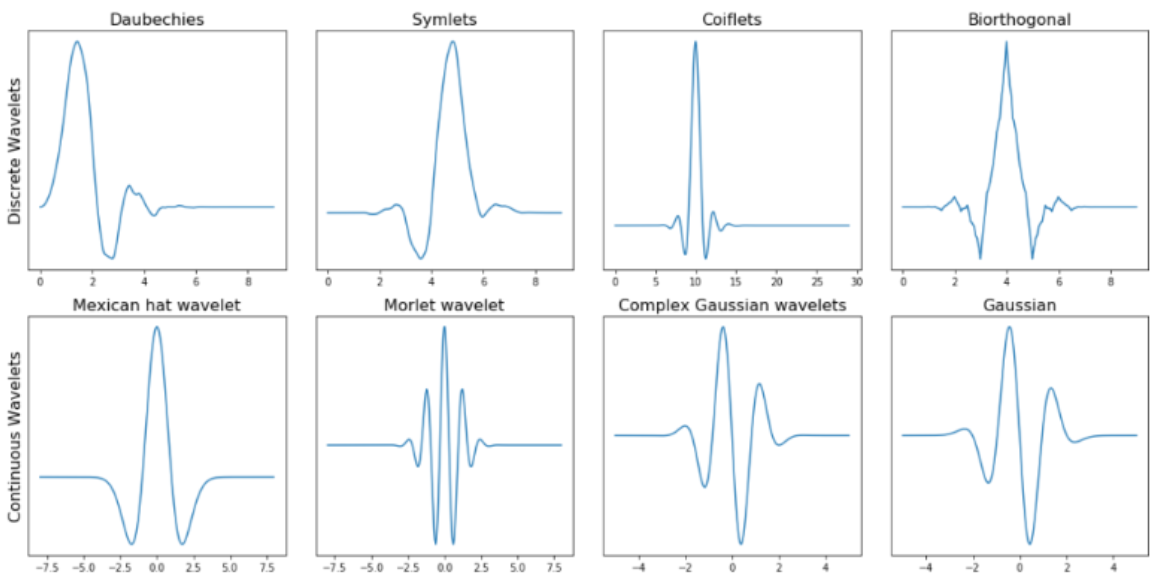
하지만 STFT 또한 푸리에 변환의 일환이기 때문에 푸리에 변환의 불확실성 원칙이라는 문제에서 자유롭지 못하다. STFT에서 윈도우 크기를 줄일수록 신호 위치를 파악하기 쉽지만 주파수의 값을 구하긴 어려워진다. 반면 윈도우 크기를 키울수록 주파수의 값을 구하긴 쉬워지지만 신호의 위치를 구하긴 어려워진다.

이를 해결하기 위한 방법으로 Wavelet Transform을 사용할 수 있다. 푸리에 변환은 싸인 형태로 신호를 반환한다. 왜냐하면 하나의 신호가 싸인 신호의 선형식으로 존재하기 때문이다. Wavelet은 싸인 형태의 신호가 아닌 다양한 형태의 신호를 사용한다.



싸인 신호와 Wavelet의 가장 큰 차이는 싸인 신호는 그 영역이 무한한 반면, Wavelet 같은 경우 특정 지역에 대한 파형을 갖는다. 이러한 특성 때문에 Wavelet은 푸리에 변환과 달리 시간적인 특성을 반영할 수 있다.

Wavelet은 여러 형태의 파형을 제공한다. 따라서 여러 Wavelet 중 가장 좋은 결과 값을 구해내는 Wavelet을 선택하면 된다. 다음은 wavelet의 종류이다.



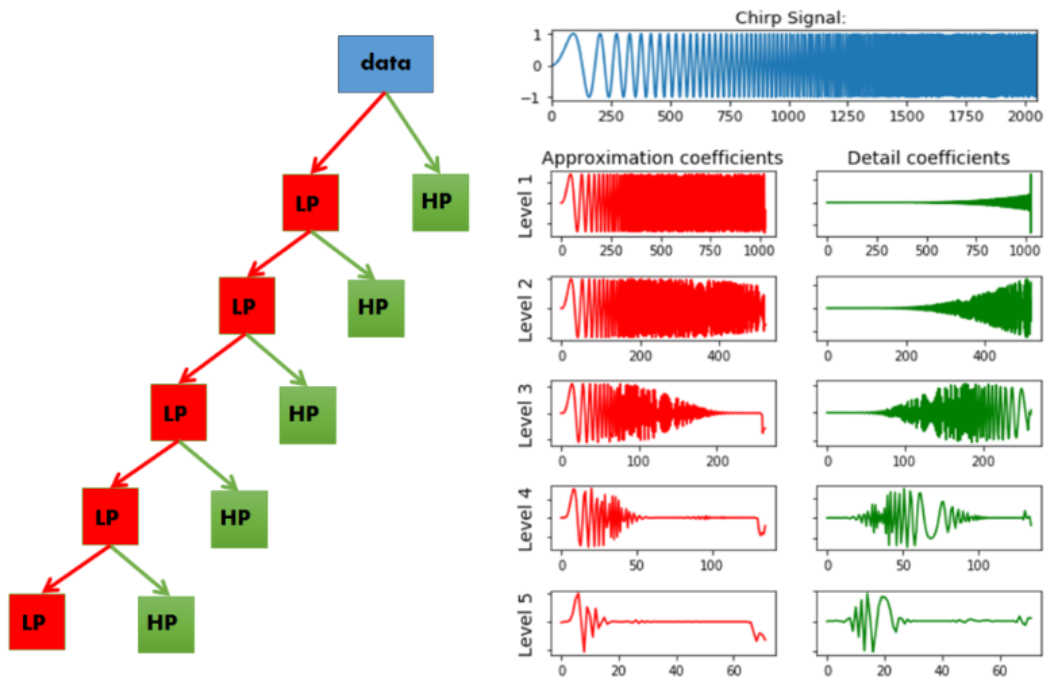
**[DWT]**

Wavelet은 연속형태 혹은 이산형태로 나타난다. 해당 보고서에서는 이산 형태의 DWT만 다루도록 하겠다.

DWT는 filter-bank 형태로 실행된다. 여기서 filter-bank은 high-pass와 low-pass filter를 활용하여 신호를 효율적으로 여러 가지의 주파수 밴드 형태로 나누는 것을 의미한다.

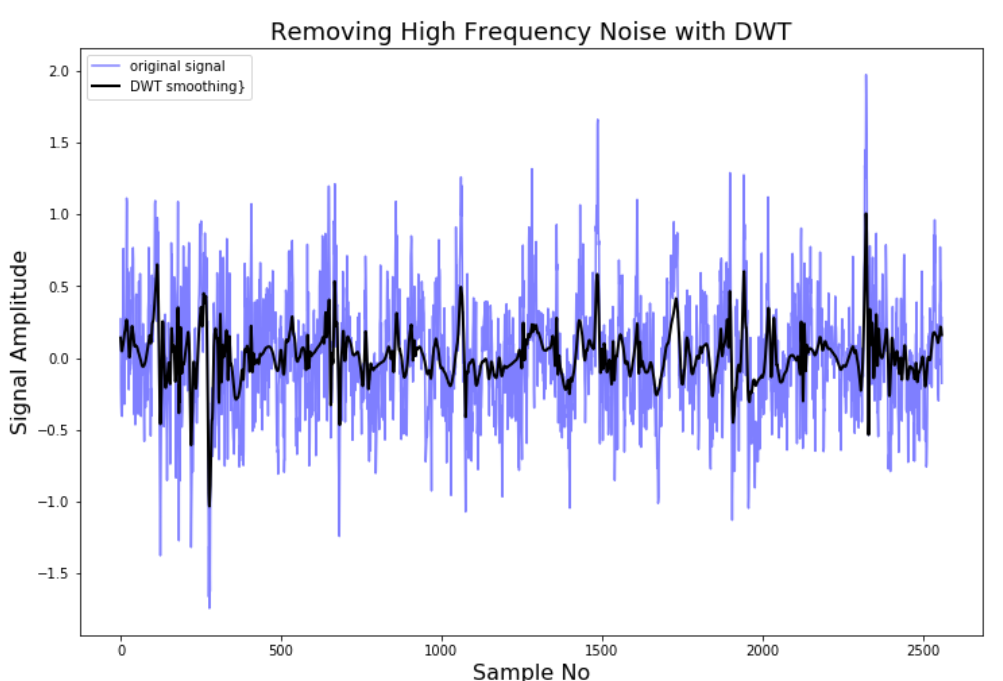
DWT를 신호에 적용할 때, 가장 작은 scale 값에서부터 시작한다. Fa = Fc/a 식에 따르면 scale 값이 작을수록 주파수 값이 커지므로 처음에 가장 높은 주파수 값을 분석하는 것이라고 할 수 있다. 두 번째 스테이지에서는 scale 값이 2배 커지게 된다. 따라서 가장 높은 주파수의 절반에 해당하는 값을 분석하게 된다. 이런 식의 계산은 최대 분해 정도를 다 다를 때까지 진행된다.

예를 들자면, 처음 신호의 주파수가 1000Hz라고 했을 때, 첫 번째 stage에서는 신호를 low-frequency 부분과(0-500Hz) high-frequency(500Hz-1000Hz) 부분으로 나뉘게 된다. 두 번째 stage에서는 low-frequency 부분의(0-500Hz)를 0-250Hz와 250-500Hz로 나뉜다. 이런 식으로 진행되다 신호의 길이가 Wavelet의 크기 보다 작아질 때까지 진행된다. 이를 시각화하면 다음과 같다.



**[DWT를 활용한 신호 분해]**

Dwt가 베어링 filtering에 활용되는 방식은 딥러닝에서 Auto-Encoder가 사용되는 방법과 유사하다. Pywt.dwt()함수를 사용하여 분해한 신호들을 다시 원본 신호를 회생시키는 과정에서 불필요한 신호들을 제거할 수 있다. 다음은 dwt를 활용하여 신호를 처리하기 전과 후를 나타내는 그림이다.



**3.1 유사성 모델 활용(similarity model)**

비슷한 동작을 보이는, 유사하거나 다른 구성요소의 Run-To-Fail 데이터를 담은 데이터 베이스가 있다면 유사성 모델을 활용하여 잔여 수명(RUL)을 예측할 수 있다. Nasa Turbofan 데이터 셋은 총 100개의 기계의 Run-To-Fail 데이터가 있기 때문에 유사성 모델을 활용하여 잔여 수명 예측을 할 수 있다.

유사성 모델은 이전 데이터와 현재 데이터의 유사성을 학습하고 예측을 수행하는 방법이다. 이전 데이터와의 유사성을 기반으로 시스템의 상태를 추론하고, 잔여 수명 예측을 한다.

**3.1.1 유사성 모델의 Feature Selection**

**[추세를 활용한 Feature selection]**

모델을 생성할 때 모델에 적합한 특성을 선택하여 분석을 하는 것이 중요하다. 유사성 모델을 사용할 때는 추세를 활용하여 특성을 추출하였다. 각 센서 값과 시계열과의 상관관계를 구한 후 상관관계가 적은 값들을 제거하는 과정을 거쳤다.

**3.1.2 유사성 모델 분석**

Nasa Turbofan에서 잔여수명 값은 시간이 지날수록 선형적으로 줄어들기 때문에 선형 열화 모델을 활용하였다. 선형 열화 모델을 만들기 위해선 크게 다음과 같은 과정을 거치게 된다.

**[유사성 모델의 단계]**

1. 센서 측정 값들을 가지고 단일 건전성 지표로 결합을 한다.
2. 센서 값을 신호처리 및 전처리 방법을 통해 노이즈를 제거한다
3. 융합된 데이터를 다차방정식에 피팅하여 유사성 모델을 완성한다.

해당 데이터 셋의 잔여 수명이 선형적으로 감소하기 때문에 전체 센서 값을 독립변수로, 잔여 수명을 종속 변수로 놓고 선형 회귀를 돌렸다. 이때 나오는 계수들을 가중치로 활용해 전체 센서 측정 값들을 대표할 수 있는 하나의 단일 건전성 지표를 만든다. 이후, 만든 지표를 활용하여 다차곡선 형태의 모델 예측 값을 만든다.

* 1. **건전 지표 활용(Health Indicator model)**

많은 경우, 정비 일정이 보수적이라면 기계가 고장이 날 경우는 드물다. 이는 Run-To-Failure 데이터가 없을 때가 많다는 것을 의미한다. 하지만 규정된 임계 값은 주어질 수 있다. 임계 값을 하나의 건전지표로 활용할 수 있다. 건전지표를 활용한 잔여 수명 예측 방법은 지수적 감소 모델(Exponential degradation Model)을 만든 후 몇 일 만에 임계 값을 넘어설 것인지를 예측하는 방법이다.

* + 1. **건전 지표 모델 Feature Selection**

건전 지표를 만들어 feature selection을 할 때는 단조 특성(Monotonicity)을 구하여 적합한 특성을 추출하는 방식을 사용한다.

단조 특성(Monotonicity)은 어떤 변수나 특성이 다른 변수나 특성과의 관계가 항상 증가하거나 감소하는 경향을 가지는 것을 의미한다. 즉, 변수 X가 증가할 때 Y도 증가하거나 감소하는 관계를 말한다. 이는 해당 특성 값이 고장이 진행되고 있는 지를 확인할 수 있는 지표이다. 만일 고장이 진행되고 있으면 단조 특성을 갖게 되고 Y 값은 1이 된다. 고장이 진행되고 있지 않으면 단조 특성을 갖지 않게 되고 Y 값은 0이 된다.

* + 1. **건전 지표 모델 분석**

지수적 감소 모델은 시스템의 성능 또는 수명이 시간에 따라 지수적으로 감소한다고 가정한다. 해당 모델은 시스템 성능 및 수명이 시간에 따라 지수적으로 감소하고 있을 때 적합하다. Nasa Turbofan 데이터 셋 같은 경우, 여러 센서 값이 주어졌으므로 전체 센서 값을 대표할 수 있는 데이터를 뽑아내는 과정이 필요하다. 이때 사용되는 것이 PCA이다. 센서 데이터에 PCA를 적용하여 전체 센서 값들을 대표할 수 있는 하나의 주성분을 추출할 수 있다. 해당 주성분과 시간과의 관계를 분석하여 지수 함수의 형태를 하고 있는 지 확인한다. 만일, 주성분과 시간이 지수 관계에 있을 시, 주성분의 시간에 따른 점수가 지수적으로 감소한다고 가정 할 수 있다. 이 가정이 성립할 시, 주성분 점수를 PC(t)라고 표시하고 시간에 따라 지수적 감소 모델(Exponential Degradation Model)을 생성할 수 있다.

[Exponential Degradation 모델을 활용한 잔여 수명 예측 모델 단계]

1. 단조 특성을 사용하여 예지 진단을 위해 특징의 이점을 수량화한다.
2. PCA를 사용하여 건전성 지표를 생성한다.
3. 잔여 수명(RUL) 추정을 위해 지수 성능 저하 모델 적합하기
   1. **생존 모델 생성하기(Survival model)**

생존 모델은 시간에 따른 사건 발생을 모델링 하는 통계적 방법 중 하나이다. 생존 모델은 이벤트 발생과 생존 시간 사이의 관계를 모델링 한다. 이벤트는 주로 시스템 또는 장치의 고장, 결함 또는 시스템 실패를 나타내며, 생존 시간은 해당 이벤트까지의 시간을 의미한다. 일반적으로 생존 모델은 수명 데이터가 주어질 때 사용할 수 있다.

생존 예측 관련 연구는 크게 세가지로 나뉜다.

1. Non-parametric: 데이터의 feature와 생존 시간 분포 정보를 사용하지 않는 방법이다. 분포 정보를 알 수 없을 때 유용하지만 예측이 부정확 할 수 있다.
2. Semi-parametric feature: 정보를 활용하지만, 생존 시간 분포 정보를 사용하지 않는 방법이다. Cox Proportional Hazard 모델이 해당한다
3. Parametric: 생존 시간 분포가 존재한다고 가정하고 회귀 모델로 생존 시간을 예측하는 기법이다. 지수 분포, 베이블 분포, 로지스틱 분포, 정규 분포 등이 있다.

**[Kaplan-Meier 추정 모델]**

Kaplan-Meier은 생존 분석 방법 중 생존과 이탈을 예측하는 가장 단순한 방법이다. Kaplan-Meier 생존 분석은 관찰 시간이 짧을 때부터 긴 순서로 나열해놓고 각 사건이 발생한 시점에서의 사건이 일어날 확률을 계산하는 비모수적 방법이다. 비모수 추정기이기 때문에 데이터 분포에 대한 초기 가정을 할 필요가 없다. 각 시점마다 구간 생존율을 구하여 이들을 누적시킴으로써 생성된 누적 생존 확률을 추정한다.

예를 들어, 지금까지 신규 가입 고객 10억명 중 20%가 한 달 내 이탈하는 양상을 보였다면, 오늘 가입한 천만 명 중 20%는 한 달 내 이탈할 것이라는 단순 예측을 할 수 있다.

**[Cox 비례 위험 모델]**

Kaplan-Meier 모델은 하나의 타겟 특성 이외의 다른 factor들을 통제할 수 없다는 점에서 한계를 가지고 있다. 이를 보완한 것이 Cox 모델이다.

Cox 비례 위험 모델은 공변량과 생존 함수를 결합한 모델이며, 다른 변수 간의 관계를 고려한다. 본 모델은 각 변수의 위험 비율을 계산하여 이벤트 확률에 미치는 영향을 평가한다.

시간 종속 변수에서 어떤 시점 t까지 생존한 사람이 직후에 사망할 조건부확률은 다음과 같이 나타낼 수 있다.

h(t∣X=x)=h 0​(t)exp(x T β)

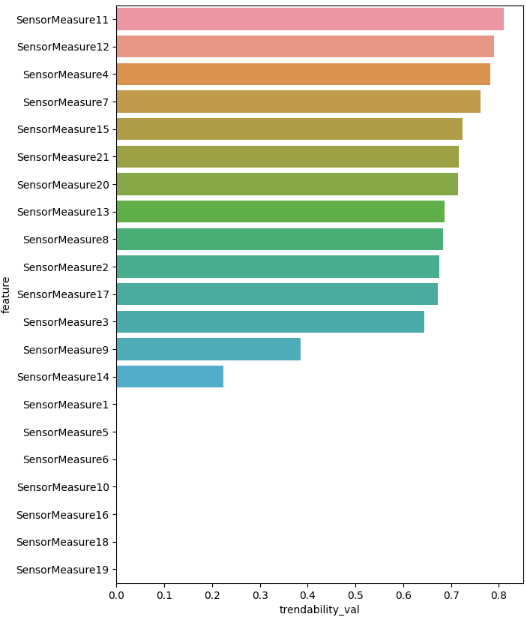
여기서 β는 각 공변량들의 계수로 구성된 벡터이고 h 0​(t)는 공변량들이 0일 때 기저위험함수이다.

위험함수 Cox의 비례위험 모형은 두 가지 가정을 따른다.

1. 위험비가 일정하다는 비례위험가정
2. 관측치는 서로 독립적이며, 공변량이 위험 함수에 선형 곱셈 효과를 갖는 지수함수를 따라 특정 시점의 생존함수가 위험비에 따라 지수함수로 나타난다는 것이다.
3. **실험**
   1. **유사성 모델(Similarity Model)실험**

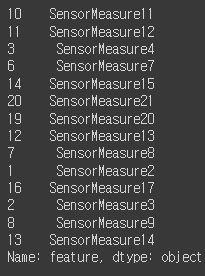
**4.1.1 Feature Selection**

**[각 센서별 추세성 확인]**



각 센서별 수세성을 barplot으로 그려보았다.

이 중 trendability value 값이 0.2보다 큰 값을 선택하였다.

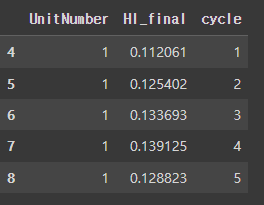


**4.1.2 데이터 전처리**

불필요한 노이즈를 제거하기 위해 이동평균법, 이산 웨이블릿 변환 기법 두 가지 방법을 모두 적용해보았다.

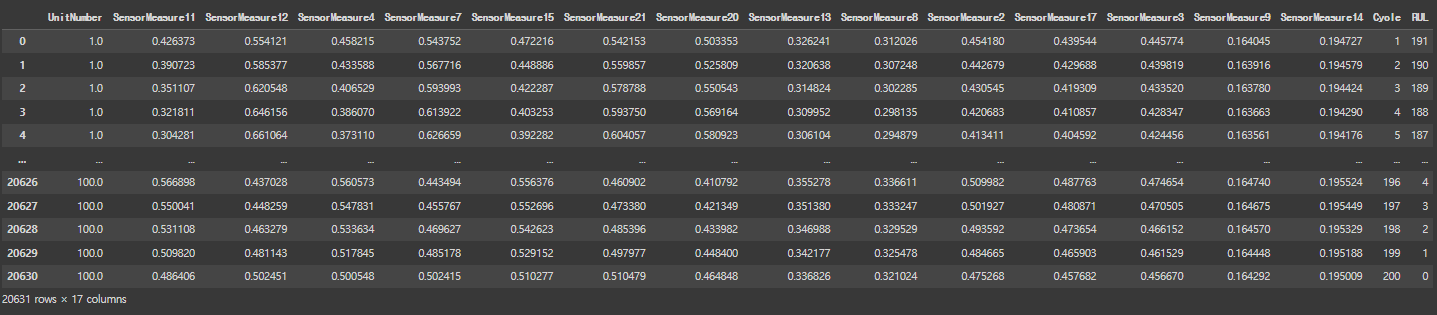
**[이동 평균법을 활용한 데이터 전처리]**

잔여 수명 값(RUL)을 Min Max Scaling을 통해 정규화를 해준 후 “HI”(Hazard Index) 값을 구한 후 그 값을 이동 평균법을 활용하여 데이터 노이즈를 제거하는 과정을 거쳤다. 이후 해당 값을 “HI\_final” 값에 저장하였다.



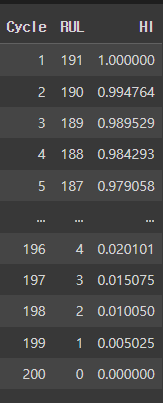
**[DWT를 활용한 데이터 전처리]**

Threshold를 0.63으로 두고 wavelet 종류를 db4로 설정한 후 DWT 신호 처리를 진행하였다.

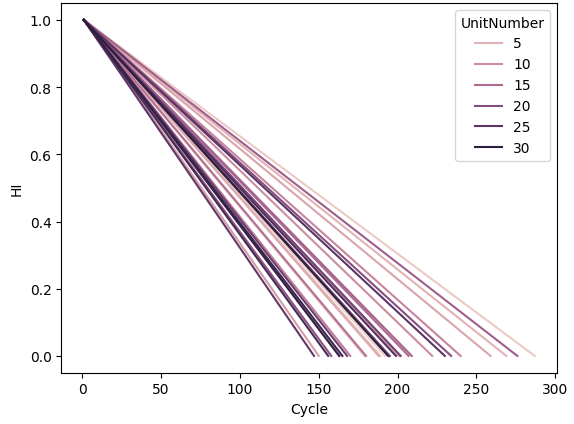


**4.1.3 대표 선형 열화 모델 생성**

잔여 수명 값(RUL)을 Min Max Scaling을 통해 정규화를 해준 후 “HI”(Hazard Index) 칼럼을 생성한 후 값을 넣었다.

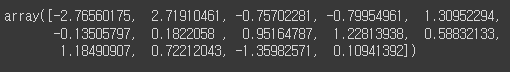


이를 각 기계의 고유 번호에 대해 그래프를 그려보면 다음과 같다.

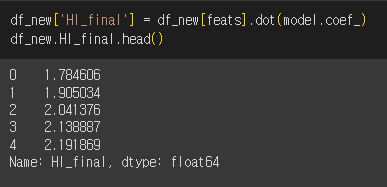


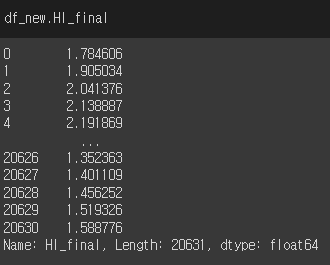
각 기계들의 고유 번호에 대해 lineplot으로 시각화하면 선형적 열화 모델을 따른다는 것을 알 수 있다.

전체 센서 값들을 대표할 수 있는 하나의 열화 모델을 만들기 위해 추출된 특성을 독립변수로 놓고 “Hi” 값을 종속변수로 놓고 선형 회귀를 돌렸다. 해당 회귀 모델의 coefficient 값은 다음과 같다.



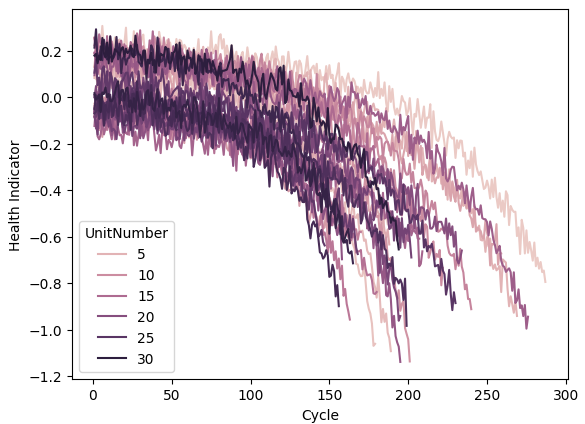
구한 선형 회귀 모델의 coefficient 값을 센서들의 데이터와 내적하여 모든 데이터를 대표할 수 있는 선형 열화 지수 “Hi\_final”을 만든다.

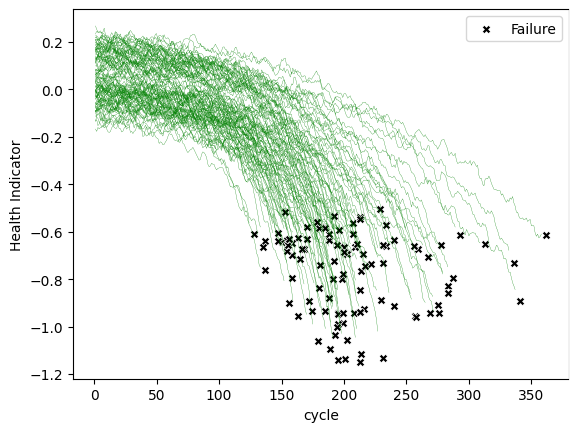




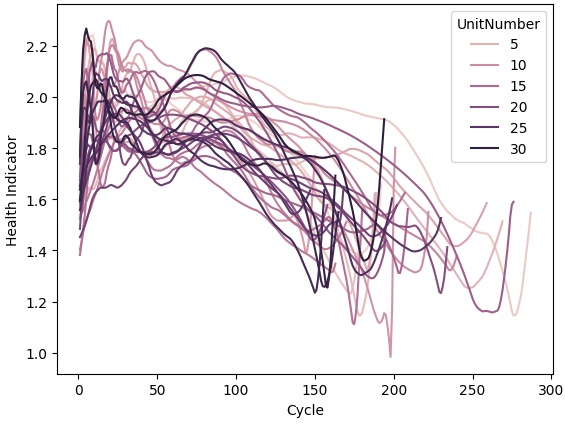
다음은 x 값을 Cycle로 놓고 y 값을 HI\_final로 놓은 후 시각화 한 그래프이다.

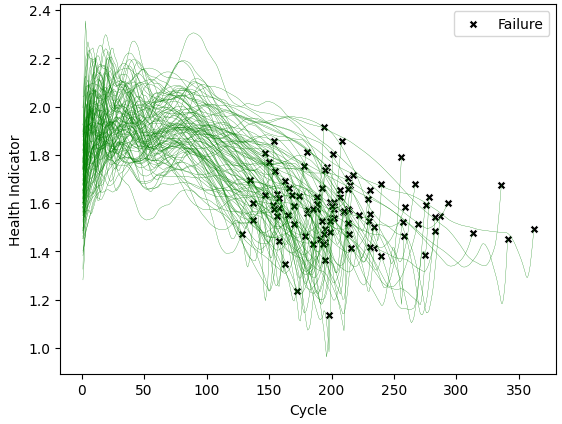
<이동 평균을 활용한 그래프>





<DWT를 활용한 그래프>



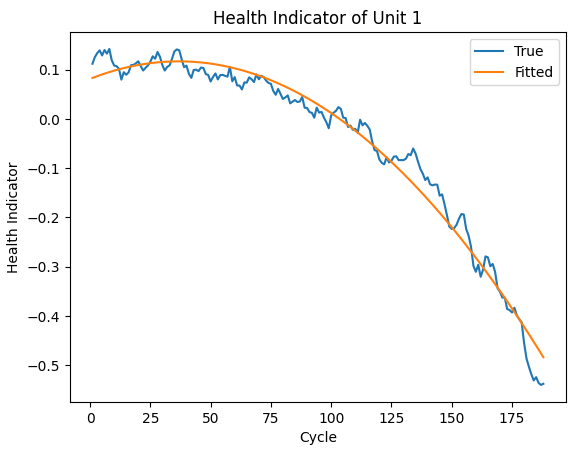


**4.1.4 융합된 데이터 다차방정식에 피팅하기**

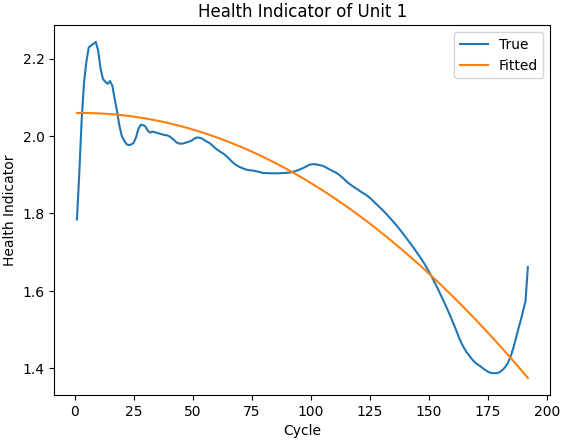
구한 “Hi\_final” 지수를 가지고 다차방정식을 생성한다. 다차 방정식이 wavelet을 통과한 데이터와 적합이 되도록 차수를 맞춰줘야 한다. 전체적인 그래프의 추이를 보고 크게 2가지 차수(이차와 사차)를 선택하여 실험하였다. 이차 방정식을 선택한 이유는 그래프의 전체적인 추이가 점진적으로 감소하기 때문이다. 사차 방정식을 선택한 이유는 그래프가 wavelet을 통과하면서 두 번의 굴곡이 있는 사차 방정식 형태를 띠기 때문이다.

아래 그래프는 각각 기계 고유 번호가 1인 기계의 이차곡선 형태의 열화모델과 사차 곡선 형태의 열화모델의 그래프이다.

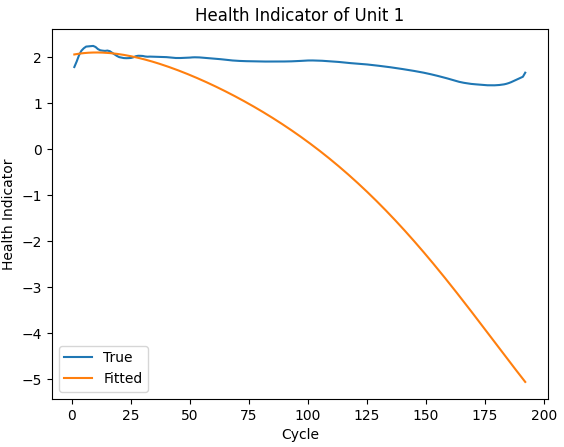
<이동평균법- 2차 곡선 형태>



<DWT- 2차 곡선 형태>

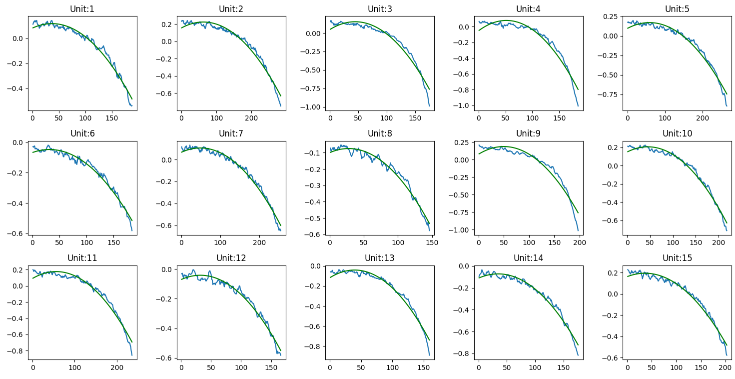


<DWT- 4차 곡선 형태>

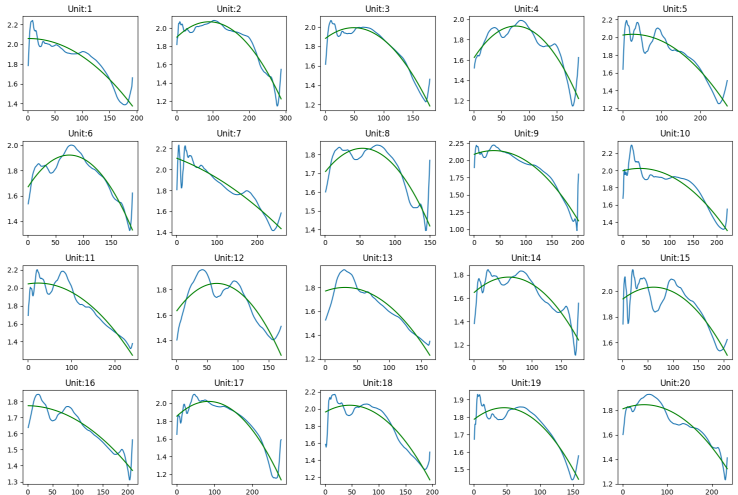


고유번호가 1인 기계만을 비교하면 이차곡선 형태가 더 잘 fitting이 되는 것을 확인 할 수 있다. 다른 모든 기계들에 대해서도 이차곡선이 더 좋은 성능을 보이는 지 확인하였다.

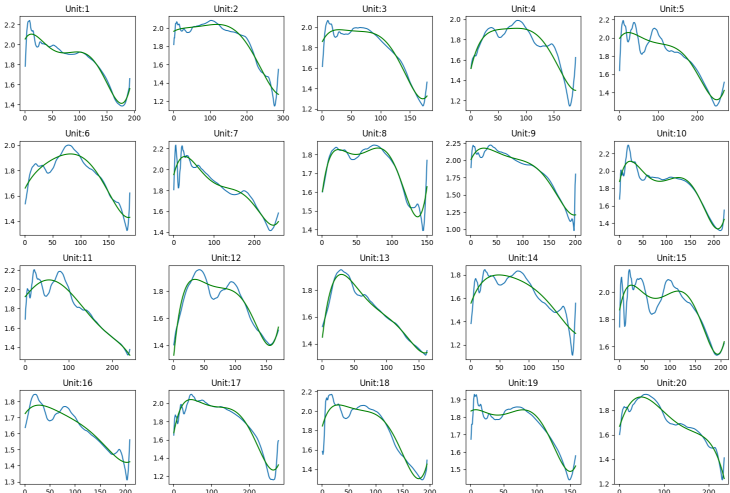
<이동평균법- 2차 곡선 형태>



<DWT- 2차 곡선 형태>



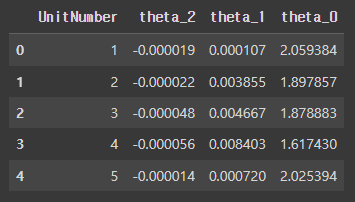
<DWT- 4차 곡선 형태>



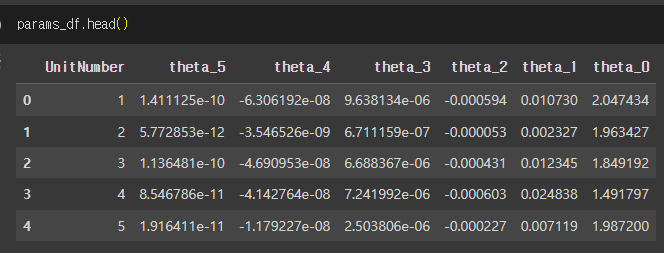
위 그래프만을 보고는 이차 곡선과 사차 곡선 중 어떤 것이 더 좋은 지 구분하기 힘들다. 따라서 생성한 모델의 잔여 수명 예측 값과 실체 잔여 수명과의 오차가 적은 것을 선택하는 방식을 택했다.

이후 각 기계의 고유번호와 각 기계별 파라미터들을 하나의 데이터 프레임 params\_df에 저장하였다. 해당 파라미터들은 주기에 따른 HI의 예측을 계산하는 데 사용된다.

<2차 곡선의 파라미터 값>

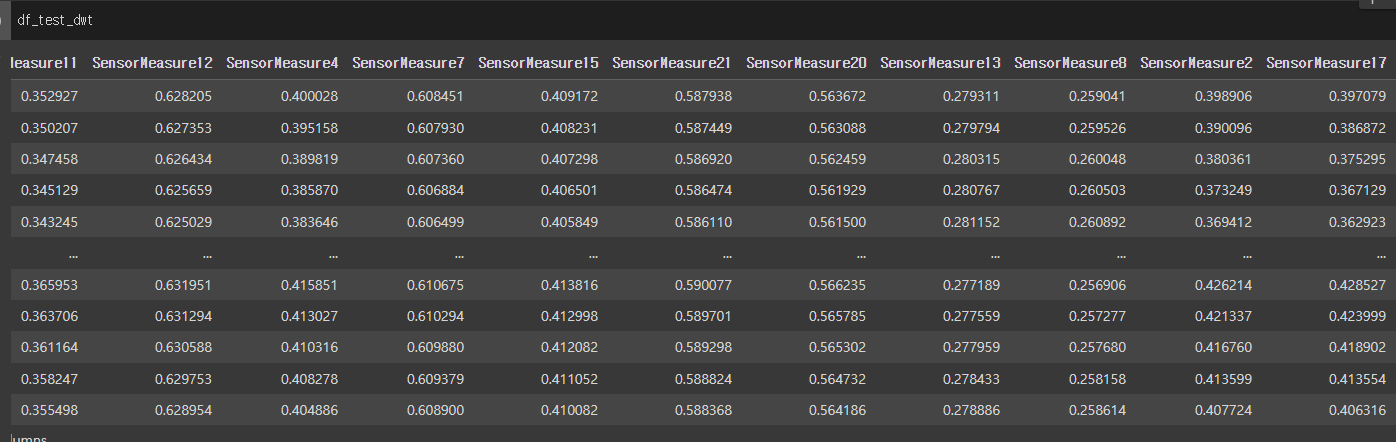


<4차 곡선의 파라미터 값>

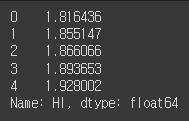


**4.1.5 테스트 데이터 준비**

테스트 데이터 셋에 대해서도 훈련 데이터 셋과 동일하게 dwt를 통해 신호처리를 하였다.



이후 훈련 데이터 셋에서 생성한 계수를 그대로 테스트 데이터 셋에 사용해 “HI”(Hazard Index) 변수를 생성하였다.



4.1.5 유사성 점수 생성하기

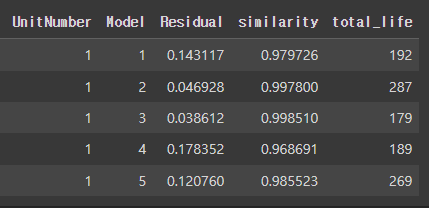
테스트 데이터 셋에서 HI(Hazard Index)를 생성하였으니, 유사성 점수를 계산하여 더 적합한 모델을 찾아내야 한다. 유사성 점수를 구하기 위해서는 우선 각 기계의 HI 예측 값을 구해야 한다.

HI의 예측 값을 구하기 위해선 다음과 같은 과정을 거쳐야 한다.

1. 테스트 데이터 셋의 cycle 값을 구한다
2. 훈련 데이터 셋에서 생성한 모델의 파라미터 값(param\_df의 파라미터)와 cycle(테스트 데이터 셋의 사이클)을 통해 다차방정식을 생성한다.
3. 생성한 다차방정식의 값이 바로 HI의 예측 값, Pred\_HI이다.

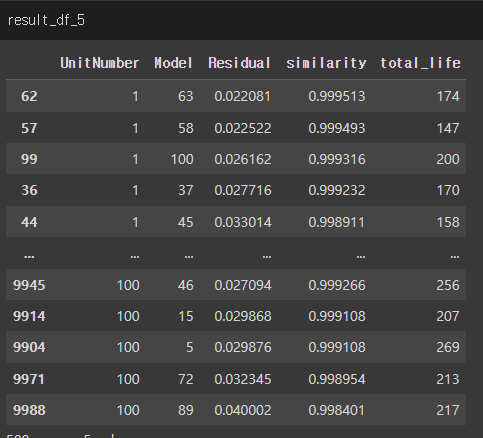
Pred\_HI 값을 구했으면 이를 실제 HI 값과 비교하여 잔차를 구할 수 있다. 잔차는 예측과 관측 값 간의 차이를 의미한다. 구한 잔차를 통해 유사성 점수를 구하는 것은 쉽다. 잔차의 제곱을 exponential 시키면 된다. 잔차가 작을수록 유사성 점수가 높아지게 된다. 즉 모델의 예측이 관측값과 더 가까울수록 유사성 점수가 높다고 할 수 있다.

UnitNumber, Model, Residual, similarity, total\_life 등 구한 값을 df\_test\_fit 데이터에 저장한다.

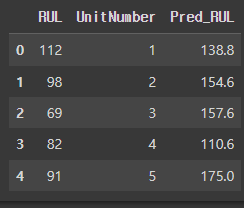


**4.1.5 잔여 수명 예측 값 구하기**

df\_test\_fit 데이터 프레임에서 각 기계(UnitNumber) 마다 유사성 점수가 5번째로 높은 것까지 추출하여 result\_df\_5라는 새로운 데이터 프레임을 만들었다.

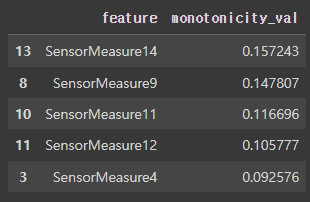


이후 result\_df\_5에서 구한 각 기계 별 total\_life의 평균과 테스트 데이터의 기계 별 사이클의 평균의 차이를 통해 잔여 수명 예측 값을 구한다.

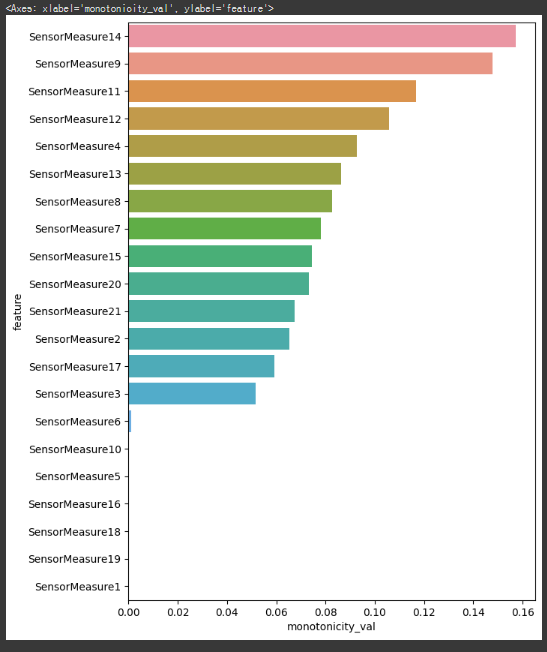


* 1. **건전성 지표 모델(Health Indicator)실험**
     1. **Feature Selection**

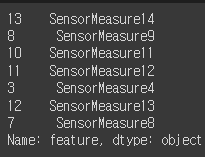
각 센서들의 단조 특성 값을 구할 수 있는 함수를 설정한 뒤 각 센서들의 단조 특성 값을 구하였다.



시각적으로 한 눈에 볼 수 있도록 barplot을 통해 시각화하였다.



단조 특성이 0.8 이상인 것들만 추출하였다.



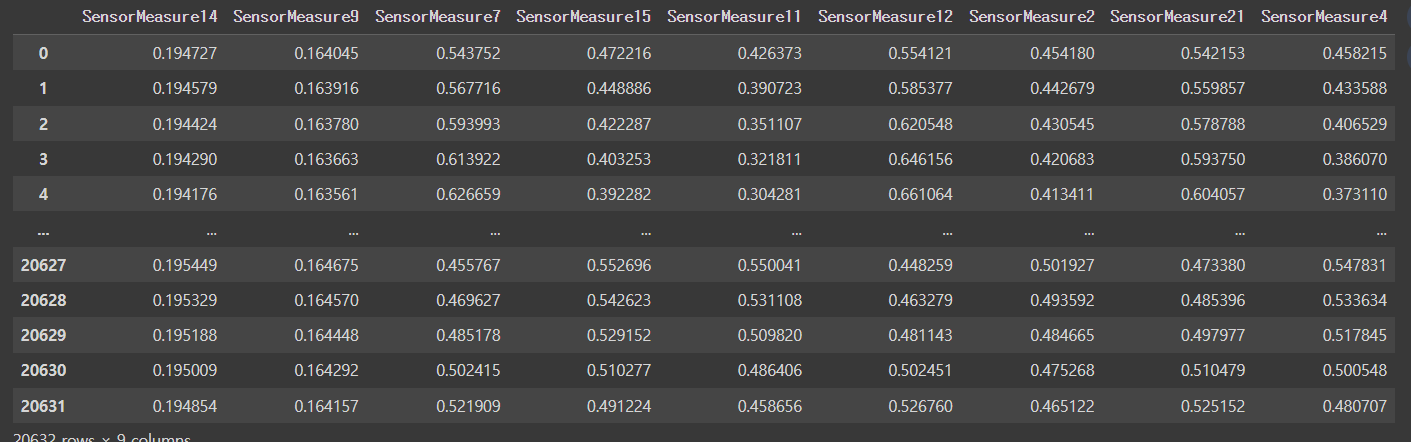
* + 1. **데이터 전처리**

**[이동 평균법을 활용한 데이터 전처리]**

불필요한 노이즈를 제거하기 위해 이동평균법, 이산 웨이블릿 변환 기법 두 가지 방법을 모두 적용해보았다.

**[DWT를 활용한 데이터 전처리]**

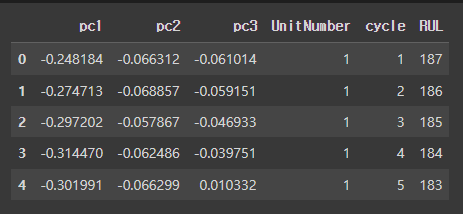
Threshold를 0.63으로 두고 wavelet 종류를 db4로 설정한 후 DWT 신호 처리를 진행하였다.



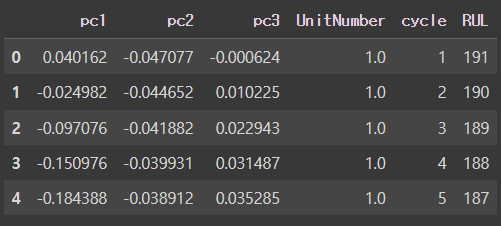
* + 1. **PCA**

주성분 분석을 통해 21개의 센서 값을 대표할 수 있는 주성분을 뽑아낸다. 주성분 분석 결과 pc1이 pc2, pc3에 비해 설명력이 더 높은 것을 확인할 수 있다.

<이동 평균법 pc>

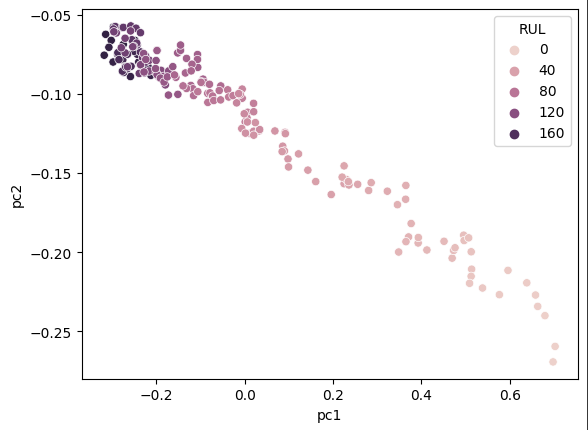


<dwt pc>



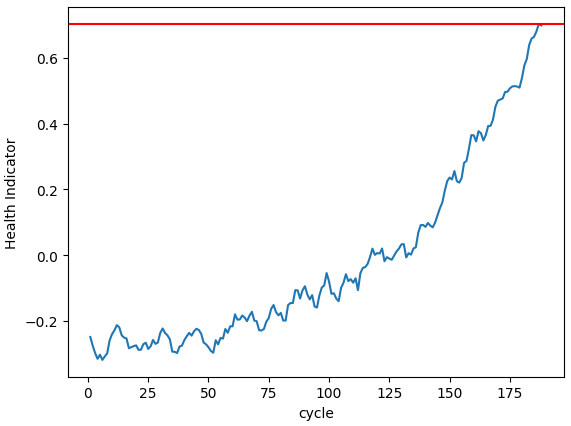
처음 두 개의 주성분으로 구성된 공간에서 데이터를 시각화 하였다.

<이동 평균법 pc>

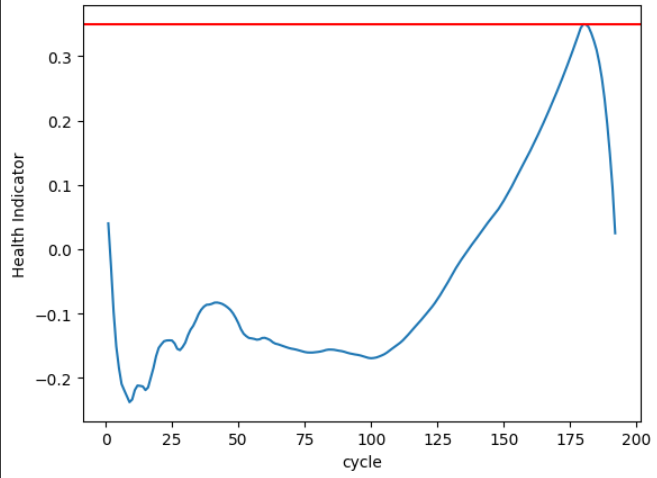


다음은 기계 고유 번호가 1인 기계의 주성분 값을 시간에 따라 나타낸 것이다. 파란색 선은 주성분 값을 나타내고 빨간색 선은 주성분의 최대값을 의미한다. 주성분과 시간이 지수 관계에 있음을 확인할 수 있다.

<이동 평균법>

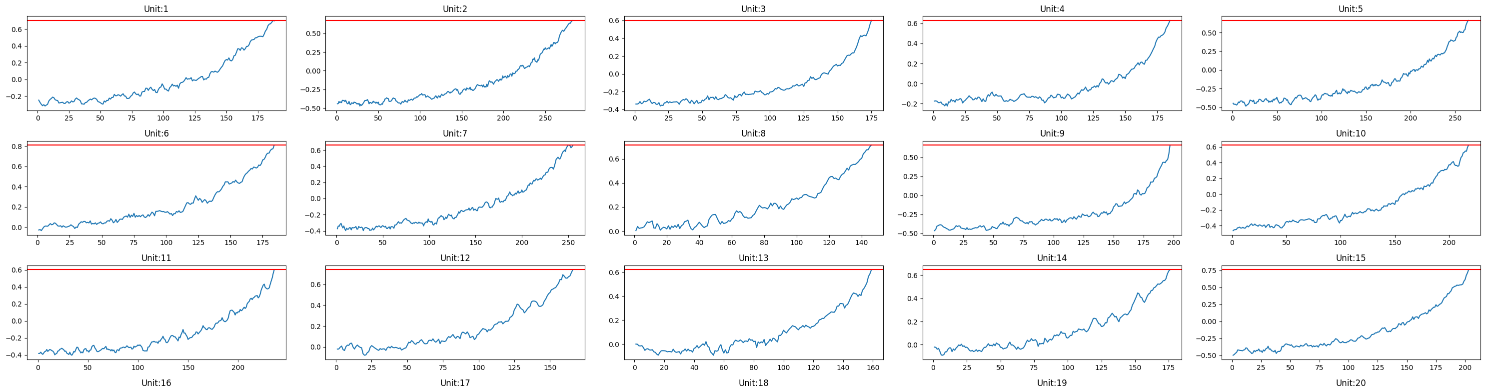


<dwt>

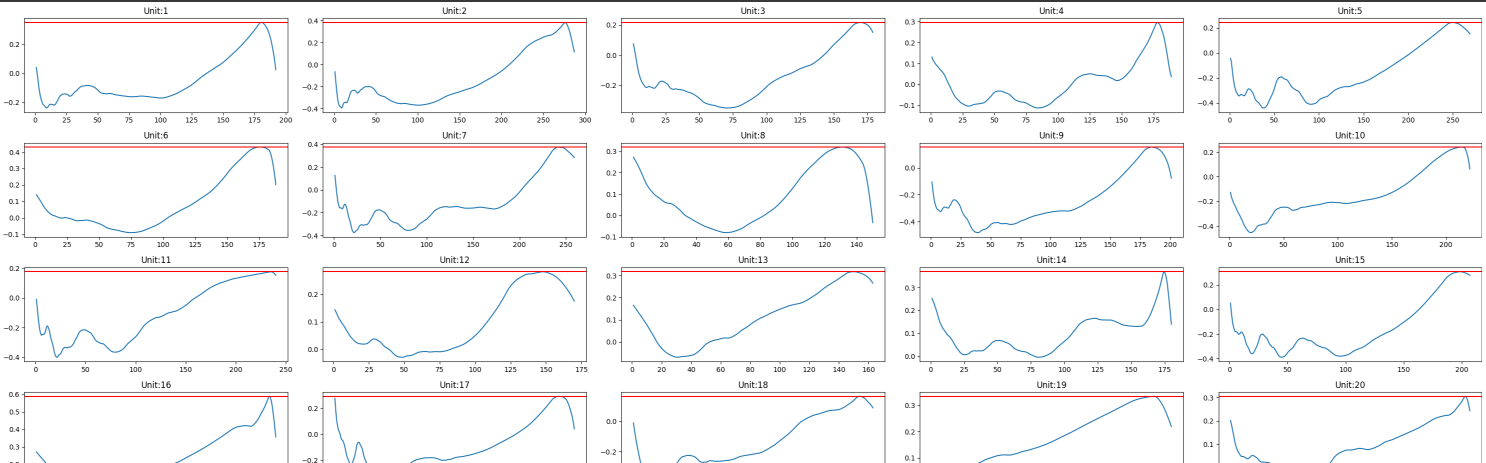


다른 기계 주성분 추이도 확인한 결과 주성분과 시간은 지수관계에 있는 것을 확인할 수 있었다. 따라서 지수적 감소 모델(Exponential degradation model)을 생성할 수 있다.

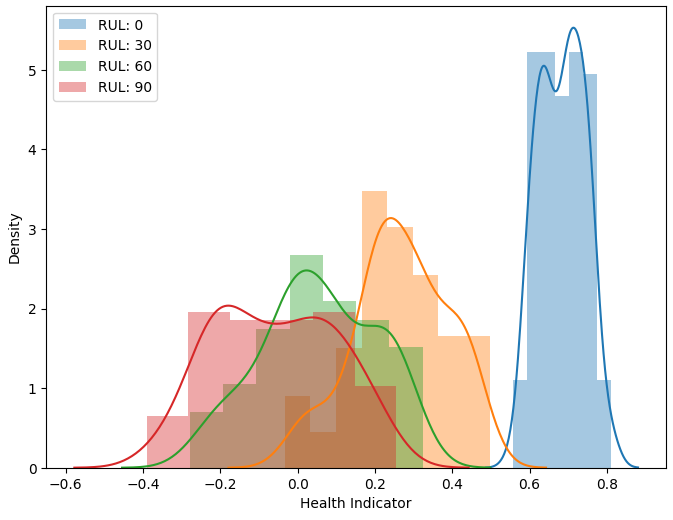
<이동 평균>



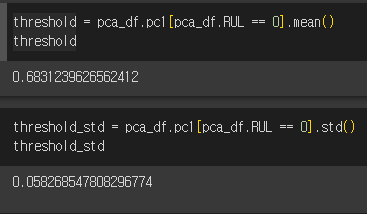
<dwt>



다음은 pc1 값을 주성분으로 했을 때 RUL(잔여 수명) 값을 도식화 한 것이다. RUL 값이 0일 때 확실하게 분류가 되는 것을 확인할 수 있다.

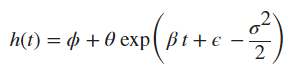


따라서 이를 threshold로 설정하였다.



* + 1. **지수 성능 저하 모델**

지수 성능 저하 모델은 다음과 같이 정의된다.



여기서 h(t)는 건전성 지표로, 시간의 함수이다. *Φ* 는 절편 항으로, 상수로 간주된다. *θ* 와 *β*는 모델의 기울기를 결정하는 임의 파라미터이다. 여기서 *θ*는 로그 정규분포이고 *β*는 가우스 분포이다. 각 시간 스텝 *t*에서 *θ*와 *β*의 분포는 *h*(*t*)의 최선 관측 값을 기준으로 사후 확률로 업데이트 된다. *ϵ* 은 *N*(0,*σ*2) 을 따르는 가우스 백색 잡음이다.

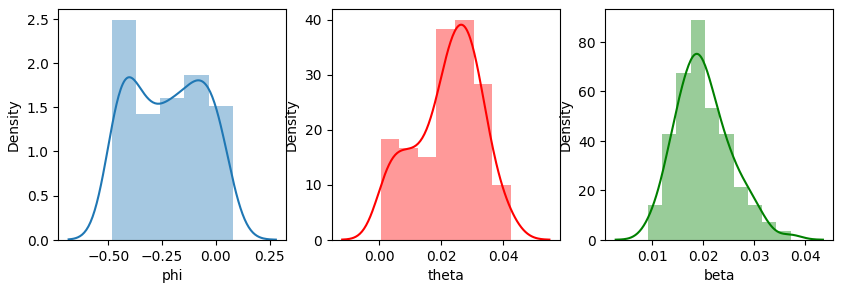
해당 식을 데이터에 적용하기 위해 간소화 시키면 다음과 같은 수식을 세울 수 있다.

*h*(*t*) = *Φ* + *θ\**exp(*β*\*cycle)

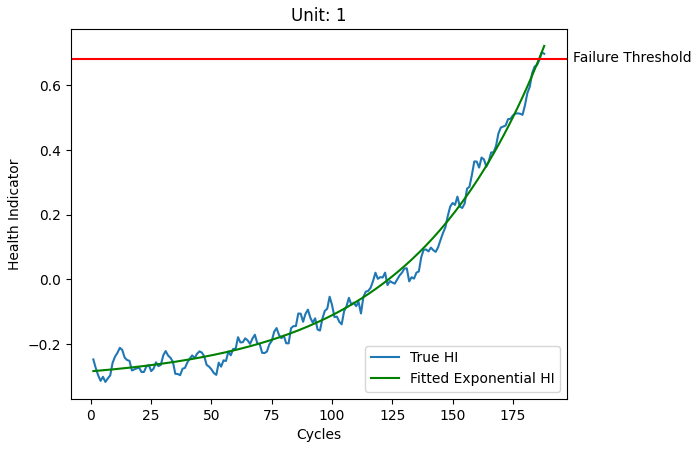
각 기계 별로 주성분을 바탕으로 지수 성능 저하 모델을 돌린 후 나온 파라미터 값을 exp\_params\_df 데이터 프레임에 저장하였다.



나온 파라미터들을 distplot 형태로 시각화하면 다음과 같다.

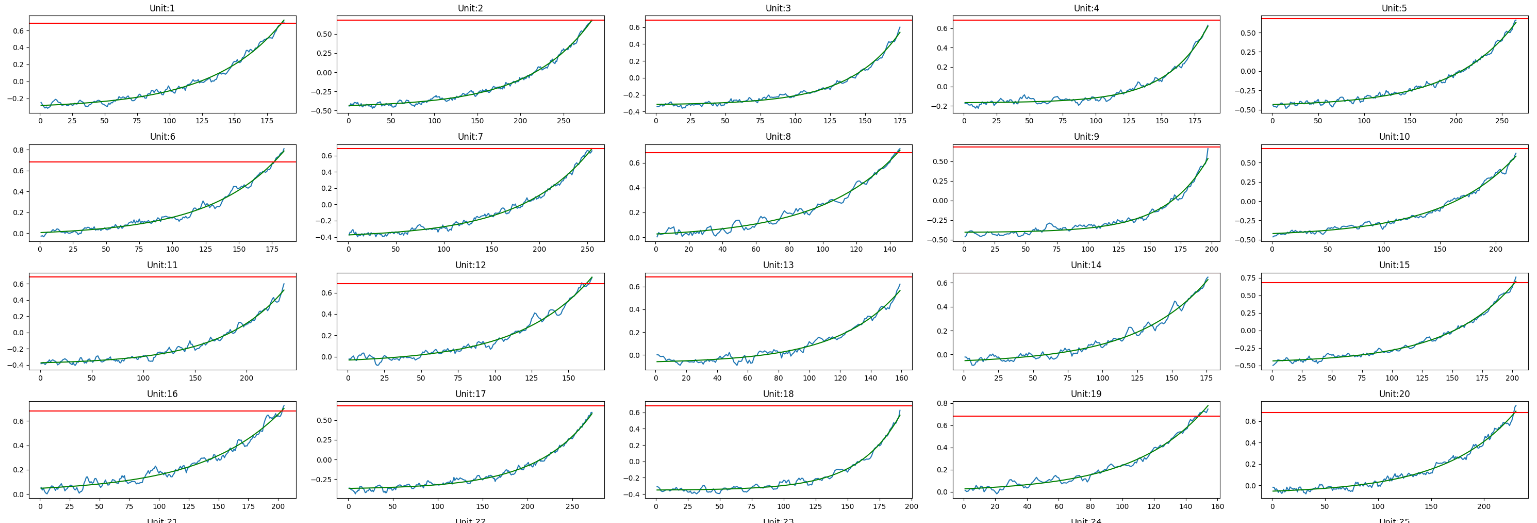


생성한 지수 성능 저하 모델이 잘나오는 지 확인하기 위해 1번 기계의 주성분 값과 지수 성능 저하 모델을 시각화하였다. 시각화 결과, 지수 성능 저하 모델이 잘 피팅이 되는 것을 알 수 있다.

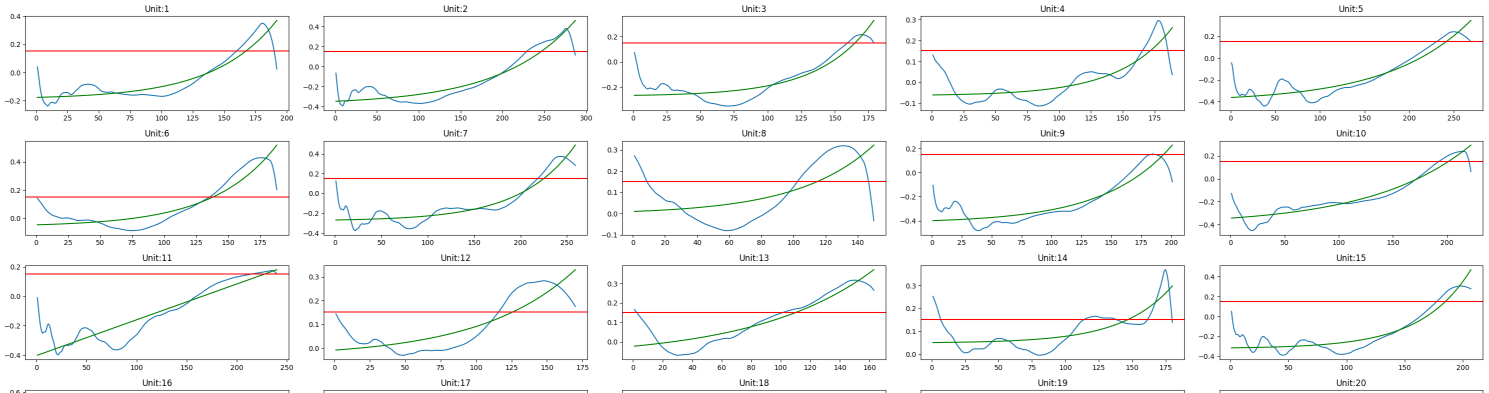


이를 다른 기계에 대해서도 적용하여 그래프 추이를 확인하였다.

<이동평균법>

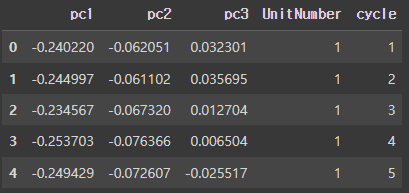


<dwt>

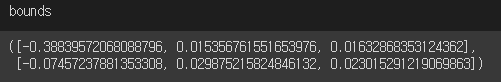


* + 1. **테스트 데이터 셋**

테스트 데이터 셋에 대해서도 PCA를 진행하고 PC1을 건전 지표로 설정하였다.



앞서 지수 성능모델을 만드는 과정을 통해 생성한 파라미터들에 대해 백분위수를 활용하여 경계(bound)를 정의하였다. 하한선을 25%, 상한선을 75%로 설정하여 백분위수 범위를 설정하였다.



**4.2.5 잔여 수명 예측 값 구하기**

앞서 구한 파라미터, bound 값, 테스트 데이터 셋의 건전 지표 값(주성분 값)을 활용하여 잔여 수명 예측 값을 구할 수 있다.

잔여 수명 예측 값을 구하기 위해서는 다음과 같은 과정을 거쳐야 한다.

1. 테스트 데이터 셋의 주성분 값
2. 테스트 데이터 셋의 cycle 값
3. 지수 성능 저하 모델의 파라미터의 경계 값
4. 테스트 데이터 셋의 주성분 값과 cycle 값, 파라미터의 경계값(bound) 값을 활용하여 지수 성능 저하 모델을 생성한다.
5. 생성한 모델을 잔차를 최소화 시키는 방향으로 최적화 시킨 후 최적화 된 파라 미터를 구한다. (phi, theta, beta)
6. 구한 최적의 파라미터를 통해 전체 사이클을 구한다. Total\_cycles = log(threshold-phi)/theta)/beta
7. 전체 사이클과 테스트 데이터 셋의 사이클의 최대 값의 차이가 잔여 수명 예측 값이 된다.



* 1. **생존 모델(Survival Model)실험**
     1. **데이터 전처리**

“breakdown”이라는 새로운 열을 생성하였다. 해당 열은 각 기계가 고장을 일으켰는지 여부를 나타낸다. 각 기계의 마지막 사이클에서 고장을 일으키기 때문에 해당 주기의 데이터 값을 1로 설정한다.

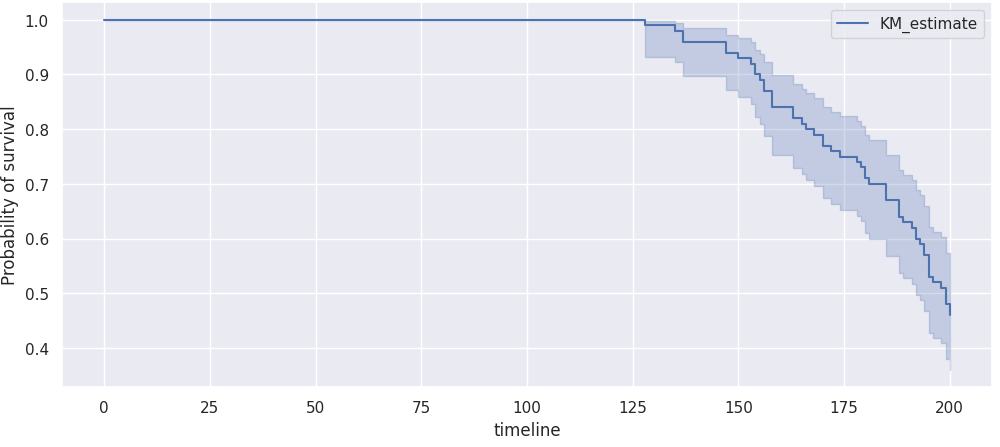
“start”라는 새로운 열을 생성한 후 time\_cycles에서 1을 뺀 값을 저장하였다.

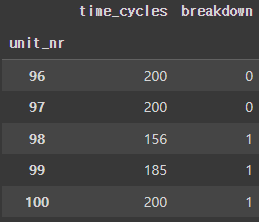
이후 타임사이클이 200개 이하인 데이터만을 추출하였다.



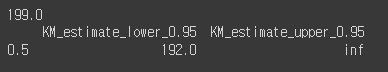
* + 1. **Kaplan-Meier 방법**

Kaplan-Meier 생존 분석을 수행하고 결과를 시각화하였다.



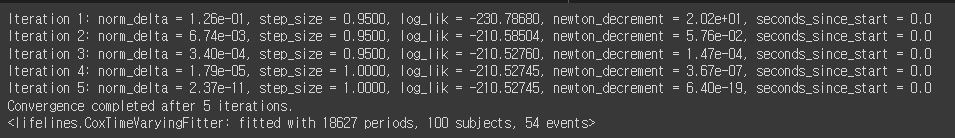


Kaplan-Meier 생존 분석 결과 중앙 생존 시간의 신뢰 구간을 계산하였다.

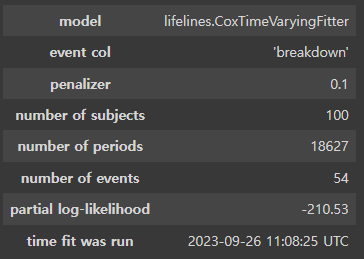


* + 1. **Cox Time-Varying Model 방법**

Cox 시간 변화 모델을 사용하여 데이터 셋의 생존 예측 모델을 구축하였다. Cox 시간 변화 모델은 시간에 따른 위험과 개별적인 요소들 간의 관계를 모델링한다.

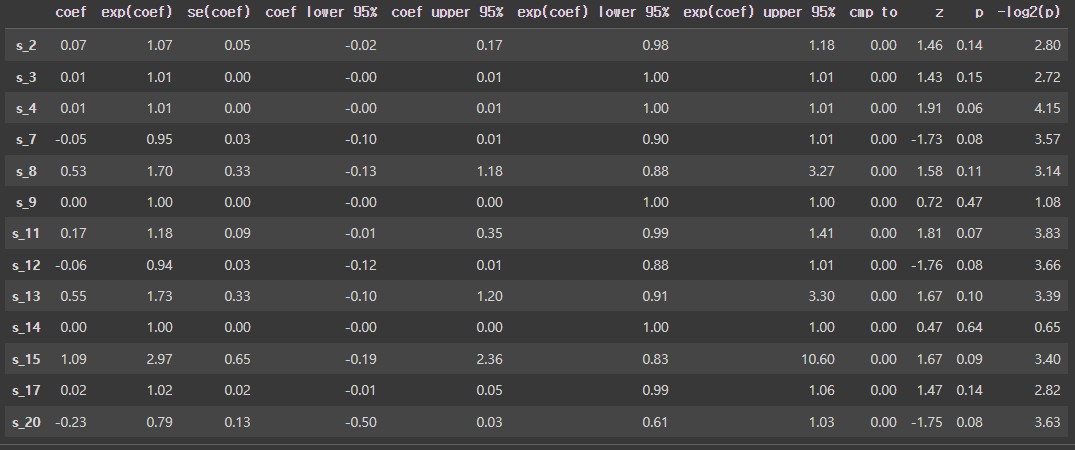


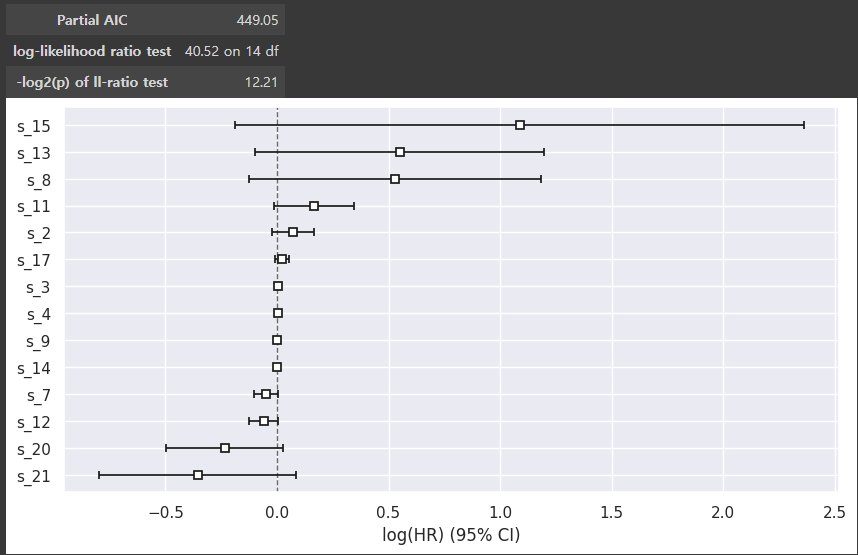
Cox 시간 변화 모델의 요약 정보를 출력하고 모델의 그래프를 시각화하였다. Cox 시간 변화 모델의 요약 정보가 담겨있다. Partial Log-likelihood는 모델의 적합도를 평가하는 중요한 지표 중 하나이다. 값이 높을수록 모델이 데이터를 얼마나 잘 설명하는 지를 나타낸다.



Cox Time Varying model의 요약 정보에 대한 설명:

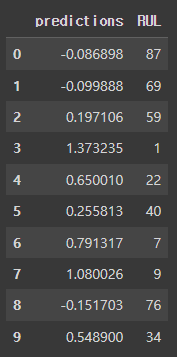
1. Coef: 모델의 각 입력 변수(설명 변수)에 대한 계수(coefficients)를 나타낸다. 이 값들은 해당 변수가 생존 분포에 미치는 영향을 나타낸다.
2. exp(coef): 계수(coef)의 지수값은 위험 비(risk ratio)를 나타낸다. 이 값이 1보다 크면 해당 변수가 위험을 증가시키는 요인으로 작용하며, 작으면 위험을 감소시키는 요인으로 작용한다.
3. z: z-값은 각 계수에 대한 표준화된 값으로, 계수가 표준 오차에 비해 얼마나 큰지를 나타낸다. 큰 z-값은 변수가 중요하다는 것을 의미한다
4. p: p-값은 계수가 통계적으로 유의미한지 여부를 나타낸다. 작은 p-값은 해당 변수가 유의미하다는 것이다.





* + 1. **지수 성능 저하 모델(Exponential Degradation Model)을 활용하여 잔여 수명 예측**

Cox 시간 변화 모델(“ctv”)를 사용하여 각 엔진에 대한 로그 부분 위험 값을 예측한다. 예측된 결과 값을 prediction 데이터 프레임에 “predictions”라는 변수로 저장한다. 이후 실제 RUL 값을 데이터 프레임에 추가한다.



Cox 시간 변화 모델을 사용하여 전체 학습 데이터셋에 대한 로그 부분 위험 값을 예측하고, 이를 데이터프레임에 추가하는 부분을 설명한다. 이렇게 한 후 각 주기(기록)에 대한 예측된 위험 값을 포함하는 새로운 열('hazard')을 생성한다.



Cox 시간 변화 모델에서도 지수 저하 성능 모델이 활용된다. 해당 지수 성능 저하 모델의 식은 다음과 같다.

a \* np.exp(-b \* z)

이때 지수 저하 성능 모델에 사용되는 파라미터 값은 로그 부분 위험 값과 RUL 값이다.

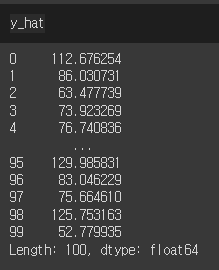
* + 1. Cox Time-Varying Model을 활용하여 잔여 수명 예측 값 구하기

앞서 구한 로그 부분 위험 값(“hazard”)을 사용하여 지수성능 저하 모델을 학습하고 최적의 파라미터 값을 구한다. 이 모델을 사용하면 로그 부분 위험 값에 기반하여 잔여 사용 예측할 수 있다.

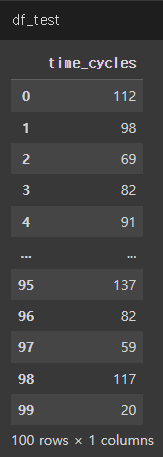
잔여 수명 예측 값을 구하기 위해서는 다음과 같은 과정을 거쳐야 한다.

1. Ctv 함수를 통해 로그 부분 위험 값을 구한다.
2. 구한 로그 부분 위험 값을 지수 저하 모델에 적용한다.
3. 지수 저하 모델로 통해 나온 예측 값과 실제 예측 값의 차이를 구한다

<로그 부분 위험 값을 통해 만든 지수저하 모델의 예측 값>



<실제 잔여수명 값>



1. **결과**

Nasa Turbofan 데이터 셋을 잔여 수명 예측에 사용되는 대표적인 3가지 방법을 사용하여 잔여 수명 예측 값을 구하였고 그 값을 실제 잔여 수명 값과의 차이를 구해 Nasa Turbofan에 가장 좋은 성능을 보이는 모델을 구할 수 있었다.

Nasa Turbofan 데이터 셋 특성 상 각 센서 값의 데이터에 노이즈가 끼어 들어 있기 때문에 적절하게 데이터 전처리를 할 필요가 있다. 본 분석에서는 이동평균법과 이산 웨이블릿 변환 두 가지 방법을 통해 분석을 진행하였다.

데이터 셋 안에 실제 잔여 수명 값이 주어졌고 구현한 3가지 방법 모두 잔여 수명 값을 예측할 수 있기 때문에 예측 값과 실제 값의 차이를 나타내는 지표인 RMSE를 평가 지표로 활용하였다.

다음은 데이터를 전처리 했을 때 이동평균법을 사용했을 시, 유사성 모델, 건전성 지표 모델, 생존 모델의 rmse 값과 dwt를 사용했을 시, 유사성 모델, 건전성 지표 모델, 생존 모델의 rmse 값을 구한 것이다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | 유사성 모델 | 건전 지표 모델 | 생존 모델 |
| 이동 평균법 | Rmse | 23.0331 | 31.838 | 26.58 |
| Mse | 530.528 | 1013 | 706.54 |
| Dwt | Rmse | 48.55 | 60.54 | 45.467 |
| Mse | 2357.482 | 1576.49 | 706.54 |

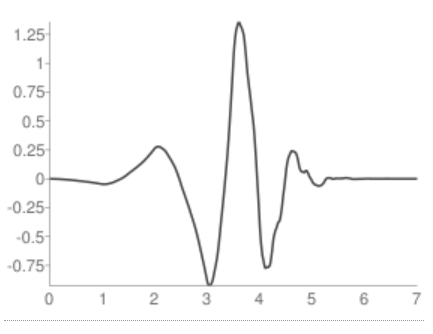
분석 결과 이동 평균법을 통해 데이터를 전처리 한 후 유사성 모델을 하는 것이 가장 성능이 좋았다.

1. **결과에 대한 해석**

**[DWT는 부적합한 데이터 전처리 방법이다]**

분석 결과 전반적으로 데이터 전처리를 DWT로 했을 때 RMSE, MSE 값 모두 이동 평균법에 비해 많이 나오는 것을 확인할 수 있다. 이는 Nasa TurboFan의 데이터의 노이즈가 크지 않고 전체적인 그래프 추이가 우상향하는 방향으로 증가하기 때문이라고 유추할 수 있다.

Dwt를 적용한 3가지 잔여 수명 예측 방법 모두 wavelet 중에 Daubechies 형태의 wavelet을 컨볼류션(convolution)해서 데이터 전처리를 한다. 직관적으로 Nasa Turbofan 엔진의 데이터 값이 눈에 띄게 변동성을 가지지 않기 때문에 보다 부정확한 값이 나왔다고 판단할 수 있다.



<Daubechies wavelet 파형>

**[데이터가 특성에 따른 결과]**

이동 평균법을 통해 데이터를 전처리한 3가지 방법들을 비교해 보면 유사성 모델이 가장 성능이 좋게 나타나고 생존 모델, 건전 지표 모델 순으로 성능이 좋은 것을 알 수 있다.

이는 Nasa Turbofan 데이터 셋의 특성에 기인한 것이다. Nasa Turbofan 데이터 셋을 시각화한 그래프에서도 알 수 있듯이, 데이터 셋에 눈에 띄는 변동성이 없다. 유사성 모델과 생존 모델, 건전 지표의 가장 차이점은 유사성 모델은 통계적인 추정이 사용되지 않는 반면, 다른 모델들은 통계적 추정이 사용된다는 점이다. 생존 모델 같은 경우, 지수 저하 모델에 넣을 파라미터 값인 로그 부분 위험 값을 구하는 과정도 Cox Time-Varying Model을 통해 신뢰도 구간을 설정하여 값을 구한다. 또한 지수 저하 모델 그 자체로도 파라미터에 정규분포의 추정 값인 *β*값이 사용된다. 건전지표 또한 지수 저하 모델을 활용하기 때문에 통계적 추정이 사용된다. 반면 유사성 모델 같은 경우 선형 회귀를 통해 전체적인 센서들을 대표할 수 있는 모델을 구한 후, 해당 계수들을 데이터에 대해 내적하여 값을 구하는 것이기 때문에 잔여 수명 예측 값을 구하는 과정에서 통계적 추정이 사용 되지 않는다.

Nasa Turbofan과 같이 변동성이 적은 데이터 같은 경우, 통계적 추정에 의한 예측보다 실제 값을 구하는 방식(수리형 모델을 사용하는 방식)이 더 좋은 성능을 보인다는 인사이트를 얻을 수 있다.

위 분석 결과를 바탕으로 PSYENCE에 인사이트를 제공하자면, 변동성이 적은 모델 같은 경우, 실제로 값을 구하는 방식이 더 좋은 성능을 보이지만, 실제 현장 데이터들은 대다수 변동성이 높을 가능성이 크기 때문에 생존 모델 및 건전지표 성능 모델을 활용하는 것을 더 권장한다.

1. **출처:**

참고 데이터 셋:

Nasa Turbofan Dataset/FD001

<https://data.nasa.gov/Aerospace/CMAPSS-Jet-Engine-Simulated-Data/ff5v-kuh6>

참고 문헌:

<Matlab>

Mathworks, <https://kr.mathworks.com/help/predmaint/ug/feature-selection-for-remaining-useful-life-prediction.html>

Mathworks, <https://kr.mathworks.com/help/predmaint/ug/feature-selection-for-remaining-useful-life-prediction.html>

Mathworks, <https://kr.mathworks.com/help/predmaint/ug/wind-turbine-high-speed-bearing-prognosis.html>

Mathworks, <https://kr.mathworks.com/company/newsletters/articles/three-ways-to-estimate-remaining-useful-life-for-predictive-maintenance.html>

<참고 논문>

Mohamad Danish Anis, [Towards Remaining Useful Life Prediction in Rotating Machine Fault Prognosis: An Exponential Degradation Model], IEEE

Ayman Mahmoud Ahmed [Prediction of Battery Remaining Useful Life on Board Satellites Using Logical Analysis of Data], IEEE