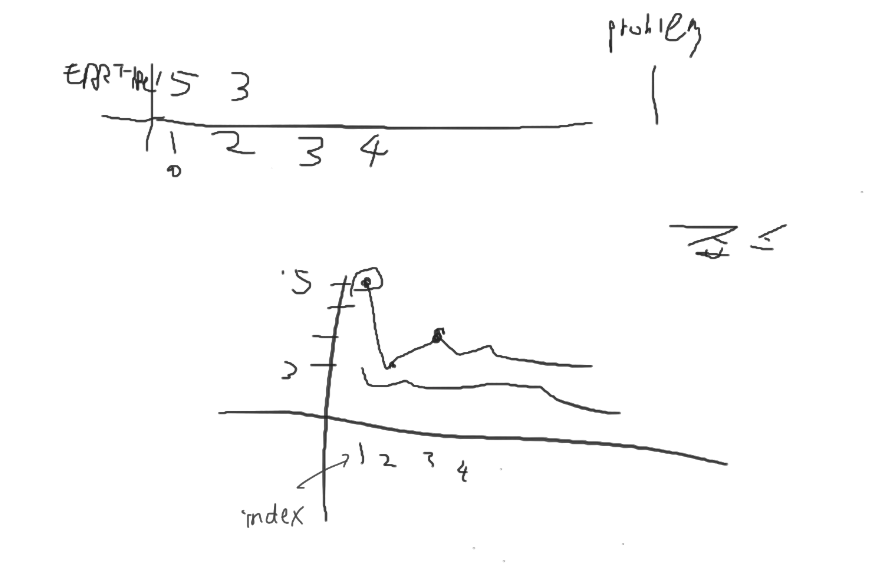
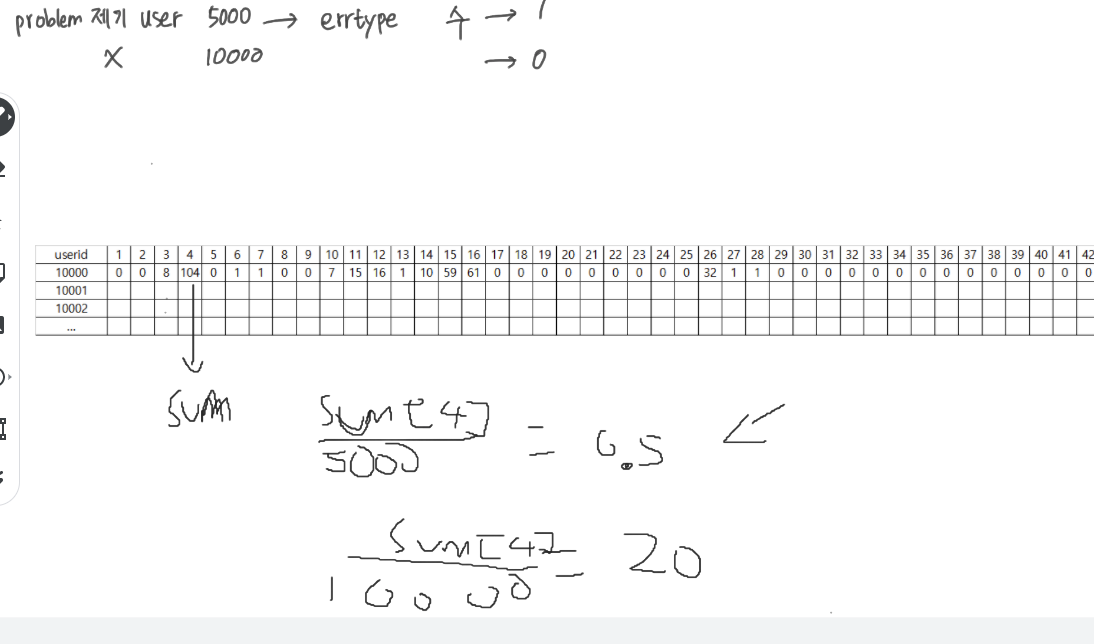
<https://jamboard.google.com/d/1AMeETim-g-4KmIoPFKFhgiT1Yr3eNmzS6yFGMtY-KYQ/viewer?f=7>

성진이 하이

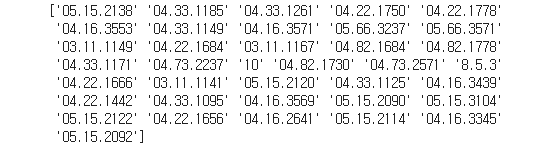
1.err type sequence를 그래프로 나타내기( sequence가 겹치는 지점이 많으면 두꺼워지게 해보기) <https://www.python2.net/questions-540662.htm> (성진)



2. 

fwversion.unique

펌웨어버전 종류



3. 문자열검색으로 일치하는 subsequence를 찾아보면어때

그거를 비교하는거야 problem 안생긴 애들의 subsequence랑

4. 중요한 idea중 하나는 현재 problem이 일어난 시간까지의 err type과 code만 가지고 모델에 사용하는데, problem 일어난 시간 이후에서 발생한 err type 과 code도 모델에 반영해야 auc 올라간다네. (dacom 토론 페이지)

5. 펌웨어 버전이 5라고 할때 4의 버전에서 있었던 에러가 똑같이 남아있을 수도 있다.

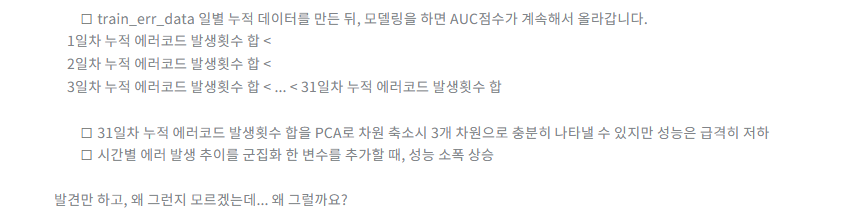
즉 에러가 안 고쳐져있다면 같은 errtype 과 code때문에 문제가 발생할 수 도 있다.

6. 뭐 fwver10에서 10 errtype이 s errcode로 문제를 확정적으로 일으킨다 하면 임의로

이 데이터를 복제해서 모델에 영향이 많이 가도록하는방법

7. 스몰 데이터식 접근부터 해보는게 맞다. 빅데이터 잘 하지도 못하는듯.. ㅜ

8.

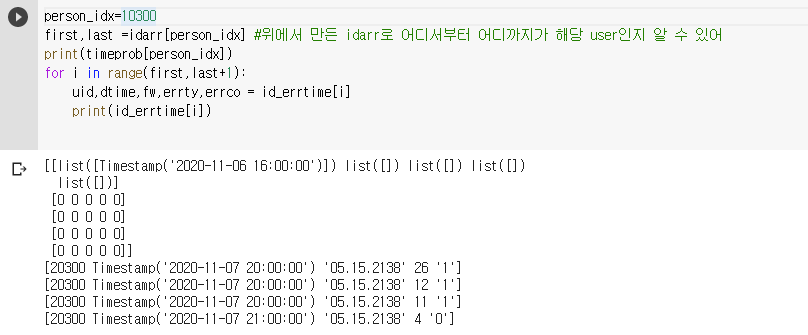


9.

\*데이터 정리\*

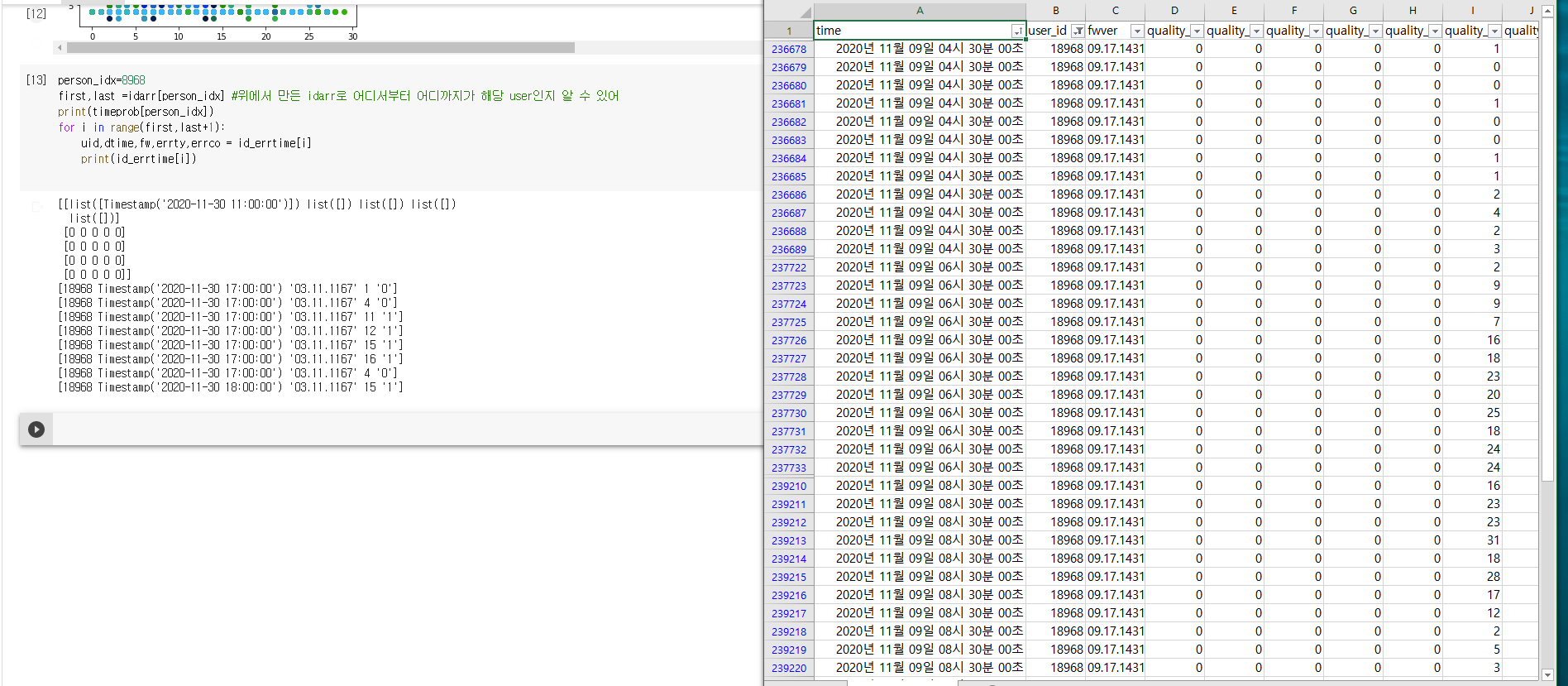
<https://docs.google.com/spreadsheets/d/16ED1yeyZiTEQMQ-A3fiuMk3rvtEINQdqgKUuYisP58E/edit?usp=sharing>

10050 -> 2020-11-13 01시에 문제제기함.

10. 

아 얘는 퀄리티 데이터도없다 11월 6일에 문제제기 해놓고

그전까진 에러타입없고,.

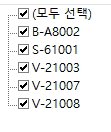
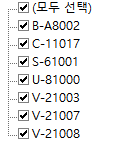
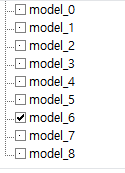


얘 또한 퀄리티데이터없다. 18968. problem 제기하고 err type이 있었지만, quality data의 fwver가 너무달라.

11. 흥미로운 데이터 펌웨어 ‘10’ 에서는 무조건 errtype 5로만 나타나는듯.

또 errcode는 다 다른경우도 있는데, 왼쪽은 문제제기 안한 애들, 오른쪽은 문제제기 한 애들

이들은 모두 모델 6 이며 (펌웨어 10인 애들은) C-11017 과 U-81000은 거의 영향이 없다고 봐도 될듯.



12. 14279 선생님은 ㅋㅋ

11월 13일 11시에 문제를 제기하셨다.

근데 펌웨어버전이

처음에 05.15.2092 쓰시다가 (이 버전에서 14279아저씨가 유일한 데이터) , 05.15.2138 쓰신다. 펌웨어가 바뀐 시점은 11월 12일 23일까지가 05.15.2092 고, 05.15.2138은 11월 13일 00시부터이다. 그리고 새 펌웨어를 쓰신지 11시간만인 11월 13일 11시에 문제제기하셨다. 이분 퀄리티 데이터 존재 다만 11월 24일 16시 10분에 05.15.2138 데이터가 12개 존재한다.

05.15.2138 에러시퀀스는

12, 11, 4, 4, 15, 4, 4, 4, 4, 4, 16, 26, 4, 4, 4, 4, 4, 15, 16, 4, 4, 19, 18, 4, 4

에러코드는

'1', '1', '0', '0', '1', '0', '0', '0', '0', '0', '1', '1', '0', '0', '0', '0', '0', '1', '1', '0', '0', '1', '1', '0', '0'

[[list([Timestamp('2020-11-13 11:00:00')])

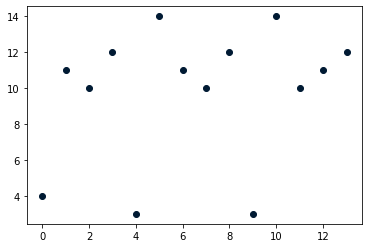
list([Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 22:00:00'), Timestamp('2020-11-12 23:00:00'), Timestamp('2020-11-12 23:00:00'), Timestamp('2020-11-12 23:00:00'), Timestamp('2020-11-12 23:00:00'), Timestamp('2020-11-13 00:00:00'), Timestamp('2020-11-13 00:00:00'), Timestamp('2020-11-13 03:00:00'), Timestamp('2020-11-13 05:00:00'), Timestamp('2020-11-13 05:00:00'), Timestamp('2020-11-13 05:00:00'), Timestamp('2020-11-13 05:00:00'), Timestamp('2020-11-13 05:00:00'), Timestamp('2020-11-13 05:00:00'), Timestamp('2020-11-13 05:00:00'), Timestamp('2020-11-13 05:00:00'), Timestamp('2020-11-13 05:00:00'), Timestamp('2020-11-13 05:00:00'), Timestamp('2020-11-13 05:00:00'), Timestamp('2020-11-13 05:00:00'), Timestamp('2020-11-13 11:00:00'), Timestamp('2020-11-13 11:00:00'), Timestamp('2020-11-13 11:00:00'), Timestamp('2020-11-13 11:00:00'), Timestamp('2020-11-13 11:00:00'), Timestamp('2020-11-13 11:00:00'), Timestamp('2020-11-13 11:00:00'), Timestamp('2020-11-13 11:00:00')])

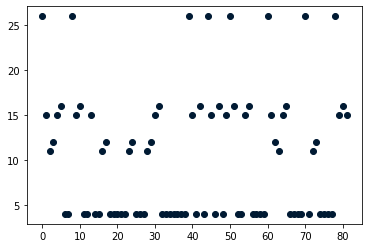
list([4, 4, 11, 12, 5, 4, 4, 3, 13, 12, 27, 28, 14, 7, 6, 14, 4, 4, 12, 11, 4, 4, 15, 4, 4, 4, 4, 4, 16, 26, 4, 4, 4, 4, 4, 15, 16, 4, 4, 19, 18, 4, 4])

list(['0', '0', '1', '1', 'Q-64002', '0', '0', '2', '1', '1', '1', '1', '1', '1', '1', '1', '0', '0', '1', '1', '0', '0', '1', '0', '0', '0', '0', '0', '1', '1', '0', '0', '0', '0', '0', '1', '1', '0', '0', '1', '1', '0', '0'])

list(['05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2092', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138', '05.15.2138'])]

12. fwver == [04.16.3345] 버전에서

이게 문제 안일어난 친구 13726(23726번 이 실제 uid)

이게 문제일어난 친구의 err sequence 13860(23860 이 실제 uid)

2명밖에없다. 문제 발생시간은 Timestamp('2020-11-06 21:00:00'이고

[23860 Timestamp('2020-11-05 20:00:00') '04.16.3345' 26 '1']

[23860 Timestamp('2020-11-05 20:00:00') '04.16.3345' 15 '1']

[23860 Timestamp('2020-11-05 20:00:00') '04.16.3345' 16 '1']

[23860 Timestamp('2020-11-05 21:00:00') '04.16.3345' 15 '1']

[23860 Timestamp('2020-11-07 10:00:00') '04.16.3345' 26 '1']

[23860 Timestamp('2020-11-07 10:00:00') '04.16.3345' 15 '1']

[23860 Timestamp('2020-11-07 10:00:00') '04.16.3345' 16 '1']

[23860 Timestamp('2020-11-07 10:00:00') '04.16.3345' 4 '0']

그 인근의 에러는 이거밖에 없다. 즉 에러발생한지 하루만에 문제를 제기했다.근데 확실히 문제 안일어난 친구에 비해 15번 과 4번에러가 엄청많이발생. ( 그래프 맨아래 점 자주 등장하는게 4번임)

그리고 이 유저는 퀄리티데이터도 존재안해.)

그리고 문제제기 한 이후에 많은 err type과 code를 만들었어.

또 23860의 train\_err.csv 데이터를 보면,

11-12 21시 데이터에 active가 뜬 뒤부터

errtype 23 ‘standby’ 부터 다음 errtype 23 ‘standby’까지 마치 반복되는 모양새

이 왼쪽에 칼라풀한 그래프가 23 ‘standby’ 부터 다음 23’standby’가 나타날때까지의 표야

잘 보면 나타나는 순서가 조금 다를 뿐 반복되는 errtype들이 존재한다.

40 31 15 16의 등장횟수정도만 차이나고 다른건 다 비슷해.

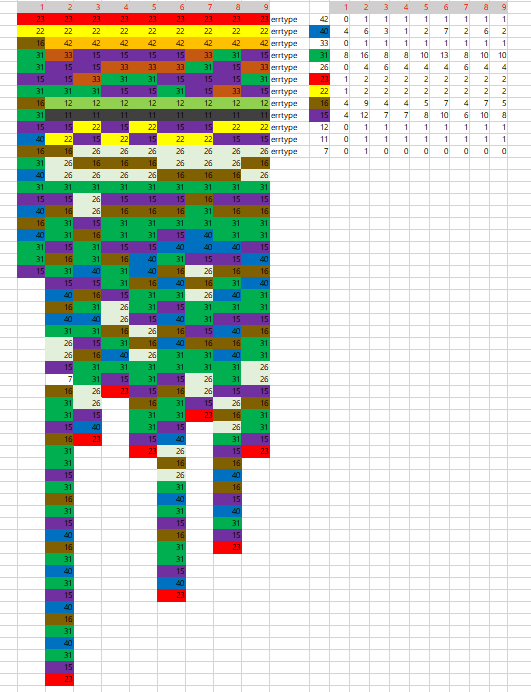
물론 이거는 문제제기 시간보다 뒤의 이야기야. 즉 이거는 정상 패턴인가봐. (문제제기를 안해도 될만한 패턴)

한 user의 error pattern이 길다면 주로 23번 error type때문인듯

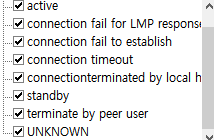
민수가 9번에서 보여준 10050에서도

connection timeout이 계속 발생하다가 (23번 errtype)

[AI경진대회.xlsx](https://o365skku-my.sharepoint.com/:x:/g/personal/bsj805_o365_skku_edu/EW0kVkKpJSlCvDTX3J8KZRABmh2fMEgiFCfHKgDF4JDaRw?e=P3CQpk) 여기 해놨음



14. err type 23은 ‘connection timeout’ 이거나 ‘terminate by user’, ‘active’, ‘standby’ ,



15. 10050 은 11월 13일에 01시에 문제를 제기한다.

23 errtype때문에 err type이 계속 반복된다.

connection timeout이 주로 발생.

<https://drive.google.com/file/d/1wFnMLJwSPGovuCVolIFfYLo7ylIEEY0_/view?usp=sharing>

이거 참고.

16. fwver = 05.15.2114는 문제발생 없음. 23476 한명뿐

17. fwver=04.16.2641 는 문제 일으킨 애만 존재. 21965친구인데, 문제발생시각은 20-11-09 19시야. quality data존재 x

[21965 Timestamp('2020-11-09 05:00:00') '04.16.2641' 3 '2']

[21965 Timestamp('2020-11-09 05:00:00') '04.16.2641' 7 '1']

[21965 Timestamp('2020-11-09 05:00:00') '04.16.2641' 6 '1']

[21965 Timestamp('2020-11-09 05:00:00') '04.16.2641' 14 '1']

[21965 Timestamp('2020-11-09 05:00:00') '04.16.2641' 14 '1']

[21965 Timestamp('2020-11-09 16:00:00') '04.16.2641' 26 '1']

[21965 Timestamp('2020-11-09 16:00:00') '04.16.2641' 16 '1']

[21965 Timestamp('2020-11-09 16:00:00') '04.16.2641' 4 '0']

[21965 Timestamp('2020-11-09 16:00:00') '04.16.2641' 4 '0']

[21965 Timestamp('2020-11-09 16:00:00') '04.16.2641' 11 '1']

[21965 Timestamp('2020-11-09 16:00:00') '04.16.2641' 12 '1']

[21965 Timestamp('2020-11-10 00:00:00') '04.16.2641' 26 '1']

[21965 Timestamp('2020-11-10 00:00:00') '04.16.2641' 4 '0']

[21965 Timestamp('2020-11-10 00:00:00') '04.16.2641' 15 '1']

[21965 Timestamp('2020-11-10 00:00:00') '04.16.2641' 4 '0']

[21965 Timestamp('2020-11-10 00:00:00') '04.16.2641' 22 '1']

이 <https://drive.google.com/file/d/1-G1jqm1OAWwkB_vM32zG1y8rekHlV3HN/view?usp=sharing>

18. fwver=04.22.1656

<https://drive.google.com/file/d/1-MHc5FcvBZ1w8Td4JHKdHZj5hnNlnXXM/view?usp=sharing>

이것도 19831 한명,

qualitydata가 존재하는데 11월 13일 18시 50분 (quality data기준) 04.22.1656이고

11월 19일 14시 50분 (04.22.1750)

problem이발생한 시간은 11월 23일 11시.

여기도 errcode 23 connection timeout 때매 데이터양이 엄청 길다.

|  |  |  |  |
| --- | --- | --- | --- |
| 2020-11-13 22:00 | 04.22.1656 | 16 | 1 |
| 2020-11-13 22:00 | 04.22.1656 | 4 | 0 |
| 2020-11-14 3:00 | 04.22.1656 | 15 | 1 |
| 2020-11-14 3:00 | 04.22.1750 | 36 | 8 |
| 2020-11-14 3:00 | 04.22.1750 | 33 | 3 |
| 2020-11-14 3:00 | 04.22.1750 | 15 | 1 |

여기도 펌웨어 업데이트가 밤중에 일어났네

|  |  |  |
| --- | --- | --- |
| 04.22.1778 | 5 | B-A8002 |

이런게 쭉일어나고 ( 몇백개정도)

|  |  |  |
| --- | --- | --- |
| 2020-11-22 22:00 | 04.22.1778 | 31 |
| 2020-11-23 1:00 | 04.22.1778 | 15 |
| 2020-11-23 7:00 | 04.22.1778 | 16 |
| 2020-11-23 7:00 | 04.22.1778 | 31 |
| 2020-11-23 7:00 | 04.22.1778 | 31 |
| 2020-11-23 8:00 | 04.22.1778 | 40 |
| 2020-11-23 8:00 | 04.22.1778 | 40 |
| 2020-11-23 8:00 | 04.22.1778 | 40 |
| 2020-11-23 13:00 | 04.22.1778 | 31 |
| 2020-11-23 13:00 | 04.22.1778 | 31 |

15 16 31 40 sequence에서 문제를 제기한다? 어라 펌웨어가 한번 더바뀌네

아 얘는 04.22.1750일때 계속 connection timeout이 났었어. 그러다가

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2020-11-20 4:00 | 04.22.1750 | 23 | connection timeout | |
| 2020-11-20 4:00 | 04.22.1750 | 22 | 1 |  |
| 2020-11-20 4:00 | 04.22.1750 | 23 | connection timeout | |
| 2020-11-20 4:00 | 04.22.1750 | 22 | 1 |  |
| 2020-11-20 4:00 | 04.22.1750 | 15 | 1 |  |
| 2020-11-20 4:00 | 04.22.1778 | 15 | 1 |  |
| 2020-11-20 4:00 | 04.22.1778 | 33 | 3 |  |
| 2020-11-20 4:00 | 04.22.1778 | 36 | 8 |  |

이렇게 펌웨어를 바꾸게되었다.

19. fwver=05.15.2122 는 문제없는사람만 한명. 19278 유저분인데, 23같은것도없고

유일한 펌웨어버전 사용자. 업그레이드도, 디그레이드도 안하셨네

퀄리티데이터존재.

에러데이터 : 11월 1일 2일 10시 15시 , 20일 13시, 16시 23일 10시 12시

퀄리티데이터: 11월 2일 12시, 14시, 11월 20일 13시 15시 30분, 11월 23일 10시 30분

<https://drive.google.com/file/d/1-YBgG9hEGSORw7v3gdGruX-9IIL8Pwl9/view?usp=sharing>

20.

<https://drive.google.com/file/d/1Oh_wl0UsRurYFgMPglU8nuY8q3t0gRI3/view?usp=sharing>

problem 발생한 user들 errtype별 계산인데, 작은 그룹별로 보니까 딱히 큰 의미가 있을 지는 모르겠어

<https://drive.google.com/file/d/1OpPZIcQD_gJ3_huzUNcdfPCyTx9VAs_h/view?usp=sharing>

//이건 problem 제기 안한 user들 errtype 별 계산

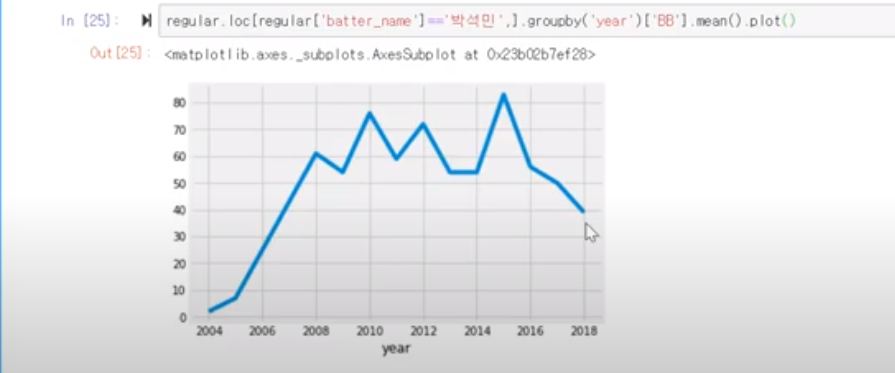
21. neural network는 작은 데이터셋에 대해서는 큰 성능을 발휘하기 힘듦

특정 값에 대해서는 가중치를 다르게 줄 수 있다.

기존에 5개의 데이터가 있는것에는 1 20개의 데이터가 있는 것에 1 의 가중치로 학습을 시켰다면,

5개의 데이터가 있는 것에는 1 20개의 데이터엔 4의 가중치로 학습을 시킬 수 있어.

22



csv 파일을 regular에 읽어온 상태. groupby command로 

이런데이터를 얻을 수 있다. 연도별 BB의 개수 where ‘batter\_name’이 박석민인 볼넷.

뒤에 .plot()을 붙이면 그래프.

두 지표사이의 correlation을 알 수 있는



pd.series.corr 이는, x와 y의 관계성이 높다 면 correlation이 높다.(y=x에 가깝다)

y=x와 점들이 멀다면 correlation이 낮다.

이를테면 뭐 경기수가 많을수록 correlation이 많다 이런 지표를 찾아낼 수ㅡ 있겠지

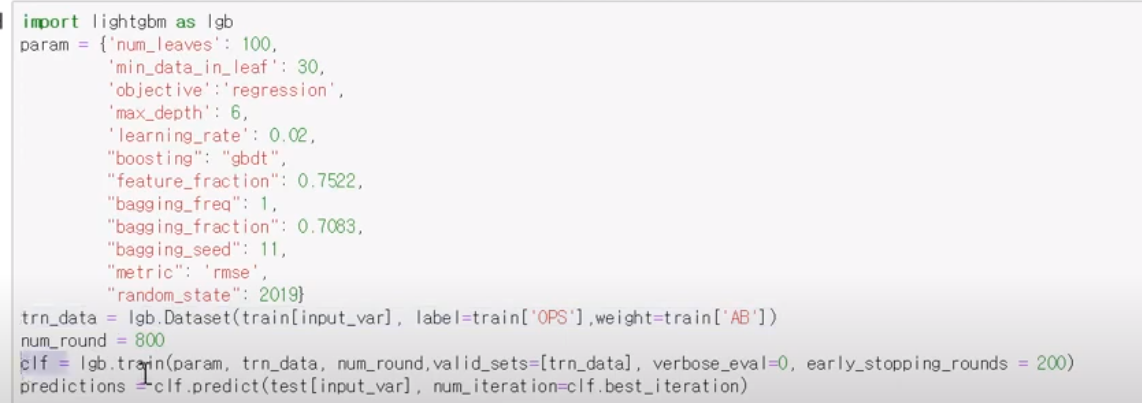
변동성이 크다면, 전년도의걸로 예측을 하기 힘들다는거지 (이전의 경험들로 다음해거를 추측하기힘들다)

그래서 데이터를 조작해서 correlation이 높은데이터를찾자

24. 컴공톡방에 올린 유튜브링크를 보면 알듯이

변동성이 높으면 (correlation이 낮으면) total값으로 , 변동성이 낮다면 (correlation이 높다면) 개개인의값으로 하는것이 맞다.

25. weight를 주는법



이런식으로 lgb인자로넣어주는거

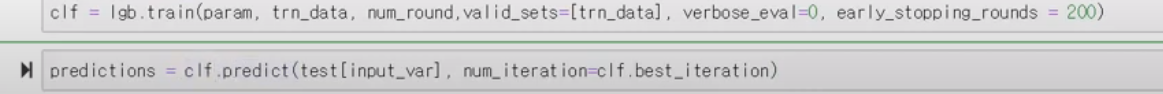
이렇게하면 weight rmse 방식의 training이 가능



num round가 epoch의 역할

valid\_set에 넣어둔 데이터는

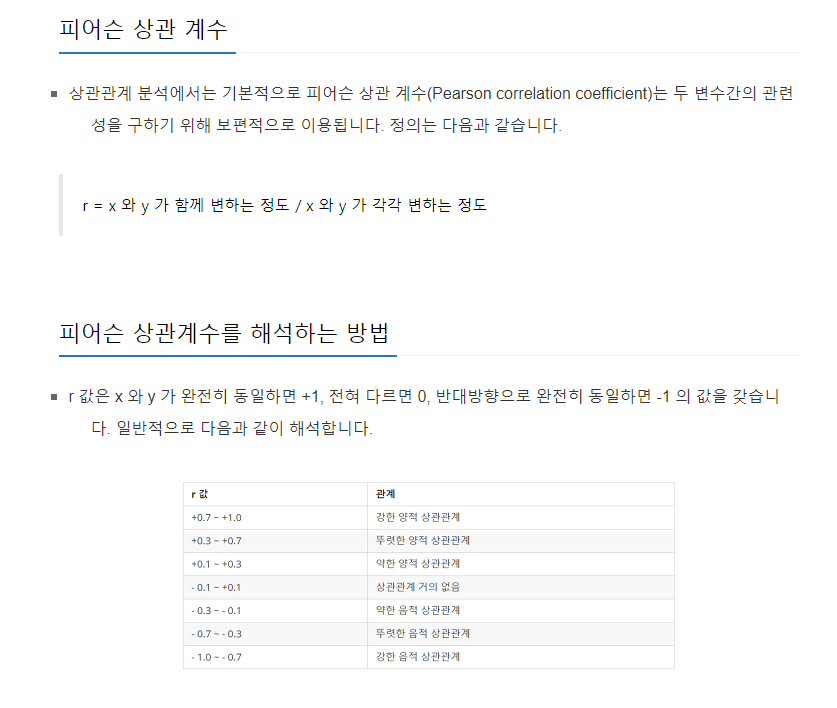
label과 비교되면서 prediction이 얼마나 맞는지.



아래처럼 predict가능하다.

일단 correlation을 찾는 노가다를 해볼만 하다.

<https://gomguard.tistory.com/173>



fwversion과 errtype 개수에 대한 correlation

fwversion과 각 quality 1~9까지의 correlation

‘fwver list’ : ‘ err type 1’ 사이 correlation\

‘fwver list’ : ‘ err type 2’ 사이 correlation\

‘fwver list’ : ‘ err type 3’ 사이 correlation\

…~42 까지

각 에러 타입별 problem에 대한 correlation

만약 해당 fwver 과 특정 err type이 0.7의 correlation가지면 어떻게 하지?

일단 에러 타입별 problem에 대한 correlation이 높다고 하면 적용하지만

상관계수 낮으면 모델링 할 때 제거하자.

이상치들이 존재하는 quality data는 정규화 과정.

상관관계에 대한 이야기는

<https://dacon.io/competitions/official/235584/codeshare/1030>

와 같이 참고.

이 상관계수를 기반으로 가장 중요시되는 feature을 도출해내자.

cntmy 배열 이랑

각 errtype list를대상으로하자. 배열은 error[userid,err(0~41)] 로 되어있으면서

errtype 간 상관관계를 알아보면

이를테면 에러 7번이 일어나면 에러 6번이 같이 일어나구

에러 11번이 일어나면 에러 12번이 같이 일어나는둥

<https://drive.google.com/file/d/1Y13hrrv_hhgFbL_VwkQtVfkTIAFC4zkB/view?usp=sharing>

[correlationerrty.xlsx](https://o365skku-my.sharepoint.com/:x:/g/personal/bsj805_o365_skku_edu/EXzLJwTWZBlChdSihTc7vxgBmmWZyWW3kvz0wZfStlir1w?e=dVjNgJ) (이게 정리된데이터)

0.7 이상의 상관관계를 가지는애들은

3번에러 : 11번, 12번,

6번에러: 7번

7번에러: 35번

11번에러: 12번

15번에러: 16번

18번에러:20번

19번에러:20번,21번

22번:23번

27번:28번(0.9999956확률로 같이 많아진다. y=x에 근접해진다. )

36번:37번 (0.9909962 확률로 같이 많아진다(비례) y=x에 근접해진다.)

음의 상관관계는없네

25.

problem을 제보한 횟수와 err type간 correlation이 제일 높은 것은 (앞에서부터 1번 errty)

-0.057792543 1.14E-05 0.005591056 -0.002141409 0.068162937 0.059969893 0.063984521 -0.013510488 0.013043991 0.015116605 0.033432359 0.036495401 0.018790269 0.025472354 0.200090176 0.202197485 0.078923282 0.261409976 0.120706637 0.289663658 0.114086119 0.087424106 0.085389468 0.017345805 0.089977852 0.13736673 0.028394467 0.028292558 0.157090665 0.085602973 0.053471193 0.117993922 0.051199596 0.074663322 0.172885792 0.173606024 0.048431302 0.035841395 0.141522467 0.078316045 0.017754968

0.28로 20 번

0.20으로 15 16 18 정도이다.

26. quality 3, quality4 은 0만나오니까 제외시켜

quality 0 이랑 2 ,5 은 NAN이있네.

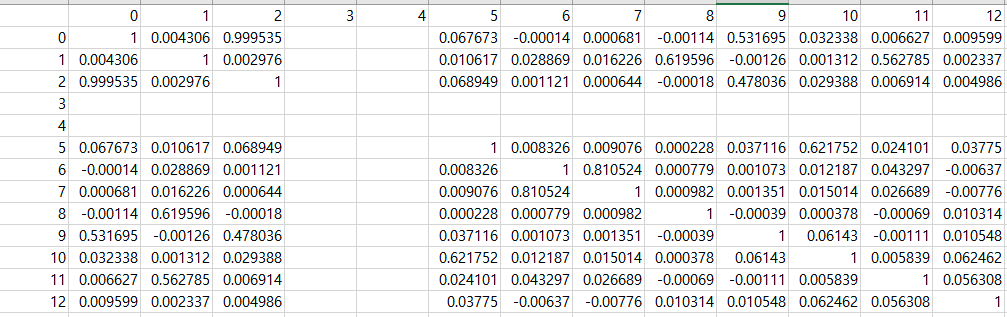
quality 78910은 튀는값존재;

quality 10은 100만넘는애가 12671, 14701, 18392, 24270

26.

[https://drive.google.com/file/d/1h3sogAEt8nQKMnbB-NC7Bqx7E4rgVlV9/view?usp=sharing\](https://drive.google.com/file/d/1h3sogAEt8nQKMnbB-NC7Bqx7E4rgVlV9/view?usp=sharing%5C)

quality간 correlation을 찾아보자



1번이 증가하면 2번이 증가한다. 6번이 증가하면 7번이증가한다, 정도의 상관관계가있다.

각 유저별 quality에 대한 상관관계 파일을 만들어보았는데 ;;1.4gb..

<https://drive.google.com/file/d/1-5268QehhfbwKVM9GwLNKscU6U-s5jx0/view?usp=sharing>

어쩐지 오래걸리더라 헐 대박;;

quality0 이랑 quality 6이 0.999 , 0.947 로 cntmy 즉 problem이 일어난 확률이랑 연관이 있나봐?

뭐지이게

아 이게 연관이 있는건 맞는거같은데 그러면?

quality data랑 problem을 제기하는거랑

다른사람들도 correlation 분석해바

왼쪽부터 quality 0 1 2 3 4 ~ 12

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.007325 | -0.0875 | 0.007024 | 0.005428 |  | 0.037151 | 0.013497 | 0.014858 | 0.033317 | 0.007354 | 0.040621 | -0.09673 | 0.025395 |

결과적으로 quality 값은 그닥 problem이 일어나는거랑은 상관이없어보여

다만 quality 6번이랑 7번은 상관관계가 0.9999확률로 있고

0번과 9번

2번과 9번도 상관관계 0.999

1번과 11번도 0.93 상관관계

<https://drive.google.com/file/d/1Tze4lxtr1EH_kboFWAXprA3-sGfd6PRQ/view?usp=sharing>

여기서 0~12가 quality 이름, 13이 problem 일으킨 횟수

27.

<https://drive.google.com/file/d/1-0-uFVzN0WLUgGbPfJlFWi7ldG2MYF8J/view?usp=sharing>

0~12 quality, 13: problem제기한횟수 14~56 errty 0~ 42

보면

quality 0 - quality 2

quality 0 -quality 9 상관관계

quality 1 -quality 11

quality 2 - quality 9

quality 5 -quality 10

quality 6- quality 7

그냥 수량으로 체크한건데,quality 0 의 count가 많을수록 quality 2의 count도 많아진다?

0.99999

parameter를 줄이는게 문젠데

quality 2, 9 ,11 , 10, 7 3, 4 데이터 0만존재.

0 1 5 6 8 12

?

err 3 - err 11

err 3-err 12

err 6- err 7

err 11- err 12

err 15- err 16

err18 - err 20

err19 -err 20

err 19 - err21

err 20 - err 21

err 22 -err 23

err 27 - err 28

err 7 -err 35

err 36 - err 37

1, 2, 3,4,5, 6, 8, 9 ,10, 13,14,15,17,18,19,22,24,25,26,27,30,31,32,33,34,36,38,39,40,41,42

이렇게하면 31개.

test\_err.csv? 에러타입만 쭉나와있는데

30000~ 45000 에러데이터가있는데

baseline에서는 errtype의 count가지고 한건데

ㄴㄴㄴ 에러 5가 몇개고 에러 10이 몇개면 problem 이다. 식의학습

errtype개수를 줄여서 ,

음음 아니면 우리가 quality data를 이용해서 error type 43 이런 걸 만들어서

이걸 input에 추가해주는거지.

feature을 늘려서하는것도 있고 ctrl+shift+ > 하면글씨키워짐

일별로 errtype 개수 count → baseline보다 점수 향상

err과 quality는 연이없어보이는ㄷ

모델에 넣을 입력변수를 생각해

그리고 각각에 대한 라벨을 입력

random forest는 input 에 null있으면 안되지만 lightgbm은괜차나

그리고 특정 에러타입조합이 예측에 도움이된다면 해당 조합을 찾아내보자.

(<https://dacon.io/competitions/official/235584/codeshare/1030> best\_Y\_ensemble\_searc)참고.

몇개의 에러타입은 problem을 내는데에 상관관계가 없을 수도 있고.

28. 문제가 발생한 시간에 가까워지면 더 높은 가중치를 둬보자

라벨링을 다시하자. 상관관계높은 데이터는 제외시켜줘도 비슷한결과를낸다는 것이 아닐까.

그리고 quality data도 생각해보면

저장된 데이터를 수집한 시점을 time으로 두고,

본인 시스템이 가동했을 때 저장해둔 데이터를 나타낸 것일것 같아.

제출점수가 81.1

0.8066887653152083

내auc점수는위.

29. firmware가 연관이있다면?

firmware가

['05.15.2138' '04.33.1185' '04.33.1261' '04.22.1750' '04.22.1778'

'04.16.3553' '04.33.1149' '04.16.3571' '05.66.3237' '05.66.3571'

'03.11.1149' '04.22.1684' '03.11.1167' '04.82.1684' '04.82.1778'

'04.33.1171' '04.73.2237' '10' '04.82.1730' '04.73.2571' '8.5.3'

'04.22.1666' '03.11.1141' '05.15.2120' '04.33.1125' '04.16.3439'

'04.22.1442' '04.33.1095' '04.16.3569' '05.15.2090' '05.15.3104'

'05.15.2122' '04.22.1656' '04.16.2641' '05.15.2114' '04.16.3345'

'05.15.2092']

f이렇게 존재하는데, 만약 이걸 숫자로 바꾼다면?

03.11.1141의 에러타입 42가 있다면 3111141+42 번의 에러타입으로 변환해봐.

그럼 firmware dependent 한 data가 생겨.

이 수치를 이용해서 training을 한다면 firmware dependent training이 가능하고

firmware dependent prediction이 가능해진다.

근데 이렇게하면 errtype의 값이 낮기때문에 중요도가 낮아지는 문제가 존재한다.

일단 input이 error sequence였으면좋겠다.

28. labli 가 현재 problem 발생한 라벨 trainli가 error sequence.

에러 시퀀스를 50개씩 잘라서 라벨링을 한다면 ? fail..

일단 점수는0.8029472980237337 이거야

제출점수는 0.8075408068

0.8037731666682972가 auc -fw데이터넣었을때 32개로.

0.8066887653152083가 auc

제출점수가 81.115984897

상관은 없는듯.

29. 길이가 너무 긴 항의 경우 가중치를 줄이자.

30. errcode가 유의미하려면 데이터가 많아야하는데,

그러니 31번 errty과 같이 errcode적은 애들로 해보자.

31. errtype 개수가 많은 애들은 (종류가 많은 애들) 가중치를 높게 둬서 training해보자

(err발생 안했는데도 등장한애들)

fw별 분류를해놓고

error type count의 평균을냈을때

평균이랑 많이 차이나는 error type이 있으면

error type 이 해당 fw에 dependent하니까

나오후 10:45

10

8

5

4

3

10 에서만 평균에비해 errtype 5의 count가

많다

42\*5= 210?

32. 해본게 problem을 일으킨 유저들의 err 데이터를 더해서 모델을만들어봤는데 그건

정확도가0.810458156

33. problem 안일으킨 친구들은

34. quality data는 그거야 추이를 이용해서 해보자

min max 차이값 이나 등등