시스템 품질 변화로 인한 사용자 불편 예지 AI 경진대회

팀명: 블랙보리

팀장 : 정진우

문제 : 유저가 해당기간안에 에러를 신고했는지 안했는지를 예측하라

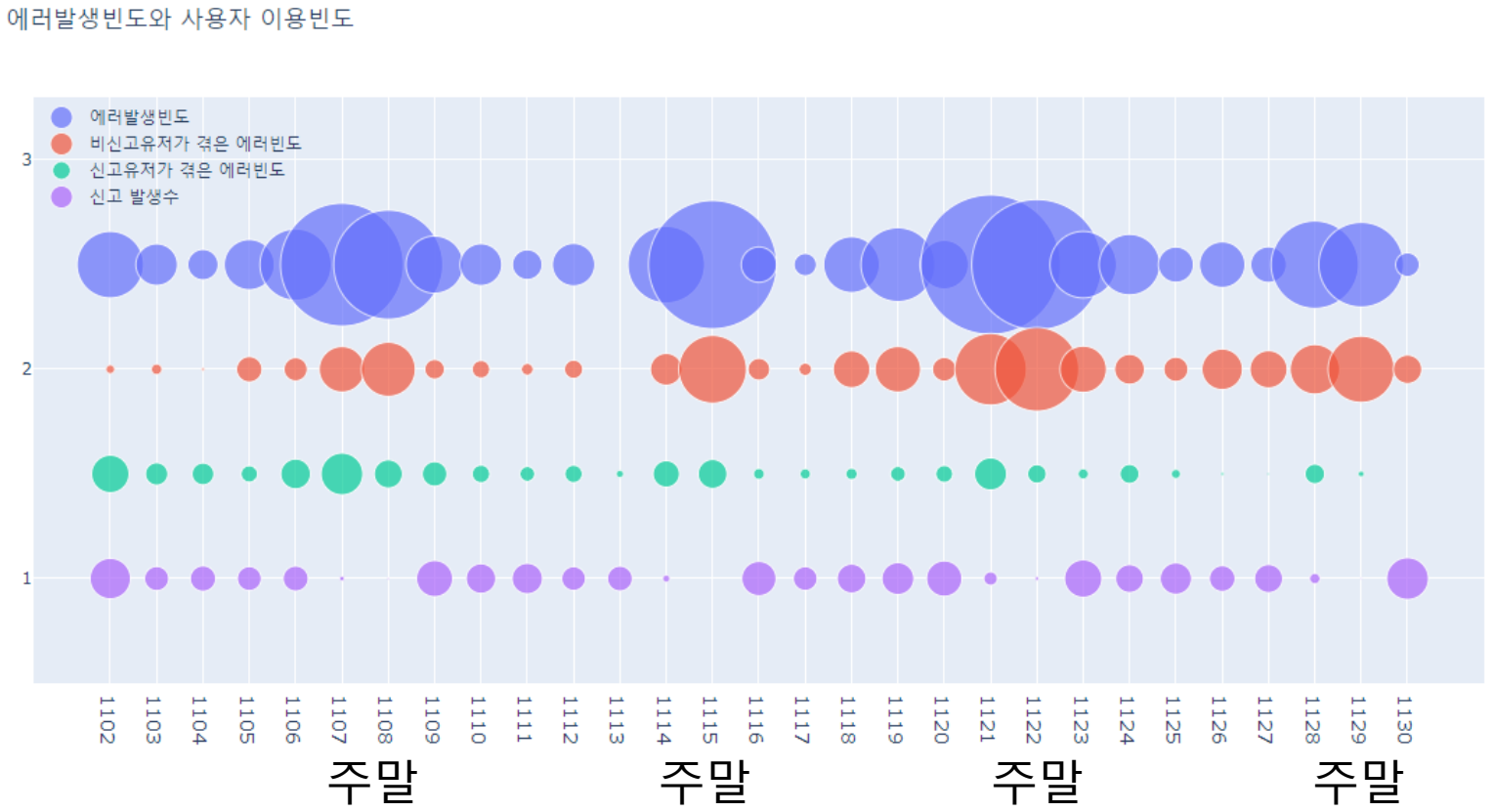
데이터 : 유저가 발생한 에러로그(시간, 종류)
시간에 따른 시스템상태

	user_id	time	model_nm	fwver	errtype	errcode
0	10000	20201101025616	model_3	05.15.2138	15	1
1	10000	20201101030309	model_3	05.15.2138	12	1
2	10000	20201101030309	model_3	05.15.2138	11	1
3	10000	20201101050514	model_3	05.15.2138	16	1
4	10000	20201101050515	model_3	05.15.2138	4	0

	time	user_id	fwver	quality_0	quality_1
0	20201129090000	10000	05.15.2138	0.0	0
1	20201129090000	10000	05.15.2138	0.0	0
2	20201129090000	10000	05.15.2138	0.0	0
3	20201129090000	10000	05.15.2138	0.0	0
4	20201129090000	10000	05.15.2138	0.0	0

0	10000	0
1	10001	1
2	10002	0
3	10003	0
4	10004	1
...	...	...
14995	24995	0
14996	24996	0
14997	24997	1
14998	24998	1
14999	24999	0

EDA

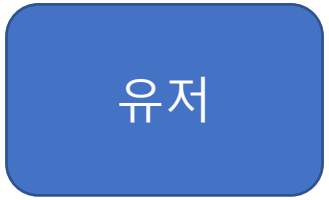


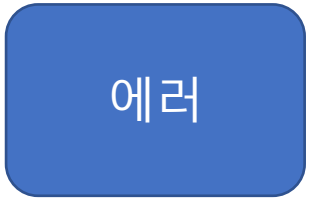
시계열 데이터의
특성을 보임

주말에 에러발생이 급증하지만
신고는 월요일에 가장 큰걸로 보아
사용자는 주로 직장인이고
주말에 사용하는 가전제품이라는
추측을 할 수 있음.

비신고유저와 신고유저의
에러발생빈도 추세가 차이가 존재

문제 해결방향





유저의 에러발생로그 요약

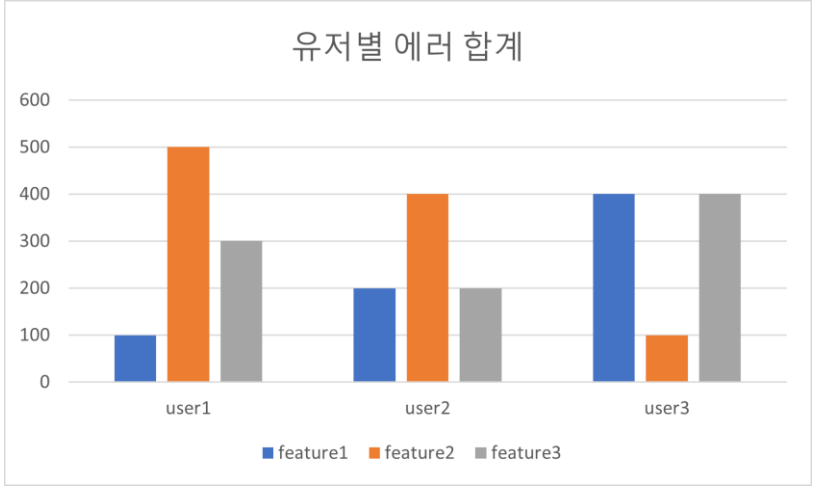
에러종류의 이산화

에러피처의 결합

최종 피처 생성

유저의 에러발생로그 요약

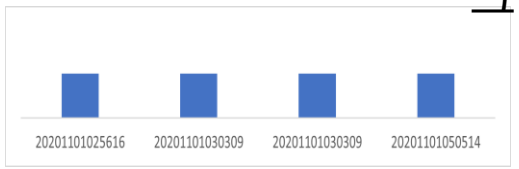
	user_id	time	model_nm	fwver	errtype	errcode
0	10000	20201101025616	model_3	05.15.2138	15	1
1	10000	20201101030309	model_3	05.15.2138	12	1
2	10000	20201101030309	model_3	05.15.2138	11	1
3	10000	20201101050514	model_3	05.15.2138	16	1
4	10000	20201101050515	model_3	05.15.2138	4	0



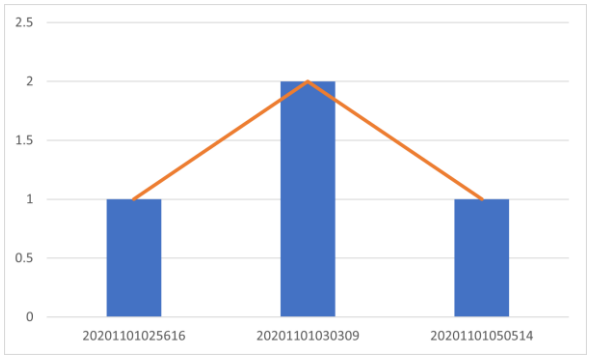
전체기간에 대한 유저의 합계는 유저의 시계열 정보를 표현하지 못한다.

유저의 에러발생로그 요약

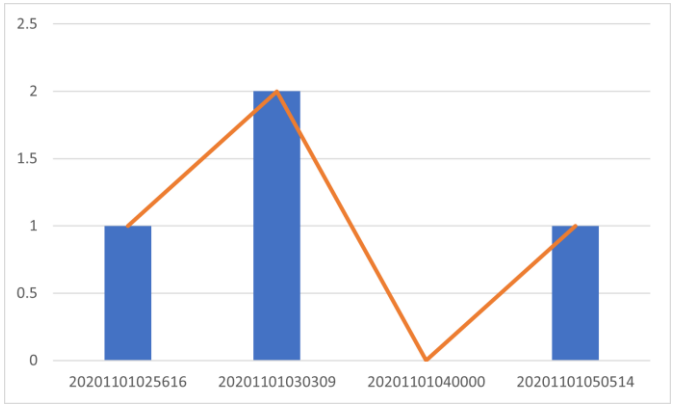
	user_id	time	model_nm	fwver	errtype	errcode
0	10000	20201101025616	model_3	05.15.2138	15	1
1	10000	20201101030309	model_3	05.15.2138	12	1
2	10000	20201101030309	model_3	05.15.2138	11	1
3	10000	20201101050514	model_3	05.15.2138	16	1



같은시간대로
그룹화



에러가 발생하지
않은 시간대를
0으로 설정

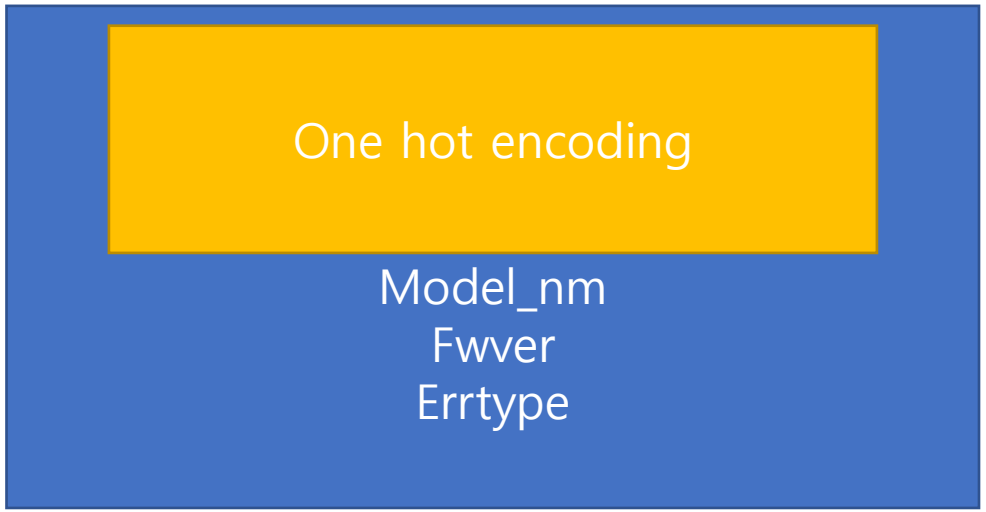


그래프를 관측하고 0인지 1인지 분류
시계열데이터의 특성을 요약한 통계량을 피처로 사용.

*유저별 이용기간의 차이가 크기때문에 딥러닝을 이용하여 요약하긴 힘듬

에러종류의 이산화

	user_id	model_nm	fwver	errtype	errcode
count	16554663	16554663	16554663	16554663	16554662
unique	15000	9	37	41	2805
top	24934	model_1	04.16.3553	23	1
freq	222186	5384491	5237816	2276515	8906967



종류가 2805나 있기때문에 one-hot encoding시
데이터의 차원이 상당히 늘어나 분석이 힘들어진다.

자연어처리에서의 TF-IDF를 적용

에러종류의 이산화

한명에게만 발생한 종류 무시
추측에 포함되지 않는 종류 무시

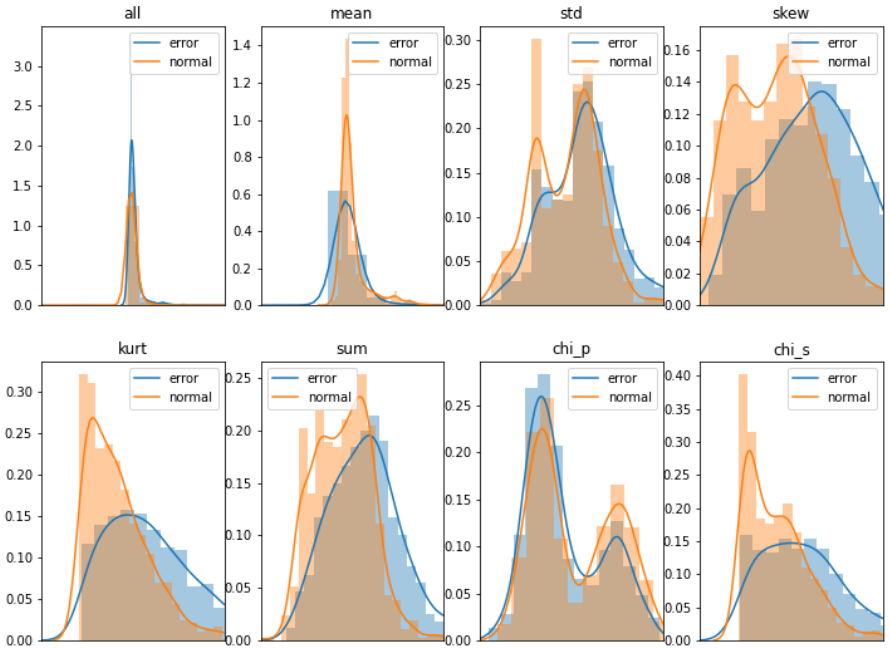
TF : 특정유저가 특정 errcode를 겪은 횟수

DF : 특정 errcode를 겪은 유저의 수

$$TF - IDF = \frac{TF}{\log(1+DF)} = \frac{\text{특정유저가 특정 } errcode\text{를 겪은 횟수}}{\text{특정 } errcode\text{를 겪은 유저의 수}}$$

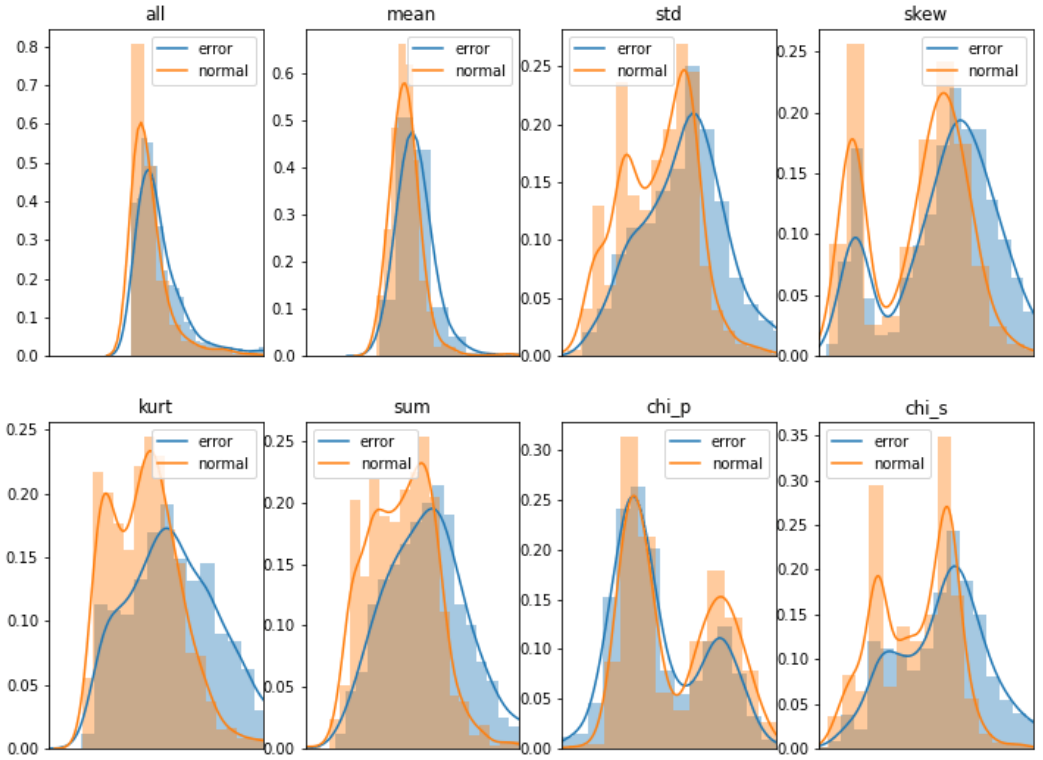
TF와 DF, TF-IDF 세가지 모두 피처로 활용

생성한 피처의 시각화 분석



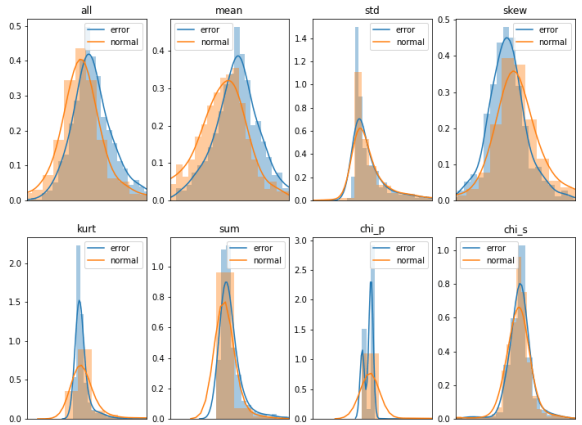
Errtype

일 별 합계만 적용했을때

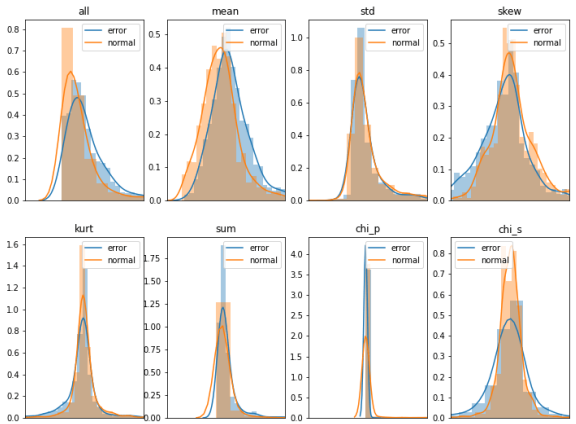


에러가 발생하지 않은 시간을
n시간 단위로 고려했을때

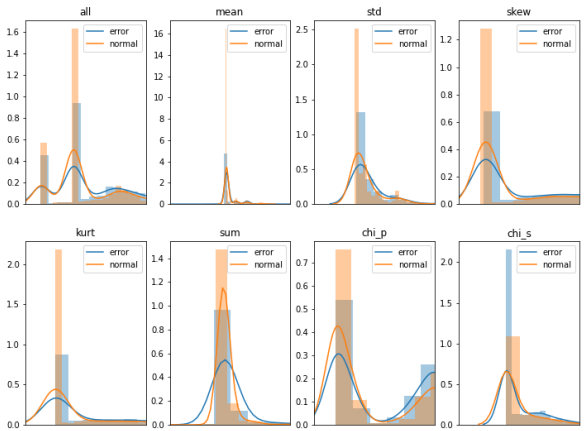
생성한 피처의 시각화 분석



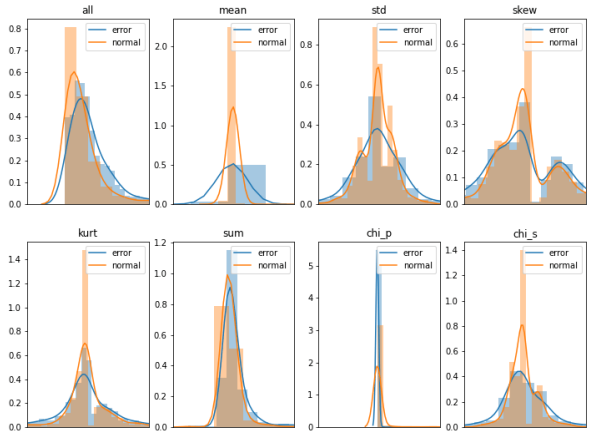
errcode



Model_nm



quality

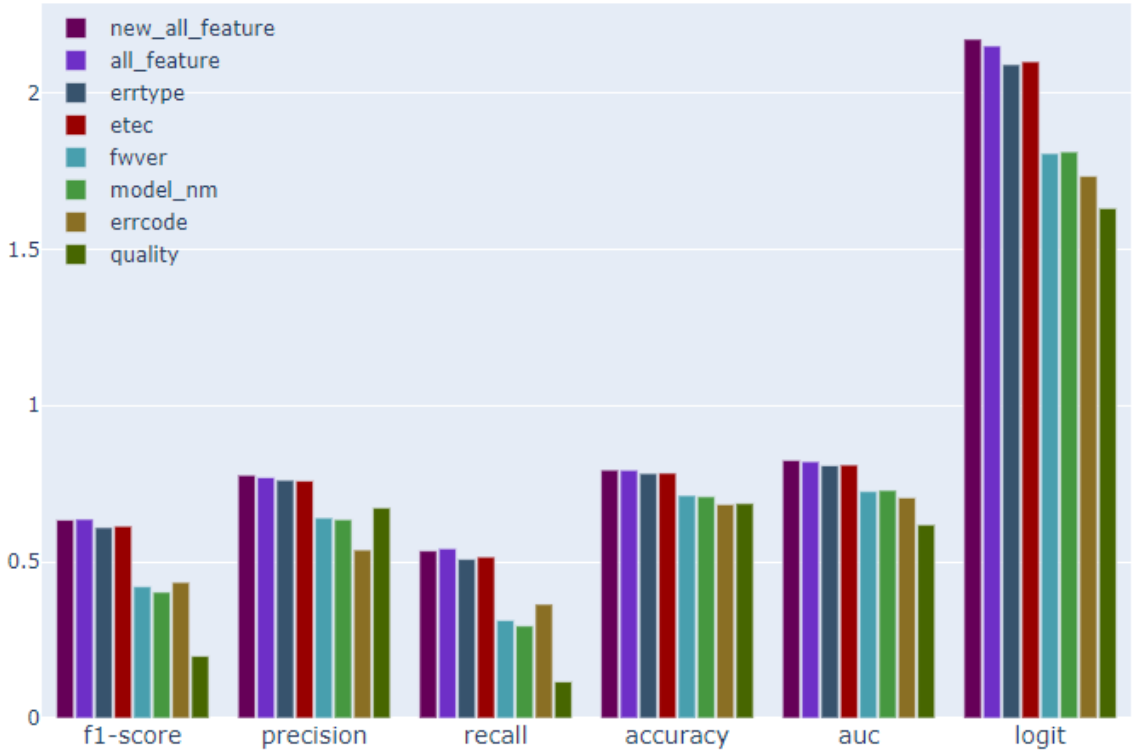


fwver

가장 집단간의 차이가 큰 시간대로 optimize 하여 분석할 수 있다.

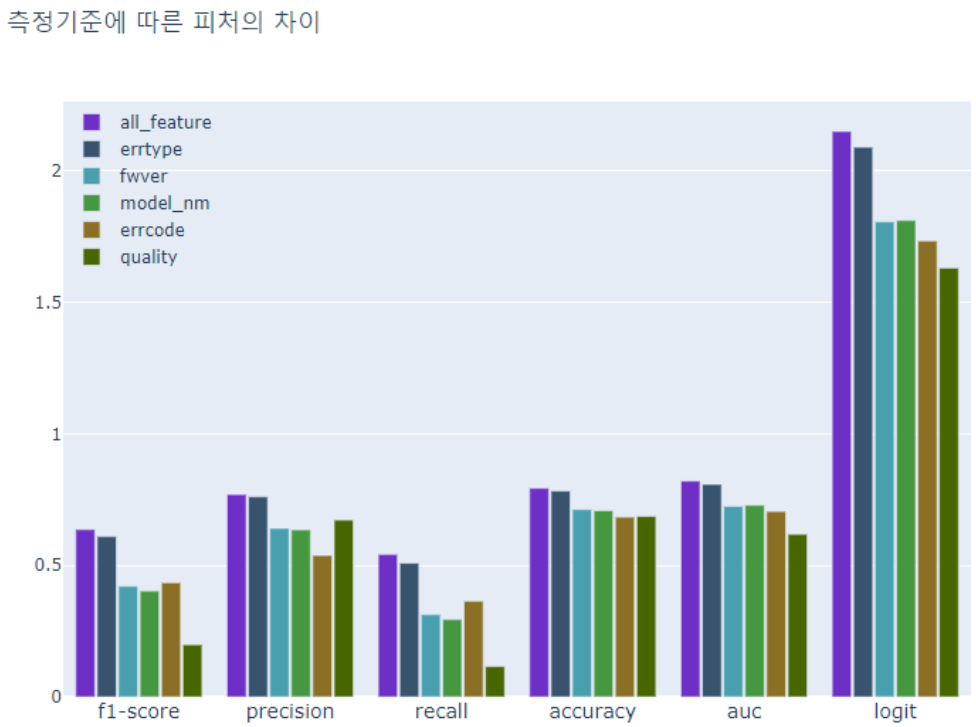
에러피처의 결합

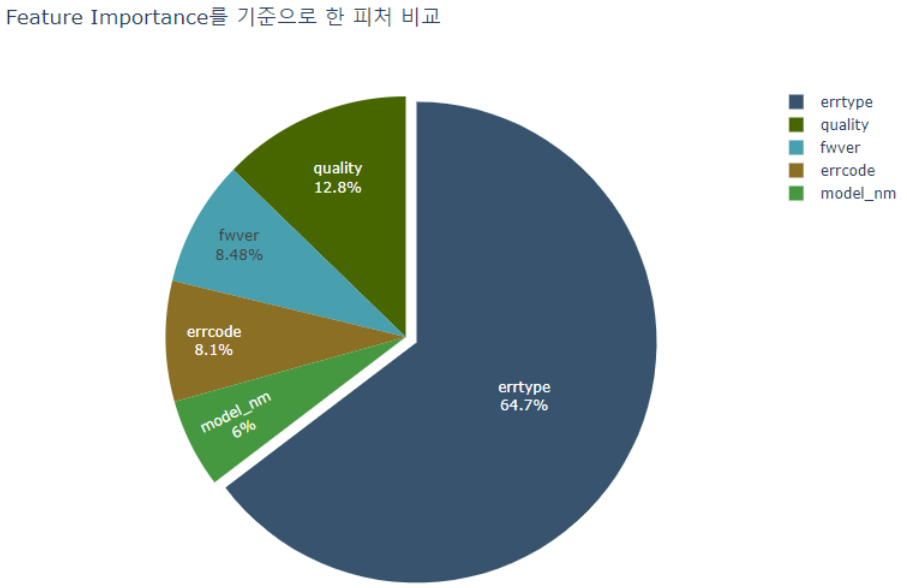
측정기준에 따른 피처의 차이



성능증가

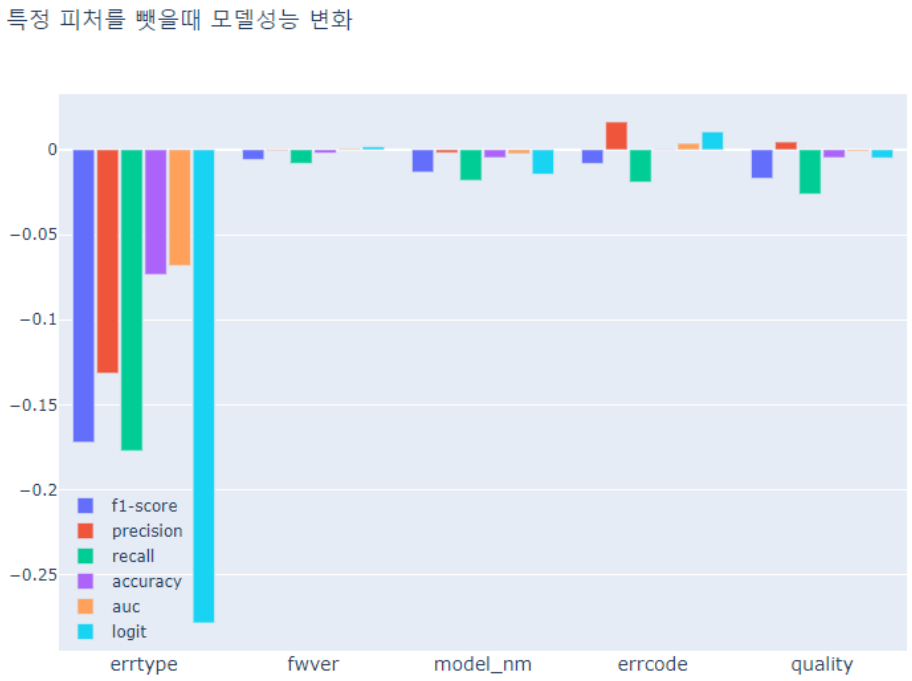
피처별 중요도 분석



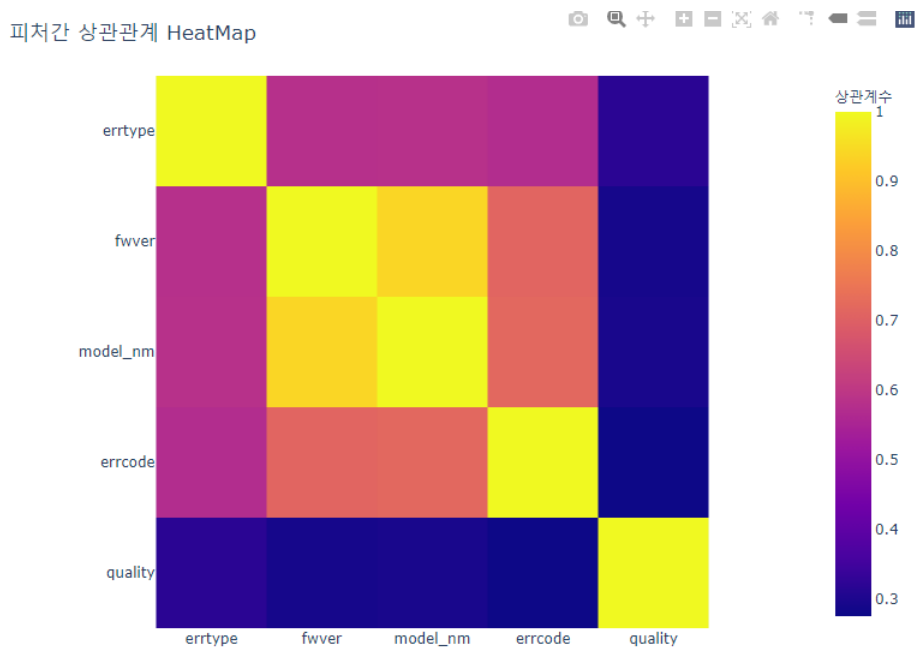


사용자의 입장에서는 errtype의 발생과 변화가 가장 중요

피처간 상관관계 분석

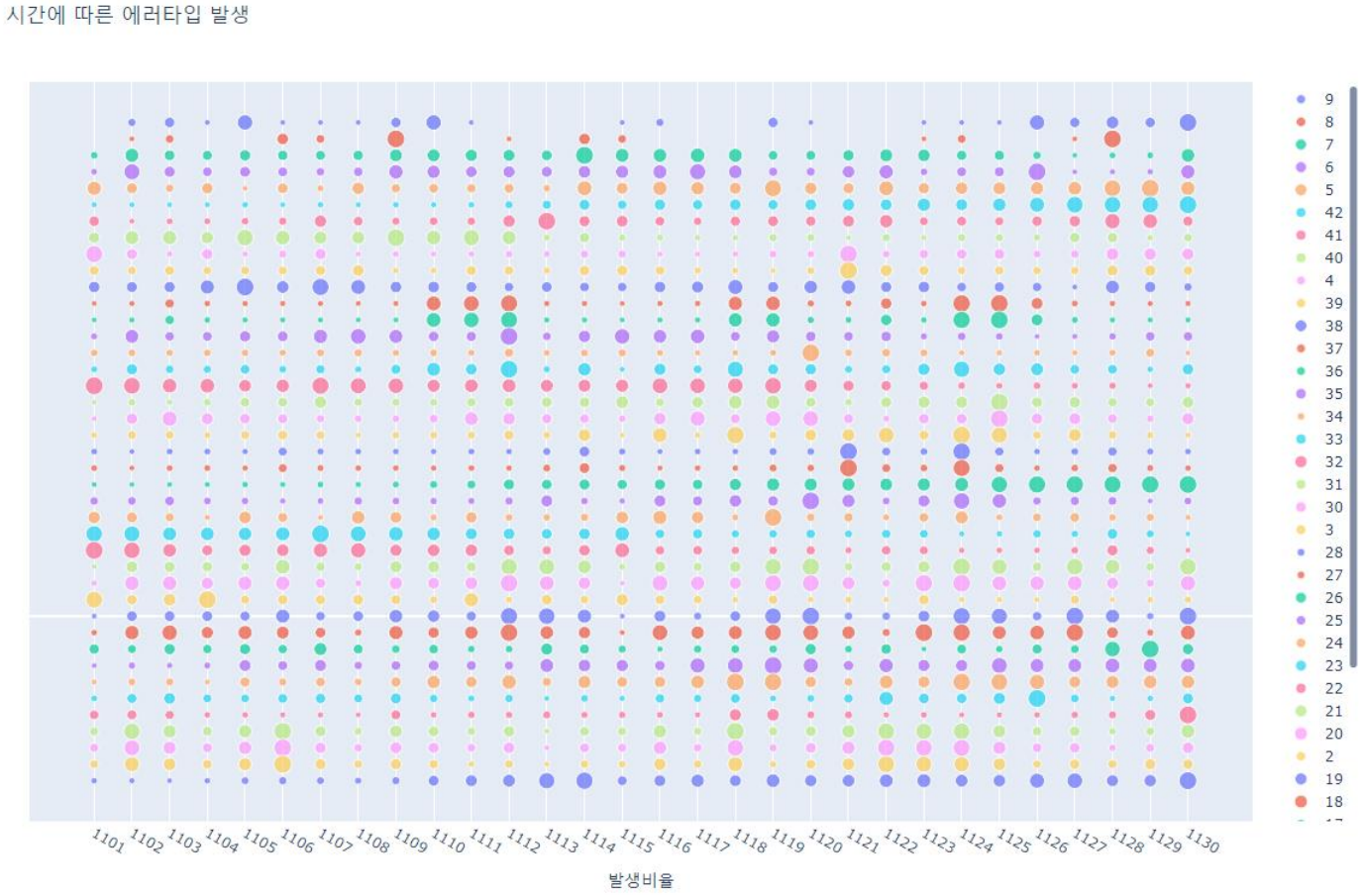


일부 피처들은 전체성능에
큰 영향이 없음



Model_nm 과 fwver은 매우 강한 상관관계가 존재
Errcode도 model_nm과 fwver과 상관관계가 크지만
성능은 좋지 않은걸로 보아 일부 종속관계에 있음

Errtype 분석



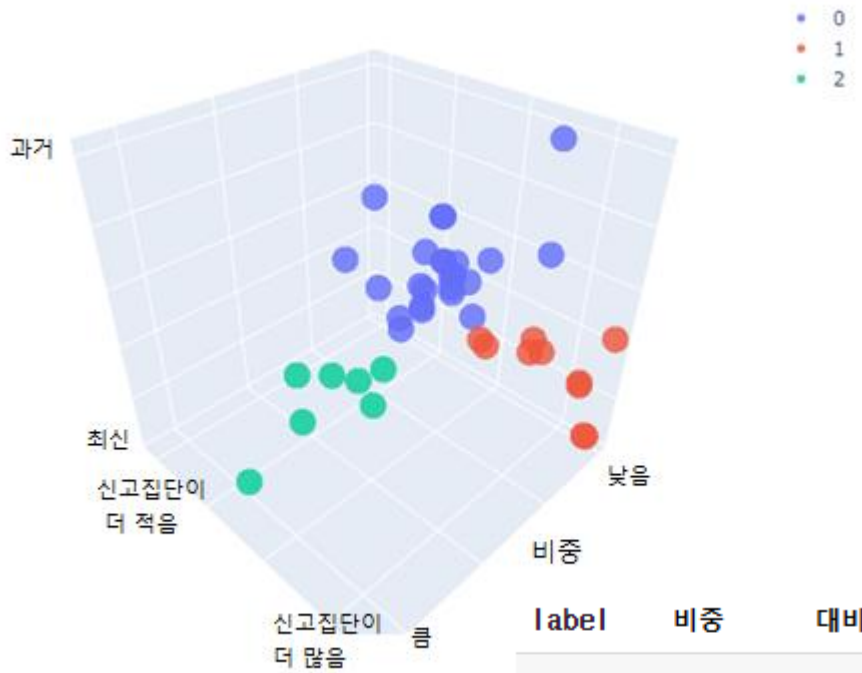
종류에 따라 다른 시계열 특성을 보이고 있다.

1.비신고 유저가 대비 신고유저가 겪은 횟수
2.전체 에러에서 차지하는 정도

3.최근에 많이 발생하는지

클러스터링을 통한 시각화 분석

Errtype분석



label	비중	대비	낮을수록 최근
0	0.292609	0.443941	1.138816
1	0.022290	0.759752	0.063888
2	0.685101	0.462082	-0.006591

0번집단 : 비신고유저가 많이 겪은 에러이다.
전체 비중이 낮고
최근엔 별로 발생하지 않았다

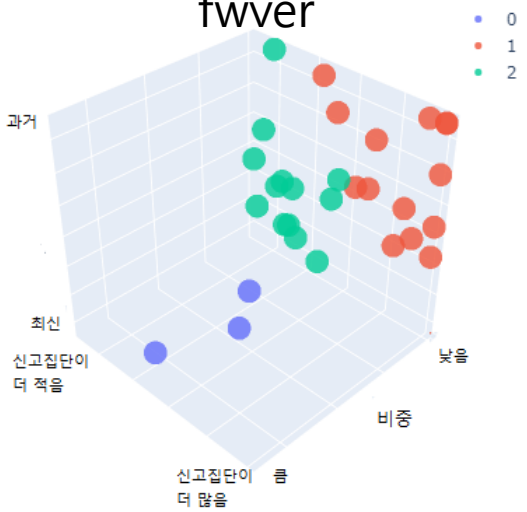
1번집단 : 신고유저가 특히 많이 겪은 에러이다.
하지만 전체 비중이 매우 낮다
비교적 꾸준히 발생하고 있다.

2번집단 : 비신고 유저가 많이 겪은 에러이다.
전체 비중이 가장 크다
최근에 더 발생하고 있다.

2번집단에 속한 errtype을 우선적으로 해결

Fwver, model_nm, errcode 분석

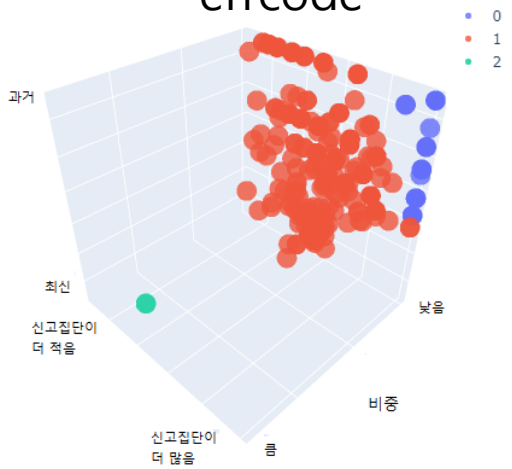
fwver



0번 : 발생빈도가 특히 크고
최근에 많이발생
1번 : 신고유저가 가장 많이 겪었지만
최근에는 거의 발생하지 않음
2번 : 신고유저가 많이 겪었지만
최근에는 거의 발생하지 않음

0번그룹 펌웨어 보완 필요

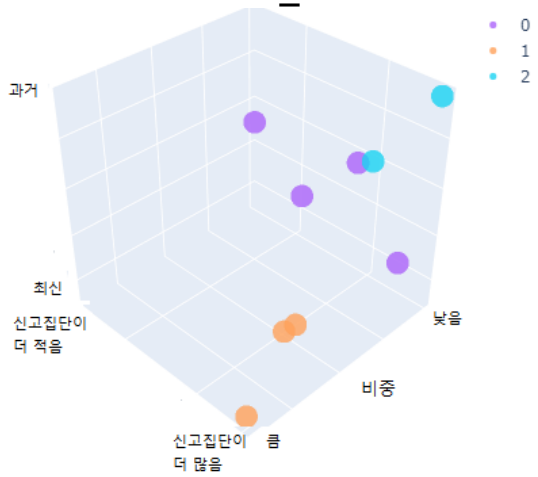
errcode



0번 : 거의 신고유저만 겪음. 하지만 비중이
매우 낮고 최근엔 거의 발생하지 않음
1번 : 비중이 크지만 최근엔 거의 발생하지
않음
2번 : 비중이 가장 크고
최근에는 많이 발생함

2번그룹 에러코드 보완 필요

Model_nm



0번 : 발생 비중이 낮고 과거에 주로
발생했다.
1번 : 비신고유저가 조금 더 많이 겼었지만
비중이 가장 크고 최근에 발생했다.
2번 : 비중이 작고
최근에는 거의 발생하지 않음

1번그룹 모델 보완 필요