LG AI Research & DACON

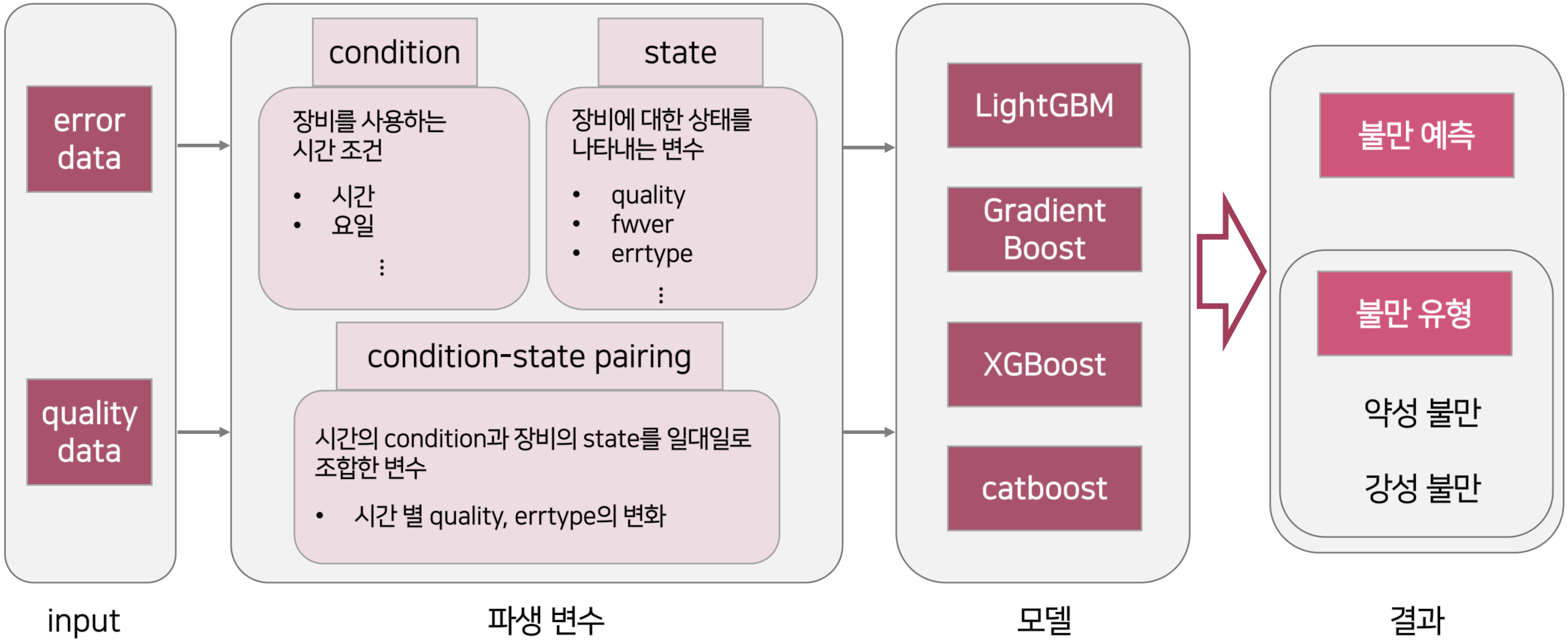
시스템 품질 변화로 인한 사용자 불편 예지 AI 경진대회

사용자의 불만 유형을 중심으로

상인동

목 차

01 분석 개요


error data
quality data
condition
장비를 사용하는 시간 조건
• 시간
• 요일
⋮
state
장비에 대한 상태를 나타내는 변수
• quality
• fwver
• errtype
⋮
condition-state pairing
시간의 condition과 장비의 state를 일대일로 조합한 변수
• 시간 별 quality, errtype의 변화
LightGBM
Gradient Boost
XGBoost
catboost
불만 예측
불만 유형
약성 불만
강성 불만
input
파생 변수
모델
결과

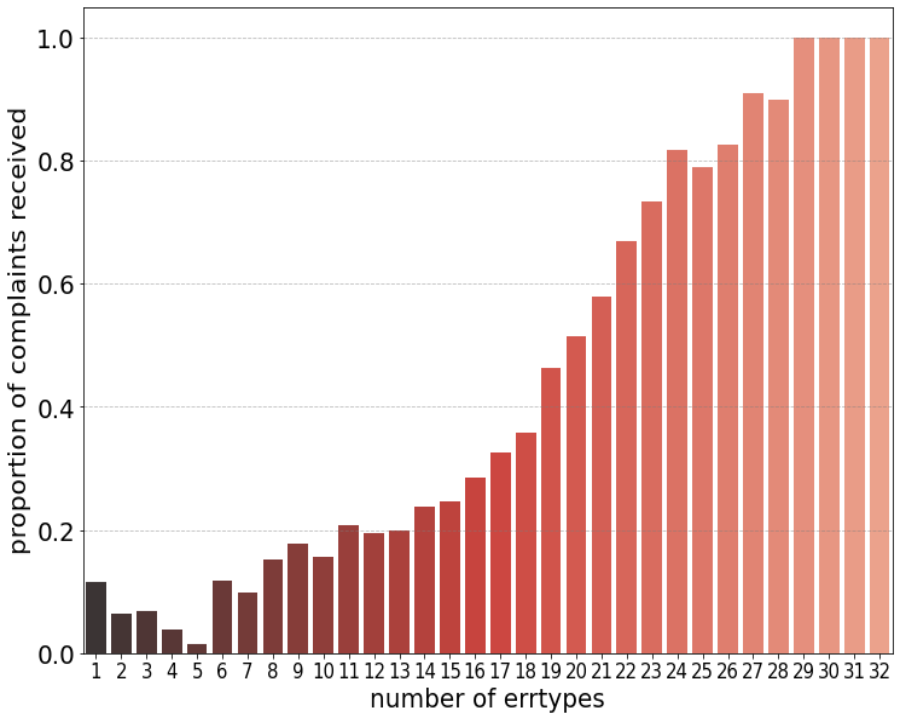
02 데이터 분석

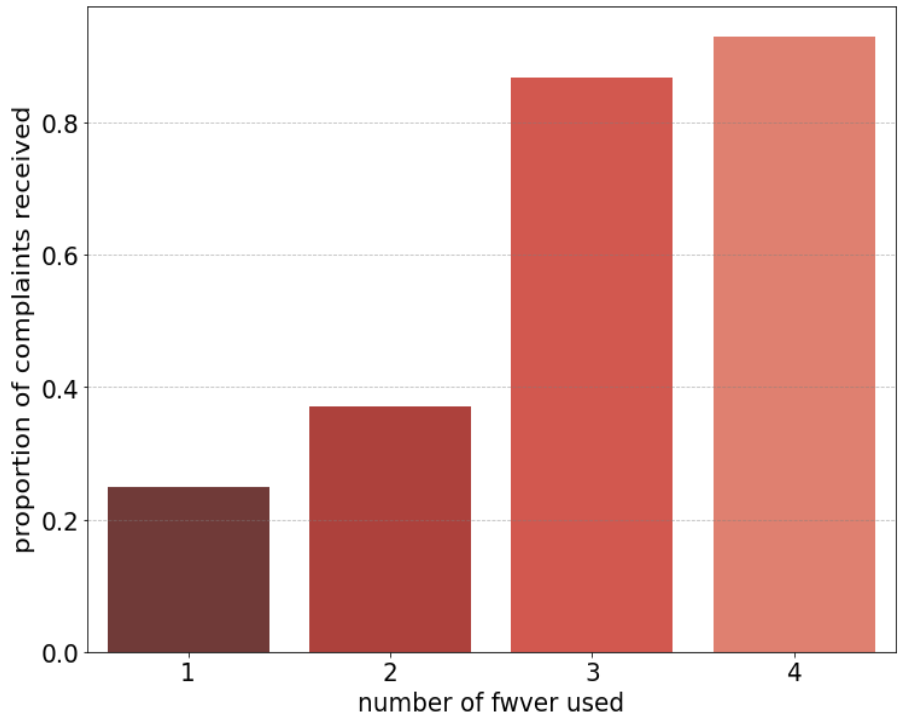
파생변수 유형

state : 장비에서 발생한 error 상태를 나타내는 그룹

quality 변수의 변화, 사용한 fwver의 수, 발생한 errtype의 수와 같이 장비에서 에러가 발생한 상태를 나타내는 변수

- user별로 발생한 errtype의 수가 커질 수록 해당 유저가 불만을 접수할 확률이 높아지는 경향
- user가 사용한 fwver의 횟수 역시, 많은 fwver을 사용한 고객일수록 불만을 접수할 확률이 높음





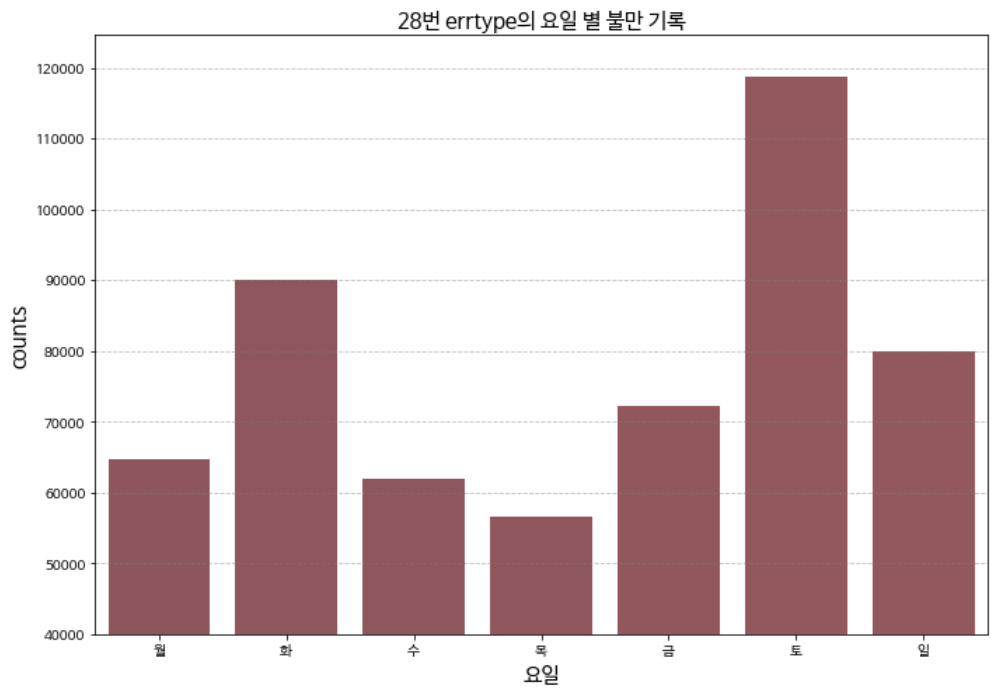
02 데이터 분석

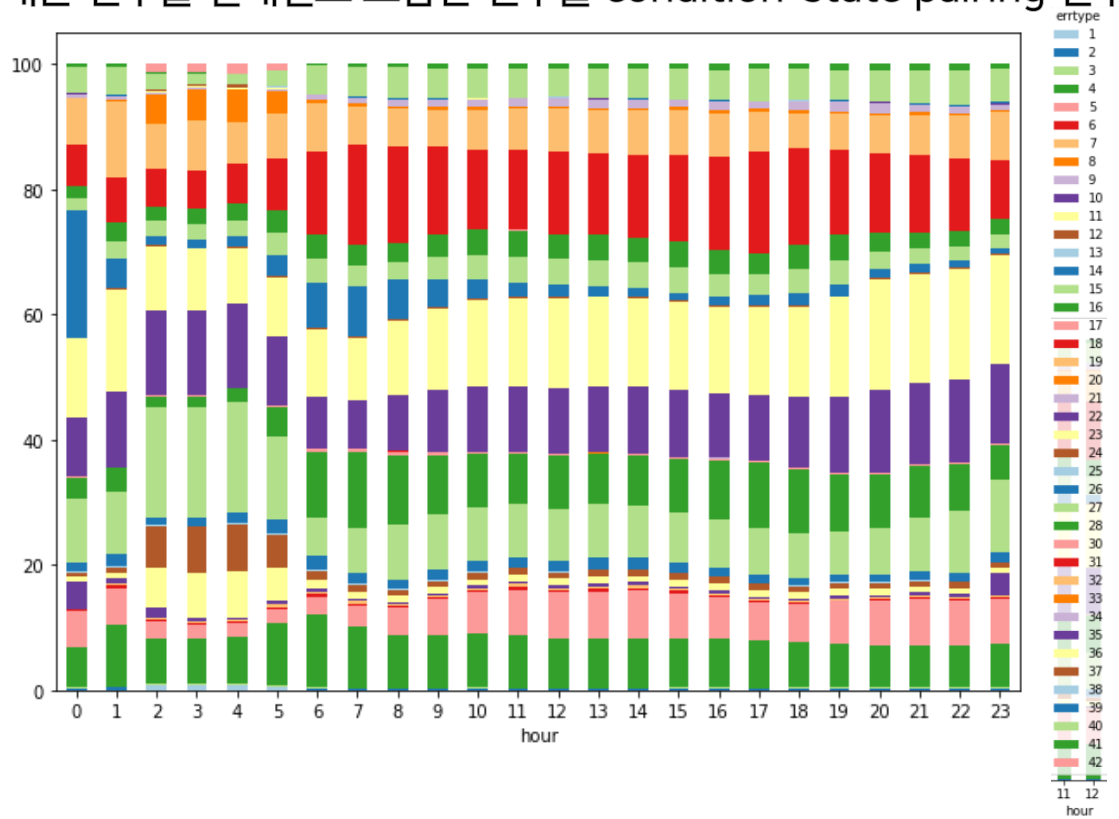
파생변수 유형

condition : 장비를 사용하는 시간 속성을 나타내는 그룹

장비를 사용하는 요일, 휴일 여부, 시간대를 나타내는 변수

- errtype에 대해 요일·시간별로 그래프를 그려본 결과 요일과 시간에 따라 자주 발생하는 errtype의 패턴 확인
- 이와 같이 시간 조건인 condition과 장비의 state을 나타내는 변수를 일대일로 조합한 변수를 condition-state pairing 변수로 설정





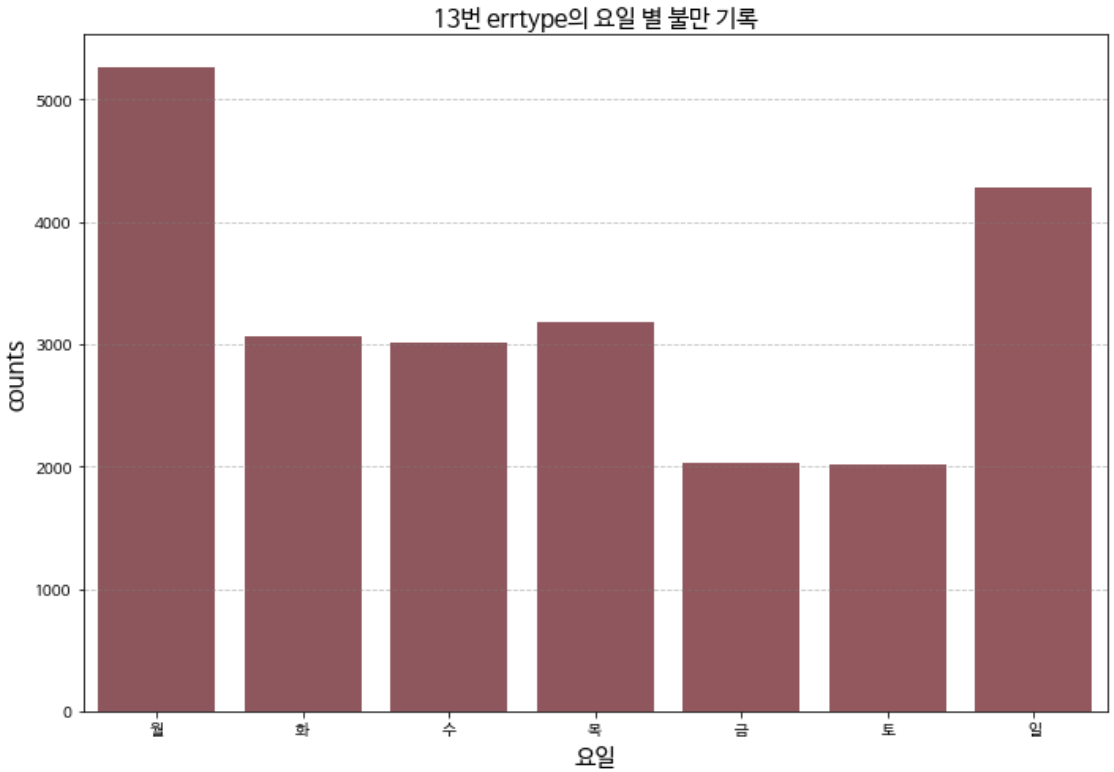
02 데이터 분석

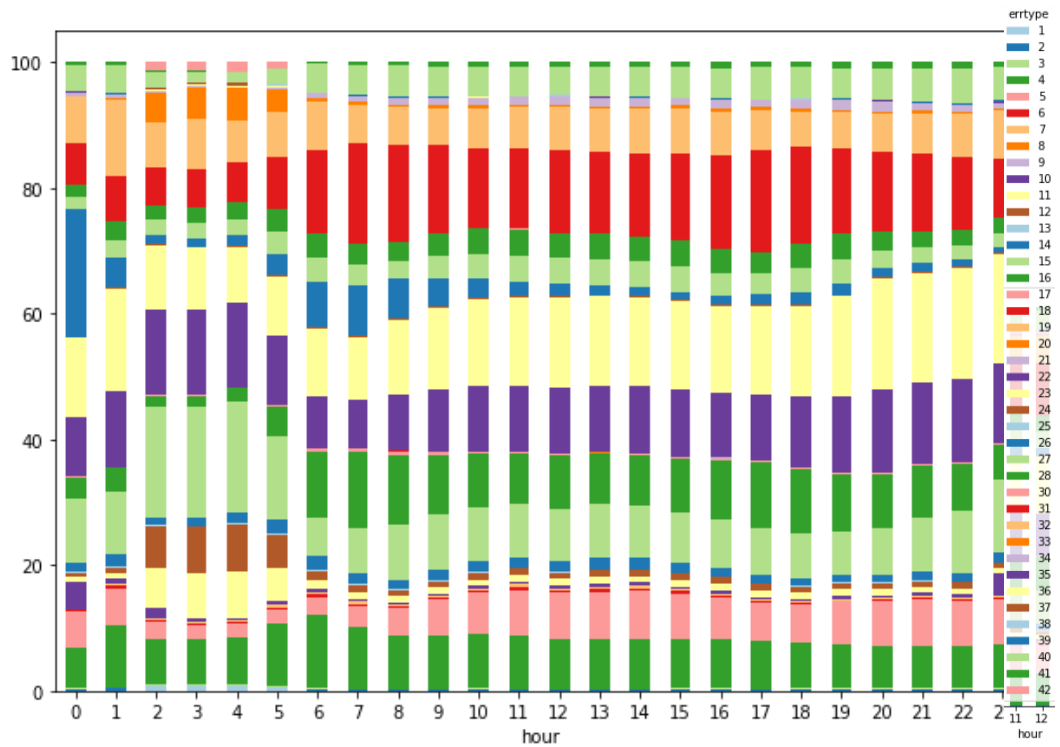
파생변수 유형

condition : 장비를 사용하는 시간 속성을 나타내는 그룹

장비를 사용하는 요일, 휴일 여부, 시간대를 나타내는 변수

- errtype에 대해 요일·시간별로 그래프를 그려본 결과 요일과 시간에 따라 자주 발생하는 errtype의 패턴 확인
- 이와 같이 시간 조건인 condition과 장비의 state을 나타내는 변수를 일대일로 조합한 변수를 condition-state pairing 변수로 설정





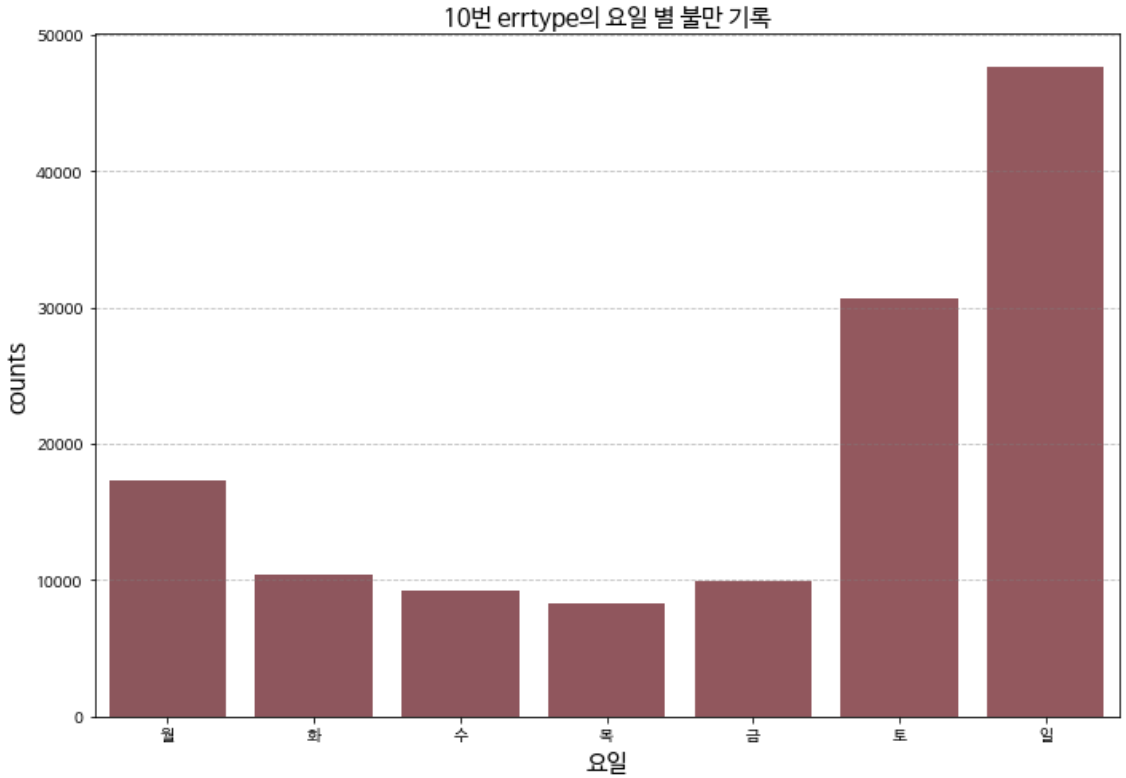
02 데이터 분석

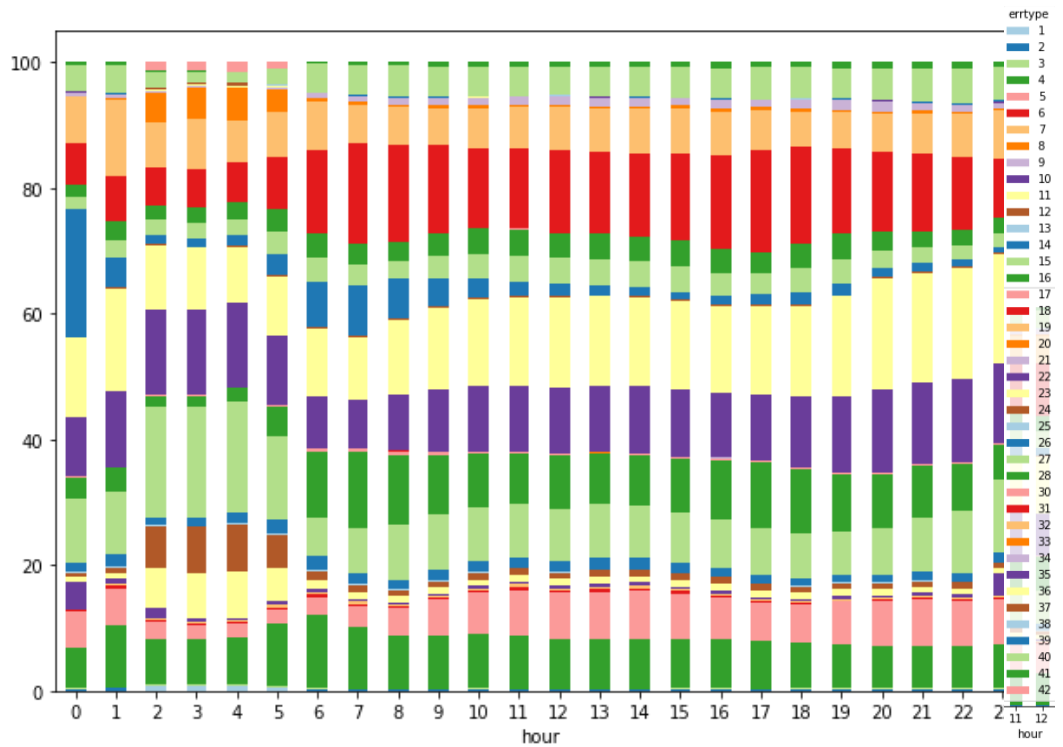
파생변수 유형

condition : 장비를 사용하는 시간 속성을 나타내는 그룹

장비를 사용하는 요일, 휴일 여부, 시간대를 나타내는 변수

- errtype에 대해 요일·시간별로 그래프를 그려본 결과 요일과 시간에 따라 자주 발생하는 errtype의 패턴 확인
- 이와 같이 시간 조건인 condition과 장비의 state을 나타내는 변수를 일대일로 조합한 변수를 condition-state pairing 변수로 설정

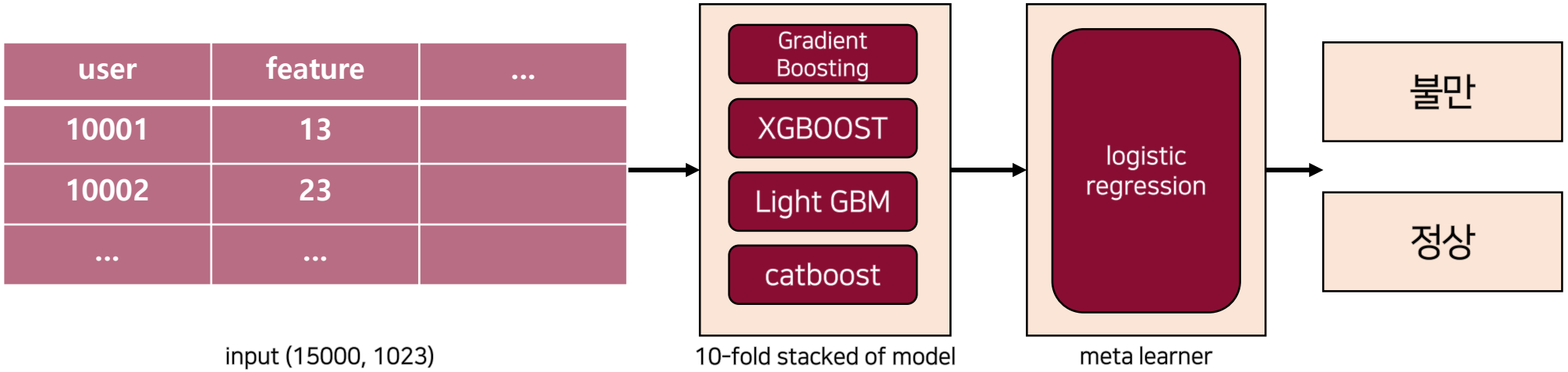




02 데이터 분석

모델링

모델 설계

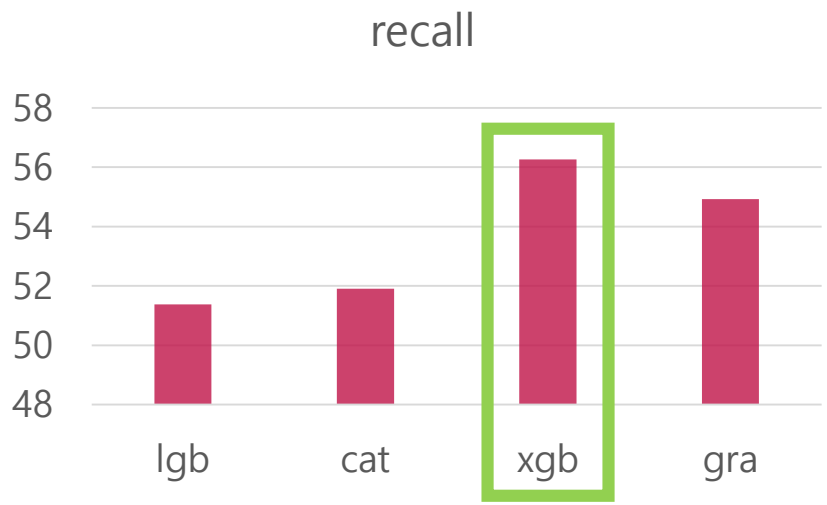


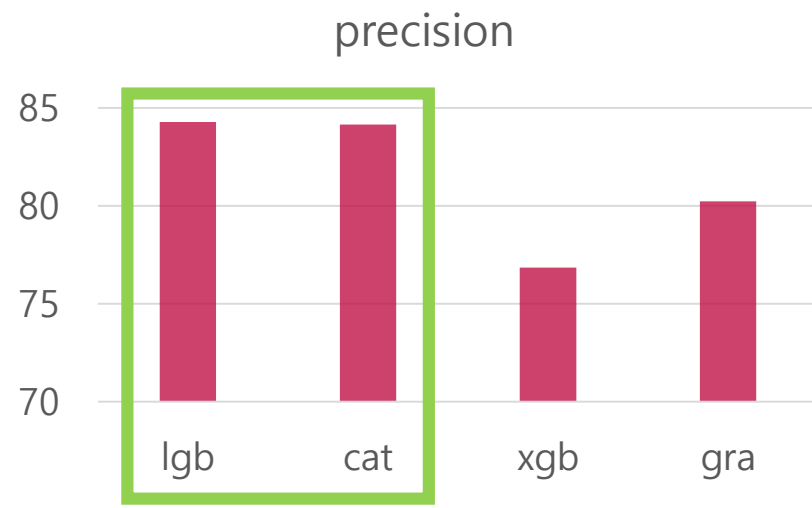
- 유저별로 시간의 정보를 담은 3차원 데이터를 2차원의 데이터로 변환
- cross validation별로 나눈 모델의 성능폭이 크기 때문에 10-fold로 나눈 모델을 사용하여 stacking
- 각각의 모델의 결과값을 고려할 수 있도록 logistic regression모델의 meta learner를 사용

02 데이터 분석

모델링

모델 별 성능 (불만 유저 기준)





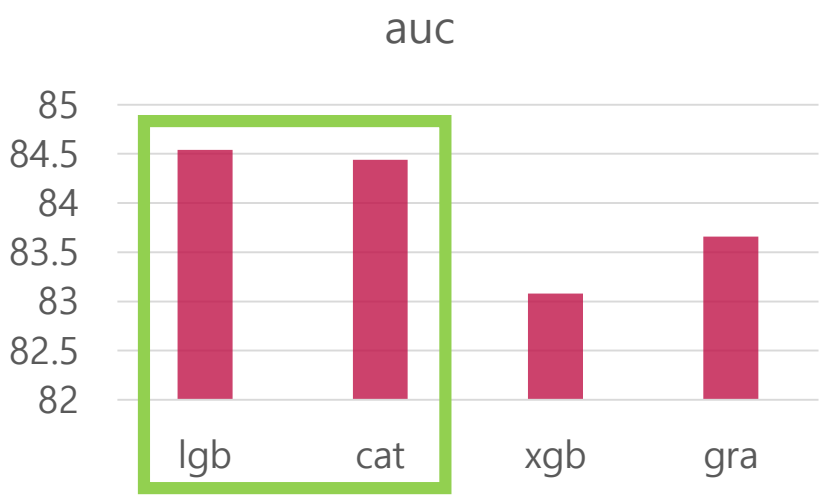
- recall이 높다는 것은 실제 불만 유저 중에 모델이 얼마나 많이 맞췄는가로 해석
- 모델이 불만 유저에 대한 기준을 넓게 잡을수록 recall값이 높음
- xgb는 불만에 대한 기준을 넓게 보고 판단
- xgb모델은 **약성 불만을 잘 잡아내는 모델**로 판단

- precision이 높다는 것은 모델이 불만이라고 예측한 것 중에 실제 불만 유저를 얼마나 많이 맞췄는가로 해석
- 모델이 불만 유저에 대한 기준을 좁게 잡을수록 precision값이 높음
- lgb, cat은 불만에 대한 기준을 좁게 보고 판단
- lgb, cat은 **강성 불만을 잘 잡아내는 모델**로 판단

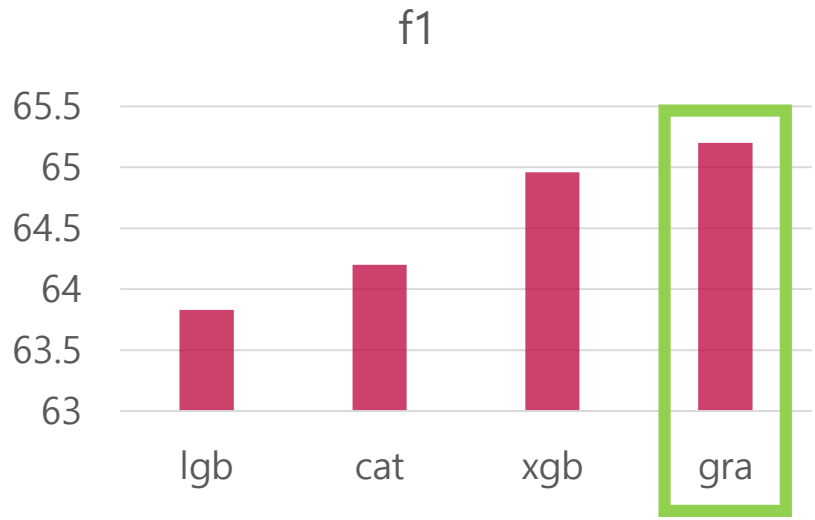
02 데이터 분석

모델링

모델 별 성능 (불만 유저 기준)



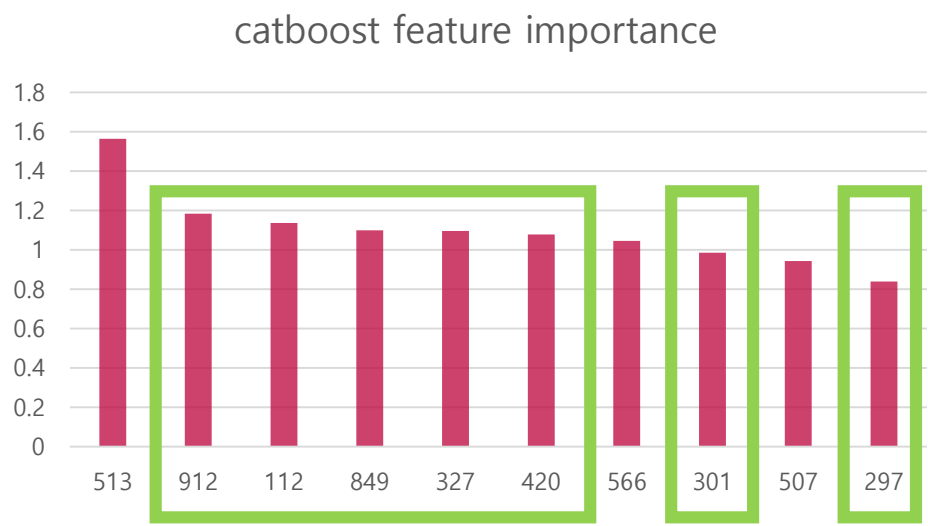
- 불만과 정상의 Optimal cut point를 찾는 auc로 경계값을 얼마나 잘 찾았는지로 해석
- 불만 뿐만 아니라 정상인 유저들도 잘 분류해야 높은 성능을 보일 수 있다고 판단
- 강성 불만에 강점을 보인 lgb와 cat은 정상인 유저들도 잘 분류하기 때문에 auc가 높게 나타남

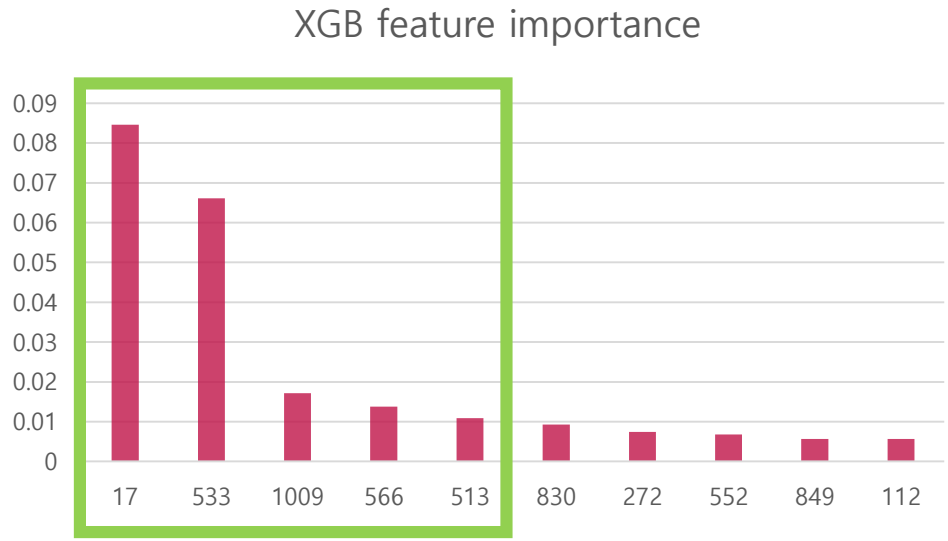


- recall과 precision에서 고른 성능을 보인 GBC가 f1 score가 제일 높게 나타남
- 강성과 약성 불만인 유저 분류의 성능을 동시에 파악할 수 있는 지표
- Gradient boosting은 불만 유형을 분류함에 있어 안정적인 모델

02 데이터 분석

모델 기반 불만 유형 분석 : 불만 유형





catboost를 이용한 강성 불만 유형

- 상위 변수 중요도의 크기 차이가 크지 않아,
  다양한 변수에 영향을 받은 모델
- 상위 10개의 변수 중요도를 가지는 변수는 주로
  condition 그룹 또는 condition-state pairing 그룹
- condition을 포함한 변수가 주요 변수로 작용했기 때문에
  시간별 state의 변화량에 따라 불만이 누적된 고객일수록
  강한 불만을 제기할 가능성이 높음

XGB를 이용한 약성 불만 유형

- XGB는 catboost와는 다르게
  변수 사이의 변수 중요도의 차이가 큼
- 상위 5개의 변수 중요도를 가지는 변수는 모두 state 그룹
- 변수 중요도가 높은 변수들은 대부분이 state 변수이고,
  특정한 에러타입(18) 발생과 같이 장비의 특정한 상태에
  문제가 있는 경우 약한 불만을 제기할 가능성이 높음

02 데이터 분석

모델 기반 불만 유형 분석 : 불만 유형 예제

강성 불만(precision based)

user_id	catboost(proba)	XGB(proba)	18번 장비 errtype	unique error type 개수	…
4363	0.8855	0.4912	0	13	…

- 4363번 유저의 경우 18번 장비 errtype이 나타나지 않고, unique error type개수를 판단하여 강성 유저를 잘 맞추는 catboost가 불만으로 분류

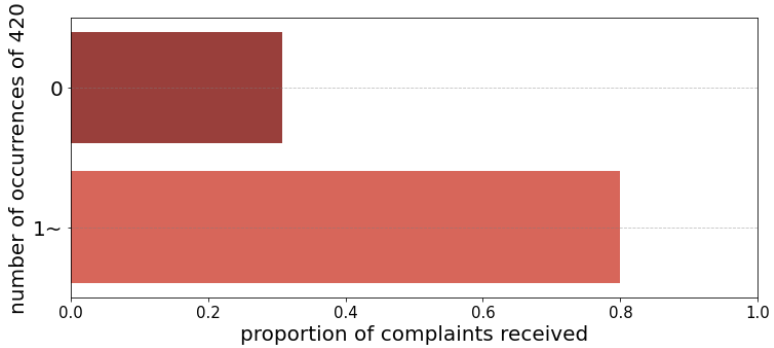
약성 불만(recall based)

user_id	catboost(proba)	XGB(proba)	18번 장비 errtype	unique error type 개수	…
1054	0.4588	0.7811	1	18	…

- 1054번 유저의 경우 약성 유저에서 발생한 18번 장비 errtype이 발생하였으므로 약성 유저를 잘 맞추는 XGB가 불만으로 분류

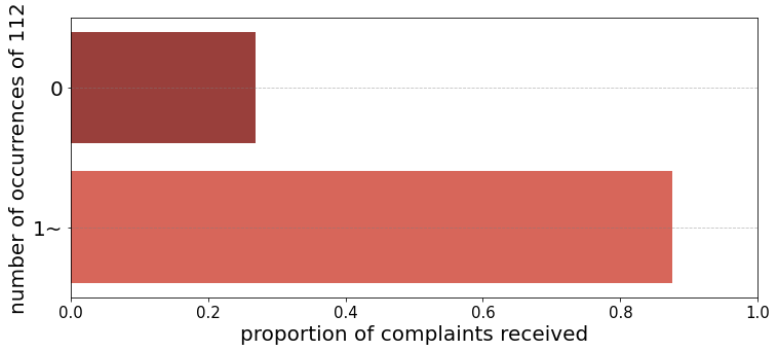
02 데이터 분석

모델 기반 불만 유형 분석 : catboost feature



420: split_codework

- 새벽 2~5시 사이에 발생하는 'terminate by peer user' errcode 발생 횟수
- 새벽시간의 에러코드임에도 불구하고 한 번이라도 발생하면 불만을 제기하는 고객의 비율이 80% 정도로 늘어남



112: split_typework

- 9~19시 사이에 발생하는 errtype 18의 발생 횟수
- 18번 에러타입은 발생 한번으로도 불만 제기율이 높아지는데, 활동시간인 9~19시 사이에 발생하면 불만 제기율이 더 높아짐
- 활동시간에도 영향을 받는 에러 타입임을 확인

quality_5	정상	불만
최솟값	0	0
평균	16.55	130.63
최댓값	13299.73	138329.2
분산	34279	6177233

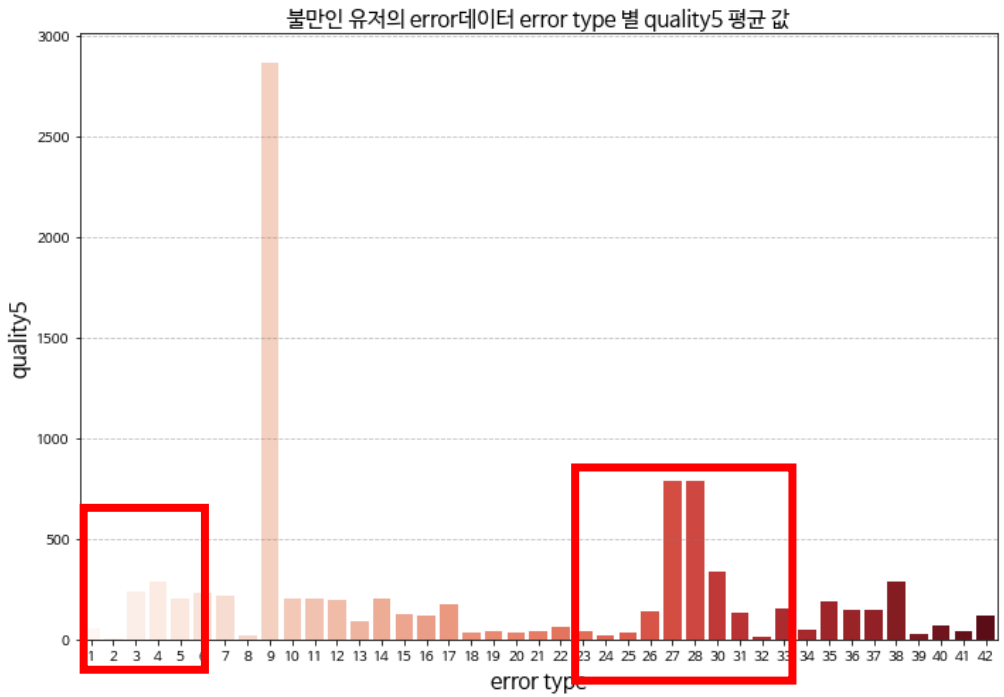
507: split_qt_int5

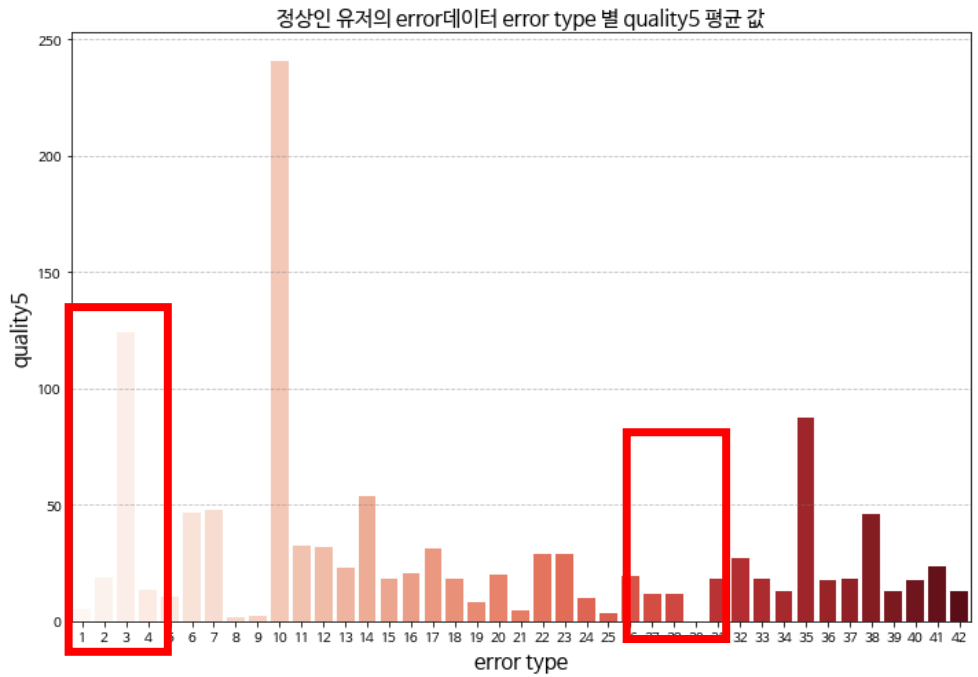
- 유저 별 quality_5의 평균 값
- quality_5의 값이 높게 나타나면 강한 불만도가 나타날 것으로 판단

02 데이터 분석

모델 기반 불만 유형 분석 : catboost feature

error data와 quality data간의 관계를 quality5의 값을 통해 확인



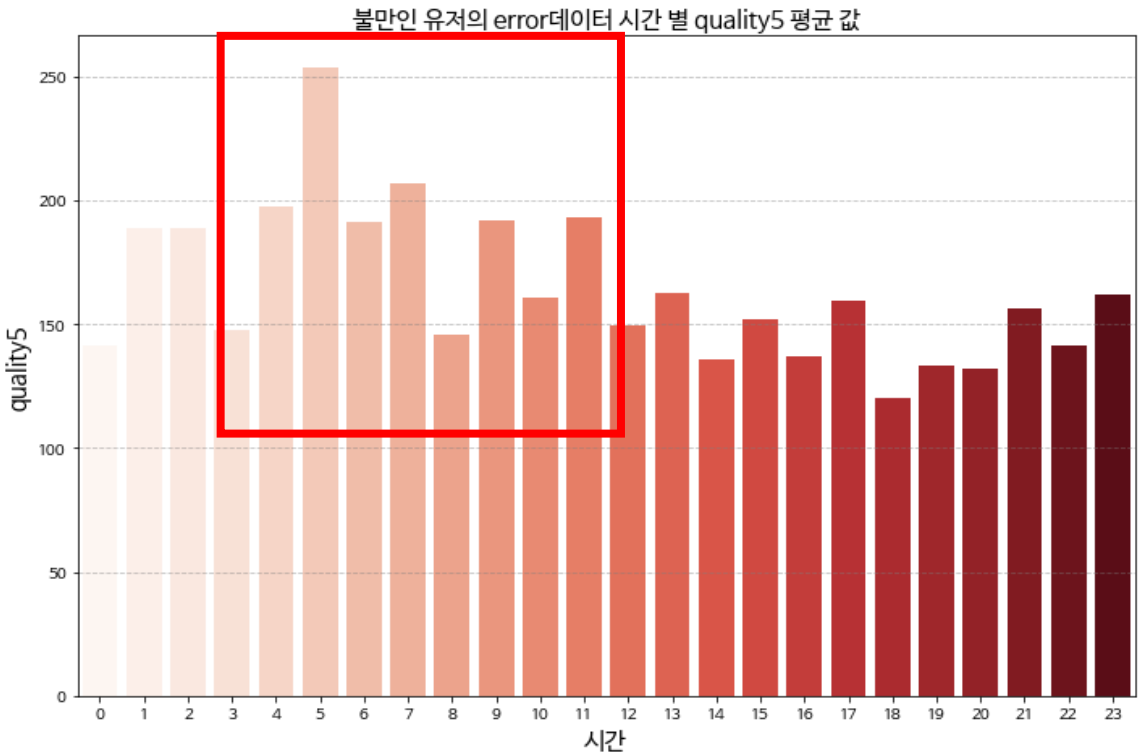


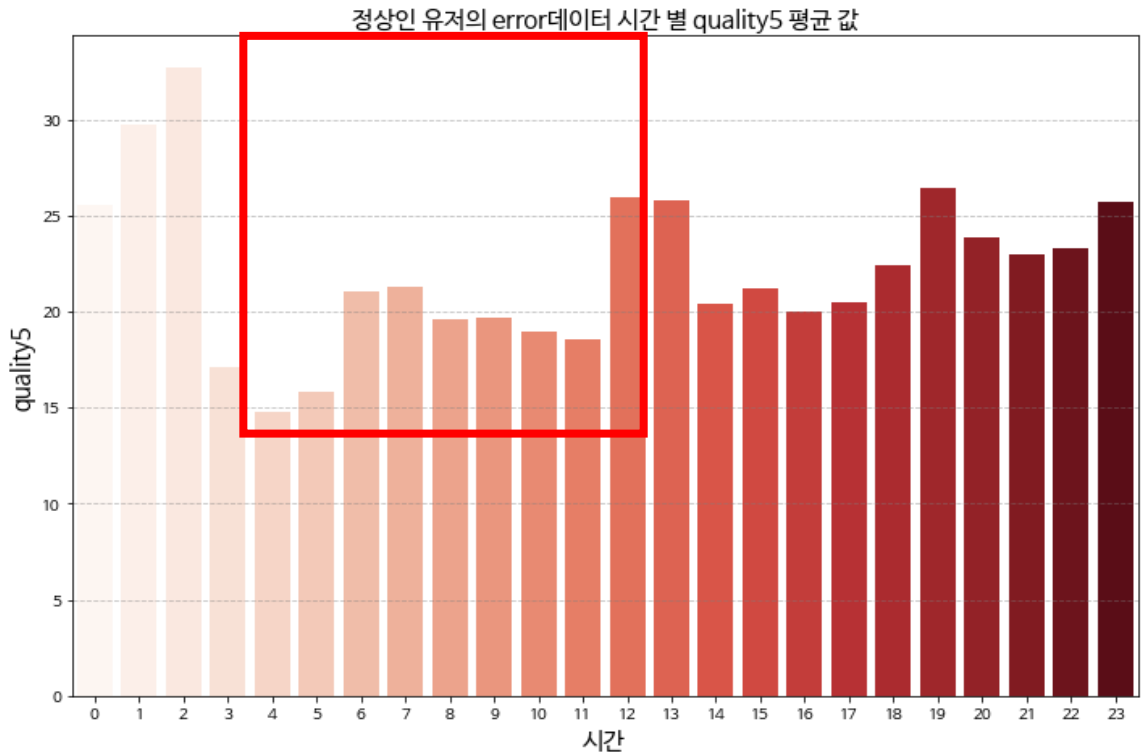
- error type별로 quality5의 수치를 확인했을 때 특정 error type은 정상과 불만의 차이가 확연히 드러남
- 9,27,28,30 error type은 불만인 유저의 경우에서 높은 비율로 발생
- 오히려, 정상인 유저에서 자주 발생하는 3, 10, 35과 같은 error type도 확인
- 시스템에 치명적인 오류를 발생하는 error와 정상 작동 되고 있지만 경고성 error를 발생하는 유형으로 구분된 것으로 추측

02 데이터 분석

모델 기반 불만 유형 분석 : catboost feature

error data와 quality data간의 관계를 quality5의 값을 통해 확인

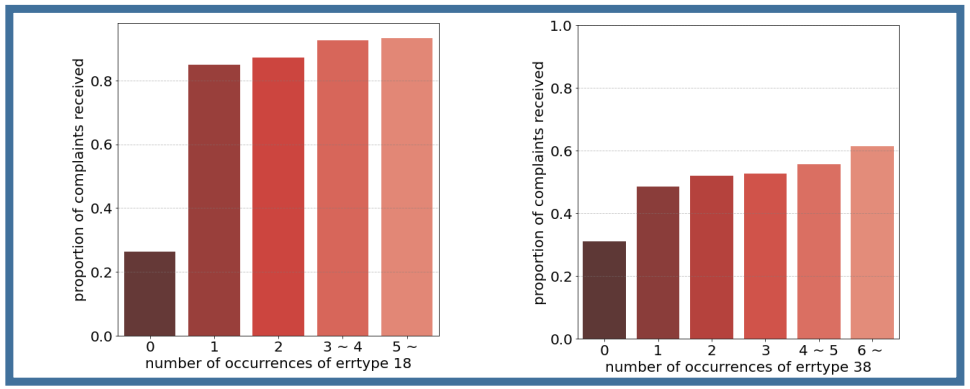


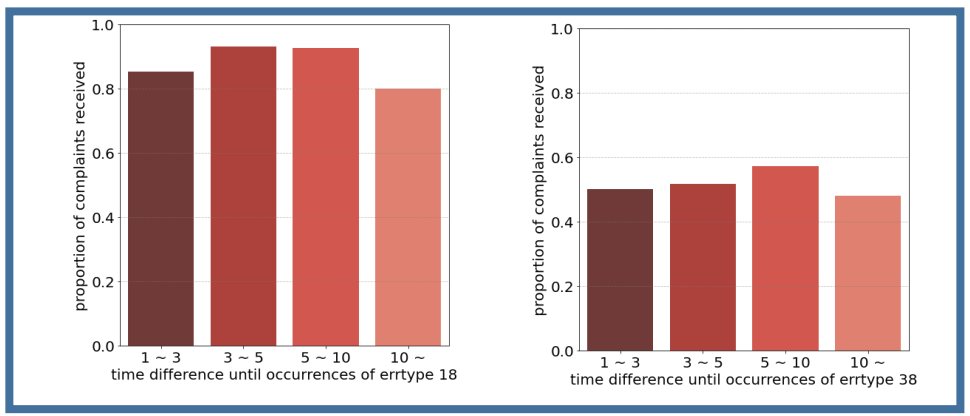


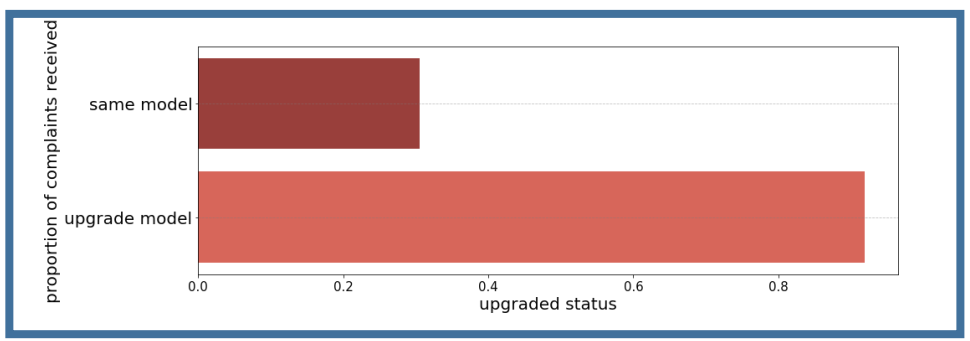
- 전반적으로 quality5의 평균값이 불만인 유저에서 높게 나타나는 것을 확인
- 불만인 유저의 경우 새벽 부터 점심시간 전까지 error data의 quality5의 값이 타 시간대보다 높은 것을 확인
- 저녁과 밤시간대에서는 정상인 유저가 quality5의 값이 타 시간대보다 높은 것을 확인

02 데이터 분석

모델 기반 불만 유형 분석 : XGB feature







17: split_type

- 한 유저당 18번 errtype이 발생한 횟수
- 다른 errtype과 비교하여 한번만 발생하여도 불만을 제기하는 확률이 높음

533: split_typediff

- 18번 errtype이 발생했을 때, 이전 errtype 발생 시점과의 시간 차이를 나타낸다.
- 발생 간격은 너무 짧거나 긴 경우에는 오히려 불만을 제기할 확률이 상대적으로 낮음
- 3시간에서 10시간과 같은 애매한 간격은 사용자의 불만을 가중시키는 것을 확인

566: model_nm

- 두번째로 사용한 모델 명을 나타내는 변수
- 모델을 업그레이드 했음에도 불구하고 에러가 발생하면 불만을 제기하는 확률이 높음

03 불만 유형을 통한 비즈니스 분석

비즈니스 인사이트 및 솔루션 제시

잠재 불만 유저

condition에 따라 불만을 표출하는 강성 유저

시스템 변화 분석을 통한 고객이탈 방지 목표

- condition에 따른 장비 사용량 예측을 통해 '이동 기지국'과 같은 커버리지 능력 확장 필요
- 버전 업데이트 시 문제 상황 별 솔루션 가이드 제공
- 시스템 품질 변화 예측을 통해 트래픽 분산 시스템 개발
- 자주 발생하는 특정 에러를 해결하는 자가치유 시스템 개선

state에 따라 불만을 표출하는 약성 유저

선별적 조치를 통해 만족 고객으로의 전환 필요

- 사용자 유형 뿐만 아니라 지역, 시간과 같은 외부요인을 고려하여 모델 교체 시기를 사용자에게 선제적으로 제안
- Main Controller와 같은 quality5 부품의 자가 진단 키트 제공
- 품질 저하가 예상되는 펌웨어에 대해 일괄적인 업데이트 실시
- 만족 고객의 시스템 품질 분석을 통한 자동 품질 최적화 제공 가능

03 불만 유형을 통한 비즈니스 분석

결론

결론

- 시스템 품질(state) 변화(condition)에 따른 사용자 불편 예지 모형 구축
  - 시스템의 품질과 장비의 상태를 의미하는 state 그룹과 사용자의 시간 환경을 의미하는 condition 그룹의 파생변수 활용
  - 다양한 불만에 대응하기 위한 4가지 머신러닝 모델 활용

- 사용자의 불만 유형을 강성 불만과 약성 불만으로 분류
  - 변수 중요도를 통해 동일한 불만에 대한 다양한 요인 탐색
  - 강성 불만과 약성 불만을 함께 고려할 수 있는 앙상블 모델 개발
  - 잠재 불만 유저의 유형 분류에 활용 가능

- 사용자의 불만 유형에 따른 불편 감소 방향 제시
  - 강성 불만과 약성 불만을 나누어 솔루션을 제시함으로써 저비용 고효율의 불편 조치 가능
  - 고객 이탈방지와 고객 만족도 상승의 효과를 기대

04 팀원 소개 및 역할 분담

전명준

연세대학교 산업공학과

- 모델 아키텍처 설계
- error data 파생 변수 생성
- 모델 성능 평가 해석

박민영

성균관대학교 통계학과

- 모델 파라미터 튜닝
- quality data 파생 변수 생성
- 사용자 불만 유형 분석

원하연

성균관대학교 통계학과

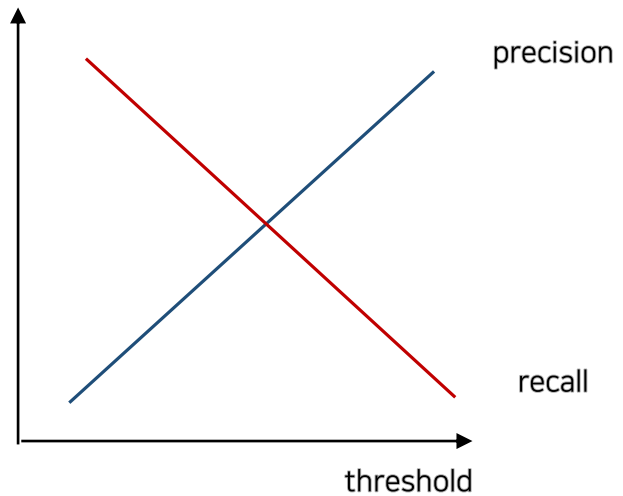
- 데이터 탐색 및 시각화
- error data & quality data 관계 분석
- 발표 자료 정리

Q̄nA

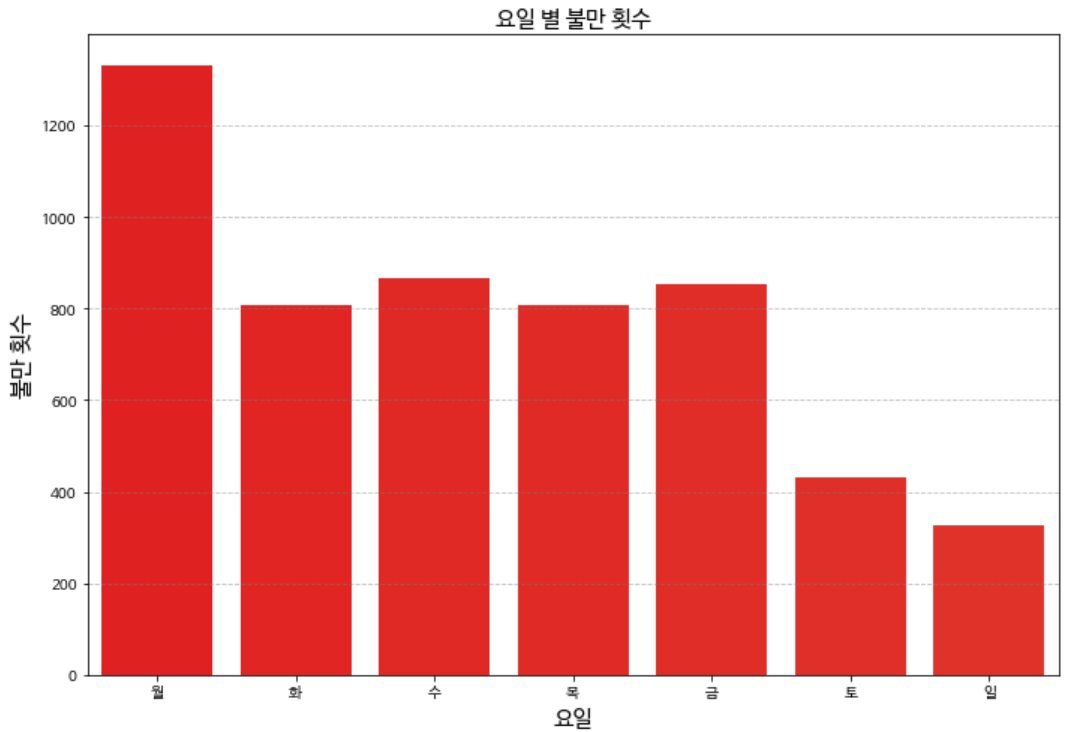
Appendix recall과 precision을 이용한 불만 유형

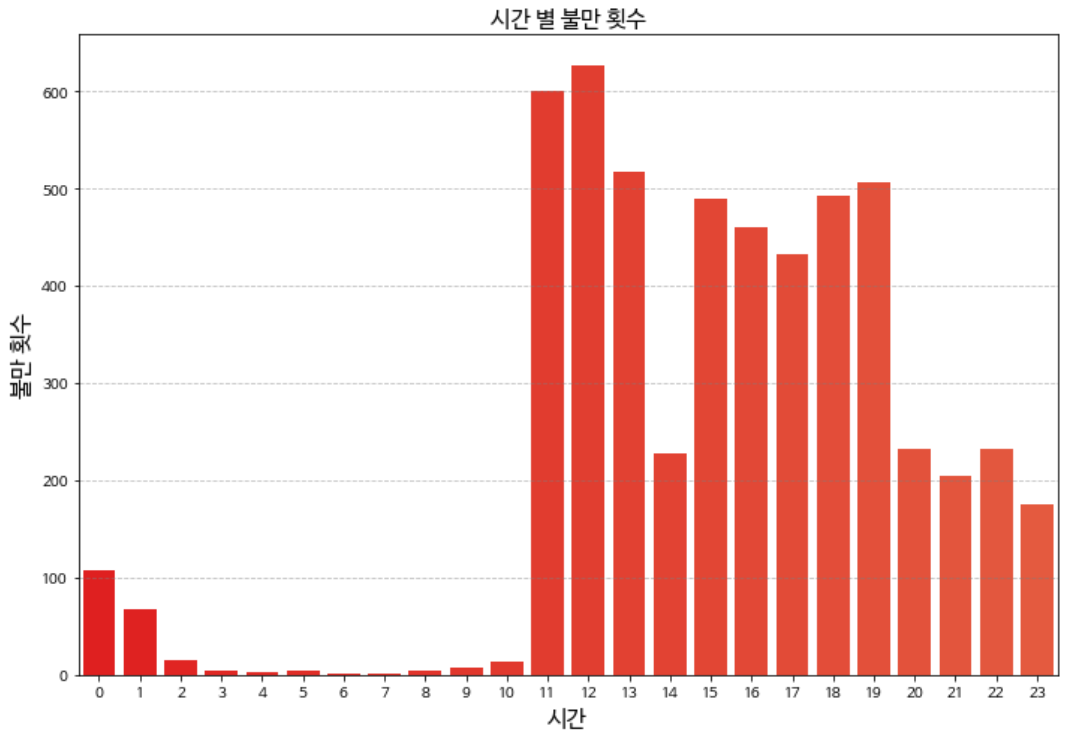
recall과 precision에 집중한 이유

- AUC는 모든 경우의 threshold에 대해 안정적으로 좋은 결과가 나올 때 높은 값을 유지하고, f1 score의 경우 recall과 precision의 균형이 잘 맞춰진 경우에 높은 값을 가진다.
  즉, AUC와 f1 score가 높은 경우 실제 라벨을 안정적으로 맞추는 모형이라 볼 수 있다.
- recall과 precision의 경우는 우측 그림과 같이 trade-off 관계를 가지는데,
  주로 imbalanced data 상황에서 recall이 중요하게 사용된다.
- recall은 실제1(불만)을 라벨로 가지는 데이터 중 1로 예측된 비율을 뜻하여,
  약한 positive를 얼마나 잘 잡아내는지를 평가하는 척도이다.
- 따라서 recall이 높은 XGB모형은 약한 불만을 잘 잡아내는 모형으로 볼 수 있고,
  precision이 높은 cat은 그 반대 의미인 강한 불만을 잘 잡아내는 모형으로 볼 수 있다.

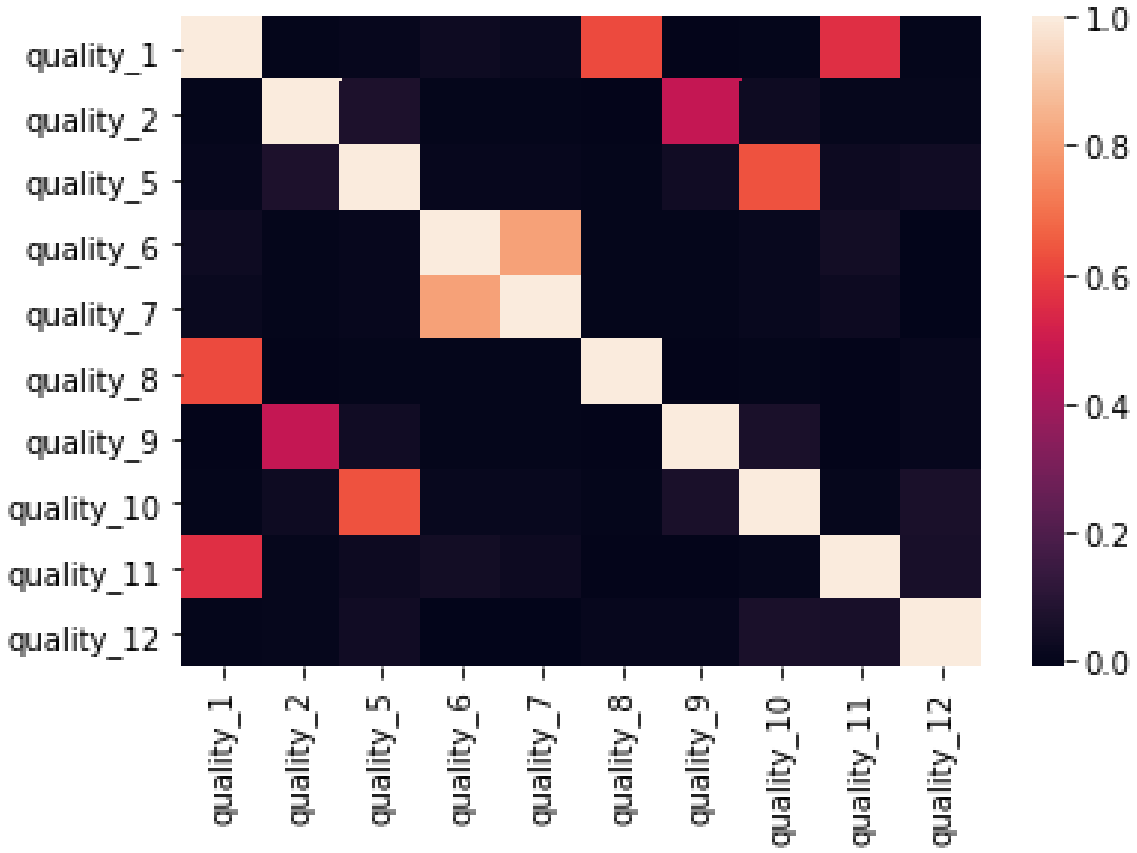


Appendix train 데이터 내 불만 시점





Appendix quality data 내 상관관계



quality	분산
1	0.2268
2	49075
5	5574183
6	925
7	95472
8	1.68
9	1621044
10	334293144
11	0.17
12	0.11

Appendix 한계점 및 해결방안 제언

- **문제에 비해 학습 데이터셋이 무거움**
  - 유저에 대한 시간 정보가 담긴 3차원의 데이터를 2차원으로 압축하다 보니 학습 데이터의 dimension 자체가 큼
  - 가지고 있는 데이터셋에 비해 맞추고자 하는 정답 자체의 속성이 너무 적다고 생각
  - 성능이 조금 떨어지더라도 중요하다고 판단되는 변수를 활용
  - 정답의 유형을 다시 분류한다면 불만 예측의 성능 또한 올라갈 것으로 판단
- **뿐만 아니라 모델에 대한 학습 시간도 길다**
  - cv간 성능 폭이 컸기 때문에 오히려 각 cv 별로 오버피팅을 시켜 cv별 모델을 활용
  - 모델의 능력치를 최대한으로 활용하자고 생각, label이 일관성이 떨어지는 점을 보완
  - 에러가 발생하는 로그 데이터 뿐만 아니라 정상 데이터도 주어지는 일정 간격의 로그 데이터를 활용한다면 시간 속성 정보를 활용하기 쉬울 것임
  - 예를들어 시간속성과 시스템 속성을 동시에 가지는 3차원 데이터셋을 통해 lstm, transformer 계열의 딥러닝 모델을 활용할 수 있음
- **문제가 발생하는 사전 시점과 사후 시점의 정보가 혼재**
  - 대회의 task는 불만을 분류하는 문제 임으로 해당 문제는 고려하지 않음
  - 사전, 사후 시점을 고려할 수 있는 문제라면 T+1시점의 예측을 통해 불만 또는 고장을 예측할 수 있을 것으로 판단