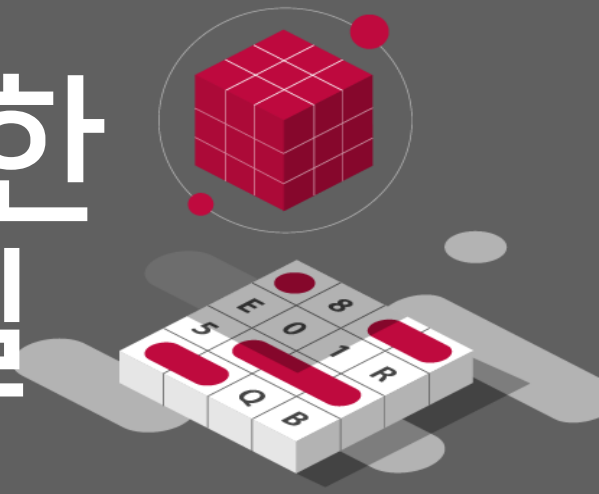
시스템 품질 변화로 인한 사용자 불편 예지 모델



[Team asdf] – 김명선, 이지훈

LG AI Research

DACON
DATA TO VALUE

팀 구성 및 역할

성명	담당업무	학력	수상 내역
김명선	- 팀장 - error 데이터 분석 - 사용자 불만 원인 분석 - 예측 모델 설계	GIST 전기전자컴퓨터공학부 석사과정	▪ 전력 데이터 활용 신서비스 개발 경진대회 최우수상 ▪ 대구 빅데이터 분석 경진대회 장려상 ▪ AI특화 창업 경진대회 '꿈꾸는아이' 3등상
이지훈	- Quality 데이터 분석 - Quality 데이터와 error 데이터 관계 해석 - 사용자 불만 원인 분석 - 비즈니스 분석		

Index

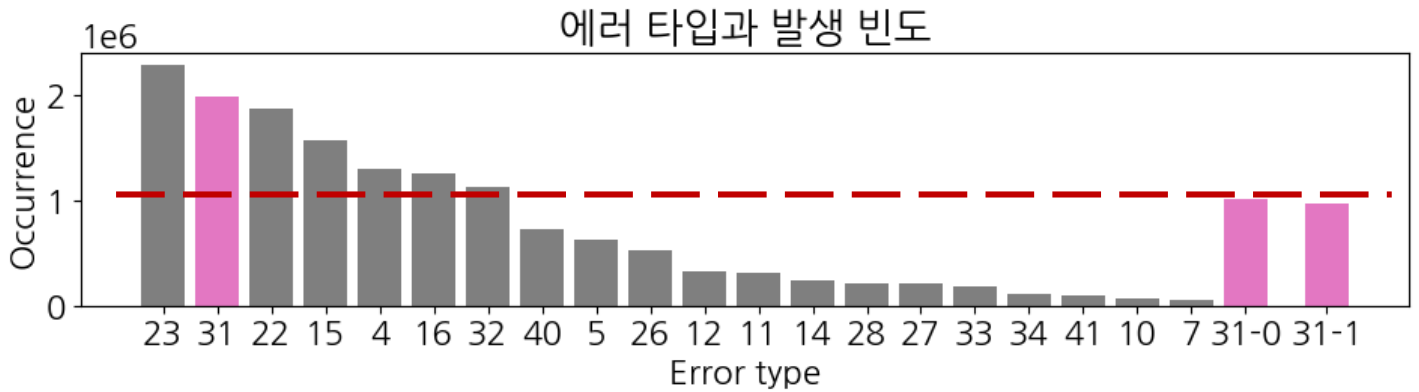
데이터 분석 (EDA)

- Error 데이터 분석
- Quality 데이터 분석
- Error-quality 관계 분석

Error 데이터 분석 – Error type 과 error code

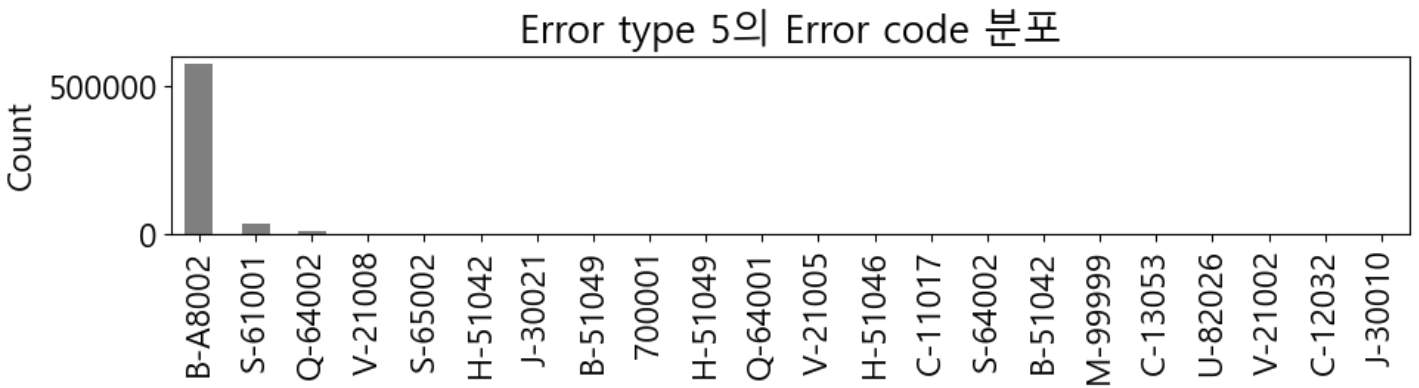
Error type 세분화의 필요성

- 41개의 error type
- 특정 error type보다 높은 빈도를 보이는 code가 다수 존재



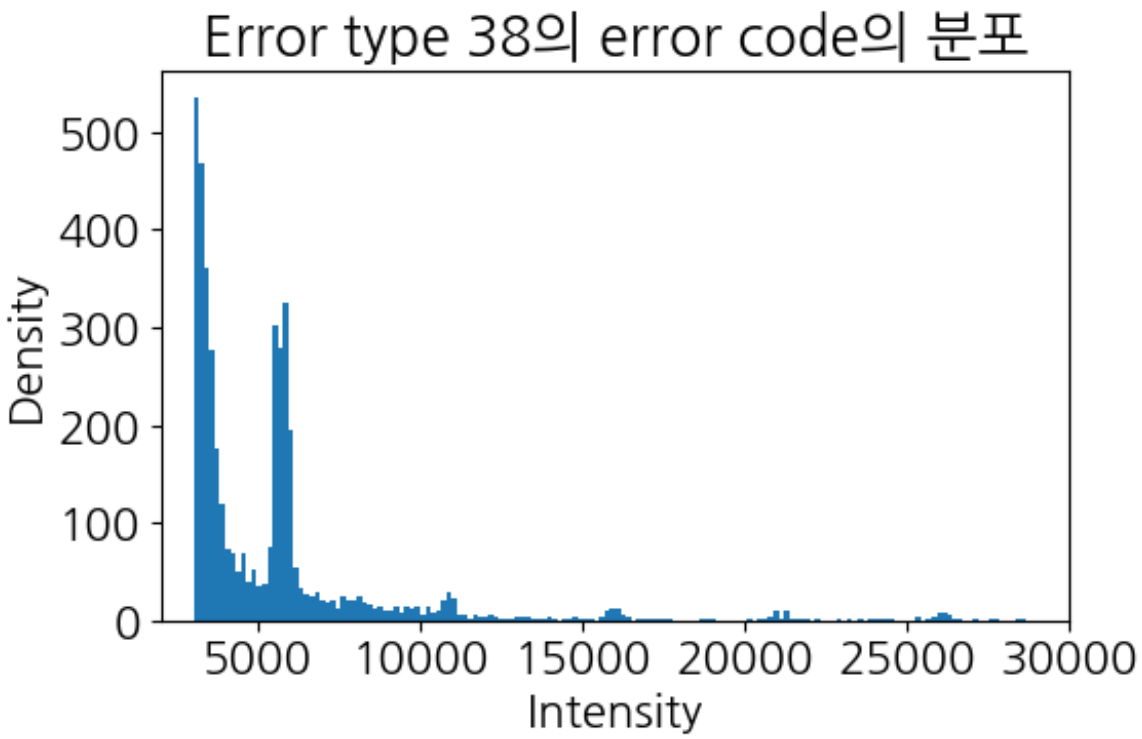
Error code 필터링의 필요성

- 빈도수가 낮은 error code가 다수 존재



Error 데이터 분석 – Error type 과 error code

Error type 38의 error code



- 5000개 이상의 정수
- 30001부터 4875823까지 값을 가짐
- Numerical variable로 추측됨
- 이외의 error type의 error code는 categorical variable로 추정됨

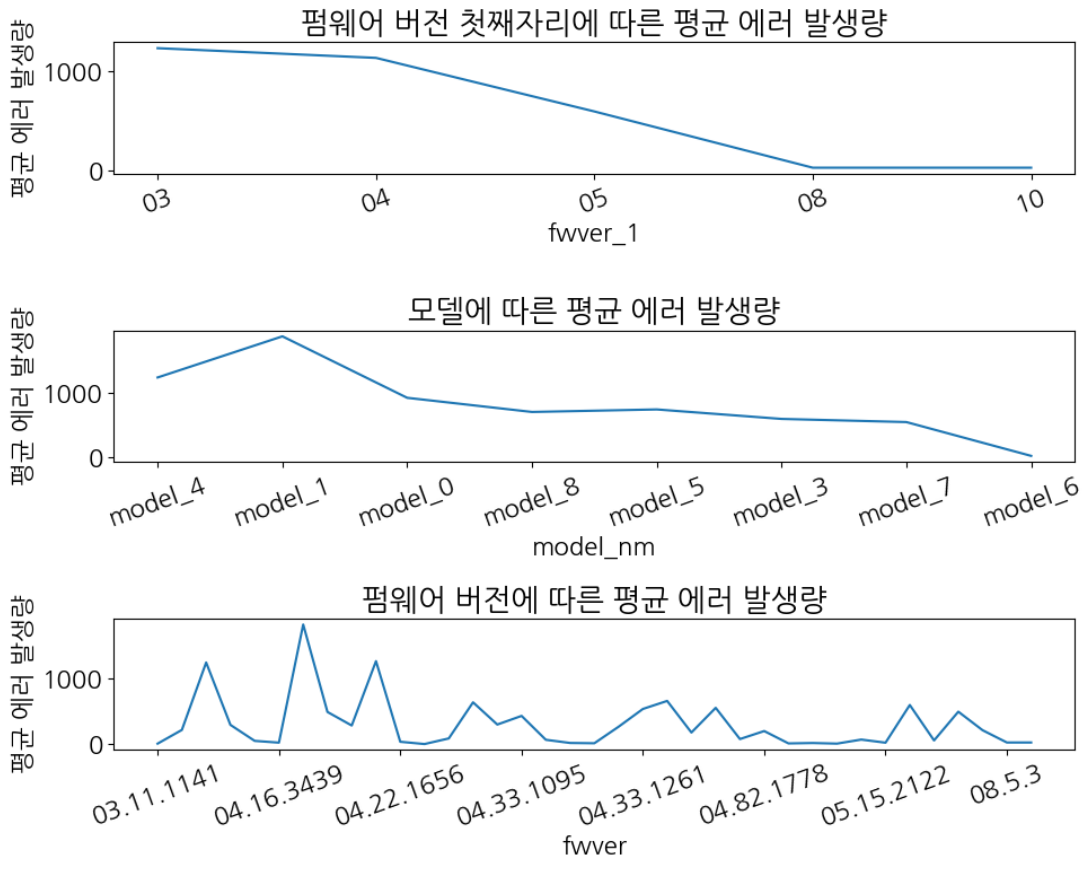
Error 데이터 분석 – model_nm과 fwver

Model_nm과 fwver에 따른 error 발생량

- 버전이 업데이트 되어감에 따라 평균 error 발생량이 낮아지는 추세
- Model_nm 순서

순서	Fwver	Model_nm
1	03.11.x	4
2	04.16.x	1
3	04.22.x	0
4	04.33.x	2
5	04.73.x	8
6	04.82.x	5
7	05.15.x	3
8	05.66.x	7
9	8.x, 10	6

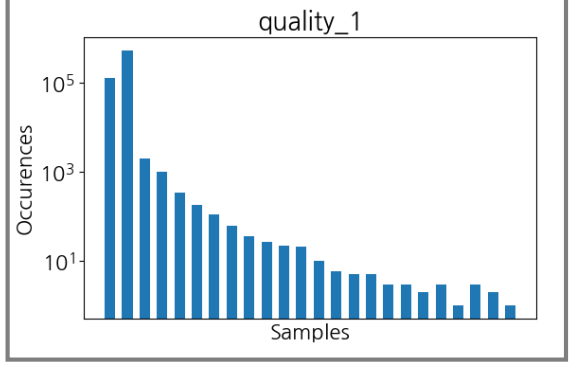
세분화



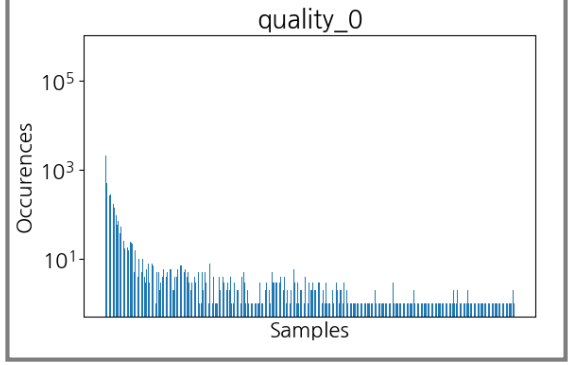
Quality 데이터 분석 – Quality 분포

Quality의 타입 분류

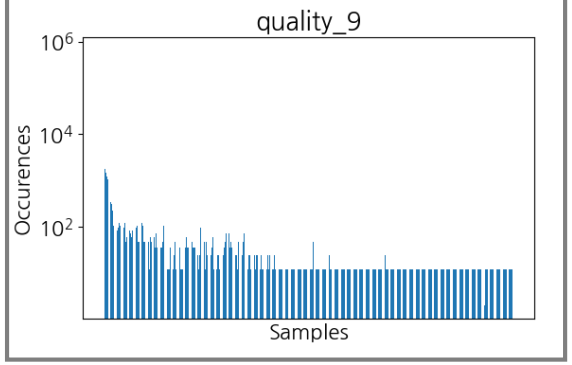
Type A.



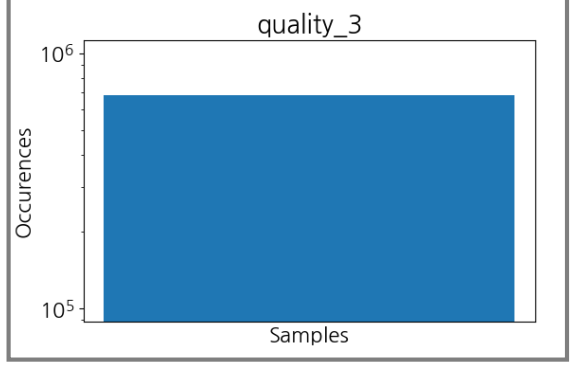
Type B.



Type C.



Type D.



- Type A. (Quality - 1, 8, 11, 12) : 특정 quality 값을 포함, categorical 변수로 추정
- Type B. (Quality - 0, 2, 5, 6) : Quality 값 중 -1을 포함, 지수 분포로 근사 가능
- Type C. (Quality - 7, 9, 10) : Quality 값 중 -1을 포함하지 않음, 지수 분포로 근사 가능
- Type D. (Quality - 3, 4) : 하나의 quality 값(0)만 확인 가능

Quality 데이터 분석 - Quality 분포

Quality 통계적 수치

Type	A				B				C			D	
Quality	1	8	11	12	0	2	5	6	7	9	10	3	4
mean	0.0	0.1	0.0	0.0	12.0	12.0	103.4	3.0	30.0	36.9	859.6	0	0
max	0.3	0.4	0.0	0.0	95.2	96.8	1417.3	30.9	78.5	204.9	3300.0	0	0
std	0.0	0.1	0.0	0.0	13.6	13.8	228.2	6.5	22.1	55.8	911.9	0	0

- Type B, C는 상대적으로 높은 수치를 보임

Quality 중 -1의 발생 빈도

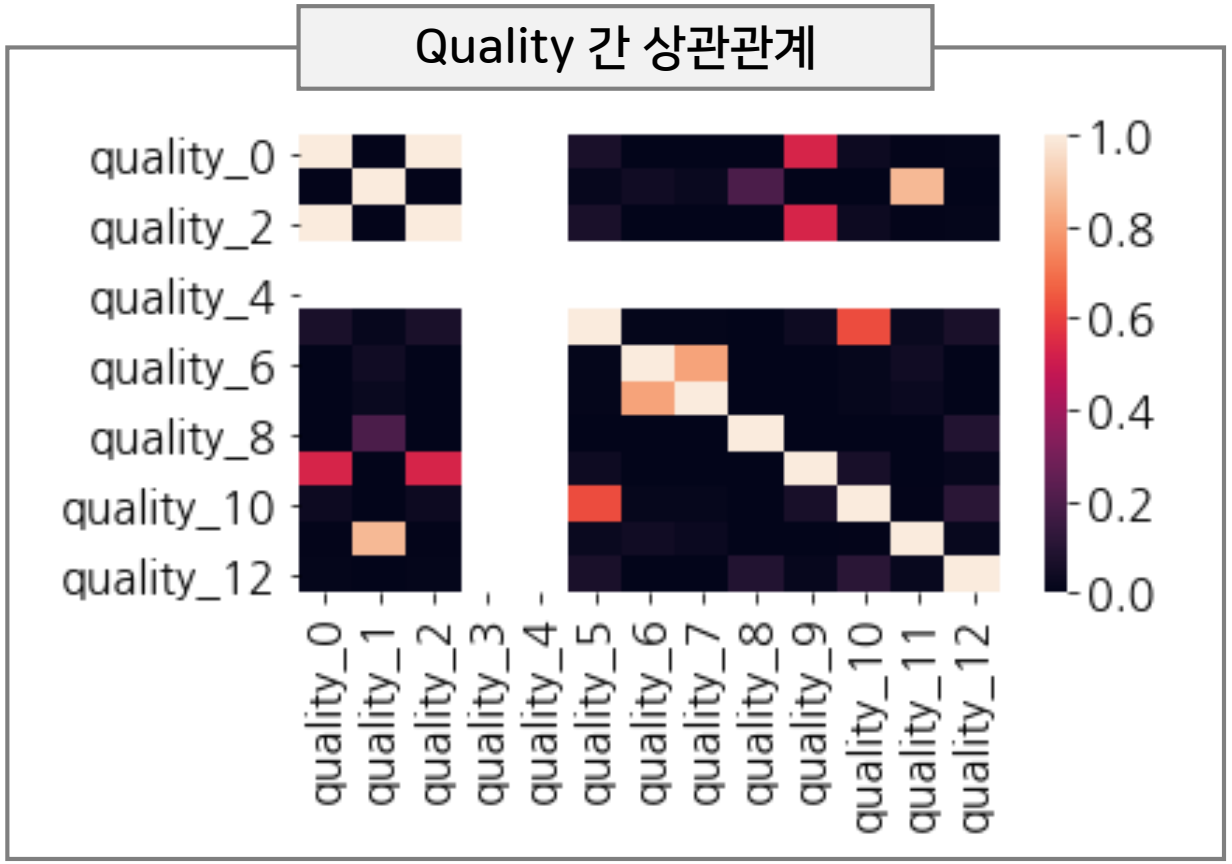
- Quality 0, 1, 2, 5, 6, 11에서 동시간대에 발생
- 한 사용자당 평균적으로 23.5회 발생, 최대 1273회 발생

Quality 데이터 분석 – Quality 간 상관 관계 도출

Quality 간 높은 상관성

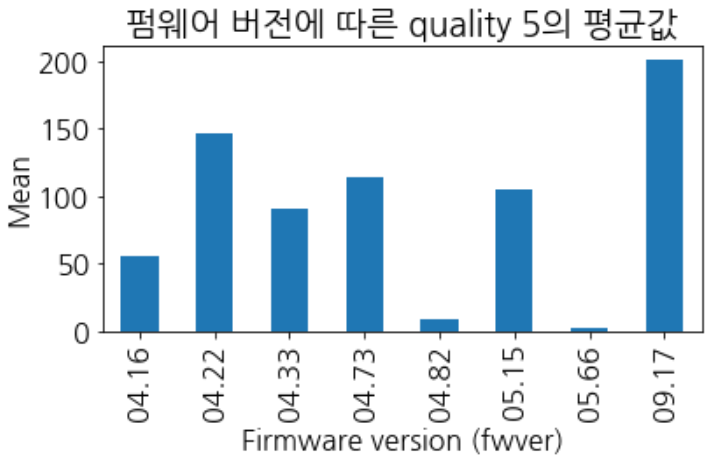
- 상관성이 높은 Quality 간 유사한 특성을 가지는 측정 지표로 추정

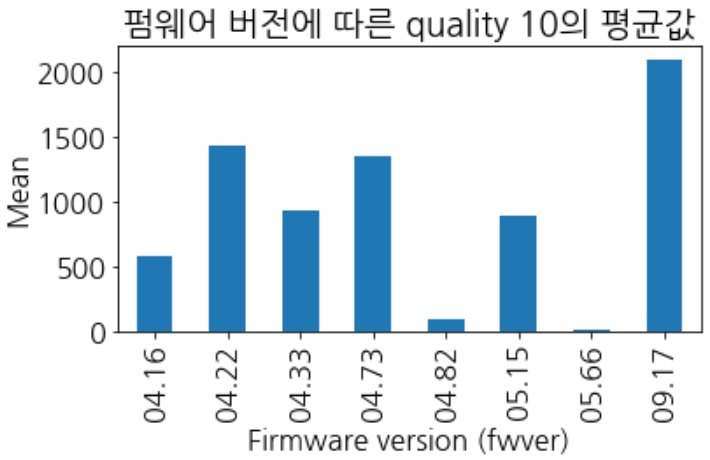
  - ✓ Quality Corr. > 0.5: (0 & 9), (2 & 9)
  - ✓ Quality Corr. > 0.6: (5 & 10)
  - ✓ Quality Corr. > 0.8: (1 & 11), (6 & 7)
  - ✓ Quality Corr. > 0.9: (0 &2)

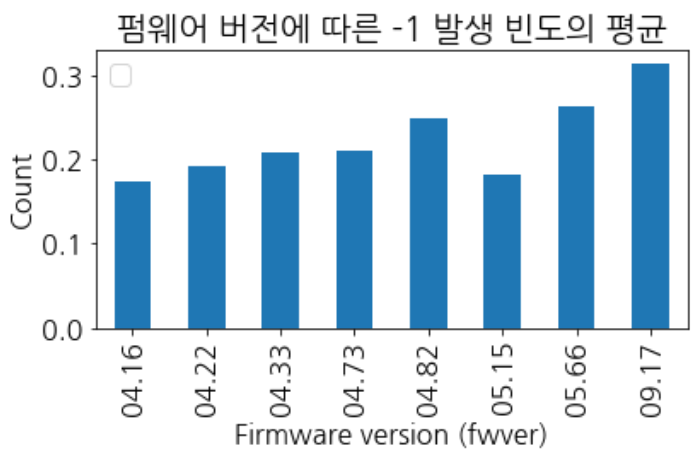


Quality 데이터 분석 – Fwver에 따른 quality 통계

Fwver에 따른 quality의 통계적 수치 확인







- Fwver에 따라 quality 5, 10의 평균값 경향이 유사하게 나타남
  - Quality 5와 10은 유사한 지표로 추정됨
- Fwver가 증가할수록 -1의 발생 빈도가 대체적으로 높아짐
- 특히, fwver 09.17에서 quality 평균과 -1의 발생 빈도가 가장 높게 나타남

Quality 데이터 분석 – Quality의 엔트로피 접근 방법

Exponential entropy

- 비 식별 데이터에 관한 정보량 해석
  - 2시간 단위 측정 데이터 (12 samples)
  - 개별 quality 값의 지수 분포를 확인하여
    근사 값을 이용함으로써 엔트로피 값을 도출함

- Exponential probability distribution

$$f(x) = \lambda e^{-\lambda x}, \qquad for\ x \geq 0$$

- Exponential entropy

$$H(X) = -\int_0^\infty \lambda e^{-\lambda x} \log \lambda e^{-\lambda x}\, dx$$

$$= -\log \lambda \int_0^\infty f(x)\, dx + \lambda E[X]$$

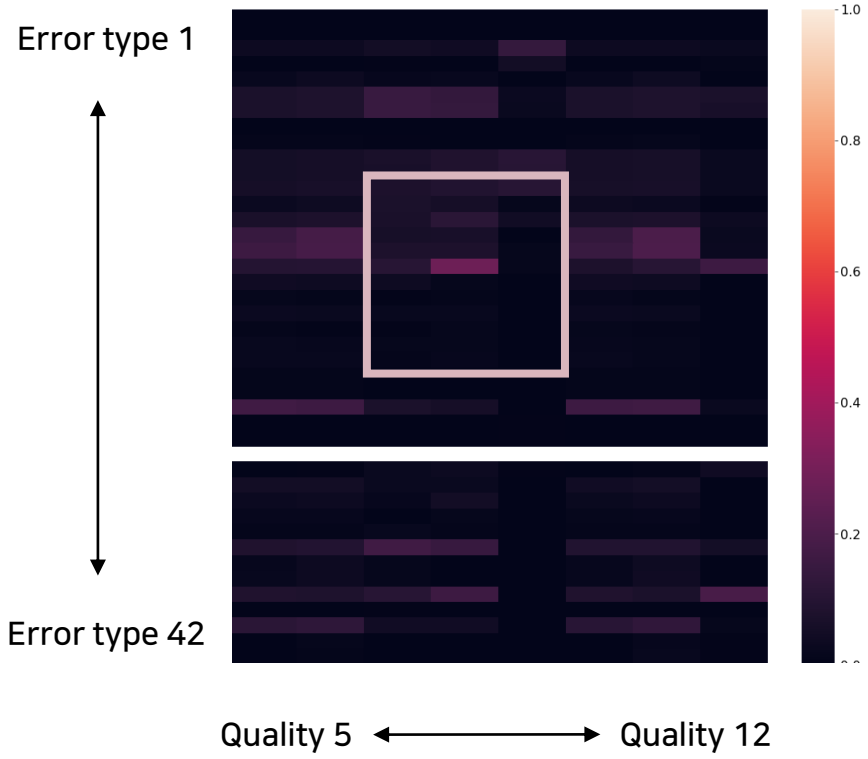
$$= -\log \lambda + 1 \ \ [nat]$$

✓ λ: $Random\ variables$
✓ X: $Exponential\ distribution$

Error-quality 관계 분석 – Quality의 엔트로피 해석

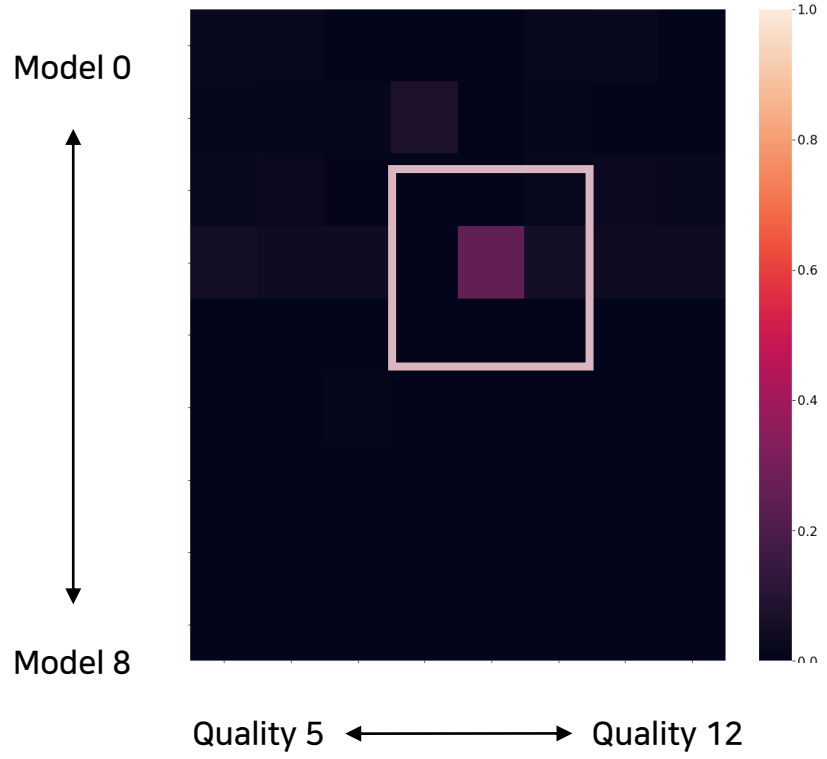
Err type과 quality의 관계 해석

- Error type 17번 & Quality 8번 (Corr. ≒ 0.4)



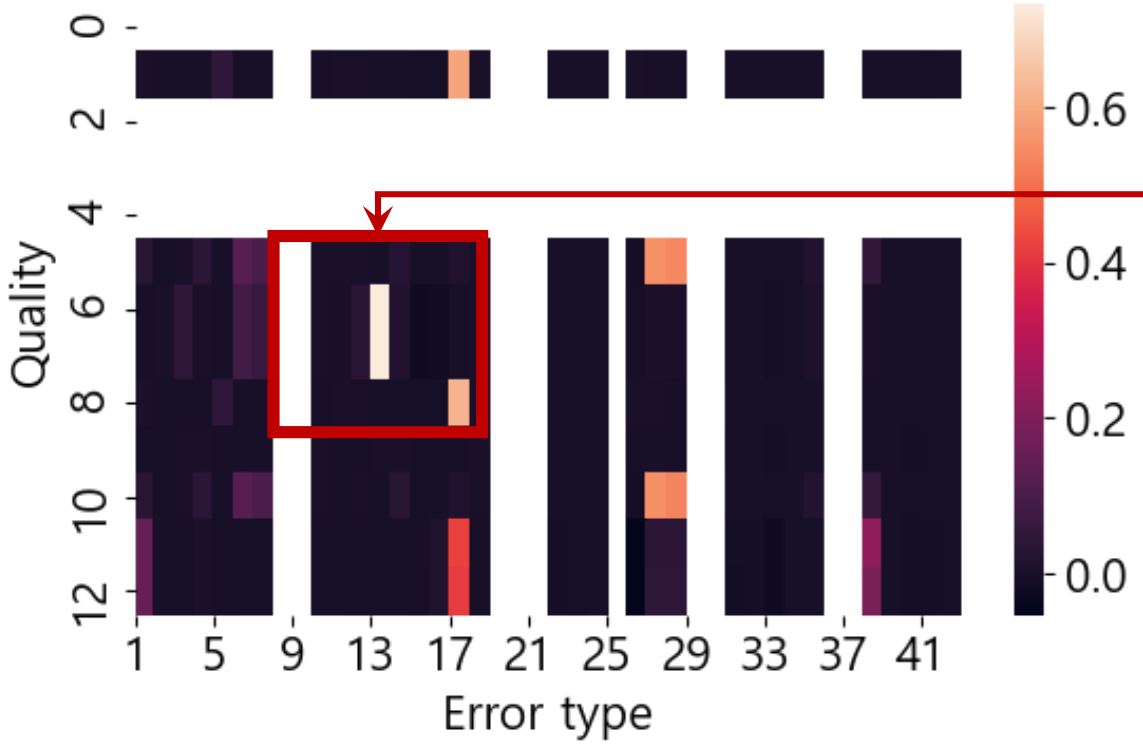
Model_nm과 quality의 관계 해석

- Model 3번 & Quality 9번 (Corr. ≒ 0.4)



Error-quality 관계 분석 – 상관 관계

Quality와 error type의 상관관계



- Quality의 일별 평균과 error type의 일별 발생량의 상관 관계 분석
- Error type 13과 quality 6, 7의 상관 관계가 높음 (0.73)
- 해당 error type을 측정하는 지표가 6, 7로 판단됨
- Error type별 주관하는 quality 측정 지표가 존재할 것으로 추정됨

사용자 불만 접수 원인 분석 (Problem)

- 요약
- Error 분석
- Version 분석
- Quality 분석

사용자 불만 접수 원인 – 요약

ERROR

- ✓ 높은 빈도의 error 발생
- ✓ 특정 error type 발생
- ✓ 특정 error code 발생
- ✓ Error type 38 내 높은 error code 발생

VERSION

- ✓ 특정 model_nm, fwver 의 소유
- ✓ Model_nm, fwver의 업그레이드/다운그레이드/변경

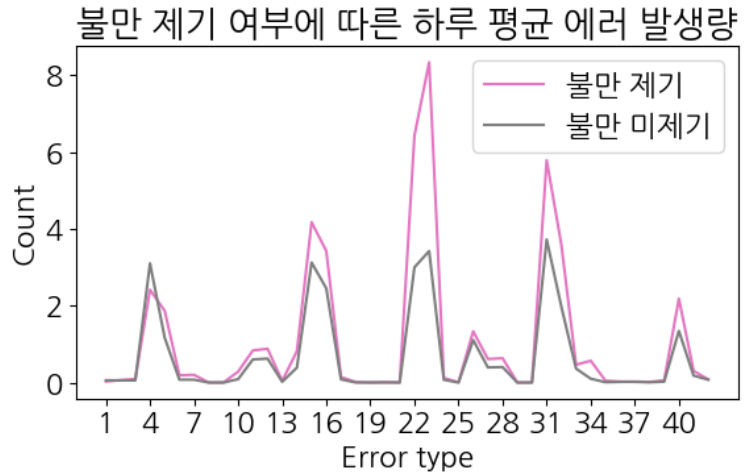
QUALITY

- ✓ -1 값 발생
- ✓ Quality의 높은 수치
- ✓ 2시간 이내에 quality 변화

사용자 불만 접수 원인 – Error 분석

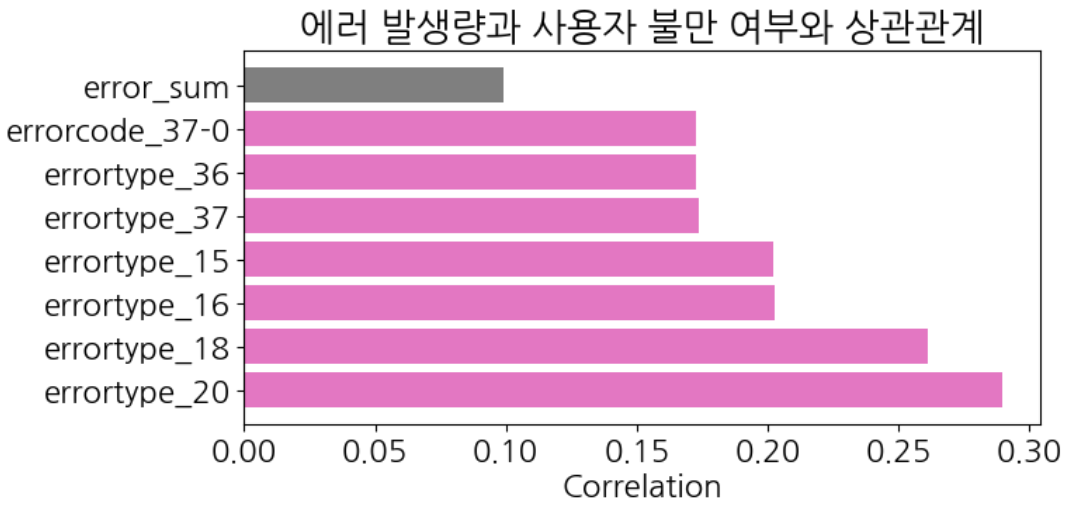
하루 평균 error 발생량 분석

- 불만을 제기한 사용자의 하루 평균 error 발생량이 제기하지 않은 사용자보다 대부분의 error type에서 높음



Error와 불만 여부의 상관관계 분석

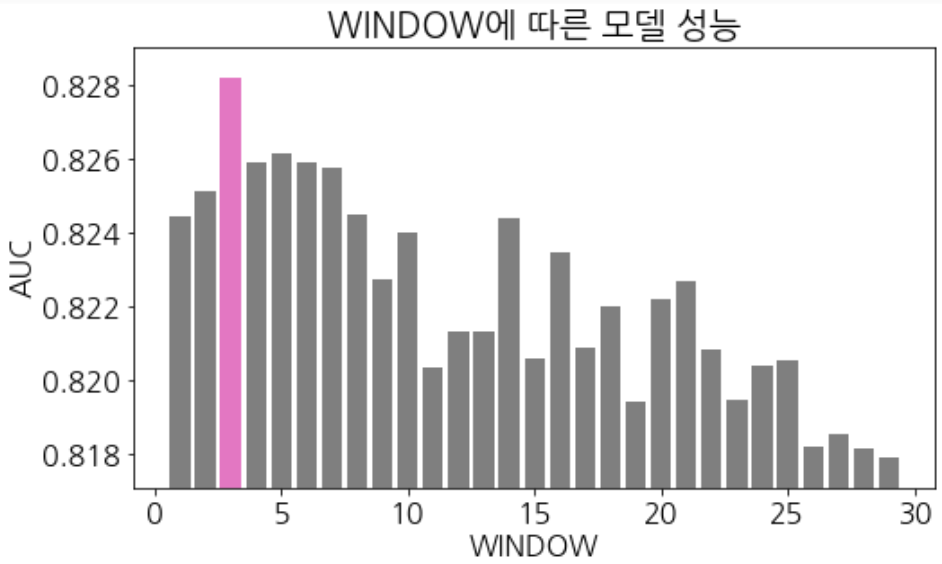
- 모든 error type의 합계보다 error type이나 code를 개별적으로 보는 것이 불만 여부와 상관성이 높음
- 특히, error type 20이 불만 접수와 가장 상관성이 높음
- Error 37의 경우 Error code 0이 사용자 불만 여부의 주된 요인으로 추정됨



사용자 불만 접수 원인 – Error 분석

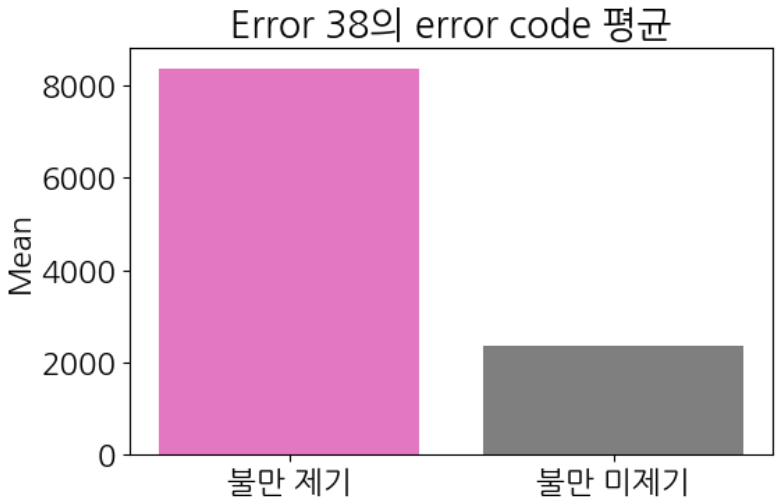
3일간의 error 발생량

- 한 달 동안 발생한 error의 총 합보다 3일의 window마다의 합계가 관련성이 큼
- 산발적 error 발생 대비 error 발생의 집약도가 클 수록 영향력을 보이는 것으로 추정됨



Error type 38의 error code 평균

- 불만을 제기한 사용자의 error type 38의 code 평균이 제기하지 않은 사용자에 비해 상당히 큼

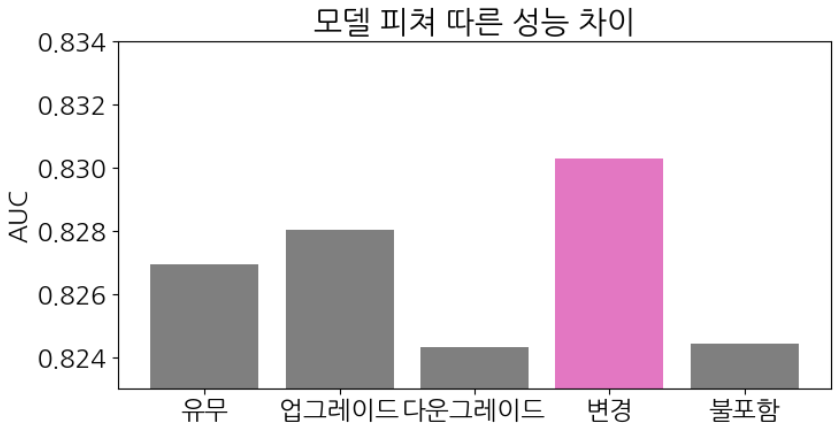


사용자 불만 접수 원인 – Version 분석

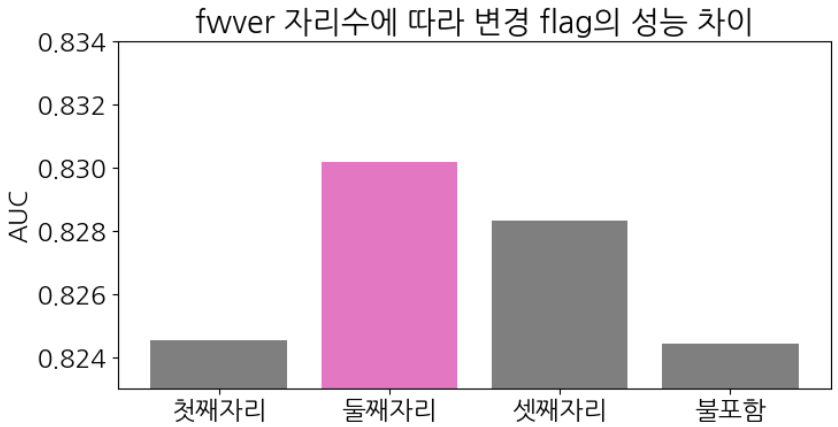
Fwver 변경에 따른 불만 제기 비율

펌웨어 첫째자리 업그레이드	펌웨어 첫째자리 다운그레이드	펌웨어 첫째자리 변경	모델 업그레이드	모델 다운그레이드	모델 변경	펌웨어 업그레이드	펌웨어 다운그레이드	펌웨어 변경
80%	68%	68%	95%	67%	91%	39%	66%	39%

Fwver 피쳐에 따른 성능 차이



- Model_nm는 변경이 가장 높은 영향력을 보임

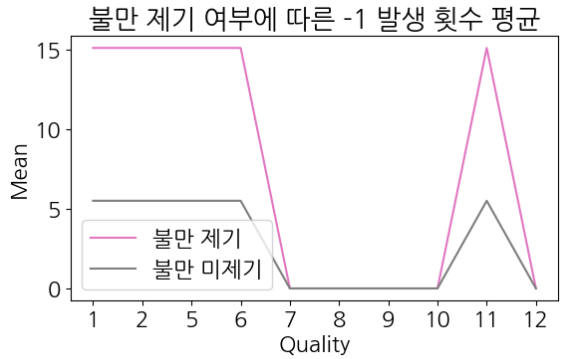


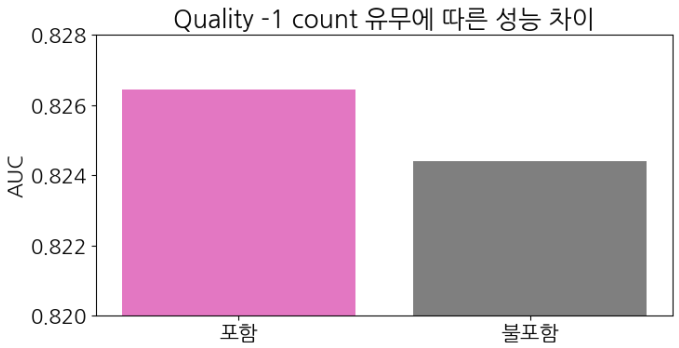
- 둘째 자리에 해당하는 model_nm 변경 flag가 fwver 변화 중 가장 높은 영향력을 보임

사용자 불만 접수 원인 – Quality 분석

Quality의 -1 발생 빈도

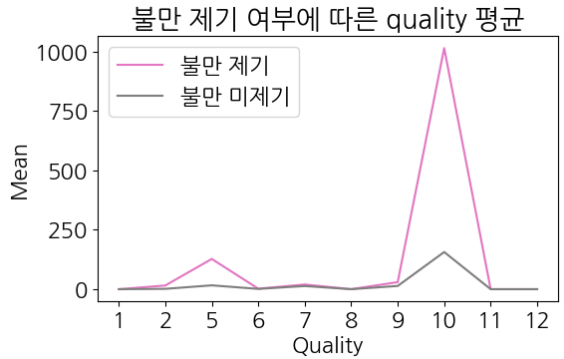
- 불만 제기 사용자가 -1이 높은 빈도로 발생함

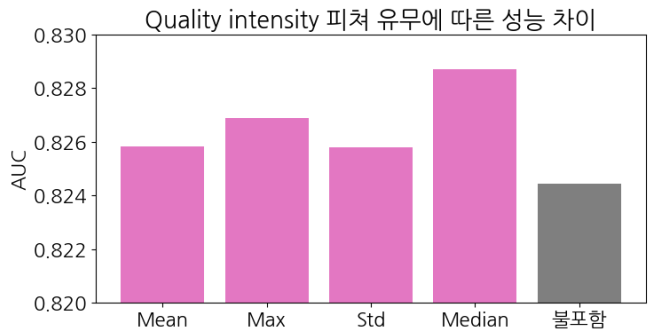




Quality의 높은 수치

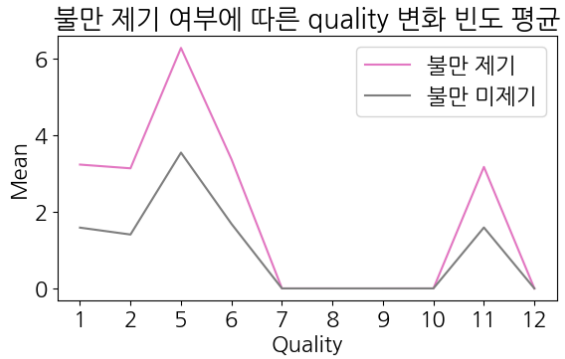
- 불만 제기 사용자가 quality 값이 크게 나타남

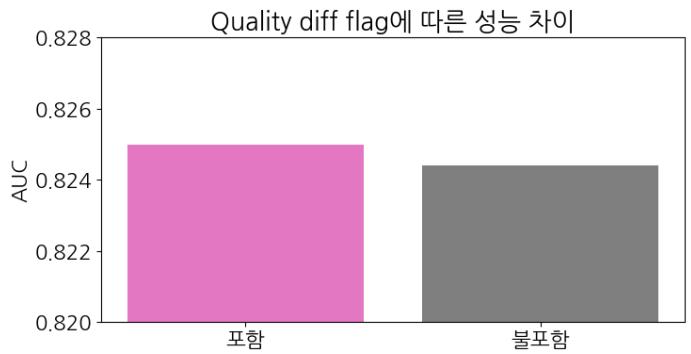




2시간 내 quality 변화

- 불만 제기 사용자의 2시간 내 quality 변화 빈도가 높음





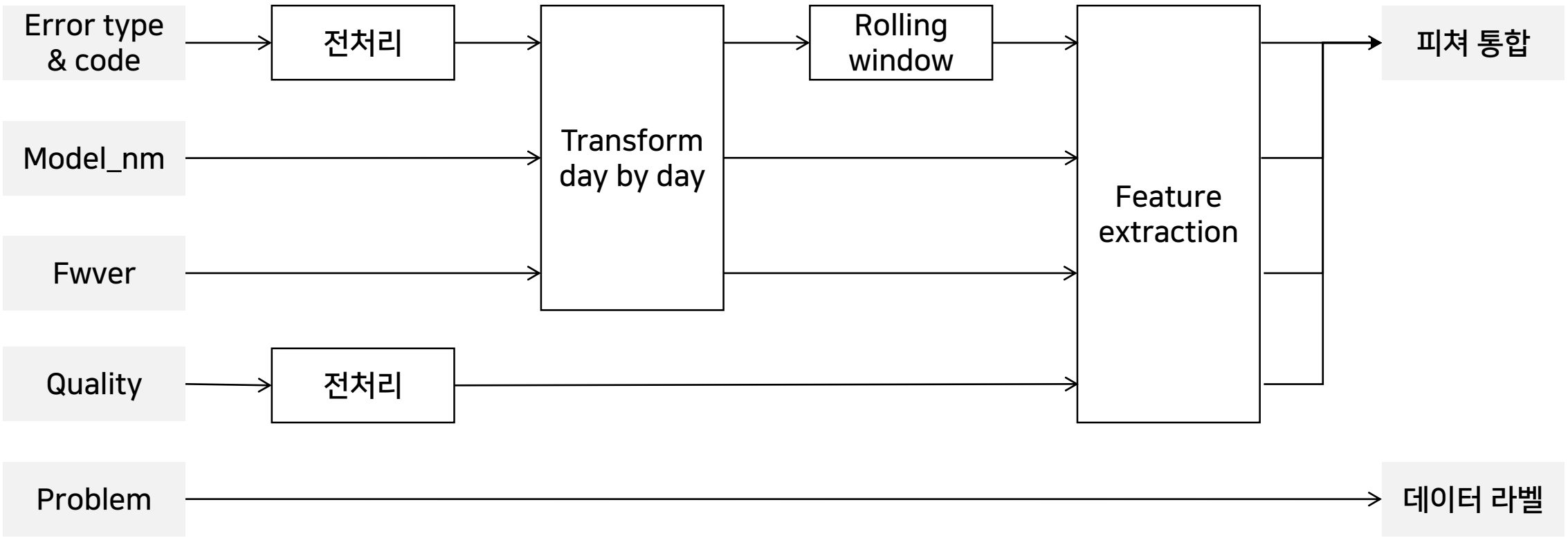
결과 분석

(Result)

- 제안 모델
- 결과 및 결론
- 고찰
- 비즈니스 분석

분석 내용에 기반한 제안 모델 – 구성도

Flow chart



분석 내용에 기반한 제안 모델 – 세부 내용

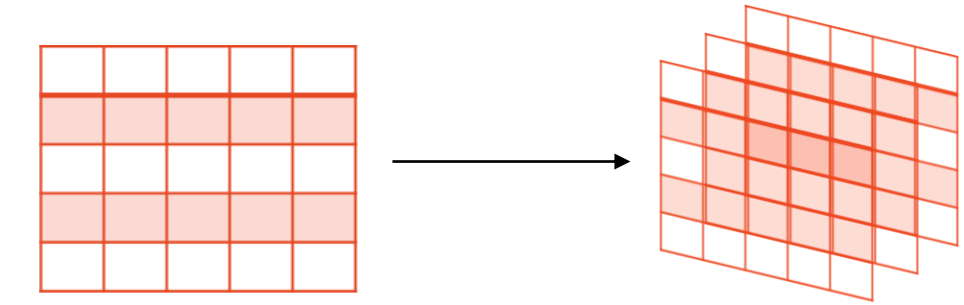
Transform day by day

- 초 단위의 로그 데이터를 일 단위로 변경함

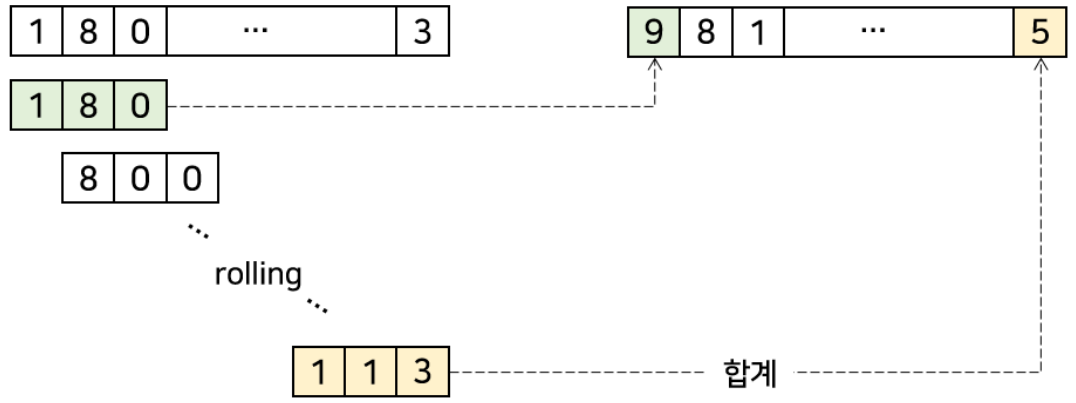
Rolling window

- 일별 데이터를 3일의 window 크기로 누적함

적용 예시



변경 전 (로그, 컬럼)　　변경 후 (사용자, 일, type)



분석 내용에 기반한 제안 모델 - XAI

Explainable AI (XAI) - LIME

- 국소적으로 해석 가능한 모형으로 근사하여 예측 결과를 설명하는 알고리즘
- 예측 결과 신뢰도와 모형 신뢰도를 고려함
- 학습 모델 당 최고 성능 30개 피쳐를 분석함

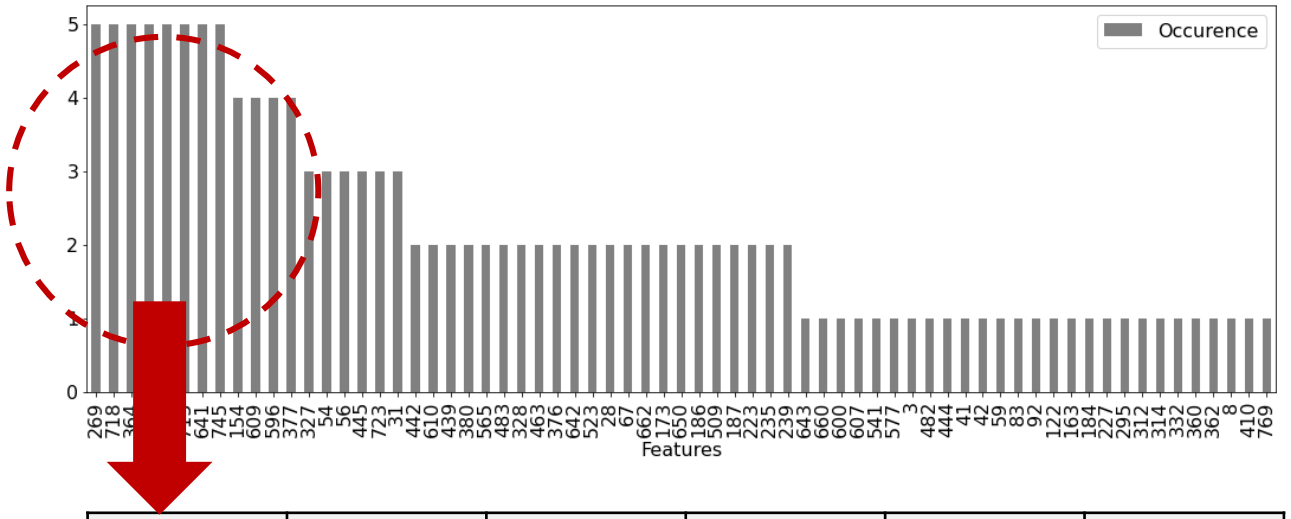
  - 성능 순위: No. 158, 718, 269, 745

  → ① Error type max (no.18)

  → ② Model_nm 업그레이드

  → ③ Error type max (no.30)

  → ④ Quality median (no.5)



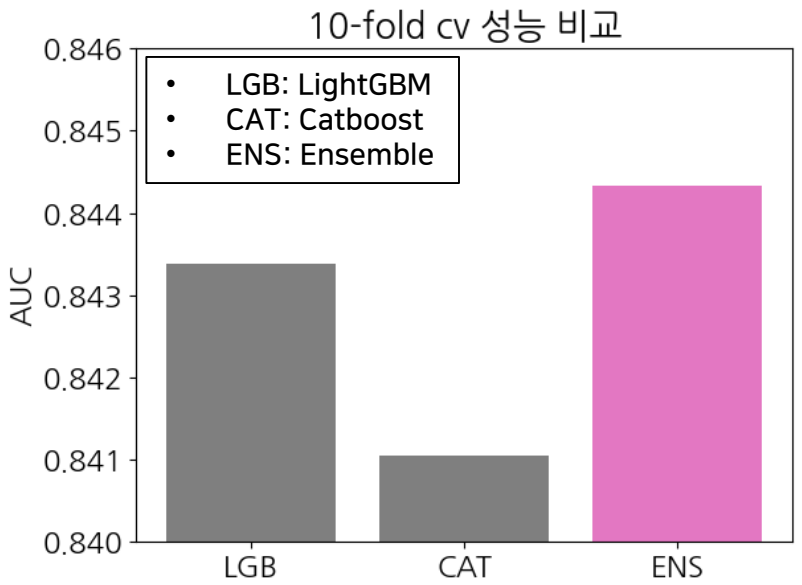
순위	Model1	Model2	Model3	Model4	Model5
1	718	718	718	718	269
2	715	269	715	269	718
3	269	745	745	745	715
4	745	642	269	715	745

[Ref. Ribeiro, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD. 2016.]

결과 및 결론

최종 모델 학습 결과

- 10-fold cross validation score: 0.8443
- Leaderboard score: 0.8479

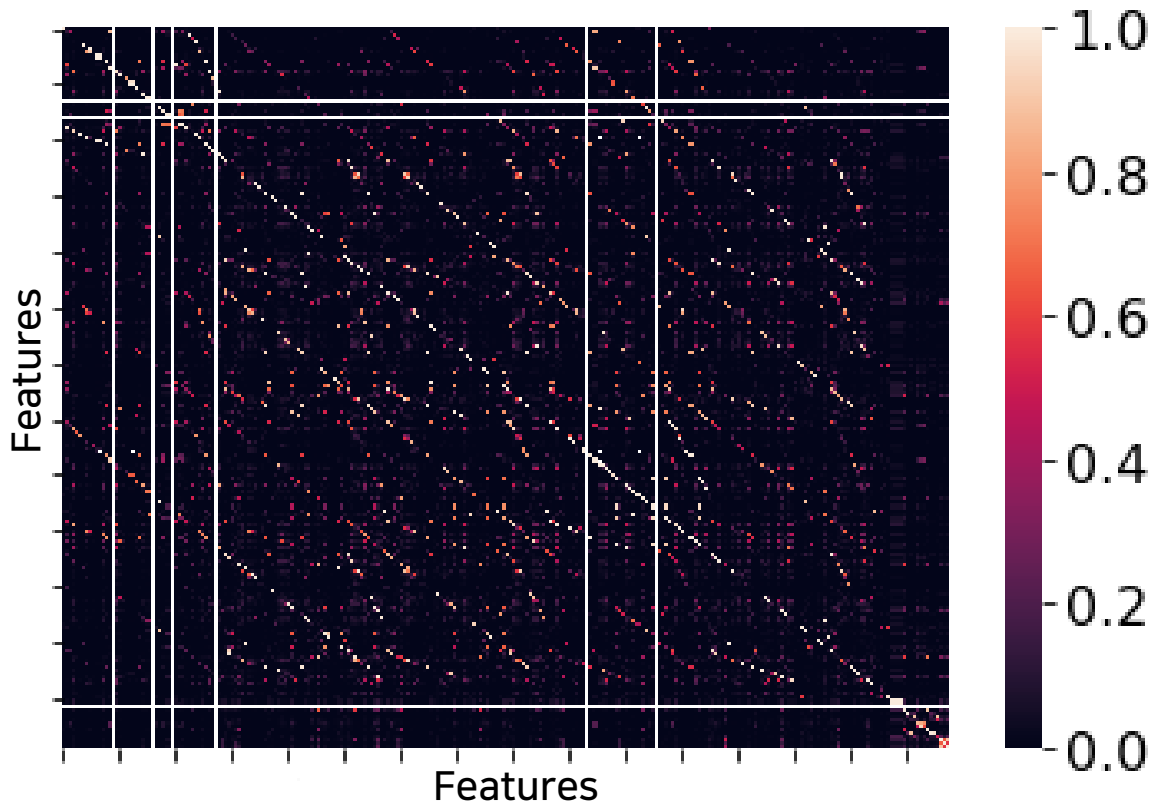


결론

- Error code를 활용할 시 error type만 사용했을 경우보다 많은 정보를 추출 가능
- Quality 값의 크기와 변화가 error의 발생과 사용자 불만 제기 여부에 연관성을 보임
- 9가지의 사용자 불만 접수 원인을 규명함
  - 3일 단위의 에러 합계가 영향력이 큼
- 통계적 특성 분석을 통해 영향력 있는 특징을 추출하고 모델에 반영하여 안정적인 제안 모델을 구축함
  - 블랙박스 모델의 설명을 위해 XAI 해석 도입

고찰

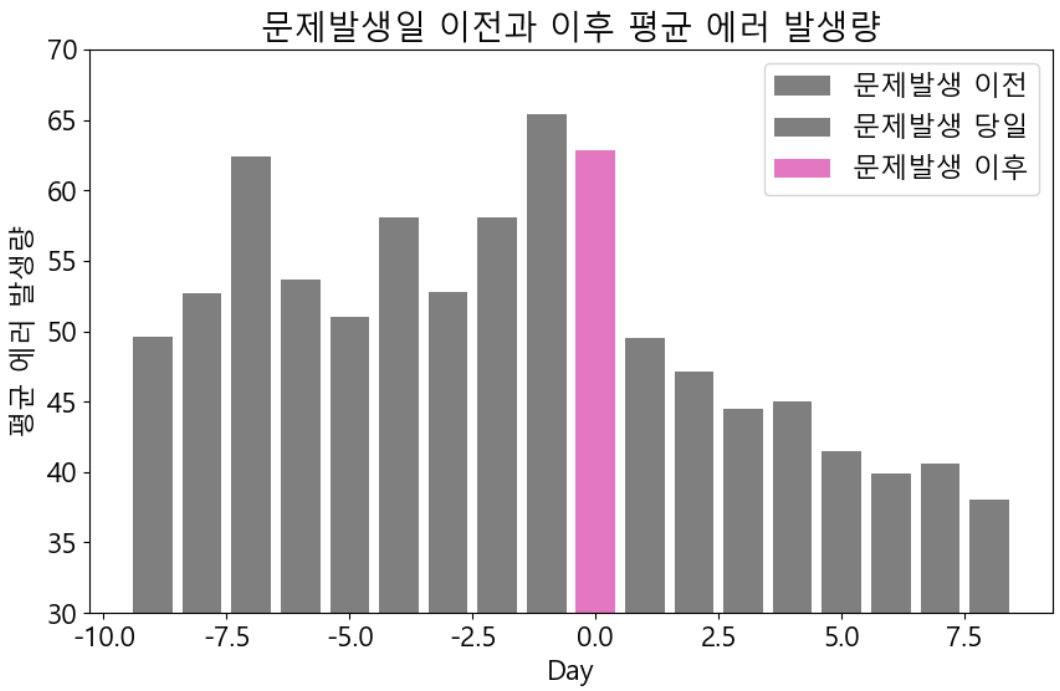
추출한 특징 간 상관성 분석 및 향후 개선 방안



- Feature selection 적용의 필요성
- 추가적인 feature extraction 방안
- 데이터 추가 및 처리 방안

비즈니스 분석

데이터 분석을 통한 사전·사후 고객 관리 필요성



➢ 예상시나리오

- 불만 접수 시간 전 3일 내: 기기 불량 발생 예상
- 불만 접수 시간 후 3일 내: 기기 AS 센터 방문 예상

- 데이터 분석 기반 고객 사전 관리의 필요성
  - 불만 접수 하루 전에 error 발생량이 가장 높음
  - 불만 접수 전에 이를 감지할 수 있다면 불만 발생량을 현저히 낮출 수 있을 것이라 기대됨

- 데이터 분석 기반 고객 사후 관리의 필요성
  - 불만 접수 시간에서 발생하는 예상 시나리오 고려
  - 잔여 5일 (4~8일)은 AS 센터 조치 받은 후 추가적인 기기 고장 등이 발생하는 것으로 간주
  - 고객 관리 기간 확대 및 선 조치 필요

E.O.D

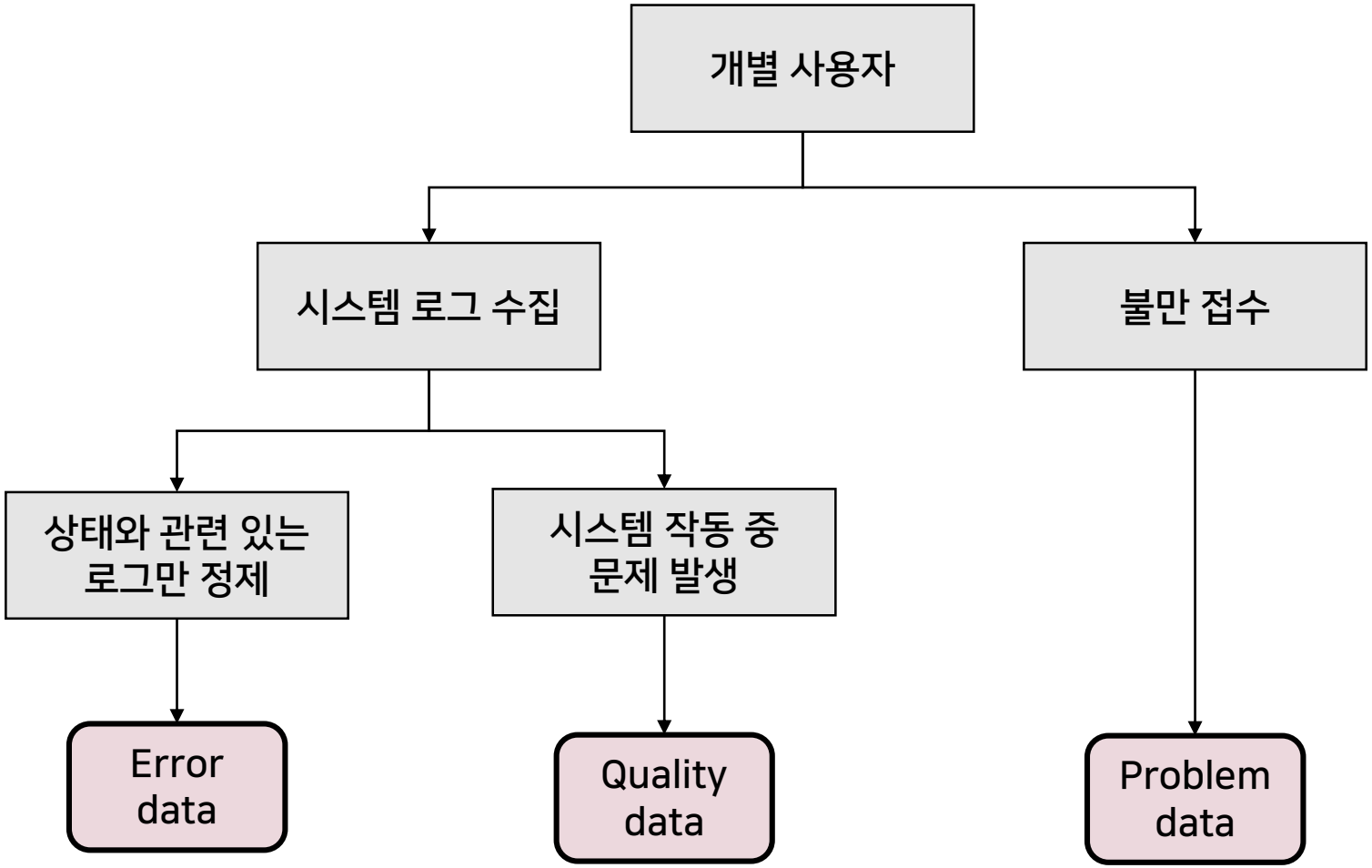
Thank You.

Appendix.

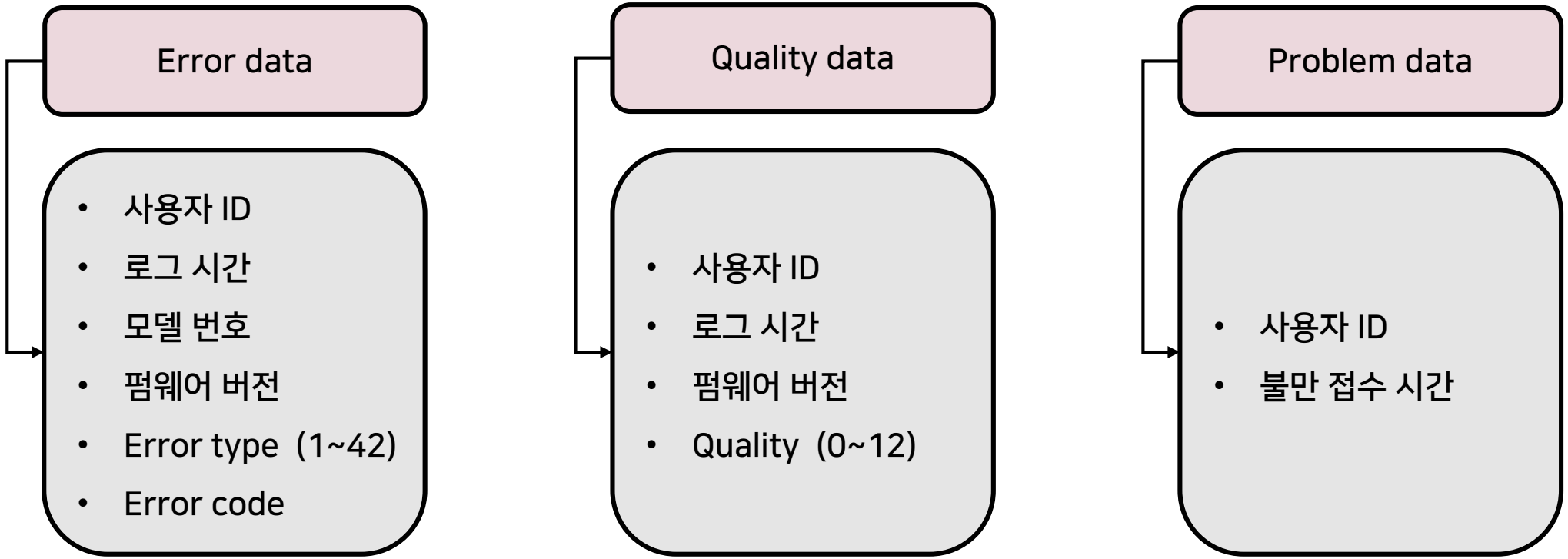
예측 및 분석 대상

주제	시스템 품질 변화로 사용자에게 불편을 야기하는 요인 진단
배경	다양한 장비/서비스에서 일어나는 시스템 데이터를 통해 사용자의 불편을 예지
목적	데이터를 통해 사용자가 불편을 느끼는 원인 분석 및 불편 요인 파악
평가 지표	AUC

데이터 요약 – 데이터 간 관계성 파악



데이터 요약 – 데이터 간 관계성 파악



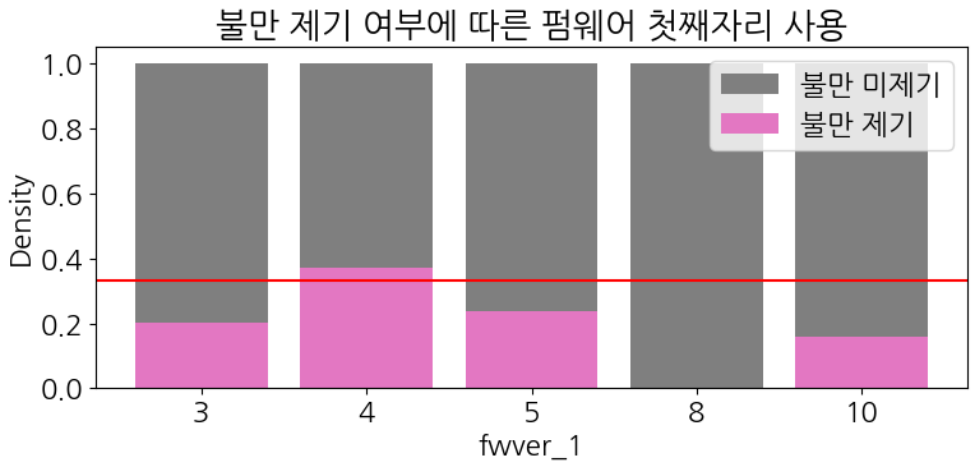
Error 데이터 분석 – Error type 과 error code

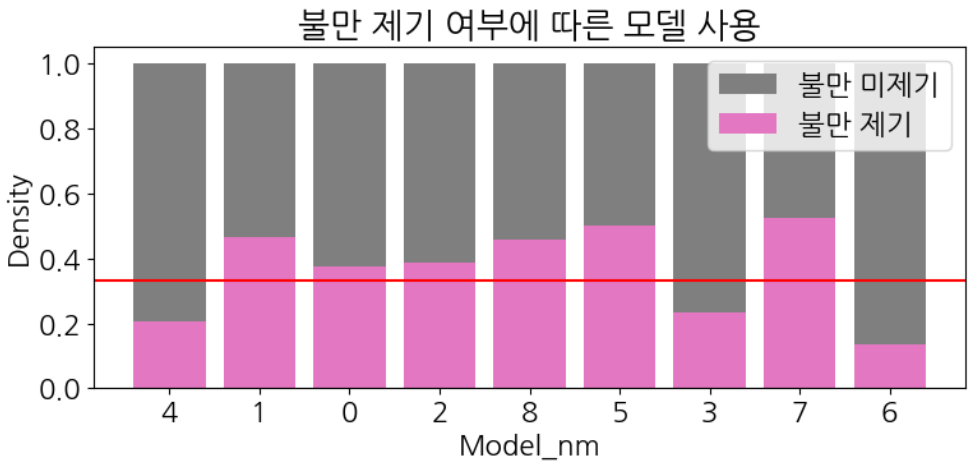
Error code 전처리

Error type	Error code	처리 방법
1, 5, 9	0, P-41001 등	숫자, 영문자 분리하여 처리 (ex. P-41001: P)
2, 3, 4, 6, 7, 14, 17, 30, 31, 33, 34, 37, 39, 40, 42	5개 이하의 숫자	각 코드를 개별적으로 처리
8	20, PHONE_ERR, PUBLIC_ERR	각 코드를 개별적으로 처리
23	Active, connection timeout 등 connection 코드	Fail, timeout, terminate, Active, standby
25	L2CAP connection cancelled 등 connection 코드	Fail, timeout, terminate, cancel, 숫자
32	55개의 숫자	양수와 음수로 나눠 처리
38	2653개의 숫자	Numeric 변수로 처리
10, 11, 12, 13,15, 16, 18, 19, 20, 21, 22, 24, 26, 27, 28, 35, 36, 41	단일 error 코드	-

사용자 불만 접수 원인 – VERSION 분석

Model_nm 별 불만 발생





분석 내용에 기반한 제안 모델 – 세부 내용

Feature extraction

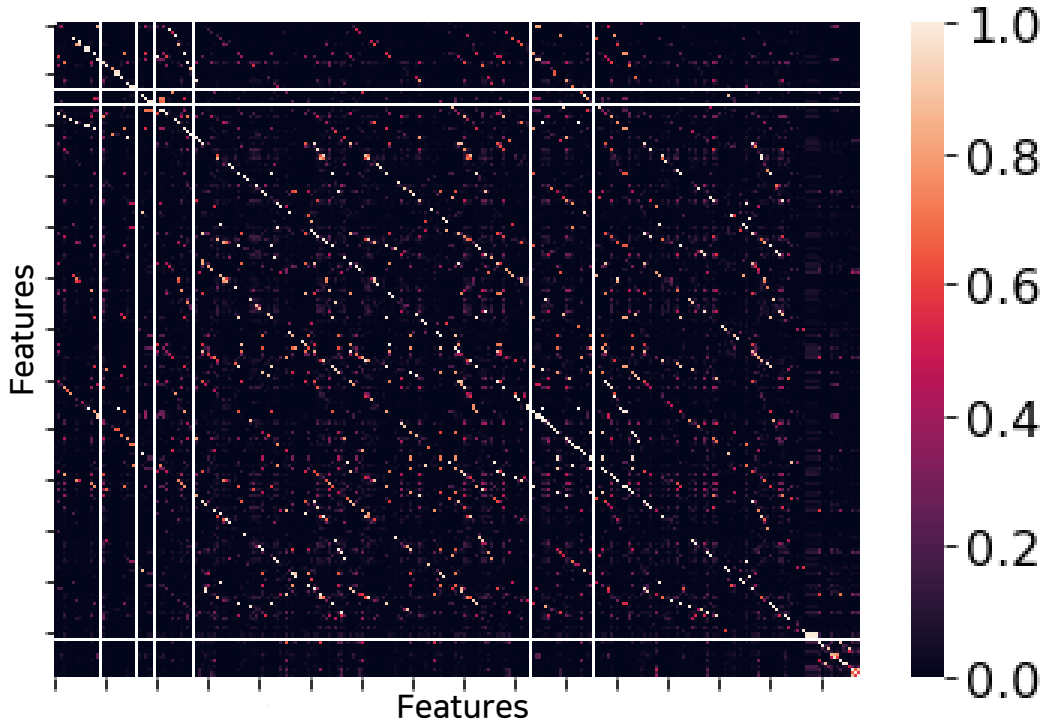
데이터	사용 피쳐
Error	▪ Error type 및 code 발생량의 min, max, mean, median, standard deviation ▪ Error type 38의 error code의 합계 포함
Model_nm	▪ Model_nm 변경, 업그레이드, 다운그레이드 flag ▪ Model 소유 유무 ▪ 시작과 끝에 소유한 Model_nm
Fwver	▪ 첫째 자리만 사용 ▪ Fwver 변경, 업그레이드, 다운그레이드 flag ▪ 시작과 끝에 소유한 fwver
Quality	▪ 수치의 mean, max, median, standard deviation ▪ '-1' 발생 빈도

피쳐 번호	피쳐 종류
1-98	Error code min
99-141	Error type min
142-239	Error code max
240-282	Error type max
283-380	Error code mean
381-423	Error type mean
424-521	Error code median
522-564	Error type median
565-662	Error code standard deviation
663-705	Error type standard deviation
706-714	Model_nm
715-719	Model_nm 변경
720-724	Fwver 변경
725-734	Quality '-1' 발생 빈도
735-774	Quality mean, standard deviation, max, median

고찰

추출한 특징 간 상관성 분석 및 향후 개선 방안

- **Feature selection 적용의 필요성**
  - Collinearity 관련 특징 중복을 확인함
  - 모델 복잡도 개선 가능성을 확인함



- **추가적인 feature extraction 방안**
  - 범주형 변수 처리의 세분화 과정을 진행하여 의미 있는 특징을 생성함
  - 시계열 데이터의 이점을 활용하여 주파수 도메인에서 특징을 생성함
  - AI 기반 압축 모델을 활용하여 특징을 생성함 (e.g. Auto-encoder)

- **데이터 추가 및 처리 방안**
  - Quality 데이터의 결측치 관련 보간법을 활용하여 정보 손실을 최소화함
  - Error type 및 quality의 HW/SW 특성을 확인하여 분석 과정에 반영함
  - 기기 불량에 직접적인 영향을 미칠 수 있는 외부데이터를 활용하여 반영함 (e.g. 온도)