2021 연세 빅데이터 경진대회

TEAM 원효데싸

남승지, 엄상준, 유은영, 조수연

Contents

Contents

분석개요

분석 목적

1. 인형 사용이 노인 사용자의 **건강** 및 **삶의 질** 향상에 미치는 효과성 검증
2. 로그데이터 시계열 분석을 통한 인형의 **사용 행태** 이해

분석 과정

Survey 모델링

1. **데이터 전처리**
- 데이터 선정 및 통합
- 추가 정제

2. **EDA**
- 정제된 데이터의 시각화
- 분석 insight 도출

3. **모델 선정 및 분석**
- 지도학습 모델링

(RandomForest, LGBM)

주별 사용량 예측 모델링

1. **데이터 전처리**
- 데이터 선정 및 통합
- 추가 정제

2. **EDA**
- 정제된 데이터의 시각화
- 분석 insight 도출

3. **모델 선정 및 분석**
- 지도학습 모델링

(RandomForest, Huber Regressor)

사용자 이탈 생존분석

1. **데이터 전처리**
- 데이터 선정 및 통합
- 추가 정제

2. **EDA**
- 정제된 데이터의 시각화
- 분석 insight 도출

3. **모델 선정 및 분석**
- 생존분석

(CoxPHFitter, LogNormalAFT)

Contents

Survey Modeling - Survey

설문데이터_200709(연대 빅데이터).xlsx → 인형 사용의 사전 및 사후 설문문항 + 인구통계학적인 정보

Aa 카테고리	변수명
기본 정보	기관, 나이(age), 성별(sex), 수급여부, 세대구성, 제공서비스, 주택타입, 배우자, 자녀, 자녀수, 아들수, 딸수, 청결, 식사, 공공방문, 종교, 종교유무, 왕래여부
건강	치매, 우울증, 만성복약, 고립, 거동불편, 소리반응, 인형관심, 모니터링, 건강관심
인형	인형 아이디(doll_id), 인형 사용시간(기상, 아침, 점심, 저녁, 취침)
복용약	아침식전_복용, 아침식후_복용, 점심식전_복용, 점심식후_복용, 저녁식전_복용, 저녁식후_복용, 취침전_복용, 치매_약, 뇌졸증_약, 혈압_약 우울증_약, 고지혈증_약, 당뇨_약, 신경과수면제_약, 복용약 개수(med_count)
우울증 설문조사	사전 설문조사 우울증 점수(psy_before, psy_before_cat), 사후 설문조사 우울증 점수(psy_after, psy_after_cat)
생활관리 설문조사	사전 설문조사 생활관리 점수(life1_before, slife_before, slife_before_cat), 사후 설문조사 생활관리 점수(life1_after, slife_after, slife_after_cat)
인형 사용 만족도 설문조사	만족도 점수(doll_score)
파생변수	머리 쓰다듬 횟수(stroke), 손 버튼 누름 횟수(hand_hold), 등 두드림 횟수(knock), 인체 감지 횟수(human_detection), 체조 실행 횟수(gymnastics), 퀴즈 실행 횟수(brain_timer), 약 복용수(drug_consume), 푸쉬 알림수(emrg_cnt), 하루 약 복용 비율(drug_consume_rt), 사용일수(yn)

Survey Modeling - Survey

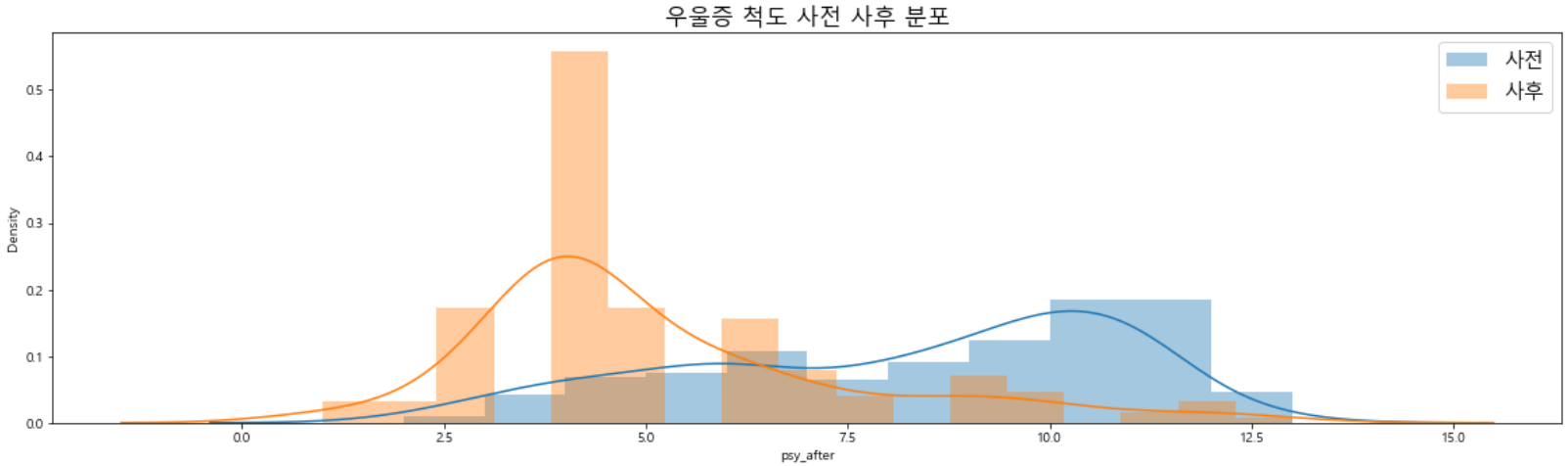
- 생활관리 설문지: 24점 만점, **높은 점수**가 긍정적

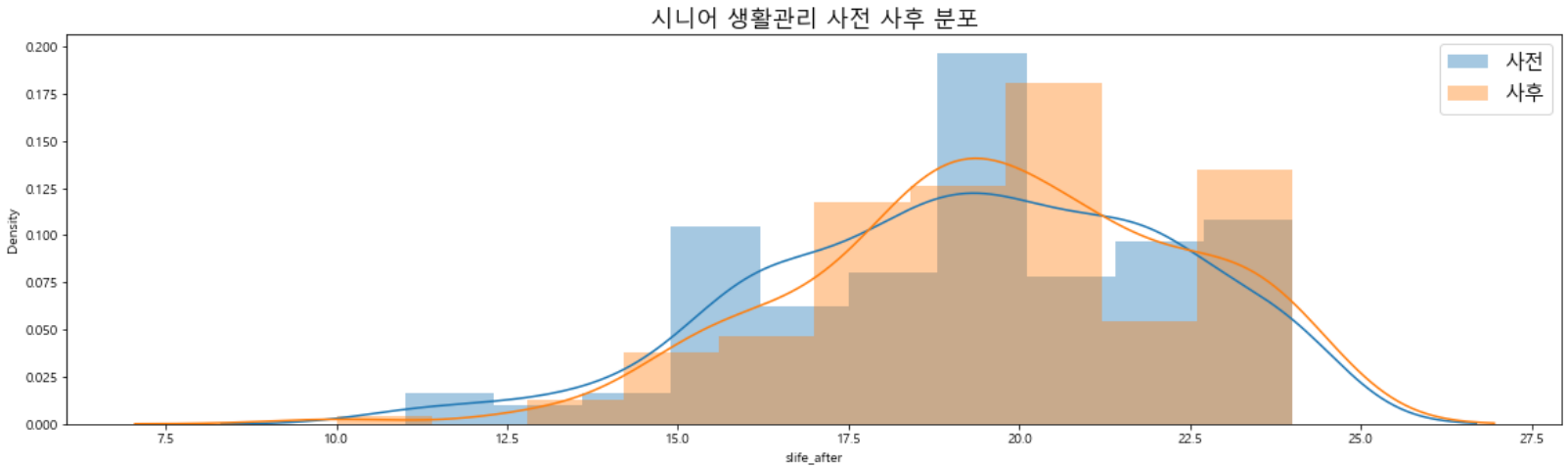
질문	평가
1. (기상/취침)어르신께서는 매일 일정한 시간에 기상/취침을 하고 계십니까?	1. 매일 규칙적 2. 때에 따라 3. 일정 하지 않음
2. (환기)어르신께서는 매일 환기를 하고 계십니까?	1. 매일 규칙적 2. 가끔 필요시 3. 거의 하지 않음
3. (약 먹기)어르신께서는 매일 정확한 시간에 약을 복용하고 계십니까?	1. 매일 정확한 시간에 약을 복용하고 있음 2. 가끔 약 먹는 것을 잊을 때가 있음 3. 거의 매번 약 먹는 것을 잊음
4. (식사)어르신께서는 매일 규칙적으로 세끼 식사를 하십니까?	1. 매일 규칙적 세끼 식사하고 있음 2. 가끔 식사시간 놓침 3. 거의 매번 식사시간 놓침
5. (산책)어르신께서는 산책을 얼마나 자주 하십니까?	1. 매일 규칙적 2. 가끔 필요시 3. 거의 하지 않음
6. (체조)어르신께서는 체조를 얼마나 자주 하십니까? (* 실내 운동도 포함)	1. 매일 규칙적 2. 가끔 필요시 3. 거의 하지 않음
7. (긍정적사고) 어르신께서는 얼마나 자주 긍정적인 생각을 하십니까?	1. 매일 긍정적 2. 가끔 부정적인 생각이 들 때가 있음 3. 긍정적인 생각을 거의 하지 않음
8. (사회적 관계 맺기) 어르신께서는 얼마나 자주 다른 사람들과 접촉(전화, 만남 등)을 하고 싶은 생각이 드십니까?	1. 매일 다른 사람들과 접촉하고 싶음 2. 가끔 다른 사람들과 접촉하고 싶음 3. 다른 사람들과 접촉하고 싶지 않음

- 우울증 설문지: 15점 만점, **낮은 점수**가 긍정적

질 문 내 용		
1. 현재의 생활에 대체적으로 만족하십니까?	예	아니오
2. 요즈음 들어 활동량이나 의욕이 많이 떨어지셨습니까?	예	아니오
3. 자신이 헛되이 살고 있다고 느끼십니까?	예	아니오
4. 생활이 지루하게 느껴질 때가 많습니까?	예	아니오
5. 평소에 기분은 상쾌한 편이십니까?	예	아니오
6. 자신에게 불길한 일이 닥칠 것 같아 불안하십니까?	예	아니오
7. 대체로 마음이 즐거운 편이십니까?	예	아니오
8. 절망적이라는 느낌이 자주 드십니까?	예	아니오
9. 바깥에 나가기가 싫고 집에만 있고 싶습니까?	예	아니오
10. 비슷한 나이의 다른 시니어들보다 기억력이 더 나쁘다고 느끼십니까?	예	아니오
11. 현재 살아있다는 것이 즐겁게 생각되십니까?	예	아니오
12. 지금의 내 자신이 아무 쓸모없는 사람이라고 느끼십니까?	예	아니오
13. 기력이 좋으신 편이십니까?	예	아니오
14. 지금 자신의 처지가 아무런 희망이 없다고 느끼십니까?	예	아니오
15. 자신이 다른 사람들의 처지보다 더 못하다고 느끼십니까?	예	아니오

설문조사 분포


우울증 척도 사전 사후 분포
사전
사후
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
psy_after


시니어 생활관리 사전 사후 분포
사전
사후
Density
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
27.5
slife_after

Survey Modeling - 분석 개요

STEP 1

사전 사후 모두
진행된 사용자 대상
&
인형 사용행태
데이터 구축

>>

STEP 2

Train / Test Split

Train: **128**명
Test: **32**명

>>

STEP 3

Numeric data
Scaling
&
Feature Selection

>>

STEP 4

다양한 지도학습
모델링 진행

- 우울증: **Random Forest**
- 생활관리: **LightGBM**

Survey Modeling - 전처리

인구통계학적 정보

Dummies

- 인구통계학적 기본 정보 중 **더미화** 시킬 수 있는 변수 처리 (수급 여부 등)

Test id

- Test id를 제외하고 **유효한 Doll id**를 가지고 있는 유저들만 선정

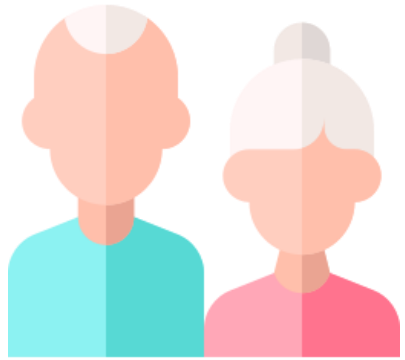
복용약

- 사람들이 대표적으로 많이 먹는 약들을 **복용 여부**로 처리 (치매, 고지혈증 등)

분석 대상

- 사전 & 사후 설문조사 **모두 존재** 하는 사용자

총 160명



Survey Modeling - 전처리

로그데이터

파생변수 생성	설문조사 데이터 + 로그데이터
설문조사 기간	사전 (2019.09.09) ~ 사후 (2019.12.20)
Doll_id	설문조사(우울증 & 생활관리) 데이터와 통일

Sum	Stroke Hand_hold Knock Human_detection Gymnastics Brain_timer Drug_consume Emrg_cnt
설정 대비 실제 복용 비율	Drug_consume_rt
사용일수	90일 중 인형 사용일수

Survey Modeling - 전처리

Y 변수

- 점수 자체가 가진 의미보다, 인형 사용으로 인한 생활관리 또는 우울증 **개선 여부**가 더 중요하다고 판단

사전점수 – 사후점수 = 변화점수

우울증: 0초과인 경우 1(개선)로 Coding (낮을수록 좋기 때문)

- 0: 57명
- 1: 103명

생활관리: 0미만인 경우 1(개선)로 Coding (높을수록 좋기 때문)

- 0: 71명
- 1: 89명

Survey Modeling - 전처리

인구통계학적 정보

1. 기본 정보 : 더미화
2. 인형 : Doll_id (6자리)
3. 복용약 : 40개 이상(치매 등)
4. 설문조사

- 사전 & 사후 설문조사 모두 존재 하는 사용자

로그데이터

- 설문조사 기간
  사전설문조사 (2019.09.09)
  사후설문조사 (2019.12.20)
- Survey data : Doll_id
- 파생변수 생성
- 사용일수(yn) 변수 생성

Y 변수

1. 우울증

- 긍정변화 여부 (Binary)

2. 생활관리

- 긍정변화 여부 (Binary)

Survey Modeling - 전처리

사용변수

Aa 카테고리	변수명
기본 정보	기관, 나이(age), 성별(sex), 수급여부, 세대구성, 제공서비스, 주택타입, 배우자, 자녀, 자녀수, 아들수, 딸수, 청결, 식사, 공공방문, 종교, 종교유무, 왕래여부
건강	치매, 우울증, 만성복약, 고립, 거동불편, 소리반응, 인형관심, 모니터링, 건강관심
인형	인형 아이디(doll_id), 인형 사용시간(기상, 아침, 점심, 저녁, 취침)
복용약	아침식전_복용, 아침식후_복용, 점심식전_복용, 점심식후_복용, 저녁식전_복용, 저녁식후_복용, 취침전_복용, 치매_약, 뇌졸증_약, 혈압_약 우울증_약, 고지혈증_약, 당뇨_약, 신경과수면제_약, 복용약 개수(med_count)
우울증 설문조사	사전 설문조사 우울증 점수(psy_before, psy_before_cat), 사후 설문조사 우울증 점수(psy_after, psy_after_cat)
생활관리 설문조사	사전 설문조사 생활관리 점수(life1_before, slife_before, slife_before_cat), 사후 설문조사 생활관리 점수(life1_after, slife_after, slife_after_cat)
인형 사용 만족도 설문조사	만족도 점수(doll_score)
파생변수	머리 쓰다듬 횟수(stroke), 손 버튼 누름 횟수(hand_hold), 등 두드림 횟수(knock), 인체 감지 횟수(human_detection), 체조 실행 횟수(gymnastics), 퀴즈 실행 횟수(brain_timer), 약 복용수(drug_consume), 푸쉬 알림수(emrg_cnt), 하루 약 복용 비율(drug_consume_rt), 사용일수(yn)

Survey Modeling - 전처리

Train / Test Split

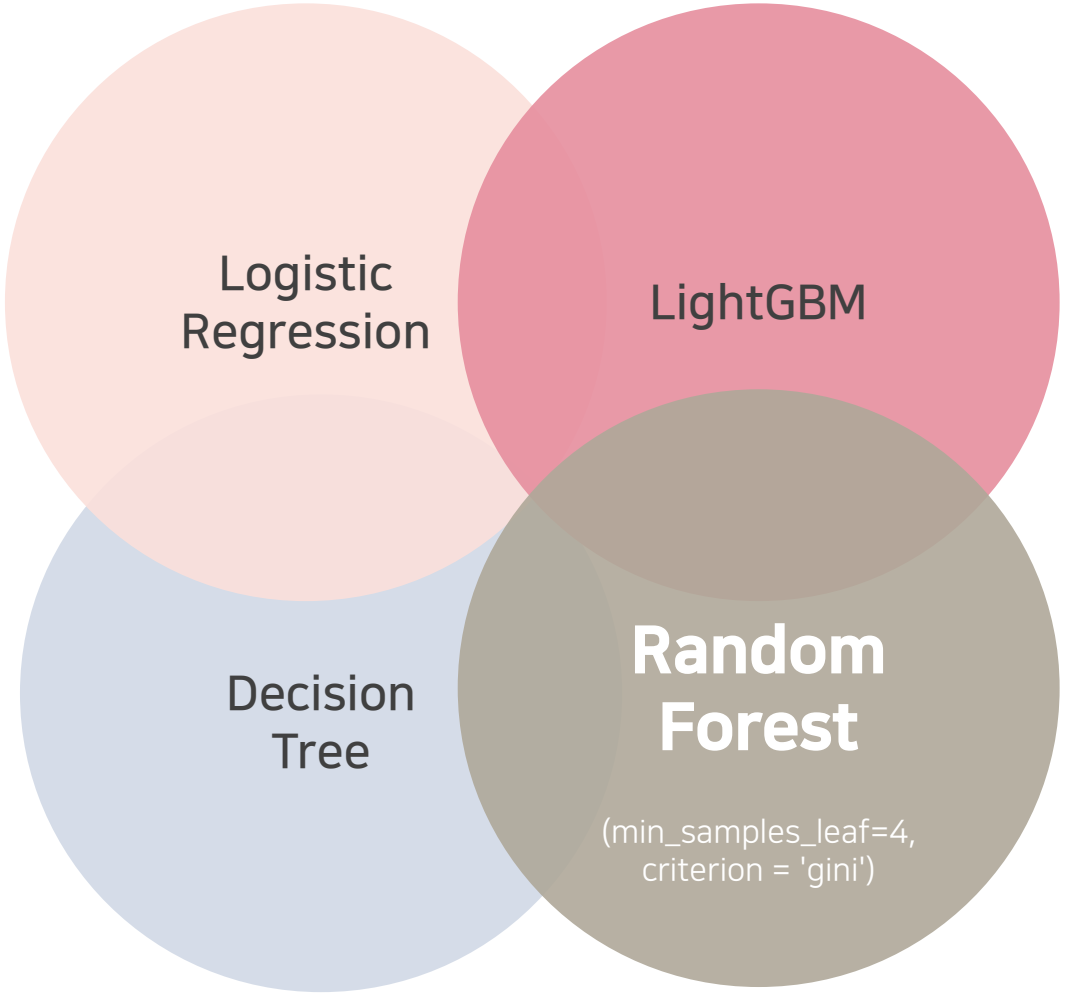
- Test 비율: 0.2
- **층화추출**로 분리

Feature selection

- Correlation 분석 → **0.7 이상** 변수들 중 다른 변수들이랑 상관관계가 높은 변수 제외

Survey Modeling – Modeling(우울증)

1) 우울증 모델링: 우울증 설문 점수 개선 여부로 Classification



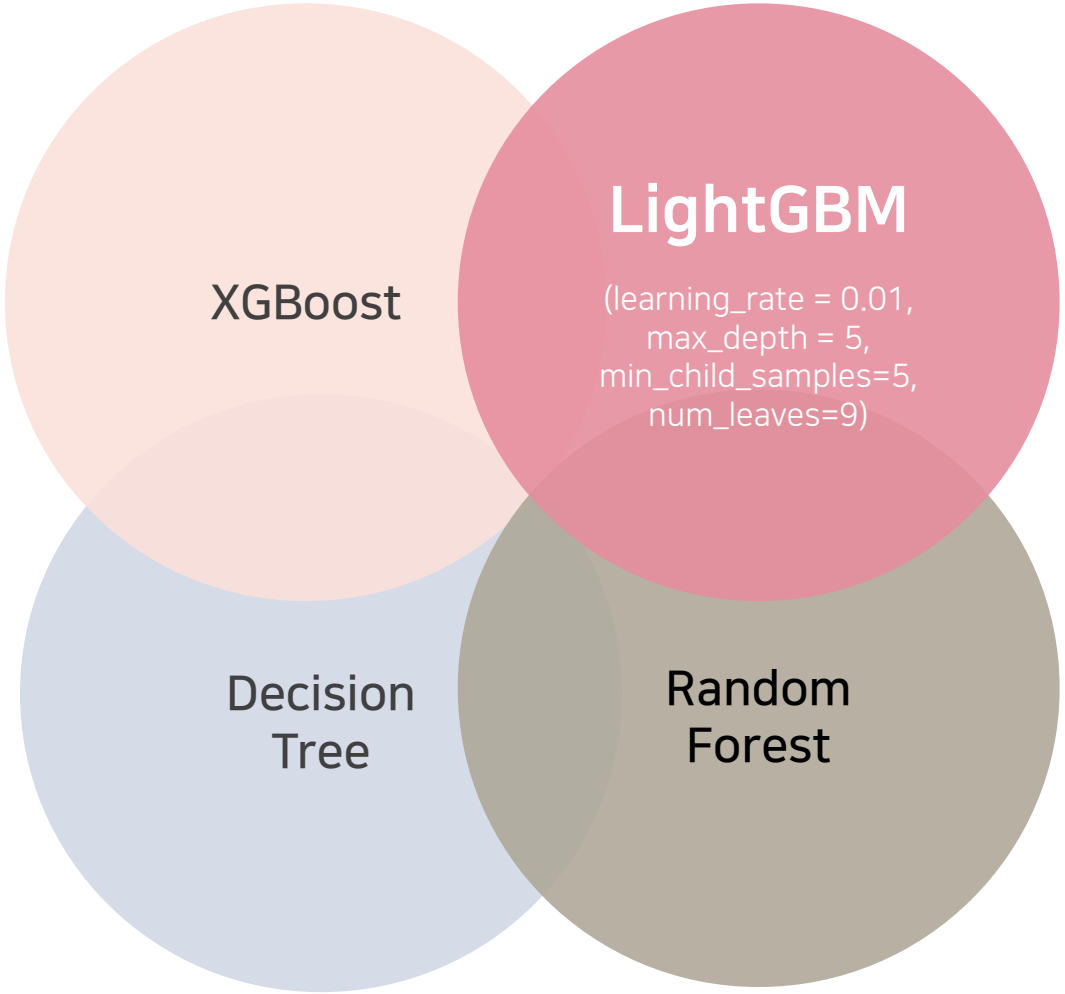
Model	Train Accuracy(CV)
RandomForest	**0.824**
LGBM	0.810
DecisionTree	0.766
Logistic 회귀	0.727

➔ RandomForestClassifier 선택

Test Accuracy: 0.781

Survey Modeling – Modeling(우울증)

2) 생활관리 모델링: 생활관리 설문 점수 개선 여부로 Classification



Model	Train Accuracy(CV)
LGBM	**0.795**
RandomForest	0.774
XGBoost	0.753
DecisionTree	0.734

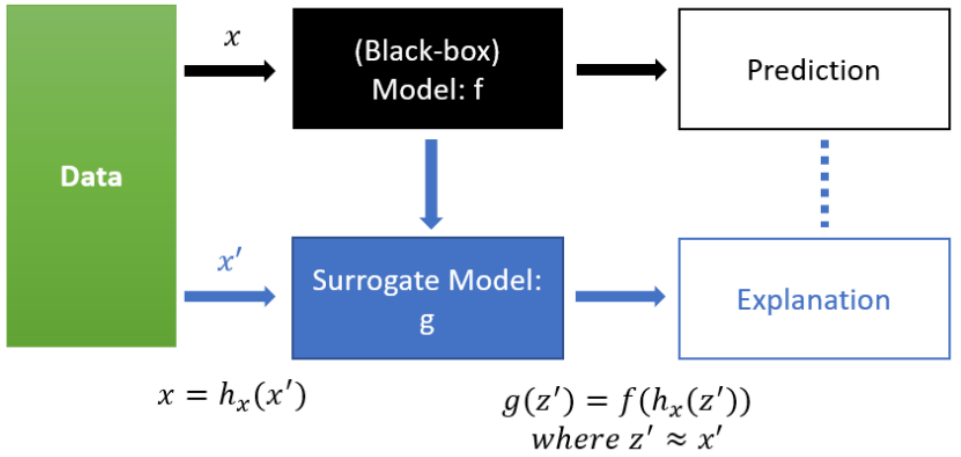
➔ LGBMClassifier 선택

Test Accuracy: 0.75

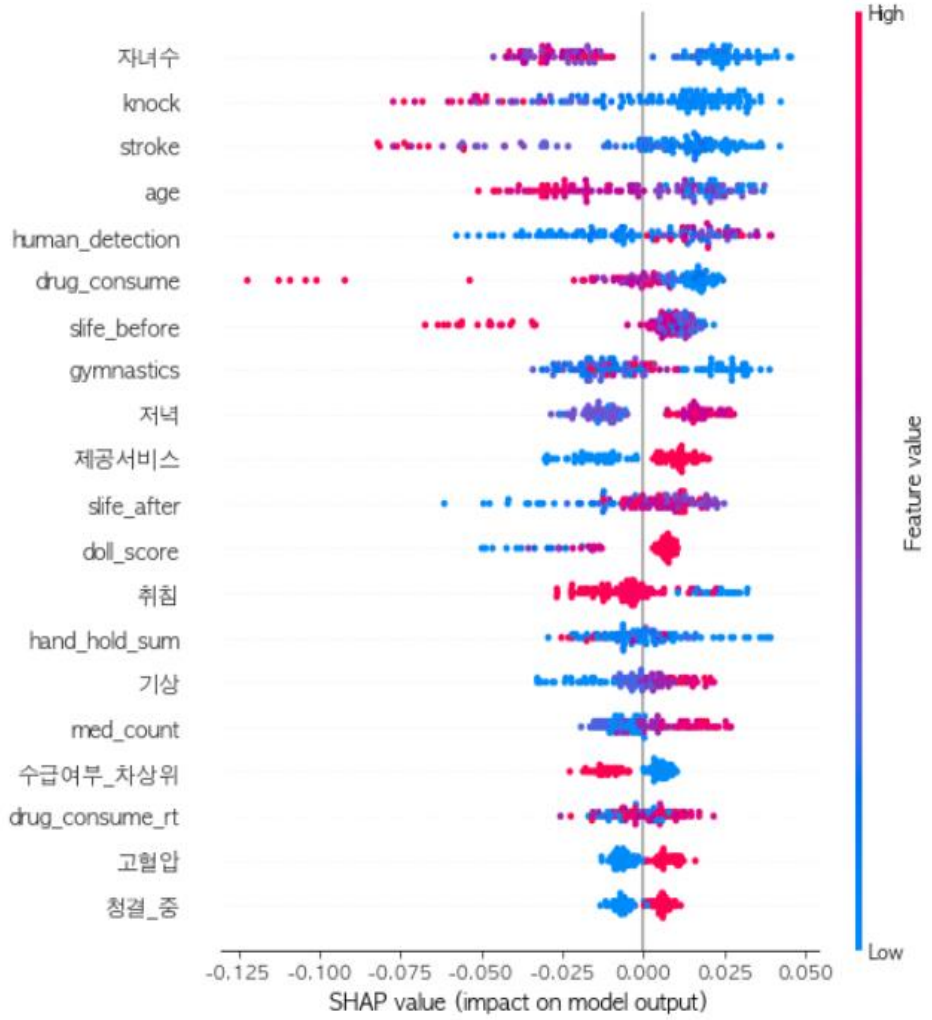
* LightGBM: Boosting 계열 알고리즘의 일종.
리프 중심으로 트리를 분할하여 예측 오류 손실을 최소화함.

SHAP (Shapley Additive exPlanations)

- **원인 인자**를 찾고, 얼마나 결과에 영향을 주었는지 파악해야 할 때
- 예측에 대한 **각 특성의 기여도 계산**



Summary Plot



우울증 개선에 긍정적인 영향

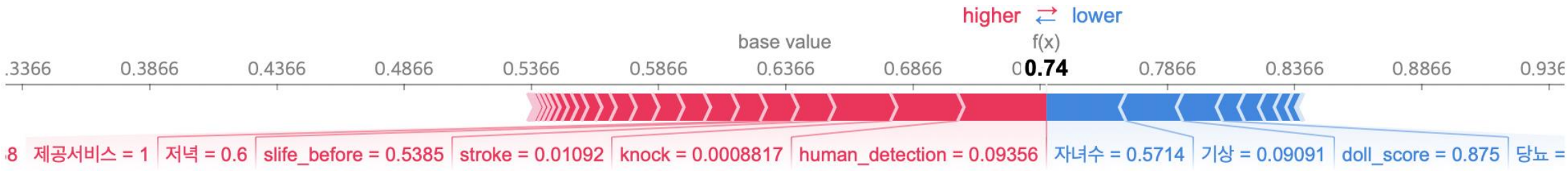
- **인형 만족도** 점수가 높을수록 긍정적
- Human_detection(**인체 감지**), med_count(**약 복용 횟수**) 값이 높을수록 긍정적
- **생활관리 사후 설문점수**가 높을수록 긍정적

우울증 개선에 부정적인 영향

- **나이**가 많을수록 부정적
- **자녀 수**가 많을수록 부정적

Individual Force Plot

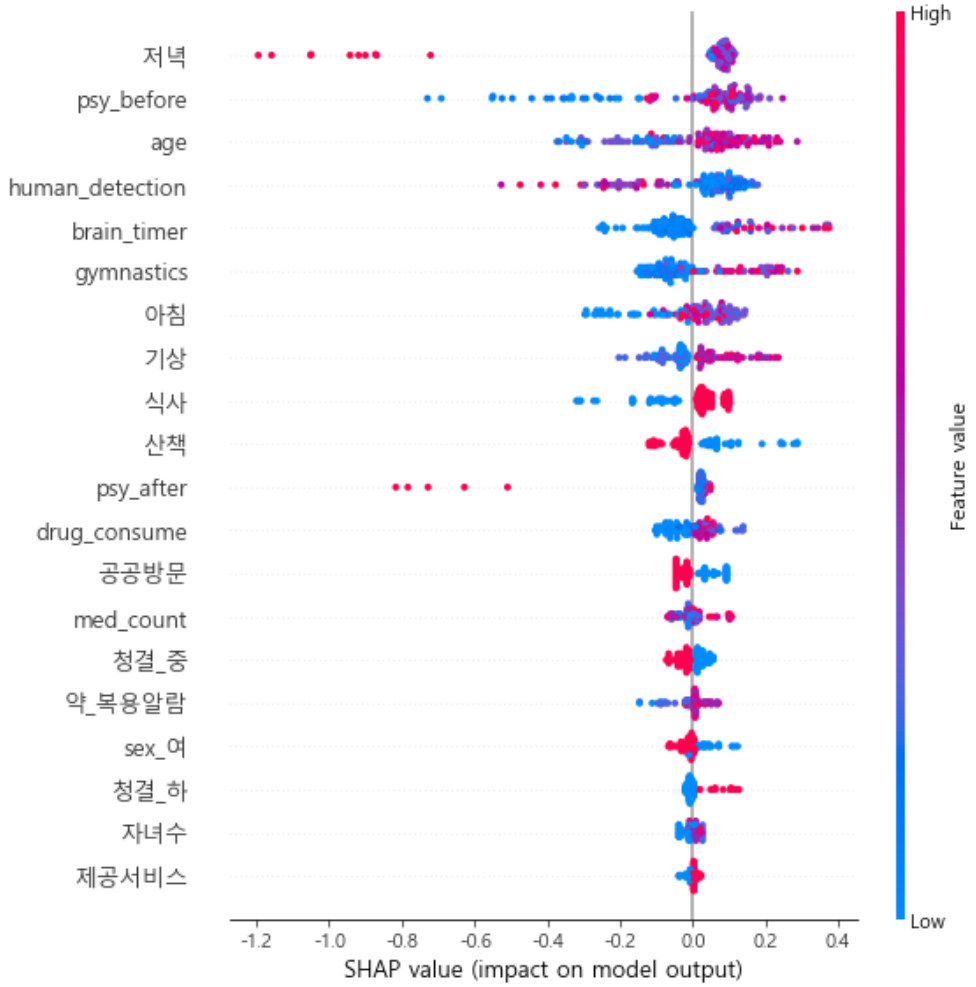
- 어느 한 유저에 대한 각 변수들의 영향을 나타냄



- Predict 결과 : 1
- 확률 : 0.74
- 해당 유저의 우울증이 개선되었을 것이라고 예측
- 부정적인 영향 : 자녀수 > 기상 > doll_score > 당뇨 …
- 긍정적인 영향 : human_detection > knock > stroke > …

Survey Modeling – 결과해석(생활관리)

Summary Plot



생활관리 개선에 긍정적인 영향

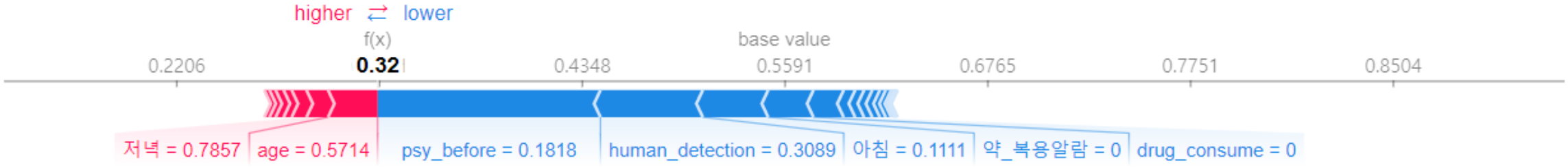
- **우울증 사전설문 점수**가 높을수록 긍정적
- **나이**가 많을수록 긍정적
- **퀴즈, 체조 실행 횟수**가 많을수록 긍정적
- **약 복용 횟수**가 더 많을수록 긍정적

생활관리 개선에 부정적인 영향

- **저녁 식사 시간**이 늦을수록 부정적
- **우울증 사후설문 점수**가 높을수록 부정적

Individual Force Plot

- 어느 한 유저에 대한 각 변수들의 영향을 나타냄



- Predict 결과 : 0
- 확률 : 0.32
- 해당 유저의 생활관리가 개선되지 않았을 것이라고 예측
- 부정적인 영향 : psy_before > human_detection > 아침 > 약_복용알람 …
- 긍정적인 영향 : age > 저녁 > …

Survey Modeling – 활용방안

Useful Results

- **퀴즈/체조** 기능: 생활관리 개선에 긍정적
- **인형 사용 전 우울증 정도**가 심할수록 생활관리 개선이 잘된다는 결과

활용방안

1) **기능 광고**: 광고 시에 퀴즈, 체조 기능의 효과 강조

2) **광고 대상 타겟팅**: 우울증 정도가 심한 사용자일수록 개선될 가능성이 높음을 선전

3) **기능 발전 및 개선**: 현재는 인형에 내장된 치매예방퀴즈 사용

Ex) 퀴즈 기능 맞춤화: 어플에서 보호자가 퀴즈 주제 및 내용을 선택할 수 있도록 개선

Contents

주별 사용량 예측 Modeling - EDA

사용일수에 따른 인형 사용 기능 양상 차이

A user

- 사용 시작일: 2020년 2월 8일

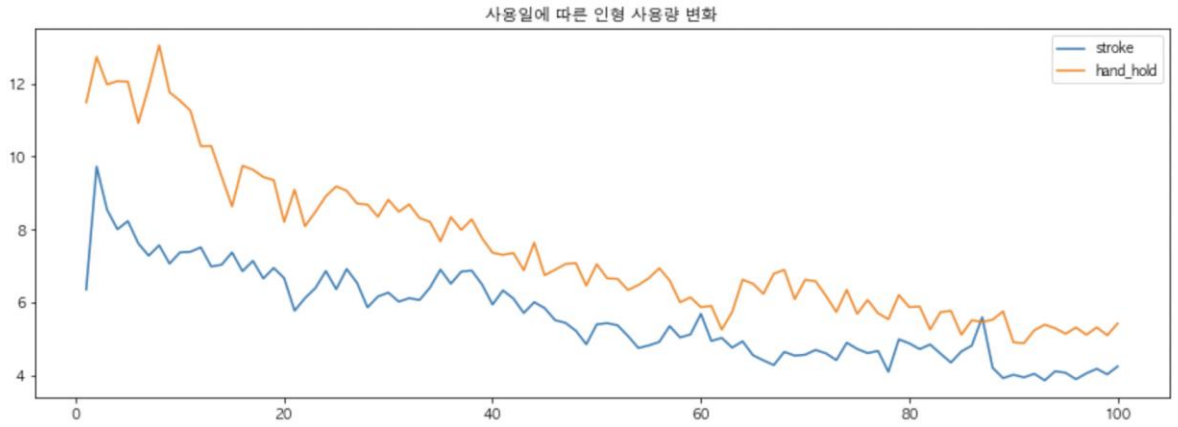
B user

- 사용 시작일: 2021년 1월 13일

서로 다른 사용 시작일

사용시작일을 1일로 맞추어
사용일에 따른 인형 사용량 변화 확인

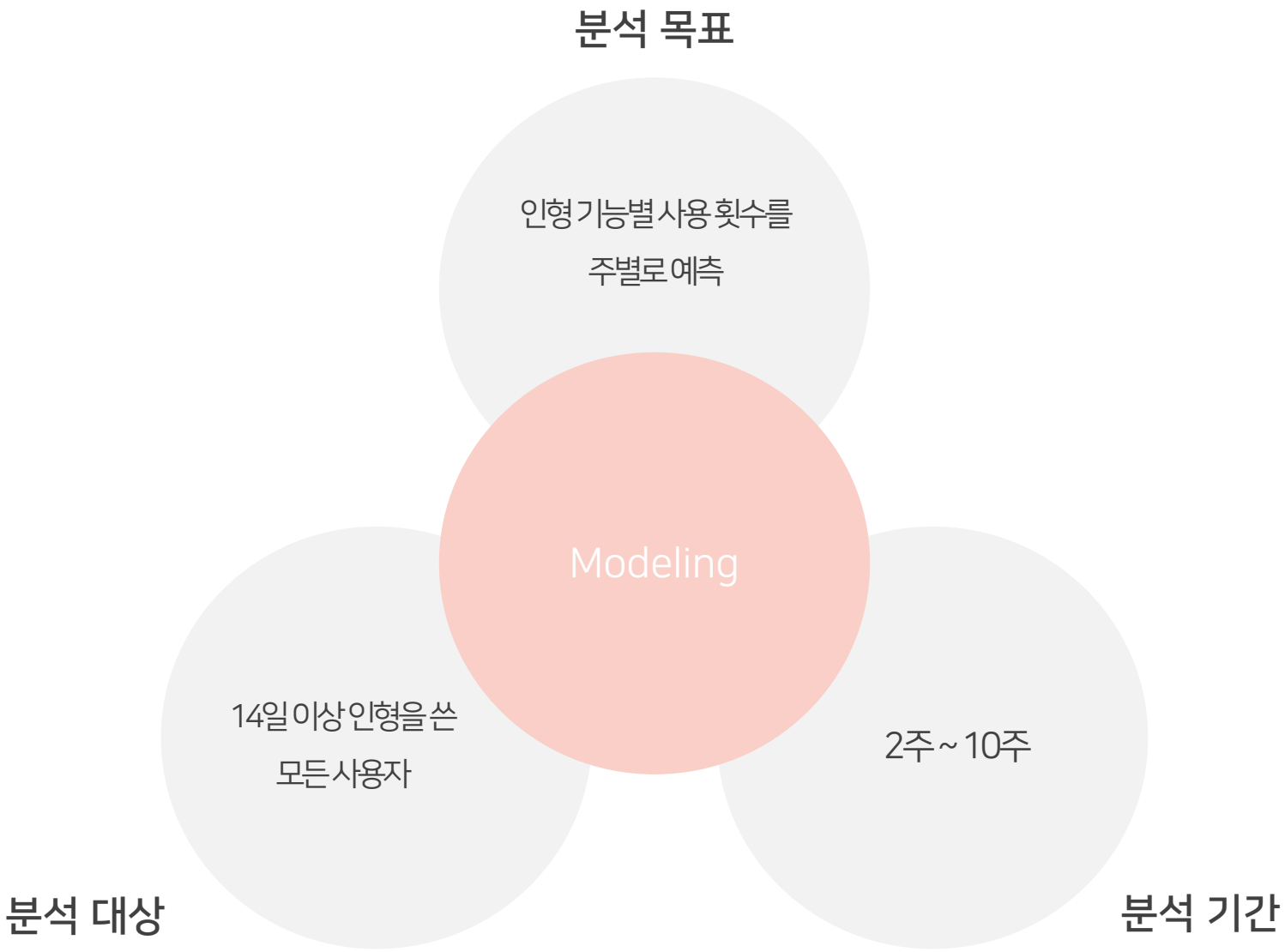
사용일에 따른 인형 사용량 변화



사용일에 따른 변화가 있음을 확인

전 주의 인형 사용 데이터를 활용하여
금주의 **인형 사용량**을 예측해보기로 함.

주별 사용량 예측 Modeling - 전처리



주별 사용량 예측 Modeling - 전처리

주간 사용 횟수

주별 예측을 위하여 기능마다 **일주일 간의 사용 횟수**를 합함

ex) 7일동안 quiz 기능을 3번 사용했으면 3

전주의 데이터

주차마다 번호를 매겨 t주차 사용 예측을 위해서 t-1주 로그 데이터를 사용

Variables

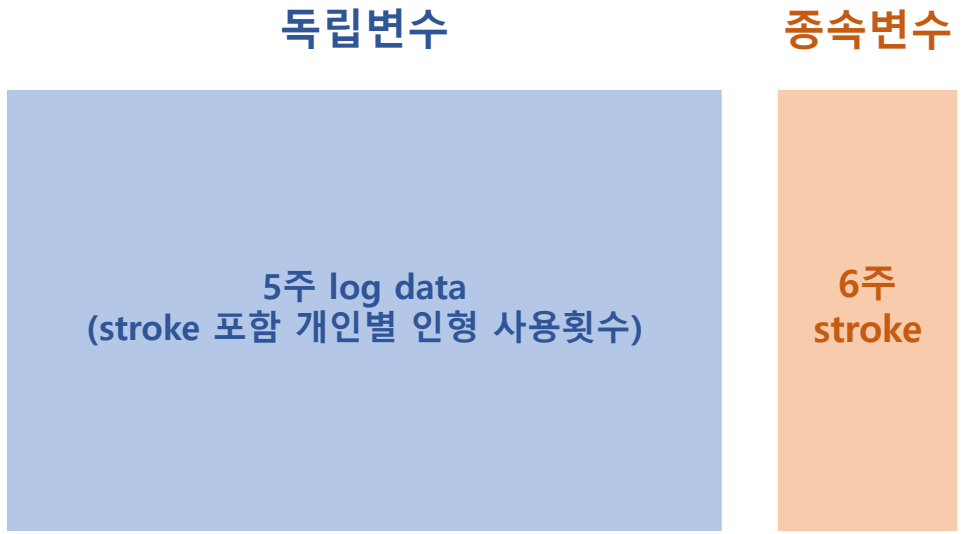
- **독립 변수**

: [stroke, hand_hold, knock, consume_cnt, story, religion, music, english, rememberance, quiz, gymnastics_y]

- **종속 변수**

: [stroke, hand_hold, knock, quiz, music, gymnastics]

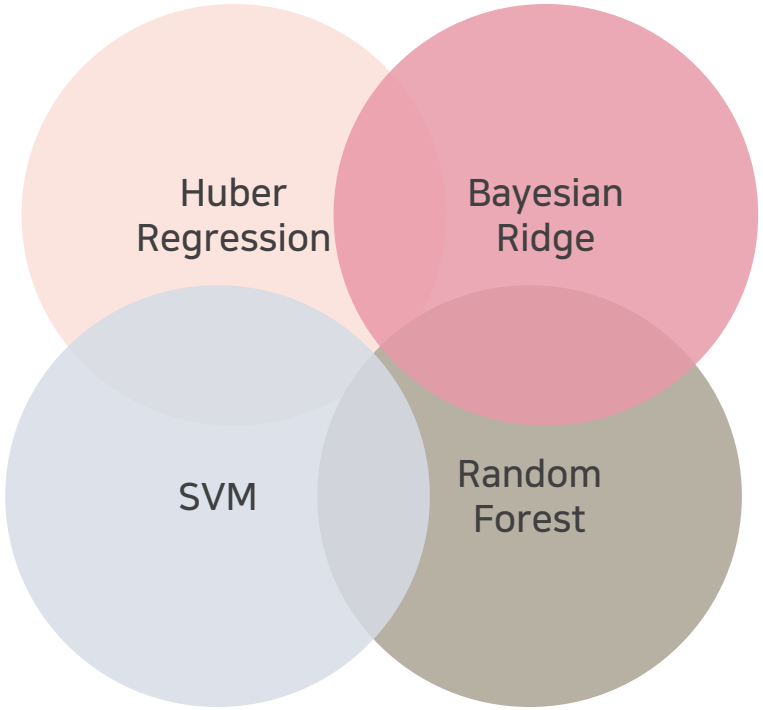
예시 (stroke)



주별 사용량 예측 Modeling - Modeling

(Robust Linear Regression의 일종)

Candidates



Cross Validation

: MAE, RMSE, MAPE

Random Forest Regressor

VS

Huber Regressor

Random Forest Regressor

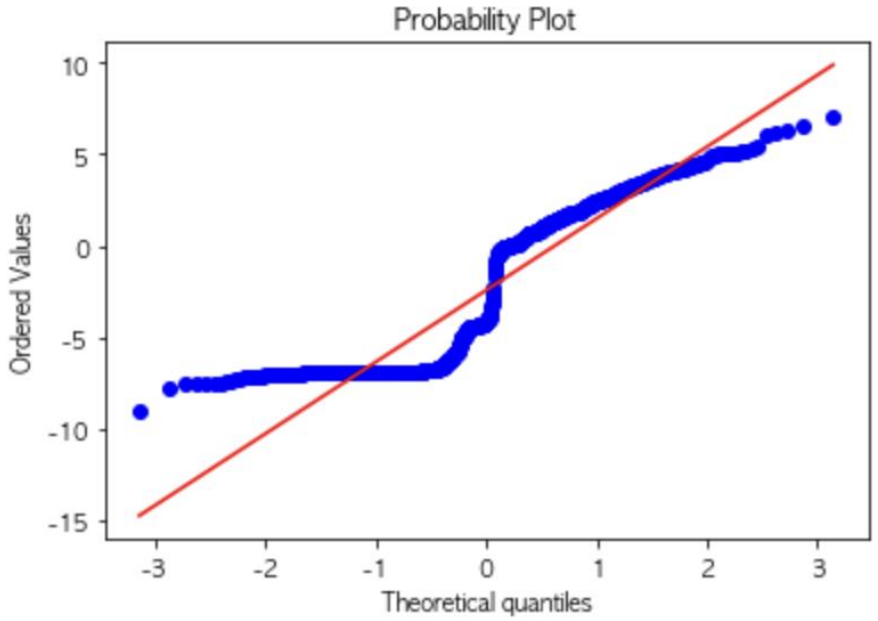
	2	3	4	5	6	7	8	9	10
stroke	53.8027	35.1510	24.1996	28.6823	31.4870	18.5069	18.6719	16.1945	14.1407
hand_hold	88.4625	49.9199	53.3108	34.6231	41.3097	27.9836	25.4731	21.9437	31.1946
knock	60.6780	46.7193	42.4346	44.1509	48.5962	31.9082	37.3695	54.1600	53.0432
quiz	2.6413	1.7429	1.6792	1.5366	1.2284	1.2358	1.5011	1.0978	0.9241
music	9.6245	5.1039	6.0264	4.5700	5.8960	2.3750	3.1950	2.5528	2.8240
gymnastics_y	3.4034	1.9898	2.1340	1.7540	1.4517	1.3407	1.3802	1.4236	1.3637

Huber Regressor

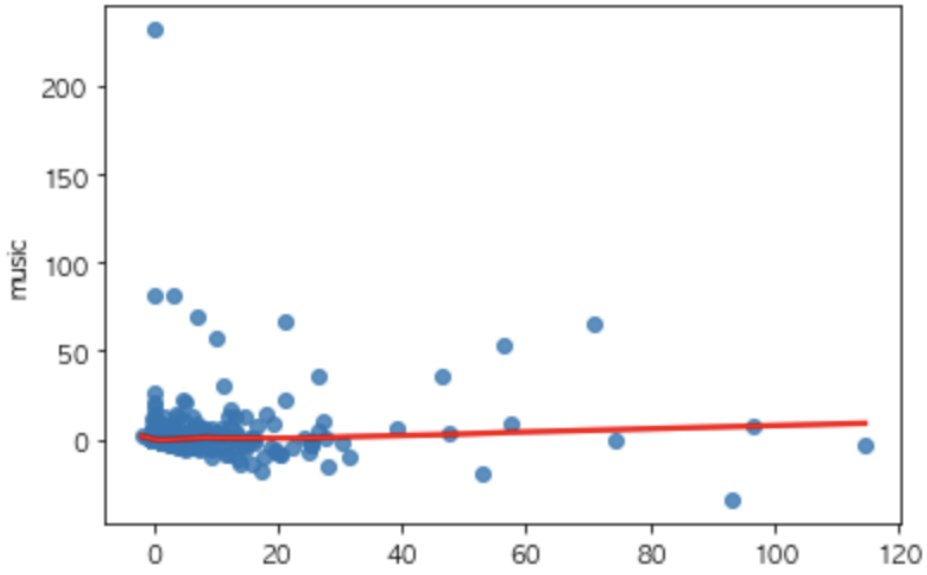
	2	3	4	5	6	7	8	9	10
stroke	37.1724	29.0602	19.2310	19.1379	26.8023	16.0330	18.2896	12.4346	13.1526
hand_hold	66.3600	39.2459	46.6682	30.2492	31.6932	23.8114	21.5812	19.8011	23.1742
knock	43.9356	36.0795	35.2805	39.0782	41.1106	24.9938	33.4878	38.1447	44.7953
quiz	2.1808	1.2994	1.3286	1.0676	0.9208	0.9783	1.3706	0.8595	0.7292
music	6.7943	4.4443	5.2313	4.0123	4.4493	1.8993	2.4224	1.9176	2.1090
gymnastics_y	2.5379	1.5661	1.7182	1.3548	1.2608	1.0672	1.0912	1.0293	1.1406

주별 사용량 예측 Modeling – Huber Regressor

Huber Regressor



QQ-plot



잔차도

주별 사용량 예측 Modeling - 결과 해석

1) 주차 경과에 따른 Error 감소

	2	3	4	5	6	7	8	9	10
stroke	53.8027	35.1510	24.1996	28.6823	31.4870	18.5069	18.6719	16.1945	14.1407
hand_hold	88.4625	49.9199	53.3108	34.6231	41.3097	27.9836	25.4731	21.9437	31.1946
knock	60.6780	46.7193	42.4346	44.1509	48.5962	31.9082	37.3695	54.1600	53.0432
quiz	2.6413	1.7429	1.6792	1.5366	1.2284	1.2358	1.5011	1.0978	0.9241
music	9.6245	5.1039	6.0264	4.5700	5.8960	2.3750	3.1950	2.5528	2.8240
gymnastics_y	3.4034	1.9898	2.1340	1.7540	1.4517	1.3407	1.3802	1.4236	1.3637

전반적으로 주차가 지날수록 Error(MAE)가 감소
→ 사람들의 사용이 **안정화** 됨에 따라 모델링 성능 향상

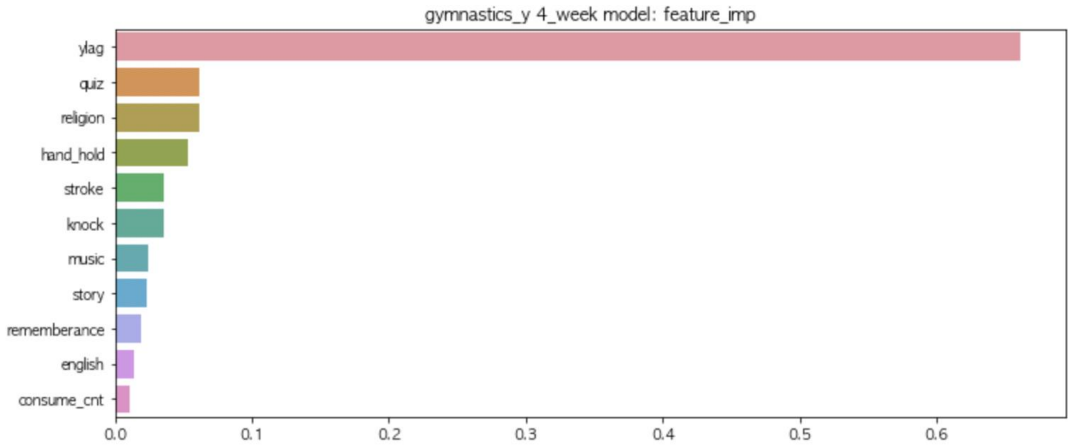
- 기능별로 MAE가 상이

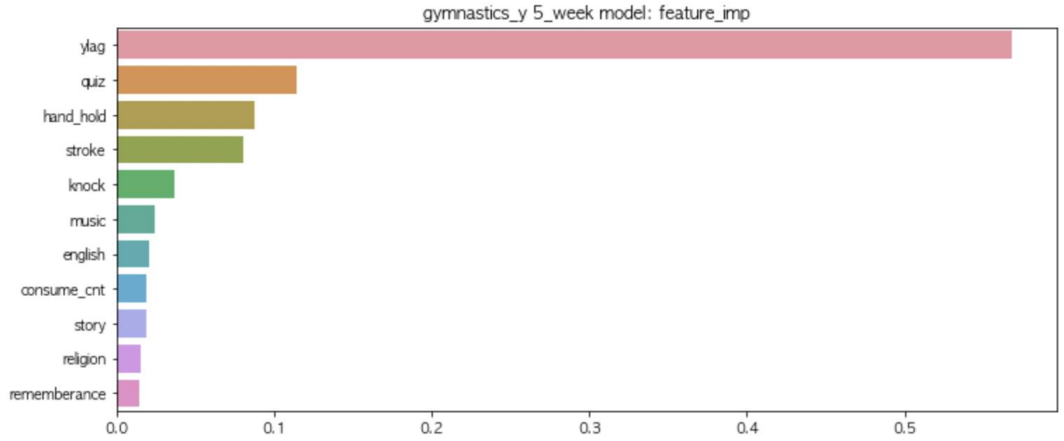
	std
stroke	138.6715
hand_hold	152.8391
knock	228.9916
quiz	6.3950
music	19.2449
gymnastics_y	6.0323

원본 데이터의 Standard Deviation을 확인하였을 때,
원본 데이터의 분산 차이에서 비롯되었음을 확인.

주별 사용량 예측 Modeling - 결과 해석

2) Feature Importance





➔ 대부분의 모형에서 **ylag 변수**가 중요

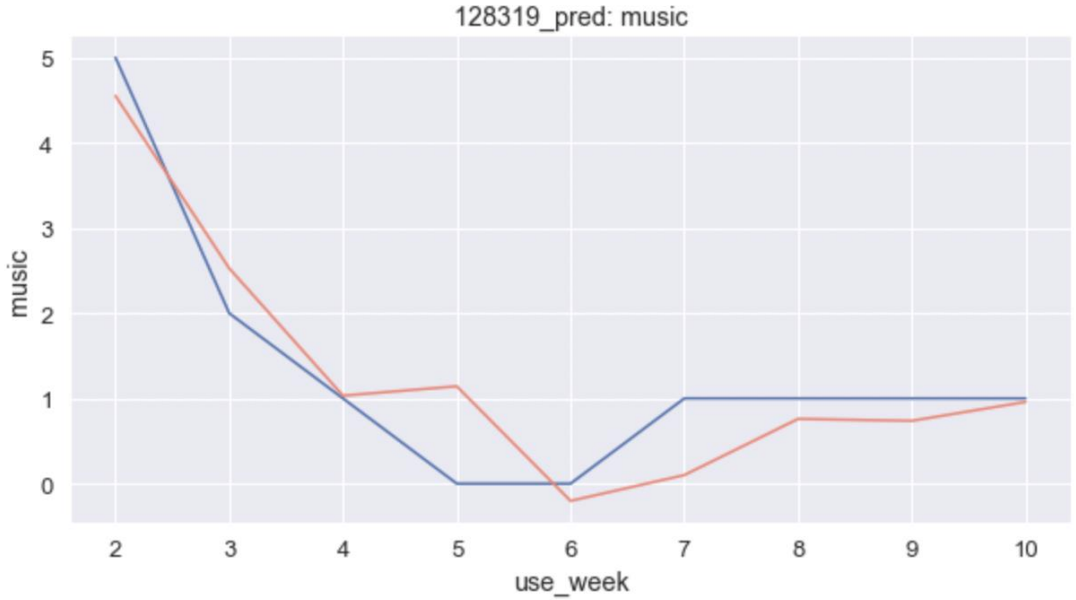
* 주차에 따른 Feature Importance의 차이는 크지 않음

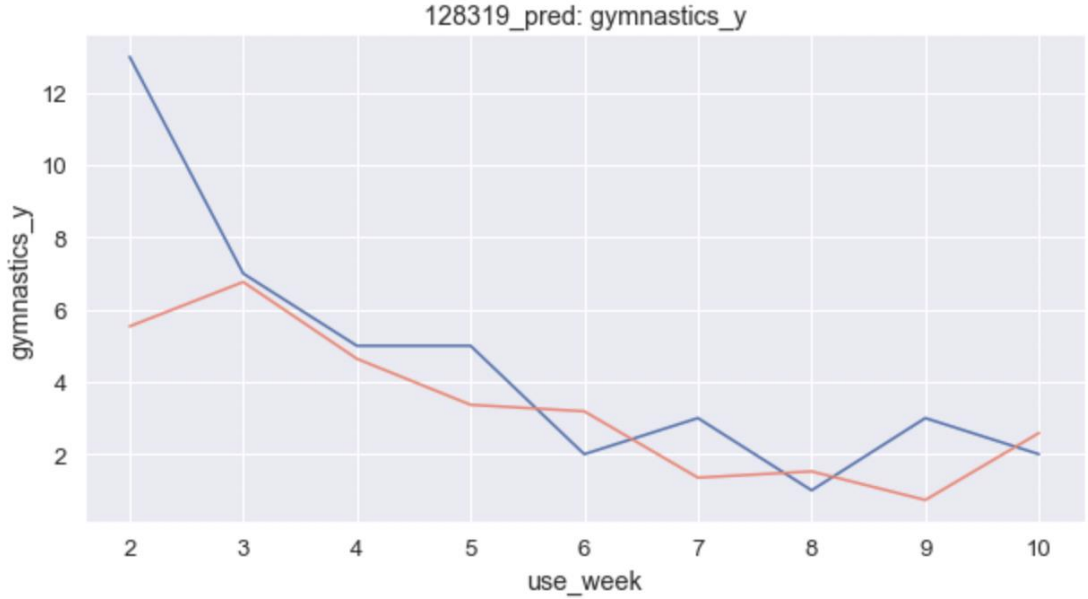
주별 사용량 예측 Modeling - 결과 해석

3) 개인별 예측 예시

Doll ID: 128319

music(음악) 기능 & gymnastics_y(체조) 기능 사용 예측





True

Predict

주별 사용량 예측 Modeling - 활용방안

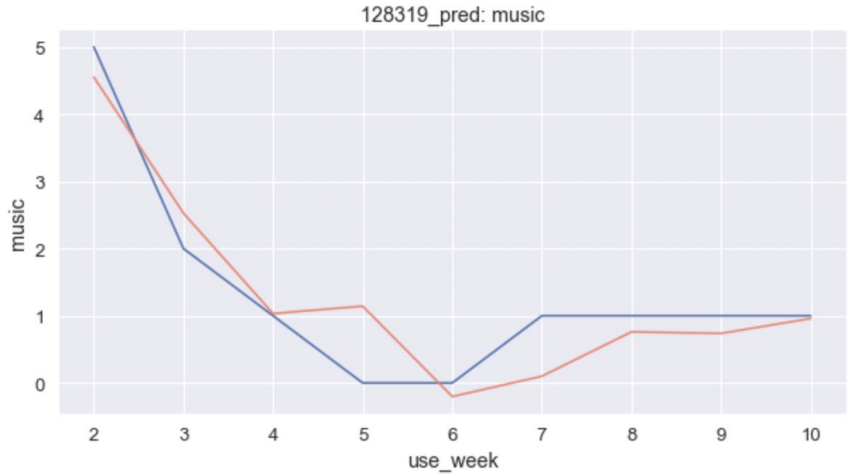
Our Models

- 인형 기능별 사용 횟수를 **주별로 예측**

개별 사용자

- 각 사용자들에게 주별 모델 적용

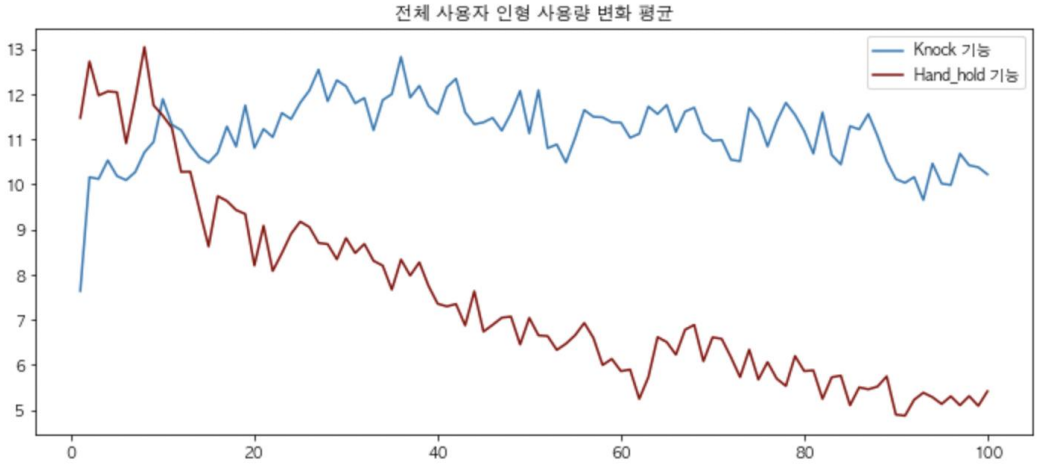
→ **개별 사용량** 사전 예측 가능



전체 사용자

- **전반적으로** 사용량이 낮아지는 기능 예측 가능

→ 해당 기능에 대한 보완책 마련



Contents

사용자 이탈 생존분석 - EDA

A user

- 사용 시작일: 2020년 5월 8일

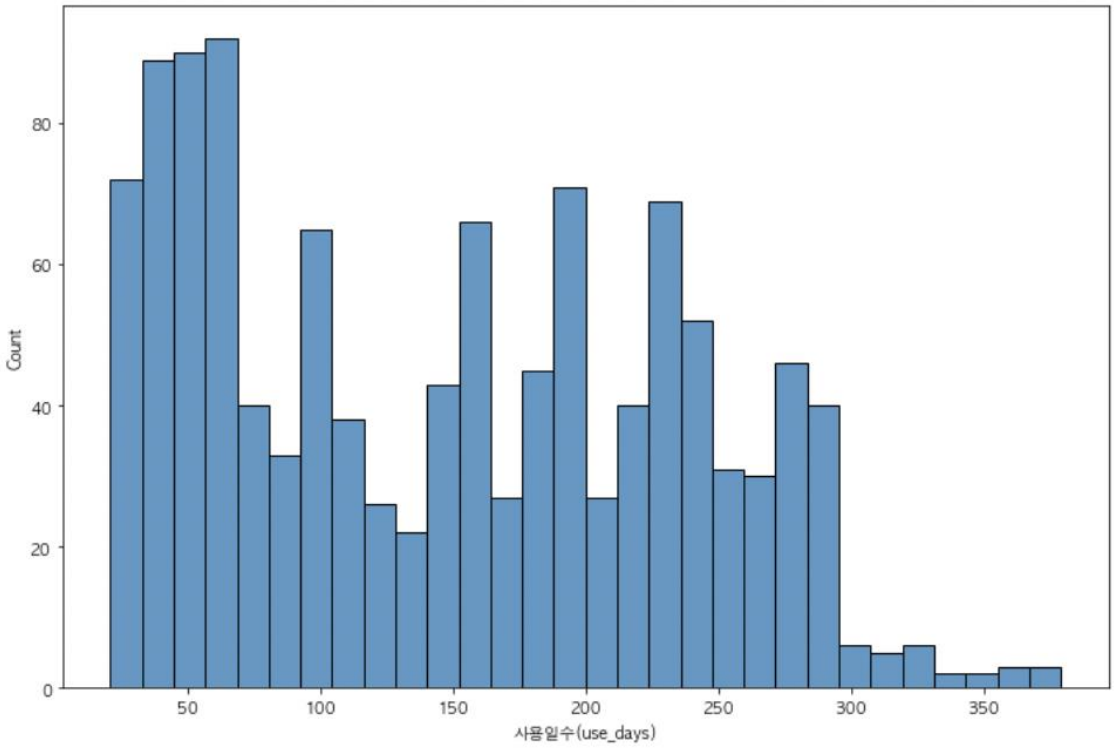
B user

- 사용 시작일: 2021년 1월 13일

서로 다른 사용 시작일

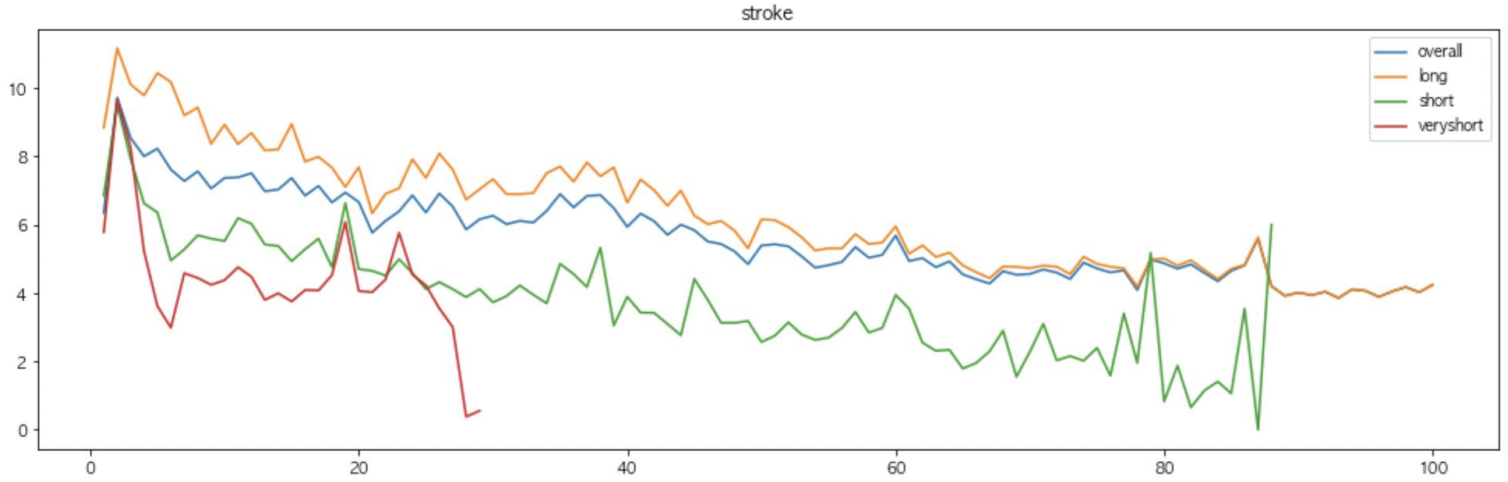
사용시작일을 1일로 맞추어
효돌 인형 **사용일수** 계산

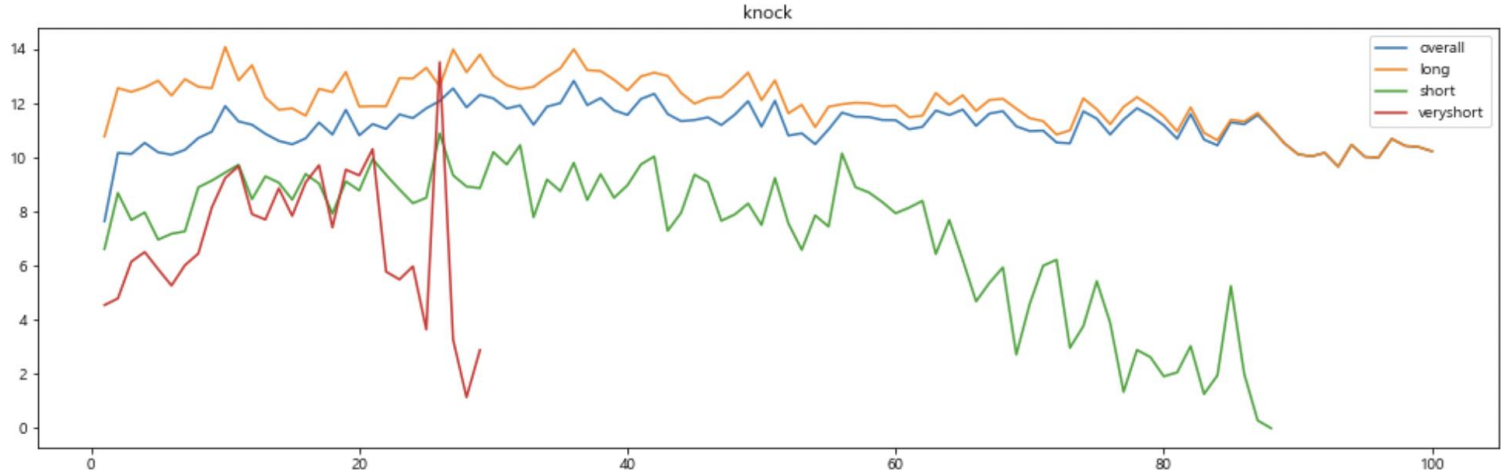
효돌 사용일수(use_days)



사용자 이탈 생존분석 - EDA

사용일수에 따른 기능별 사용 양상





long: 90일 이상

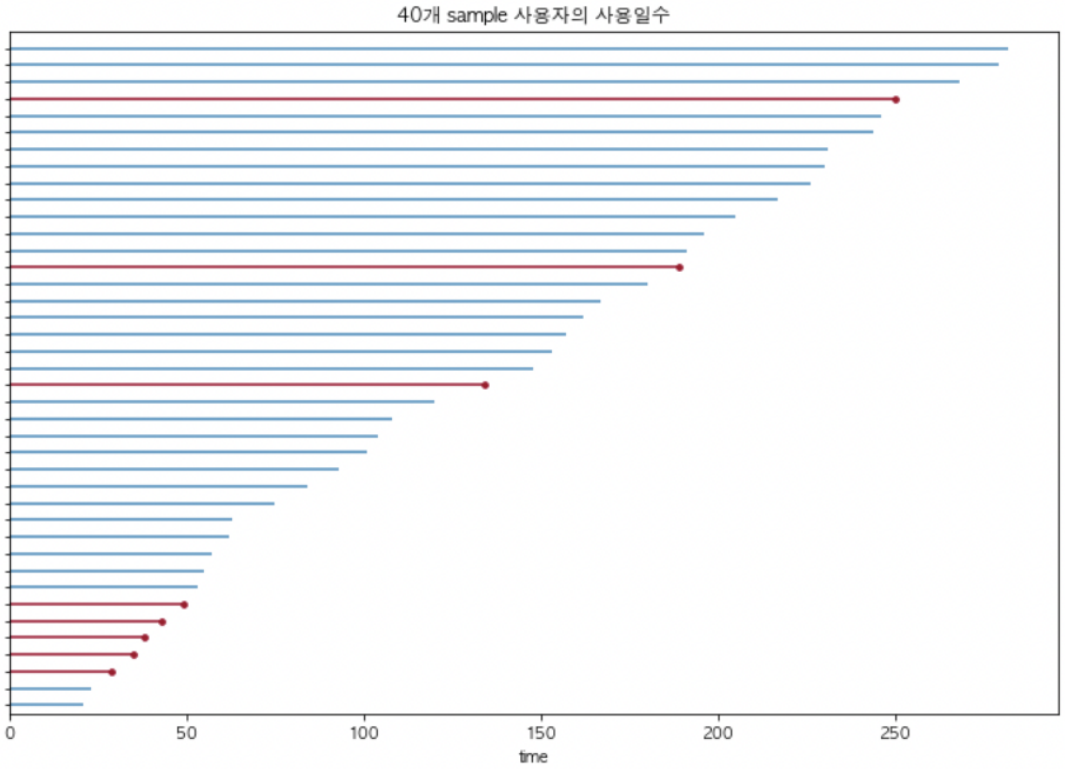
short: 15일 이상 90일 미만

very short: 15일 이상 30일 미만

사용일수에 따라
기능별 사용 양상이 다름을 확인

사용자 이탈 생존분석 - EDA

Censored Data



인형사용을 **중단**한 사용자

→ 정확한 이탈 시점을 알 수 있음

현재 **계속 사용 중**인 사용자

→ 정확한 이탈 시점을 알 수 없음
→ **중도 절단**(right-censored)된 생존 데이터 형태

생존분석 진행

- **사건**(Event): 효돌 인형 사용 중단
- **생존시간**: 인형 사용 기간
- 중도절단 **원인**: 현재 계속 사용

사용자 이탈 생존분석 - EDA

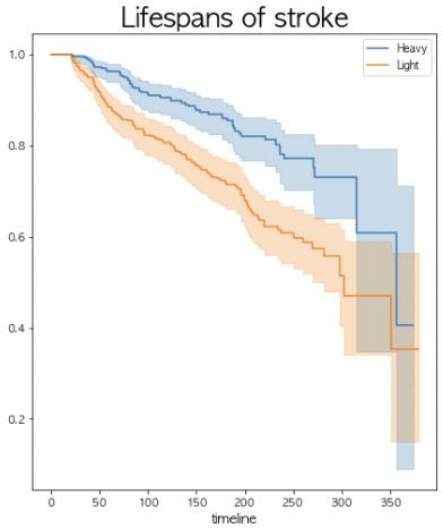
Kaplan Meier Plot

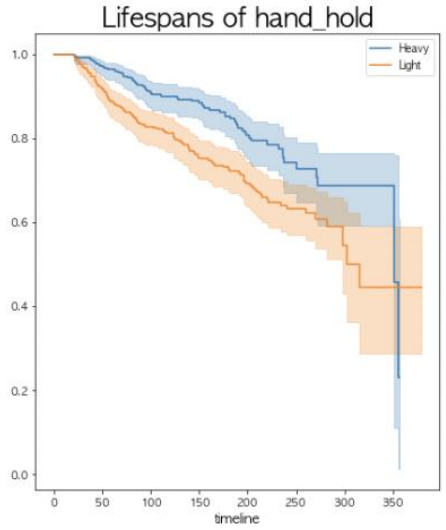
- 독립변수에 따라 **survival function**에 차이가 있는지 확인

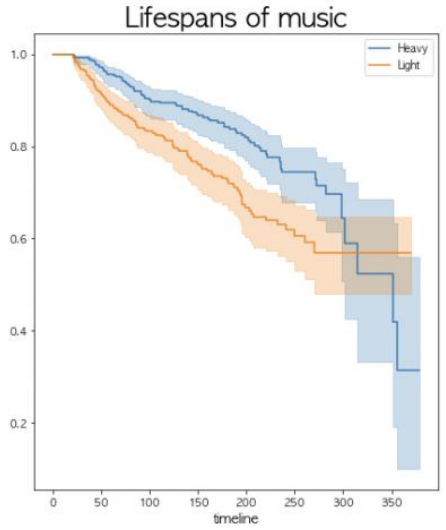
- 각 변수의 median을 기준으로 0, 1로 더미화하여 **Kaplan Meier Plot**을 그림

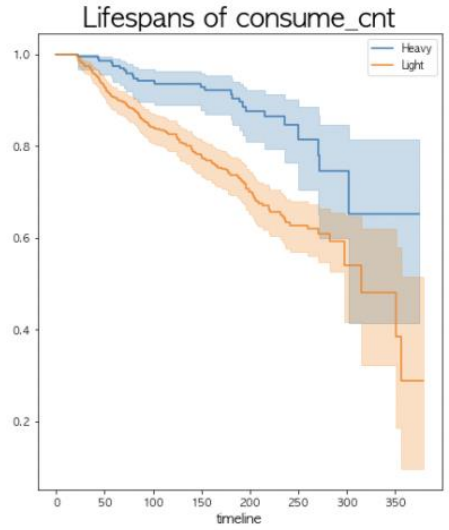
  (단, median 값이 0인 경우 75% quantile을 기준으로 함)

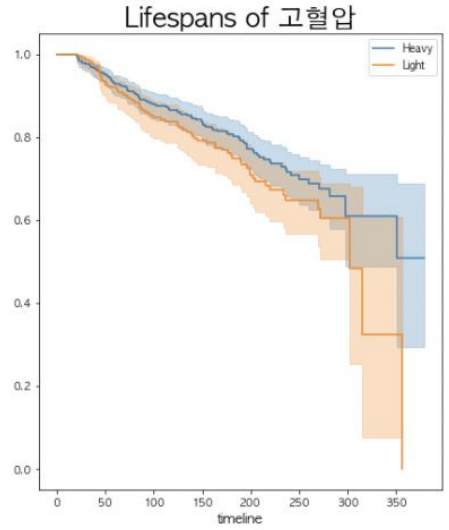
- 차이가 보이는 경우(두 곡선이 차이가 나는 경우)가 많았으나 차이가 뚜렷하지 않은 변수들도 존재







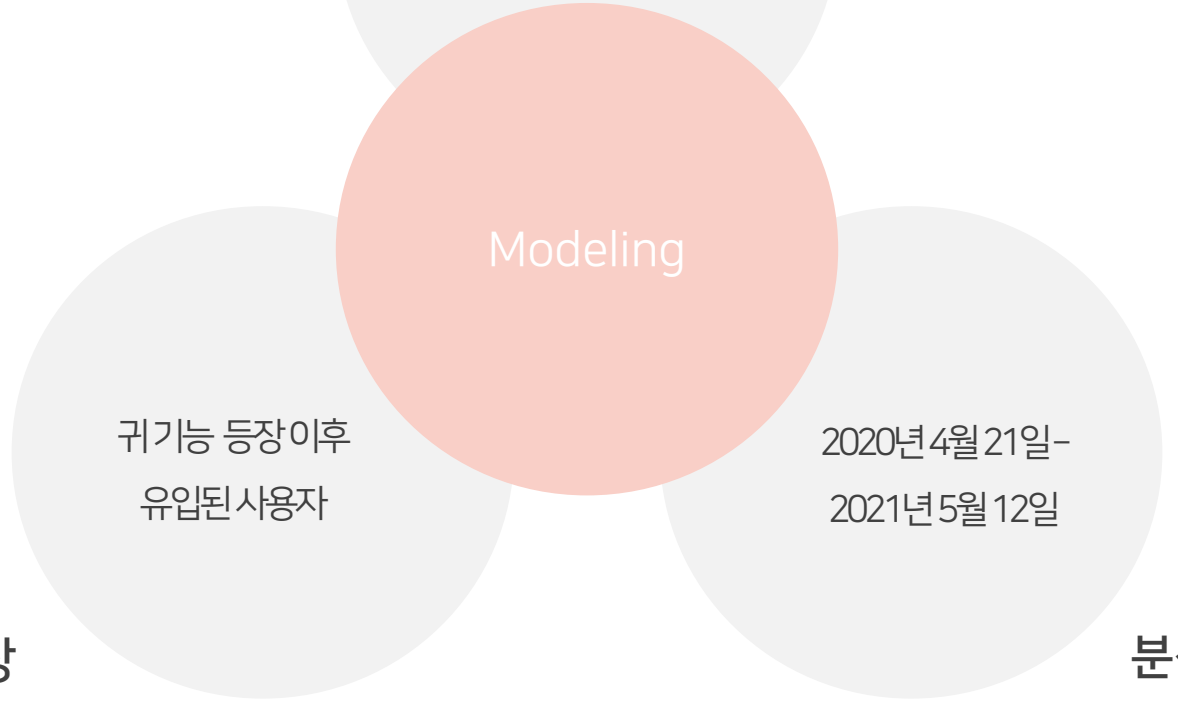




사용자 이탈 생존분석 - 전처리

분석 목표

3주 간의 데이터를 이용하여
인형 사용 이탈 시점 예측



분석 대상

귀기능 등장 이후
유입된 사용자

분석 기간

2020년 4월 21일 -
2021년 5월 12일

사용자 이탈 생존분석 - 분석 개요

STEP 1

사용일수 및
인형 **사용행태**
데이터 구축

>>

STEP 2

Train / Test Split

Train: **826**
(censored- 642, 77%, Not censored- 184, 23%),

Test: **355**
(censored- 276, 77%, Not censored – 79, 23%)

>>

STEP 3

Numeric data
Scaling
&
Feature Selection

>>

STEP 4

생존분석 모델링 진행

- Log-normal AFT
- Cox Proportional Hazard

사용자 이탈 생존분석 - 전처리

사용변수

기본 정보

성별, 종교, 기상, 아침, 점심, 저녁, 취침, 환기, 산책, 고지혈, 고혈압, 당뇨
(from 인형 기본 정보 테이블(scc_doll))

Log Data

인형 머리 쓰다듬 횟수(stroke), 손버튼 누름 횟수(hand_hold), 등 두드림 횟수(knock), 체조 실행 횟수(gymnastics), 퀴즈(brain_timer), 이야기 수행 횟수(story), 종교말씀 수행 횟수(religion), 트로트 수행 횟수(music), 회상 수행 횟수(remembrance)
(from 인형 5분 주기 접속 로그 테이블(log_doll), 귀 프로그램 수행 로그(scc_ear_function_log))

약 복용

약복용 기록 횟수(consume_cnt)
(from 인형 약복용 로그 테이블(log_doll_drug_consume))

3주간의 데이터를 Sum

사용자 이탈 생존분석 - Modeling

후보 모형

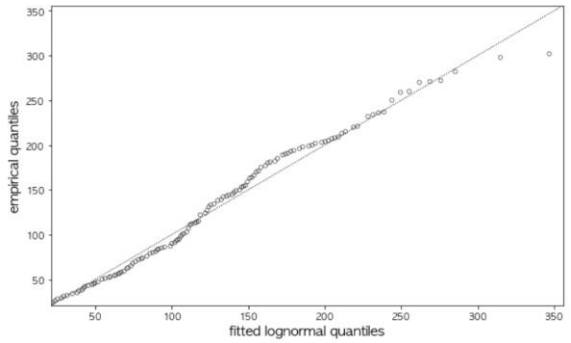
Log-normal Accelerated Failure Time

- **Parametric** Model

각 변수가 생존시간을 가속 또는 감속한다고 가정하는 선형 모형
noise term이 log-normal 분포를 따른다고 가정한다.
생존시간 자체에 대한 설명변수의 효과를 모형화한다.

$$\log(t_i) = XB + \sigma\epsilon_i$$

- 가정사항: Log normal 분포를 따르는지 체크



따르는 것으로 보임

Cox Proportional Hazard

- **Semi - Parametric** Model

생존시간에 대해 분포를 가정하지 않는 비모수적 특정을 가지고 있지만,
회귀계수를 추정한다는 점에서 모수적 형태를 가진 모형이다.
Relative hazard(위험함수)에 대한 설명변수의 효과를 모형화한다.

$$h(t|x) = h_0(t)e^{XB}$$

- 가정사항: Schoenfeld residuals를 이용한 proportional hazard 가정 체크

		test_statistic	p	-log2(p)
brain_timer	**km**	2.37	0.12	3.02
	rank	0.56	0.45	1.14
consume_cnt	**km**	0.05	0.82	0.29
	rank	0.00	0.99	0.01
english	**km**	1.37	0.24	2.05
	rank	0.29	0.59	0.76

가정사항 만족하지 못한
당뇨, 종교 변수 존재

사용자 이탈 생존분석 – 모형결과

사용한 변수

- 총 11개 변수: 손버튼 누름 횟수(hand_hold), 등 두드림 횟수(knock), 인형 머리 쓰다듬 횟수(stroke), 트로트 수행 횟수(music), 약 복용 횟수(consume_cnt), 퀴즈(brain_timer), 이야기 수행 횟수(story), 회상 수행 횟수(remembrance), 취침, 기상

* 변수 선정 기준: 모형에 각 독립변수 하나만 적합하여 concordance-index를 기준으로 평가 (scikit-learn의 feature selection 방법)

평가 지표

- 데이터를 7:3으로 train, test 데이터로 나누어 **Concordance index** 평가

C-index	Train C-index	Test C-index
Log- normal AFT	0.66	0.71
Cox-proportional hazard	0.66	0.69

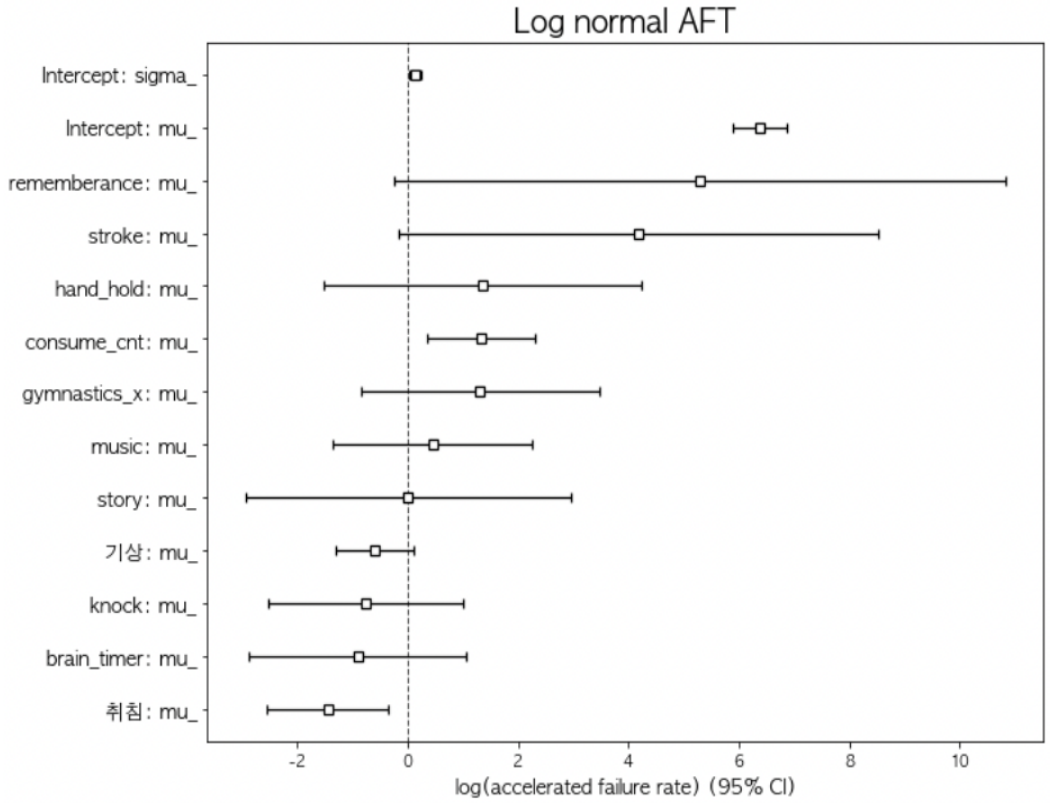
* Concordance index: 생존분석모형에서 관측된 실제 사용일수와 예측된 위험도와의 **랭크 상관관계**

사용자 이탈 생존분석 – 모형결과(Log-normal AFT)

유의한 변수

* 괄호 안의 값: p-value

- 약 복용 횟수(<0.0005), 취침(0.03), 인형 머리 쓰다듬 횟수(0.04), 기상(0.07)



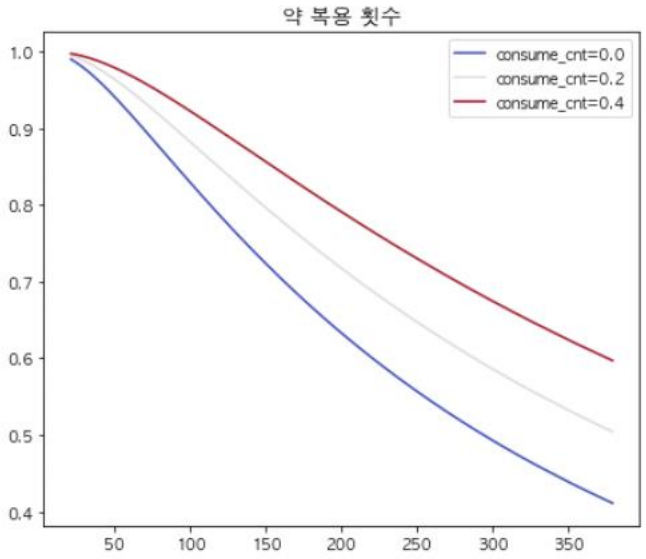
변수	계수	Scaling 보정 계수
약복용횟수(consume_cnt)	1.33	0.0233
취침	-1.45	-0.145
머리 쓰다듬 횟수(stroke)	4.17	0.0004549

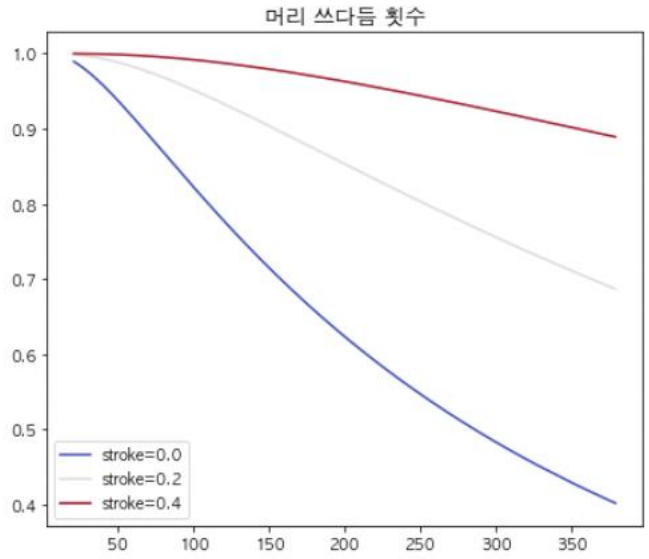
- 다른 변수가 같은 값을 가질 때, 약 복용 횟수가 1번 늘어나면, exp(0.0233) = 1.023만큼, 즉 생존확률이 **2%** 늘어난다.
- 다른 변수가 같은 값을 가질 때, 취침 설정시간이 1시간 늦어지면, exp(-0.145) = 0.86만큼, 즉 생존확률이 **14%** 줄어든다.
- 다른 변수가 같은 값을 가질 때, 머리 쓰다듬는 횟수가 100회 늘어나면, exp(0.04549) = 1.0465만큼, 즉 생존확률이 **4%** 늘어난다.

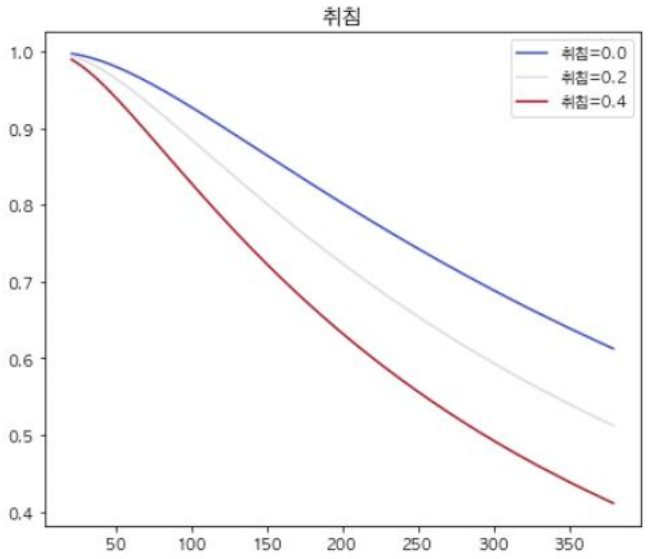
사용자 이탈 생존분석 – 모형결과(Log-normal AFT)

Partial Effect Plot

- 최종 모형에서 p-value가 0.05 이하인 변수들에 대해 **Partial effect plot**을 그리고, 이를 통해 각 변수 값이 달라짐에 따라 survival function의 차이가 있는지 확인할 수 있다.
- 적합한 모형에서 다른 변수가 모두 같은 값을 가질 때, 해당 변수의 값이 달라짐에 따라 survival function의 차이를 통해 partial effect를 확인한다.







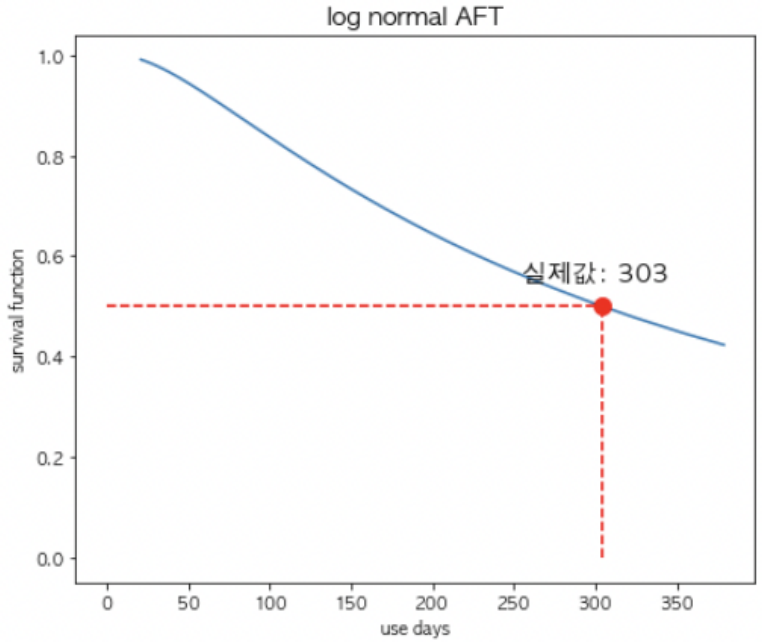
* 각 변수는 min-max scaling을 해주었기 때문에 0과 1사이의 값을 가진다.

사용자 이탈 생존분석 – 모형결과(Log-normal AFT)

개별 사용자 Survival Function 예측

Doll Id: 128464

hand_hold	1.0000
knock	1.0000
stroke	0.0000
music	64.0000
consume_cnt	9.0000
gymnastics_x	0.0000
brain_timer	1.0000
story	0.0000
rememberance	0.0000
취침	23.0000
기상	7.0000



Survival function의 median 값이 실제값과 유사함을 확인할 수 있다.

사용자 이탈 생존분석 – 활용방안(1)

Useful Results

위험도에 따라 어떤 사용자가 먼저 인형 사용을 관둘 것인지 예측 가능

ex) 49 일차 survival function

doll_id	133697	132073	129538	129654	132202	133003	132245	130932	133061	133745	131937	132343
survival_function	0.8621	0.8773	0.8823	0.8937	0.8947	0.8954	0.8967	0.8971	0.8996	0.9001	0.9007	0.9010

활용방안

- 사용자 위험도에 따라 인형 사용을 **모니터링** 할 사용자 선별 가능

사용자 이탈 생존분석 – 활용방안(2)

활용방안

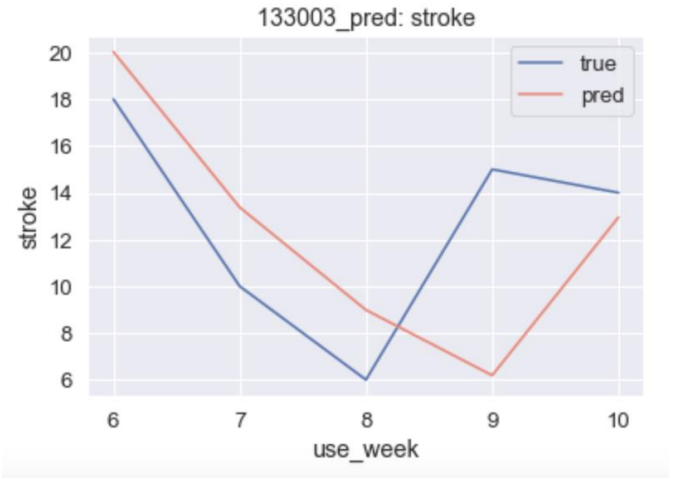
1) 사용자 이탈 예측 모델을 통하여 **인형 사용을 관둘 위험도가 높은** 사용자를 선별

2) 유의한 변수들에 대하여 선정된 대상자들의 현재까지의 사용행태를 살펴보고, **주별 예측 모형**을 통해 앞으로의 사용행태를 예측하였을 때, 어떤 기능이 문제인지 파악.

→ 사용자 **개별 위험도**에 따라 어떤 기능 사용에 대해 조치를 취할지 파악 가능

ex) 49 일차 survival function

doll_id	133697	132073	129538	129654	132202	133003	132245	130932
survival_function	0.8621	0.8773	0.8823	0.8937	0.8947	0.8954	0.8967	0.8971



stroke 값이 빠르게 감소

Contents

결론 - 의의

의의

인형 사용의 긍정적 효과 입증

: Survey Modeling을 통해 인형을 사용하는 것이 노인분들의 우울증과 생활관리 개선에 긍정적인 상관관계가 있음을 확인.

사용 주차별 사용자 행태 파악

: 사용 주차별 예측 Modeling을 통해 시간의 흐름에 따라 사용자의 행동 패턴이 어떻게 변하는지 파악할 수 있는 모델을 생성.

조기 이탈자 파악

: 사용자 이탈 생존분석을 통해 조기에 사용을 중단할 사람들을 파악하고 관리할 수 있는 모델을 생성.

추후 발전 가능성

: 추후에 데이터를 추가하여 모델에 반영하면 더 정확한 결과들을 얻을 수 있음.

결론 – 한계점 및 제언

한계점

✓ **데이터의 불완전성**

ex) 응급푸쉬알림(emergency push)

: 불가능한 값을 가진 데이터들의 존재로 인하여 데이터의 신뢰성이 떨어지거나, 모델링에 사용하지 못한 변수들이 존재

✓ **데이터의 부족**

: 특정 기능들의 데이터 부족으로 인하여 모델링에서 성능 저하 발생 ex) 귀 기능 등

제언

추가 설문조사

: 효돌의 구체적인 성능을 입증하기 위하여 추가 설문조사 진행

Error 데이터 판단

: Error 데이터를 판단할 수 있는 장치를 마련한다면, 추후에 더 정확한 분석이 가능할 것

Reference

- Art B. Owen (2006), A robust hybrid of lasso and ridge regression.

  https://statweb.stanford.edu/~owen/reports/hhu.pdf

- Peter J. Huber, Elvezio M. Ronchetti, Robust Statistics Concomitant scale estimates, pg 172

- Davidson-Pilon, (2019). lifelines: survival analysis in Python. Journal of Open Source Software, 4(40), 1317,

  https://doi.org/10.21105/joss.01317

- Collett, D. (2014). Modelling survival data in medical research (3rd ed.). Chapman & Hall/CRC.

- Majeed, Abdul-Fatawu. (2020). Accelerated Failure Time Models: An Application in Insurance Attrition.

Q & A


감사합니다.