게임유저 잔존가치를 고려한 고객 이탈 예측 모형

Big Contest 2019

라이온킹
김미소, 김정환, 김현준, 조예린

Contents

1. Lineage에 대한 이해

모델 구축 전 게임에 대한 이해

1. Lineage에 대한 이해

리니지 기본 정보

1998년 9월부터 서비스한 장수 **MMORPG 온라인 게임**으로, 한국 MMORPG 게임의 기반이 되었다.

현재까지도 **모바일로의 IP확장, 리마스터** 등을 통하여 주로 **중장년층 게이머**들에게 많은 사랑을 받고 있다.

1. Lineage에 대한 이해

모델 구축 시 고려한 리니지 Key Point

풍성한 사회적 활동 컨텐츠

- 혈맹을 중심으로 한 공성전, 전쟁 등이 활발하며,
  유저들이 게임속에서 하나의 이야기를 써나갈 수 있음(ex. 바츠해방전쟁)
- 혈맹과 같은 사회적 활동은 게임 잔존율에 많은 영향을 준다는
  선행 연구들이 존재함[1]

[혈맹들 간의 공성전]

아이템 강화(enchant) 시스템

- 과금을 통해 무기, 갑옷과 같은 아이템을 강화할 수 있음
- 강화된 '진명황의 집행검' 등 최상급 아이템은
  억대가 넘는 가격으로 인해 사회적 이슈가 되기도 했음

[진명황의 집행검]

참고: [1] '온라인 게임 내 유저 간 상호작용 분석을 통한 유저 이탈 예측기(문주연, 2018)', '온라인 게임 내 최고 레벨 유저의 이탈 분석(박건우, 2017)'

1. Lineage에 대한 이해

모델 구축 시 고려한 리니지 Key Point

미니게임(낚시 등)

- 유령의 집, 낚시, 무한대전 등 미니 게임에 참가함으로써
  각종 소모형 아이템, 장비 및 아데나 획득
- 낚시의 경우, 리뉴얼 후 경험치 효율이 사냥을 앞서는 경우가 존재해
  많은 유저들이 낚시에 시간을 투자함

[낚시 중인 유저들]

지배의 탑

- 각기 다른 서버의 유저들이 모이는 통합 전장으로,
  각 서버의 고레벨 유저들의 경쟁이 일어나는 곳
- 각 층마다 보스가 등장하며 매주 보스 몬스터에 대한 공헌도를 집계하여
  서버 별로 공헌도 측정 후 보상 지급

[지배의 탑의 관문 오만의 탑]

1. Lineage에 대한 이해

공홈 커뮤니티로 파악한 이탈 요인

과도한 과금 요소

- 캐릭터의 성장시키는데 필요한 과금에 피로감을 느껴 이탈

부족한 사회활동

- 게임에 접속해도 컨텐츠를 함께 즐길 유저가 부족하여 이탈

진입장벽으로 인한 신규 유저 이탈

- 과금으로 인한 진입장벽과 더불어 고레벨 유저들의 막피, 썰자, 사냥터 통제 등으로 레벨이 낮은 신규 유저가 잔존하기 어려움

2. EDA 및 데이터 전처리

탐색적 분석에 기반한 Preprocessing

2. EDA 및 데이터 전처리

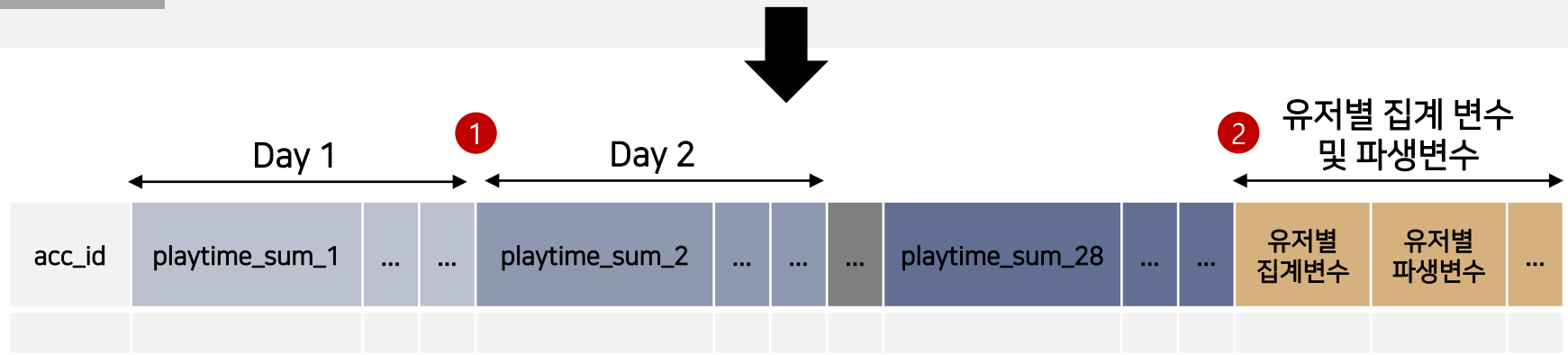
0. 데이터프레임 구조

5개의 데이터 셋에서 각 acc_id 별로 시계열 데이터 정제와 파생 변수 생성을 통해 데이터프레임을 만들었다.

- Activity
- Payment
- Combat
- Pledge
- Trade

1. 시계열 특성 고려 변수 (Time Series Variable)
2. 유저별 집계 변수 및 파생변수

- 시계열적 특성을 고려하여 각 acc_id 별 데이터를 변수명_1, 변수명_2 식으로 1일차, 2일차, ..., 28일차까지의 시계열 컬럼 생성
- 여러 개의 char_id를 지닌 유저가 있음을 고려하여, acc_id 별로 모든 캐릭터의 데이터 합산
- 게임에 접속하지 않아 생긴 결측치의 경우, 0으로 대체

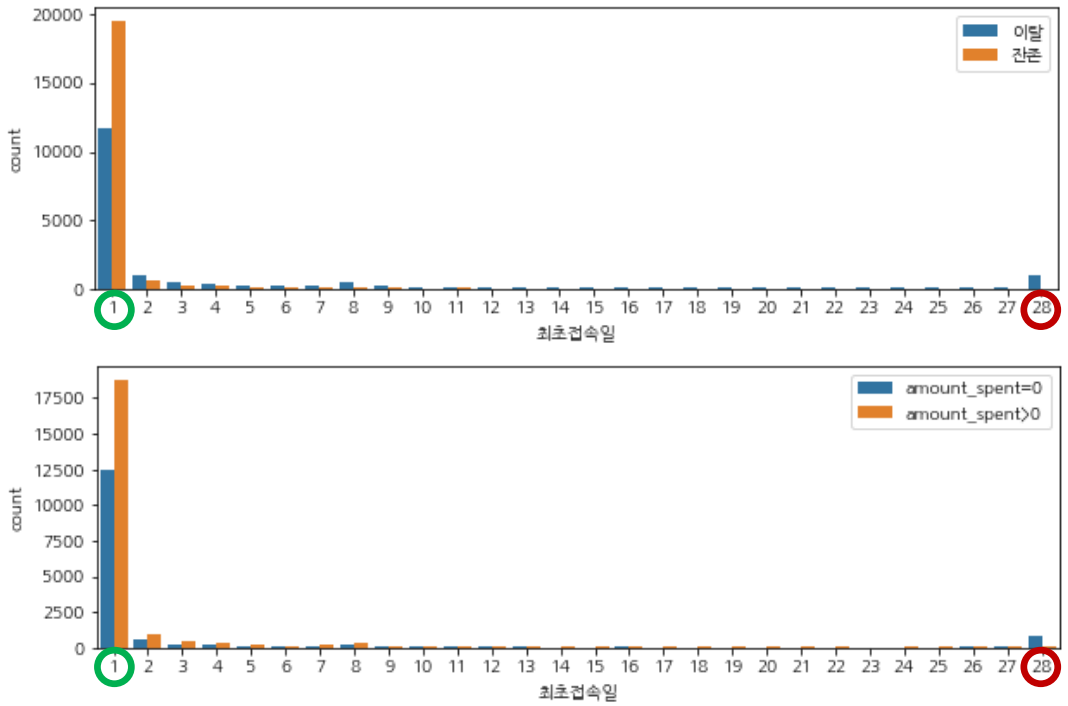


2. EDA 및 데이터 전처리

1. Activity - 파생변수

최초접속일에 따라 잔존/이탈 유저의 비율과 과금/비과금 유저의 비율이 다르다는 점에서 이와 관련한 파생변수를 만들었다.

[최초접속일 별, 이탈/잔존 유저 count와 과금/비과금 유저 count]



최초접속일이 이를 수록
잔존 유저와 과금 유저가 더 많이 존재

최초접속일: acc_id가 최초로 접속한 day

최초접속주차: 28일을 7일 기준으로 나누었을 때, 최초 접속한 주차

첫주차접속일 & 4주차접속일: 1주차&4주차 접속일 수 count

캐릭터별총접속일: acc_id별 유저가 가진 모든 캐릭터의 접속일 총합

ID별총접속일: acc_id별 접속일 수

최초접속이후비접속횟수: 최초 접속 이후 비접속한 일 수

접속28일캐릭터수: 28일동안 매일 접속한 캐릭터의 수

첫주플레이시간평균 & 사주플레이시간평균: 1주차&4주차 playtime 평균

최대연속접속일수: 연속으로 접속한 기간 중 최대 기간의 일 수

최대연속미접속일수: 연속으로 비접속한 기간 중 최대 기간의 일 수

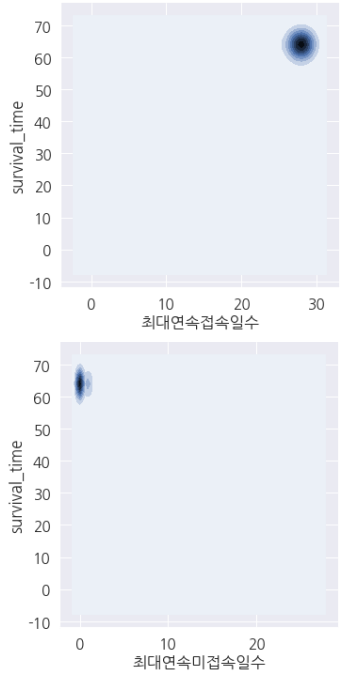
day_playpattern: 접속패턴으로 Kmeans Clustering한 결과 군집 번호

2. EDA 및 데이터 전처리

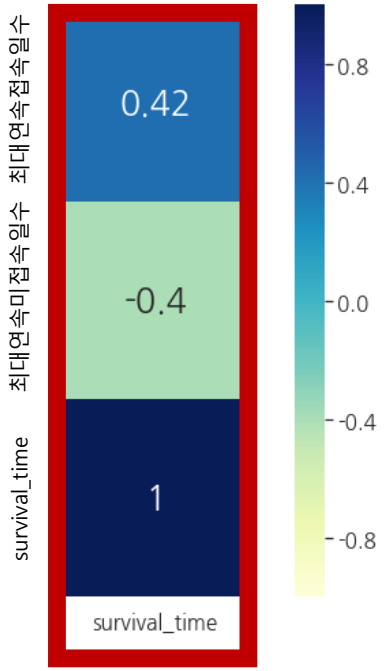
1. Activity - 파생변수

유저들의 접속패턴이 survival time에 영향을 준다는 점에서 이와 관련한 파생변수를 만들었다.

[Joint plot]



[상관계수 히트맵]



최대연속접속일수가 높을수록 survival time이 긴 경향이 있었으며,
최대연속미접속일수가 높을수록 survival time이 짧은 경향이 있었음

최초접속일: acc_id가 최초로 접속한 day

최초접속주차: 28일을 7일 기준으로 나누었을 때, 최초 접속한 주차

첫주차접속일 & 4주차접속일: 1주차&4주차 접속일 수 count

캐릭터별총접속일: acc_id별 유저가 가진 모든 캐릭터의 접속일 총합

ID별총접속일: acc_id별 접속일 수

최초접속이후비접속횟수: 최초 접속 이후 비접속한 일 수

접속28일캐릭터수: 28일동안 매일 접속한 캐릭터의 수

첫주플레이시간평균 & 사주플레이시간평균: 1주차&4주차 playtime 평균

최대연속접속일수: 연속으로 접속한 기간 중 최대 기간의 일 수

최대연속미접속일수: 연속으로 비접속한 기간 중 최대 기간의 일 수

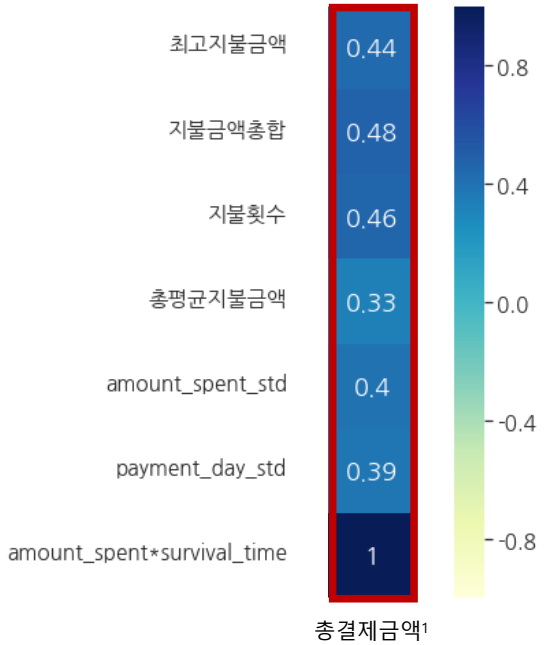
day_playpattern: 접속패턴으로 Kmeans Clustering한 결과 군집 번호

2. EDA 및 데이터 전처리

2. Payment - 파생변수

유저들의 결제금액을 예측하는데 과거 과금액과 과거 결제 주기와 같은 결제 패턴이 중요하다고 고려되어 이와 관련한 파생변수를 만들었다.

[payment 파생변수와 총결제금액[1] 간 상관계수 히트맵]



과거 결제 패턴을 드러내는 파생변수가 Train 데이터 기간 동안의 총결제금액과 높은 상관계수를 보임

최고지불금액: acc_id별 가장 많이 지불한 금액

지불금액총합: acc_id별 28일동안 지불한 금액의 총액

지불횟수: acc_id별 28일동안 지불 횟수

총평균지불금액: acc_id별 평균 지불 금액

amount_spent_std: acc_id별 지불 금액의 표준편차

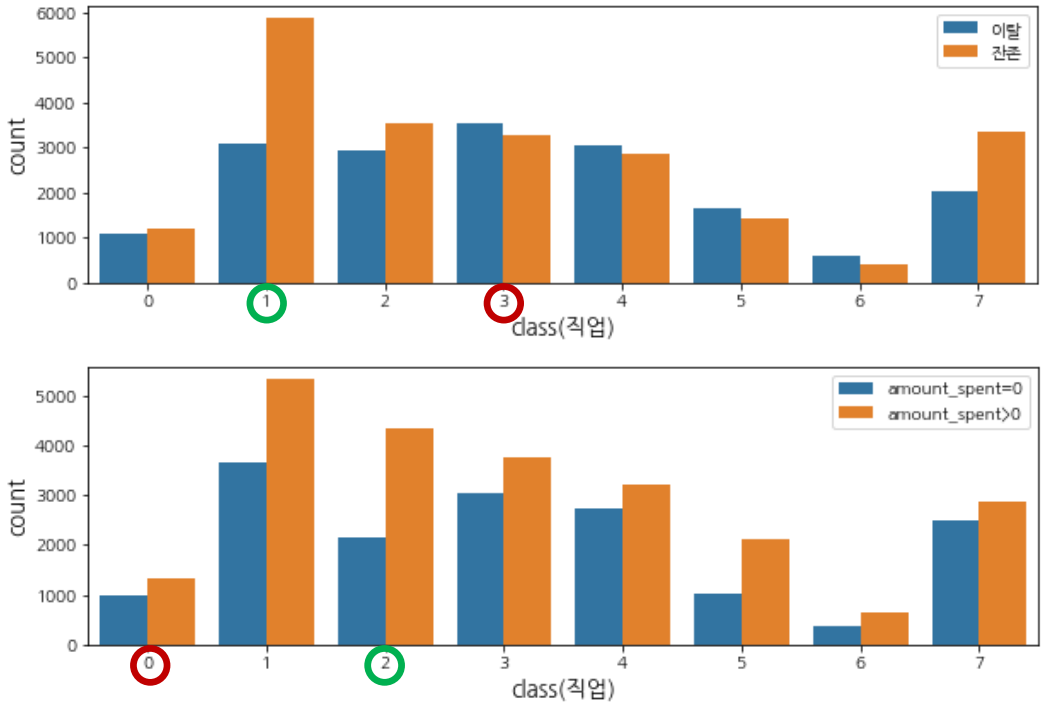
payment_day_std: acc_id별 지불 day 수의 표준편차

참고: [1]총결제금액 = amount_spent * survival_time

3. Combat - 파생변수

유저가 플레이하는 캐릭터의 직업과 레벨에 따라 survival time과 amount spent의 양상이 다르다는 점에서
이와 관련한 파생변수를 만들었다.

[class(직업)별, 이탈/잔존 유저 count와 과금/비과금 유저 count]



기사(1), 전사(7), 요정(2) 직업의 경우 잔존 유저가 더 많았으며
요정(2), 용기사(5), 군주(0) 직업의 경우 과금 유저의 비율이 더 많음

class_0

... : 플레이 시간이 가장 긴 캐릭터의 직업(해당:1, 이외:0)

class_7

level_0

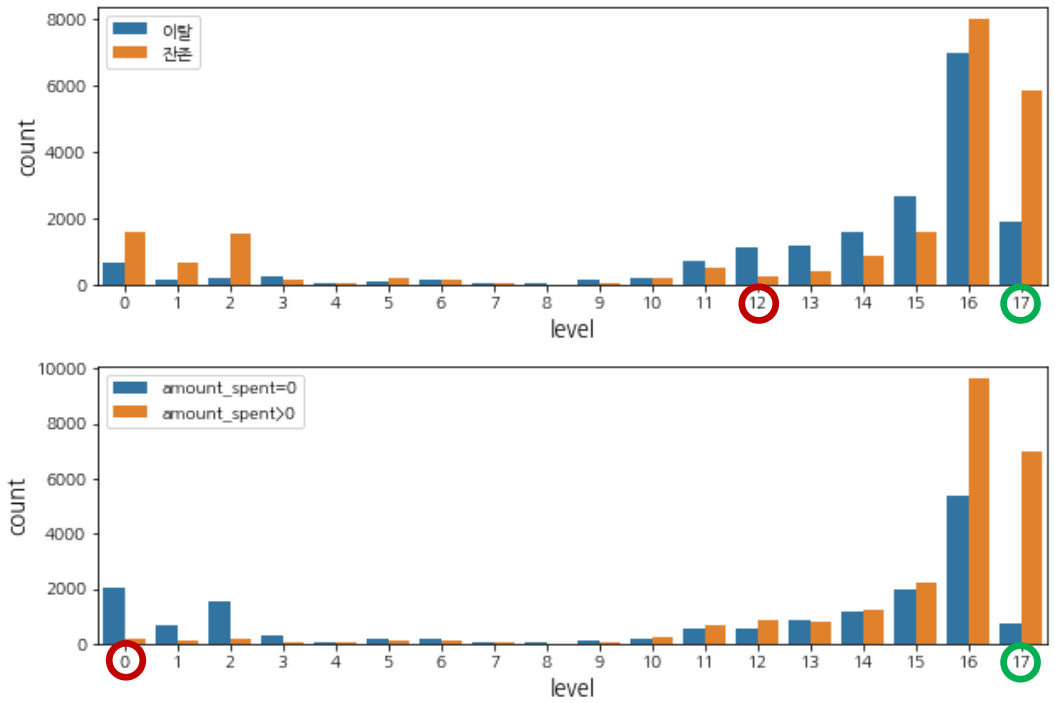
... : 플레이 시간이 가장 긴 캐릭터의 레벨(해당:1, 이외:0)

level_17

3. Combat - 파생변수

유저가 플레이하는 캐릭터의 직업과 레벨에 따라 survival time과 amount spent의 양상이 다르다는 점에서
이와 관련한 파생변수를 만들었다.

[level별, 이탈/잔존 유저 count와 과금/비과금 유저 count]



중간 레벨에서 이탈이 주로 발생하며
높은 레벨일수록 과금 유저가 더 많은 비율로 나타남

class_0

... : 플레이 시간이 가장 긴 캐릭터의 직업(해당:1, 이외:0)

class_7

level_0

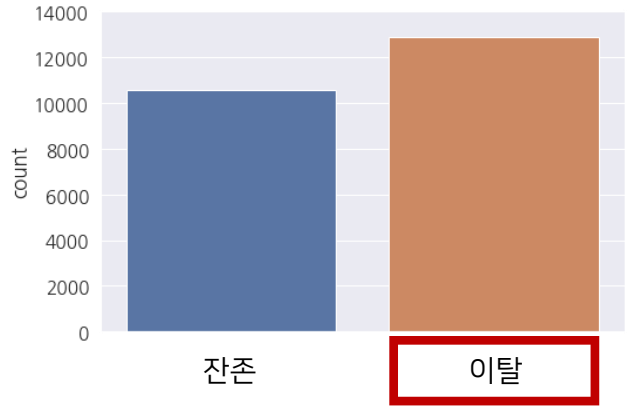
... : 플레이 시간이 가장 긴 캐릭터의 레벨(해당:1, 이외:0)

level_17

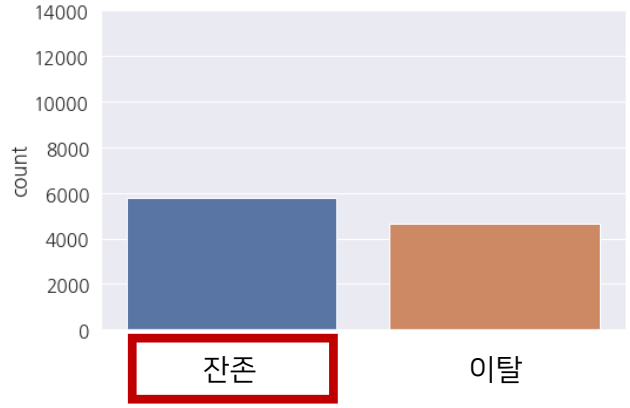
4. Pledge - 파생변수

유저가 속한 혈맹의 규모가 유저의 잔존/이탈 여부와 상관이 있다는 점에서 이와 관련한 파생변수를 만들었다.

[평균[1] 미만 혈맹원수를 지닌 혈맹에서 활동하는 유저의 잔존/이탈 여부 count]



[평균 이상 혈맹원수를 지닌 혈맹에서 활동하는 유저의 잔존/이탈 여부 count]



혈맹원수: acc_id가 가입한 혈맹의 규모를 대변하는 변수

acc_id별 가장 playtime이 긴 pledge_id

acc_id	pledge_id	혈맹원수
5	25467	10.0
8	21094	40.0
17	1956	26.0
20	21479	34.0
21	4647	2.0

이 Table에서
해당 pledge_id를 가지는 acc_id 갯수

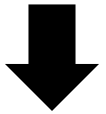
참고: [1]평균: 약 34명

5. Trade – 데이터 전처리

기존 Trade 데이터 셋은 거래별 데이터로 이루어져 있어, 각 acc_id 별로 집계하여 데이터를 정제하였다.

[예시]

day	time	type	server	source acc_id	source char_id	target acc_id	target char_id	item_type	item amount	item_price
7	21:13:05	1	ag	11439	385109	48152	34247	enchant_scroll	4.793968e-08	NaN



acc_id	개인상점 소스	교환창 소스	개인상점 타겟	교환창 타겟	소스거래 횟수	타겟거래 횟수	accessory	adena	armor	enchant scroll	etc	spell	weapon
11439	1	34	0	1	35	1	0	1	8	1	27	0	0

개인상점소스/교환창소스/개인상점타겟/교환창타겟: acc_id별 유형별 거래 총 횟수

소스거래횟수/타겟거래횟수: acc_id별 소스로서의 거래 총 횟수/타겟으로서의 거래 총 횟수

Accessory/.../weapon: acc_id별 거래한 아이템별 거래 총 횟수

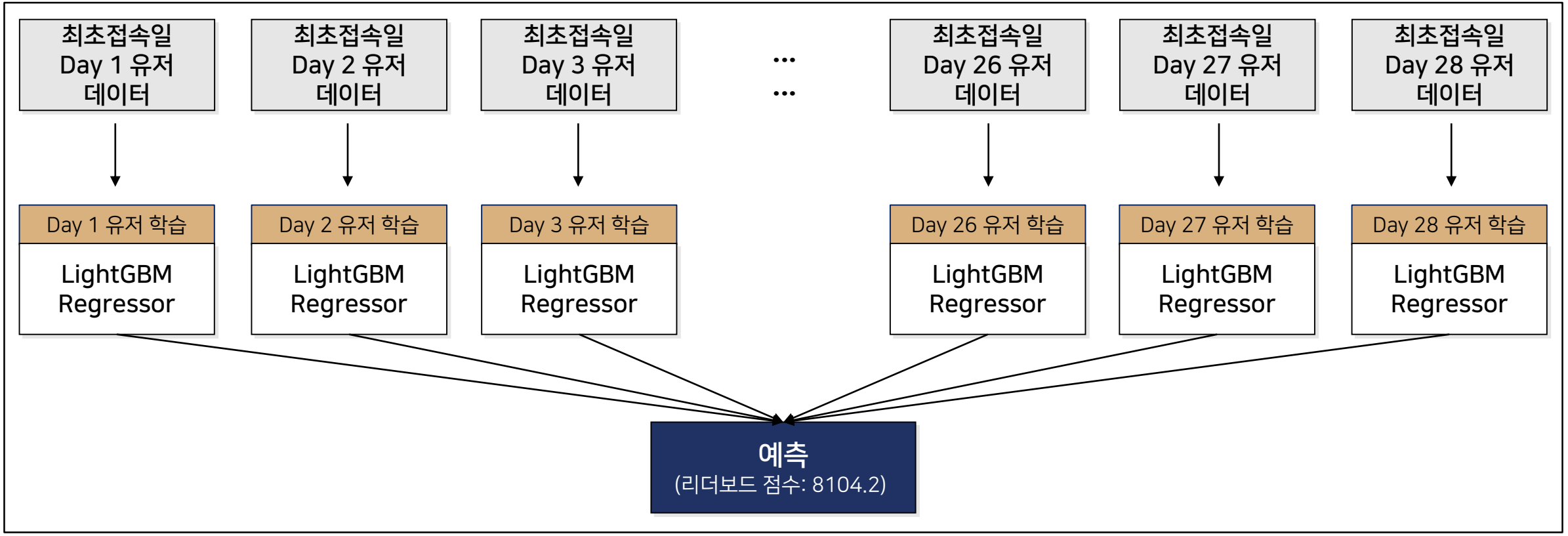
3. 1차 Modeling

1차 모델링과 발견한 문제점

3. 1차 Modeling

1차 모델: 최초접속일 일일 기준 유저 그룹핑 모델

1일, 2일, ..., 28일 등 최초접속일 기준으로 유저를 그룹지어 모델링을 시도하였으나, 낮은 성능을 보였다.



- 최초접속일이 유저들의 생존과 과금에 영향을 준다는 점을 기반으로 모델 고안
- 기존 데이터프레임을 각 유저의 최초접속일 별로 나눈 뒤, 각각 다른 모델(총 28개의 모델)로 학습을 진행

3. 1차 Modeling

1차 모델의 문제점 ①

1차 모델은 1)그룹별 데이터 수의 큰 편차, 2)survival time의 overpredict & amount spent의 underpredict 문제를 지닌다.

그룹별 유저 수

최초접속일 Day 1 유저 데이터	최초접속일 Day 2 유저 데이터	최초접속일 Day 3 유저 데이터	…	최초접속일 Day 26 유저 데이터	최초접속일 Day 27 유저 데이터	최초접속일 Day 28 유저 데이터
31262	1604	721	…	154	170	1045

- 최초접속일로 유저를 그룹화했을 때, 그룹별 데이터의 수의 편차가 크고
  일부 경우는 데이터의 수가 설명변수의 개수(약 470개)보다 적어 차원의 저주(curse of dimensionality)문제 발생
- 이 경우에 잡음(noise)이 발생되어 모델의 정확도가 감소할 수 있고 과적합(Over-fitting)이 일어날 가능성이 높음

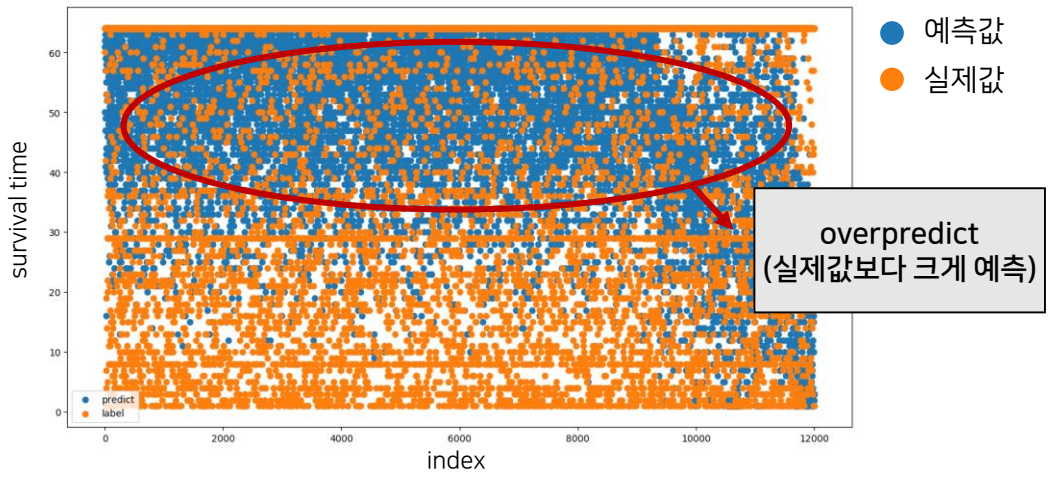
3. 1차 Modeling

1차 모델의 문제점 ②

1차 모델은 1)그룹별 데이터 수의 큰 편차, 2)survival time의 overpredict & amount spent의 underpredict 문제를 지닌다.

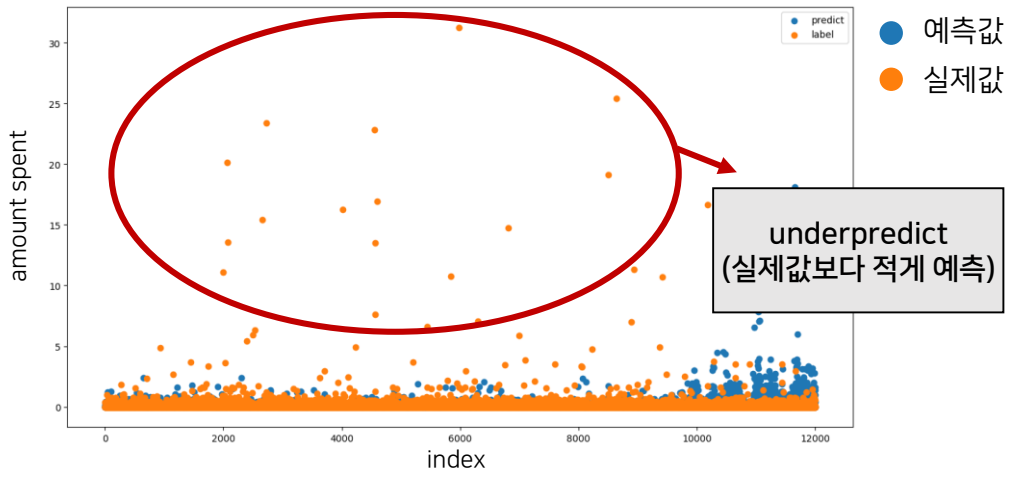
survival time 예측 결과

[survival time의 실제값과 예측값의 scatterplot]



amount spent 예측 결과

[amount spent의 실제값과 예측값의 scatterplot]



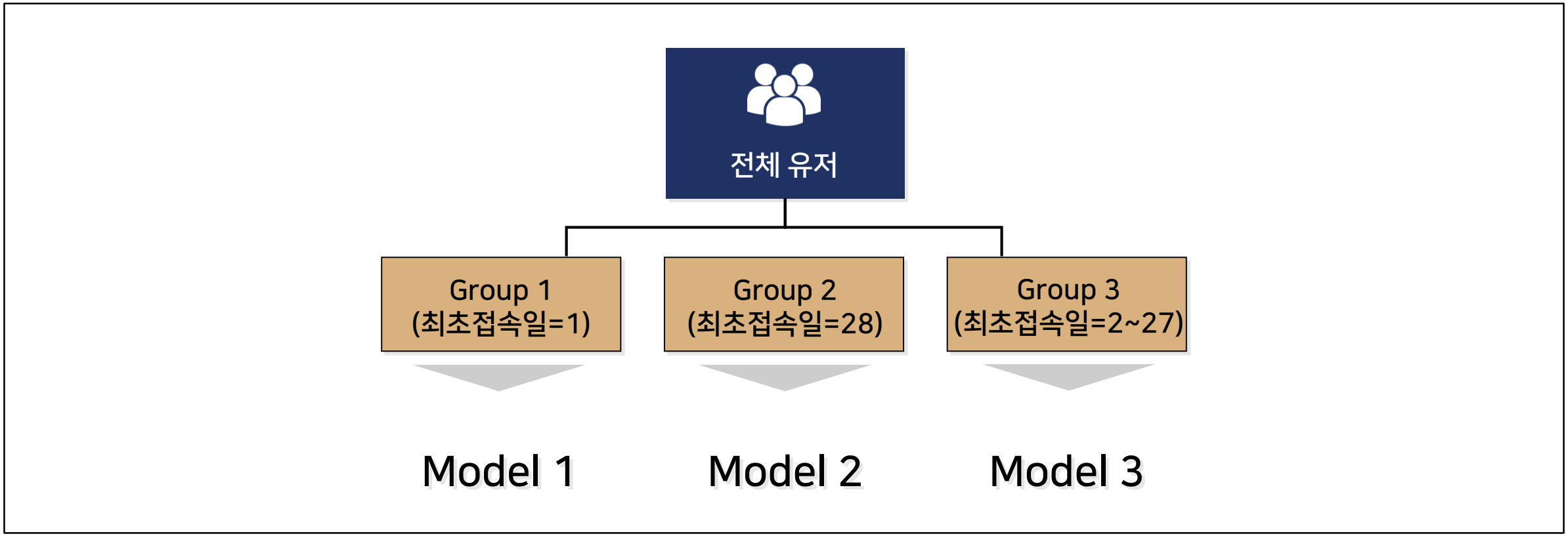
- survival time은 overpredict(실제값보다 크게 예측), amount spent는 underpredict(실제값보다 작게 예측)하는 문제점이 드러났다.
- 이는 '생존기간이 짧은 과금 유저'를 예측하지 못해, 잔존가치가 높은 유저에게 적절한 인센티브 정책을 시행하는데 걸림돌이 된다.

4. 2차 Modeling

1차 모델링의 문제를 개선한 2차 모델링

문제점 ①의 해결 방법: 세 그룹으로 유저 그룹핑

최초접속일을 기준으로 삼되 '최초접속일=1', '최초접속일=28', '최초접속일=2~27' 세 그룹으로 다시 그룹핑하여
그룹별 데이터 수의 큰 편차 문제를 해결하였다.
이를 통해 각 그룹 별로 survival time 예측 모델을 구축하였다.



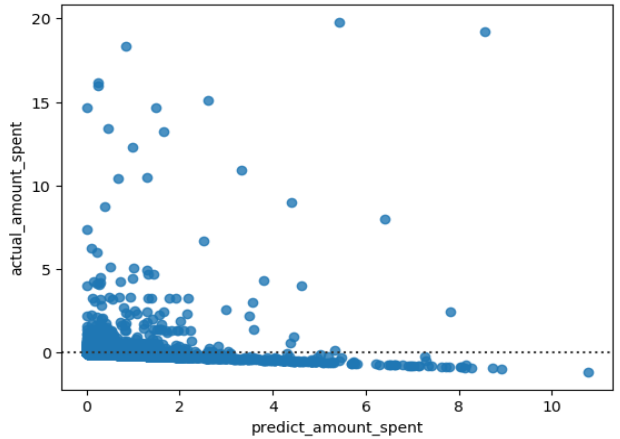
4. 2차 Modeling (2)1차 모델링 문제점 ②해결책: Quantile loss function 적용

문제점 ②의 해결 방법: 문제상황 분석

1차 모델이 survival spent의 overpredict, amount spent의 underpredict 경향을 띄는 이유는
잔차(y-yhat)가 일정하지 않은 분포(Non-constant variance)를 보여 이분산성 문제를 가지고 있고, 잔차가 정규성을 만족하지 않기 때문이다.
이 때 기본적인 MSE loss function[1] 만을 활용하면, 1차 모델처럼 '생존기간이 짧은 과금 유저'를 예측하지 못한다.

amount spent의 Non-constant variance

[amount spent residual plots]



- amount spent의 잔차의 경우 일정하지 않은 분포 (Non-constant variance)를 보임

잔차의 정규성 검정 결과

[Jarque Bera 검정 통계량]

survival time 잔차	amount spent 잔차
1129.04	26818100.46

- survival time과 amount spent 모두 Jarque Bera 검정통계량이 95% 임계치 5.99보다 크므로, 정규 분포를 따른다는 귀무가설을 기각한다.
- 이 경우 잔차가 비정규분포의 영향을 덜 받는 로버스트 추정량을 사용하면 더욱 효과적이다.

참고: [1]MSE(Mean Squared Error)

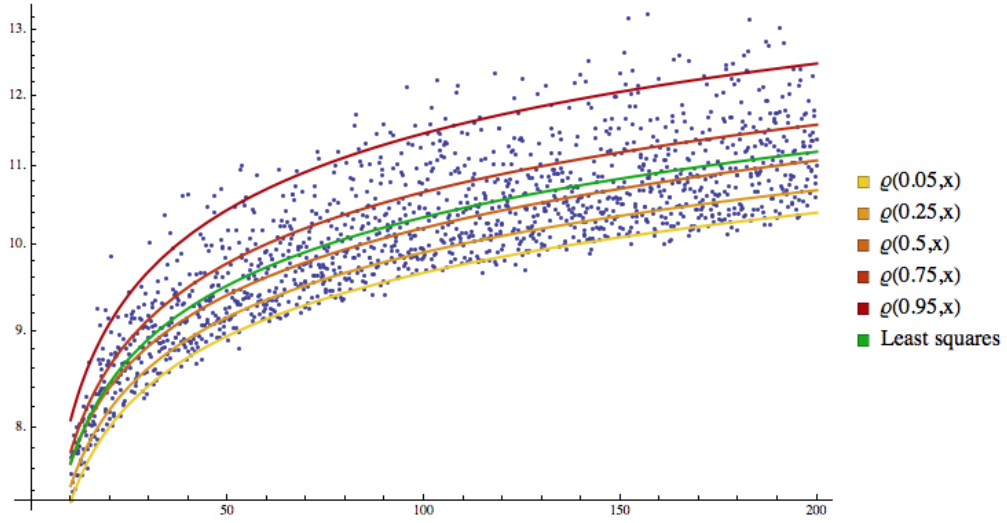
4. 2차 Modeling (2)1차 모델링 문제점 ②해결책: Quantile loss function 적용

문제점 ②의 해결 방법: Quantile loss function 적용

Quantile loss function은 잔차간 일정하지 않은 분포를 띄는 경우, 입력한 quantile에 따라 융통성있게 interval prediction을 가능케 한다.
또한, quantile loss function을 사용한 추정량은 이상점이나 잔차의 분포에 민감하게 반응하지 않는 로버스트 성질을 갖는다.
따라서, 이 loss function을 적용하여, 전략적으로 중요도가 높은 '생존기간이 짧은 과금 유저'를 예측할 수 있도록 한다.

Quantile loss function 특징 및 활용 방향

[지정된 Quantile 따라 달라지는 prediction과 MLS 기반 prediction 비교]



- Quantile(α)를 얼마로 정하지는 지에 따라,
  Overpredict 혹은 Underpredict에 주는
  페널티의 강도를 조절할 수 있다.
  (ex. α =0.25인 경우 Overpredict에 더 많은 페널티 부여)

- '생존기간이 짧은 과금 유저'를 보다 잘 예측하기 위해,
  모델에 Quantile loss function을 적용하고
  survival time 예측시 **낮은 quantile(α)**을,
  amount spent 예측시 **높은 quantile(α)**을 지정하였다.

4. 2차 Modeling (2)1차 모델링 문제점 ②해결책: Quantile loss function 적용

문제점 ②의 해결 방법: 문제 해결 중요성

문제점 ②를 해결하는 것은 아래와 같은 리스크를 피하기 위해 중요하다.

'생존 기간이 짧은 과금 유저' 예측 실패 리스크

survival time의 overpredict는 실제보다 예상 생존 기간을 길게 예측하여 적절한 대처 시기를 놓친다.

- 이미 유저가 이탈한 후에는 incentive 제공 불가능
- 이탈 이전 적절한 시점에 incentive 제공이 필요

'소 잃고 외양간 고친다'

amount spent의 underpredict는 과금 유저를 무과금/저과금 유저로 오판단하게 한다.

- 과금 유저에게 부족한 양의 incentive 제공
- incentive가 부족하여 유저가 이탈한 경우, incentive를 통한 기대이익 증진의 가능성이 없어짐

'부족하면 안하느니만 못하다'

참고: [1]MSE(Mean Squared Error)

5. 최종 예측 모델

5. 최종 예측 모델 (1) Key point ①: Feature Selection

Key point ①: Feature Selection

RFE, Boruta와 같은 자동화된 방법 뿐만 아니라 유튜브, 리니지 공식 홈페이지 및 커뮤니티 등에서 얻은 정보를 고려하여 다각적인 Feature Selection을 진행한 결과, 최종적으로 모델 예측 성능을 높일 수 있었다.

제거한 변수 목록[1]

random_attacker_cnt_pledge_sum

random_defender_cnt_pledge_sum

combat_play_time_pledge_sum

non_combat_play_time_pledge_sum

random_attacker_cnt_combat_sum

random_defender_cnt_combat_sum

- 대부분 전투 관련 변수가 관련이 없다고 파악되었다.
- 이 변수들을 제거했을 때 모델 성능이 향상되었다.

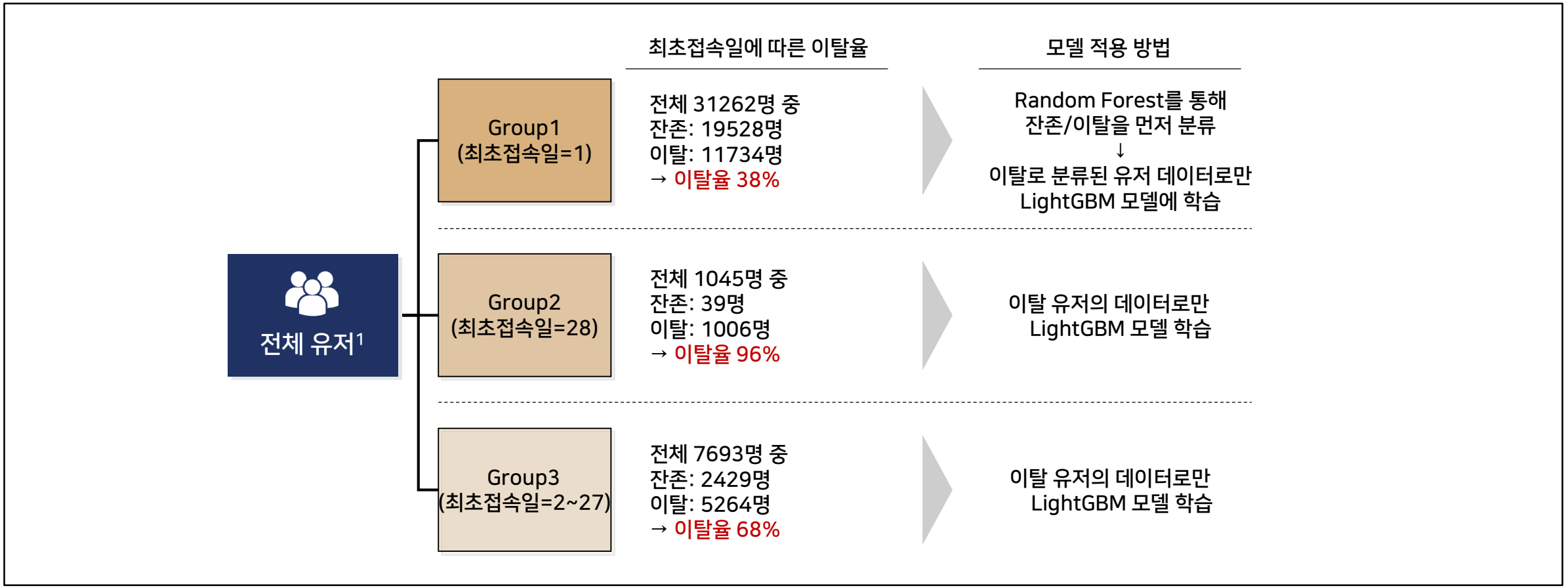
YouTube

참고: 재현성 파일에서 '파트명_preprocess_2.csv'는 RFE를 통해 선택된 변수만으로 이루어진 데이터프레임을 만든 결과이며, '파트명_preprocess_1.csv'은 '파트명_preprocess_2.csv'에서 외부에서 얻은 정보를 고려해 위 6개의 변수를 제거한 데이터프레임이다.

5. 최종 예측 모델 (2) Key point ②: 세 그룹으로 유저 그룹핑

Key point ② : 유저 그룹핑

EDA 결과, 과거 28일 중 최초접속일에 따라 유저의 이탈 비율이 크게 달라짐을 알 수 있었다.
따라서 survival time 예측 모델에서는 다음과 같이 유저를 3개의 그룹으로 나눈 뒤, 각 그룹별로 다른 모델을 적용시켜 학습하였다.



참고: [1]전체유저:Train 데이터 셋의 유저 40000명

5. 최종 예측 모델 (3) Key point ③: LightGBM & Ensemble

Key point ③ : Quantile loss function & Ensemble(모델 간 앙상블)

Loss function을 Quantile loss function으로 설정하고 Quantile(α) 값을 조정하였으며,
모델 간 앙상블을 통해 최종 모델을 구축하였다.

1. **LightGBM의 loss function을 기존의 'regression'에서 'quantile'로 변경**

$$L_\gamma(y, y^p) = \sum_{i=y_i<y_i^p} (\gamma - 1).|y_i - y_i^p| + \sum_{i=y_i \geq y_i^p} (\gamma).|y_i - y_i^p|$$

2. **Quantile(α) 값 조정을 통해 survival time의 overpredict, amount spent의 underpredict 방지**
   - Grid Search를 통해 각 종속 변수를 잘 예측하는 Quantile(α) 값 튜닝

3. **'잔존/이탈 분류한 후 회귀한 모델'과 '분류하지 않고 회귀한 모델'을 앙상블**
   - 최종 survival time 예측값: 두 모델 예측값의 max값[1]
   - 최종 amount spent 예측값: 4:6의 가중 평균값

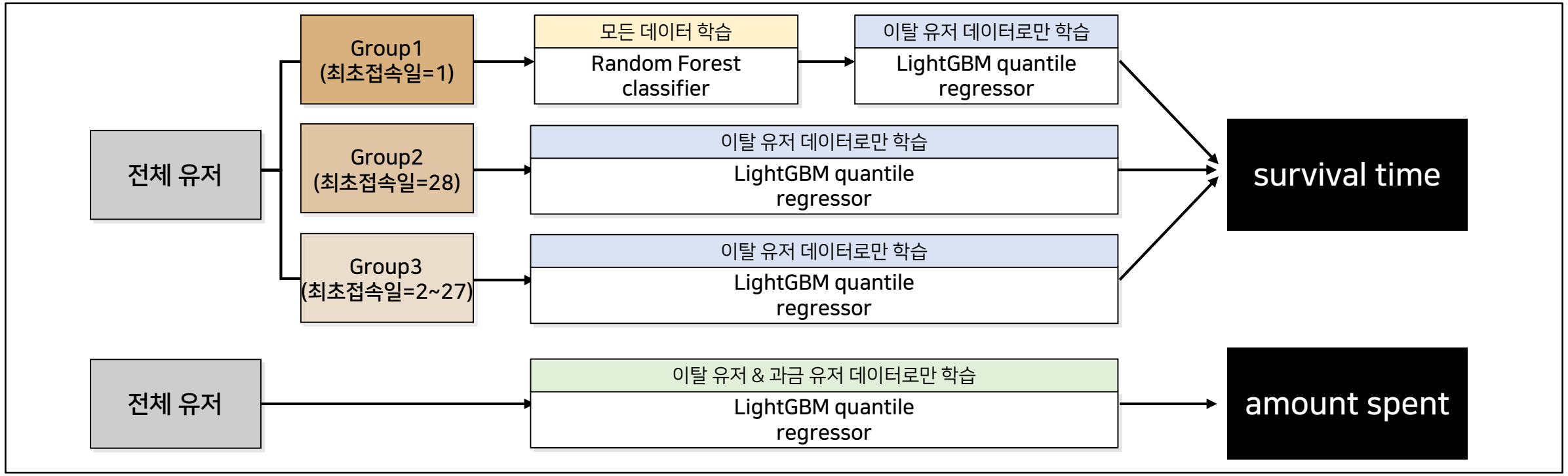
최종 모델 score: 15944점 (리더보드 기준)

참고: [1]Max 값으로 정한 이유: 슬라이드 #35 참고

5. 최종 예측 모델 (4) 모델 학습 프로세스

Ensemble: 첫번째 모델 A 학습 과정

survival time은 유저 3그룹으로 나누어 학습하고, amount spent는 모든 유저에 대해 학습하였다.



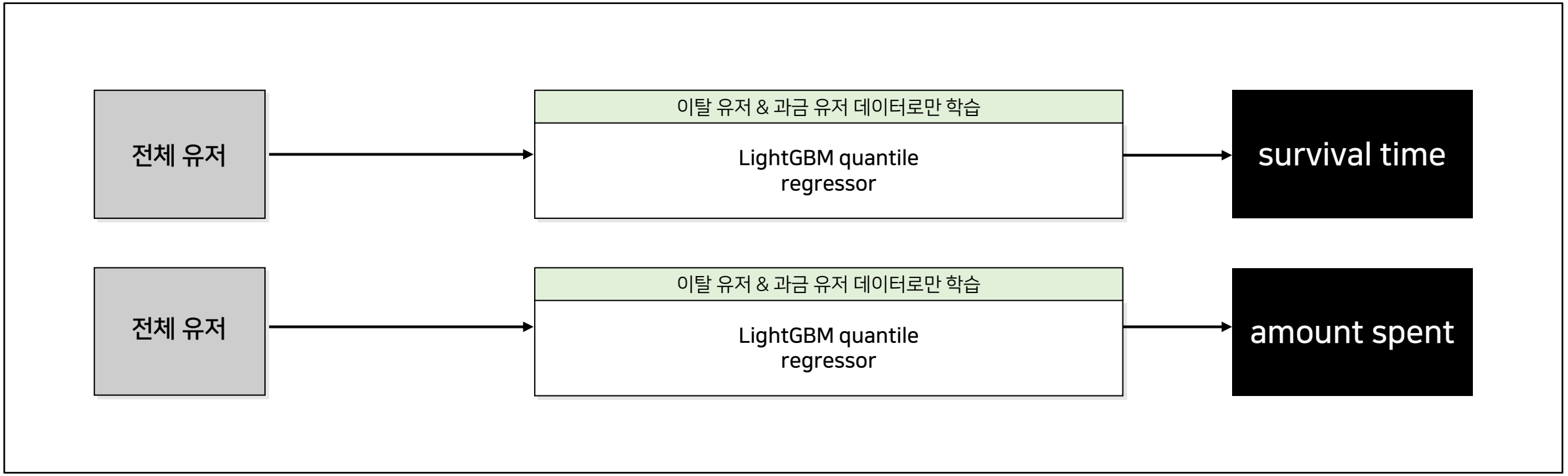
[이탈 유저 & 과금 유저 데이터로만 학습한 이유]

- 전체 데이터를 사용해 학습한 결과 이탈 유저/과금유저를 잘 선별하지 못하는 문제 발생
- 프로젝트의 목적이 이탈 유저와 과금 유저를 찾는 것이기 때문에 전체 모집단에서 이탈 유저/과금 유저 데이터로만 샘플링
- 샘플링 된 데이터로 학습한 결과 모델의 성능이 개선됨

5. 최종 예측 모델 (4) 모델 학습 프로세스

Ensemble: 두번째 모델 B 학습 과정

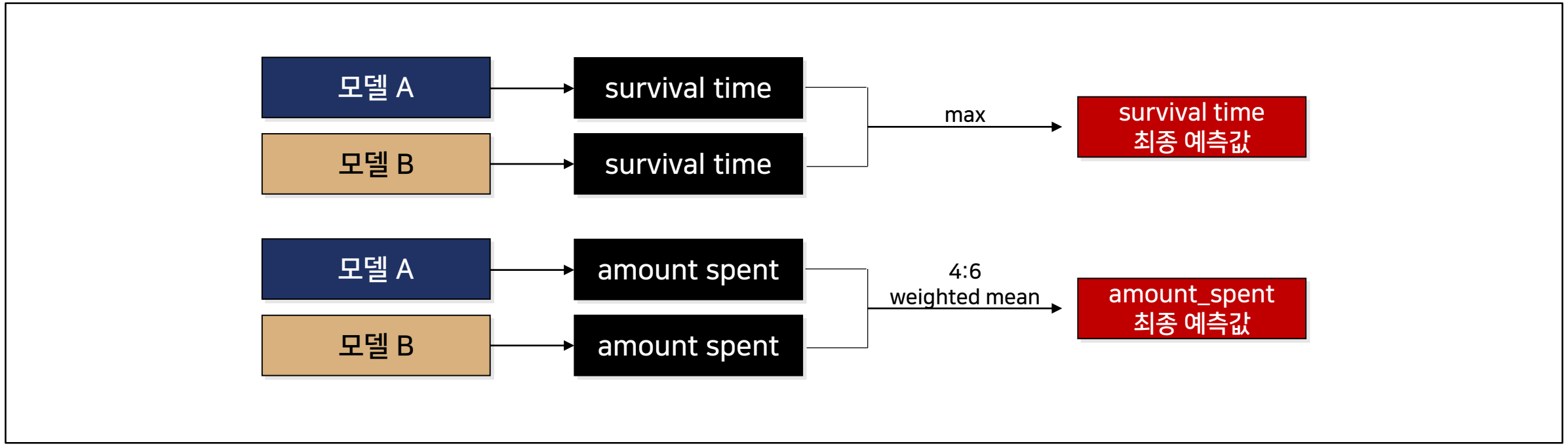
survival time, amount spent 모두 모든 유저에 대해 학습하였다.



5. 최종 예측 모델 (4) 모델 학습 프로세스

Ensemble: 모델 A & 모델 B

모델 A와 모델 B의 앙상블을 통해 최종 예측 모델을 구축하였다.



- survival time 예측에서 모델 B는 분류를 하지 않고 바로 회귀를 한 모델로 **잔존 유저의 survival time을 정확히 64로 예측하지 못한다.**
- 하지만 잔존/이탈을 먼저 분류하고 회귀를 한 모델 A에서는 1차적으로 잔존 유저로 분류가 되면 이들의 survival time은 64가 되게 모델링이 된다. 이를 통해 survival time의 예측값이 64인 케이스가 다수 존재하므로 좀더 현실적인 모델이 된다.
- 따라서 이 두 모델의 max값을 가져가는 형태로 앙상블을 하여 **survival time 예측값이 64인 케이스를 보존**한다.

6. 결과 분석

모델의 해석 및 결론

6. 결과 분석 (1) 모델 해석

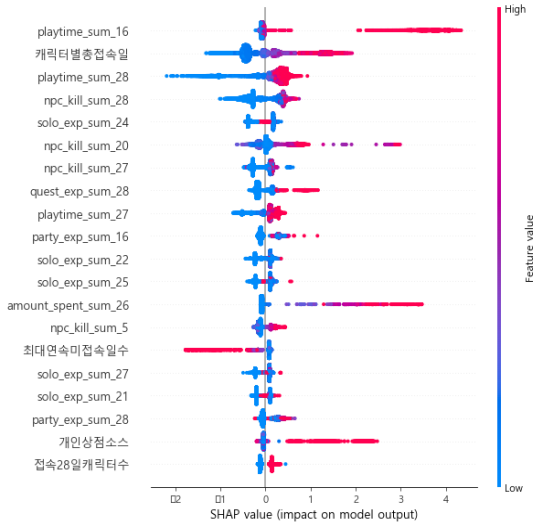
SHAP를 통한 모델 해석

모델은 예측력 뿐만 아니라 해석도 중요하기에, SHAP를 이용하여 모델에서 사용된 변수를 분석하였다.
최종모델을 Test1 데이터 셋과 Test2 데이터 셋에 각각 적용하여 변수들의 SHAP Value를 구하였다.
아래는 SHAP Value가 높은 변수들 순으로 정렬하여 시각화한 것이다.

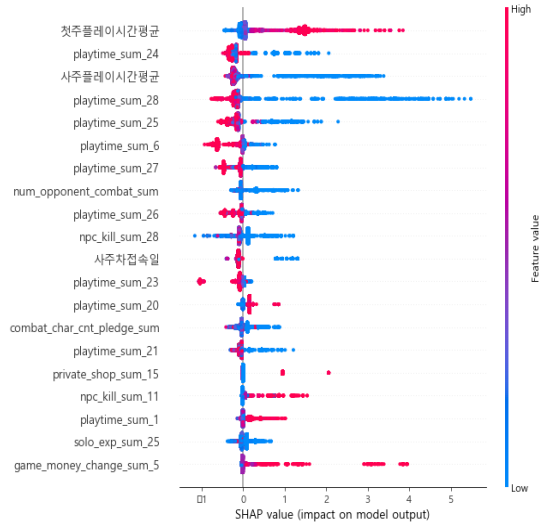
*SHAP Value: 각 Feature가 모델의 output에 영향을 미치는 정도를 의미

Test1

survival time

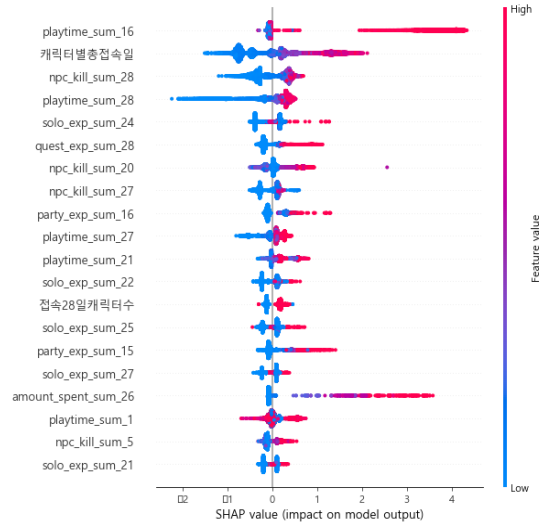


amount spent

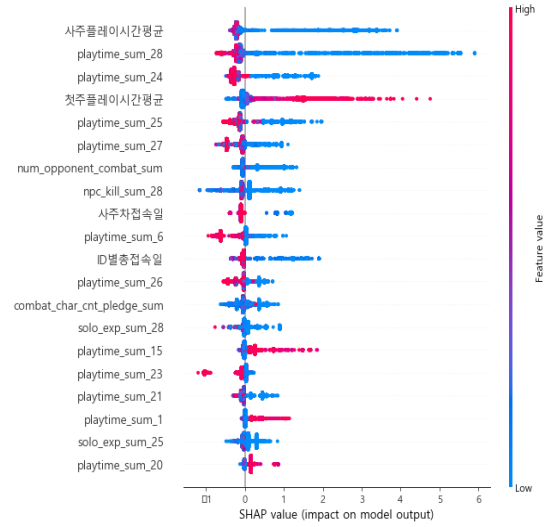


Test2

survival time



amount spent



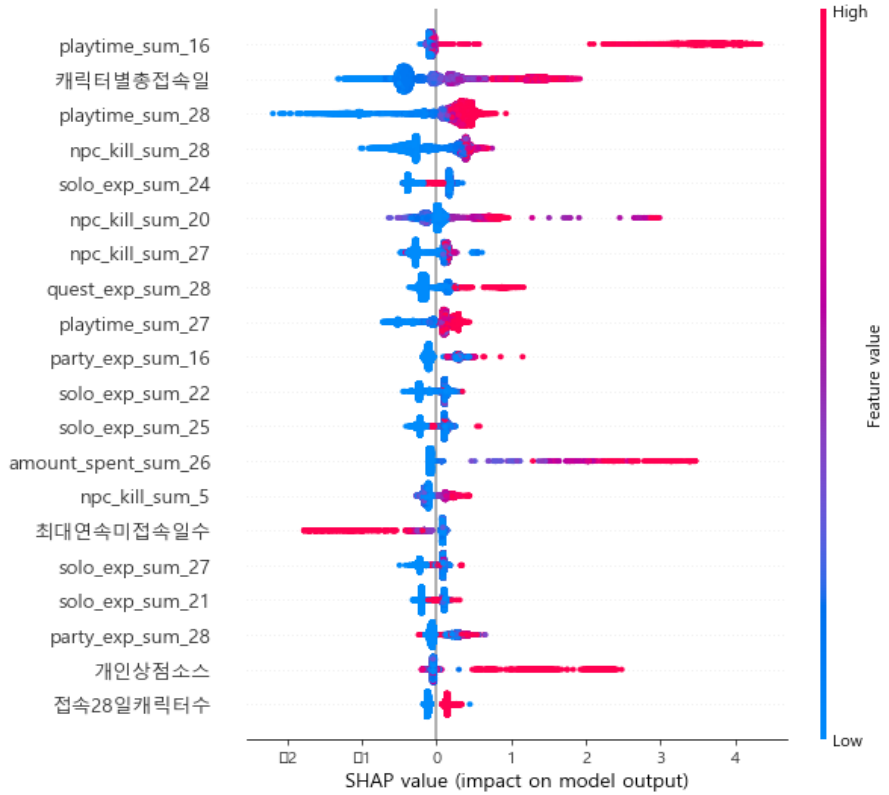
- Test1과 Test2의 SHAP Value를 비교해본 결과 중요한 변수는 유사하다.
  이를 통해 최종모델이 <u>시간 흐름에 강건한 모델</u>이라는 것을 알 수 있다.

6. 결과 분석 (1) 모델 해석

survival time 예측변수의 SHAP Value 분석

Group1(최초접속일=1) 기준, Test1 데이터 셋의 예측변수 SHAP Value는 아래와 같다.

[Group1 survival time 예측변수 SHAP Value]



1. **대부분의 변수가 최근 날짜(28)에 가까울 수록 SHAP value 증가**
   - 시간을 고려하여 최근 기간의 데이터를 집중적으로 분석해야 한다.

2. **'playtime'과 '과거 접속일 관련 변수', 'exp 관련 변수', 'npc_kill' 등 activity 관련 변수들이 전반적으로 중요**

3. **'개인상점소스'는 survival time과 뚜렷한 양의 상관 관계**
   - 개인상점을 통해 아이템을 여러 번 판매할 정도로 재력이 높은 유저는 잔존 확률이 높을 것이다.

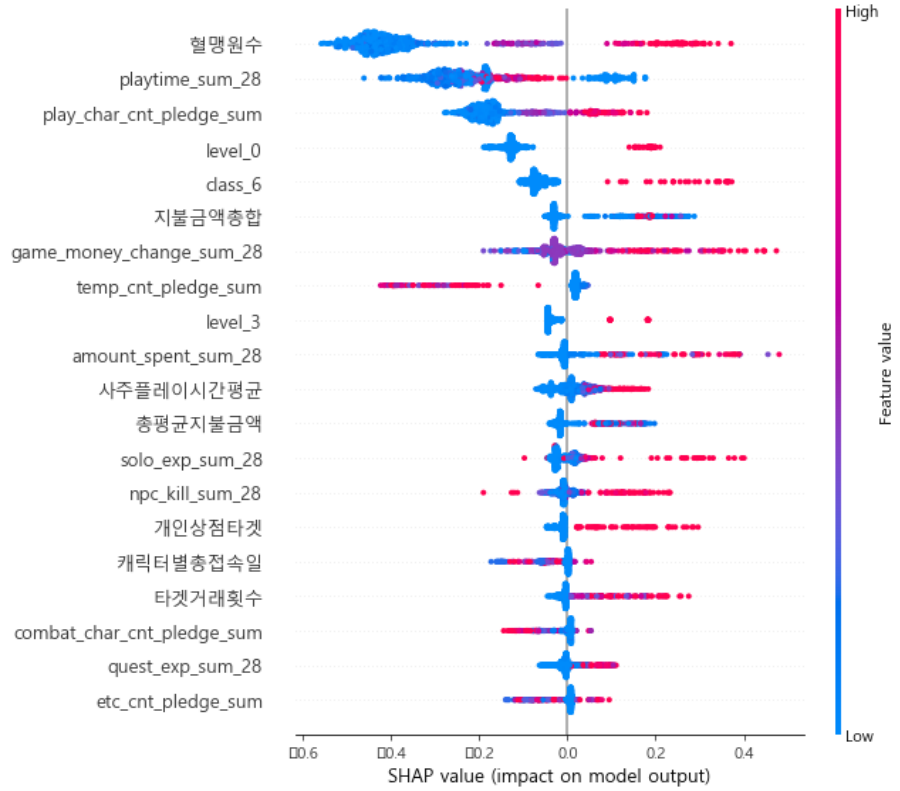
Group3(최초접속일=2~27)인 유저들은
Group1(최초접속일=1)과 비슷한 양상을 보였음

6. 결과 분석 (1) 모델 해석

survival time 예측변수의 SHAP Value 분석

Group2(최초접속일=28) 기준, Test1 데이터 셋의 예측변수 SHAP Value는 아래와 같다.

[Group2 survival time 예측변수 SHAP Value]



1. **Group2는 Group1과 현저히 다른 양상을 보임**
   - Group1은 개인의 활동과 관련된 변수의 중요도가 높지만, Group2는 사회적 활동(혈맹)과 관련된 변수의 중요도가 높다.

2. **'혈맹원수', 'play_char_cnt_pledge_sum' 등 사회성 관련 변수의 중요도가 높음**
   - 양의 상관 관계를 띄며 사회적 활동이 많을 수록 잔존 확률이 높다는 것을 볼 수 있다.

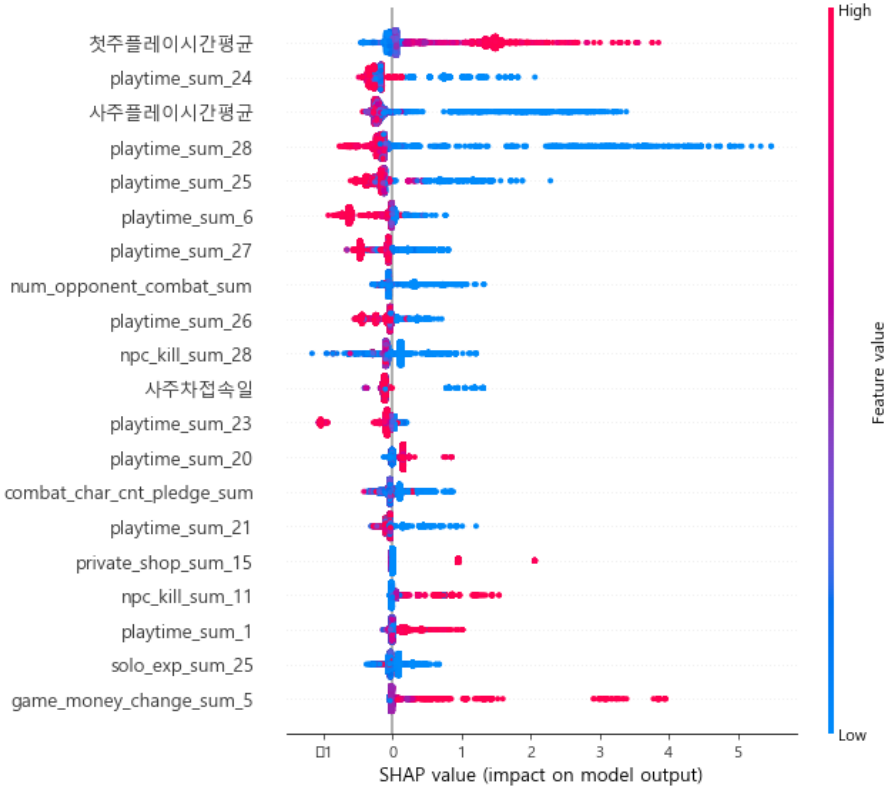
3. **경제활동이 활발할수록 잔존 확률이 높음**
   - '개인상점타겟' 변수는 survival time과 양의 상관 관계를 지닌다.
   - 'game_money_change' 변수의 중요도가 높음

6. 결과 분석 (1) 모델 해석

amount spent 예측변수의 SHAP Value 분석

모든 유저 기준, Test1 데이터 셋의 예측변수의 SHAP Value는 아래와 같다.

[amount spent 예측변수 SHAP Value]



1. **survival time과 마찬가지로 'playtime'과 '접속일', 'exp' 변수가 예측에 많은 영향을 끼침**
   - 위 세 변수는 잔존 가치 산출에 중요한 변수들이다.

2. **survival time과 달리 amount spent에서는 'num_opponent_combat_sum' 등 전투 관련 변수가 높은 영향을 끼침**
   - 그러나 인과 관계를 보여주지는 않는다.
     (과금을 이미 충분히하여 더 할 것이 없는 유저가 전투를 많이 하고 다닐 수 있음)

3. **최근 날짜의 playtime 변수가 survival time과는 양의 상관 관계, amount spent와는 음의 상관 관계를 가짐**
   - 최근에 많이 플레이한 사람은 잔존할 확률이 높지만 과금은 적게한다.
     따라서 잔존가치가 낮다.

6. 결과 분석 (2) 결론

모델 해석 요약

- **28일에 가까운 변수가 1일에 가까운 변수보다 변수 중요도가 높음**
  - 시간을 고려하여 최신 데이터에 비중을 높게 두어 분석하여야 한다.
- **28일에 처음 접속한 유저들은 그 전에 접속한 유저들과 현저히 다른 경향을 보임**
  - 그들을 분류하여 따로 모델을 만드는 것이 좋다.
- **전체적으로 playtime 변수가 제일 중요한 변수임**
  - playtime은 survival time과 amount spent 모두에서 중요한 변수로 선정된다.
- **과거의 접속 패턴, 경험치 획득 변수가 survival time, amount spent에 큰 영향을 끼침**
  - 이 변수들은 잔존 가치 산출에 중요한 역할을 할 것이다.
- **혈맹과 같은 사회적, 거래와 같은 경제적 활동이 survival time에 높은 상관관계를 보임**
  - 앞서 공홈 커뮤니티로 추론한 원인이 모델 해석에서 드러난다.

결론

이탈 징후 유저들을 판별하고
적절한 인센티브를 제공하기 위해서

1)특정 기간 내 유저의 첫 접속일 이후
게임 투자 시간 변동에 주목하여야 하며,

2)신규/복귀 유저들을 위한
지속적인 사회적, 경제적 컨텐츠의 관리가 필요함


THE CROSS RANCOR
Lineage
E.O.D.