

Analysis 분야 챔피언리그

# 리니지 고객 활동 데이터를 활용하여 잔존 가치를 고려한 이탈 예측

---

팀 초코송이

강민수, 권홍욱, 안재일, 이태우, 이한송

# Index

- I. EDA & Data Preprocessing
- II. Amount Spent Modeling
- III. Survival Time Modeling
- IV. Conclusion



I. EDA & Data Preprocessing

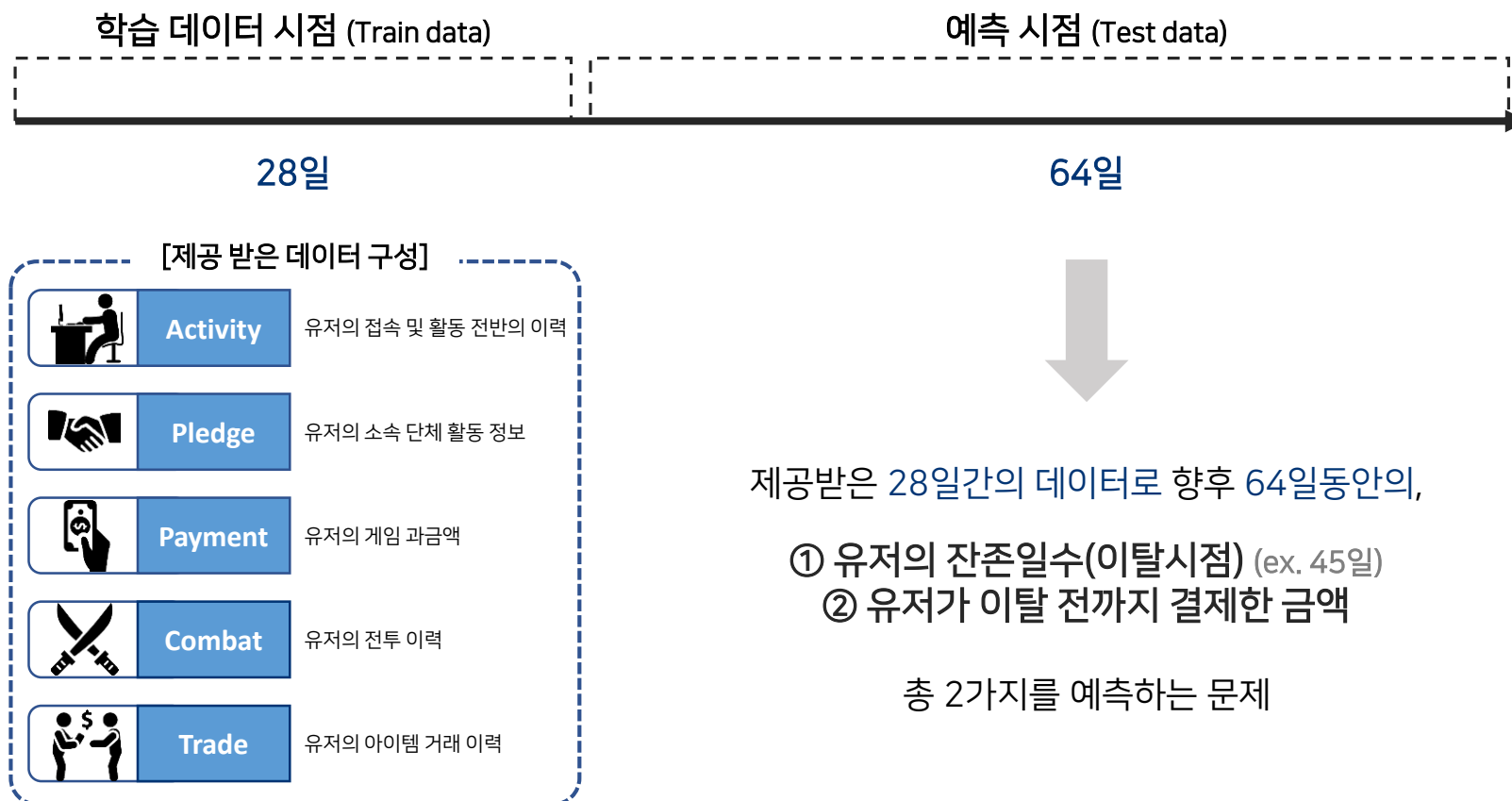
II. Amount Spent Modeling

III. Survival Time Modeling

IV. Conclusion

## ① 문제 정의

분석 과제 : 온라인 게임 유저 활동 데이터를 활용하여 기대이익을 고려한 **결제 및 이탈 예측** 모형 개발



## ① 문제 정의

분석 과제 : 온라인 게임 유저 활동 데이터를 활용하여 **기대이익을 고려한 결제 및 이탈 예측** 모형 개발

$$\text{기대이익} = 30 * \left\{ \left( \frac{1}{e} \right)^{\frac{\text{Survival time error}^2}{2 * 15^2}} * \text{amount spent label} * \text{gamma} - (\text{amount spent predict} * 0.01) \right\}$$

\* gamma,

if) survival time predict = 64, gamma = 0

if) amount spent label = 0, gamma = 0

else)

$$\text{gamma} = 0 \quad \frac{\text{amountspent predict} - \text{amountspent label} * 0.1}{\text{amountspent label} * 0.9} \quad 1 \rightarrow \text{amountspent predict}$$

if) survival time predict = 64, 예측비용 = 0

	비결제	결제
이탈	기대이익=0	
잔존	기대이익=0	기대이익=0

‘결제-이탈’ 유저에게서만 기대이익이 발생하는 것을 확인

즉, 회사의 입장에서는 **돈을 지불하지만, 이탈하는 유저들을 찾아 잔존기간을 늘려 오랫동안 돈을 지불하도록** 만드는 것이 목표

## ① Word Cloud

- 리니지 이탈 원인을 파악하기 위해, 리니지 실제 유저 반응을 **텍스트 분석** 진행
- 리니지 커뮤니티 게시판에서 '그만두다', '접는다'라는 검색어로, 총 1600개의 Data 수집

[텍스트 마이닝 결과를 워드 클라우드로 시각화]



[실제 유저가 작성한 게시물]

"1시간에 경험치 5%씩 오를까 말까 그냥 접는 방법밖에 없나요?"

"서버 상황이 여의치 않으면 게임을 접는 사람이 늘고 있습니다. ... 남은 사람들은 위해서 캐릭터 밸런스라던지 서버사람이 없는 곳은 합쳐 버리는게 ..."

"리니지라는 게임은 20년 된 게임으로서 레벨 격차가 너무 심해서 그 격차를 따라가보려고 한다면 일반 직장인으로서 감당 할 수 없는 시간과 캐쉬 과금을 해야만 합니다. 결국 그 장벽에 부딪혀 복귀했다가 죽어라 열렘만 하다 다시 접는 반복이 이루어 집니다."

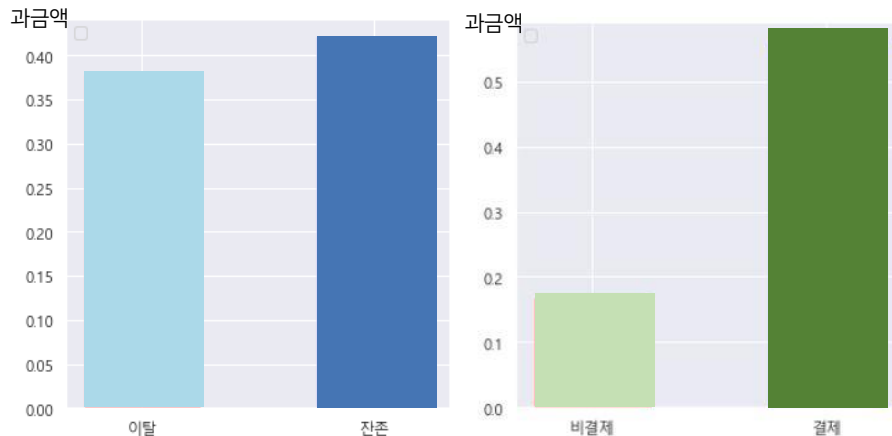
출처: 리니지 공식 커뮤니티 자유 게시판  
(<https://lineage.plaync.com/board/free/list>)

- '접다(그만두다)'에 관련된 통한 워드 클라우드 확인 결과, '시간', '과금', '레벨', '서버' 등이 높은 빈도를 보임
- 이를 통해 게임을 접는(그만두는) 행위와 관련된 요소로 해당 변수들이 관련성이 있다고 판단 후 파생변수 생성
- 특히, 일반적인 요소가 아닌 '**서버**' 또한 높은 빈도를 나타내는 것으로 보아, 서버 역시 잔존 여부와 관련이 있다고 유추

## ① Label과의 관계

- 64일간의 잔존일수와 결제금액이 있는 정답 테이블을 바탕으로, 유저의 잔존여부와 결제여부에 따른 변수별 차이를 비교
- Label에 따라 이탈과 잔존, 결제와 비결제 그룹을 구분하고 해당 그룹 간의 28일간의 '과금액' 변수의 평균값 비교

['과금액' 변수에 대한 잔존여부, 결제여부 그룹별 평균값 비교]



- ✓ 동일 변수이지만, 잔존여부에 따른 차이는 크지 않으나, 향후 미래의 결제여부에 있어서는 큰 차이를 보임

## ② Feature Importance

- 제공받은 Raw data 전체를 별도의 전처리 과정 없이 하나의 테이블로 통합하여 random forest로 변수중요도를 확인

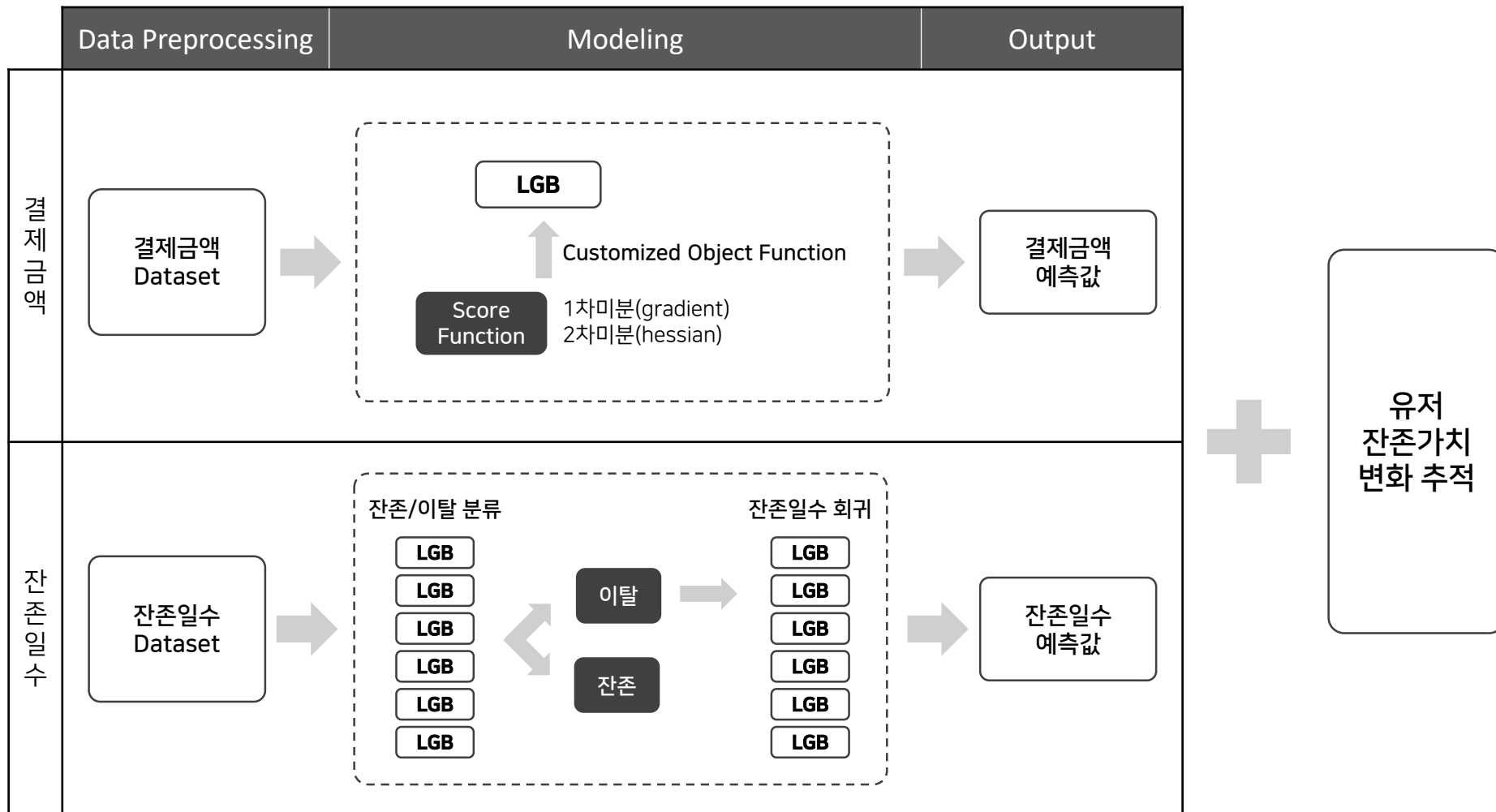
[Random Forest Feature Importance]

중요순위	결제금액		잔존일수	
	변수	중요도	변수	중요도
1	questexp	0.0084	playtime	0.0044
2	playtime	0.0051	cls0	0.0043
3	soloexp	0.0034	soloexp	0.0029
4	fishing	0.0033	maxlevel	0.0027
5	maxlevel	0.0027	npckill	0.0026
6	npckill	0.0020	moneychng	0.0024
7	moneychng	0.0016	privateshop	0.0023
8	minlevel	0.0010	questexp	0.0018
9	privateshop	0.0007	charcnt	0.0013
10	partyexp	0.0006	fishing	0.0012

- ✓ 결과적으로 잔존일수와 결제금액 예측에서 동일한 변수가 서로 다른 중요도를 가짐을 확인

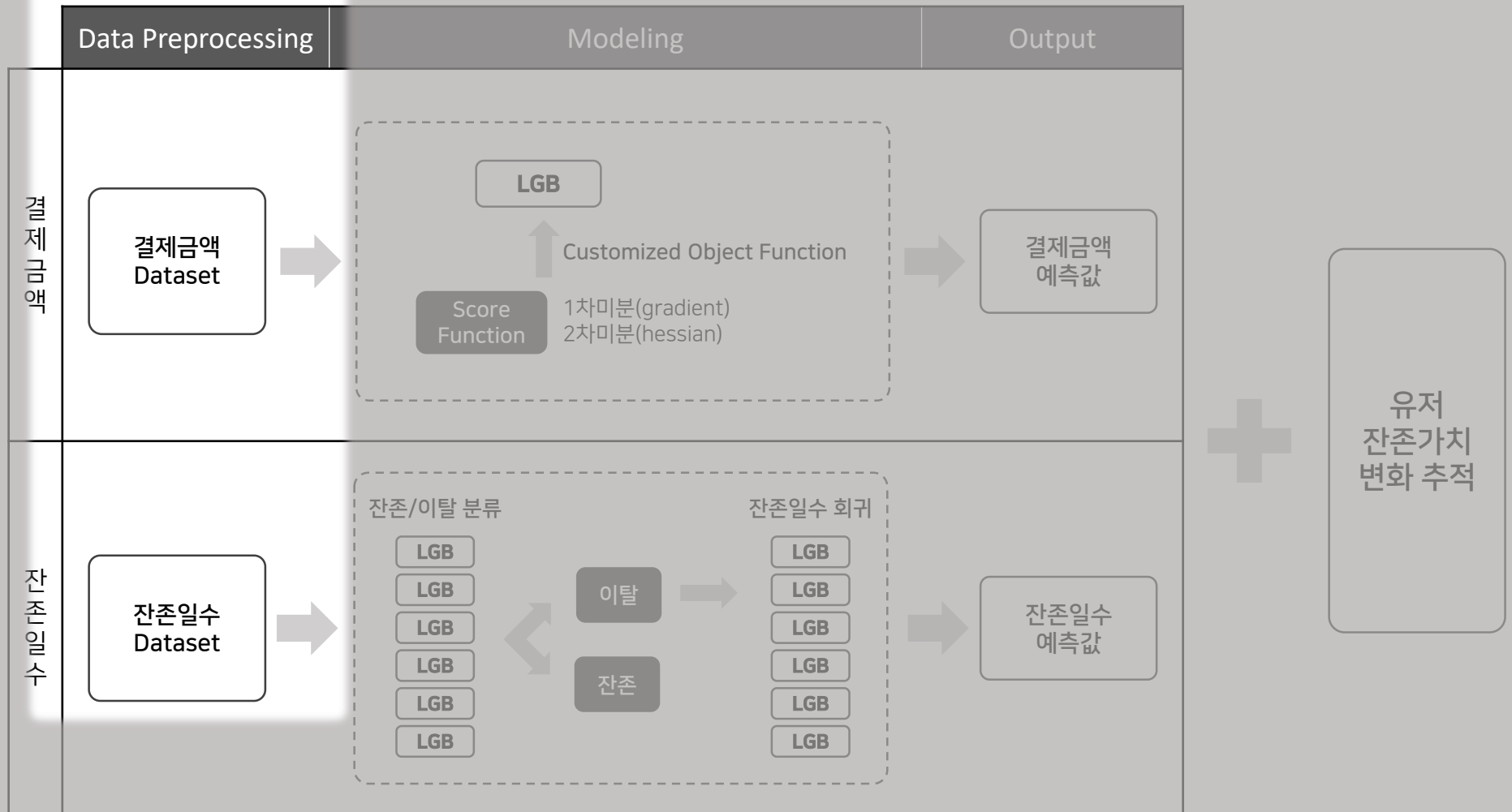
- ✓ 잔존일수와 결제금액의 특성이 다르기 때문에, 별도의 데이터 셋과 모델을 구성하여 분석 진행

## ① Pipe Line





## ① Pipe Line

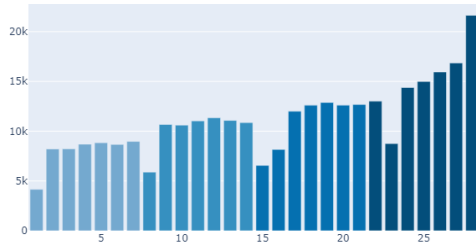


## ① 주차별 변수 생성

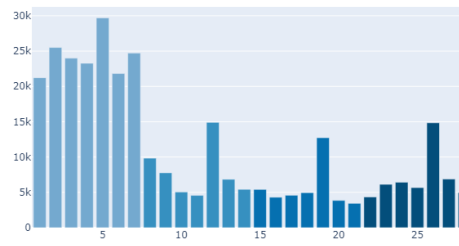
Daily

[Raw data의 fishing, same\_pledge\_cnt, playtime 변수의 28일 간의 변화 추세]

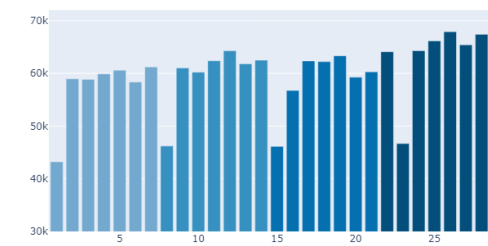
[Fishing]



[Same\_pledge\_Cnt]



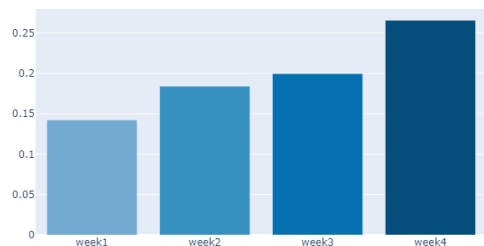
[Playtime]



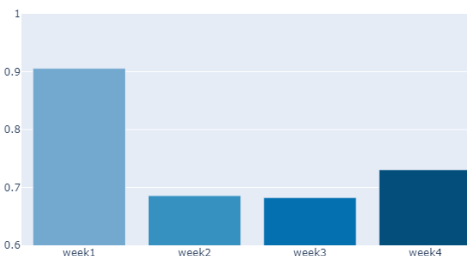
Weekly

[일주일 단위로 세 변수의 수치를 통합한 주차별 변수의 변화 추세]

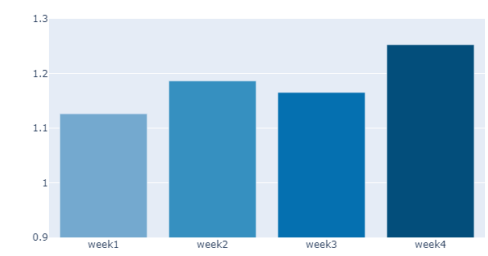
[Fishing]



[Same\_pledge\_Cnt]



[Playtime]



- raw data에서는 서버점검일과 주말로 추정되는 특정일에 수치가 급변하는 패턴이 7일주기로 반복되기 때문에, 그대로 학습할 경우 모델이 시간 순서를 혼동할 가능성 존재
- 이에 각 유저의 일자별 데이터를 일주일 단위로 통합하여 주차별 변수를 생성

## ② User Clustering

[Train data의 feature를 바탕으로 40,000명의 전체 유저에 대해 **군집화** 진행]

	Features										Labels				
	Death	Revive	Fishing	Solo exp	Party exp	Pledge cnt	Playtime	Source cnt	Source price	Private shop	잔존 기간	잔존 확률	일평균 결제금액	누적 결제금액	결제 확률
군집1	0.26	0.24	0.15	1.38	1.23	0.14	1.10	0.92	0.01	1.16	35.88	0.29	0.10	3.11	0.58
군집2	0.00	0.00	0.00	0.01	0.00	0.00	3.10	5.77	0.91	4.43	59.38	0.88	0.02	0.62	0.16
군집3	0.03	0.03	0.07	0.22	0.06	0.01	0.28	0.34	0.02	0.32	28.39	0.25	0.24	2.40	0.51
군집4	0.09	0.07	0.40	0.27	0.09	0.05	1.58	0.58	0.02	0.49	49.25	0.60	0.10	4.45	0.68
군집5	0.48	0.47	0.65	0.10	0.03	2.91	2.00	1.42	0.01	1.83	59.45	0.86	0.22	12.69	0.97

- 군집화 결과, 각 군집들은 **Feature 간 명확한 특성 차이**를 보임
  - ✓ 예를 들어, 2번 군집의 경우 게임 플레이와 관련된 변수들은 모두 최저값에 해당하지만, playtime과 거래 관련 변수에 있어서는 최대값에 해당. 즉, 2번 군집의 유저들은 게임 플레이는 하지 않고 상행위를 주로 하는 '전문상인' 계정으로 추정 가능
- 또한, 군집화 결과를 바탕으로 각 군집 별 **label 값 또한 명확한 차이**를 보임
  - ✓ 잔존기간(survival\_time)과 일평균결제금액(amount\_spent), 잔존여부와 결제여부를 바탕으로 계산한 잔존확률과 결제확률, 일평균 결제금액에 잔존일수를 곱한 누적결제금액 모두 군집별로 큰 차이를 보임

[User Cluster Feature Engineering]

Acc_id	Features	cls
2	...	3
5	...	4
8	...	3
17	...	1
⋮	⋮	⋮



Acc_id	Features	cls_1	cls_2	cls_3	cls_4	cls_5
2	...	0	0	1	0	0
5	...	0	0	0	1	0
8	...	0	0	1	0	0
17	...	1	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

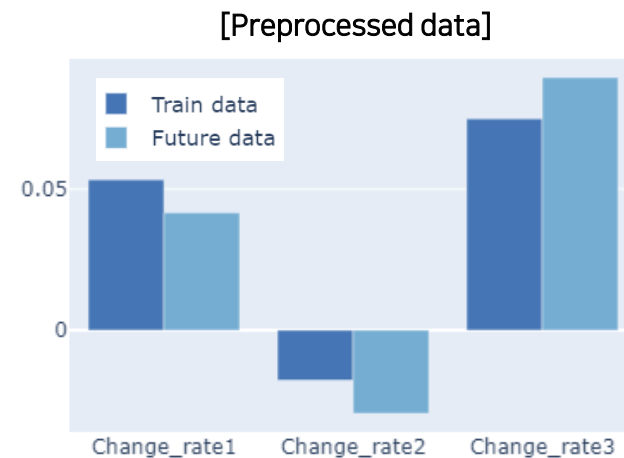
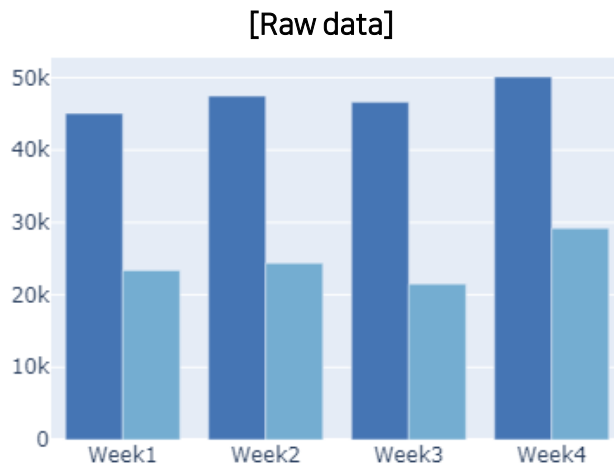
## ③ Concept drift

“데이터와 패턴, 규칙은 시간이 지남에 따라 변화하기 때문에 당시에 만들어진 모델은 시간이 지나면서 쓸모 없게 된다. 기계학습과 데이터 마이닝에서 이 현상을 ‘**Concept drift**’라고 한다.”

Žliobaitė I., Pechenizkiy M., Gama J. (2016) An Overview of Concept Drift Applications



- Concept drift에 의해서 현재 분석하고 있는 User들의 특성이 미래의 시점에서 바뀐다면, 현재의 데이터로 학습한 모델로 특성이 바뀐 미래의 데이터를 통해 예측 정확도가 떨어질 수 있음.
- 이에 학습데이터의 28일간의 변화율과 비율변수 등 상대적인 값들을 활용한 **비율변수를 추가**함으로써 학습데이터에 Overfitting 되는 문제를 보완



## ④ 파생변수 생성

- 주차별 playtime 기울기의 변화량에 대한 파생변수 생성

변수명	의미	식(방법)
playtime_change_w1_w2	week1, week2 사이의 Playtime 기울기	$\frac{\text{playtime\_week2} - \text{playtime\_week1}}{\text{Playtime\_week1}}$
playtime_change_w2_w3	week2, week3 사이의 Playtime 기울기	$\frac{\text{playtime\_week3} - \text{playtime\_week2}}{\text{Playtime\_week2}}$
...	...	...
playtime_change_w1_w4	week1, week4 사이의 Playtime 기울기	$\frac{\text{playtime\_week4} - \text{playtime\_week1}}{\text{Playtime\_week1}}$

- 유저 행동 패턴 별로 Clustering한 파생변수 생성

변수명	의미	식(방법)
cls_#	해당 유저가 분류된 군집 번호	activity, combat, pledge, trade 테이블을 바탕으로 K_means 군집화 후 더미 변수 생성

## ④ 파생변수 생성

- playtime 대비 획득한 경험치 파생변수 생성

변수명	의미	식(방법)
solo_exp_per_pt	투자한 시간 대비 solo_exp 취득량	$\frac{\text{solo\_exp} + 0.01}{\text{Playtime} + 0.01}$
party_exp_per_pt	투자한 시간 대비 party_exp 취득량	$\frac{\text{party\_exp} + 0.01}{\text{Playtime} + 0.01}$
quest_exp_per_pt	투자한 시간 대비 quest_exp 취득량	$\frac{\text{quest\_exp} + 0.01}{\text{Playtime} + 0.01}$

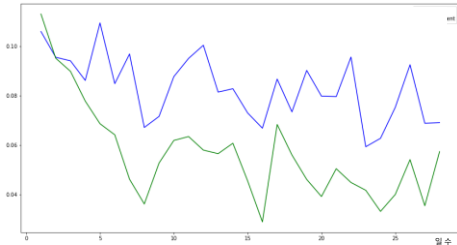
- 투자한 금액 대비 획득한 경험치 파생변수 생성

변수명	의미	식(방법)
solo_exp_per_as	투자한 금액 대비 solo_exp 취득량	$\frac{\text{solo\_exp} + 0.01}{\text{pay\_mean} + 0.01}$
party_exp_per_pt	투자한 금액 대비 party_exp 취득량	$\frac{\text{party\_exp} + 0.01}{\text{pay\_mean} + 0.01}$
quest_exp_per_pt	투자한 금액 대비 quest_exp 취득량	$\frac{\text{quest\_exp} + 0.01}{\text{pay\_mean} + 0.01}$

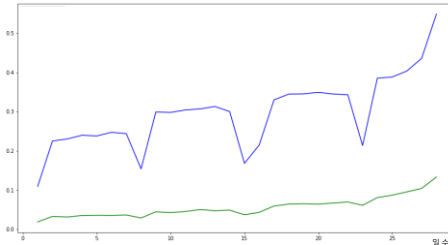
### ① Activity Table 변수들을 활용한 파생변수 생성

- 결제, 비결제 유저간 변수 특성 차이 비교 분석

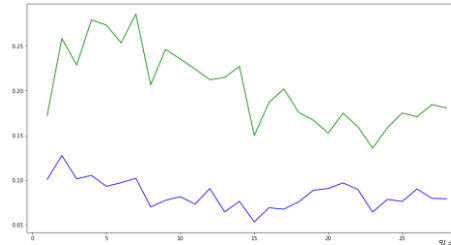
[유저 특성에 따른 quest exp 추세]



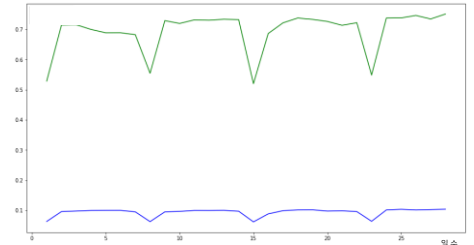
[유저 특성에 따른 fishing 추세]



[유저 특성에 따른 party exp 추세]



[유저 특성에 따른 private shop 추세]

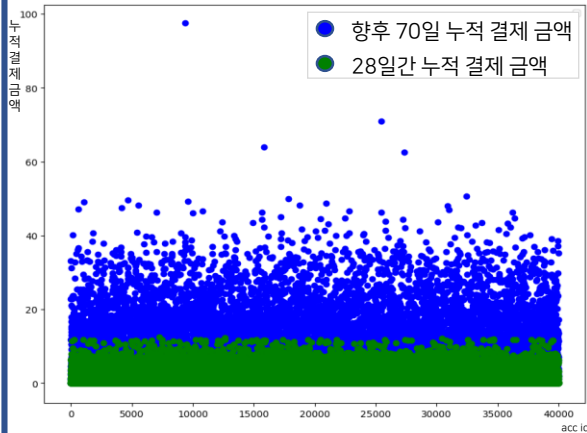


- 과금을 한 유저들이 quest exp 값과 fishing 시간이 더 많고, 과금을 하지 않은 유저들이 party exp 값과 private shop 시간이 더 많음
- 비교 분석한 변수들에 대한 파생변수 생성

변수명	의미	식(방법)
party_exp_mean	유저 평균 파티 경험치 획득 수치	$\frac{1}{\text{Number of days}} \times \sum_{i=1}^{28} \text{party\_exp}_i$
fishing_mean	유저 평균 낚시 시간	$\frac{1}{\text{Number of days}} \times \sum_{i=1}^{28} \text{fishing}_i$
quest_exp_mean	유저 평균 퀘스트 경험치 획득 수치	$\frac{1}{\text{Number of days}} \times \sum_{i=1}^{28} \text{quest\_exp}_i$
private_shop_mean	유저 개인 상점 운영 시간	$\frac{1}{\text{Number of days}} \times \sum_{i=1}^{28} \text{private\_shop}_i$

### ② Payment 파생변수 생성

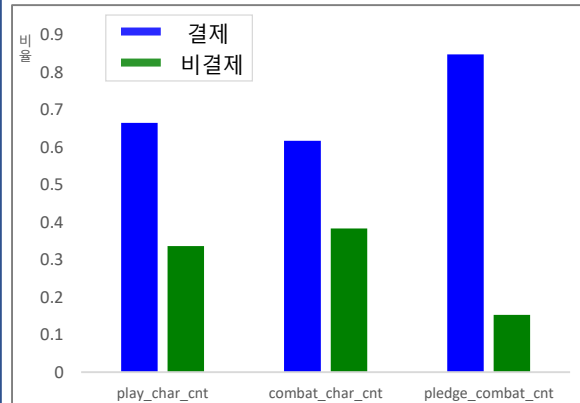
[유저 별 28일, 향후 70일 누적결제금액 관계]



- 누적 결제 금액의 경우, 대체적으로 유저 별 **28일간과 향후 70일간 누적결제금액이 비례 관계**임을 알 수 있음.

### ③ Pledge 파생변수 생성

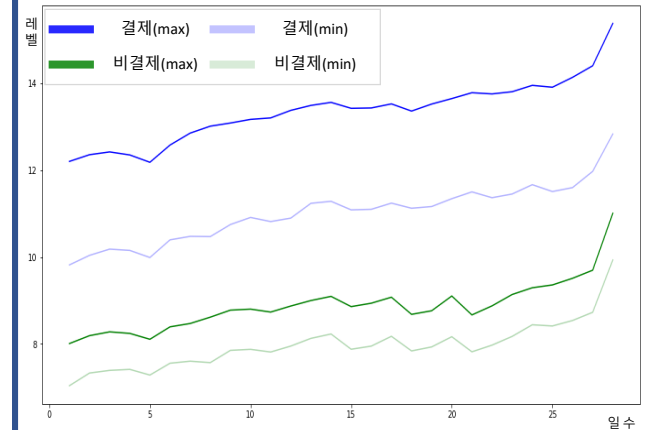
[유저 특성에 따른 혈맹의 성향 차이]



- play\_char\_cnt, combat\_char\_cnt, pledge\_combat\_cnt의 비율이 타 혈맹보다 높은 혈맹에 과금을 한 유저들이 많이 속해 있다는 것을 확인 가능.
- ✓ **전투를 보다 많이 하고, 공격적인 성향**을 갖고 있는 혈맹에 과금을 한 유저가 많다는 점을 확인 가능.

### ④ Combat 파생변수 생성

[유저 특성에 따른 각 일자별 Min / Max Level 추세]



- 유저들이 일자별 접속한 캐릭터들의 레벨 최솟값, 최댓값 확인 결과, 모두 증가하는 추세를 발견
- 과금을 한 유저들의 캐릭터 레벨의 최솟값, 최댓값 모두 과금을 하지 않은 유저보다 높음
- ✓ 과금을 한 유저들의 **캐릭터 최소 레벨, 최대 레벨이 더 높**다는 것을 알 수 있음.



### ⑤ Server Clustering

- 리니지는 일반적인 'PvP서버' 뿐만 아니라, '전투 특화 서버' 및 'Non-PvP 서버' 등 다양한 서버를 제공함
- 각 서버는 서로 다른 다양한 특징을 보이고 유저의 성향을 반영함

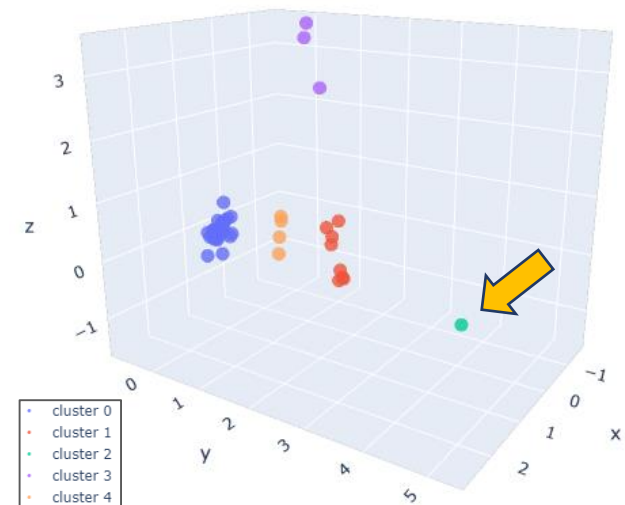
#### Factor Analysis

	개인플레이(x)	협동플레이(y)	거래활동(z)
playtime	80	-51	16
npc_kill	84	-47	-21
solo_exp	78	62	5
rich_monster	96	-18	-5
exp_recovery	86	-31	7
fishing	80	-56	1
private_shop	76	-62	-9
party_exp	46	82	7
combat_char_cnt	16	72	-61
pledge_combat_cnt	-57	33	-56
pledge_cnt	-64	30	-56
item_amount	28	-35	60
item_price	7	-51	41
random_attacker_cnt	0	7	-54
temp_cnt	46	13	-63

서버의 특성을 3차원으로 축소한 결과,

- 개인플레이 요인:** 솔로 경험치나 낚시, 개인상점 등 독립적인 플레이와 관련된 요인
- 거래요인:** 아이템 거래와 관련 있는 요인
- 협동플레이 요인:** 파티경험치와 혈맹 관련 변수와 연관성 높은 요인

#### K-Means Clustering

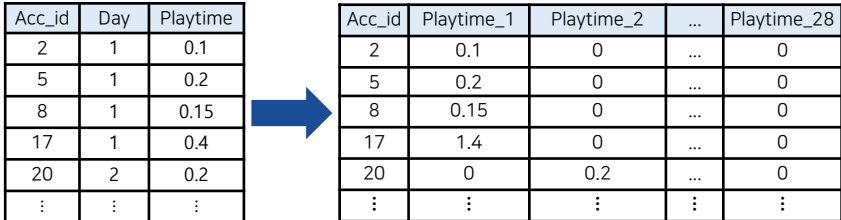


- Elbow point를 활용하여 최적의 군집 수가 5개임을 확인 후, K-Means clustering 진행
- 군집화 결과를 Factor Analysis 결과와 결합하여 3차원으로 시각화 한 결과, **군집들이 특성 차이를 보이며 군집화가 이루어졌음을 확인 가능**
- 예를 들어, y축(협동플레이 요인) 값이 상대적으로 큰 군집에 포함된 서버 'bd'는 파티경험치 평균값이 타 서버에 비해 4배부터 300배 까지 큰 값을 보이며 협동 관련된 활동들이 활성화 된 서버임을 확인할 수 있음

40개의 서버를 특징적 차이를 보이는 **군집 5개로 구분**하여 **서버 변수를 추가**

### ⑥ 파생변수 생성

- 일별 Playtime 파생변수 생성

변수명	의미	식(방법)
playtime_day1 ~ playtime_day28	일 단위 유저별 플레이 시간	

- 서버 특성 별로 Clustering한 파생변수 생성

변수명	의미	식(방법)
Server	Server Cluster 0, 1, 2, 3, 4	activity, combat, pledge, trade 테이블 바탕으로 서버를 Cluster 5개로 K_means

- 누적결제금액 파생변수 생성

변수명	의미	식(방법)
total_as	총 결제 금액	총 결제 일수 * 한 결제 당 평균 결제액


### ⑥ 파생변수 생성

- 투자한 금액, 시간 대비 레벨 변화량 파생변수 생성

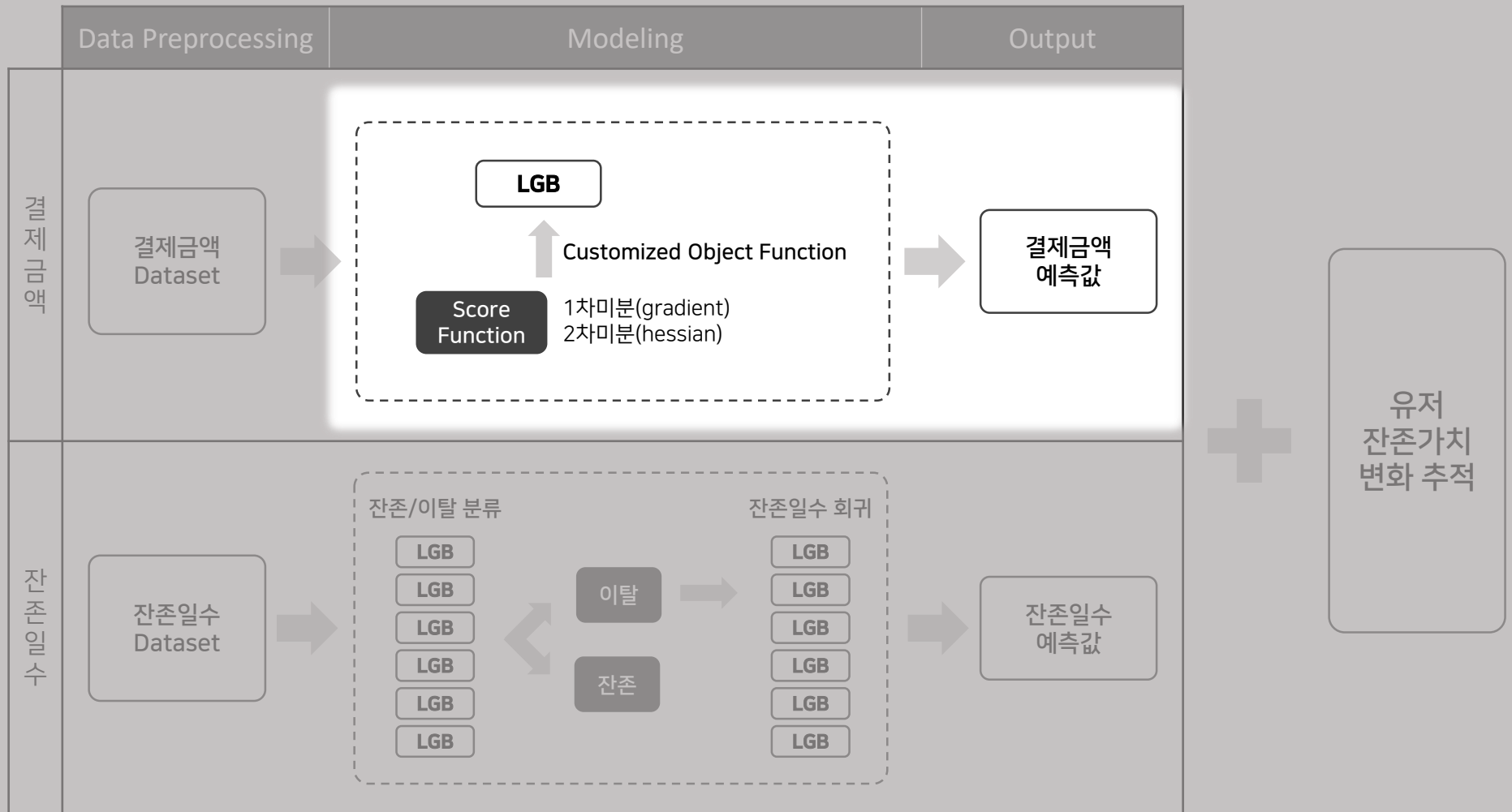
변수명	의미	식(방법)
level_change_pay_weighted	투자한 금액 대비 레벨 변화량	$\frac{\max_{level} - \min_{level} + 0.01}{pay\_mean + 0.01}$
level_change_time_weighted	투자한 시간 대비 레벨 변화량	$\frac{\max_{level} - \min_{level} + 0.01}{meanPlaytime + 0.01}$

- 혈맹의 특성에 따라 파생변수 생성

변수명	의미	식(방법)
pled_active_group	혈맹 소속원들의 활발정도	$\frac{combat\_char\_cnt + 1}{play\_char\_cnt + 1}$
pled_active_war	혈맹의 전투 Volume	$np.log(pledge\_combat\_cnt * temp\_cnt * etc\_cnt)$
pled_active_meet	혈맹 내 전투 활발 정도	$\frac{same\_pledge\_cnt + 1}{pledge\_combat\_cnt + 1}$
pled_aggressive	혈맹원들의 공격성 정도	$\frac{random\_attacker\_cnt + 1}{random\_defender\_cnt + 1}$

- 
- I. EDA & Data Preprocessing
  - II. Amount Spent Modeling**
  - III. Survival Time Modeling
  - IV. Conclusion

## ① Pipe Line

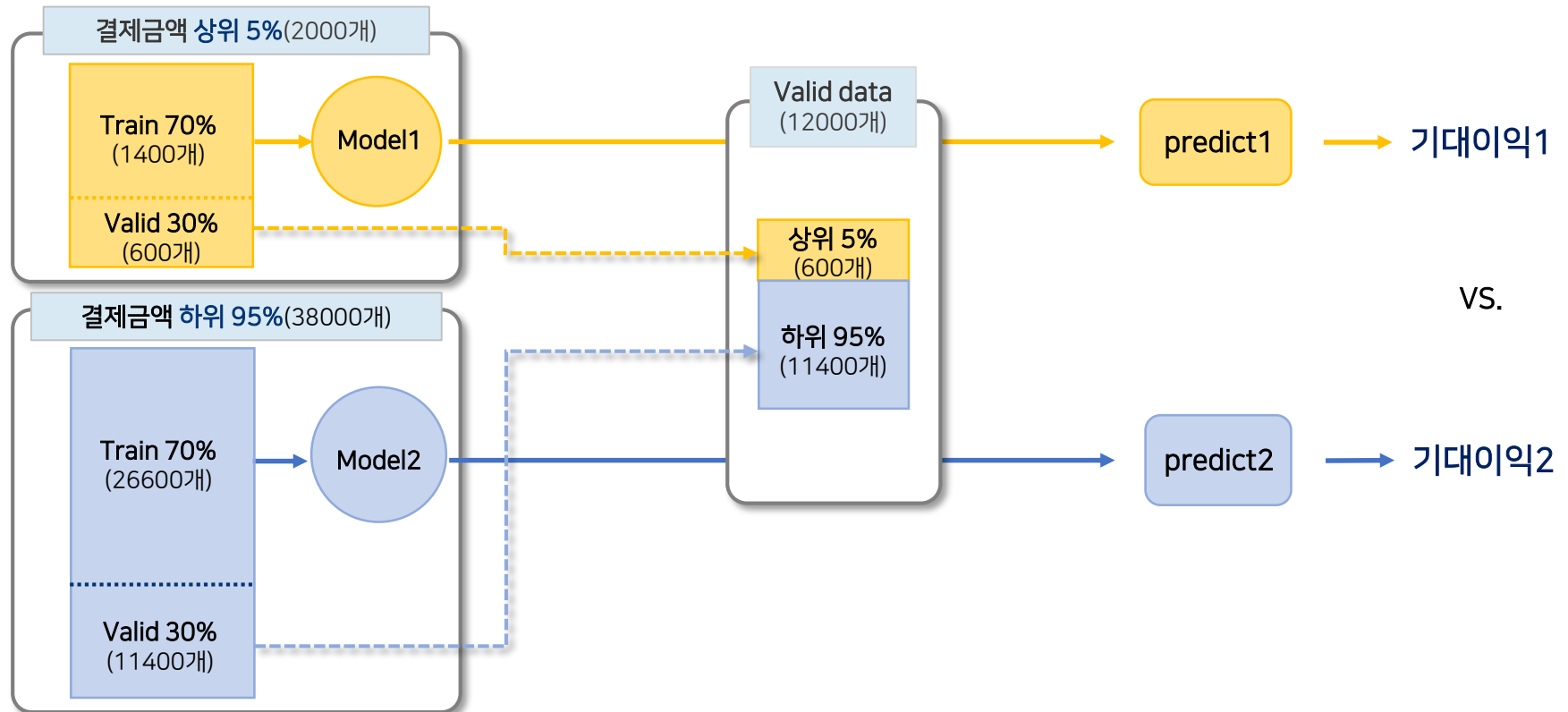


### ① 모델링 설계 배경과 가설 검증

*"long-term loyal customers의 사용자 당 CLV(Customer Lifetime Value)는 non-loyal customer의 CLV보다 약 300 배 높고, 이는 loyal customer 한 명이 이탈하는 것을 방지하면 300명의 non-loyal customers가 이탈하는 것을 방지 할 수 있는 것과 비슷한 예상 이익을 얻을 수 있음"*

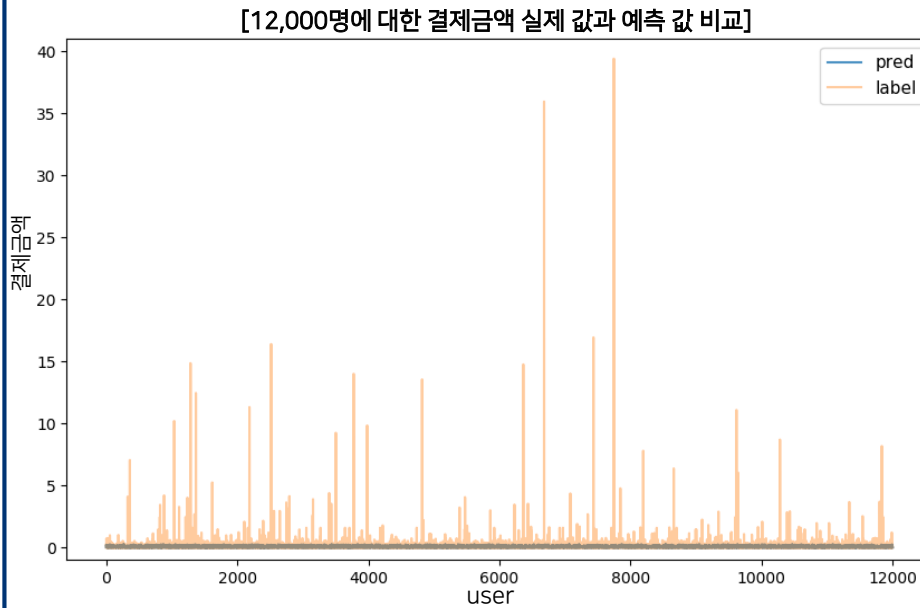
E. Lee et al, "Profit Optimizing Churn Prediction for Long-term Loyal Customer in Online games," in IEEE Transactions on Games.

- 해당 내용을 검증하기 위해, **결제금액 상위 5%와 하위95%**의 data를 학습한 Model1, 2 생성
- 각각의 validation data를 합쳐서 하나의 validation data를 생성하고, Model1, Model2의 **기대 이익을 계산하여 비교**



### ① 검증 결과

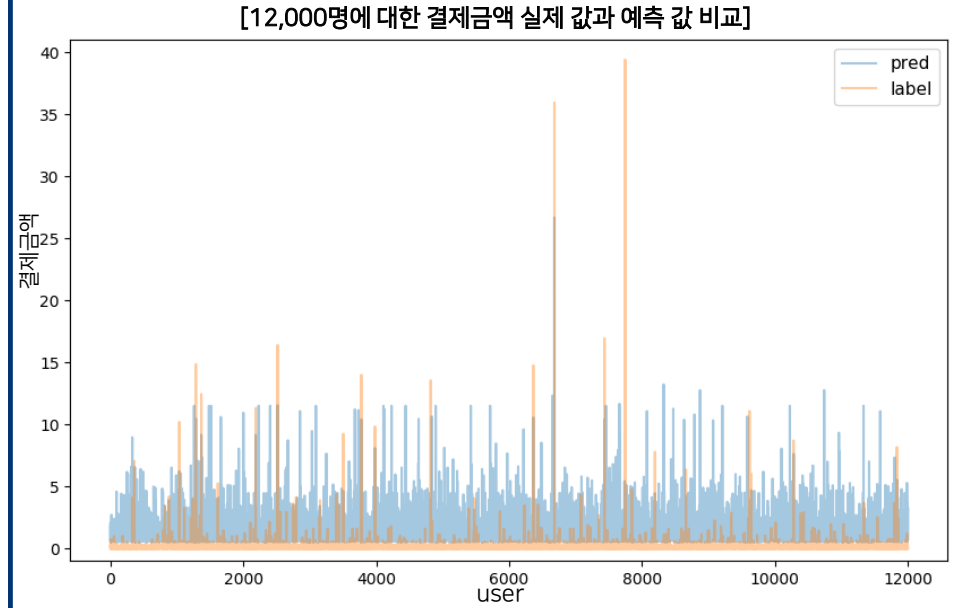
결제금액 하위 95% data(38000개) 학습 모델



기대이익 = 8,529

실제 Validation set의 최대값이 약 39인 것에 비해,  
예측값이 모두 0.3 미만으로 예측됨

결제금액 상위 5% data(2000개) 학습 모델



기대이익 = 18,211(약 2.14배 증가)

예측값이 모두 0.3 이상으로 예측됨  
기대이익에 중요한 높은 결제금액 또한 상대적으로 높게 예측

✓ 95%(3만8천명)의 non-loyal 유저보다 5%(2천명)의 loyal 유저를 학습할 때 기대 이익이 약 2.14배 증가함

**“기대이익을 높이기 위해서는 모든 유저가 아닌 결제금액이 높은 유저를 잘 예측해야 한다”**

#### ① 결제금액 상위 5%의 data만 학습할 시 문제점

- 상위 5%의 data 수(2,000개)가 전체 data 수(4만개)에 비해 적어서 **과적합 위험**
- 학습이 안 된 하위 95%에 대해서 예측비용이 커짐에 따라 **기대이익이 작아짐**  
(기대이익 = 잔존가치 x 잔존율 - 예측비용)

#### ② 학습 모델의 목적 함수를 결제금액에 대한 기대이익 공식으로 Customizing

- 기대 이익**  
=  $(30 \times \text{결제금액 실제값}) - (0.3 \times \text{결제금액 예측값})$
- 기대 이익의 제공**  
=  $(900 \times \text{결제금액 실제값의 제공}) + (0.09 \times \text{결제금액 예측값의 제공}) - (18 \times \text{결제금액 실제값} \times \text{결제금액 예측값})$ 
  - 결제금액 예측값에 대한 1차 미분**  
=  $(0.18 \times \text{결제금액 예측값}) - (18 \times \text{결제금액 실제값})$
  - 결제금액 예측값에 대한 2차 미분**  
= 0.18

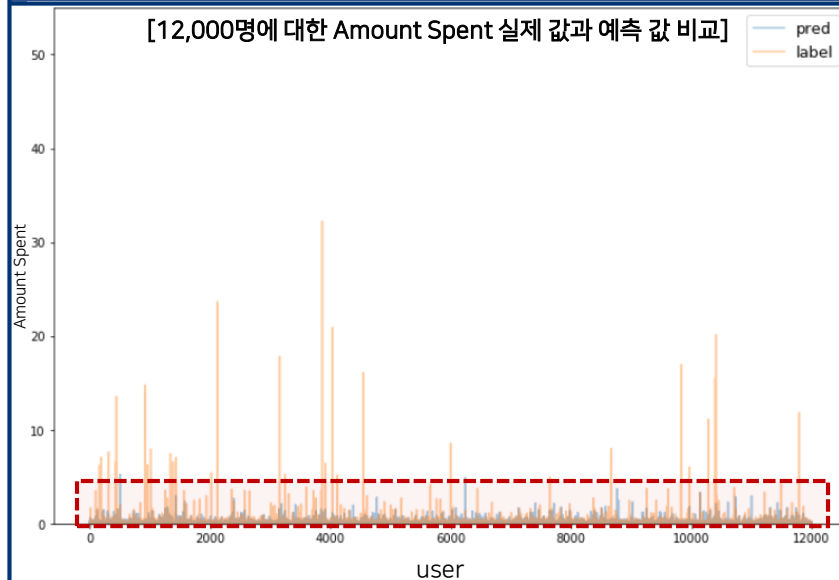
목적 함수  
(Object function)

- 가정1: (잔존기간 예측값  $\neq$  64) and (잔존기간 실제값  $\neq$  64) and (잔존기간 예측값 = 잔존기간 실제값)
- 가정2: (결제금액 실제값  $\neq$  0) and (결제금액 예측값  $\geq$  결제금액 실제값)



### ① 기대이익에 기반한 Customized 모델 성능 평가

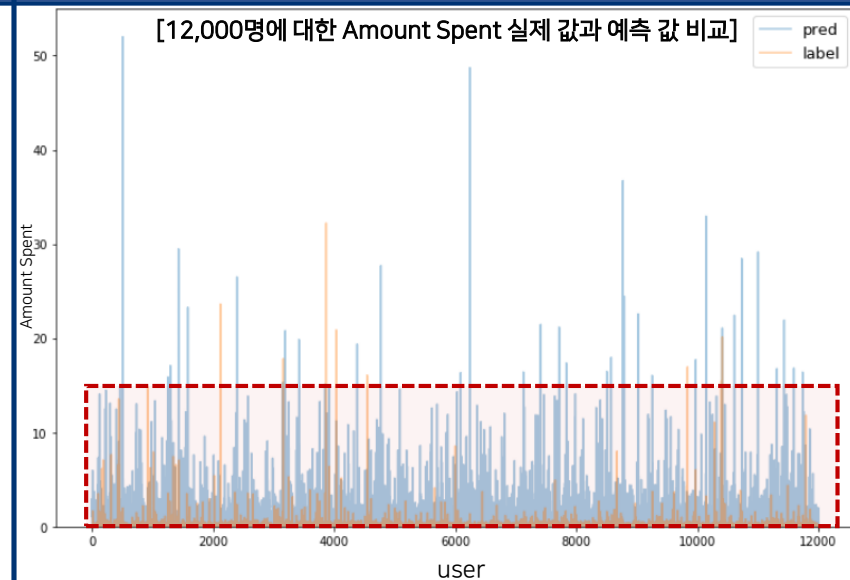
Lightgbm 목적 함수: minimize MSE



**기대 이익: 15,847**  
(잔존기간은 정확히 예측했다고 가정)


- MSE를 줄이려면 모든 유저에 대해 잘 학습해야 함
- 전체의 99%가 0~1.2 사이에 밀집되어 편향된 분포를 보임
- 따라서 모델은 실제 amount spent에 비해 값을 낮게 예측하는 경향이 있음 (0~5.2 사이에 분포)

Lightgbm 목적 함수: maximize 기대이익

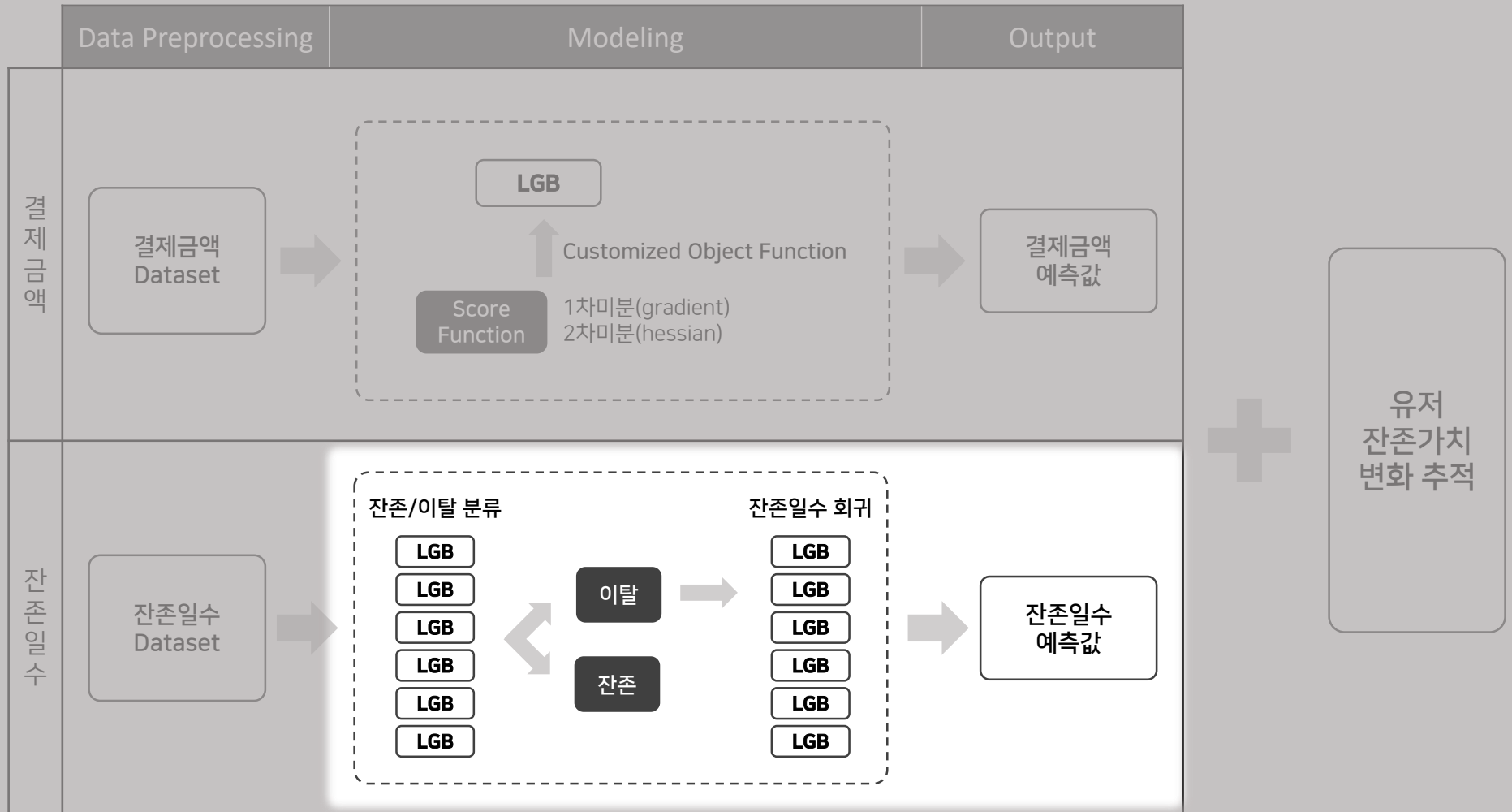


**기대 이익: 31,543 (약 2배 증가)**  
(잔존기간은 정확히 예측했다고 가정)

- 모든 유저에게 동일한 가중치를 주기 보다 **기대이익이 높은 유저에게 가중치**를 주어 학습함
- 모델은 amount spent를 0~52 사이에서 고르게 예측함
- 상위 5%의 data만 사용했을 때 보다 **예측비용**을 감소시킴

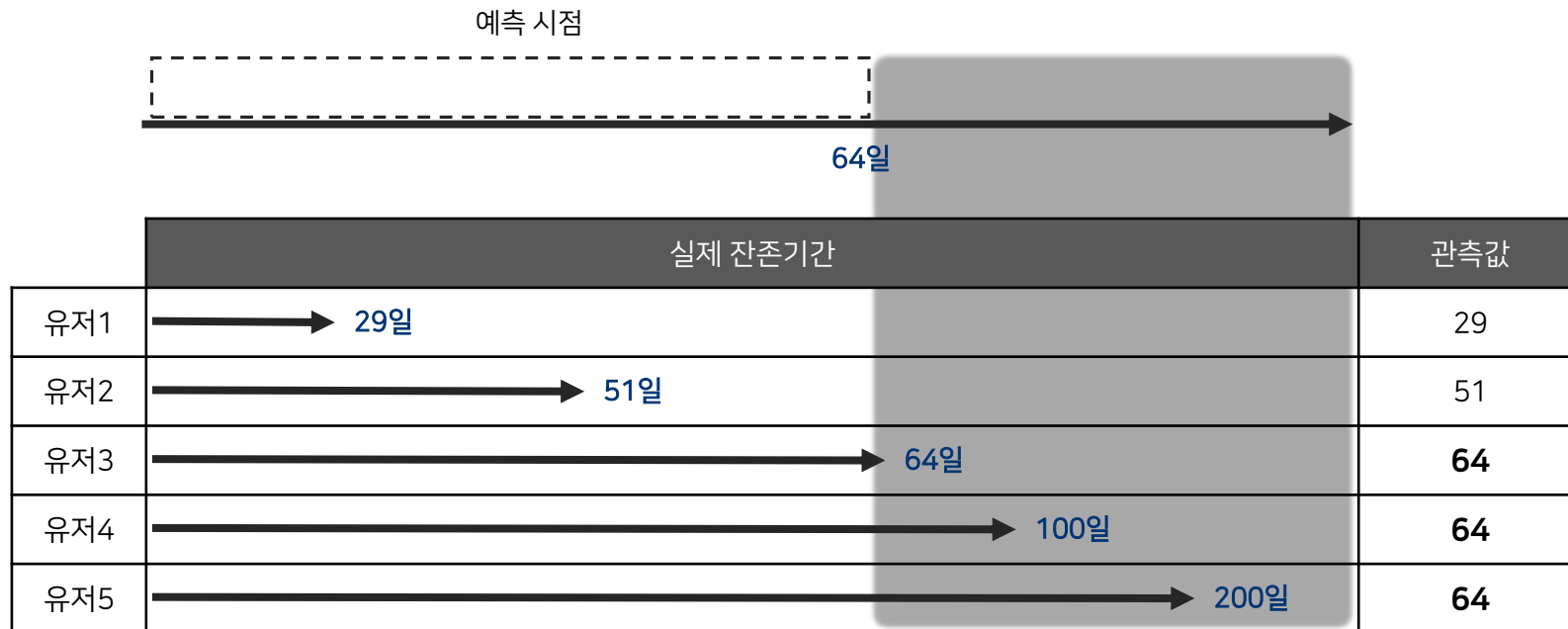
- 
- I. EDA & Data Preprocessing
  - II. Amount Spent Modeling
  - III. Survival Time Modeling**
  - IV. Conclusion

## ① Pipe Line



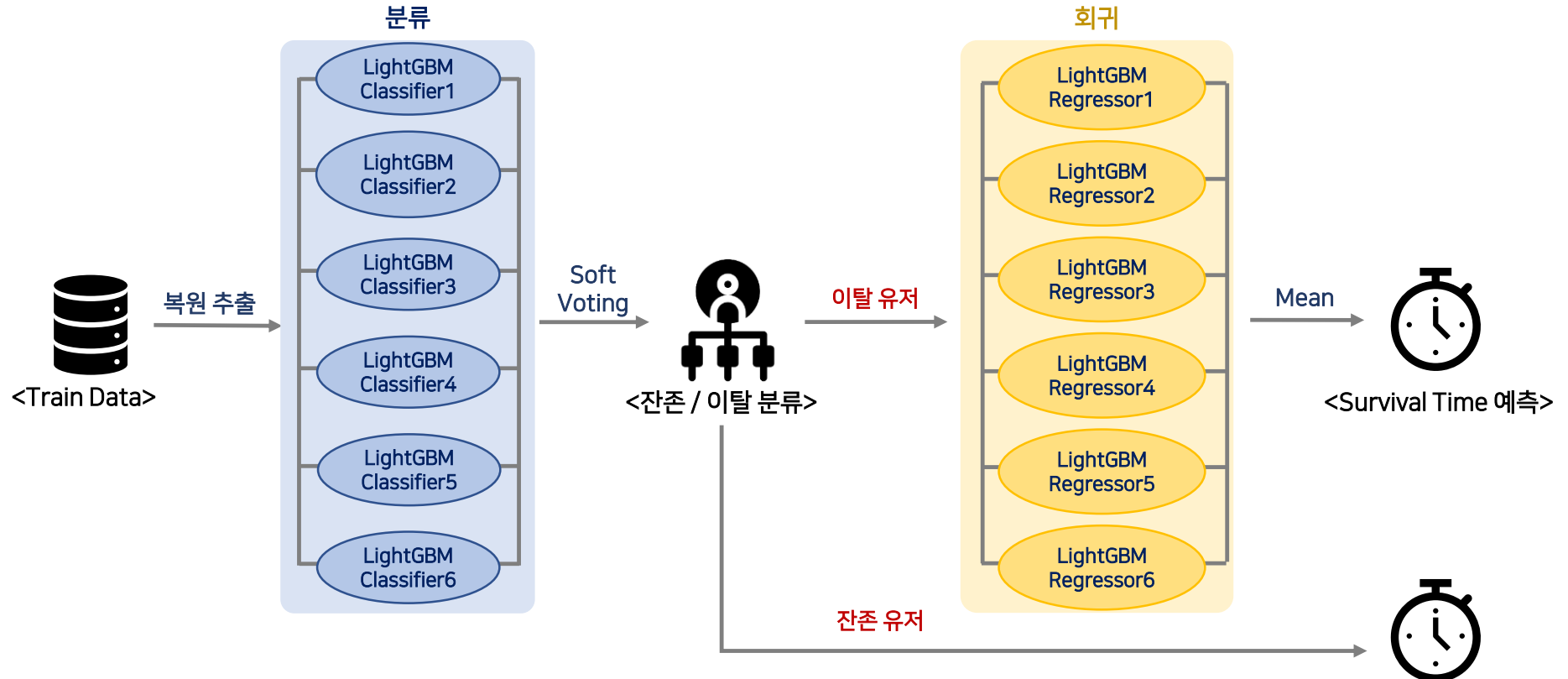
## ① 잔존기간(이탈시점) 정의

잔존기간 : 예측기간 64일 동안 유저가 잔존한 기간. 즉, 이탈한 날짜와 동일한 의미를 가짐



실제 64일만 잔존한 사용자와, 향후에도 생존한 사용자의 값이 동일하게 64로 나타남.  
즉, 1~63의 숫자는 그 자체의 값을 의미하지만, **64의 경우 '잔존'이라는 포괄적 의미를 담고 있음**

### ① 앙상블 모델 + 복원 추출

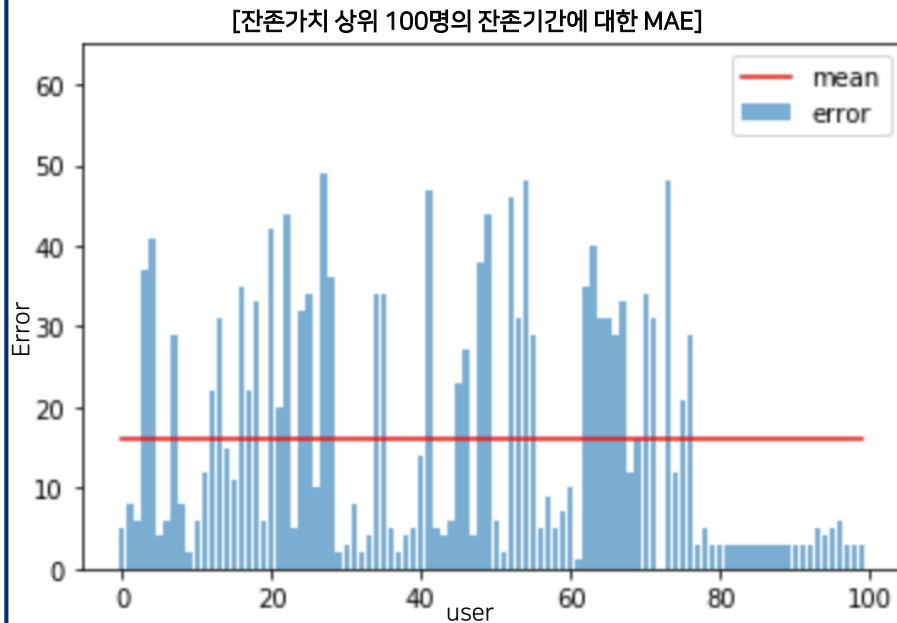


- 분류 모델과 회귀 모델에 모두 **앙상블 모델링 기법**을 이용
- 앙상블 모형을 **복원 추출**하여 학습함
- 잔존/비잔존 예측을 하는 **분류** 앙상블 모델을 이용하여 잔존 유저를 걸러 냄

- 분류 모델의 경우 Soft Voting, 회귀 모델의 경우 Mean 방식으로 각 모델들의 결과를 종합
- **회귀** 앙상블 모델을 이용하여 비잔존 유저로 예측된 유저들의 생존기간을 예측함

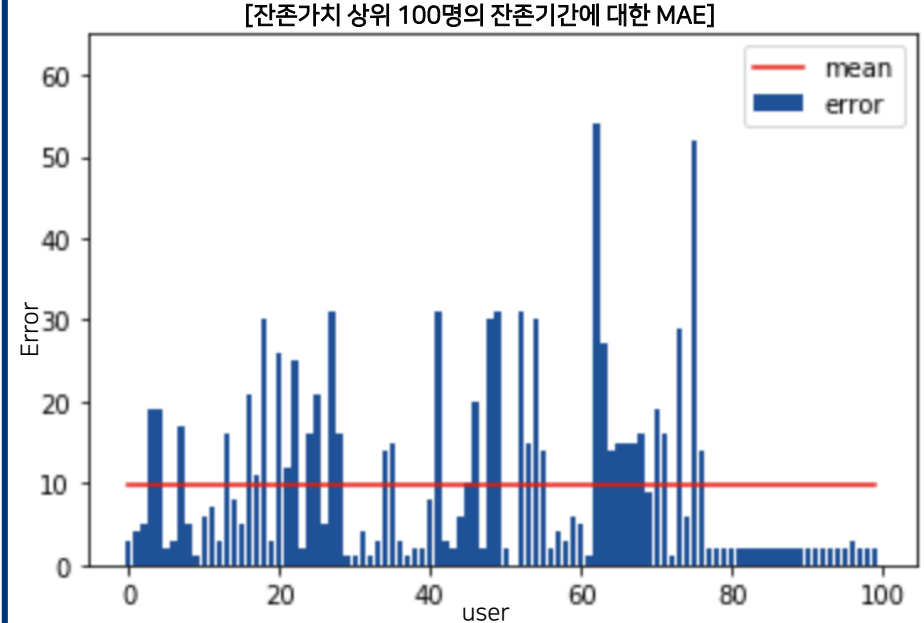
#### ① 회귀 모델 vs 분류 + 회귀 모델

회귀 모델만 사용



✓ 기대 이익: 14,527  
(결제금액은 정확히 예측했다고 가정)

잔존/비잔존 분류 모델 + 회귀 모델 사용



✓ 기대 이익: 16,258 (약 2000점 증가)  
(결제금액은 정확히 예측했다고 가정)

- 잔존기간에서 64는 단순히 64일 뿐만 아니라 잔존의 의미를 담고 있음
- 잔존가치가 높은 상위 100명의 예측값과 Label의 차이인 Error를 확인했을 때 분류 후 회귀 모델이 더 정확히 예측
- Test 데이터(12000개)를 예측한 결과로 기대 이익을 비교한 결과 분류 후 회귀 모델의 점수가 더 높음

### ① 가중치(weight)를 부여한 모델 학습

잔존 기간(survival time)을 예측하는 데 가중치 사용

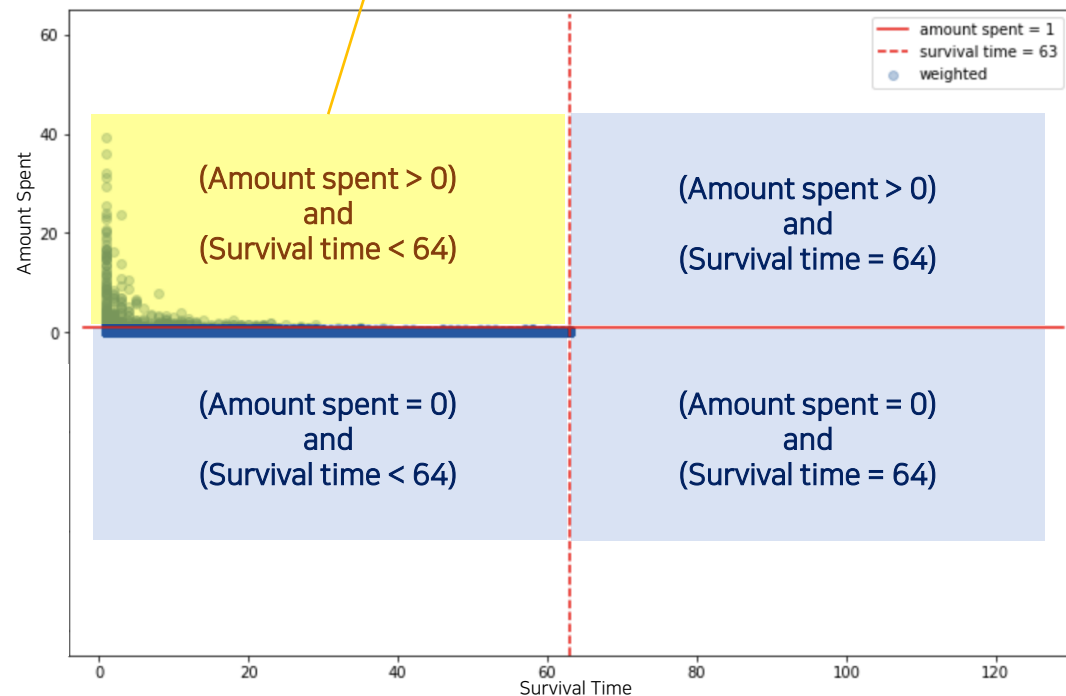
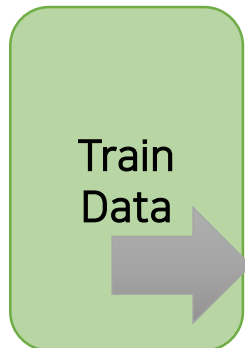
금(amount spent > 0) 유저 이면서 이탈(survival\_time < 64) 유저에게

중치 사용

잔존기간이 큰 유저의 Survival Time을 정확히 예측하는 것이 중요함

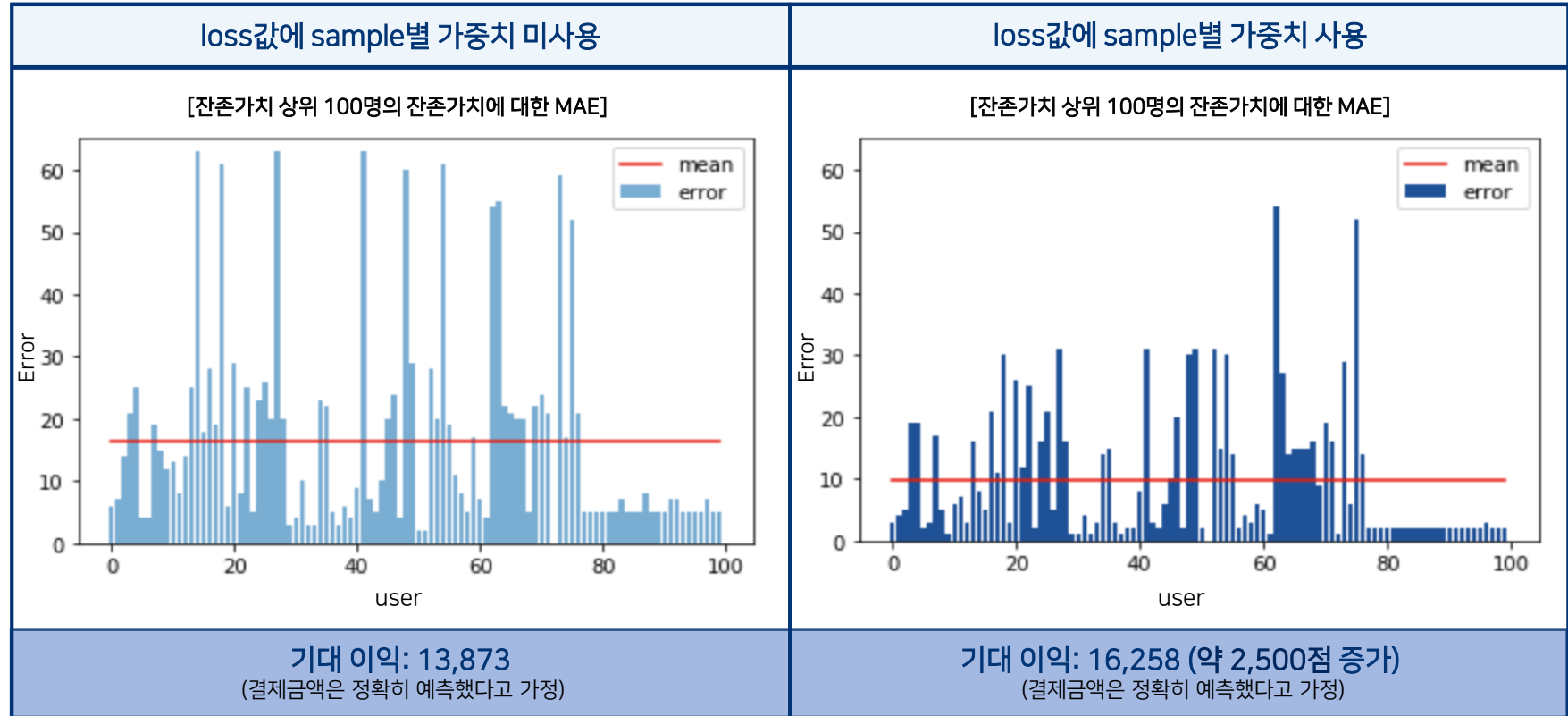
중치를 amount spent에 비례하게 부여

가중치 부여 (가중치 = Amount Spent × 자연상수(e))




[가중치를 부여 받는 유저 분포]

### ① 가중치 사용 유무에 따른 모델 비교

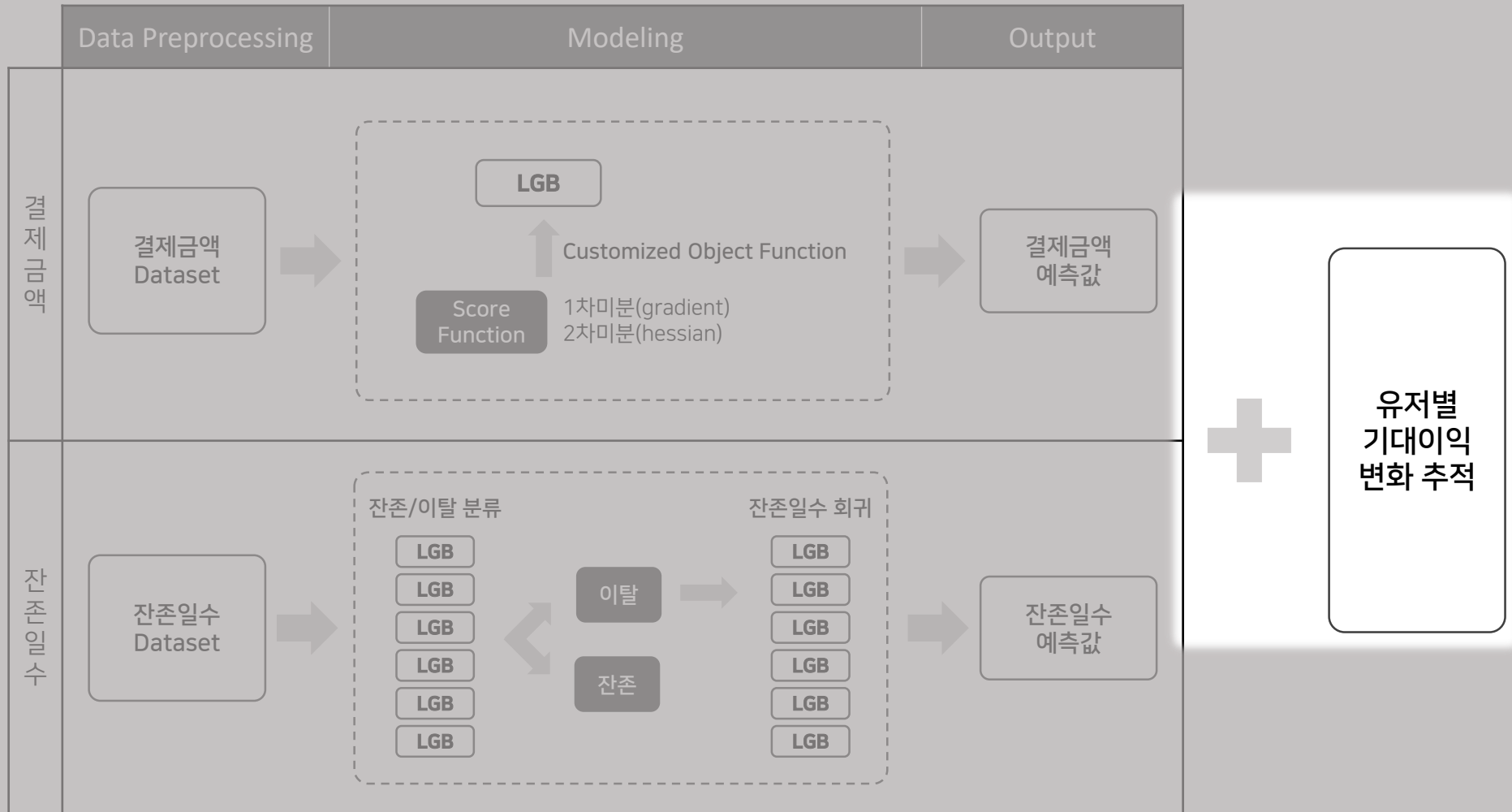


- 가중치를 사용했을 때(444.67)에 비해, 가중치를 사용하지 않았을 때 MSE(399.39)가 더 낮게 예측
- 반면, 기대이익은 가중치를 사용한 경우가 더 높게 산출
- 기대이익이 높은 유저의 잔존기간을 정확하게 예측하는 방향으로 학습을 진행해야 함



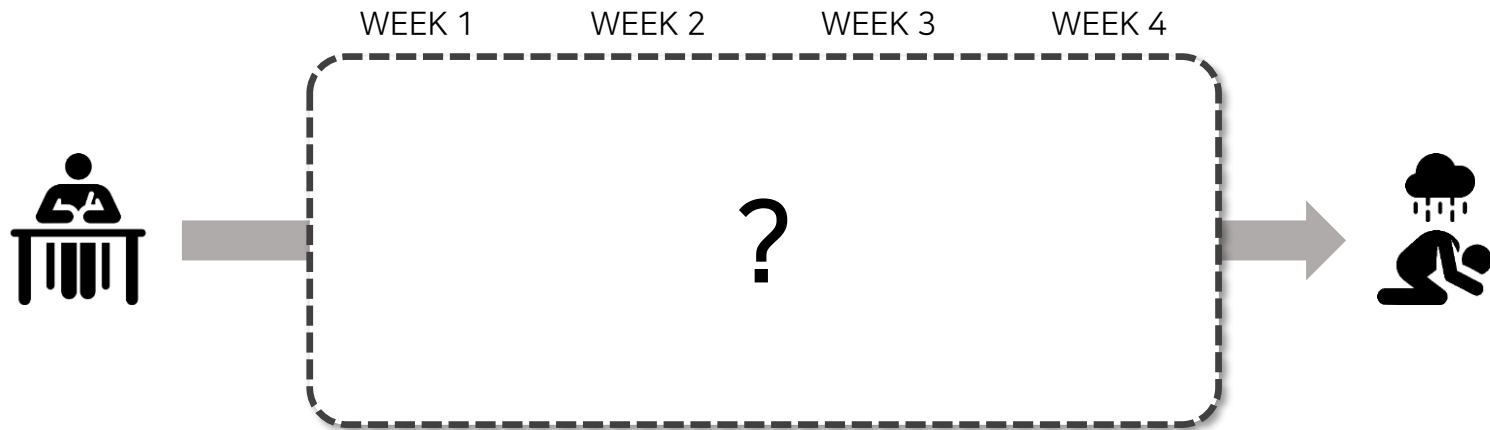
- 
- I. EDA & Data Preprocessing
  - II. Amount Spent Modeling
  - III. Survival Time Modeling
  - IV. Conclusion**

## ① Pipe Line



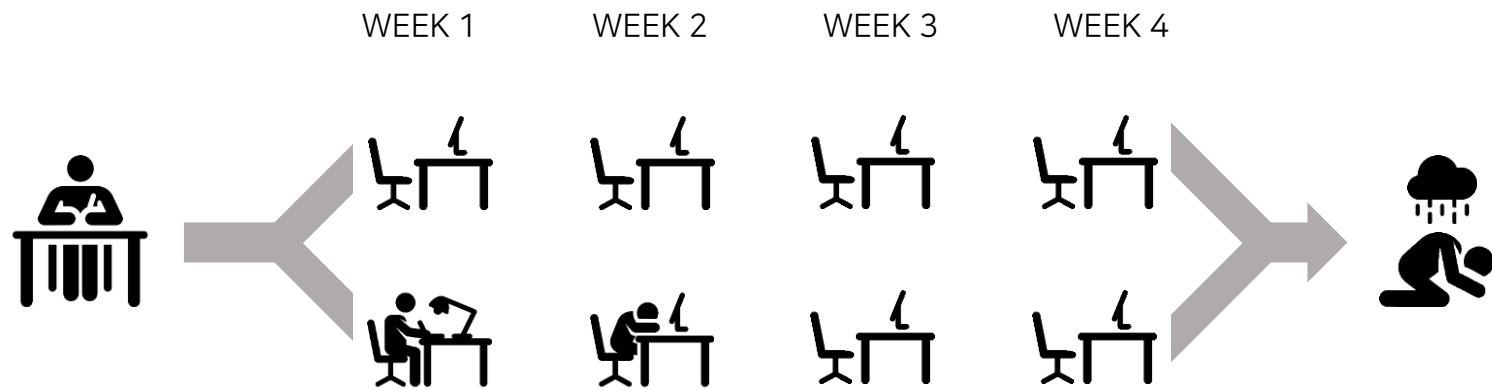
## ① 모델이 예측한 방식

[예시] 한 달 간의 데이터로 학생의 성적을 예측하는 문제



- ✓ 지금까지의 모델은 **28일간의 데이터**를 바탕으로 **최종적으로 미래에** 시험을 잘 볼 것인지에 대한 예측.  
즉, 28일동안 변화하는 해당 시점의 예측성적이 어떻게 바뀌는지 확인할 수 없는 **Black Box**

## ② 이탈징후 포착



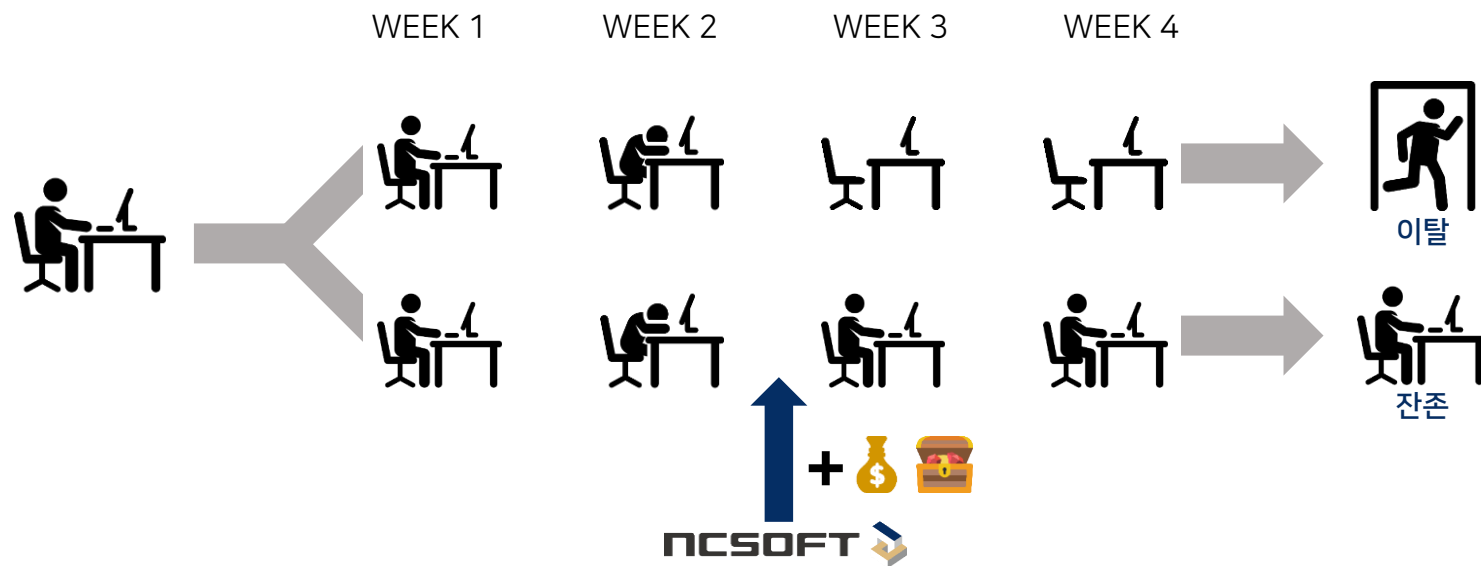
- ✓ 즉, 예측 대상이 최종적으로 시험을 망쳤다고 할 때,
    - ① 한 달 내내 공부를 열심히 하지 않아서 망친 경우
    - ② 초반부에는 공부를 열심히 하였지만, 도중에 공부를 열심히 하지 않은 경우
- 두가지 경우를 구분할 수 없음

## ② 이탈징후 포착



- ✓ 성적이 떨어질 것으로 예상되는 학생을 사전에 판별하여 선생님이 개입하였다면, 도중에 공부를 포기하는 것을 예방 가능

## ② 이탈징후 포착



- ✓ 마찬가지로, 기대이익이 낮아질 것으로 예상되는 사용자를 사전에 판별할 수 있다면 프로모션 등을 통해 기대이익이 떨어지지 않게 유지 가능

## ① 클러스터링 개요

- 학습에 사용되는 전체 유저의 특성을 단순화하고자 군집화를 진행

[Train data의 feature를 바탕으로 40,000명의 전체 유저에 대해 군집화 진행]

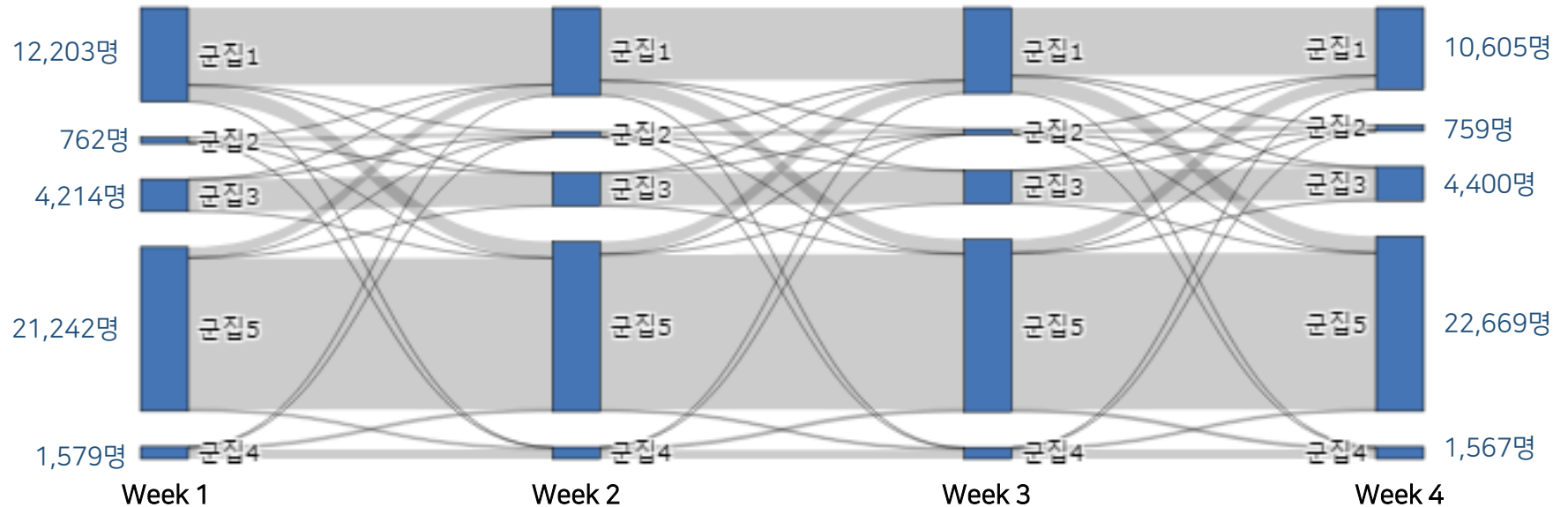
	Labels					군집 설명
	일평균 결제금액	결제 확률	누적 결제금액	잔존 기간	잔존 확률	
군집1	0.19	0.52	2.50	30.45	0.26	결제금액이 높고 잔존기간은 짧아 이탈 확률이 높음 (기대이익에 있어 큰 잔존가치를 차지하는 중요한 집단)
군집2	0.09	0.37	1.91	58.34	0.86	결제금액관련 지표가 모두 낮고 잔존기간과 잔존확률은 높음
군집3	0.01	0.14	0.57	59.22	0.88	결제금액관련 지표가 낮고 잔존확률은 높음 (기대이익에 있어 상대적으로 중요도가 낮은 집단)
군집4	0.22	0.97	13.00	59.80	0.86	일평균결제금액과 잔존기간, 잔존확률 등 모든 지표가 높은 집단
군집5	0.11	0.70	4.78	49.59	0.61	대부분의 지표에서 중간값을 가지는 집단

\* 기대이익에 있어서 중요한 군집의 우선순위 : 군집1 > 군집4 > 군집5 > 군집2 > 군집3

- ✓ 시간의 변화에 따른 유저가 군집 간 이동하는 경로를 추적하여 유저의 기대이익의 변화 추세를 확인하고  
변화 추세 별로 구분되는 기대이익 변화의 징후를 파악하고자 함

## ② Sankey Diagram 시각화

[4주 간 Train Data 40,000명의 유저에 대한 Sankey Diagram 결과]



✓ 앞선 각 군집별 label 값의 지표를 확인한 결과, **기대이익**에 있어서 중요한 군집의 **우선순위**는 다음과 같음

- 군집1 > 군집4 > 군집5 > 군집2 > 군집3

✓ 기대이익을 높이기 위해 주목해야 할 시나리오

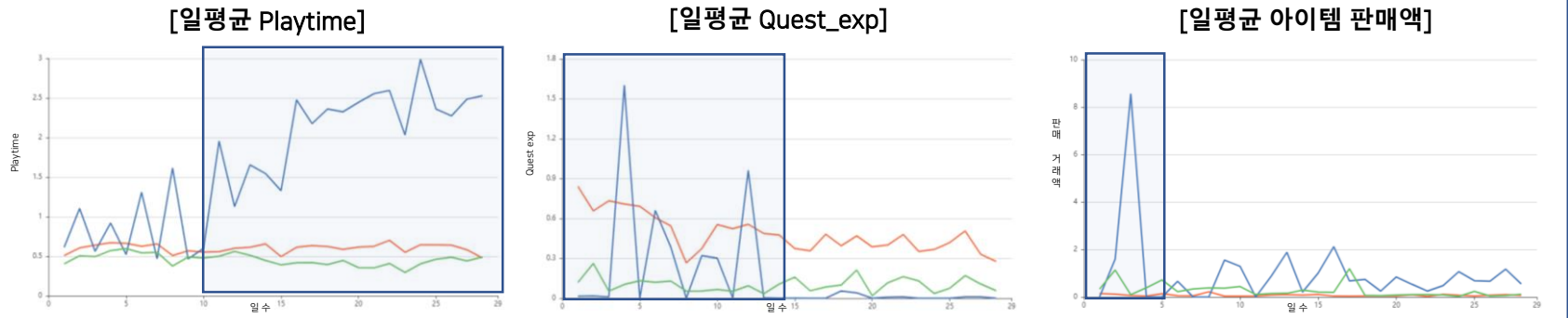
시나리오	Week1	Week2	Week3	Week4
S1) 기대이익이 높은 군집을 유지하는 유저	군집1	군집1	군집1	군집1
S2) 기대이익이 높은 군집에서 낮은 군집으로 이동하는 유저	군집1	군집1	임의의 군집	군집3
S3) 기대이익이 낮은 군집에서 높은 군집으로 이동하는 유저	군집3	임의의 군집	군집1	군집1



## ③ 시나리오 간 주요 변수 비교

■ 기대이익하락(S2) ■ 기대이익유지(S1) ■ 기대이익증가(S3)

Train



Test




- 기대이익이 낮아지는 그룹(S2)의 경우, **일평균 Playtime**이 큰 변동성을 보이며 전반적으로 증가하는 추세를 보임
- **일평균 Quest\_exp** 에서 기대이익이 낮아지는 그룹(S2) 은 다른 유저군과 달리 특정 일자에 수치가 급등하는 패턴을 보임
- 기대이익이 낮아지는 그룹(S2) 은 특정 일자에 **일평균 아이템 판매액**이 급증하는 변동성을 보임

### ① 모델링 결과

- **기대이익이 높은 고객**의 결제금액과 잔존기간을 정확히 예측 하는 것이 기대이익을 높이는 방법임
- 단순한 MSE loss function으로 학습을 시키게 될 경우 실제 값과 예측값의 차이인 데이터 전체의 오차가 줄어 들지만, 기대이익이 낮은 유저들과 잔존가치가 높은 유저들의 구별없이 오차가 줄어드는 방향으로 학습하기 때문에, 정작 중요한 **기대이익이 높은 유저들의 예측력이 떨어질 수 있음**
- 우리에게 중요한 것은 기대이익이 낮은 유저들에 대한 예측력이 떨어지더라도 **기대이익이 높은 유저들에 대한 예측력을 높여 기대이익을 최대화**하는 것임
- 따라서, 결제금액 학습 모델에는 **기대이익을 loss function으로 구현**하여 기대이익을 최대화하는 방향으로 학습을 진행함
- 그리고 잔존기간 학습 모델에는 잔존가치가 높은 유저의 loss값에 가중치를 결제금액에 비례하게 부여하여, 기대이익이 높은 유저들의 잔존기간을 잘 예측하는 방향으로 학습을 진행함
- 결과적으로, 기대이익이 높은 유저들의 결제금액과 잔존기간을 잘 예측할 수 있게 되어 높은 기대이익을 얻는 모델이 학습됨

### ② 유저 클러스터 변화 추적 결과

- 모든 유저에게 동일하게 집중하기 보다 기대이익이 높은 유저들을 우선순위로 타겟 하는 것이 중요
- 유저들을 특성에 따라 군집화 했을 때, 군집 별 일정한 패턴을 보이고, 시간에 따라 그 특성이 변화할 수 있음
- 시간에 따른 유저의 군집 변화 양상 분석 시, 기대이익이 높은 유저 군에서 이탈 징후를 보이는 유저군으로 변화하는 유저들에 대해서는 타 유저들에 비해 Activity, Trade 등의 **활동관련 데이터에서 변동성이 크다는 특징**을 발견할 수 있음
- 따라서 이러한 유저들에 대해 미리 징후를 파악하여 이탈 위험군 분류 및 조치가 필요함



감사합니다