

# <게임 유저 이탈 분석 및 예측 모형 설계>

## 0. Intro

- 주제: 유저들의 로그 데이터를 활용하여 이탈 유저 분석 및 예측하기
- 분석대상: Blade & Soul 2018년 OpenData
- 선정이유: 게임 오픈 API를 포함하여, 공개된 게임 데이터 중 가장 방대하고 자세하게 제공된 데이터

## 1. Data

- 규모
  - 학습 데이터: 10만 명의 게임 활동 데이터
  - 평가 데이터: 4만 명의 게임 활동 데이터
- 종류: 게임 내에서 수행하는 주요 활동량을 유저별로 1주일 단위로 집계
  - train\_activity: 접속 주/일수, 플레이 시간, 재화 획득량, 던전 입장 횟수, 채팅 횟수....
  - train\_payment: 해당 주 총 결제액
  - train\_party: 파티 구성원 리스트, 파티 생성/종료 주,일,시간
  - train\_guild: 문파원 아이디 리스트
  - train\_trade: 거래된 아이템 수량, 종류...
- 보안: 계정 아이디는 해싱을 통해 마스킹 처리, 통계량은 표준화를 통해 변환
  - 표준화된 데이터로 합, 평균 값을 구할 시 왜곡 일어날 수 있으므로 minmax, robust scaling
- 유저 이탈 라벨
  - week: 1주 이내 이탈
  - month: 2~4주 이내 이탈
  - 2month: 5~8주 이내 이탈
  - retained: 잔존

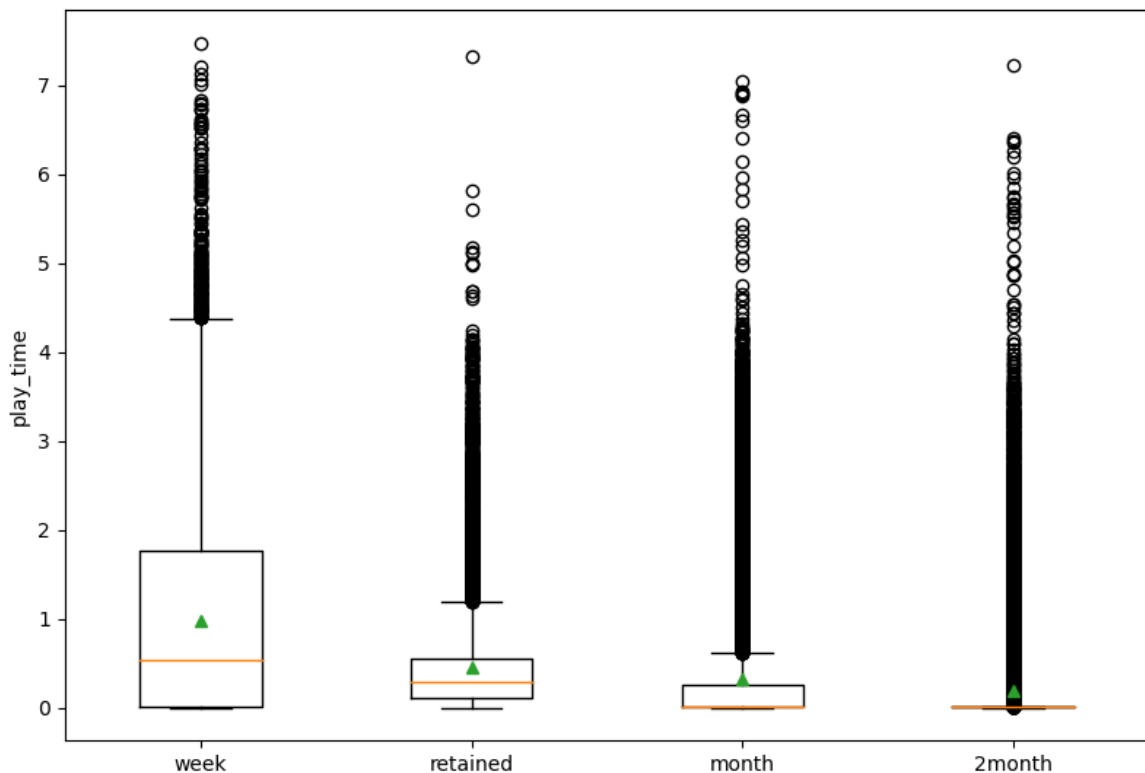
## 2. EDA

- 계정 별 8주간 총 플레이 시간 확인

- 총 플레이 시간 결과, 평균/중위값 둘 다 retained-week-month-2month 순으로 많았다.

	retained	week	month	2month
mean	0.982391	0.444441	0.318746	0.186074
min	0.000030	0.000007	0.000003	0.000002
max	7.470863	7.316027	7.038576	7.228317
50%	0.536930	0.288132	0.007996	0.004348

라벨 별 총 플레이 시간 Box plot (Δ의 경우 평균값)



- 8주간 접속한 일수 확인

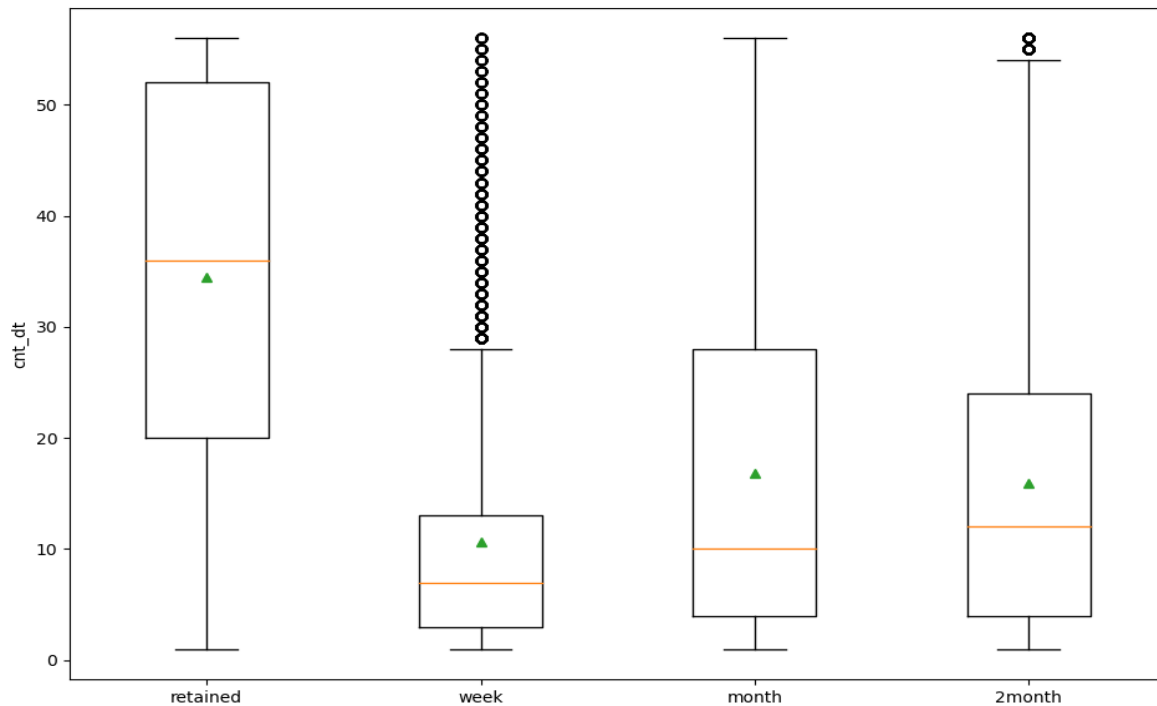
- 총 접속 일수 결과, 플레이 시간과 동일하게 retained의 평균/중위값 접속일수가 가장 많았다. 하지만 week의 경우, 플레이 시간과 달리 접속 일수는 가장 낮음을 보였다.

-> 이는 week 라벨의 경우, 데이터 수집 기간인 8주동안 신규 접속하였기 때문에 다른 라벨의 기존 유저보다 접속일수는 낮게 나왔지만, 처음 게임을 접하였기 때문에 다른 라벨보단 좀 더 오래 게임을 플레이 하였을 거라고 유추하였다.

(이후 게임 플레이 후, 본인의 취향과 맞지 않아 바로 이탈했을 것)

	retained	week	month	2month
mean	34.454240	10.640000	16.841520	15.92708
min	1	1	1	1
max	56	56	56	56
50%	36.000000	7.000000	10.000000	12.000000

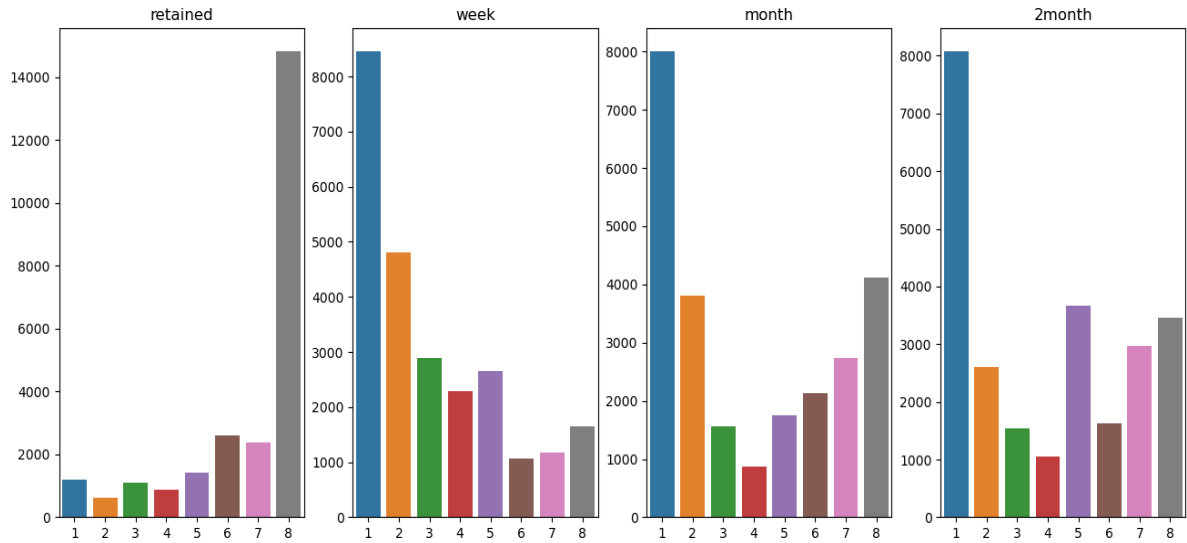
라벨 별 총 라벨 별 접속일수( $\Delta$ 의 경우 평균값) [단위: 일]



- **cnt\_week: 접속한 주 개수**

- 8주간 접속한 총 몇 주간 접속하였는지 확인하였다. 확실히 retained의 경우, 계속 남아 있는 유저이기 때문에 8주간 꾸준히 접속한 유저가 매우 많았다. 하지만 이탈한 유저의 경우 1주만 접속하고 이탈한 유저가 많았지만 꾸준히 접속한 유저도 꽤 있었다.

라벨 별 접속한 주 개수

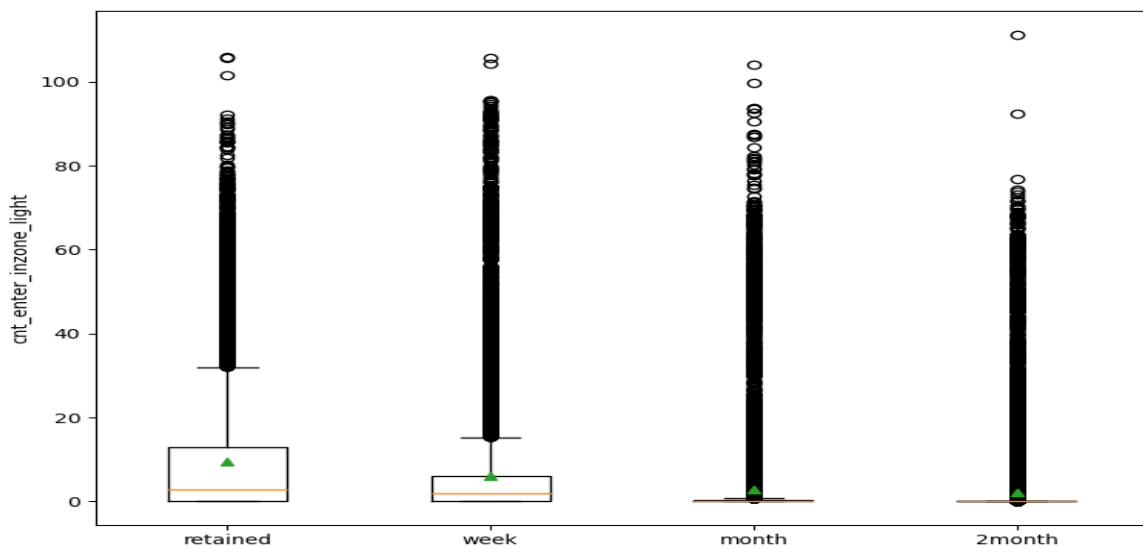


● 인턴, 레이드: 8 주간 인턴, 레이드 입장 횟수

- 인턴, 레이드도 숙련, 라이트, 솔로로 나누어져있다. 숙련, 라이트, 솔로 인턴이냐에 따라 평균 입장 횟수가 라벨별로 다르게 나왔다.
- 라이트 인턴의 경우 평균 retain-week-month-2month 순으로 많았다.

	retained	week	month	2month
mean	9.535658	6.054323	2.803785	2.124740
min	0	0	0	0
max	105.875000	105.750000	104.062500	111.250000
50%	2.875000	1.812500	0	0

라벨 별 총 라이트 인턴 입장 횟수( $\Delta$ 의 경우 평균값)



- 하지만 숙련 인턴의 경우 평균 retain-2month-month-week 순으로 많았다. week 라벨의 유저가 라이트 인턴에 비해 입장 횟수가 적다. 이는 접속일수에서 말했듯이 week 라벨의 경우 신규 접속한 유저가 다소 많기 때문에 라이트 인턴은 입장하기 쉬웠지만 숙련 인턴의 경우 입장하기 어려웠을 것이다

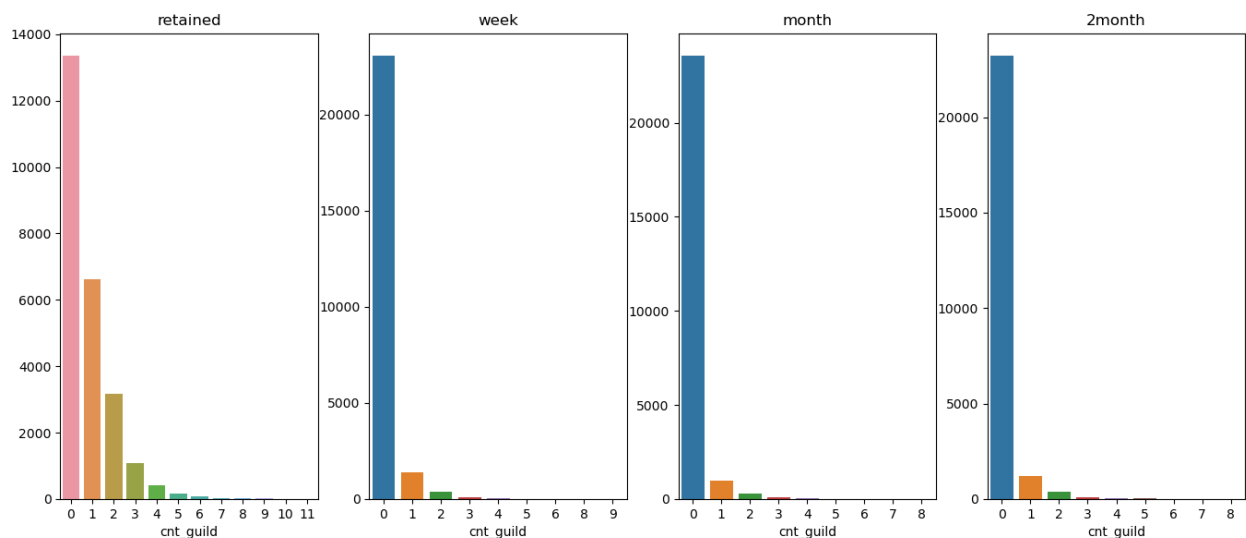
라벨 별 총 숙련 인턴 입장 횟수

	retained	2month	month	week
mean	2.998677	0.176477	0.103839	0.057343
min	0	0	0	0
max	166.611789	74.016388	82.107248	53.339745
50%	0	0	0	0

### ● 길드 가입 여부

- 라벨 별 가입한 길드의 개수 분포를 확인하였다. 이탈하지 않은 retained의 경우, 확실히 이탈한 유저들에 비해 가입한 길드의 개수가 많았다. 하지만 week, month, 2month의 경우 길드를 가입하지 않은 (0) 유저가 매우 많았고 분포도 크게 차이가 나지 않았다.

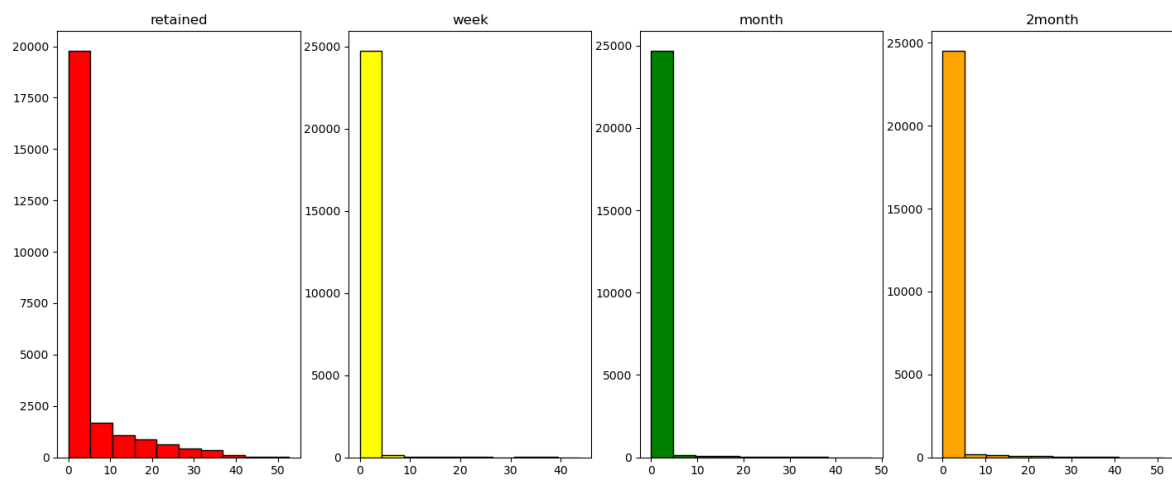
라벨 별 길드 가입 개수 분포



- 결제 정보

- 8주간 총 결제액 확인하였다. 결제를 하지 않은 (0) 인 유저가 많았지만 retained 유저의 경우 결제한 유저가 어느정도 있었다. 하지만 나머지 week, month, 2month 경우에는 큰 차이가 없었다.

	retained	2month	month	week
mean	3.997946	0.370088	0.241355	0.183446
min	0	0	0	0
max	52.734390	51.363296	47.812514	43.945325
75%	3.414419	0	0	0



### 3. Modeling

#### 1. RandomForest

- 8주간 활동내역의 합을 데이터로, 총 43개의 feature를 사용하여 모델 학습  
(플레이시간, 길드, 채팅, 거래, 던전, 경험치 등)
- GridSearch를 사용하여 하이퍼파라미터 튜닝

##### 1-1)결과

###### - Train

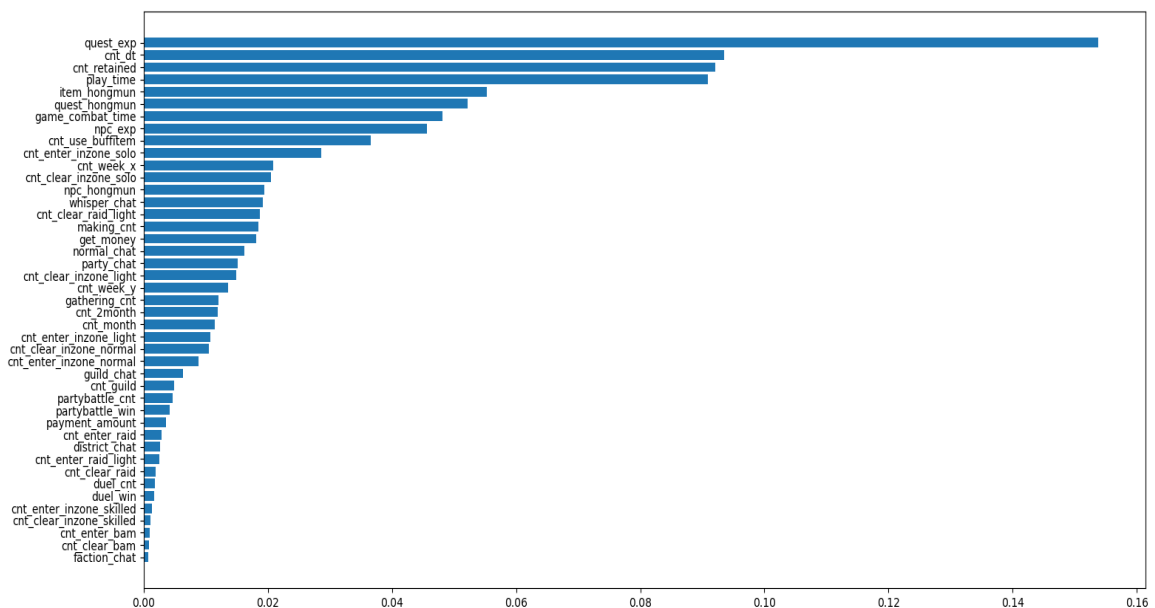
pred \ true	retained	week	month	2month	total	accuracy
retained	16066	226	324	2134	18750	85.69%
week	113	18246	213	178	18750	97.31%
month	414	678	11961	5697	18750	63.79%
2month	824	328	1743	15855	18750	84.56%

###### - Test

pred \ true	retained	week	month	2month	total	accuracy
retained	4977	205	186	882	6250	79.63%
week	269	5495	357	129	6250	87.92%
month	381	533	3115	2221	6250	49.84%
2month	707	236	880	4427	6250	70.83%

- Cross\_val\_score : 0.71853

###### - features 중요도



## 2 .ModelingStacking

- 여러 예측 모델에서 나온 결과값을 다시 학습하여 새로운 예측 모델 생성
- DecisionTree, RadomForest, KNN, AdaBoost 모델의 예측 결과값으로 다시 학습

### 1-1)결과

#### - Train

<div>pred true</div>	retained	week	month	2month	total	accuracy
retained	14612	675	525	2938	18750	77.93%
week	662	16554	1196	338	18750	88.29%
month	1148	1655	9102	6845	18750	48.54%
2month	2147	694	2401	13508	18750	72.04%

#### - Test

<div>pred true</div>	retained	week	month	2month	total	accuracy
retained	4457	736	342	715	6250	71.31%
week	162	5612	391	85	6250	89.79%
month	252	698	3056	2244	6250	48.90%
2month	414	558	1066	4212	6250	62.40%