랜덤포레스트를 활용한 유저 이탈 예측

- 리니지 게임 데이터를 중심으로

발표자 | 장선영
Date | 2022.09.30

Contents

1. 문제정의 및 데이터 수집

1.1 리니지란?

리니지란?

엔씨소프트에서 10년 이상 서비스 되고 있는 RPG 게임으로서, 한때 한국 최고 매출을 달성한 게임.

전투 및 전쟁, 생활, 거래, 혈맹(길드) 등 다양한 데이터가 존재하여, 게임 내 유저들의 다양한 상호작용을 관찰할 수 있다.

기본적으로 PVP(유저간 전투)기반의 게임으로, 소위 '막피'(막무가네 PK)가 자유롭고, 친한 유저를 끌어들여 복수가 가능하다.

이에 따라 '혈맹' 이라는 사회적 집단이 실제 인맥이 되기도 한다.

(아래) 실제 리니지 공성전 스크린샷

1. 문제정의 및 데이터 수집

1.2 데이터 수집

데이터 수집 경로

NC soft 에서 공개한 '게임유저 잔존가치를 고려한 고객 이탈예측" 데이터 사용

주요 데이터 항목

- 일일 활동 데이터(경험치 획득, 플레이시간 등)
- 전투데이터
- 거래데이터
- 소속 혈맹 데이터
- 일별 결제 금액 데이터

train_label.csv	대상 유저들의 생존 기간 및 평균 결제 금액
train_activity.csv	대상 유저의 **캐릭터별** 활동 이력
train_combat.csv	대상 유저의 **캐릭터별** 전투 이력
train_pledge.csv	대상 유저 **캐릭터별** 소속 혈맹 전투 활동 정보
train_trade.csv	대상 유저의 **캐릭터별** 거래 이력
train_payment.csv	대상 유저의 일별 결제 금액

1. 문제정의 및 데이터 수집

1.3 문제정의

문제정의

게임 내 활동 데이터를 활용한 이탈 예측(이진 분류) 모형 개발

이탈?

64일 이후에도 게임 접속 여부로 생존 측정 (7일간 미접속 시 이탈)

64일 이후에도 생존 시 (미이탈: 0)

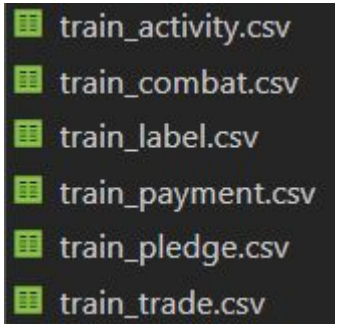
64일 미만 (이탈: 1)

2. EDA & 전처리

2.1 데이터 집계

- 모든 전처리 이전, 데이터가 분산되어있어 분석의 편의성을 위해 하나의 데이터셋으로 모아주는 것이 필요
- 다만 데이터가 관찰 일자 / 유저 계정 / 계정 내 캐릭터으로 데이터가 구별되어 있어 통합하는데 어려움이 존재.
- 유저 이탈 예측이므로 유저 아이디 기준으로 집계하고, 시계열 데이터는 '하루평균'으로 간주하여 집계

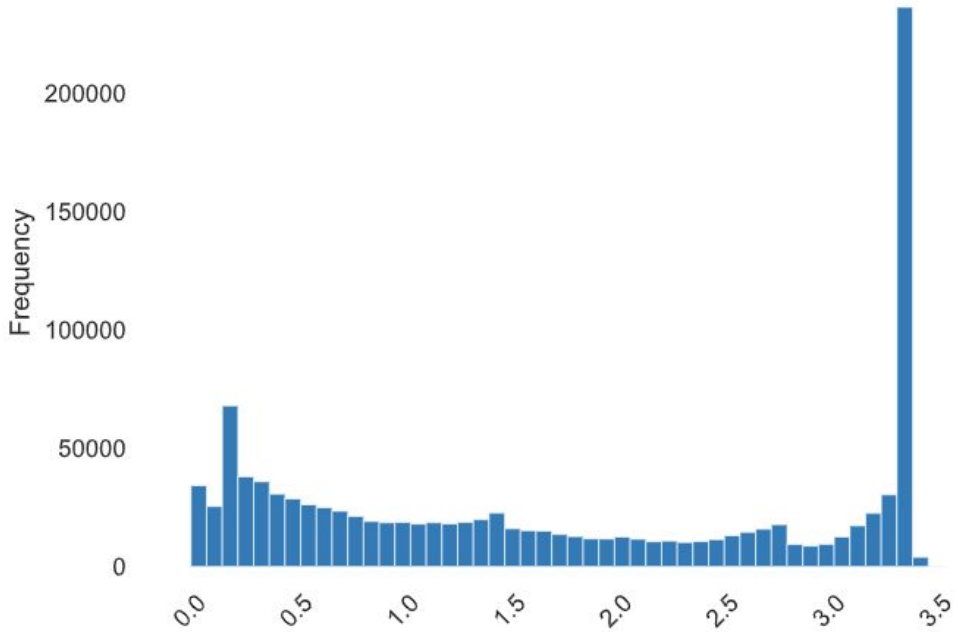
day	acc_id	char_id	server	playtime	npc_kill	solo_exp	party_exp	quest_exp	rich_monster	death	revive	exp_recove
1	75001	397380	aa	1.442	0.000	0.000	0.000	0.000	0	0.000	0.000	0.0
1	75001	216231	aa	0.283	2.248	0.047	0.000	0.000	0	0.000	0.000	0.0
1	75711	308769	aa	1.037	2.957	0.322	0.167	0.003	1	0.246	0.247	0.0
1	72230	387177	aa	0.229	4.042	0.099	0.000	0.002	0	0.000	0.000	0.0
1	34253	339862	aa	1.088	0.597	0.003	0.000	0.000	0	0.000	0.000	0.0



2. EDA & 전처리

2.2 데이터 분포 확인

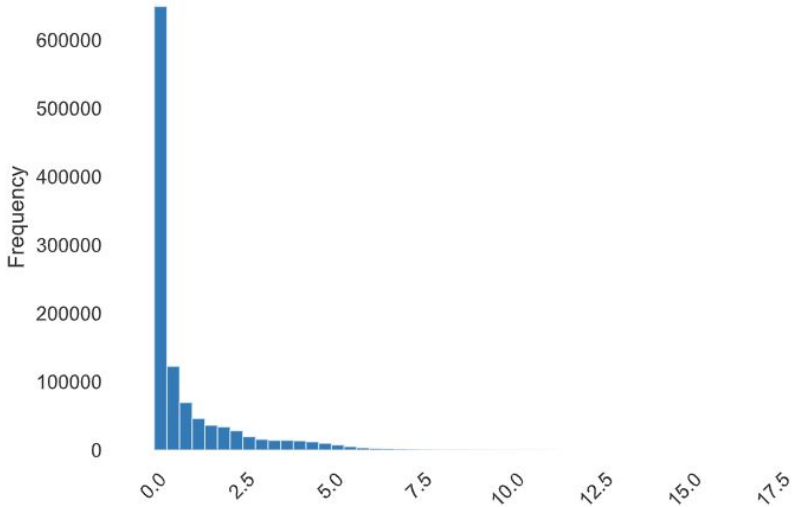
특성 개수가 많아 주요한 특성의 분포만 소개
playtime 항목



2. EDA & 전처리

2.2 데이터 분포 확인

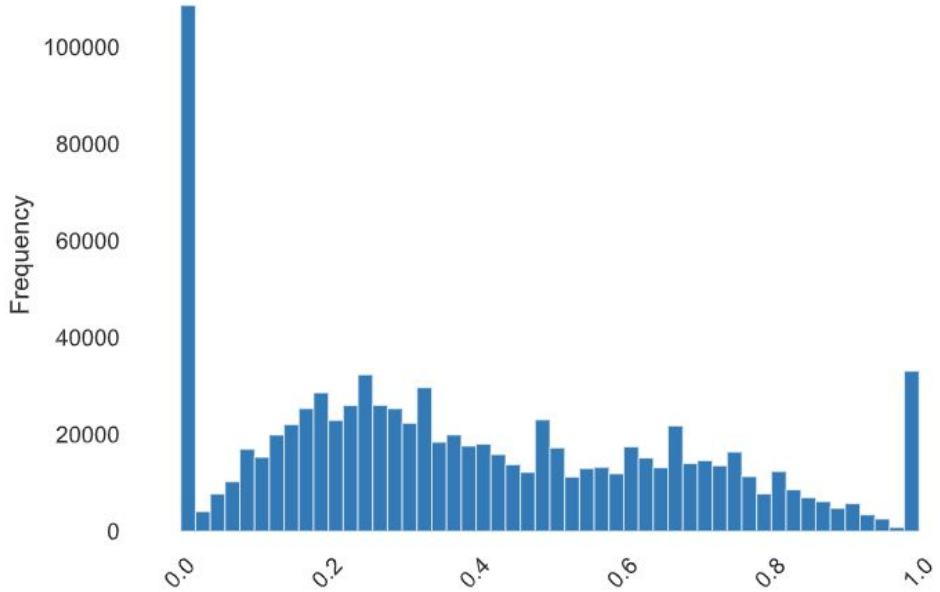
npc kill 항목



2. EDA & 전처리

2.2 데이터 분포 확인

혈맹활성화 정도(feature engineering)



2. EDA & 전처리

2.3 전처리

이상치 | 게임 내에는 다양한 유저가 존재

결측치 | 0으로 대체

중복값 | 없음

2. EDA & 전처리

2.4 피쳐 엔지니어링

피쳐끼리의 상관성을 고려하여 아래와 같은 피쳐를 생성

pledge_activated_rate : 혈맹데이터_전투에 참여 구성원 수 / 혈맹 구성원 수
-> 유저 소속 혈맹 내 전투 참여 비율을 통해 혈맹의 활성화 정도 확인

total_exp
-> 퀘스트, 솔로플레이, 파티플레이 경험치를 모두 합산한 항목

freq_selling
-> 거래 데이터_판매 횟수

freq_buying
-> 거래 데이터_구매 횟수

2. EDA & 전처리

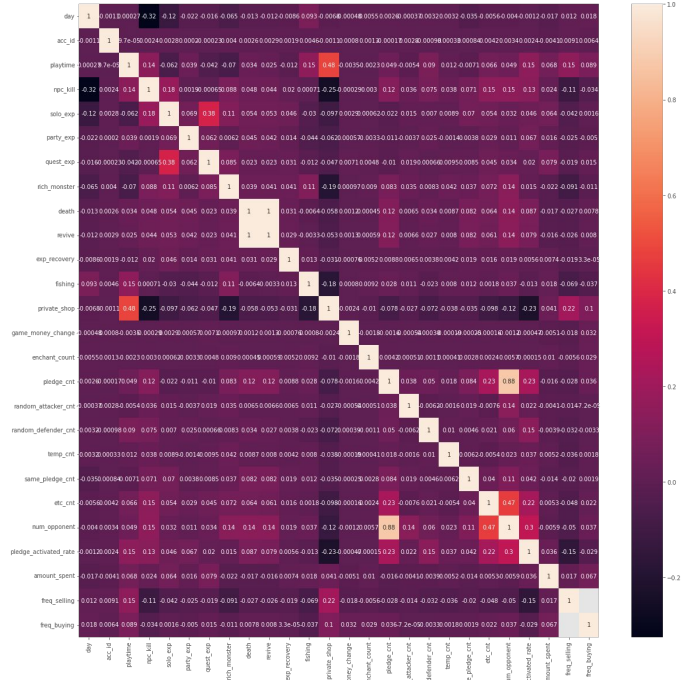
2.5 상관관계

타겟끼리의 상관관계가 높은 항목

- 플레이타임과 개인상점 운영시간
- 개인상점 운영시간과 소속 혈맹의 활성화 정도
- 전투 참여횟수와 전투 상대 캐릭터 수

개인상점 운영시간, 전투 상대 캐릭터 수 피쳐 제외

타겟과의 상관관계는 플레이타임이 높게 나타남.



3. 모델링

3.1 프로토타입 비교

대표적인 분류 알고리즘
Tree : Random forest, 부스팅 기법 중 XGBoost
Linear model : Logistic Regression

평가 지표
AUC score를 활용

성능 비교
5-fold cross validation

선정 모델
랜덤 포레스트 모델

```
roc_score of baseline:  0.5
xgb : 0.81 roc with std : 0.00
rf : 0.82 roc with std : 0.00
logi : 0.74 roc with std : 0.00
```

3. 모델링

3.2 하이퍼 파라미터 튜닝 & 최종 평가

Randomized Search 를 활용하여 결과적으로 0.02 만큼의 성능을 향상

테스트 데이터셋에 대한 최종 평가 성능

roc auc score : 0.74

정확도 : 75%

4. 모델해석 및 결론

4.1 feature importance / permutation importance

예측에 영향을 준 상위 feature

1. playtime
2. total_exp
3. npc_kill
4. pledge_activated_rate

4. 모델해석 및 결론

4.2 의의 및 보완점

의의

- 예측 모형을 통하여 유저의 이탈을 사전에 방지

보완점

- 시계열 데이터인 점을 활용할 필요
- 사회 연결망 분석을 통한 모델링
- 클러스터링을 통한 '잠재적으로 가치있는 유저' 를 대상으로 이탈 예측

감사합니다