Profit Optimizing Churn Prediction for Long-Term Loyal Customers in Online

0. Abstract

- 온라인 게임을 성공적으로 운영하기 위해서는, 고객들과의 관계가 중요하다.
  - 새 고객을 유치하는 것보다 **기존 고객의 이탈을 막는 것이 비용 측면에서 유리**하기 때문.

- 따라서, 우리는 **이익을 극대화하는 이탈 예측 방식을 제안**한다.
  - 첫번째로, 예측의 대상을 정한다.
    - 온라인 게임의 LTV 분포는 매우 편향되어있어, 소수의 유저가 대부분의 매출을 발생시킨다.
    - 따라서, **득이 되는 충성 고객에 집중**하는 것이 비용 측면에서 유리하다.
  - 두번째로, 모델의 threshold를 조정한다.
    - 예측모델의 threshold를 조정하여, 예상되는 **이탈의 정확도가 아닌 '이익'을 극대화**하는 것이 역시 비용 측면에서 유리하다.
  - 우리는 위 제안을 Aion에 적용하였고, 해당 방법이 모든 유저를 대상으로 하는 것보다 비용 측면에서 효율적임을 보일 것이다.

1. Introdution

- 이탈에 대해 다양한 연구가 진행되었으나, 이를 온라인 게임에 적용하기 위해서는 **몇가지 이슈를 고려**해야 한다.
  - 첫번째로, 게임에서의 이탈은 전통적인 형태의 서비스 기반 구독 시스템과 다르다.
    - 형식적인 탈퇴 절차 없이, 계정을 그대로 둔 채로 이탈하기 때문이다.
    - 따라서, 오탐과 미탐으로 인한 비용을 줄이기 위해 **이탈을 올바르게 정의하는 것이 중요**하다.
  - 두번째로, 이탈 **예측의 대상이 올바르게 정의**되어야 한다.
    - 신규 게임과 달리, 오래 서비스한 게임 타이틀에서는 충성 고객의 이탈을 막는 데에 집중하는 것이 더 효율적이다.
    - 게임 내에 악영향을 미치는 유저를 필터링해야 더 나은 결과를 얻을 것이다.
  - 세번째로, 이탈을 방지하여 얻게 되는 **잠재적 이익이 정량적으로 측정**되어야 한다.
    - 이탈을 막기 위해 들이는 비용이 그로 인한 이익을 초과할 수 있기 때문이다.

- 위 사항을 고려한 이탈 예측 프로세스의 특징은 아래와 같다.
  - 첫번째로, **"장기 충성 고객"을 선별**하여 이탈 예측 대상으로 삼았다.
    - 이를 위해, 각 고객들에게 인게임 행동과 과금 패턴을 기반으로 한 충성 등급을 매겼으며, 대략 **6개월 간의 충성 등급을 분석에 사용**했다.
  - 두번째로, 이탈 예측은 **이익을 극대화하는 방향으로 최적화**되었다.

- 이 논문의 key contribution은 아래와 같다.
  - 첫번째로, 온라인 게임 영역에서 예상 이익을 고려한 첫번째 이탈 예측 사례이다.
  - 타 영역의 비용-이익 분석기법이 온라인 게임 도메인 적용을 위해 수정되었다.
  - 50만 유저의 1년 6개월 분의 게임 내 행동을 분석했고, 게임 내의 사회적 이벤트가 이탈에 중요함을 밝혀냈다. 또한, 인게임 행동의 트렌드와 변동성도 이탈 예측의 중요 요소라는 것을 알 수 있었다.

2. Related Literature

- 데이터마이닝을 활용한 이탈 분석은 다양한 영역에서 활발히 연구되는 주제이다.
  - 이하 생략...

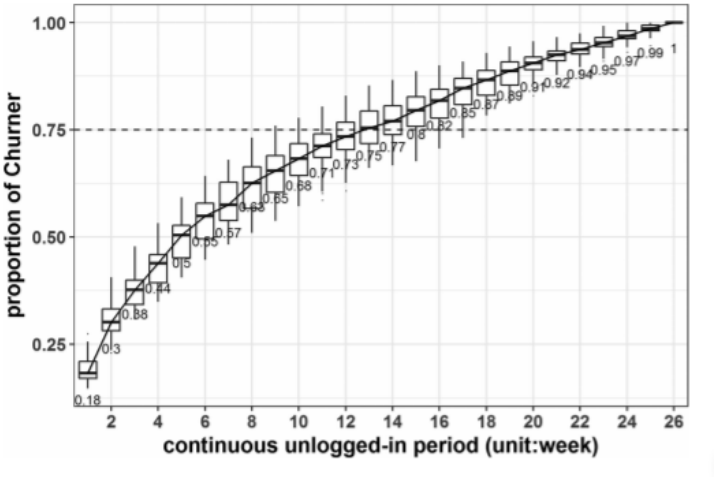
3. Dataset

- 2008년 런칭한 MMORPG인 Aion의 데이터셋을 활용 (2014년 11월 ~ 2017년 7월의 게임 행동 로그)

- A. **데이터 집계 기준**이 되는 시간 단위
  - 다수의 유저가 1주일 단위 플레이 패턴을 보였다.
  - Aion은 매주 수요일에 업데이트가 이루어져, **수요일 ~ 차주 수요일의 주차별 집계**가 모든 통계량의 기준이 되었다.

- B. **이탈의 정의**
  - **'미접속 기간'**을 활용하여 이탈을 정의
    - 만약 해당 기간을 너무 길게 잡는다면, 이탈 방지로 얻는 기대 수익이 감소할 것이다. (이탈을 확인하는 시점이 너무 늦어, 이탈을 막지 못할 것이다.)
    - 반대로, 해당 기간을 너무 짧게 잡는다면, 짧은 휴식 후 돌아온 유저들이 이탈자로 분류되어 의미없는 비용이 발생할 것이다. (이탈하지 않을 유저에게 이탈방지비용 지불)
    - 따라서, 이탈예측의 이익을 극대화하려면 이탈 유저를 정의하는 적정 기간을 정하는 것이 매우 중요하다.
  - Aion에서의 이탈 정의
    - Aion에서는 6개월 단위로 메인 컨텐츠를 업데이트, 이에 반응하지 않을 경우 완전한 이탈로 간주한다.
    - 그러나 6개월 미접속을 기준으로 이탈을 정의한다면 그 기간이 너무 길어지므로, 이러한 예측이 어느 정도 확실시되는 이탈 시점을 찾을 필요가 있다.
    - P_n(Churn) = F(26)/F(n)
      - P_n(Churn) = 0.75가 되는 시점을 elbow로 간주, n=13일 때 해당 기준을 만족함.

- 26주가 아닌 13주 미접속을 이탈의 기준으로 타협.



4. Methodology

- A. 예측 대상 선별
  - 잔존 시 이익을 발생시킬 것으로 기대되는 장기 충성고객을 특정하기 위해, **주차별로 충성 등급을 할당**하였다.
  - 그리고, **주차별 충성등급의 변화를 30주 동안 추적하여 상위 등급을 충분히 오래 유지하는 사용자 그룹을 선택**하였다.

  - 충성 등급 할당
    - 게임 내 다양한 행동(경험치 획득, 아이템 획득, 재화 거래, 컨텐츠 경험, 파티 경험등)을 유저별/주차별로 집계, 9개 클러스터로 구분 (K-means clustering)
    - 각 클러스터의 특성은 각 지표의 평균치를 통해 비교
      - 클러스터 1~5에서는 인게임 행동과 ARPU가 정비례
      - 클러스터 6,7에서는 인게임 행동과 ARPU가 반비례 (Bot으로 간주)
      - 클러스터 9는 Bot detection을 피하기 위한 그룹 어뷰저 (Bot으로 간주)
      - 클러스터 8은 매우 적은 수의 유저, 특정 행동 수가 많음 (아웃라이어로 간주)
    - 이후, 클러스터 1~5의 유저들은 충성등급 1~ 4등급으로 차등 부여, 클러스터 6,7,9의 유저들은 충성등급 5등급(가장 낮은 부정적 등급), 클러스터 8의 유저들은 충성등급 6등급(논외) 부여, 이와 별개로 1주 휴면일 경우 충성등급 9등급을 부여 + 신규일 경우 충성등급 0등급을 부여(첫 1주차)
      - **1등급~ 4등급(정상 유저그룹)까지는 1등급에 가까울수록 인게임 행동도 많고 과금량도 많음**

  - 장기 충성고객 판별
    - 모든 유저는 30주간의 주차별 등급 시퀀스가 존재한다.
    - **장기 충성고객을 추출하기 위해, 시퀀스 클러스터링을 사용했다.**
      - 각 시퀀스 간 거리 : optimal matching algorithm
      - 클러스터링 알고리즘 : hierarchical clustering algorithm
      - 두 알고리즘의 계산 비용을 줄이기 위해, 30주 중 최소 1회 이상 1~2등급을 받은 유저 중 일부만을 사용 (7540개의 시퀀스)
    - 클러스터링 결과, '장기 충성고객'에 해당하는 그룹이 특정되었고, 이를 쉽게 구분하는 필터링 룰을 만들기 위한 분석이 진행되었다.
    - 결과적으로, 유저 일부를 분석하여 만든 룰로 전체 유저 중 장기 충성고객을 추출했다. (논문 상에서도 이는 시스템 리소스 제약을 해결하기 위한 트릭임을 밝힘)

- **실제 룰은 매우 단순하다. (5,6등급을 받은 적이 없으면서 1~3등급이 대부분이며, 최소 10주 이상 플레이한 유저)**

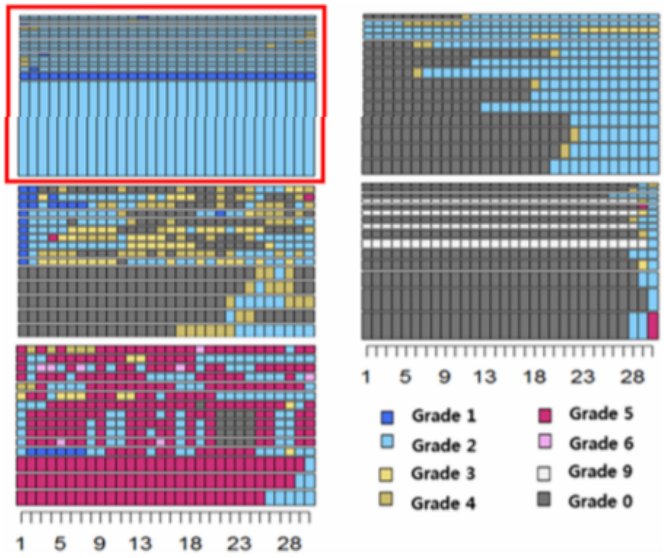


Fig. 2. Categorizing customers via sequence clustering. Users of red box (left top) are identified as long-term loyal customers because they are given a grade 1 or 2 for most periods.

- B. 특성 전처리 (Feature Engineering)
  - 이탈예측모델에 사용된 특성들은 아래와 같다.
  - 1) 사회적 관계 - 친구와 길드(Legion)
    - **친구의 이탈, 길드의 해체와 같은 부정적 사회 경험은 유저의 이탈 확률에 영향을 주었다.**
  - 2) 사회적 관계 - 파티
    - **파티 관련 행동 또한 이탈 확률에 영향을 주었다.**
    - 관련 특성을 추출하기 위해, social network 분석이 진행되었다. 과정은 아래와 같다.
      - 파티 네트워크 생성 (node : 유저, edge : 1주일간 최소 2회 이상 파티 관계)
      - 파티 네트워크를 여러 개의 커뮤니티로 분해 (Modularity 기준)
      - 각 커뮤니티 네트워크에 대해 각 지표 산출
        - Clustering coefficient
        - Graph density
        - Community size
  - 3) 인게임 행동의 트렌드와 변동성
    - **플레이타임 또는 플레이 빈도의 변화 또한 유저의 이탈과 높은 상관관계를 보였다.**
    - 특히, 장기 충성고객은 게임에 꾸준히 접속하고 플레이하는 경향을 보였다.
      - 따라서, 이들은 갑자기 이탈하기보다는 게임에 서서히 흥미를 잃을 것이라는 가설을 세웠다. (**게임 접속 또는 행동 횟수가 단계적으로 감소**할 것으로 가정)
    - 트렌드와 변동성을 통계적으로 측정하기 위한 방법은 아래와 같다.
      - 1) **MACD**(Moving average convergence divergence)
        - 4주 Moving Average와 12주 Moving Average의 차이
        - 이 값이 양수/음수일 경우, 각각 증가/감소 트렌드를 나타냄
      - 2) **Coefficient of variation** (변동계수, 표준편차를 평균으로 나눈 값)
        - 이 값이 높을수록, 해당 유저의 게임 내 행동이 불규칙함을 나타냄
    - 모든 특성들은 정규화되었음

- C. 이익 평가 (Profit Evaluation)

- 잔존 관련 캠페인이 시행될 경우, 기대되는 유저 당 수익은 아래의 도식으로 계산된다.

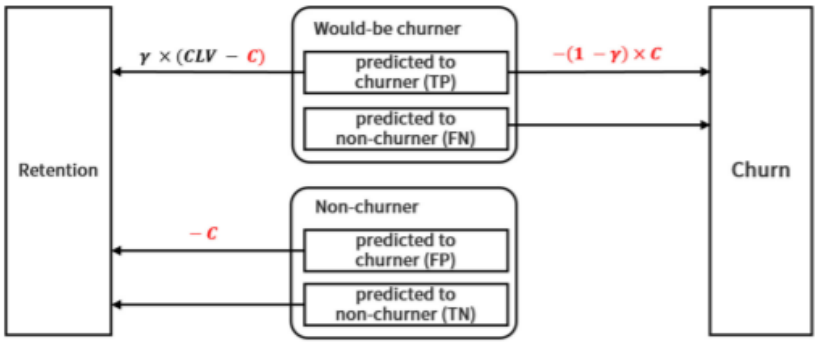


Fig. 3. Expected profit and cost upon application of the churn prediction model.

- 위 도식에서 CLV는 기대되는 유저의 생애가치(Lifetime Value)이다.
- C는 캠페인 비용이다. (C는 보통 잠재적 이탈 유저들의 마음을 돌릴 할인 쿠폰, 보상 아이템의 가치이다.)
- r은 캠페인을 통해 이탈에서 잔존으로 전환되는 유저의 비율이다.
- 분류 모델에서, TP와 FP의 비율은 유저가 이탈했는지 아닌지를 판단하는 데에 사용되는 기준(threshold)에 따라 달라진다.
    - 따라서, TP와 FP 모두 threshold t에 대한 함수로 표현되며, 기대수익 역시 t에 대한 함수로 표현된다.
    - **Profit(t) = CLV{r*TP(t)} - C{TP(t) + FP(t)}**
    - **(돌아온 일부 유저로 인해 얻은 이득에서 캠페인 비용을 제한 값)**

- CLV는 유저의 잔존기간과 단위기간 평균 수익을 곱한 값이다.
    - (유저별 value가 아닌 전체 유저의 평균적인 value)
    - 그러나 유저의 잔존기간을 정확히 측정하기는 어렵기 때문에, 무한급수를 활용하여 계산하게 된다.

$$\mathrm{CLV} = \sum_{i=0}^{\infty} (\mathrm{ARPU} \times r^i) = \frac{\mathrm{ARPU}}{1-r}.$$

    - 위 수식에서 ARPU는 유저당 평균 수익이며, r은 잔존율을 뜻한다.
        - 위 식에서는 단위시간당 ARPU와 r이 상수임을 가정한다.
        - 하지만, 주차별 ARPU와 r은 실제로는 기간에 따라 크게 변동하므로, 안정적인 값을 찾기 위해 아래와 같은 방법을 사용했다.
            - 1주~15주의 window별 ARPU와 r을 계산하고, 각 window별 ARPU와 r들의 분산을 구해, 그 분산이 가장 작아지는 window로 각 값을 계산한다.
            - 이를 통해, 13주 단위의 ARPU와 r을 사용하기로 결정했다.
    - 더 나아가, **개별 유저 각각의 CLV** 또한 계산하였다.
        - 예측모델이 적용되었을 때, 보다 정확한 이익 계산을 하기 위함이다.
        - 이를 위해, ARPU가 아닌 **최근 13주 동안의 개별 유저 과금액**을 이용하였다.
        - 유저 그룹 또한 장기 충성고객 그룹과 그렇지 않은 그룹으로 나누어 잔존율을 계산하였다.
        - 자세한 과정은 6장에 기술되어있다.

5. EXPLORATORY DATA ANALYSIS

- A. 인게임 행동의 트렌드와 변동성
    - 플레이타임 : 이탈자 그룹에서는 플레이타임이 서서히 감소하는 경향을 보임.
    - MACD, CV : 이탈자 그룹의 플레이타임이 감소 추세이며 변동성이 심함.

- B. 파티 활동
    - 이탈자 그룹의 네트워크 크기가 더 크고, 밀도가 낮다.
    - 이탈자들은 랜덤 유저들과 파티를 맺으며, 잔존유저들은 친구들과 파티를 맺는 경향이 있음을 알 수 있다.

- C. 길드(Legion)
    - 이탈자 그룹의 길드 가입률이 더 낮고, 가입되었더라도 활성도가 낮으며, 길드 이동 또한 더 잦다.

6. EXPERIMENTS

- A. 전체 유저 이탈 예측 vs. 충성 고객 이탈 예측
    - 2017년 3월 ~ 6월 사이에 1회 이상 접속한 유저를 대상으로 했다.
        - 그리고 해당 유저의 장기 충성고객 여부, 이탈자 여부를 태깅했다.
        - CLV는 장기 충성고객 그룹과 나머지 그룹을 나누어 계산했다.
    - **장기 충성고객은 전체 유저의 2.4%이며, 그 중 11%만이 이탈했다.**
        - 장기충성고객의 CLV per user가 나머지 그룹에 비해 300배 높아, **한 명의 장기충성고객 유저 1명의 이탈을 막는 것이 그렇지 않은 유저 300명의 이탈을 막는 것과 같은 효과**라고 말할 수 있다.
    - 전체 유저로 이루어진 데이터셋과, 장기 충성고객으로만 이루어진 데이터셋을 따로 생성하여 모델링을 진행했다.
        - 모델 학습에는 Random forest, XGBoost, generalized boosting regression이 사용되었다.
        - 모든 이탈자가 잔존 유저로 전환되었을 때의 기대수익을 극대화하는 것이 목적이다. 그러나, 이를 검증하기 위해 실제 라이브 서비스에 이탈 방지 캠페인을 적용할 수 없었으므로 전환율과 캠페인 비용을 세밀히 조정하여 모델간 차이를 비교했다. (해당 도메인 전문가의 도움으로, 적절한 값을 할당하였다.)
        - 결과적으로, **2개의 데이터셋과 3개의 학습모델을 통해 6개의 예측모델 비교**가 가능해졌다.

- 모델링 결과

			Test set I			Test set II		
			RF	XGB	GBM	RF	XGB	GBM
# of true positive users			63,778	63,249	61,932	212	213	201
# of false positive users			3,917	4,307	4,713	568	615	586
# of true negative users			22,233	21,843	21,437	1,558	1,511	1,540
# of false negative users			10,072	10,601	11,918	62	61	73
accuracy			0.8601	0.8509	0.8337	0.7375	0.7183	0.7254
precision			0.9421	0.9362	0.9293	0.2718	0.2572	0.2554
recall			0.8636	0.8565	0.8386	0.7737	0.7774	0.7336
F1 score			0.9012	0.8946	0.8816	0.4023	0.3866	0.3789
AUC			0.9358	0.9264	0.9067	0.8296	0.8129	0.8159
expected profit	γ:0.1	C:0	704	503	488	5,192	5,169	4,908
		C:0.01	27	-173	-178	5,184	5,161	4,900
		C:0.1	-6,065	-6,253	-6,177	5,114	5,086	4,830
		C:1	-66,991	-67053	-66,157	4,412	4,341	4,121
	γ:0.05	C:0	352	251	244	2,596	2,585	2,454
		C:0.01	-325	-424	-422	2,588	2,576	2,446
		C:0.1	-6,417	-6,504	-6,421	2,518	2,502	2,375
		C:1	-67,343	-67,305	-66,401	1,816	1,757	1,667
	γ:0.01	C:0	70	50	49	519	517	491
		C:0.01	-607	-625	-618	511	509	483
		C:0.1	-6,699	-6,705	-6,616	441	434	412
		C:1	-67,625	-67,506	-66,596	-261	-311	-296

  - **전체 유저를 대상으로 한 모델의 score가 더 좋지만, 기대 이익은 장기 충성고객을 대상으로 한 모델의 경우가 훨씬 높다.**
    - 그 이유는 무엇일까?
    - 전체 유저를 대상으로 한 모델은 CLV가 극히 낮은 유저들의 이탈은 잘 포착하면서도 장기 충성고객의 이탈은 포착하지 못하기 때문이다. (아웃라이어로 인식)
    - 반면, 장기충성고객을 대상으로 한 모델은, 그 정확도는 낮을지라도 그 중의 이탈을 포착하기 때문에 기대 이익 또한 높아지게 된다. (CLV가 높은 유저의 이탈을 포착)
    - 또한 전체 유저를 대상으로 할 때의 캠페인 비용이 훨씬 크므로, 장기 충성고객 대상 모델이 더 높은 이익과 더 낮은 비용(=더 적은 캠페인 대상자 선별)을 가져오게 된다.

- B. 이탈기준 최적화 (Threshold Optimization)
  - 위에서도 밝혔듯이, 높은 CLV를 갖는 유저의 이탈을 최대한 많이 포착하는 것이 가장 중요하다.
  - 따라서, 이탈유저를 잔존유저로 잘못 판단하는 경우를 줄이는 편이 도움이 된다. (그로 인해 잔존유저를 이탈유저로 잘못 판단하는 경우가 늘어나더라도)
  - 이를 위해, 이탈의 기준이 되는 확률값을 낮추는 방향으로 조정했다.
    - 실험 결과, 이탈 기준점을 0.5보다 낮게 조정하는 것이 더 큰 기대이익을 가져왔다.

7. DISCUSSIONS

- 한계와 개선방안을 다룰 것이다.

- A. Binary Classifier의 한계
  - 이탈로 잘못 분류된 잔존유저의 행동을 추적했다.
  - 그 결과, 해당 그룹 또한 명시적 이탈이 아니더라도 등급이 하락하는 등의 부정적 동향을 보였다. (예측시점 이후 플레이타임이 감소하는 경향 또한 확인되었다.)
  - 따라서, 유저의 상태를 보다 세분화하여 multi-class classification을 진행하는 것이 보다 정확한 분석결과를 가져올 수 있다.
  - 또한, 유저의 서비스 생애수명을 regression하는 것도 좋을 것이다.
    - 다만, 잔존 유저가 관측 시점 이후 언제 이탈할지를 알 수 없기 때문에 생애수명을 실제로 라벨링하기는 어려웠다.
    - 이러한 관점에서 생존분석기법을 적용하는 것이 도움이 될 것이다.

- B. 비용 최적화
  - 캠페인 비용 C가 커질수록, 전환률 r 또한 올라갈 것이다.
    - 따라서, r은 C에 대한 logistic function으로 표현될 수 있다.
    - 다수 유저 집단에 대해 A/B test가 가능하다면, 함수의 파라미터가 보다 정밀히 측정될 수 있을 것이다.

8. CONCLUSION

- 온라인 게임 분야에서 기대 수익을 고려한 이탈예측은 이 연구가 최초이다.
- 이 연구의 특징은 크게 3가지로, 아래와 같다.
  - 유저의 접속 패턴을 분석하여 이탈을 정의
  - 높은 수익을 가져다주는 장기 충성 고객을 분류하고, 그를 예측 대상으로 활용
  - 유저 당 기대수익을 비용-이익 분석으로 비교하고, 이를 예측 모델에서 최적화
- 이탈자의 유의미한 특성을 발견하였다.
  - 사회적 행동, 접속시간 및 행동 횟수의 단계적 감소