Purchase Prediction in Free Online Games via Survival Analysis

I. INTRODUCTION

이번 논문에서 주로 참고한 3개의 논문은 다음과 같습니다.

- From Non-paying to Premium: Predicting User Conversion in Video Games with Ensemble Learning(A. Guitart - 2019)
- Understanding Player Engagement and In-Game Purchasing Behavior with Ensemble Learning(A. Guitart - 2019)
- Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles(Á. Periáñez - 2016)

2번째 논문은 과금을 0과 1의 레이블로 표시한 후 분류의 문제로 접근한 방법입니다.

논문의 저자는 이러한 방법들은 "이탈이 언제 발생할 것인가?"와 "특정 시점에서의 이탈 확률"에 대한 설명력이 부족하기 때문에
이를 보완하기 위하여 생존 분석 방법을 이용했다고 말하고 있습니다.

prior work의 간단한 설명을 읽어보면서 가장 흥미가 갔던 부분은 첫 번째 논문이었습니다.
해당 논문은 **non-pu 유저가 pu 유저의 변환을 생존 분석을 통해 예측한 논문**인데,
과금 연구의 핵심 두 가지 중 하나인 pu conversion을 생존 분석을 통해 예측 했다는 것에서 현재 다룰 수 있는 데이터로 적용이 가능하다면 큰 의미가 있다고 생각합니다.

II. SURVIVAL MODELS

생존 분석을 진행하기 위해 저자는 event와 생존 시간을 다음과 같이 정의하였습니다.

- Event: player stops paying for the game
- T: keep paying time

사실 이 부분이 논문을 읽으면서 가장 의문이 들었던 부분인데, 뒤에 나올 cox 모델은 Event를 사망으로 놓고 위험률을 예측해주며 이를 1에서 차감한 확률을 생존 확률이라고 정의하고 있습니다.
이것을 위의 Event에 적용해보면 위험률이란 Event가 일어날 확률인 pu가 non-pu로 전환할 확률을 의미하며 생존 확률은 non-pu가 계속 non-pu일 확률을 의미하게 됩니다.
이러한 정의에 맞춰서 저자는 **pu만을 대상자로 선정하고 분석을 진행**하였는데, Event를 과금으로 두고 생존 시간인 T를 무과금 기간으로 두는 것이 더 직관적이라는 생각이 들었습니다.

저자가 과금에 대한 예측을 위해 사용한 생존 분석 방법은 크게 세 가지 입니다.

1. Kaplan-Meier Estimator

카플란 마이어는 가장 기본적인 생존 분석 방법이며 비모수 방법에 속합니다.
이 방법은 event가 일어난 개체와 그렇지 않은 개체의 시점별 비율의 곱을 활용해서 다음과 같이 생존 함수를 표현하고 있습니다.

$$S(t) = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i})$$

(ti: 최소 한 사람이 과금을 그만 둔 시점
di: ti에 과금을 그만 둔 유저의 수
ni: ti에 과금을 계속하고 있는 유저의 수)

2. Cox Proportional Hazards Model

이 모델은 현재 생존 분석에서 가장 널리 사용되는 모델이며 준모수 방법에 속합니다.
해당 모델을 통해 공변량들에 대한 다양한 해석과 더불어 "위험률"을 예측하며, 이를 위한 hazard function을 다음과 같이 사용합니다.

$$\lambda(t|x_k) = \lambda_o(t) e^{\sum_{i=1}^{p} \beta_i x_{k,i}}$$

(λo(t): baseline hazard function
βi: partial likelihood method로 구해진 covariates)

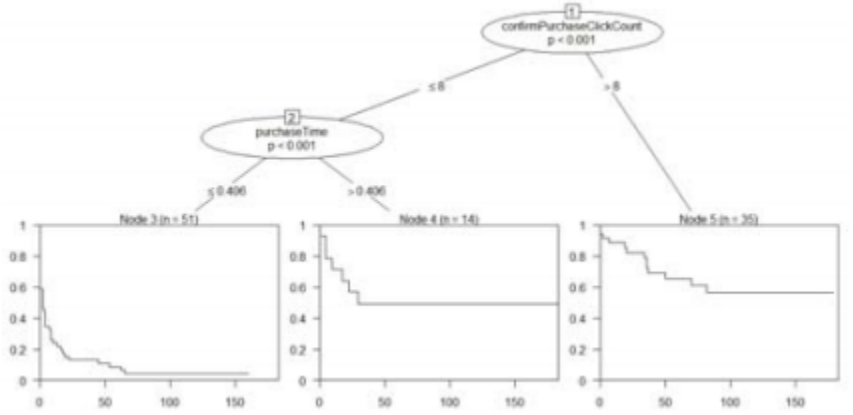
3.Conditional inference Survival Forest

이 모델은 decision tree와 random forest의 로직을 생존 분석에 적용한 모델입니다.

random forest가 단일 decision tree를 base learner로 하여 ensemble한 방법인 것과 동일하게, 단일 survival tree를 base learner로 하여 이것들을 ensemble 한 방법이 Conditional inference Survival Forest입니다.

논문에서는 가지를 뻗어나가는 조건을 kaplan-meier를 이용한다고 언급되어 있고 자세한 방법은 소개하지 않고 있는데, 방법을 생각해 본다면 log-rank test를 이용해서 두 군의 차이가 최대로 되는 변수와 threshold를 선택하여 가지를 뻗어나가는 방법이라고 생각합니다.

논문에서 100개의 subsample을 이용해서 예시를 든 그림은 아래와 같습니다.



III. FREE ONLINE GAME DATA SETS

1.Data selection

본 논문에서는 Yoozoo Games의 Game of Thrones Winter is Coming이라는 PC 게임의 데이터를 활용하여 실험을 진행하였으며, 환경 조성을 위해 몇 개의 제약 조건을 설정하였습니다.

우선, 과금 이탈 유저의 조건을 60일로 설정하였습니다. 그 이유는 과금을 멈춘지 60일 이상이 되는 유저 중 95% 이상이 영영 과금을 하지 않았기 때문이라고 말하고 있습니다.

또한, 데이터는 2019/02/20 ~ 2019/10/28의 기간 중 10,000명을 random sampling을 진행하였다고 합니다.

마지막으로, 구매 확률의 예측은 구매 후 수 일 이내에 충분히 예측이 가능했기 때문에 time window를 과금 후 1주일로 설정하여 실험을 진행했다고 저자는 말하고 있습니다.

2.Feature selection

저자는 몇 개의 prior work를 참고하여 feature의 범주를 과금 유저의 **플레이 행동 특성**과 과금 행동 특성 두 가지로 나누었습니다.

1) Players' play behavior related features:

- Players' time spending.
- Players' login days count.
- Players' in-game mouse click counts.
- Players' role level.

2) Players' purchase behavior related features:

- Players' total payment amount.
- Players' average payment amount.
- Players' purchase frequency.
- Players' remaining soft/hard in-game currency.

-soft in game currency: 게임으로부터 획득 가능한 currency
-hard in game currency: 현금으로 획득 가능한 currency

- Three types of players' mouse click count when processing a purchase.

-구매 과정에서의 클릭
-구매를 확정할 때의 클릭
-구매를 취소할 때의 클릭

- The ratio of mouse click count for confirming the purchase compared to mouse click count for canceling the purchase.

IV. RESULTS AND DISCUSSION

본 논문에서는 두 가지의 생존 모델을 활용을 세 가지의 시나리오를 설정하고 실험을 진행하였습니다.

- Players' Play Behavior
- Players' Purchase Behavior Related Features
- Players' Play & Purchase Behavior Related Features

각 모델의 시나리오 별 성능을 평가하기 위해 Brier score를 사용하였습니다. 해당 score는 낮으면 낮을수록 좋은 성능이라고 해석합니다.

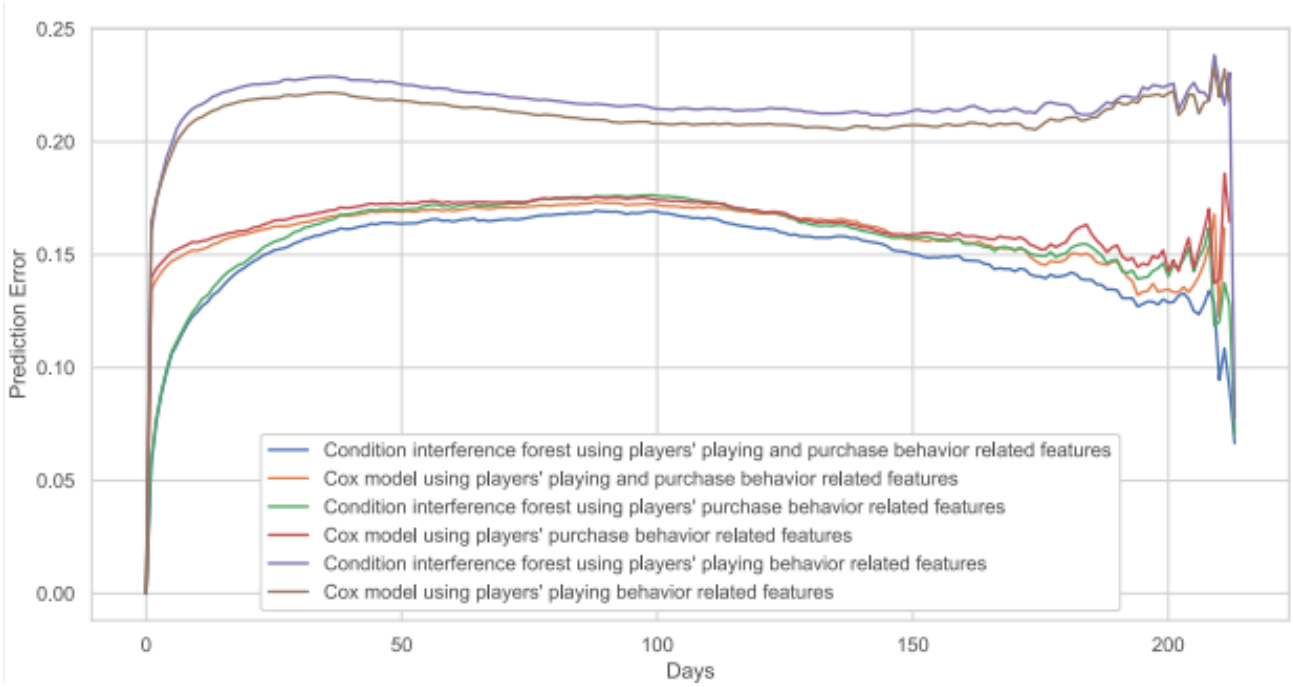
다음은 실험의 결과를 나타낸 표입니다.

TABLE I
INTEGRATED BRIER SCORES OF SURVIVAL MODELS SELECTING DIFFERENT PROPOSED FEATURES

Survival Models	Features		
	Players' Play Behavior Related Features	Players' Purchase Behavior Related Features	Players' Play & Purchase Behavior Related Features
Conditional inference Survival Forest	0.217	0.158	0.150
Cox Proportional Hazards Model	0.211	0.163	0.161

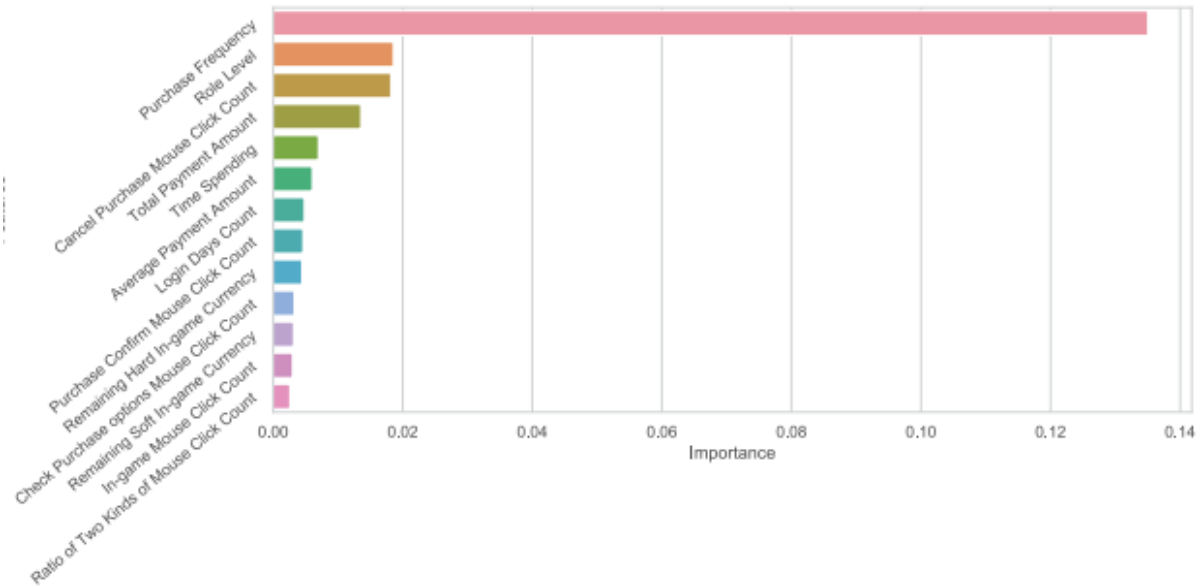
TABLE 1을 보면 플레이와 과금에 대한 feature 모두 사용한 survival forest 모형이 가장 좋은 성능을 보이는 것을 알 수 있습니다.

다음은 10 bootstrap cross-validation을 활용하여 누적 예측 오류를 구한 그래프입니다.



위 그래프에서도 플레이와 과금에 대한 feature 모두 사용한 survival forest 모형이 전체 기간에서 가장 낮은 예측 오류를 보여주며 가장 우수한 성능을 보이는 모델이라는 것을 말해주고 있습니다.

다음은 가장 좋은 성능을 보여줬던 survival forest로 구한 feature importance에 관한 정보입니다.



구매 빈도가 압도적으로 큰 중요도를 갖으며 Role level이 두 번째로 중요한 feature로 선정 되었습니다. prior work에서는 총 과금액 또한 중요한 feature로 선정되었는데 이번 논문에서는 그렇지 못하다는 결과를 보여주고 있습니다.

또한, 구매와 관련된 클릭이 꽤나 중요한 feature이며, 유저의 경제적 상태를 나타내는 soft/hard currency는 중요하지 않은 feature라는 결과를 보여주고 있습니다.

Review

개인적으로 이번 논문은 아쉬운 부분이 많았습니다.
IEEE는 ML/DL 분야에서 꽤나 저명한 저널로, 논문 reviewer들의 accept 조건이 상당히 까다롭다고 알고 있지만, 이번 논문에서는 논리나 방법론에 대한 설명, 데이터와 실험 설계에 대한 설명이 부족하다고 느꼈습니다.

우선, 보통은 statistic 저널 수준까지는 아니더라도 데이터 셋에 대한 전반적인 설명과 수치들을 제시하여 데이터에 대한 타당성을 확보하게 되는데 이 부분에 대한 설명도 찾아볼 수 없었습니다.
그리고 방법론을 선택하게 된 이유와 evaluation metric 선정에 대한 이유를 수식과 함께 설명하며 타당성을 확보하는데 이번 논문에서는 이러한 과정도 생략되어 있었습니다.
또한, 더욱 다양한 환경에서 다양한 metric으로 선택한 방법론들의 우수성을 입증해야 하지만 이러한 부분에서도 미흡하다는 생각이 들었습니다.
마지막으로 생존 분석에서 가장 중요하고 첫 번째로 다루어져야 할 중도 절단에 대한 설명이 없었으며, 통계적 방법론을 선택했음에도 불구하고 결과에 대한 해석이 부족했고 이러한 결과가 나오게 된 원인에 대한 탐구 또한 부족하다는 느낌을 받았습니다.

생존 분석이라는 특수한 분석 방법과 게임 산업의 과금 연구라는 어려움이 더해졌기 때문에, 이러한 연구 자료가 있다는 것 자체도 감사해야 할 일이지만 한편으로는 아쉬움이 많이 남는 논문이었습니다.

추가: 정식 논문이 아니었다고 하네요