이화여자대학교 대학원
2018학년도
석사학위 청구논문

생존 분석을 활용한 모바일 온라인 게임 사용자 이탈 예측

통 계 학 과
김 나 현

2019

생존 분석을 활용한 모바일 온라인 게임 사용자 이탈 예측

이 논문을 석사학위 논문으로 제출함

2019 년 1 월

이화여자대학교 대학원

통 계 학 과 김 나 현

김 나 현 의 석사학위 논문을 인준함

지도교수 송종우 ____

심사위원 송종우 ____

이외숙 ____

차지환 ____

이화여자대학교 대학원

목 차

표 목 차

그 림 목 차

논 문 개 요

고객 이탈 예측은 고객의 잠재된 가치를 발굴함으로써 기업의 수익을 극대화하는 데에 목적이 있어 다양한 분야에서 연구되고 있는 주제이다. 본 논문에서는 NCSOFT사의 모바일 MMORPG 리니지M 유저 이탈에 관한 예측 모형을 제시하고자 생존 분석 모형을 이용하여 분석하였다. 이때 최적 모형을 결정하기 위하여 3가지 측면에서 비교 분석을 실시하였다. 첫째로는 생존 기간을 분석에 사용할 데이터의 기간에 따라 두 종류로 정의하고 이에 따른 예측력의 차이를 보았다. 둘째는 설명변수를 대부분의 MMORPG가 지닌 컨텐츠들로 구성하여 다른 게임에도 적용이 가능한 일반적인 변수만 고려하였는지, 리니지M 고유의 특징을 반영한 설명변수도 함께 고려하였는지에 따라 예측력의 차이를 비교하였다. 마지막으로는 Cox 비례 위험 모형, 랜덤 포레스트-생존 모형 중 어떤 모형이 더 좋은 예측치를 제공하는지에 관해서도 비교하였다. 그 결과 일반적 설명변수와 컨텐츠를 반영한 설명변수를 함께 삽입하여 최근의 자료만으로 구축한 반응변수를 랜덤 포레스트-생존모형을 사용하여 예측하는 것이 가장 좋다는 결론을 얻었다. 본 논문에서 제시한 최종 모형을 사용하면 이탈 위험 유저를 예측할 수 있어, 이로부터 이탈 위험군의 특징을 파악해 사업 운영 방안 개선책 수립에 도움을 줄 수 있다. 또한 모형을 통해 예측된 생존 기간으로부터 고객 생애 가치 또한 추정할 수 있어 마케팅 비용의 산정에도 유리할 것이다.

Ⅰ. 서론

고객 이탈 예측은 다양한 산업에서 중요시되는 분야로, 예측을 통해 이탈 위험 고객들의 잠재된 가치를 발굴함으로써 새로운 고객을 유치하는 것보다 적은 비용으로 수익을 창출하여 기업의 이윤을 극대화 하는 데에 목적을 둔다. 그 중에서도 온라인 게임의 사용자 이탈에 관한 연구들은 이미 수 차례 진행되어 왔다. 강성민 외 1인(2008)은 사용자들이 선호하는 장르에 따라 이탈 성향에 차이가 존재하므로 이를 참고하여 게임을 개발 및 운영해야 한다고 했으며, 손정민 외 2인(2014)은 신규 고객과 기존 고객에 따라 이탈 방지를 위하여 관리해야 할 요인들을 제시하고 이에 대한 실무적 시사점을 제공하였다. 또한 문주연 외 2인(2018)은 온라인 게임 내에서 유저간의 사회 활동이 이탈에 미치는 영향에 관하여 분석하고 이에 대한 예측 모형을 제시하였으며, 김지경 외 1인(2004)은 관계 지속 기간을 종속 변수로 삼아 고객이 게임을 이용하기 시작한 시점으로부터 이탈하는 시점까지의 기간에 영향을 미치는 요인들을 분석한 바 있다.

이에 본고는 선행 연구들의 아이디어를 종합 및 확장하여 NCSOFT사의 모바일 MMORPG (Massively Multiplayer Online Role-Playing Game; 대규모 다중 사용자 온라인 롤플레잉 게임) 리니지M의 게임 로그를 활용하여 이용자 이탈 예측 모형을 제시하고 이탈에 영향을 미치는 요인들을 도출하였다. 분석에는 이탈의 대상이 되는 상품이 MMORPG라는 특성을 고려하여 이용자들의 인구통계학적 요인보다는 게임 플레이 패턴을 활용하였다. 이를 통해 게임 내 협동 과정에서 느낄 수 있는 소속감과 성취감, 소속 집단의 높은 명성으로부터 얻는 자기만족감과 과시, 가상 세계 속에서의 새로운 역할 등을 특징화할 수 있다. 따라서 본 논문에서는 이러한 게임 내 컨텐츠 이용에 관한 정보를 활용하여 최적 이탈 예측 모형을 제시하고자 한다.

최적 모형을 결정하기 위해 고려되는 사항은 크게 세 종류로 나뉜다. 첫째는 반응변수로 생존 기간을 분석에 사용할 데이터의 기간에 따라 두 종류로 정의하고

이에 따른 예측력의 차이를 볼 것이다. 둘째는 설명변수를 대부분의 MMORPG에 적용 가능한 일반적인 변수만 고려한 것인지, 이에 추가적으로 대상 게임의 컨텐츠적 특징을 반영한 변수를 포함시킨 것인지에 따른 예측력의 차이를 보고자 한다. 마지막으로는 어떤 모형을 사용하여야 효과적인 이탈 예측이 가능할 것인가에 관하여 살펴볼 것이다. 구체적으로 분석에 활용된 모형은 랜덤 포레스트-생존(Random Forest-Survival)모형과 Cox 비례 위험 모형(Cox Proportional Hazard Model)이다. 이때, 이탈 시점은 생존 확률이 0.5미만이 되는 날로 정의하였다. 또한 최적 모형을 결정하는 기준은 예측력이며 이에 대한 평가 지표로는 특정 시점을 고정한 후, 해당 시점 이전에 이탈상태인지 잔존상태인지 분류한 오차행렬에서 계산된 F1-Score를 사용하였다.

Ⅱ. 분석 자료 설명

A. 자료의 내용

분석에 사용된 자료는 2017년 6월 21일 출시된 모바일 MMORPG '리니지M'의 게임로그로 NCSOFT 데이터베이스 시스템에서 직접 추출하였다. 게임로그란 유저들의 로그인 시점부터 로그아웃 시점까지의 모든 행동이 기록된 자료로, 현실에서는 정량화하기 힘든 행동들까지 가상 현실 속 정보로 기록되고 있다. 본 논문에서는 리니지M 출시 후 각종 컨텐츠들이 안정화되고 새로운 서버 및 직업을 출시한 시점인 2017년 11월 29일을 시작으로 이후 2주간의 플레이패턴을 관찰하여 각 유저 별 이탈 여부 및 시점과 해당 시점의 이탈 확률을 예측하였다. 이때 분석의 대상은 플레이 패턴 수집 기간의 시작일인 2017년 11월 29일 이전에 계정을 생성하였으면서 관측기간인 2017년 11월 29일부터 2017년 12월 12일 사이에 하루라도 게임에 접속한 이력이 있는 유저로 한정하였다. 단, 불법 프로그램 사용 등으로 인해 게임의 약관에 의거한 제재내역이 존재하는 계정은 분석 대상에서 제외하였다.

B. 변수 선택 과정

2.1 이탈의 정의

온라인 게임의 경우, 다른 산업 분야에서의 이탈 형태인 해지·해약 등과 달리 고객들의 이탈은 계정을 삭제하기보다는 단순히 접속을 하지 않는 방식으로 일어난다. 따라서 얼마나 오랜 기간 동안 연속으로 접속하지 않은 상태를 이탈로 간주할 것인가에 대한 정의가 우선적으로 이루어져야 한다. 본 분석의 대상 게임인

‘리니지M’의 경우, 대다수의 유저들이 매일 접속하는 모바일 게임이라는 특성을 감안하여 14일 연속 미접속 시 해당 계정이 이탈한 것으로 정의하였다. 이때, 14일 이상 미접속을 유지하다 복귀한 유저에 대해서는 이탈하지 않은 것으로 처리하였는데, 이는 온라인 게임에서는 이탈 후 복귀하는 경우와 플레이 상태를 지속적으로 유지하는 경우의 구분이 다소 모호하기 때문이다.

2.2 반응변수

분석의 목적은 유저들의 이탈 시점 및 해당 시점의 이탈 확률을 예측하는 것이며 데이터의 절단 시점은 2018년 1월 9일이다. 그렇다면 반응변수는 시작 시점이 각 유저들의 계정 생성 시점이고 끝 시점이 이탈 시점 혹은 절단 시점인 생존기간이 된다. 이에 따라 본 분석에서 최대 생존 기간은 게임 출시일인 2017년 6월 21일부터 절단 시점인 2018년 1월 9일인 203일이 된다. 이는 유저들이 게임을 시작한 이래로 이탈 혹은 절단되기 전까지 생존한 기간이라는 의미를 지니게 되며 이 반응변수가 정의된 방식에 대한 도식은 아래 Figure 2.1과 같이 나타낼 수 있다. 그런데 이 경우, 위 2.1의 이탈에 대한 정의에 따라 오랜 기간 접속하지 않다가 최근에 다시 접속한 유저의 생존 기간을 처음 계정 생성 시점부터 절단시점이라고 정의하게 되는데 이러한 유저의 미접속 이력을 고려하지 않은 채 정의한 생존 기간이 적절한가에 대한 의문점이 제기될 수 있다. 따라서 이러한 상황을 고려하여 계정 생성 시점을 생존 기간의 시작 시점으로 정의한 반응변수 A와 더불어 시작 시점 또한 절단시킨 반응변수를 생각해볼 수 있다.

이에 플레이패턴 관측 기간인 2017년 11월 29일을 시작 절단 시점으로 설정하여 생존 기간이 최소 1일부터 최대 42일까지인 반응변수 B를 정의하였다. 그러면 이 반응변수의 의미는 관측 시작 이래로 각 유저들이 이탈 혹은 절단되기 전까지 생존한 기간이 되며, 이 반응변수의 정의 방식에 관한 도식은 Figure 2.2에 나타나있다. 그런데 이때, 현재 시점인 2017년 12월 13일이 아닌 관측 시작 시점인 2017년 11월 29일을 시작 절단 시점으로 설정한 이유는 Figure 2.2의 첫 사례와 같이 15일째인 2017년 12월 13일에는 접속하지 않았으나 며칠 전까

지는 접속한 경우를 분석 대상에 포함시키기 위함이다. 다시 말해, 반응변수 정의 시 시작 절단 시점을 현재 시점으로 설정한다면 앞서 언급한 유저들은 현재 시점에 접속하지 않았으므로 생존 기간은 0이 되어 분석 대상에서 제외되기 때문에 이러한 유저들의 14일 이내 향후 재접속 가능성을 고려하기 위하여 관측 시작 시점을 시작 절단 시점으로 설정하였다.

Figure 2.1. Response Variable A

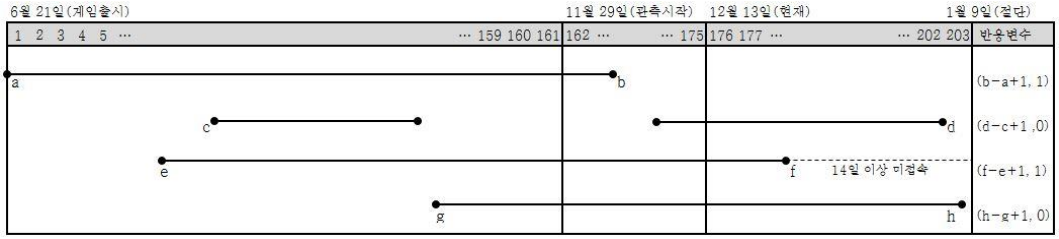
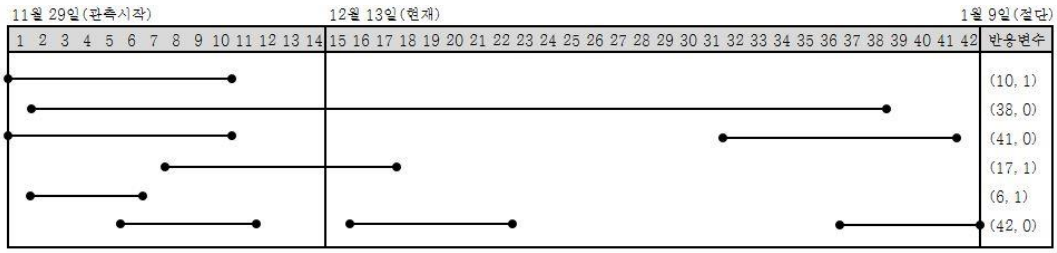


Figure 2.2. Response Variable B

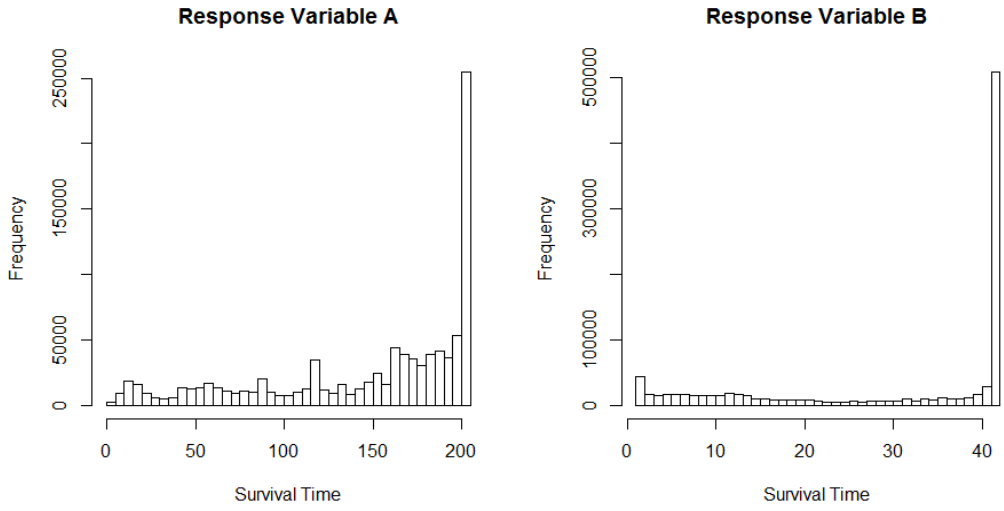


위 Figure 2.1과 Figure 2.2의 반응변수에 앞의 숫자는 생존 기간이고 뒤의 숫자는 이탈일 경우 1, 절단일 경우 0인 절단 여부이다. 쉽게 말해 생존 기간은 시작시점부터 마지막 접속일까지의 기간이며 마지막 접속일이 2017년 12월 27일 이후인 경우에는 가진 자료가 이탈의 기준인 14일보다 적어 이탈여부를 알 수 없으므로 절단된 데이터이다. 이때, 향후 미래 데이터에 적용할 시에는 현재 시점을 플레이 패턴 수집기간이 끝난 직후라 놓으면 된다.

본 논문에서는 시작 시점이 각 유저들의 계정 생성일인 반응변수 A와 시작 절단 시점을 관측 기간의 시작 시점으로 적용시켜 새로이 정의한 반응변수 B 두 종류에 대한 분석을 진행하고 비교할 것이다. 각각의 반응 변수에 대한 히스토그램

은 Figure 2.3과 같으며, 중도 절단된 데이터와 그렇지 않은 데이터의 비율은 각각 66.39%, 33.61%이다.

Figure 2.3. Histogram of Response Variable



2.3 설명변수

예측을 위해 분석에 사용되는 플레이 패턴을 나타내는 설명 변수는 크게 두 가지로 나뉜다. 한 가지는 특정 게임에 국한되지 않고 MMORPG라면 수행할 수 있는 활동들이기 때문에 대부분의 MMORPG의 모델링에 적용 가능한 일반적 변수이고, 또 다른 하나는 대개 해당 컨텐츠를 가진 게임과 가지지 않은 게임이 나뉘거나, 본 논문의 분석 대상 게임인 리니지M만의 컨텐츠가 반영된 변수이다. 본 분석에서는 일반적 변수만을 설명변수로 적합한 모형과 일반적 변수에 컨텐츠 반영 변수를 포함한 전부를 설명변수로 적합한 모형을 각각 비교하여 이탈 예측에 어떤 변수들이 중요한 지에 관해서도 알아보고자 한다. 각 변수들의 값은 아래 설명에서 특별한 언급이 없는 한 플레이 패턴 관측 기간인 2주 간의 총 합으로 계산하였다.

2.3.1. 일반적 변수

1) 혈맹 이력 존재 여부

MMORPG는 단순한 플레이뿐만 아니라 유저들이 모임을 갖거나 친구를 맺는 등의 사회 활동도 게임을 즐기는 데 중요한 요소이다. 특히 모임의 경우 게임에 따라 클랜, 길드 등으로 불리는데, 본 분석의 대상 게임인 리니지M의 경우 이를 '혈맹'이라 칭한다. 하나의 캐릭터는 하나의 혈맹에만 가입할 수 있으며 탈퇴 또한 원할 때에 할 수 있다. 이 변수는 관측 기간 동안 한번이라도 혈맹에 가입상태인 캐릭터로 접속하였을 경우 1, 그렇지 않을 경우 0을 갖는다. 이 변수의 모집단 내 구성비율은 1인 유저가 70.5%, 0인 유저가 29.5%로 분석 대상에 속하는 유저들 중 단 하루도 혈맹 가입 상태가 아니었던 유저들은 30%미만이었다.

2) 총 접속 시간

관측 기간인 2주 동안 게임에 접속해 있었던 누적 시간을 분단위로 집계하였다.

3) 캐릭터 수/서버 수

대부분의 MMORPG는 여러 개의 서버를 가지고 있으며 하나의 계정으로 각 서버마다 여러 개의 캐릭터를 생성할 수 있다. 이 변수들은 관측 기간 동안 한 번이라도 접속했던 캐릭터의 수와 서버의 수를 나타내며 모집단 하에서 각 변수들의 분포는 아래 Table 2.4와 같다.

Table 2.4. Proportion of the Number of Characters and Servers

캐릭터 수	1	2	3	4 이상
비율	41.97%	31.83%	17.58%	8.61%
서버 수	1	2	3	4 이상
비율	87.90%	9.72%	1.75%	0.63%

4) 달성한 출석체크 최고 보상단계

많은 게임들은 지속적인 접속을 장려하기 위하여 매일 접속하여 출석체크를 할 때마다 보상을 제공한다. 주로 오랜 기간 출석을 하면 더 좋은 보상을 제공하는 경우가 많다. 따라서 이 변수는 플레이 패턴 관측기간 동안 달성했던 출석체크의 가장 높은 보상 단계를 나타낸다.

5) UI상점 아데나/다이아 소비량

게임에는 일반적으로 두 종류의 재화가 있는데, 하나는 가상 세계에서의 활동을 통해 얻게 되는 재화이고 다른 하나는 현금으로 구입해야만 얻을 수 있는 재화이다. 리니지M의 경우 전자는 아데나, 후자는 다이아라고 불리며 각각 UI상점에서의 소비량을 집계하였다. 이때, UI상점이란 게임 내 소모품을 주로 판매하는 잡화상이 아닌 메뉴에서 접속 가능한 상점을 일컫는다.

6) 파티 횟수/지속 시간

앞선 1)에서 언급한 모임과 별개로 퀘스트 완료와 같은 빠른 목표 달성의 목적 등을 위해 일시적으로 그룹을 이루어 함께 게임을 하는 것을 파티라 하며, 이는 게임 내 사회 활동의 일환이라고 할 수 있다. 이와 관련한 변수로 각 계정 별 관측 기간 동안의 파티 생성, 신청, 수락 횟수를 모두 더해 파티 횟수를 생성하고 파티 지속시간을 초단위로 집계하여 예측에 고려하였다.

7) 몬스터에게 죽은 횟수

MMORPG에서 가장 흔한 활동은 몬스터를 사냥하여 경험치나 재화, 아이템 등을 획득하는 것이다. 그런데 몬스터를 사냥하다 도리어 몬스터에게 플레이어가 죽게 되는 경우가 있는데, 이러한 횟수를 집계하였다.

8) 친구 신청/수락/삭제 횟수

게임 상에서 유저들 간 친구로 등록 혹은 삭제하기 위하여 수행한 친구 신청 및 수락, 삭제한 횟수를 각각 집계하였다.

9) 혈맹 가입/탈퇴/생성/해체 횟수

앞서 1)에서 언급한 바와 같이 하나의 캐릭터는 하나의 혈맹에만 가입할 수 있으며 원한다면 직접 만들 수도 있다. 다만, 리니지M의 경우 혈맹 생성 및 해체는 특정 직업을 가진 캐릭터만 가능하다는 제약이 있다. 이러한 혈맹의 가입, 탈퇴, 생성, 해체 횟수를 각각 집계하였다.

10) 메인 퀘스트 완료 횟수

퀘스트란 게임 내에서 해결해야 할 일종의 미션으로 꼭 수행하지 않아도 게임을 즐길 수는 있지만 해결한 경우 경험치나 재화 등을 얻기 때문에 대부분의 유

저들은 레벨 상승이나 성취감 등을 위해 이를 수행한다. 이때 서사를 갖고 순차적으로 진행되는 퀘스트를 메인 퀘스트라 하고, 이를 완료한 횟수를 집계하였다.

11) 창고 맡기기/되찾기 횟수

캐릭터는 게임 내에서 자주 필요한 아이템들을 가방의 개념인 인벤토리에 저장하고 모험을 한다. 그런데 이 인벤토리에는 칸, 무게 등의 제한이 있기 때문에 유저들은 굳이 휴대하지 않아도 되는 물건들을 갖고 다니지 않고 창고에 보관하곤 한다. 이 변수들은 창고에 물건을 맡기고 다시 되찾은 횟수를 각각 집계한 것이다.

12) 강화 실패 횟수

캐릭터가 지닌 무기나 착용하고 있는 방어구들을 통틀어 장비라 하는데, 이러한 장비를 더 강하게 만드는 작업을 강화라고 한다. 이러한 강화는 100%의 확률로 성공할 수 있기도 하지만 어느 정도 이상 강한 장비를 더욱 강화시키려 하면, 실패하는 경우도 발생한다. 이처럼 장비를 강화시키려다 실패한 횟수를 집계하였다.

13) 분당 획득 경험치

게임 내에서 경험치를 얻는 방법에는 여러 가지가 있지만 대부분의 경우에는 몬스터를 사냥하여 얻는다. 따라서 분당 획득 경험치가 높다는 것은 사냥 속도가 빠르다 혹은 플레이 효율이 높다고 볼 수 있다. 이를 집계하기 위하여 관측 기간 동안 총 획득 경험치를 2)의 총 접속시간으로 나누어 변수를 생성하였다.

14) 결제 액

결제 액은 관측 기간 동안 실제 결제한 금액으로 앞선 5)에서 언급한 현금으로만 구매 가능한 재화인 다이아 소비량과는 차이가 있다. 결제를 통해 다이아를 구매했다 해도 다이아로 물건을 구매하는 것은 결제 시점이 아닐 수도 있기 때문이다. 이 변수는 관측 기간 동안 해당 유저가 결제한 금액을 원화로 집계하였다.

15) 과거 게임 이용 기간

유저 별 계정 생성 시점 이래로 플레이패턴 관측 시작 전까지의 기간을 말한다.

2.3.2. 컨텐츠 반영 설명변수

16) 전투 유형

리니지M은 플레이어 간 전투인 PvP(Player vs. Player) 전투가 가능할 뿐 아니라 매우 활발한 것이 특징이다. 이러한 PvP 전투에는 여러 유형이 있을 수 있는데, 예를 들면 집단과 집단 간의 싸움일 수도 있고 개인과 개인 간의 싸움일 수도 있다. 특히 집단과 집단 간의 싸움인 경우 앞선 2.3.1의 1)에서 언급한 혈맹 간 세력 다툼을 위한 싸움인 경우가 많다. 또한 개인 간 싸움인 경우 서로 싸움이 붙은 것일 수도 있지만 한 명의 강한 캐릭터가 다수의 약한 캐릭터들을 일방적으로 공격하며 다니는 것 일수도 있다. 이처럼 많은 PvP 전투의 유형들 중 각 유저마다 자주 하는 전투 유형이 다르므로 7개의 유형으로 나누어 카테고리 변수로 삽입하였다. 7개의 각 유형은 혈맹전투자, PvP 공격자, PvP 피해자, 단발성전투자, 동일혈맹전투자, 기타, 전투없음으로 구성되며 모집단에서 유형별 비율은 아래 Table 2.5와 같다. 각 전투 유형에 관한 보다 자세한 설명은 Appendix에 있다.

Table 2.5. Combat Type Proportion

전투 유형	혈맹 전투자	PvP 공격자	PvP 피해자	단발성 전투자	동일혈맹 전투자	기타	전투없음
비율	6.45%	0.54%	7.12%	3.41%	0.34%	1.01%	81.13%

17) 플레이어 공격 횟수/플레이어에게 죽은 횟수/플레이어를 죽인 횟수

PvP 전투에서 플레이어를 공격한 횟수와 플레이어에게 죽은 횟수, 플레이어를 죽인 횟수를 각각 집계하였다.

18) PvP 전투 횟수

플레이어 간 전투인 PvP 전투를 몇 회 하였는지 나타내는 변수이다.

19) 혈맹 출석 횟수

혈맹에 가입한 캐릭터의 경우 매일 혈맹 출석체크를 할 수 있다. 이때, 하나의 계정으로 여러 개의 캐릭터를 생성할 수 있기 때문에 각 캐릭터로 각각의 혈맹에 모두 출석하는 것이 가능하다. 이러한 경우에는 하루에 2번 이상 혈맹 출석체크를 하는 것이 가능하고, 이를 반영하여 중복 집계하였다.

20) 달성한 연속 출석 최고 보상단계

출석체크를 연속으로 특정 횟수 이상 하는 경우 연속 출석 보상을 제공하는데, 관측 기간 동안 달성했던 최고 보상 단계를 변수로 삽입하였다.

21) 자주 사용하는 변신/마법인형 등급

리니지M의 경우 사냥 시에 변신을 하고 일종의 펫 개념인 마법인형을 소환함으로써 일정 시간 동안 더욱 강해질 수 있다. 변신이나 마법인형에는 다양한 종류가 있고 각각은 등급을 지니는데 전설, 영웅, 희귀, 고급, 일반 순으로 강하다고 볼 수 있다. 이 5개의 등급 중에서 가장 자주 사용하는 변신과 마법인형이 각각 어떤 등급인지 집계하였다. 단, 관측 기간에 한 번도 해당 활동을 하지 않았을 경우에는 없음으로 할당하여 총 6개의 카테고리를 가지는 변수로 삽입하였다.

22) 변신/마법인형 횟수

앞선 21)에서 언급한 변신과 마법인형의 사용 횟수를 각각 집계하였다.

23) UI상점 변신&마법인형 뽑기에 소비한 아데나

변신이나 마법인형은 UI상점을 통해 뽑기 형식으로 구매할 수 있고, 뽑았던 종류만 사용 가능하다. 이는 UI상점에서 아데나로 매일 구입할 수 있는데 이때 소비한 아데나량을 집계하였다.

24) 획득 아데나/다이아 성수익

리니지M에는 공성전이라는 컨텐츠가 있는데, 성을 차지하기 위한 혈맹들 간의 싸움이라고 보면 된다. 이 전투에서 우승하여 성을 차지하게 되면 잡화상점과 거래소 판매액의 일부를 세금으로 얻을 수 있는데 이를 성수익이라 한다. 이 변수는 각각 관측 기간 동안 성수익으로 획득한 아데나와 다이아량을 나타낸다.

25) 텔레포트 횟수

게임 내 특정 위치로 순간 이동하는 것을 텔레포트라 하는데, 지도 상에서 원하는 위치 클릭을 통해 텔레포트 한 횟수를 집계하였다.

26) 시련던전 성공/구매 횟수

시련던전이라는 컨텐츠는 6종류이며 매일 1회씩 완료할 수 있는데, 다이아로 횟수를 구매할 경우, 각 1회씩 더 완료할 수 있다. 시도 횟수에는 제한이 없으나,

도전 가능한 시간에는 제한이 있다. 이러한 시련던전의 성공 횟수 및 구매 횟수를 각각 집계하였다.

27) 복수/조롱 횟수

PvP 전투로 인하여 플레이어에게 죽임을 당한 경우, 죽임을 당했던 유저에게 복수나 조롱을 할 수 있는데, 해당 액션을 수행한 횟수를 집계하였다.

28) 혈맹 기부 아데나/다이아 량

자신이 속한 혈맹을 더욱 강하게 만들기 위해 아데나나 다이아를 기부할 수 있는데 관측 기간 동안의 재화별 기부량을 각각 집계하였다.

29) 사망 시 아이템 손실 개수

플레이어나 몬스터에 의해 사망하면 경험치를 손실하게 되는데 이와 더불어 특정 조건 하에서는 자신이 지닌 아이템도 잃게 된다. 이때 손실한 아이템의 개수를 집계하였다.

30) 경험치/아이템 복구 횟수

앞선 15)에서 언급된 사망 시 잃은 경험치와 아이템은 사망 24시간 이내에 복구가 가능한데, 이를 복구한 횟수를 각각 집계하였다.

31) 거래소 정산 다이아

리니지M에서는 거래소를 통해 유저들간 아이템 거래를 할 수 있는데, 거래소에서 사용되는 재화는 다이아이다. 거래소를 통해 내가 지닌 아이템을 판매하여 정산 받은 다이아를 나타내는 변수이다.

32) 용병 가입 횟수

앞선 24)에서 언급한 공성전에 인력이 더 필요하다고 판단되는 경우 혈맹의 우두머리는 자신의 혈맹원은 아니지만 해당 혈맹을 위해 전투를 함께할 용병을 모집할 수 있다. 이때 자신이 용병에 가입했던 혈맹이 성을 차지하면 보상을 얻게 되는데, 이러한 용병에 가입했던 횟수를 집계하였다.

33) 우편함으로 아인하사드 받은 횟수

리니지M에서는 일정 경험치를 획득할 때마다 일정량의 아인하사드를 소모하여 경험치를 많이 얻을 수 있게 되며 이를 모두 소진하면 경험치를 조금밖에 획득

하지 못하게 된다. 이러한 아인하사드는 다양한 방법으로 얻을 수 있는데, 그 중 하나는 매일 3회 특정 시간마다 우편함을 통해 받는 것이다. 이처럼 우편함을 통해 아인하사드를 받은 횟수를 집계하였다.

34) 맵 충전 횟수

몇몇의 사냥터는 머무를 수 있는 시간이 제한되어 있는데, 이를 모두 소진한 경우 특정 아이템을 통해 이용 시간을 충전할 수 있어 이에 대한 충전 횟수를 집계하였다.

35) 일간/주간/월간 퀘스트 완료 횟수

매 주기마다 반복되는 퀘스트로 매일/매주/매월 주어지는 미션을 완료한 횟수를 각각 집계하였다.

사용된 모든 변수의 목록은 Table 2.6과 같다.

Table 2.6. Independent Variables

구분	변수명	형태
일반적 변수	혈맹 이력 존재 여부	Binary
	총 접속 시간 캐릭터 수 서버 수 달성한 출석체크 최고 보상단계 UI상점 아데나 소비량 UI상점 다이아 소비량 파티 횟수 파티 지속 시간 몬스터에게 죽은 횟수 친구 신청 횟수 친구 수락 횟수 친구 삭제 횟수 혈맹 가입 횟수 혈맹 탈퇴 횟수 혈맹 생성 횟수 혈맹 해체 횟수 메인 퀘스트 완료 횟수	Numeric

	창고 맡기기 횟수 창고 되찾기 횟수 강화 실패 횟수 분당 획득 경험치 결제 액 과거 게임 이용 기간	
컨텐츠 반영 변수	전투 유형	Category
	플레이어 공격 횟수 플레이어에게 죽은 횟수 플레이어를 죽인 횟수 PvP 전투 횟수 혈맹 출석 횟수 달성한 연속 출석 최고 보상단계	Numeric
	자주 사용하는 변신 등급 자주 사용하는 마법인형 등급	Category
	변신 횟수 마법인형 소환 횟수 UI상점 변신&마법인형 뽑기에 소비한 아데나 획득 아데나 성수익 획득 다이아 성수익 텔레포트 횟수 시련던전 성공 횟수 시련던전 구매 횟수 복수 횟수 조롱 횟수 혈맹 기부 아데나량 혈맹 기부 다이아량 사망 시 아이템 손실 개수 경험치 복구 횟수 아이템 복구 횟수 거래소 정산 다이아 용병 가입 횟수 우편함으로 아인하사드 받은 횟수 맵 충전 횟수 일간 퀘스트 완료 횟수 주간 퀘스트 완료 횟수 월간 퀘스트 완료 횟수	Numeric

Ⅲ. 분석 과정

3.1. 학습/평가 데이터 분리

본 분석에서 사용된 학습 데이터는 앞선 2장에서 언급한 대상 유저들 중, 7개의 전투 유형에 대하여 각 유형별 1000계정씩 총 7000계정을 랜덤하게 추출하였다. 평가 데이터는 전투 유형과 관계 없이 모집단의 비율에 맞도록 전체 모집단 중에서 랜덤하게 2000계정을 추출하였다.

3.2. 예측력 평가 지표

여러 모형의 비교를 위해서는 성능, 즉 예측력에 대한 평가 지표가 필요한데 본 논문에서는 시점을 플레이 패턴 관측이 종료된 지 2주 후인 2017년 12월 13일로 고정한 후, 해당 시점에서 이탈유저와 잔존유저를 분류한 결과에 대한 F1-Score를 사용하였다. 계산 시 사용되는 오차행렬 및 공식은 아래와 같다.

Table 3.1. Confusion Matrix

예측 값	실제 값	
	이탈 (1)	잔존 (0)
이탈 (1)	True Positive (TP)	False Positive (FP)
잔존 (0)	False Negative (FN)	True Negative (TN)

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F1 - Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

3.3. 반응변수에 따른 비교

앞선 2장에서 언급한 바와 같이 반응변수를 계정 생성 시점부터 측정한 반응변수 A와 관측 시작 시점부터 측정한 반응변수 B인 2가지로 정의하였다. 이 두 반응변수에 대한 모형을 각각 적합하여 이에 따른 예측력을 비교하여 더 적절한 반응변수를 선택할 것이다.

3.4. 설명 변수에 대한 비교

앞선 2장에서 언급한 바와 같이 설명변수를 일반적 변수와 컨텐츠 반영 변수의 2가지 범주로 나누었다. 이에 따라 모형에 설명변수를 일반적 변수만 삽입한 경우와 컨텐츠 반영 변수까지 함께 삽입한 경우 중 어떤 경우에 더욱 예측력이 높은지 비교할 것이다. 더불어 전체 및 각 범주 별 주요 변수를 함께 도출하고자 한다.

3.5. 사용된 모형 및 교차 검증

본 연구에서는 이탈을 죽음으로 간주하여 생존 분석 기법을 적용시켜 예측 모형에 활용하고자 한다. 사용된 모형은 랜덤 포레스트-생존(Random Forest-Survival)모형과 Cox 비례 위험 모형(Cox Proportional Hazard Model)이다. 특히 랜덤 포레스트-생존 모형은 생존 분석 기법과 통계적 학습 알고리즘이 혼합된 모형으로서 범용적으로 사용되는 랜덤 포레스트-분류(Classification) 혹은 회귀(Regression) 모형을 반응변수가 생존 기간과 절단 여부인 생존 분석에 적용시킨 것이다. 본 논문에서는 R 패키지 'randomForestSRC'를 사용하여 해당 랜덤 포레스트-생존분석을 실시하였다. 본 분석에서는 랜덤 포레스트-생존모형과 Cox 비례 위험 모형 중 어떤 이 두 모형 중 어떤 모형을 사용하였을 때 예측력이 더 높은지에 관해서 비교할 것이며 이렇게 선택된 최종 모형을 분류 모형과도 비교하여 최종 모형의 성능을 평가하고자 한다.

이에 따라 본 논문에서 고려될 3가지 측면은 각각 반응변수, 설명변수, 사용된 모형이며, 각각은 2가지씩의 경우의 수를 가지므로 총 8개의 모형에 관한 비교를 진행할 것이다. 이러한 경우의 수를 정리하여 하나의 표로 나타내면 아래 Table 3.2와 같다.

Table 3.2. All Possible Model Cases

		반응변수 A		반응변수 B	
설명변수		일반적 변수	모든 변수	일반적 변수	모든 변수
모형	Random Forest Survival	모형 1	모형 2	모형 3	모형 4
	Cox PH	모형 5	모형 6	모형 7	모형 8

이때 랜덤 포레스트-생존 모형을 사용하는 모형1부터 모형4까지의 경우, 각 모형별로 학습 데이터 내 10겹 교차검정을 30회 반복 실시하여 세부 모수를 조정할 것이다. 또한 Cox 비례 위험 모형인 모형 5부터 모형 8까지의 경우에는 AIC를 기준으로 한 Stepwise를 통해 변수 선택을 할 것이며, 그 후 평가 데이터를 예측할 것이다. 이를 통해 결정된 모형 1부터 모형 8까지 중 가장 평가 데이터 하에서 예측력이 좋은 모형을 최종 모형으로 결정한 다음, 해당 모형의 예측력과 분류 모형의 예측력을 비교할 것이다.

Ⅳ. 분석 결과

A. 교차 검증 결과

앞선 3장에서 언급된 바와 같이 랜덤 포레스트-생존(Survival)모형에 해당하는 모형 1부터 모형 4까지의 경우, 세부 모수 조정을 위하여 학습 데이터 내 10겹 교차검증을 30회씩 실시하였다. 이때 조정의 대상이 되는 세부 모수는 의사결정나무에서 가지를 나눌 때 사용되는 변수의 개수와 의사결정나무의 개수이다. 아래 Table 4.1, Table 4.2는 차례로 모형 1과 2, 모형 3과 4에 해당하는 결과표이며 평가의 기준은 앞선 3장에서 언급한 바와 같이 시점을 플레이 패턴 수집 기간 종료 후 2주가 지난 시점인 2017년 12월 27일로 고정했을 때, 해당 시점 이전 이탈유저와 잔존유저의 분류에 대한 F1-Score를 사용하였다. 단, 학습 데이터 내 10겹 교차 검증이 30회씩 이루어 졌으므로 각 표의 값은 매 시행에서 얻어진 F1-Score의 평균값이다.

Table 4.1. Result of Cross Validation (Model 1 & 2)

사용된 설명변수	나무의 개수	가지를 나눌 때 사용되는 변수의 개수		
		8	14	20
일반적 변수 (모형 1)	100	0.7660	0.7460	0.7423
	300	0.7479	0.7471	0.7449
모든 변수 (모형 2)	100	0.7637	0.7589	0.7574
	300	0.7611	0.7596	0.7579

위 Table 4.1의 결과에 따라 반응변수 A를 사용한 모형 1과 모형 2의 경우, 나무의 개수가 100개이고 가지를 나눌 때 사용되는 랜덤한 후보 변수의 개수가 8개인 랜덤 포레스트-생존 모형을 사용하는 것이 좋을 것으로 보인다.

Table 4.2. Result of Cross Validation (Model 3 & 4)

사용된 설명변수	나무의 개수	가지를 나눌 때 사용되는 변수의 개수		
		8	14	20
일반적 변수 (모형 3)	100	0.7510	0.7459	0.7469
	300	0.7516	0.7481	0.7449
모든 변수 (모형 4)	100	0.7627	0.7606	0.7604
	300	0.7649	0.7619	0.7614

위 Table 4.2의 결과에 따라 반응변수 B를 사용한 모형 3과 모형 4의 경우, 나무의 개수가 300개이고 가지를 나눌 때 사용되는 랜덤한 후보 변수의 개수가 8개인 랜덤 포레스트-생존 모형을 사용하는 것이 좋을 것으로 보인다.

B. 모형의 해석

4.1. 변수 중요도 그림

앞선 A절에서의 교차 검증으로 얻은 모형 1부터 모형 4까지 각각의 최적 세부 모수를 토대로 해당 모형을 적합했을 때 변수 중요도 그림을 그려보면 어떤 변수들이 상대적으로 얼마나 예측에 중요한지에 관하여 알 수 있다. 각 모형 별 상위 10개의 주요 변수에 대한 변수 중요도 그림을 사용된 설명변수에 따라 살펴봄으로써 각 범주별 주요 변수를 찾아내고자 한다. 따라서 일반적 변수만 사용된 모형 1과 모형 3의 변수 중요도 그림을 차례로 Figure 4.3, Figure 4.4에 나타내었다.

Figure 4.3. Variable Importance Plot (Model 1)

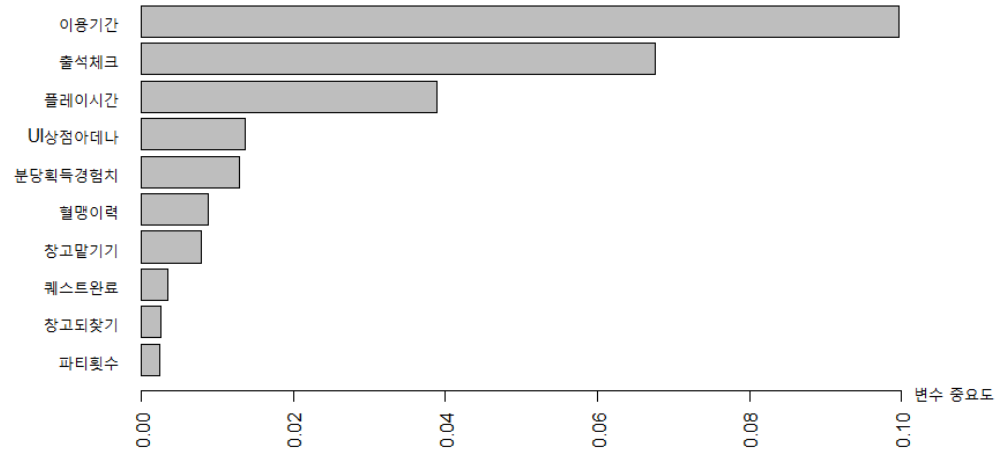
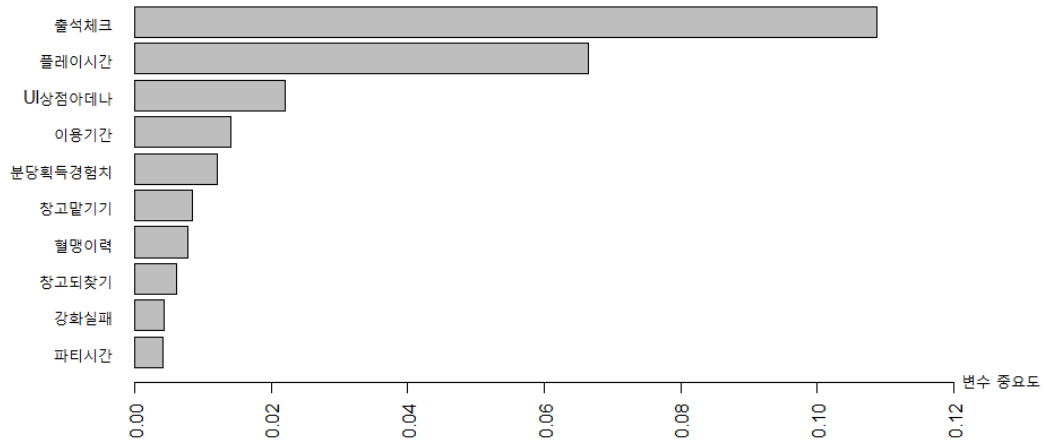


Figure 4.4. Variable Importance Plot (Model 3)



일반적 변수만 삽입한 모형인 모형 1과 3에 해당하는 변수 중요도 그림인 위 Figure 4.3과 Figure 4.4를 살펴보면 공통적인 주요 변수가 달성한 출석체크 최고 보상 단계, 총 플레이 시간, UI상점에서 소비한 아데나량이라는 것을 알 수 있다. 특히 출석체크와 플레이 시간이 다른 변수에 비해 중요한 정도가 크다는 것을 알 수 있다. 이때 모형 1에서 가장 중요하다고 나타난 변수인 계정 생성 후부터 플레이패턴 관측기간까지의 기간을 의미하는 이용기간은 각 유저들의 최소 생존

기간을 나타내는 변수와 같은 작용을 하기 때문에 얻어진 결과라고 볼 수 있다.

마찬가지로 모든 변수를 삽입한 모형 2와 모형 4에 해당하는 변수 중요도 그림을 아래 Figure 4.5와 Figure 4.6에 나타내었다.

Figure 4.5. Variable Importance Plot (Model 2)

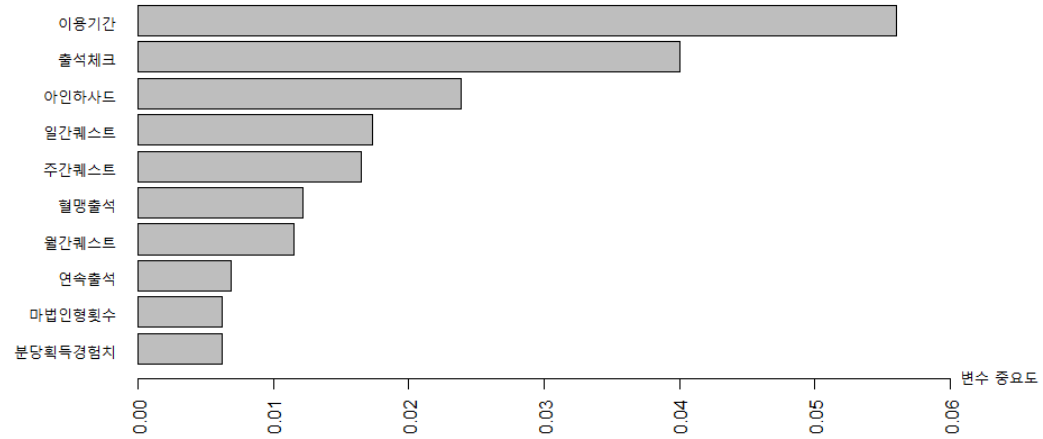
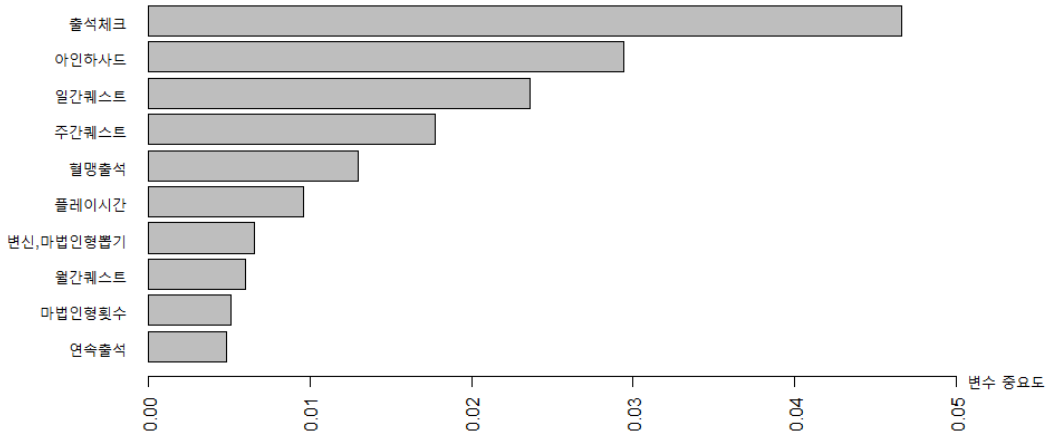


Figure 4.6. Variable Importance Plot (Model 4)



일반적 변수에 컨텐츠 반영 변수를 포함하여 모두 삽입한 모형인 모형 2와 4에 해당하는 변수 중요도 그림인 위 Figure 4.5와 Figure 4.6을 살펴보면 공통적인

주요 변수가 달성한 출석체크 최고 보상 단계, 우편함으로 아인하사드를 받은 횟수, 일간 및 주간 퀘스트 등 반복되는 퀘스트라는 것을 알 수 있다. 특히 위 결과와 마찬가지로 출석체크가 다른 변수에 비해 중요한 정도가 크다는 것을 알 수 있으며, 모형 3에서 가장 중요하다고 나타난 변수인 이용기간은 앞서 모형 1에서 언급한 마찬가지의 이유로 인해 얻어진 결과라고 볼 수 있다. 위와 같은 결과에 따라 설명변수의 성격에 따른 주요 변수를 정리하면 아래 Table 4.7과 같다.

Table 4.7. Important Predictors (Model 1-4)

일반적 변수	컨텐츠 반영 변수
달성한 출석체크 최고 보상 단계 총 접속 시간 UI상점 아데나 소비량	우편함으로 아인하사드 받은 횟수 일간, 주간 퀘스트 등 반복적 퀘스트 완료 횟수

4.2. 부분 의존도 그림

앞선 4.1에서 얻은 주요 변수 중, 각 모형 별 상위 3개의 주요 변수에 대하여 부분 의존도 그림을 그려봄으로써 각 변수들이 생존 확률에 미치는 비선형적인 영향을 알아보고자 한다. 이때 부분 의존도 그림의 세로 축은 예측력 평가 시점인 2017년 12월 27일의 생존 확률을 의미한다. 또한 앞 절에서와 같이 삽입된 설명변수에 따라 각 경우에 따른 주요 변수를 알아보기 위하여 모형 1과 3에 관한 부분 의존도 그림을 먼저 살펴보면 Figure 4.8, Figure 4.9와 같이 나타난다.

Figure 4.8. Partial Dependence Plot (Model 1)

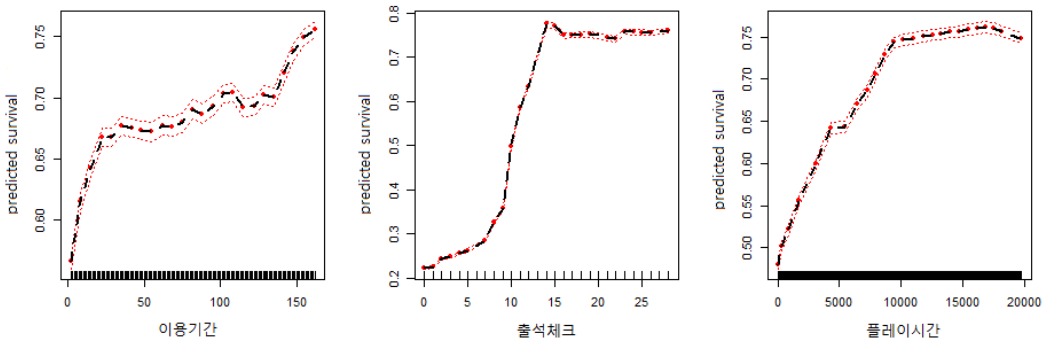
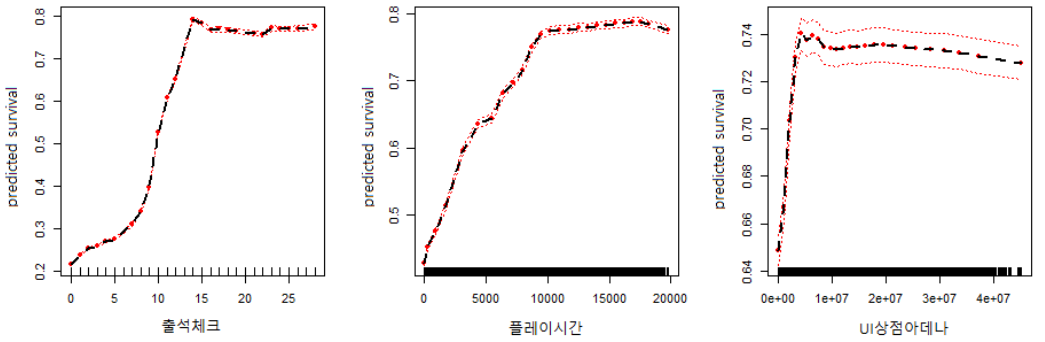


Figure 4.9. Partial Dependence Plot (Model 3)



위의 부분 의존도 그림을 살펴보면 두 모형에서 공통적으로 출석체크 보상 단계가 약 9~14인 구간에서 생존 확률이 급증하며 플레이 시간의 경우 이용 시간이 늘어남에 따라 생존 확률이 증가하지만 약 160시간(9600분, 일 평균 약 11.4시간)이상인 경우에는 생존 확률에 크게 영향이 없다는 것을 관찰할 수 있다.

이용기간에 관해서는 계정을 일찍 생성한 유저의 생존 확률이 높은 경향이 있으며, UI상점에서 소비한 아데나의 경우에는 일정량 소비 시까지는 생존 확률이 급증하나 그 이후부터는 오히려 서서히 감소하는 경향이 있다는 것을 볼 수 있다.

마찬가지로 모든 변수를 삽입한 모형 2와 모형 4에 해당하는 주요 변수 상위 3개에 대한 부분 의존도 그림을 아래 Figure 4.10과 Figure 4.11에 나타내었다.

Figure 4.10. Partial Dependence Plot (Model 2)

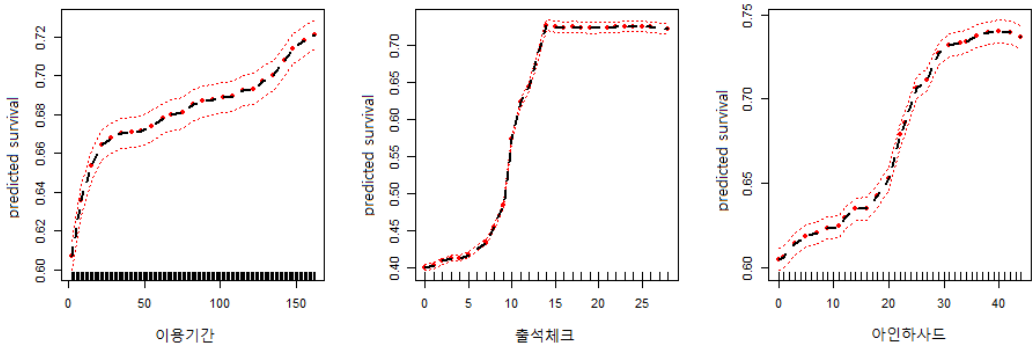
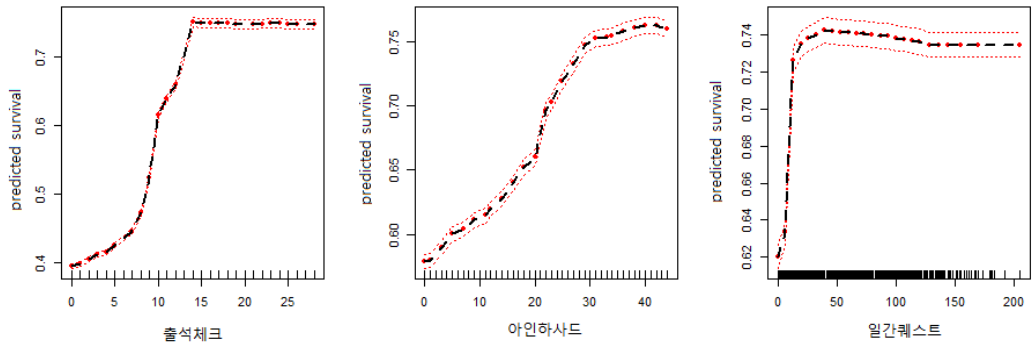


Figure 4.11. Partial Dependence Plot (Model 4)



위의 부분 의존도 그림을 살펴보면 두 모형에서 공통적으로 앞선 모형 1, 모형 3 과 마찬가지로 출석체크 보상 단계가 약 9~14인 구간에서 생존 확률이 급증한다는 것을 볼 수 있다. 더불어 우편함으로 아인하사드를 받은 횟수 또한 수령 횟수가 증가함에 따라 생존 확률이 증가하는 것이 공통적으로 나타난다.

이용 기간의 경우 모형 1과 마찬가지로 계정을 일찍 생성한 유저의 생존 확률이 높은 경향이 있으며, 일간 퀘스트의 경우 약 30회 이상 완료 시 까지는 생존 확률이 급증하나 그 이후로는 거의 변동이 없는 것을 확인할 수 있다.

C. 선택된 변수

Cox 비례 위험 모형을 적용한 모형 5부터 모형 8의 경우, 처음에 삽입된 모든 설명변수를 사용할 것이 아니라 AIC를 기준으로 Stepwise방법을 통해 변수를 선택한 후 축소된 모형을 사용할 것이다. 이에 따라 각 모형 마다 선택된 변수들만을 나열하였으며 진하게 표시된 변수는 유의수준 0.01하에서 유의한 변수이다. 이 절에서도 앞 절과 마찬가지로 삽입된 설명변수의 성격 별로 중요한 변수들을 도출하기 위하여 모형 5와 7에서 선택된 변수들을 Table 4.12, Table 4.13에 각각 나타내었다.

Table 4.12. Selected Predictors (Model 5)

혈맹이력존재여부	**총접속시간**	**캐릭터수**	**달성한출석체크 최고보상단계**
UI상점 아데나소비량	**파티횟수**	**몬스터에게 죽은횟수**	친구수락횟수
친구삭제횟수	친구신청횟수	**혈맹가입횟수**	혈맹생성횟수
메인퀘스트 완료횟수	**창고되찾기횟수**	**강화실패횟수**	**분당획득경험치**
결제액	**과거게임이용기간**		

Table 4.13. Selected Predictors (Model 7)

혈맹이력존재여부	**총접속시간**	**캐릭터수**	서버수
달성한출석체크 최고보상단계	**UI상점 아데나소비량**	**파티횟수**	**몬스터에게 죽은횟수**
친구삭제횟수	친구신청횟수	혈맹가입횟수	혈맹생성횟수
혈맹탈퇴횟수	**메인퀘스트 완료횟수**	**강화실패횟수**	**분당획득경험치**
결제액	**과거게임이용기간**		

위 표를 살펴보면 일반적 설명 변수만을 모형에 삽입한 다음 변수 선택을 하면 공통적으로 혈맹 이력 존재 여부, 총 접속 시간, 캐릭터 수, 달성한 출석체크 최고 보상 단계, 파티 횟수, 몬스터에게 죽은 횟수, 메인 퀘스트 완료 횟수, 강화 실패 횟수, 분당 획득 경험치, 결제 액, 과거 게임 이용기간이 공통적으로 선택된 유의수준 0.01하에서 유의한 변수임을 알 수 있다.

마찬가지로 일반적 변수에 컨텐츠 반영 변수를 포함시켜 적합한 다음 변수 선택을 한 모형 6과 모형 8에 대하여 변수 선택 결과를 나타내면 아래 Table 4.14, Table 4.15와 같다.

Table 4.14. Selected Predictors (Model 6)

혈맹이력존재여부	**총접속시간**	캐릭터수	서버수
달성한출석체크 최고보상단계	UI상점 아데나소비량	친구수락횟수	친구삭제횟수
친구신청횟수	혈맹가입횟수	혈맹탈퇴횟수	혈맹생성횟수
메인퀘스트 완료횟수	**창고되찾기횟수**	**강화실패횟수**	**분당획득경험치**
결제액	**과거게임이용기간**	**전투유형**	**플레이어공격횟수**
달성한연속출석 최고보상단계	자주사용하는 변신등급	**변신횟수**	**마법인형소환횟수**
UI상점변신& 마법인형뽑기에 소비한아데나	**텔레포트횟수**	**시련던전성공횟수**	시련던전구매횟수
혈맹기부아데나량	혈맹기부다이아량	사망시아이템 손실개수	**경험치복구횟수**
아이템복구횟수	**거래소정산다이아**	용병가입횟수	**우편함으로아인 하사드받은횟수**
일간퀘스트 완료횟수	**주간퀘스트 완료횟수**		

Table 4.15. Selected Predictors (Model 8)

혈맹이력존재여부	**총접속시간**	캐릭터수	달성한출석체크 최고보상단계
친구삭제횟수	친구신청횟수	혈맹가입횟수	혈맹탈퇴횟수
혈맹생성횟수	혈맹해체횟수	**메인퀘스트 완료횟수**	**창고되찾기횟수**
강화실패횟수	**분당경험치획득량**	**결제액**	**과거게임이용기간**
전투유형	**플레이어공격횟수**	플레이어에게 죽은횟수	**달성한연속출석 최고보상단계**
자주사용하는 마법인형등급	자주사용하는 변신등급	**변신횟수**	**마법인형소환횟수**
UI상점변신& 마법인형뽑기에 소비한아데나	획득다이아성수익	**텔레포트횟수**	**시련던전성공횟수**
시련던전구매횟수	**혈맹기부아데나량**	혈맹기부다이아량	**경험치복구횟수**
아이템복구횟수	**거래소정산다이아**	용병가입횟수	**우편함으로아인하사드받은횟수**
일간퀘스트 완료횟수	**주간퀘스트 완료횟수**		

위 표를 살펴보면 총 접속시간, 메인퀘스트 완료 횟수, 창고 되찾기 횟수, 강화 실패 횟수, 분당 획득 경험치, 결제 액, 과거 게임 이용 기간, 전투 유형, 플레이어 공격 횟수, 달성한 연속 출석 최고 보상 단계, 변신 횟수, 마법인형 소환 횟수, 시련던전 성공 횟수, 혈맹 기부 아데나량, 경험치 복구 횟수, 거래소 정산 다이아, 우편함으로 아인하사드 받은 횟수, 일간퀘스트 완료 횟수, 주간 퀘스트 완료 횟수가 공통적으로 선택된 유의수준 0.01하에서 유의한 변수임을 알 수 있다.

그렇다면 4개의 모형 모두에서 선택된 변수이면서 유의수준 0.01하에서 유의한 변수는 일반적 변수 중에서도 중요한 변수라고 볼 수 있다. 이에 해당하는 변수는 총 접속시간, 메인퀘스트 완료 횟수, 강화 실패 횟수, 분당 획득 경험치, 결제 액,

과거 게임 이용기간이다. 특히 총 접속 시간의 경우 앞선 B절의 랜덤 포레스트-생존 모형에서도 주요 변수로 판별된 점으로 미루어 보아 이탈 예측에 매우 중요한 변수라는 것을 알 수 있다. 이 절에서 공통적으로 선택되었으면서 유의수준 0.01하에서 유의한 변수를 정리하면 아래 Table 4.16과 같다.

Table 4.16. Important Predictors (Model 5-8)

	일반적 변수만 삽입	모든 변수 삽입
선택된 변수	총접속시간 메인퀘스트완료횟수 강화실패횟수 분당획득경험치 결제액 과거게임이용기간	
	혈맹이력존재여부 캐릭터수 달성한출석체크최고보상단계 파티횟수 몬스터에게죽은횟수 창고되찾기횟수	전투유형 플레이어공격횟수 달성한연속출석최고보상단계 변신횟수 마법인형소환횟수 시련던전성공횟수 혈맹기부아데나량 경험치복구횟수 거래소정산다이아 우편함으로아인하사드받은횟수 일간퀘스트완료횟수 주간퀘스트완료횟수

D. 예측 모형 평가

앞선 절에서 모형 1부터 모형 8까지 각 모형 하에서 예측력을 최대로 할 수 있도록 교차 검증 및 변수 선택을 실시하고 모형의 주요 변수를 살펴보았다. 이에 따라 얻은 결과를 바탕으로 모형 1에서 모형 4까지의 경우 교차 검증을 통해 정해진 세부 모수를, 모형 5에서 모형 8까지의 경우 변수 선택 후 AIC를 최소로 하도록 축소된 모형을 사용하여 예측 성능을 평가하였고 그 결과를 아래 Table 4.17과 같이 정리하였다.

Table 4.17. Model Performance Comparison

	랜덤 포레스트-생존				Cox 비례 위험 모형			
	반응변수 A		반응변수 B		반응변수 A		반응변수 B	
	일반적	모두	일반적	모두	일반적	모두	일반적	모두
	모형 1	모형 2	모형 3	모형 4	모형 5	모형 6	모형 7	모형 8
TP	548	568	540	**570**	585	579	526	539
FP	246	242	218	**244**	564	537	240	212
FN	122	102	130	**100**	85	91	144	131
TN	1084	1088	1112	**1086**	766	793	1090	1118
Precision	0.6902	0.7012	0.7124	**0.7002**	0.5091	0.5188	0.6867	0.7177
Recall	0.8179	0.8478	0.8060	**0.8507**	0.8731	0.8642	0.7851	0.8045
Accuracy	0.816	0.828	0.826	**0.828**	0.6755	0.686	0.808	0.8285
F1-Score	0.7486	0.7676	0.7563	**0.7682**	0.6432	0.6384	0.7362	0.7586

위와 같은 결과표를 살펴보면, 전체적으로 잘못 분류된 것들 중 실제로는 잔존이나 이탈로 예측된 False Positive가 실제로는 이탈이나 잔존으로 예측된 False Negative보다 꽤 많은 것을 볼 수 있다. 이로 인해 정확도(Precision)는 떨어졌으나 재현율(Recall)이 높아졌으며, 이러한 결과의 공정한 비교를 위해 조화평균인 F1-Score를 평가의 기준으로 사용하였다. 따라서 F1-Score를 기준으로 하면 모형 4, 즉 반응변수 B를 컨텐츠를 반영한 변수까지 모두 삽입한 랜덤 포레스트-생존 모형으로 예측하는 것이 가장 좋은 예측치를 얻을 수 있는 방법이라는 결론을 얻고 이를 최종 모형으로 결정하였다.

E. 분류 모형과의 비교

앞 절에서는 생존 분석을 사용한 모형들에 관해서 다루었으며 그 결과를 바탕으로 최종 모형이 결정되었다. 이를 분류 모형과 비교해 보고자 최종 모형인 모형 4에 활용된 랜덤 포레스트-생존 모형과 가장 유사한 분류 모형인 랜덤 포레스트-분류 모형을 사용하여 동일한 분석을 진행하였다. 이때 사용된 R 패키지는 randomForest이며, 사용된 반응변수는 이탈 여부로 앞 절에서 예측력 평가 시점이었던 2017년 12월 27일을 기준으로 이탈유저와 잔존유저를 각각 1, 0으로 할당한 값이다. 참고로 가진 데이터가 2018년 1월 9일에서 절단된 데이터이므로 이 데이터의 이탈 여부의 값이 구조상 생존 분석 시의 절단 여부와 동일하다. 나무의 개수와 가지를 나눌 때 고려되는 변수의 개수는 앞 절에서 결정된 최종 모형의 값과 동일하게 각각 8과 300을 사용하였다. 평가 데이터 하에서의 결과를 앞 절에서 결정한 최종 모형의 결과와 함께 정리하면 아래 Table 4.18과 같다.

Table 4.18. Performance Comparison: Survival vs. Classification

	Classification		Survival
	일반적변수	일반적+컨텐츠변수	최종 모형 (모형 4)
TP	520	522	570
FP	193	169	244
FN	150	148	100
TN	1137	1161	1086
Precision	0.7293	0.7554	0.7002
Recall	0.7761	0.7791	0.8507
Accuracy	0.8285	0.8415	0.828
F1-Score	0.7520	0.7671	0.7682

위 표를 살펴보면 분류 모형과 생존 분석 모형의 예측력 차이가 거의 없으며 이 경우 생존 모형의 예측력이 더 좋다. 따라서 한정적인 정보만을 제공하는 분류 모형보다 시점에 따른 확률을 제공해주는 생존 모형을 사용하는 것이 더 효율적이라고 볼 수 있다.

F. 최종 모형의 개선

앞 절에서 결정된 최종 모형은 반응변수 B와 모든 변수를 사용하여 랜덤 포레스트-생존으로 예측한 모형이다. 그런데 이 경우 설명변수를 굳이 모두 사용해야 하는 지에 관한 의문이 제기될 수 있다. 따라서 앞 절에서 도출된 주요 변수들만 사용하여 모형을 적합하여 예측력이 상승하거나, 비슷하다면 효율성을 위하여 삽입되는 변수를 줄이는 것이 좋을 것이다. 이에 따라 최종 모형인 모형 4에 사용된 반응변수 B와 랜덤 포레스트-생존 모형을 사용하되, 사용된 모든 변수들 중 주요 변수로 판별된 일부의 변수만 사용하여 모형을 적합하였다. 이때 사용된 변수의 목록은 Table 4.19와 같으며 변수 선별 기준은 Table 4.7과 Table 4.16에 있는 변수들이다.

Table 4.19. Independent Variables for Reduced Model

구분	변수명	형태
일반적 변수	혈맹 이력 존재 여부	Binary
	총 접속 시간 캐릭터 수 달성한 출석체크 최고 보상단계 UI상점 아데나 소비량 파티 횟수 몬스터에게 죽은 횟수 메인 퀘스트 완료 횟수 창고 되찾기 횟수 강화 실패 횟수 분당 획득 경험치 결제 액 과거 게임 이용 기간	Numeric
컨텐츠 반영 변수	전투 유형	Category
	플레이어 공격 횟수 달성한 연속 출석 최고 보상단계 변신 횟수 마법인형 소환 횟수	Numeric

	시련던전 성공 횟수 혈맹 기부 아데나량 경험치 복구 횟수 거래소 정산 다이아 우편함으로 아인하사드 받은 횟수 일간 퀘스트 완료 횟수 주간 퀘스트 완료 횟수	

반응변수 B에 대하여 위 Table 4.19의 변수들만을 사용해 랜덤 포레스트-생존 모형에 대하여 가지를 나눌 때 사용되는 랜덤한 변수의 개수가 8개이고 나무의 개수가 300개인 모형을 적합하였다. 이는 최종 모형인 모형 4에 사용된 것과 설명변수의 구성만 다른 모형이며 이에 대한 평가 데이터 하에서 예측력을 모형 4의 예측력과 비교하여 정리하면 아래 Table 4.20과 같다.

Table 4.20. Performance Comparison: Final vs. Reduced Model

	최종 모형 (모형 4)	주요 변수로만 구성된 모형
TP	570	541
FP	244	197
FN	100	129
TN	1086	1133
Precision	0.7002	0.7331
Recall	0.8507	0.8075
Accuracy	0.828	0.837
F1-Score	0.7682	0.7685

위 Table 4.20을 살펴보면 변수를 선택한 경우, 이 평가 데이터 하에서는 변수를 선택하지 않은 모형인 모형 4보다 예측력이 조금 더 우수한 것을 알 수 있다. 비록 그 차이가 크지는 않지만 설명변수의 개수가 55개에서 25개로 줄어들었다는 점에서 강점을 갖는데, 그 이유는 게임 로그로부터 이탈 예측 분석에 사용할 데이터를 구축하는 데에 시간과 비용이 많이 들기 때문이다. 따라서 향후 관측 기간 및 예측 시점을 달리하여 이탈 예측을 할 때, 본 논문에서 제시하는 주요 변수로만 구성된 모형을 사용하면 보다 효율적인 분석이 가능할 것이다.

Ⅴ. 결론

본 논문에서는 3가지 사항에 대한 비교를 통해 최적 이탈 예측 모형을 제시한다. 첫째는 반응변수의 정의에 따른 비교, 둘째는 사용된 설명 변수의 성격에 따른 비교, 마지막으로는 사용된 모형에 따른 비교이다.

유저들의 생존 기간을 계정 생성시점부터 마지막 접속일로 정의한 반응변수 A와 플레이 패턴 관측 시점부터 마지막 접속일로 정의한 반응변수 B를 비교한 결과, 반응변수 B로 적합한 모형의 예측력이 더 좋다는 결과를 얻었다.

설명변수로 고려된 변수들의 특성은 두 종류로 나뉜다. 대부분의 MMORPG가 지닌 컨텐츠들로 구성된 일반적 변수들과 리니지M의 특징을 반영하여 이를 바탕으로 만들어진 리니지M 컨텐츠 반영 변수들이다. 본 연구에서는 일반적 변수들로만 적합한 모형과 일반적 변수들에 리니지M의 컨텐츠를 반영한 변수를 함께 삽입한 모형에 대한 비교를 통해 이탈에 영향을 미치는 주요 요인들에 대하여 알아보았다. 대표적으로 예측에 중요했던 변수로는 일반적 변수의 경우 출석체크 보상 단계, 컨텐츠 반영 변수로는 우편함으로 아인하사드를 받은 횟수가 있었다. 또한 예측력을 높이기 위해서는 일반적인 변수만을 사용하는 것보다는 분석 대상 게임의 자체적인 특성이 반영된 변수들을 고려하는 것이 이탈 예측에 도움이 된다는 결론을 얻었다. 뿐만 아니라, 앞선 분석에서 얻어진 주요 변수들 만으로 모형을 새로 적합하면 예측력을 유지하면서 효율성을 높일 수 있다는 결론을 얻었다. 따라서 보다 효율적인 모형의 적합을 추구하는 경우 굳이 전체 변수를 모두 사용하기 보다는 주요 변수만을 선택한 모형을 사용하여도 좋을 것으로 보인다.

모형에 관해서는 생존 분석 기법을 사용한 모형인 랜덤 포레스트-생존과 Cox 비례 위험 모형이 고려되었으며 생존 분석과 통계적 학습 기법이 혼합된 랜덤 포레스트-생존 모형이 최종 모형으로 선택되었다. 이 최종 모형과 가장 유사한 분류 모형인 랜덤 포레스트-분류 모형과의 비교를 통해 본 논문에서 제시하는 최적 모형의 장점이 확연히 드러난다. 평가 데이터 하에서 F1-Score를 기준으로 비교했을 때 예측력에 대한 차이는 크게 존재하지 않지만 시점에 따른 이탈 확률을 제

공해주기 때문에 모형이 주는 정보량이 더욱 풍부하다는 점에서 본 논문에서 제시하는 최종모형인 랜덤 포레스트-생존 모형을 사용하는 것이 더욱 효과적이라는 것을 알 수 있다.

본 논문에서 제시하는 예측 모형을 사용하면 원하는 시점에 유저 별 이탈 확률을 알 수 있어 이로부터 이탈 위험 유저를 파악할 수 있다. 따라서 이러한 결과로부터 예측된 이탈 위험군에 대한 사후 분석을 실시한다면 이들의 특징을 파악할 수 있고 이로부터 사업 운영 방안을 개선하는 데에도 도움을 줄 수 있을 것이다.

뿐만 아니라 이로부터 추정된 각 유저들의 생존 기간을 통해 각각의 유저가 생애 동안 얼마만큼의 이익을 가져다 주는지 나타내는 지표인 고객 생애 가치(Customer Lifetime Value)를 추정할 수도 있으며 이는 마케팅 비용 산정에도 활용될 수 있다.

본 논문에서 제시한 최종 모형은 생존 분석과 통계적 학습 기법이 혼합된 모형으로서 높은 예측력과 풍부한 결과를 제공해준다. 하지만 기존의 생존 분석과는 달리 게임 분야에서는 생존 분석에 앞서 죽음을 정의해야 하며, 복귀가 가능하다는 특수성을 고려해야 하기 때문에 결과 해석에 유의해야 할 것이다.

이러한 측면에서 본 연구는 휴면 상태에서 복귀한 유저, 즉 2주 이상 미접속 이후 다시 접속하여 복귀한 유저를 이탈로 간주하지 않고 분석을 실시하였다. 그러나 이러한 유저들의 생존 기간을 정의할 때에 장기 휴면 기간 이후 재접속 시점을 계정을 만든 시점으로 간주하는 방식 등 다양한 이탈에 관한 정의가 가능할 것이고, 그에 따라 예측력에 영향을 미칠 수도 있다. 덧붙여 이러한 유저들에 대해 집중적으로 분석하여 복귀유저들만의 특징을 살펴보거나 복귀에 관한 예측 모델링 또한 수립하는 것 또한 좋은 연구 주제가 될 것이다.

참 고 문 헌

강성민, 김태준 (2008). 온라인 게임에서 장르에 따른 사용자 이탈성향에 관한 연구. Entrue Journal of Information Technology, 7(1), 51-62.

손정민, 조우용, 최정혜 (2014). 온라인 게임의 고객 유형별 이탈 요인: 신규 고객과 기존 고객을 중심으로. 한국경영과학회지, 39(4), 115-136.

문주연, 김휘강, 우지영 (2018). 온라인 게임 내 유저간 상호작용 분석을 통한 유저 이탈 예측. 정보과학회 컴퓨팅의 실제 논문지, 24(9), 433-441.

김지경, 김상훈 (2004). 온라인 게임 서비스 이용 고객의 관계지속기간에 영향을 미치는 요인에 관한 연구. 마케팅연구, 19(1), 131-158.

Cox D. (1975). Partial Likelihood. Biometrika, 62(2), 269-276.

H. Akaike (1974). A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control, 19(6), 716-723.

Kitty J., Paul C., Carmine Z., Friedo W., (2008). The Analysis of Survival Data: the Kaplan-Meier Method. Kidney International, 74(5), 560-565.

John F. (2008). Cox Proportional-Hazard Regression for Survival Data.

Hemant I., Udaya B., Eugene H., Michael S. (2008). Random Survival Forest. The Annals of Applied Statistics, 2(3), 841-860.

Leo B. (2001). Random Forests. Machine Learning, 45, 5-32.

Cyril G., Eric G. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. European Conference on Information Retrieval, 345-359.

Africa P., Alain S., Anna G., Colin M. (2016). Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles. IEEE International Conference on Data Science and Advanced Analytics.

Torsten H., Kurt H., Achim Z. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework, Journal of Computational and Graphical Statistics, 15(3), 651-674.

Choi D., Kim J. (2004). Why People Continue to Play Online Games: In Search of Critical Design Factors to Increase Customer Loyalty to Online Contents. Cyber Psychology & Behavior, 7(1), 11-24.

Kim S., Choi D., Lee E., Rhee W. (2017). Churn Prediction of Mobile and Online Casual Games using Play Log Data. PLoS ONE, 12(7), e0180735.

https://cran.r-project.org/web/packages/survival/survival.pdf

https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf

https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

< 부 록 >

전투 유형은 유저들의 전투 행태에 따라 매일 7개 유형 중 하나로 할당된다. 본 연구에서는 관측기간 동안 유저 별 가장 많이 분류되었던 유형으로 전투 유형을 할당하였다. 동점 시에는 숫자가 작은 유형을 우선 순위로 할당하였다.

1. 혈맹 전투자

맹 전투는 집단과 집단 간의 싸움이다. 주로 혈맹 간 세력 다툼 등을 이유로 발생하며 혈맹 간의 전투가 특정 횟수 이상이어야 혈맹 전투자로 분류된다.

2. PvP 공격자

대개 자신의 캐릭터가 강하다는 것을 과시하기 위하여 자신과 크게 관련이 없는 약한 유저 여러 명을 공격하고 다니는 캐릭터를 말한다. 이러한 유형으로 분류되기 위해서는 혈맹 전투를 제외하고 공격한 캐릭터 수가 특정 값 이상이어야 하며, 상대방을 공격한 횟수가 상대방에게 공격 당한 횟수에 비해 높아야 한다.

3. PvP 피해자

위 2에서 서술한 PvP 공격자에게 주로 일방적으로 공격 당하여 피해를 입는 캐릭터를 말한다. 이러한 유형으로 분류되기 위해서는 개인 간의 싸움에서 주로 먼저 공격을 당한 횟수는 많지만, 상대방을 공격한 횟수는 거의 없어야 한다.

4. 단발성 전투자

위 2와 3에서 분류되는 일방적 공격에 의한 전투가 아닌 단발적으로 일어나는 전투를 주로 하는 유형을 나타낸다. 즉, 2의 PvP 공격자처럼 많은 유저들과 전투를 하지는 않으나 개인 간의 전투가 일어난 경우 단발성 전투자로 분류된다.

5. 동일혈맹 전투자

혈맹 간 싸움이 아닌 동일 혈맹 내 전투를 많이 한 경우 이 유형으로 분류된다.

6. 기타

위 5가지 유형에 해당하지 않는 전투를 하는 경우 기타 유형으로 분류된다.

7. 전투 없음

PvP 전투가 발생하지 않은 경우 전투 없음으로 분류된다.

ABSTRACT

Kim Nahyun

Department of Statistics

The Graduate School

Ewha Womans University

Churn prediction is a topic that is being explored in various fields with the aim of maximizing corporate profits by discovering the potential value of customers. In this paper, we use the survival analysis to propose optimal churn prediction model for NCSOFT's mobile MMORPG *Lineage M*. In order to determine the optimal model, we conducted three comparative analyses. First, we defined two types of the survival time according to the period of data used for analysis. Second, we built models with different composition of independent variables. The one is a model considering only generic variables that can be applied to other MMORPGs. The other is a model which also includes specific independent variables reflecting characteristics of *Lineage M* contents. Finally, we also compared the models with Cox proportional hazard model and random forest-survival. As a result, we concluded that the random forest-survival model considering all independent variables is the best for the response variable made by only with the recent data. The best model proposed in this paper helps predict users with high probability to churn, and identify the characteristics of these people. By extending this study, the customer lifetime value can be estimated more effectively from the predicted survival time, which contributes to marketing strategy and business operation plan.