[NDC 18] 우리 유저들은 언제 게임을 떠날까? '이탈 예측'을 위한 팁과 노하우

유저들이 게임을 그만둘 확률을 알아볼 수 있을까? 나아가 마음이 떠날 것 같은 유저들을 미리 알아내 이들의 마음을 돌릴 수 있을까?

유저들의 행동 패턴을 분석해 게임에서 이탈할 확률, 이탈하는 시기 등을 예측하는 '이탈 예측'. 이탈 예측은 게임 데이터 분야에선 비교적 생소한 연구 분야다.

하지만 이런 생소함과 별개로, 제대로 된 '이탈 예측'의 가치는 어마어마하다. 게임사 입장에선 새로운 유저를 유치하는 것보다, 기존 유저를 잘 케어해 게임에 남기는 것이 훨씬 더 쉽기 때문이다. 또한 유저가 이탈할 시기와 이탈 원인을 알면, 이 데이터를 바탕으로 보다 나은 콘텐츠·서비스를 제공할 수도 있다.

그렇다면 이탈 예측을 잘 하려면 어떻게 해야 할까? 엔씨소프트 데이터분석모델팀 '장윤제' 강연자가 NDC 2018에서 말한 경험과 노하우를 정리했다.

이탈이 뭘까? 이탈 예측의 첫 걸음 '학습 데이터 생성'

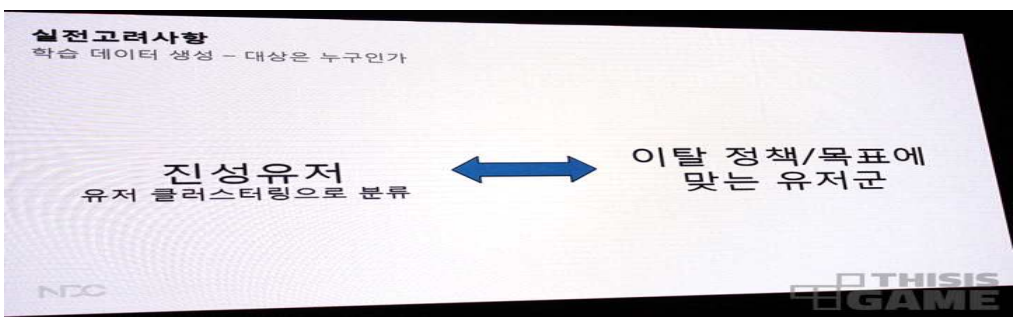
이탈 예측은 크게 ▲ 학습 데이터 생성 ▲ 데이터 가공 ▲ 모델 생성 ▲ 성능 평가 4가지 단계로 진행된다.

첫 과정인 '학습 데이터 생성'은 쉽게 말해 이탈 예측에서 체크할 '이탈'이 무엇인지, 체크하고자 하는 이탈자가 누구인지 확실히 정의하는 과정이다. 이 과정을 소홀히 할 경우, 나머지 과정이 제대로 돼도 의미 없는 결과값을 얻기 십상이다. 자신이 어떤 것을 알려고 하는지 제대로 정하지 못한 채 분석을 한 것이니까.

학습 데이터 생성은 크게 ▲ 학습 대상 ▲ 대상에 대한 정보 ▲ 레이블 3개 요소로 구성된다. 이 중 이번 강연에서 구체적으로 설명된 부분은 학습 대상과 레이블 2개 요소였다.

학습 대상은 쉽게 말해 개발자가 어떤 유저층을 대상으로 '이탈 예측'을 할 지 정하는 단계다. 이 단계에서 유저층은 좁고 상세할수록 의미 있는 결과를 얻는다. 참고로, 가장 효율적인 학습 대상은 게임에 대한 애정이 증명된 '진성 유저'다.

반대로 '모든 유저' 같이 방대한 계층을 목표로 하면 정확한 결과가 나오기도 힘들고, 나온 결과를 제대로 써먹기도 힘들다. 예를 들어 개발자가 '모든 유저'를 학습 대상으로 했을 경우, 그 안에는 BOT이나 악성 유저, 개발자의 게임과 맞지 않는 유저들도 포함된다. 이들은 이탈을 방지하기도 쉽지 않고, 방지한다고 해서 긍정적인 효과를 얻기도 힘들다. 때문에 개발자는 자사의 이탈 정책과 목표에 맞는 유저를 학습 대상으로 선정해야 한다.

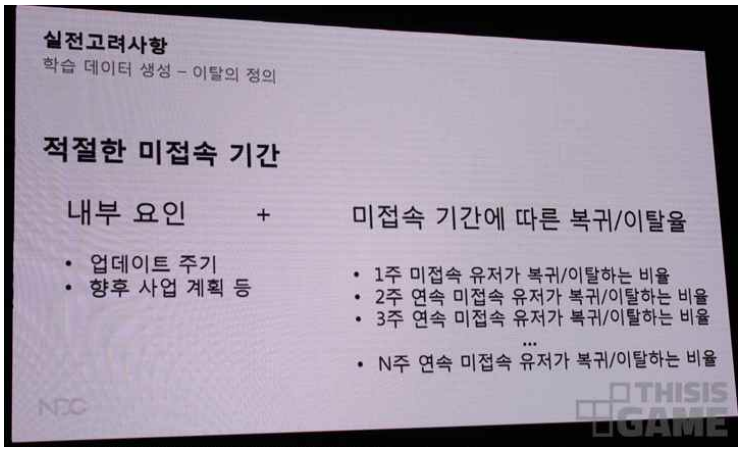


레이블은 유저의 이탈, 생존을 정의하는 기준이다. 의외로 많은 개발자가 '이탈'을 탈퇴로 생각한다. 하지만 계정을 유지하는데 특별한 비용이 들지 않는 게임의 경우, 이탈자 중 탈퇴 유저의 비율은 극소수다. 실제로 모 게임의 1년 이상 미접속 유저 중 서비스 탈퇴 유저는 0.8%에 불과했다.

그렇다면 이탈을 정의하기 가장 좋은 기준은 뭘까? 장윤제 강연자는 '연속 미접속 기간'을 꼽았다. 단, 이 연속 미접속 기간도 어떻게 설정하느냐에 따라 예측의 결과와 신뢰도가 달라진다. 극단적인 예를 들면, 이틀 연속 미접속을 이탈로 정의하면 불금에 너무 달린 탓에 토요일에 게임을 못한 유저도 졸지에 이탈자가 돼 버린다. 반대로 1년 연속 미접속을 이탈로 정의하면 유저가 게임을 떠난지 몇 달이 지난 뒤에야 이를 파악하게 된다.

게임의 장르와 플랫폼, 회사의 정책도 연속 미접속 기간을 정의하는데 큰 영향을 끼친다. 예를 들어 엔씨소프트의 <아이온>은 최대 연속 미접속 기간을 26주로 잡았다. 이는 약 6개월에 1번 대규모 업데이트를 하는 온라인 MMORPG의 특성을 반영한 설정이다. 6개월마다 대규모 업데이트로 게임이 달라지니, 이탈 유저를 파악하고 붙잡을 시간도 이에 맞춰 26주로 설정한 것.

레이블을 설정할 때 가장 중요한 것은 충분히 긴 시간을 들여 데이터를 살펴보는 것이다. 특히 게임의 경우, 요일이나 공휴일, 명절 등 특정 시기에 따라 값이 크게 변하기 때문에 긴 관점에서 데이터를 본 후 레이블을 설정해야 한다.



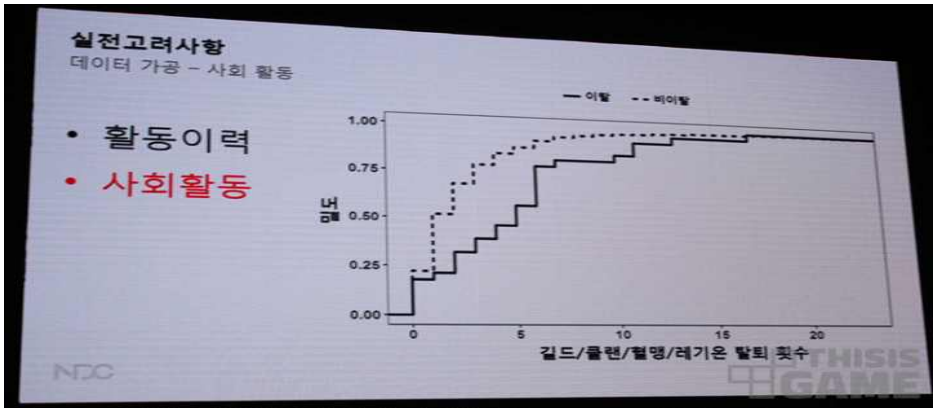
어떤 데이터를 봐야 이탈을 예측할 수 있지? 데이터 가공

데이터 가공은 학습 데이터 생성 단계에서 잡은 목표를 예측하기 위해 사용할 '데이터'를 꼽고 가공하는 단계다.

이전 단계가 애초에 기준 자체가 없어 처음부터 기준을 설정하는 것이 힘들었다면, 이번 단계는 사용할 수 있는 데이터가 너무 많아 이 중 의미 있는 데이터를 찾는 것이 과제다. 특히 게임 내 데이터는 상당수가 이탈 예측에 도움이 되기에 더더욱 까다롭다.

예를 들어 유저의 '활동 이력'을 데이터로 꼽는다면, 개발자는 유저가 게임에서 이탈할 시기가 가까워 질수록 게임 내 활동 이력이 줄어드는 것을 확인할 수 있다. '사회활동' 데이터는 이탈 시점이 가까워질수록 길드의 가입·탈퇴가 활발해지는 경향을 보여준다. 유저의 결제 이력도 분석하기에 따라서 이탈 예측에 큰 도움이 된다.

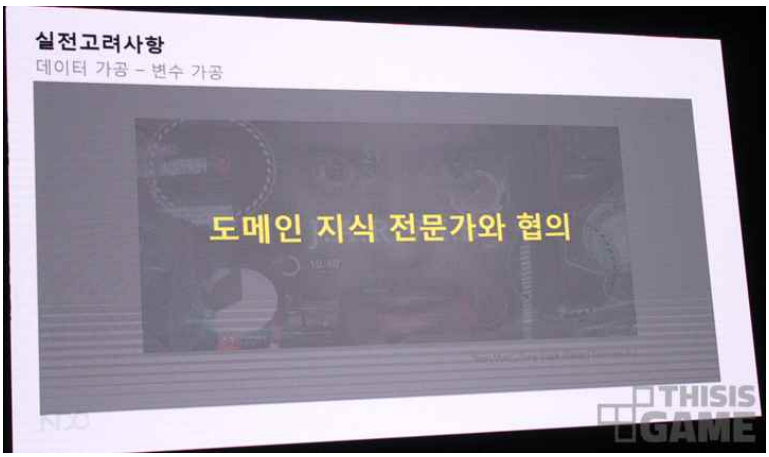
이처럼 게임 데이터는 하나 하나가 유저의 행동 경향을 보여주고 각기 다른 의미를 담고 있다. 때문에 개발자가 섣불리 다른 성질의 데이터를 함께 쓸 경우 오히려 예측에 악영향을 줄 수도 있다.



그렇다면 이탈 예측에 좋은 데이터는 어떻게 찾아야 할까. 장윤제 강연자의 조언은 간단하다. "도메인 지식 전문가와 협의하라."

도메인 지식 전문가는 쉽게 말해 매일 게임을 모니터링하고 유저들의 피드백을 수렴하는 게임 담당자다. 이들은 평소 계속 유저들의 행동을 분석하고 피드백을 체크하기 때문에, 데이터를 분석하는 사람들보다 훨씬 더 데이터의 의미와 중요성을 잘 알고 있다.

데이터를 선정하고 가공하는 것은 전문가에게 맡기라는 조언이다.



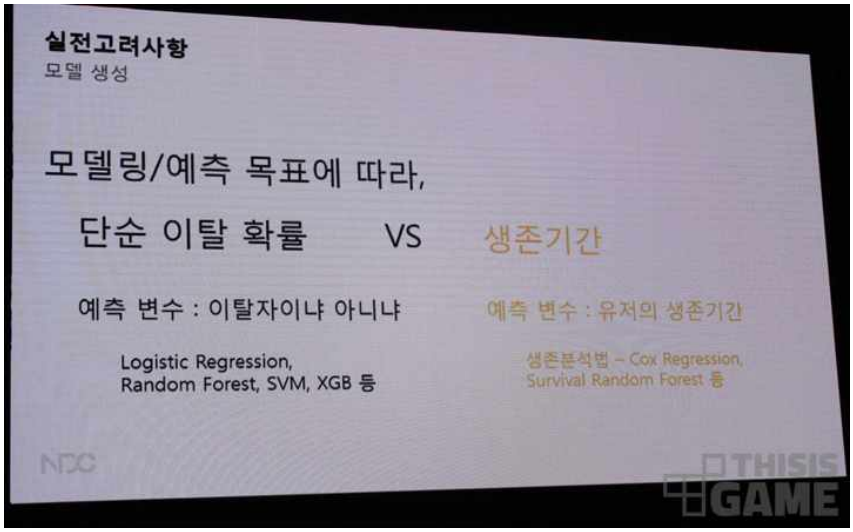
예측을 잘해도 해석을 못하면 무용지물. 모델 생성과 성능 평가

모델 생성은 앞서 설정한 정의와 데이터를 바탕으로 실제로 예측값을 낼 공식(?)을 만들고 값을 내는 과정이다.

모델의 가치는 크게 예측력과 해석력으로 나뉜다. 예측력은 말 그대로 모델로 구한 값이 실제로 얼마나 정확한지를 나타내는 기준이다. 예측력이 낮은 모델은 최초 목적인 '이탈률' 자체를 제대로 구할 수 없다.

해석력은 모델이 값을 도출했을 때, 이 값의 원인을 잘 도출할 수 있는가를 나타내는 기준이다. 해석력이 낮은 모델은 설사 예측력이 높다고 할지라도, 이탈 원인을 도출하기 힘들기 때문에 개발자가 실제로 문제를 해결하는 데 도움이 되지 않는다. 때문에 개발자는 모델을 만들 때 예측력과 해석력이 모두 높은 모델을 만들어야 한다.

모델을 정할 때 또 하나 고려해야 할 것이 예측 목표다. 예를 들어 단순 이탈 확률을 구하는 것, 그리고 유저가 얼마나 오래 게임을 할지 구하는 것은 예측 변수도 다르고 사용하는 알고리즘도 다르다. (참고로 장윤제 강연자는 두 모델 중 생존기간 기반 모델링을 더 추천한다)



마지막 단계인 '성능 평가'는 말 그대로 정확도와 같은 모델의 성능을 평가하는 단계다. 이 단계 또한 절대적인 기준은 없다. 예를 들어 같은 모델이라도 개발자가 확실히 이탈을 예측할 수 있는 유저만 찾기를 원하느냐, 아니면 최대한 많은 유저를 잠재 이탈자로 간주해 이들에게 마케팅을 하길 원하느냐에 따라 평가가 완전히 달라진다.

때문에 모델의 성능을 평가할 때는 개발자가 최초 원한 목적이 무엇인지, 그리고 해당 모델이 이를 얼마나 잘 알려줬는지를 따져야 한다. 어렵지 않은 일 같지만, 보통 이탈 예측 모델은 애매한 결과를 내놓는 경우도 많기 때문에(모델은 내일 이탈한다고 예측했는데 3일 뒤 유저가 이탈하면 이건 맞춘 걸까, 맞추지 못한 걸까?) 생각 이상으로 까다로운 작업이기도 하다.