베테랑 플레이어들의 이탈을 예측하다

(Predicting Churn: When Do Veterans Quit?)

작성자: 드미트리 노즈닌(Dmitry Nozhnin)

작성일: 2012.8.30

러시아의 MMO 배급사 이노바(Innova)에서 통화 재정과 분석을 맡고 있는 드미트리 노즈닌(Dmitry Nozhnin)이 플레이어 이탈을 예측하는 지난 기사[1] 에 이어 베테랑 플레이어들이 게임을 그만두는 시기를 예측하는 방법론을 보여준다. 플레이어들이 게임을 그만두는 때를 이삼 주 전에 알아내는데 그 정확성이 95퍼센트에 달한다. 모두 NCSoft의 <아이온(Aion)> 러시아 버전의 라이브 환경에서 진행되었다.

지난 기사[2]에서 나는 막 게임에 등록한 신입 유저들의 이탈(churn)을 예견하기 위해 우리가 개발한 프로세스를 보여주었다. 이는 유저들이 게임을 훑어보는 첫 이삼 일의 자료를 토대로 했었다. 한편 스펙트럼 저쪽 끝에는 몇 달씩 게임을 하다가 여러 이유로 게임을 접는 노련한 게이머들이 있다. 이 글에서 우리는 게임을 그만두려는 마음을 예측하는 것이 가능하다는 것과 함께, 데이터 마이닝 방법론을 공유하고자 한다.

---

[1]참조링크: http://gamasutra.com/view/feature/170472/predicting_churn_datamining_your_.php

[2]참조링크: 위와 동일

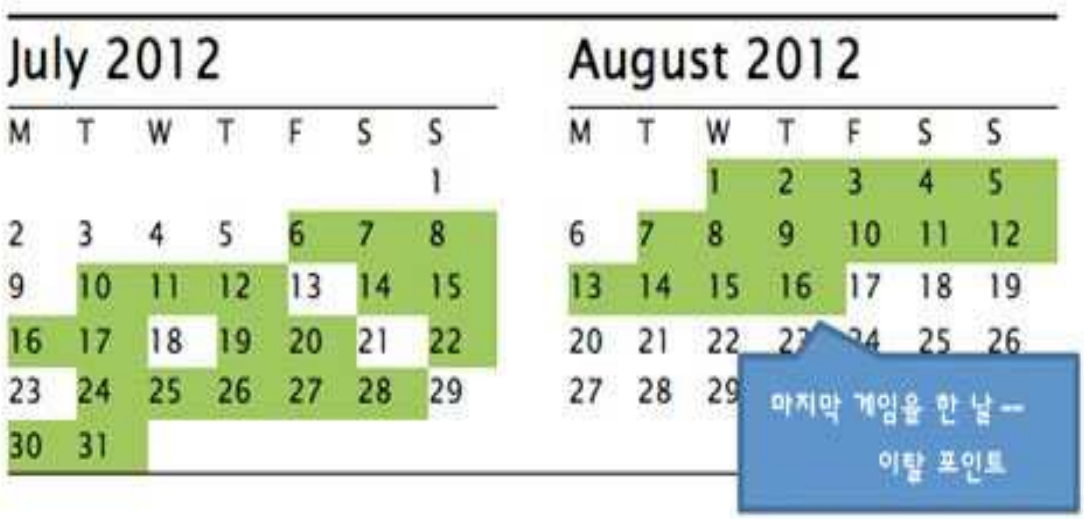
기술적인 면

첫 데이터 마이닝 프로젝트로부터 바뀐 것은 아무것도 없었다. 우리는 여전히 32기가 램 듀얼 제온 E5630 블레이드 두 개, 10 TB 콜드 스토리지와 3TB 핫 스토리지 RAID10 SAS 유닛을 돌렸다. 두 블레이드 모두 MS SQL 2008R2로, 하나는 데이터 웨어하우스로 쓰고, 다른 하나는 MS Analysis Services용이었다. 표준 MS BI 소프트웨어 스택만 사용했다.

우리 데이터셋에는 3만 8000여 명에 달하는 베테랑 플레이어들의 게임플레이를 최대 육 개월까지 보관했다.

이탈 시점을 규정하다

새로운 플레이어들의 이탈을 규정하는 것은 아주 단순하다. 그들은 몇 분 혹은 몇 시간 뒤 게임을 떠난다. 그게 전부다. 게임을 접는 날이 명확하게 규정되었고 그런 이탈 요인의 데이터 마이닝 모델은 이미 잘 구축되어 있었다. 그러나 베테랑의 경우에는 분명히 게임에서 떠났다고 규정하려면 여러 번 반복해야 했다. 우리의 첫 추정은 이랬다. 한 플레이어가 게임을 한동안 즐기고 있다. 하지만 그러다 그만두고 떠나기로 결심한다. 이 플레이어가 게임을 한 날을 녹색으로 표시하면서 우리는 이렇게 예상했다.



이탈 포인트를 규정한 우리의 추측은 간단했다. 마지막으로 게임을 한 날. 그러나, 현실은 더 복잡했다. 대다수의 플레이어들은 이렇게 행동했다.

우리가 마지막으로 그 플레이어를 본 게, 즉 8월 25일이 이탈 포인트인가? 혹은 사실상 7일을 연속으로 보지 못한 8월 16일인가? 아니면 처음으로 7일보다 더 오래 게임에 접속하지 않았던 7월 31일인가? 우리는 여러 가설을 세웠고, 단순한 가정은 잘 들어맞지 않았다. 단순한 방법으로 이탈을 규정하는 것은, 말하자면 특정 날이 마지막일 것이라고 예측하는 것은 정확도가 65퍼센트밖에 되지 않았다.

수작업으로 데이터를 점검해보니 게임에서 이탈하는 플레이어들의 대다수가 꼬리가 길다(long tail)는 것을 알 수 있었다. 두 번째 달력의 예에서 볼 수 있듯 몇 주 혹은 몇 달 동안 가끔씩 게임을 하는 것처럼 말이다. 이들은 활발하게 게임 하는 것은 실질적으로 그만두었지만 가끔 로그인을 한다. 사실 게임은 이미 접었지만 경매나 채팅을 하기 위해서이거나, 어쩌면 길드 친구(guildmate)들에게 계정을 넘겨주었기 때문에 가끔 로그인하는 것이다.

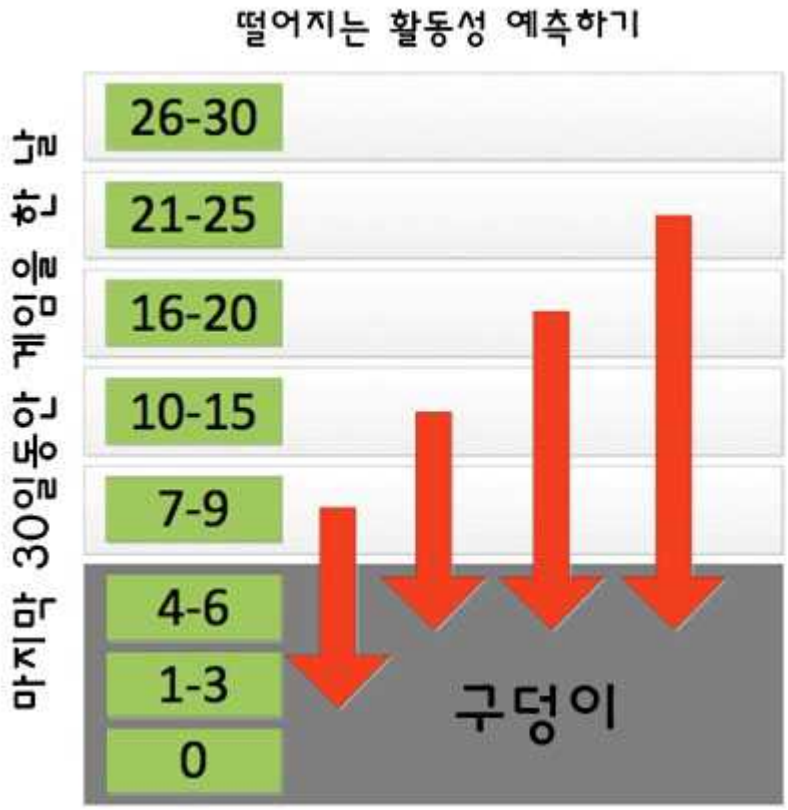
다음 단계는 플레이어의 활동이 줄어들기 시작한 날을 역추적하기 위해 실증적인 기준치를 사용해 이 꼬리를 잘라내는 것이었다. 가장 효과적인 질문(query)은 이런 것이었다. "지난 30일 동안 게임을 한 날이 9일보다 적을 때 플레이어의 마지막 날은?" 그럼에도 정확도는 80퍼센트가 안 됐고, 실증적인 규칙은 아주 캐주얼한 플레이어들에게는 적용되었지만 충성도 높은 플레이어들에게는 적용되지 않았다.

이탈 순간을 재규정하다

이 프로젝트가 성공할 수 있었던 주요 요인은 이탈 순간을 "플레이어가 게임을 떠났을 때"에서 "플레이어의 활동이 이탈 한계점 아래로 떨어졌을 때"로 재구성한 것이었다. 우리는 이미 "지난 30일 중 게임에 로그인한 날"로 정의되는 빈도 기준(Frequency metric)을 만들어두고 폭넓게 사용하고 있다. 요컨대, 이는 플레이어가 얼마나 자주 게임을 해왔는지를 의미한다. 매일, 하루 걸러, 주말에, 한 달에 단 며칠일 수 있다. 우리는 플레이어들을 플레이 빈도에 따라 다음과 같이 나눈다.

다음 단계는 그들이 이탈의 아주 높은 가능성을 가진 극도의 휴지 상태 구역인 "구덩이(The Pit)"로 떨어질 때 이탈을 재규정하는 것이다. 이 아이디어는 비즈니스 견해로 볼 때 정말 타당하다. 이탈자들을 그들이 게임을 영원히 떠나는 날 발견해내는 대신, 우리는 **일찍 발견하는 것**에, 그리고 흥미를 잃은 플레이어들을 예상하는 것에 포커스를 맞추고 있다. 그리고 몇 주 동안 그들이 계속 게임을 하도록 장려한다.

이 새로운 접근은 2주 안에 구덩이로 떨어질 7-9, 10-15, 그리고 16-20 집단의 플레이어들, 3주 안에 떨어질 21-25 집단을 예상하는 것이었다. 그래서 우리는 탄력을 잃고 있는, 다음 몇 주에 걸쳐 활동이 상당히 떨어지는 플레이어들 찾고 있다.



메트릭(metiric) 선택하기

첫 프로젝트[3]에서 얻은 핵심적인 통찰 중 하나는 이탈을 예측하는 데 일반적인 활동들이 중요하다는 것이었다.

우리는 그것이 베테랑 플레이어들을 분석 하는 데도 중요한 역할을 하리라 예상했지만, 그래도 게임과 연관된(game-specific) 활동들과 사회적 메트릭 몇 가지도 함께 시험해보기로 했다.

- 채팅 활동 – 귓속말, 길드 채팅, 공개 채팅
- 자원 채집과 제작
- PvP 및 PvE 인스턴스 방문
- 계정의 최대 캐릭터 레벨
- 남아 있는 유료 이용일수

매일 활동하는 시간과 게임하는 시간은 핵심적인 예측 요소로 예상했었고, 실제로도 그랬다.

적합한 계산법의 선택

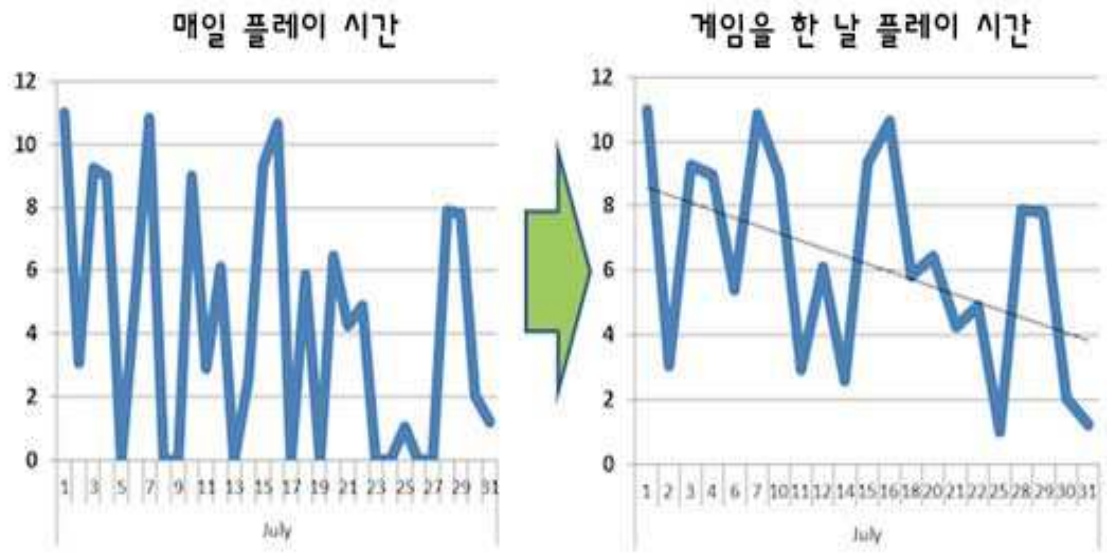
새로운 플레이어들을 분석할 때는 단 며칠 분량의 데이터만 갖고 있었다. 단순하고 즉각적인 값을 사용했다는 뜻이다. 그러나 베테랑 플레이어라고 하면 몇 주(weeks)나 몇 개월(months)을 생각한다. 따라서 시간에 따라 집적된 데이터에는 다른 접근방식이 필요하다. 이 경우에는 이동합계(moving total)와 이동 평균(moving average), 도함수(derivative)과 절편(intercept)이 유용하다.

[3] 참조링크: 위와 동일

우리는 장기간에 걸친 매일의 활동을 분석하기 위해 활동이 있었던 30일의 이동 합계를 이용했고, 그래프는 1차 방정식에 가까왔다. 데이터 마이닝 모델로 가는 실제 메트릭은 직선 근사치의 기울기와 절편이었다. 계산은 [평범한 최소제곱법(ordinary least squares method)][4]로 하며, 데이터를 준비하는 과정에서 T-SQL로 하면 꽤 쉽다.

매일 게임하는 시간을 분석하기 위해 게임을 전혀 하지 않는 휴지 상태인 날들은 근사치 적용 전에 제거했다.

[4] 참조링크: http://en.wikipedia.org/wiki/Ordinary_least_squares

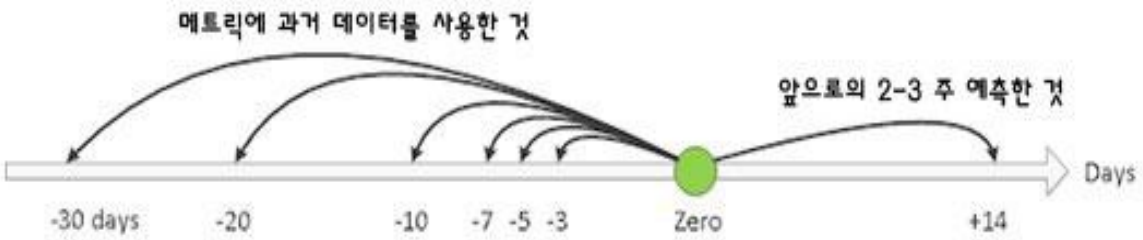


ETL 프로시져는 원점에서 다시 작성해야 했고 모든 데이터를 다시 가져와야 했지만 이 아이디어는 그럴만한 가치가 있었다. 16-20 집단을 대상으로 한, 조율되지 않은 데이터 마이닝 모델에서 나온 첫 결과가 80퍼센트에 가까운 정확성을 보였다.

마침내, 다른 집적 기간과 방법에서 약 30 메트릭으로 2-3 주 내에 구덩이로 떨어질 위기의 플레이어를 예측하는 데 있어 80~90퍼센트의 정확성을 달성했다. 이것은 꽤 인상적인 결과다. 하지만 몇 달 동안 우리가 새로운 메트릭과 방법을 시도했음에도 여기에서 벗어나지 못하고 고착 상태에 빠졌다. 결국, 돌파구는 상세한 과거 메트릭을 도입하면서 찾을 수 있었다.

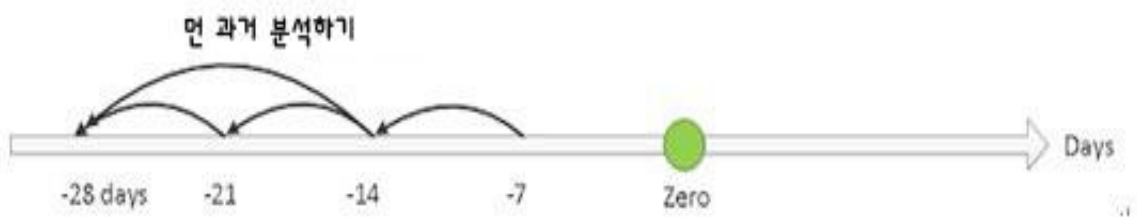
상세한 지난 경험

그때 우리 모델의 타임라인은 다음과 같았다.

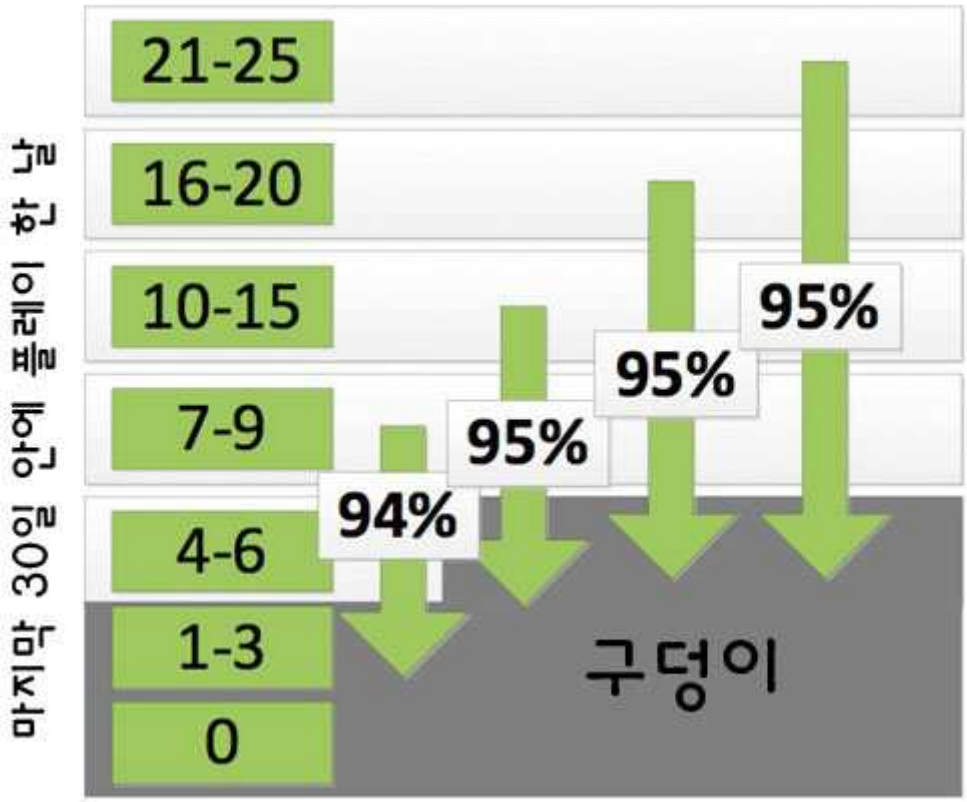


제로(Zero: 0) 지점은 우리가 미래를 예측하는 날이다. 예측은 앞서 언급한 집단(cohort)에 따라 이삼 주 앞서 만들어진다. 데이터마이닝 모델은 과거 X일 동안 계산된, 게임하는 날 플레이한 시간의 이동 평균의 1차 도함수(derivative)처럼 과거 기간 동안 다양한 메트릭으로 채워진다. 모든 메트릭은 0 지점으로부터 지난 X일을 토대로 계산되었다. 지난 사흘, 닷새, 일주일 전 등등.

신선한 아이디어는 과거의 포인트에 관련된 몇몇 메트릭을 계산하는 것이었다. 예를 들어, 우리는 *7일 동안 게임한 날마다 매일 게임 시간의 평균 움직임의 1차 도함수*를 계산할 수 있다. 하지만 이때 0 지점 전의 14일을 되돌아보면서 해야 한다. 플레이어 활동의 롱테일 효과에 대해 말한 것을 기억하는가? 본질적으로 이 아이디어는 그 꼬리의 각 부분을 상세히 해부하고 독립적인 메트릭으로 분석한 것이다. 우리는 그러한 상세한 과거 질문의 조합을 시도했다. 지난 21일 동안 7일(7, -21), (7-7), (14-14)처럼 말이다.



이 아이디어는 엄청난 승리였고 정밀함을 높여주었다. 그리고 거의 모든 집단에 있어 95퍼센트까지 수작업으로 조율한 뒤 다시 불러들였다.



블랙박스

가장 황홀했던 것은 최상의 정밀함을 갖춘 마지막 데이터 마이닝 모델이 전적으로 **단 두 개의 메트릭**(활동한 날들과 매일 플레이 시간)의 도함수와 계산을 기반으로 했다는 점이었다! 다른 세그먼트에서는 다른 도함수들도 중요하다. 21-25번 모델의 경우 우리의 상세한 과거 계산들이 모두 중요하다. 하지만 7-9 집단(cohort)의 경우, 30일 평균은 물론이고 포인트 제로 이전 3-5일의 가까운 과거의 메트릭에도 근거하고 있다. 어쨌든 신규 플레이어의 이탈을 예측할 때보다는 계산이 한참 많이 복잡하다. 아래는 최종 데이터 마이닝 모델의 예시이다.

Structure	Old Cluster 21-25	21-25 Neyro	21-25 Neyro simple
	Microsoft_Decision_Trees	Microsoft_Neural_Network	Microsoft_Neural_Network
Ap Instant Intercept	Input	Input	Ignore
Ap Instant Slope	Input	Input	Ignore
AP Sum Intercept	Input	Input	Ignore
AP Sum Slope	Input	Input	Ignore
Char Max Lvl	Input	Input	Ignore
Chat A Cnt Intercept	Input	Input	Ignore
Chat A Cnt Slope	Input	Input	Ignore
Chat A Instant Intercept	Input	Input	Ignore
Chat A Instant Slope	Input	Input	Ignore
Chat W Instant Intercept	Input	Input	Ignore
Chat W Instant Slope	Input	Input	Ignore
Chat W Intensity Intercept	Input	Input	Ignore
Chat W Intensity Slope	Input	Input	Ignore
Cluster14 Back	Input	Input	Input
Cluster21 Back	Input	Input	Input
Cluster21 Future 0 6	PredictOnly	PredictOnly	PredictOnly
Cluster7 Back	Input	Input	Input
Craft Intercept	Input	Input	Input

이 것이 어떤 미스테리한 수학공식이 들어있는 블랙박스처럼 보인다면....맞다. 신규 프레이어의 이탈을 예상하는 방법을 알아냈을 때를 돌이켜보면, 우리가 만들어낸 매우 정밀한 모델에도 불구하고 실제 이탈요인에 대해서는 거의 모른다는 점이 문제였다. 베테랑의 경우에도 마찬가지다. 이탈의 속성에 대해 인간을 이해할 수 있는 결과는 얻지 못했다. 단지 95퍼센트 정확도의 멋드러진 블랙박스뿐이다.

결과

이제 우리는 베테랑 플레이어들이 게임을 그만두려고 할 때, 이삼 주 앞서 그들의 활동이 극적으로 줄어드는 것을 예상할 수 있다. 이에 따라 우리의 커뮤니티 관리자들이 그 플레이어들을 살피고, 문제를 풀어주거나, 약간의 인센티브를 제공해서 충성도를 높일 수 있게 되었다.

이 데이터마이닝 프로젝트는 신규 플레이어의 이탈을 예상할 때보다 수학적으로 복잡하고 블랙박스처럼 내부구조는 알기 힘들며 세부 조율과 결과 검증에도 시간이 더 많이 필요하다. 하지만 95 퍼센트의 정확도와 재현율을 보인다. 게임과 직접 관련없는 메트릭이 최종 데이터마이닝 모델에 들어간다는 사실이 흥미롭다. 순전히 '활동 일수'와 '매일 플레이 시간'에서 나온 메트릭에만 의존해서 예상하는데, 이는 모든 게임에 통용되는 요소이고 어쩌면 웹서비스에도 적용할 수 있을 것이다.