type_LI_행운 지수 산출

지원자 : 이호성_961123

문제

제공된 데이터를 이용해서 다음 과제을 해결해 주시기 바랍니다.
A라는 게임에는 4% 확률로 S급 아이템을 주는 퀘스트가 있습니다. 이 퀘스트를 시도한 유저 중 운이 없는 유저는 따로 선별하여 별도의 보상을 추가로 지급하려고 합니다.

이에 첨부된 데이터를 활용하여 유저 별로 얼마나 운이 있었는지(행운지수)를 측정하는 모델을 만들고 모델의 동작원리를 분석자료로 정리해주세요.
그 모델로 측정했을 때 케어가 필요한 유저를 식별하여 해당 유저의 user_id를 분석자료와 함께 제출해주세요.

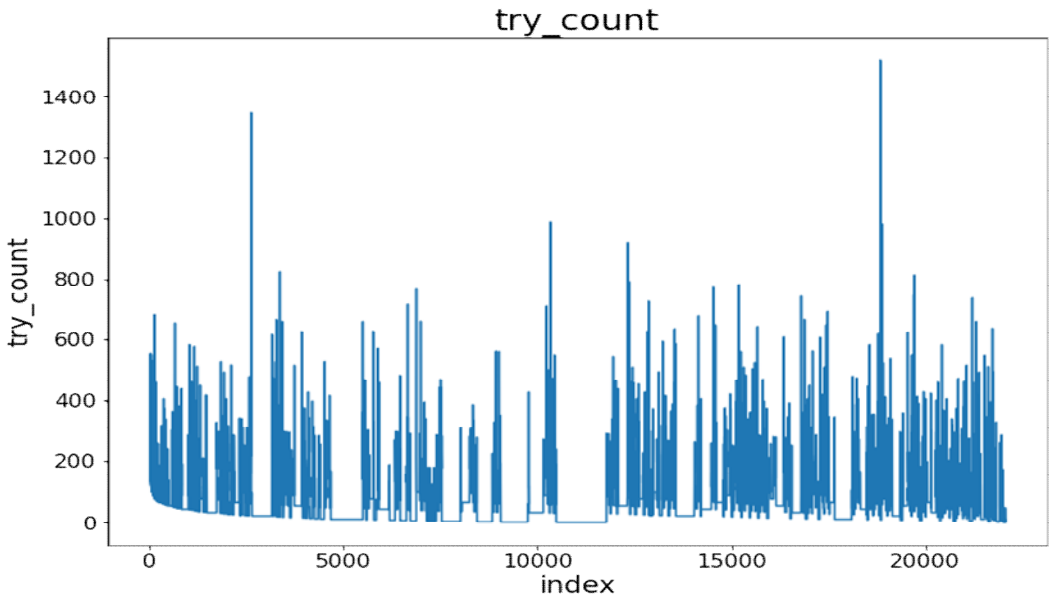
참고로 데이터는 다음과 같습니다.

filename : dataset.csv

user_id : 유저 고유 키
try_count : 퀘스트를 클리어한 횟수
get_count : S급 아이템을 얻은 횟수

dataset.csv 데이터 분석

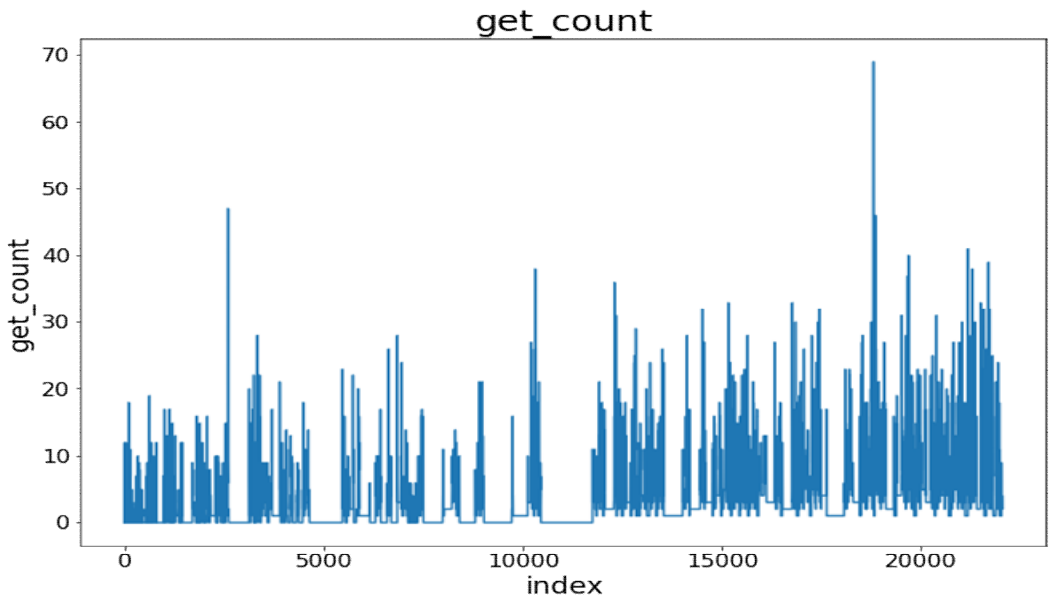
dataset.csv에는 user_id, try_count, get_count 3가지 칼럼이 존재합니다.
먼저 try_count 데이터를 시각화하여 개략적인 형태를 살펴보겠습니다.



전체 try_count plot

많은 데이터를 하나의 plot으로 그려서 정확하지는 않으나 200회 이상부터 1,000회 미만의 try_count가 많다는 것을 파악할 수 있습니다.

다음으로 get_count 데이터를 시각화한 plot입니다.



전체 get_count plot

위 플롯을 통해 대략적인 성공 횟수 및 분포가 0~40회 정도에 머무르는 것을 확인할 수 있었습니다.

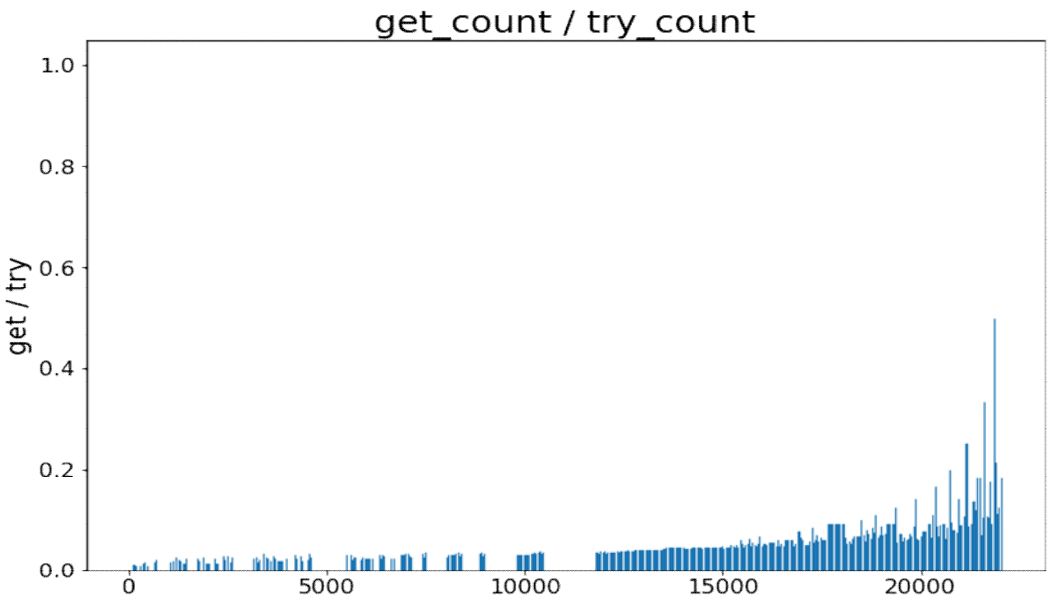
try_count 최댓값	get_count 최댓값	try_count 평균값
1518	69	67.57

try_count와 get_count의 최댓값. try_count의 평균값은 위의 표에서 확인할 수 있습니다. 그 외에도, 1,000회 이상의 try_count 값은 2개 존재하며, 그중 1,518번 시도하여 69번 성공한 18,809번 유저는 4.54%의 비율로 가장 높은 get_count 수치를 가지고 있습니다.

행운 지수 측정

1. get_count / try_count 비율로 행운지수를 측정

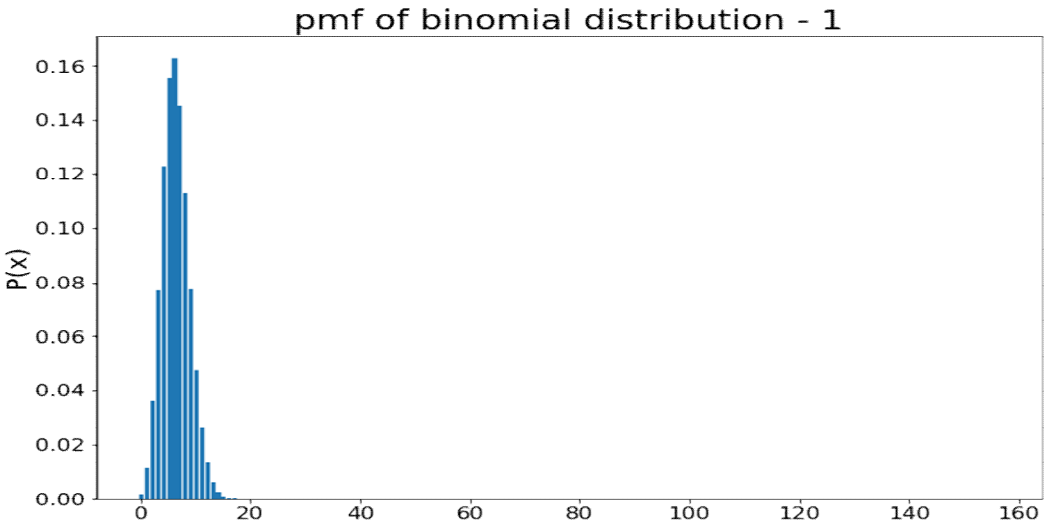
각 유저의 get_count를 try_count로 나누어 비율을 통해 행운지수를 측정해보고자 하였습니다.



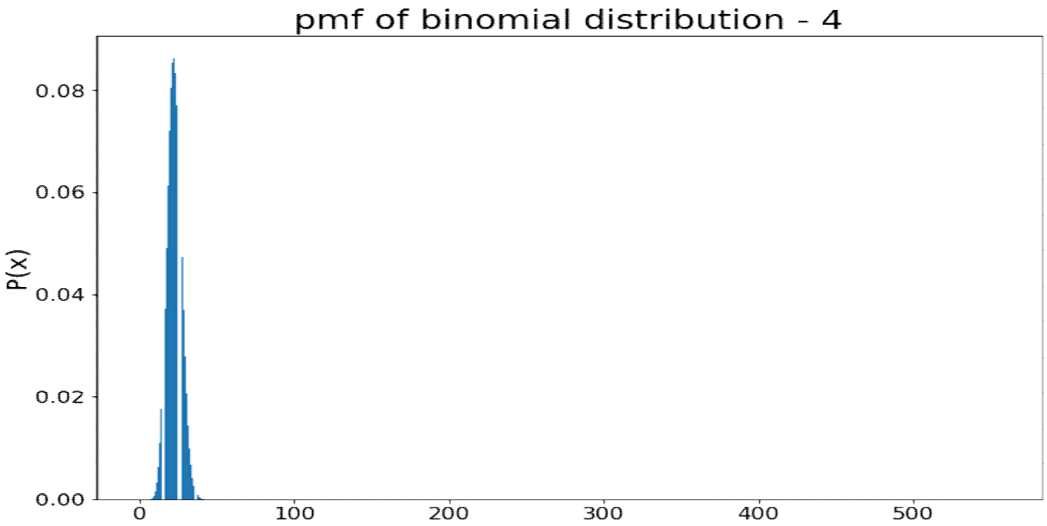
get_count / try_count plot

plot에서 전체적으로 데이터의 수치가 0.2를 넘지 않습니다. 추가로 0인 데이터가 꽤 많이 확인할 수 있습니다. 이는 get_count가 0인 한 번도 당첨되지 못한 유저들을 의미합니다.
또한 단순 비율로만 따졌기 때문에 시도를 많이 한 유저와 시도를 적게 한 유저간 구별이 어렵습니다. 20번 중 1번 된 사람과 100번 중 4번 당첨된 유저는 get / try에서의 비율은 동일하나 동일한 운이라고 하기 힘듭니다. 따라서 행운 지수를 단순 비율로 측정하기는 쉽지 않다고 판단하였습니다.

2. 이항분포로 행운 지수를 측정



1번 인덱스 이항분포에 대한 plot

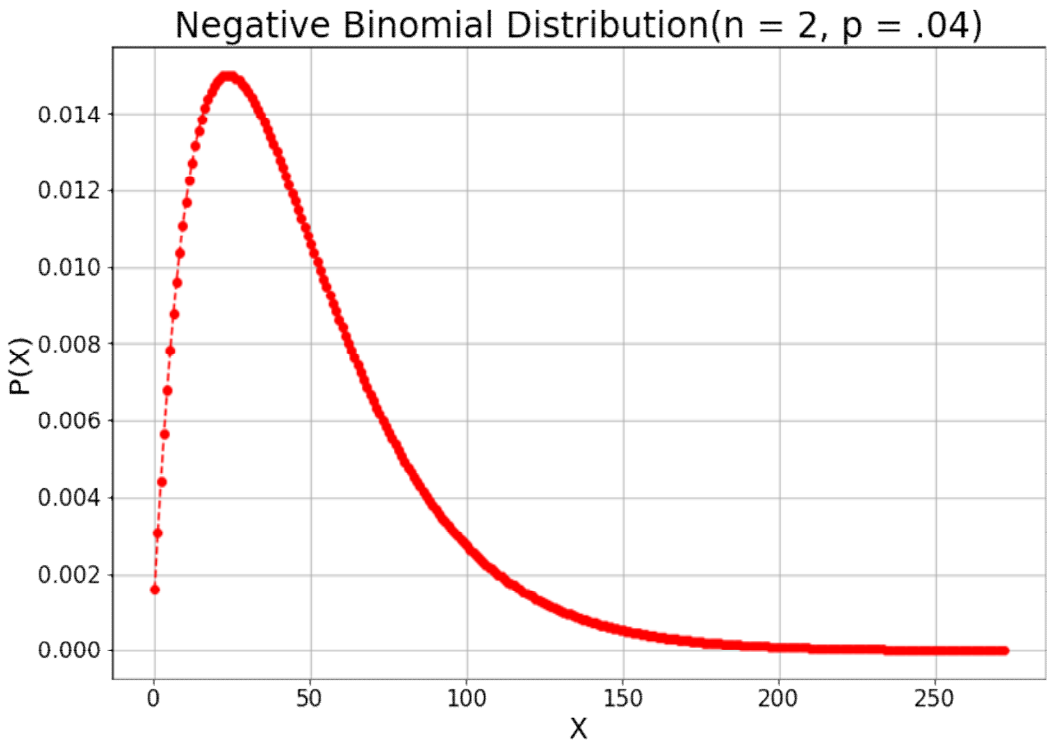


4번 인덱스에 대한 이항분포에 대한 plot

1번 유저는 try_count 값이 156, get_count 값이 0이고, 4번 유저의 경우 try_count가 555번, get_count 값이 11회입니다.
1번 유저의 경우 0회 ~ 약 15회 내의 성공확률이 높다는 것을 확인할 수 있습니다. 데이터 프레임 4번 유저의 경우 0회 ~ 약 30회 사이에서 당첨될 확률이 높은 것을 확인할 수 있습니다. try_count 횟수가 높아질수록 그래프가 더 촘촘해지며 try_count에 대한 n 회 성공 확률이 점점 낮아지는 것을 확인할 수 있었습니다.
그리고 각 유저가 0.04인 강화를 n 회 시도했을 때, k번 성공할 확률은 단순히 성공실패에 대한 확률에 대한 실 확률이므로, 이항분포 데이터가 행운지수와 관련된 큰 의미를 가지기는 어렵습니다.

3. 음이항분포로 행운 지수를 측정

이항분포에서는 n 회 시도했을 때 get_count 회 성공할 확률에 대해서 구했습니다. 그와 다른 방법으로 음이항분포를 이용하여 n 회 성공하기 위한 시도횟수에 대해서 구하도록 하겠습니다. 즉, n 회 성공을 중점으로 성공하기 위해 필요한 시도 횟수를 구하여 해당 데이터를 활용할 수 있는지 살펴보겠습니다.



0번 유저에 대한 음이항분포

다음 plot은 0번 유저에 대한 음이항 분포를 산점도로 표현한 것입니다. 이 plot에서 다음과 같은 정보를 확인할 수 있습니다.

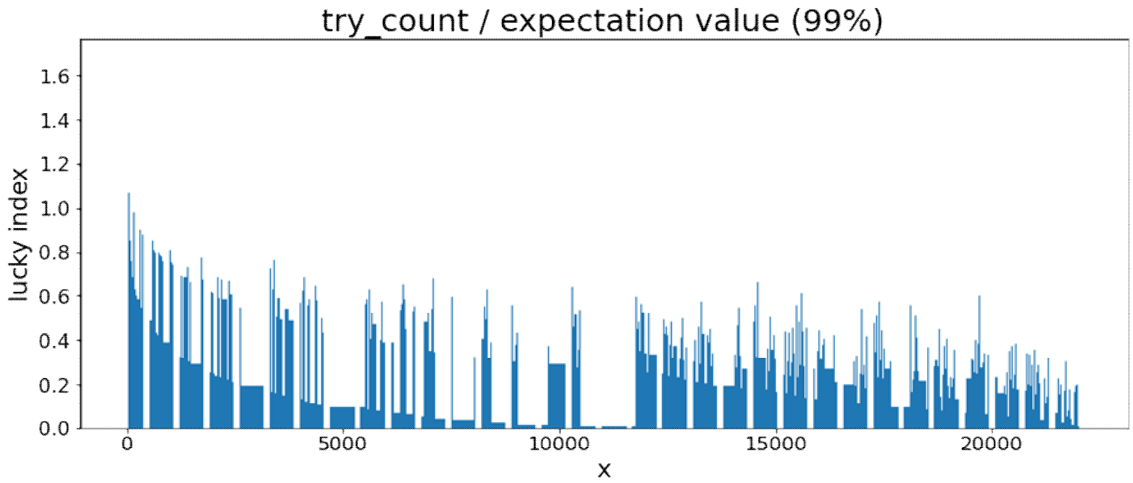
- 약 50% 확률로 2회 성공하기 위한 횟수 : 41회
- 약 80% 확률로 2회 성공하기 위한 횟수 : 73회
- 약 95% 확률로 2회 성공하기 위한 횟수 : 95회
- 약 99% 확률로 2회 성공하기 위한 횟수 : 163회

이를 통해서 각 유저의 시도횟수가 음이항분포 상 어느 위치에 있는지 대략 알 수 있습니다. 따라서 음이항분포를 사용해 각 유저에 맞는 행운지수를 구할 수 있을 것입니다.

3-1. 음이항 분포 상 99% 시도횟수로 행운지수 측정

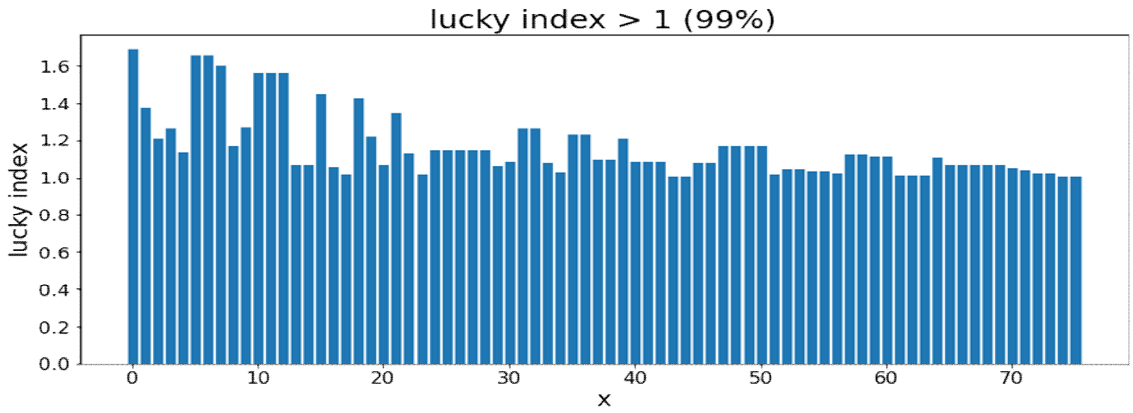
저는 다음과 같은 방법으로 행운지수를 측정할 것입니다.
1. 99% 확률로 get_count 회 뽑을 수 있는 음이항 분포상의 값들을 모두 구합니다.
2-1. try_count / get_count 회 뽑기 위한 시도횟수(99% 음이항 분포 값)로 행운지수를 나타냅니다.
2-2 get_count가 0인 경우에는 try_count / 1회 뽑기 위한 시도횟수(99% 음이항분포)로 행운지수를 나타냅니다.
3. 해당 행운 지수가 담겨있는 리스트를 데이터프레임으로 변환합니다.

이렇게 구한 행운지수를 통해 1 이하의 수치의 경우 해당 당첨 횟수의 기댓값(99%) 보다 적은 횟수로 당첨이 되었기 때문에 상대적으로 운이 좋다고 할 수 있습니다. 1 이상은 당첨 횟수 기댓값(99%)보다 많은 횟수로 운이 좋지 않습니다.



행운 지수 plot

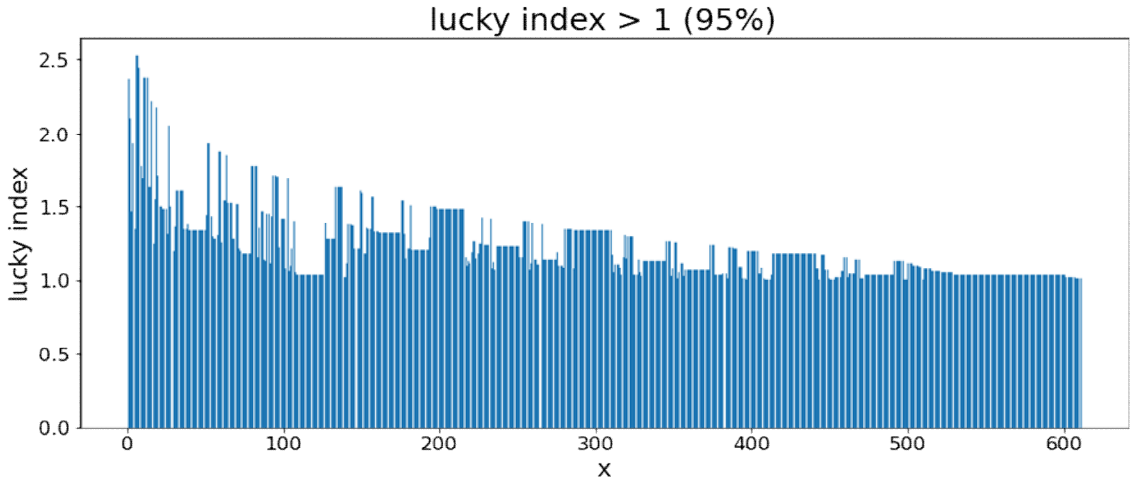
또한 행운 지수가 1 이상인 유저의 행운 지수를 나타낸 plot을 통해 99% 상위 분포 이상의 시도횟수를 가진 유저가 77명이라는 것을 확인할 수 있으며 99% 이상 시도 횟수보다 1.6배 이상 시도한 유저가 있다는 것을 파악할 수 있습니다.



lucky index > 1

3-2. 음이항 분포 상 95% 시도횟수로 행운지수 측정

이번에는 음이항 분포상 95%의 시도횟수를 기준으로 행운지수를 구해보겠습니다. 구하는 과정은 99%에서 수행했던 과정과 동일합니다.



95% 기준 lucky index > 1

95% 행운지수가 1 이상인 유저는 672명이었으며, 95% 시도횟수보다 2.5배 이상 시도했던 유저가 존재하는 것을 확인할 수 있습니다. 그리고 99%에서의 index > 1 과 형태가 매우 유사합니다. 99%에서 구한 77명은 95%의 672명에 대해서 약 11.4%이며 95%의 경우 상대적으로 증가량이 많다고 판단됩니다. 따라서 소수의 유저를 케어하기 위해서는 99%가 효율적입니다.

결론

행운 지수를 try_count / get_count 회 성공하기 위한 음이항 분포 상 시도횟수로 정하였습니다. 행운 지수가 커지면 커질수록 운이 좋지 않고, 반대로 작을수록 운이 좋다고 판단할 수 있습니다.
저는 99% 성공할 수 있는 시도횟수 이상 시도한 1 이상의 행운지수를 가지고 있는 유저들에 대한 케어가 필요하다고 판단되며, 이를 가지고 있는 데이터로 분류하였을 때 약 77명의 유저 ID가 확인되었습니다. 해당 유저들의 명단은 answer_99.csv 파일으로 저장하였습니다.
추가로, 좀 더 많은 유저를 케어할 수 있다면, 95%와 99% 사이에서 휴리스틱하게 판단하여 유연하게 케어가 필요한 유저를 산출할 수 있을 것입니다.