실험을 위한 지표와 종합 평가 기준

가짜연구소 인과추론팀
온라인 통제 실험 연구자로 거듭나기

2024-04-09

양유승

Chapter 7 세 줄 요약

Causal-Lab

1. 비즈니스 지표로부터 실험 지표를 만들기 위해 고려해야 할 몇 가지 사항이 있다.

2. 한 가지 지표만 보기보다는 두 가지 이상의 지표를 조합해 OEC를 만들어 실험을 해보자.

3. 상관관계는 인과관계가 아니니 OEC를 만들었더라도 막 사용하지 말자.

1. 비즈니스 지표로부터 실험에 적절한 지표 만들기

Causal-Lab

비즈니스 지표와 실험 지표

- 비즈니스 지표는 온라인 실험에 직접적으로 유용하지 않을 수 있음
- 실험에 대한 지표는 아래 세 가지 고려할 사항이 있음

측정 가능 Measurable	• 단기(실험 기간)에 측정할 수 있어야 하고 계산 가능해야 함 • 하지만 구매 만족도와 같이 측정하기 어려운 것도 있음
귀속 가능 Attributable	• 변형군에 지표값을 귀속시킬 수 있어야 함 • 하지만 예를 들어 어떤 처리(treatment)가 앱 충돌을 유발하는지 분석을 하려고 하는데, 해당 데이터는 third-party에서 들고 있어서(ex. 구글 앱스토어) 측정할 수 없는 경우도 있음
민감하고 시기적절 Sensitive and timely	• 실험 지표는 시기적절하게 중요한 변화를 감지할 수 있을 정도로 민감해야 함. • 민감도는 기초가 되는 지표의 통계적 분산, 효과의 크기(실험에서 실험군과 대조군 사이의 델타) 및 무작위 추출 단위(사용자 등) 수에 따라 달라짐

What is crash rate?

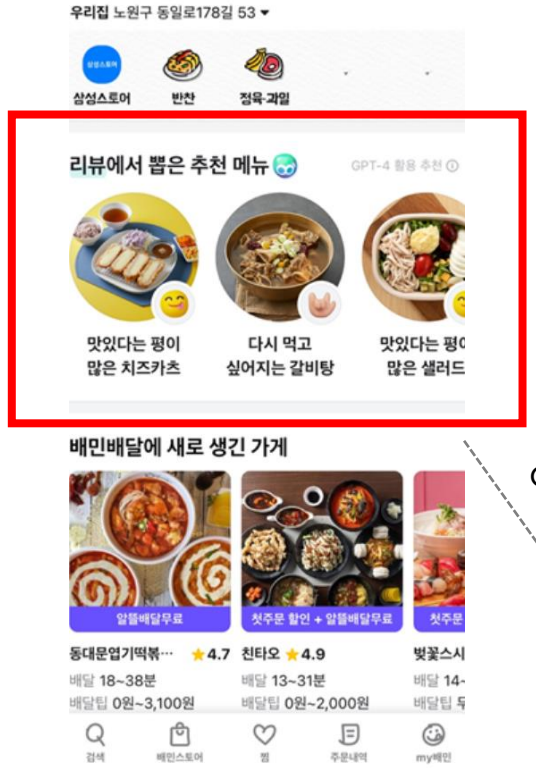
Crash rate refers to the percentage of times a particular application or system crashes or fails to function properly during a given period of time. It is usually measured as the number of crashes divided by the total number of sessions or operations performed. A high crash rate indicates that the application or system is unreliable and needs improvement. Measuring and monitoring crash rate is important for ensuring the stability and performance of software applications and systems.

1. 비즈니스 지표로부터 실험에 적절한 지표 만들기

Causal-Lab

(낮음) 민감도 (높음)

영향 미미
(냅다 가도할 수 밖에…)



사용자에게 어떤 경험을 주는지 알 수 없음

노출 (Imp)

다른 섹션에 대한 카니발 측정 불가

섹션 CTR

GOOD

페이지 CTR

섹션 클릭 후 구매수

1. 비즈니스 지표로부터 실험에 적절한 지표 만들기

Causal-Lab

- 더 중요한 것은 **'무엇을 위해 최적화'** 하고 있는가
  - Ex. 품질은 생각하지 않은 채 체류 시간을 늘리기 위해, 불필요한 기능들을 붙이고 느린 사이트를 만들어낸다면
    단기적으로는 지표는 개선되겠지만 장기적으로는 낮아진 품질로 사용자가 이탈하는 원인이 될 수 있음

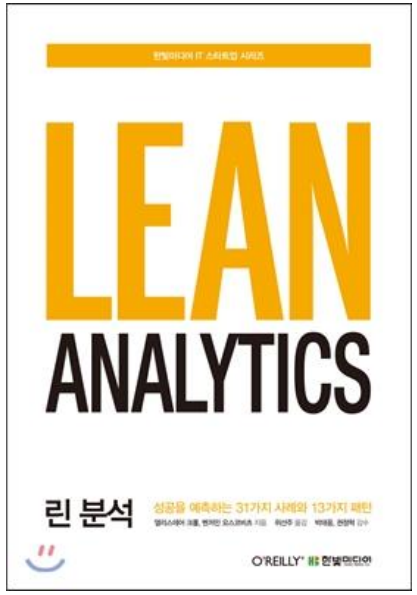
- **실험을 위한 지표 선정** 시 고려해야할 것
  - 측정 가능성, 계산 가능성, 민감도 및 적시성을 충족하는 비즈니스 목표/ 동인 및 가드레일 지표에서 일부 선정
  - 특정 기능의 움직임을 이해하는 데 도움이 되는 전체 지표보다 세분화된 지표 (ex. 전체 CTR > 기능별 CTR로 쪼개보는 것)
  - 신뢰도 가드레일 및 데이터 품질 지표
  - 진단과 디버그 지표 (문제가 나타나는 상황을 자세히 검토할 때 유용)

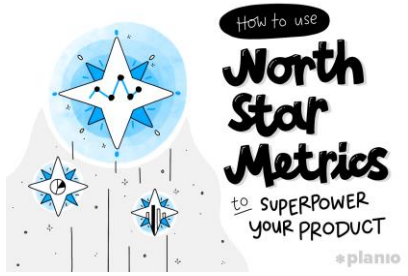
2. 주요 지표를 OEC로 결합하기

Causal-Lab

주요 지표와 OEC 결합의 필요성

- 여러 가지 목표와 동인 지표를 갖고 있는 상황이 주어졌을 때, **하나의 지표**만 선택해야 할까? 아니면 **둘 이상의 지표**를 유지해야 할까? 아니면 그것들 모두를 **하나의 조합 지표**로 결합할까?
    - 목표 달성을 위한 단일 지표는 존재하지 않고, 여러 목표 지표와 동인 지표가 있음
- 실제로 많은 조직은 여러 가지 주요 지표를 검토하고, 이들 지표들의 특정 조합을 고려할 때, 어떤 트레이드오프를 수용할지에 대한 모델을 마음 속에 가지고 있음.
    - 마음 속에 있는 트레이드 오프 모델을 표현하기 위한 해결책이 OEC



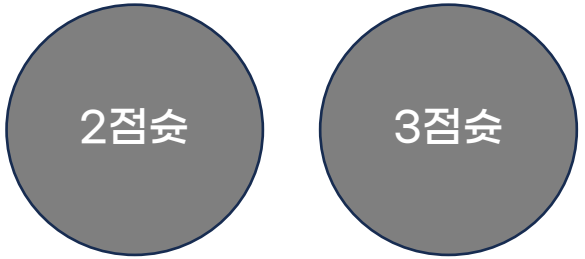


2. 주요 지표를 OEC로 결합하기

Causal-Lab

여러 지표를 가중 조합한 단일 OEC의 효과

- 많은 경우엔 여러 지표를 가중 조합한 OEC를 고안하는 것이 더 바람직한 해결책이 될 수 있음
  - OEC는 성공의 정의를 명확하게 만들고, 더 나아가 구성원들이 트레이드오프에 대해 컨센서스를 가질 수 있도록 함
  - 또한 의사결정에 일관성이 생기며, 팀이 경영진에게 보고할 필요 없이 결정을 내릴 수 있도록 함



더 높은 점수를 얻어 경기에서 이기는 게 중요하지,
3점슛을 더 많이 넣는 것이 중요하지는 않음

2. 주요 지표를 OEC로 결합하기

Causal-Lab

OEC 고안 시 고려 사항

- 결합된 지표를 조작하지 못하도록 확실하게 해야 함
- 여러 지표를 갖고 있는 경우, 각 지표를 사전 정의된 범위(EX.0~1)로 정규화하고 각 지표에 가중치를 할당해 가중합을 도출할 수 있음
    - 처음이라 결정이 어렵다면 아래 네 개의 판단 기준에 따라 결정할 수 있음

NO.	세부 사항	결정
1	모든 핵심 지표의 변화가 0(통계적으로 유의미하지 않음)이거나 양수(통계적으로 유의)고 적어도 하나의 지표가 양수	성공
2	모든 핵심 지표의 변화가 0이거나 음수이고, 적어도 하나의 지표가 음수	실패
3	모든 핵심 지표가 0	실험 검정력을 높이거나, 빨리 실패로 간주
4	일부 핵심 지표가 양수이고 일부 핵심 지표가 음수	트레이드오프를 기반으로 결정

OEC를 도출하는 것이 어렵다면?

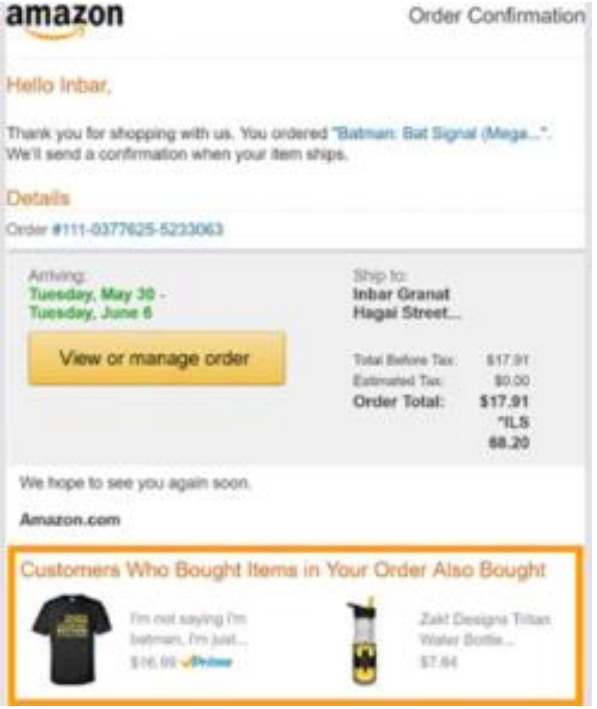
- 핵심 지표를 단일 OEC로 결합할 수 없다면, 핵심 지표의 수를 최소화
    - 지표가 너무 많으면 인지 과부하와 복잡성을 야기할 수 있고, 잠재적으로 조직은 주요 지표를 무시하는 방향으로 갈 수 있기 때문
    - 지표의 수를 줄이면 통계적으로도 도움이 됨. 대표적인 경험 법칙은 핵심 지표를 5개로 제한하는 것

2-1. OEC 예시 - 아마존 이메일

Causal-Lab

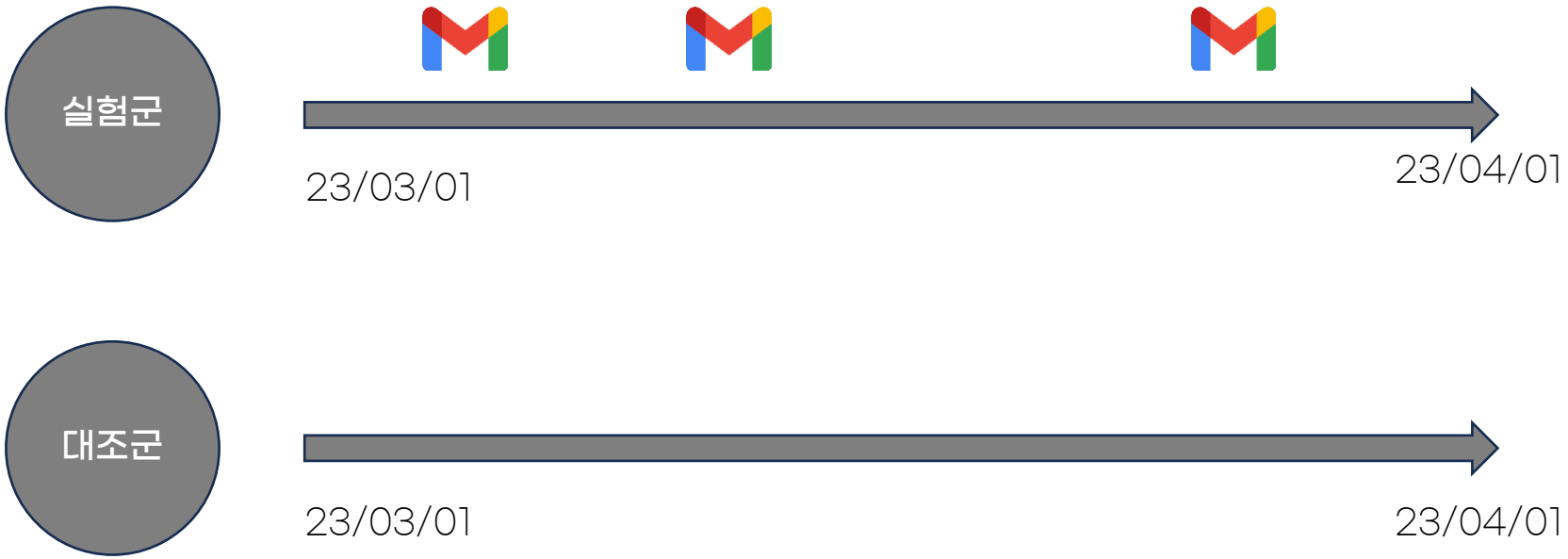
- 아마존은 다양한 조건을 바탕으로 특정 고객을 선정했고, 프로그램화된 캠페인을 기반으로 이메일 시스템을 구축함

이전에 구입한 저자의 신간 서적	• 고객이 구매한 작가의 신작 출시에 안내 이메일 발송
고객 구매 이력 기반 추천	• 고객이 구매했거나 보유하고 있다고 알려준 제품을 기반으로 추천 • 아마존 추천 알고리즘이 설정
상품 기반 추천 (Cross Pollination)	• 특정 제품 카테고리 조합에 기반해 상품을 추천 • 사람이 직접 설정 (defined by humans…)



2-1. OEC 예시 - 아마존 이메일

- 관건은 이러한 프로그램에 어떤 OEC를 사용해야 하는가
- 초기 아마존이 활용한 OEC(a.k.a fitness function)는 이메일을 클릭한 유저로부터 나온 매출을 기반으로 프로그램을 평가함
- 하지만 단순히 메일만 많이 보내면 매출이 늘어나고, 대조군 또한 시간이 지남에 따라 매출이 늘어날 것이기 때문에 적절하지 않았음



2-1. OEC 예시 - 아마존 이메일

Causal-Lab

- 사용자들이 너무 많은 이메일을 받는 것에 대해 불평하기 시작하면서 스팸화에 대한 문제점이 드러나기 시작
- 아마존의 초기 해결책은 매 X일마다만 이메일을 수령하는 제약 조건을 추가하는 것
- 하지만 이메일 캠페인 자체가 최적화의 대상이 되어버리는 문제 발생.
    - 각 프로그램이 모두 이메일을 보내고 싶어한다면 어떤 프로그램으로 보내야 할까?
    - 어떤 사용자들한테는 3일 주기로 보내고 어떤 사용자한테는 7일 주기로 보내야 할까?

2-1. OEC 예시 - 아마존 이메일

Causal-Lab

- 이후 아마존은 기존 OEC가 단기 매출에 최적화되고 있다는 걸 깨달음
- 이메일에 피로감을 느낀 사용자들이 이메일 수신을 취소하면, 아마존은 미래에 그들을 타겟팅할 기회를 잃게 됨
- 이를 막기 위해, 아마존은 사용자가 이메일 수신을 취소할 때의 사용자 생애 기회 손실에 대한 하한을 설정해 OEC를 새로 정립함

$$\text{OEC} = \left(\sum_i Rev_i - s * unsubscribe\_lifetime\_loss \right) \Big/ n$$

where:

- i ranges over e-mail recipients for the variant
- s is the number of unsubscribes in the variant
- unsubscribe_lifetime_loss is the estimated revenue loss of not being able to e-mail a person for "life"
- n is the number of users in the variant.

- 그 결과, 수신 취소의 생애 손실에 단 몇 달러만 할당하더라도 프로그램 캠페인의 절반 이상이 부정적인 결과가 나타남
  - 즉, 메일을 차라리 안 보내는 것이 장기적으로는 더 높은 매출을 회사에게 가져다 주는 것
- 더 나아가, 수신 취소가 큰 손실을 초래한다는 인식을 가질 수 있었고, 이를 바탕으로 캠페인 군을 위한 수신 취소 페이지를 따로 만들어 수신 취소로 인해 발생하는 손실을 대폭 줄일 수 있었음

2-2. OEC 예시 - 빙의 검색 엔진

Causal-Lab

- 빙은 검색 점유율과 매출을 가지고 실험 목표 달성도를 측정함
  - 실험군에서 우연치않게 알고리즘 버그로 나쁜 검색결과를 보여주게 될 때, 두 핵심 지표가 개선되는 효과를 얻을 수 있었음
    - 만족하지 못한 검색 결과로 인해 더 많이 검색하게 되고, 그로 인해 광고를 더 많이 클릭한 결과 -> 바람직하지 못한 OEC 설정

사용자 당 검색어 수

+10%

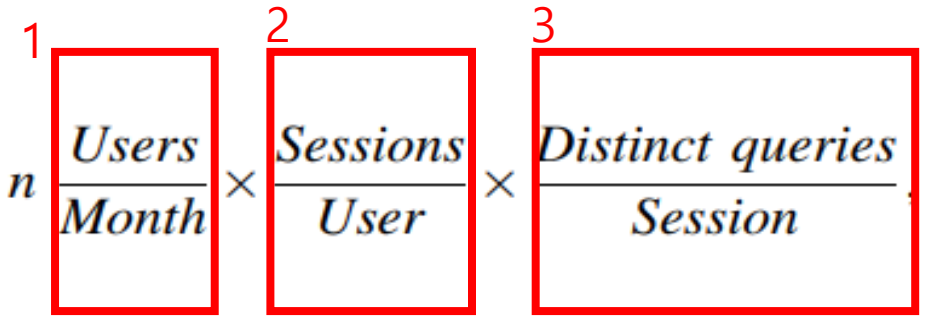
사용자 당 매출

+30%

2-2. OEC 예시 - 빙의 검색 엔진

Causal-Lab

- 빙은 월별 고유 쿼리수를 아래와 같은 방정식으로 분해해 OEC를 새로 생성함
- 검색당 매출은 광고가 사용할 수 있는 평균 픽셀 수를 제한하는 제약 조건을 추가해 측정할 수 있을 것



NO.	세부 사항
1 월별 사용자	• 만약 대조군과 실험군을 50/50으로 분할한다면 각 변형군에 속하는 사용자가 거의 같으므로 이 항을 OEC로 활용할 수 없음
2 사용자당 세션 수	• 최적화해야 하야 하는 핵심 지표 • 검색 알고리즘에 만족했다면 사용자는 빙을 더 자주 방문할 것이기 때문
3 작업별 고유 쿼리 수	• 작업은 측정할 수 없기 때문에 세션을 proxy 지표로 활용 • 더 많은 쿼리를 입력했다면 검색 알고리즘이 형편없다고 볼 수도 있지만 반대로 너무 적은 쿼리를 입력했다면 바로 작업을 포기했다고도 볼 수 있음

3. 굿하트의 법칙, 캠벨의 법칙과 루카스 비판

Causal-Lab

"Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes"
(Goodhart 1975, Chrystal and Mizen 2001)

"When a measure becomes a target, it ceases to be a good measure"
(Goodhart's law 2018, Strathern 1997).

"The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor"
(Campbell's law 2018, Campbell 1979).

루카스 비판:
모든 상관관계가 인과관계를 의미하지 않으며,
많은 조직이 OEC를 선택할 때 상관관계를 인과관계로 잘못 판단하는 것을 주의해야 함

논의해볼 거리

Causal-Lab

- OEC를 활용하는 실험 사례가 많은가? 사용자 당 세션 정도 말고, 아마존 이메일 사례처럼 더 많은 지표를 포함.
  - 일반적으로는 핵심 지표, 보조 지표, 가드레일 지표 각각에 1~2가지를 넣지 않나?

- OEC를 활용하기 어렵다면 어떻게 트레이드오프를 감안할 수 있을까?
  - ex. 실험군을 더 세분화. 특정 서비스가 잘 안 되는 지역과 잘 되는 지역 구분

- 팀마다 KPI가 다를 텐데, 어떻게 각 팀의 KPI를 포괄해 OEC를 만들 수 있을까?
  - 예를 들어 A기업이 한집배달을 새로 출시해 기존 묶음배달과 비교하고자 한다면…
    - 고객팀에서는 고객의 배달경험을 향상시키고자 할 것이고, 운영팀에서는 운영 리소스를 줄이고자 할 것
    - 지표 또한 전자는 배달시간, 배차수락율 / 후자는 배달 건당 배달처리비 등을 보고자 할 텐데…

- 잘 만든 OEC의 사례는 무엇이 있을까? 트레이드오프를 어떻게 반영했을까?

- 모두 함께 얘기해보아요~~

E.O.D