2부: ML 예측과 인과관계 분석

예제 소개 및 강의 개요

데이터와 의사결정 | 정종빈

강의 개요

1부: 불확실성과 데이터

- 의사결정 모형
- 불확실성 계량화 전략
  - 빈도주의(Frequentist)
  - 베이지안
- 최적 의사결정

2부: ML 예측과 인과관계 분석

- ML 예측모형 기초
- 의사결정과 인과관계 분석

예제 소개

“배송료 무료” 쿠폰 이벤트

예제: 배송료 무료 쿠폰 이벤트

- 온라인 쇼핑몰 “구빵” DS로 취직한 당신의 첫 프로젝트
- 작년 대히트였던 “배송료 무료 쿠폰” 이벤트 시즌2 진행!
- 비상장 기업이었던 작년에는 무작위로 쿠폰 뿌렸던 데에 반해, 상장한 올해는 좀 더 스마트하게 쿠폰을 뿌리고자 하는 경영진
- 어떻게 쿠폰 지급을하고 이벤트를 진행해야 할까?

2부 강의 목적 및 개요

선수학습: 기본적인 ML 코딩 경험 (sklearn Pipeline 활용 등)

목적:

✅ 보편적인 “ML 모델을 활용한 의사결정 접근 방식" 소개

⛔ 특정 ML 모델/알고리즘 소개

개요:

1. 예측 모형의 보편적인 고려사항 숙지
2. 불가피한 “인과관계분석"의 난제, 접근법, 흔한 함정 탐구

다음: 데이터 소개

다음 수업 수강 전 데이터를 탐구/분석해 보세요

2부: ML 예측과 인과관계 분석

데이터 소개

데이터와 의사결정 | 정종빈

데이터 소개

- 귀찮아도, 데이터를 꼼꼼히 살펴보고 꼬치꼬치 캐묻는 것이 데이터사이언티스트 업무의 반 (이상)
- 데이터는 어디서 왔는가?
- 각 항목의 정확한 정의 및 특이 사항은?

예제 상황

- 데이터는, 작년 이벤트 후 성과 분석을 위해 당시 DS가 수집
- 작년 이벤트 진행 관련 DS는 모두 상장 후 이직
- 구빵 영업 디테일
  - 고객은 비회원(guest), 일반회원, 연간(유료)회원으로 구매 가능
  - 연간 회원(subscriber)은 무조건 배송료 무료
  - 비회원/일반회원은 일정 금액 이상 구매시 배송료 무료
  - 배송료 무료 금액은 배송지역에 따라 다름 (도서산간/해외 등)
  - 해외 배송료는 배송량에 따라 상이
  - 배송료가 있고 배송료 무료 쿠폰을 보유한 경우, 자동 적용 후 소멸

데이터 개요

- 파일명: **gooppang.csv**
- 당시 행사 대상이었던 일반회원 이상의 세션 정보만 포함
- 각 행(row)은 <한 유저, 한 세션> 간 장바구니를 나타냄
  - 행사 기간 내 한 유저가 여러 세션으로 기록 가능
  - 유저를 구분할 수 있는 정보는 없음
  - 유저 관련한 개인 정보(나이, 성별 등)는 정확하지 않을 수 있음
  - (작년) 행사 기간 중 쿠폰은 한 번만 적용 가능

데이터 개요

age: int
행사 당시 유저 나이

basket: float
장바구니 총액(천원)

checkout: bool
결제(최종 구매) 여부

region: string
배송지역 코드

coupon: bool
배송료 무료 쿠폰 보유 여부

gender: string
유저 성별 (“m”: 남성, “f”: 여성)

monthly_spend: float
행사 당시 월 평균 결제량

shipping_fee: float
배송비

subscriber: bool
연간 회원 여부

tenure: int
행사 당시 유저 가입 기간(월)

기본적인 의문점 탐구 (연습문제)

배송비(shipping_fee)의 정확한 정의는?

- 배송지역별 배송료 무료 금액 적용 전? 후?
- 쿠폰 적용 전? 후?
- 연간 회원은 무조건 무료? 아니면 배송비 적용 후 “할인”?

장바구니 총액(basket)은 배송비(shipping_fee) 포함?

그 외의 다른 의문점은?

앞으로의 분석을 위해 불가피한 가정은?

다음: 연습문제 풀이

다음 수업 수강 전 연습문제에 대해 답을 각자 해보세요

2부: ML 예측과 인과관계 분석

데이터 연습문제 풀이

데이터와 의사결정 | 정종빈

기본적인 의문점 탐구 (연습문제)

배송비(shipping_fee)의 정확한 정의는?

- 배송지역별 배송료 무료 금액 적용 전? 후?
- 쿠폰 적용 전? 후?
- 연간 회원은 무조건 무료? 아니면 배송비 적용 후 “할인”?

장바구니 총액(basket)은 배송비(shipping_fee) 포함?

그 외의 다른 의문점은?

앞으로의 분석을 위해 불가피한 가정은?

Google Colab에서 진행

pandas, numpy, seaborn이 설치된 환경 어디든 실습 가능

다음: 예측 모형 기초

2부: ML 예측과 인과관계 분석

예측 모형 기초

데이터와 의사결정 | 정종빈

개요

이론: Bias-variance (tradeoff?)

실무: Train, validation(, test)

“예측 모형”의 목적

관측한 정보(데이터)를 기반으로

관측되지 않은 것(미래)에 대한 “일반화”의 시도

“예측 모형”의 “오류”

정보 자체의 불확실성 (random noise)

예측 방법으로 인해 발생하는 오류

(빈도주의적 “반복시행”의 관점에서)

“예측 모형”의 “오류”

정보 자체의 불확실성 (random noise)
→ 불가피한 오류

예측 방법으로 인해 발생하는 오류
(빈도주의적 “반복시행”의 관점에서)

“예측 모형”의 “오류”

정보 자체의 불확실성 (random noise)
→ 불가피한 오류

예측 방법으로 인해 발생하는 오류
(빈도주의적 “반복시행”의 관점에서)

- 편향(bias): 예측치가 실제값과 근본적으로 다른 정도
- 분산(variance): 예측치가 데이터의 변화에 민감한 정도

“예측 모형”의 “오류”: 극적인 예

“동전을 던져서 앞면 나올 확률 예측”

데이터: 같은 동전을 10번 던져 봄

“예측 모형”의 “오류”: 극적인 예

“동전을 던져서 앞면 나올 확률 예측”

데이터: 같은 동전을 10번 던져 봄

- 예측방법1: 첫번째 던진게 앞면이면 “100%”, 뒷면이면 “0%”
- 예측방법2: 무조건 “50%”

“예측 모형”의 “오류”: 극적인 예

“동전을 던져서 앞면 나올 확률 예측”

데이터: 같은 동전을 10번 던져 봄

- 예측방법1: 첫번째 던진게 앞면이면 “100%”, 뒷면이면 “0%”
  - Bias = 0: 빈도주의적 입장에서 반복시행시, 앞면이 나올 “확률”로 수렴
  - Variance 높음: 데이터에 따라 100% 혹은 0%
- 예측방법2: 무조건 “50%”

“예측 모형”의 “오류”: 극적인 예

“동전을 던져서 앞면 나올 확률 예측”

데이터: 같은 동전을 10번 던져 봄

- 예측방법1: 첫번째 던진게 앞면이면 “100%”, 뒷면이면 “0%”
  - Bias = 0: 빈도주의적 입장에서 반복시행시, 앞면이 나올 “확률”로 수렴
  - Variance 높음: 데이터에 따라 100% 혹은 0%
- 예측방법2: 무조건 “50%”
  - Bias 높음: 실제 앞면이 나올 확률이 50%가 아닐 수 있음
  - Variance = 0: 데이터와 무관하게 예측치(50%)가 일정함

Bias-variance tradeoff?

- 전통적 통계에서는
  - unbiased (bias = 0) 예측치를 선호
  - 수학적으로
    1. 다양한 unbiased 예측치를 찾고
    2. 그 중 Variance를 최소화

Bias-variance tradeoff?

- 전통적 통계에서는
  - unbiased (bias = 0) 예측치를 선호
  - 수학적으로
    1. 다양한 unbiased 예측치를 찾고
    2. 그 중 Variance를 최소화
- 실무에서는 많은 경우
  - Bias/variance 높든 낮든, 결과적인 예측 성능이 중요
  - → 결과적인 예측 성능을 측정하는 실증적 방법의 중요성
  - 단, bias/variance에 대한 이해가 직관적인 사고에 여전히 도움이 됨

Bias-variance tradeoff?

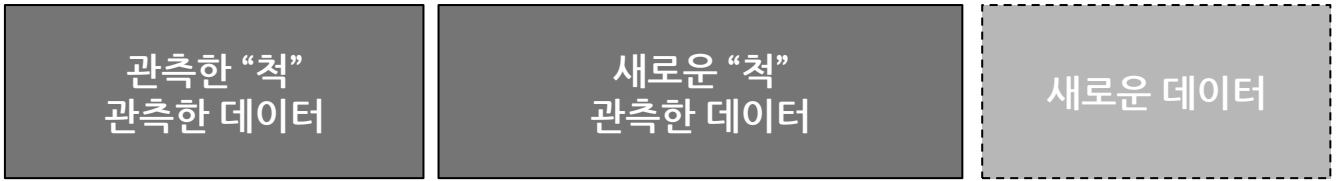
- 전통적 통계에서는
  - unbiased (bias = 0) 예측치를 선호
  - 수학적으로
    1. 다양한 unbiased 예측치를 찾고
    2. 그 중 Variance를 최소화
- 실무에서는 많은 경우
  - Bias/variance 높든 낮든, 결과적인 예측 성능이 중요
  - → 결과적인 예측 성과를 측정하는 실증적 방법의 중요성
  - 단, bias/variance에 대한 이해가 직관적인 사고에 여전히 도움이 됨

예측 성과 측정

알고 싶은 것: “관측한 데이터”를 가지고 만든 모형을 “새로운 데이터”에 적용했을 때 어떤 성능을 가지는가?

문제: “새로운 데이터”는 관측하지 않았음

예측 성과 측정



해결책: “관측한 데이터” 중 일부만 사용하고, 일부는 “관측 못한 새로운 데이터”처럼 취급

예측 성과 측정

관측한 “척” 관측한 데이터	새로운 “척” 관측한 데이터	새로운 데이터

해결책: “관측한 데이터” 중 일부만 사용하고, 일부는 “관측 못한 새로운 데이터”처럼 취급

문제: 다양한 모형의 성과를 비교해서 선택했을 때, 측정된 모형의 성과가 상향 편향

예: 주사위 두 개를 10번씩 던져서 그 중 평균이 낮은 숫자를 선택했을 때, 그 평균은?
= 3.5? < 3.5? > 3.5?

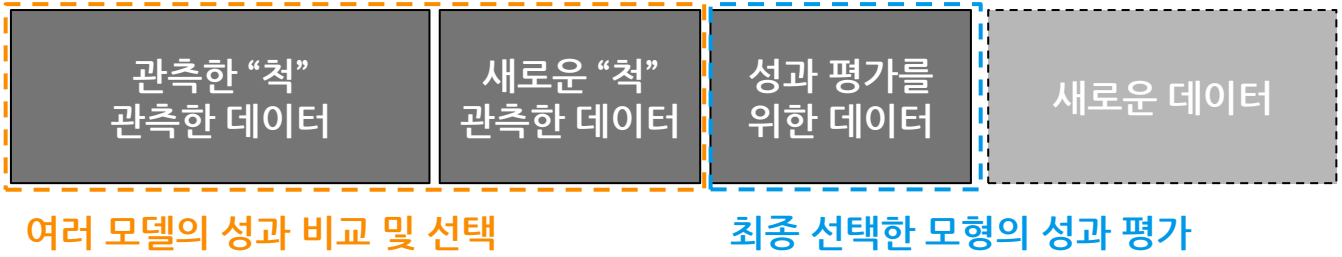
예측 성과 측정

해결책: 모델을 선택하기 위한 데이터와, 모델의 실제 성과를 측정하기 위한 데이터를 분리

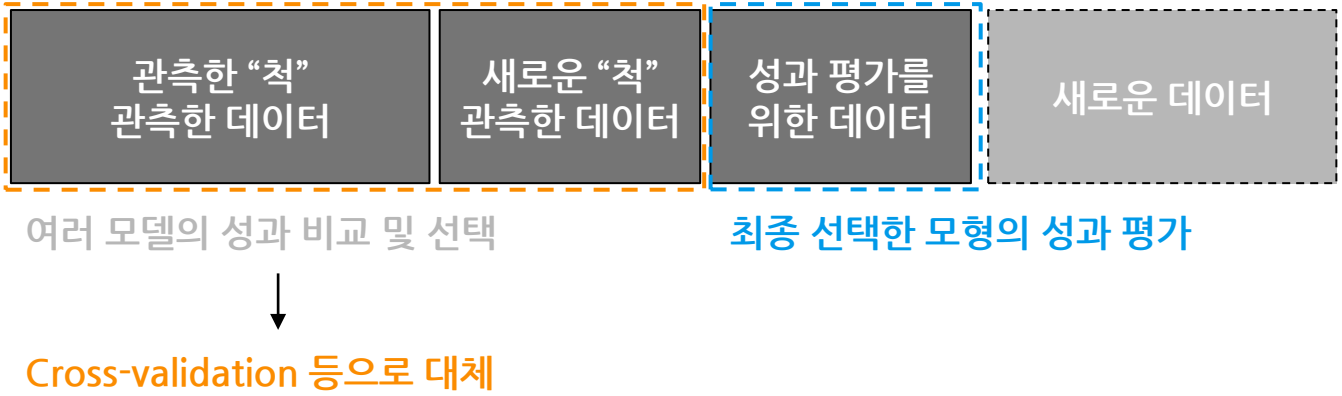
예측 성과 측정

여러 모델의 성과 비교 및 선택

예측 성과 측정


관측한 “척”
관측한 데이터
새로운 “척”
관측한 데이터
성과 평가를
위한 데이터
새로운 데이터
여러 모델의 성과 비교 및 선택
최종 선택한 모형의 성과 평가

예측 성과 측정


관측한 “척”
관측한 데이터
새로운 “척”
관측한 데이터
성과 평가를
위한 데이터
새로운 데이터
여러 모델의 성과 비교 및 선택
최종 선택한 모형의 성과 평가
Cross-validation 등으로 대체

다음: 예측 모형 실습

쿠폰 지급을 위한 예측 모형

2부: ML 예측과 인과관계 분석

의사결정 중심 데이터 분석

데이터와 의사결정 | 정종빈

“데이터, ML”을 거론하기 전에 …

- 쿠폰 지급 목적은?
  - 홍보/성장?
  - 수익률 증가?
- 예산?
  - 예산은 어떻게 산정이 됐는가? 내가 산정해야 되는가?
- 이벤트 “성공”여부의 척도는?

DS 프로젝트 성공 꿀팁

- “무슨 데이터가 있고 어떤 세련된 기법을 활용할 것인가”를 묻기 전에, “<u>어떤 결과/정보를 얻고 싶은가</u>”를 먼저 대답!
  - 그에 따라 “어떤 데이터/모형/기법”은 자연스럽게 정해짐
- 코딩 전, 기대하는 최종 결과(가설)를 먼저 구상
  - 어떤 결과가 나올지에 대해, 실제 결과를 보기 전 최대한 고민
  - 어떤 식으로 의사결정에 도움이 될지 실용적으로 고민
- 어떤 형태로든 “인과관계분석”은 피해 갈 수 없다!
  - 과학적으로 정확한 결과를 도출 할 수 있는 경우는 극히 적음
  - 하지만, 생각보다 “부정확한” 결과도 유용할 때가 많음
  - 최소한의 “인과관계분석”에 대한 이해와 내 결과의 오점은 알고 살자!

다음: 다양한 상황에서의 분석

"정답"은 없음. 목적은 "시야를 넓히는 것".

2부: ML 예측과 인과관계 분석

배송료 무료 쿠폰의 효과

데이터와 의사결정 | 정종빈

앞으로의 수업 개요

- 간단한 의사 결정 상황 제시
- 실증적인 결과 기반으로 실습
- 필요에 따라 이후 관련 이론 소개

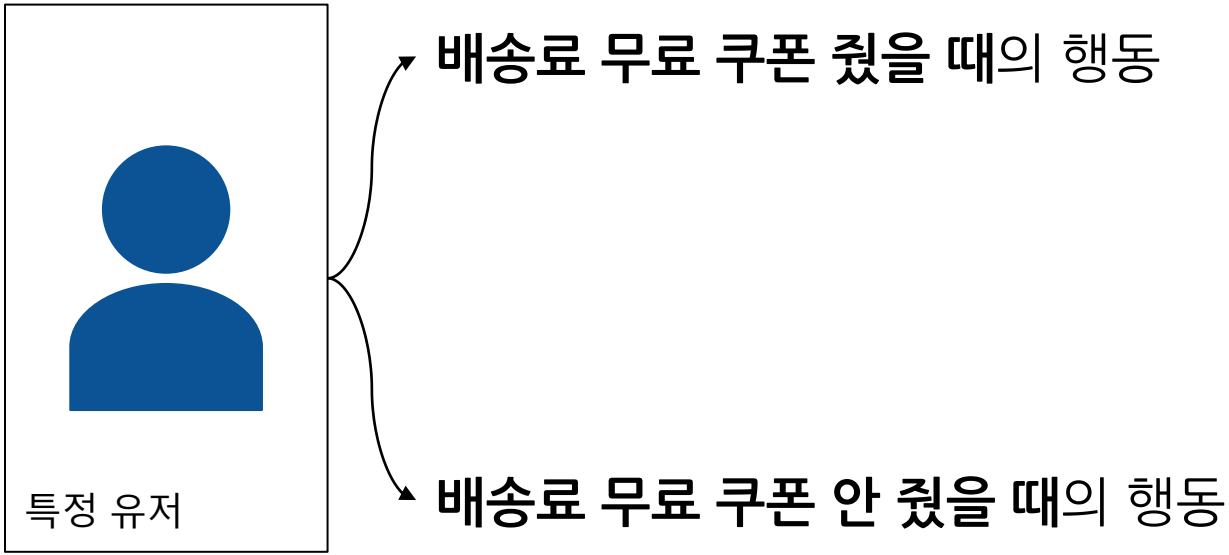
의사 결정 상황

경영진 曰: “배송료 무료 쿠폰, 그게 무슨 효과는 있어?”

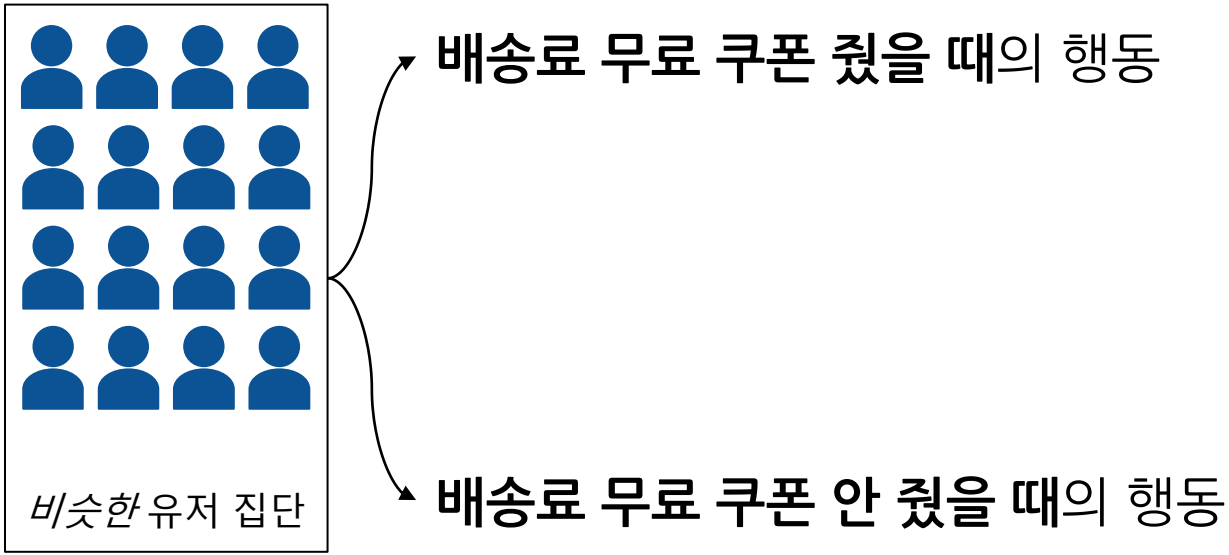
유용한 결과: 쿠폰 지금 전/후의 각종 성과 지표에 대한 예측

필요한 모형: 다양한 성과지표에 대한 쿠폰의 *효과*(인과관계!)

알고 싶은 (하지만 절대 알 수 없는) 것



(알고 싶은건 아니지만) 알 수 있는 것



Google Colab 실습

pandas, numpy가 설치된 환경 어디든 실습 가능

다음: 인과관계 분석 기초

2부: ML 예측과 인과관계 분석

인과관계 분석 기초

데이터와 의사결정 | 정종빈

인과관계분석(Causal inference)의 학풍

- Rubin-Neyman causal model
  - 통계 기반의 인과관계 분석법
  - 관찰되지 못한 결과(potential outcome)이 핵심
- Judea Pearl causality
  - 데이터/그래프 기반의 분석법
  - 원인과 결과 간의 관계에 대한 철학적 논쟁이 핵심
- 비교
  - 실증적으로 결과적인 계산/결론은 똑같음
  - 접근법/사고의 차이

“인과관계”란?

관찰되지 못한 결과(potential outcome)의 관점에서

- “키 크는 약”의 효과 → 이 약은 정말 “성장”의 원인인가?
  - 종빈이가 먹고 한 달만에 2cm 성장
  - 종빈이가 먹지 않고 한 달만에 1cm 성장
  - → “키 크는 약”은 종빈이에게 1cm 성장을 유발 (인과관계)
- 현실적인 문제:

  종빈이는 약을 먹거나 먹지 않을 수 있지만, 둘 다 할 수는 없다!

  즉, 인과관계 계산에 필요한 두 관측치 중 하나는 관찰 불가능
- (부족하지만) 직관적인 접근법
  - 전/후 관계(pre/post) → 인과관계?
  - 최대한 종빈이와 비슷한 사람 찾기

Average Treatment Effect

- 개인적인 효과(treatment effect)는 알 수 없지만
- 어느 집단에 대한 평균적인 효과는 측정 가능
  → Average Treatment Effect (ATE)

Average Treatment Effect

참가자	키 크는 약 처방? (Assignment)	약 없이 성장 (Control Outcome)	약 먹고 성장 (Treatment Outcome)
1	☑	-	2 cm
2	⊖	1 cm	-
3	☑	-	1 cm
4	⊖	1 cm	-

Average Treatment Effect

참가자	키 크는 약 처방? (Assignment)	약 없이 성장 (Control Outcome)	약 먹고 성장 (Treatment Outcome)
1	☑	-	2 cm
2	⊖	1 cm	-
3	☑	-	1 cm
4	⊖	1 cm	-

- 키 크는 약의 평균 효과(ATE)는 (2 + 1)/2 - (1 + 1)/2 = 0.5cm

인과관계 분석을 위한 중요한 가정

- Selection bias
  - 처방 여부(assignment)와 각 (potential) outcome 간의 상관관계가 없어야 함
- Stable Unit Treatment Value Assumption (SUTVA)
  - 한 참가자의 처방 여부가 다른 참가자에게 영향을 미치지 않음
  - 처방의 종류가 동일

Selection Bias: Example

유저	쿠폰 지급 여부	쿠폰 없을 때 결제 여부	쿠폰 받았을 때 결제 여부	쿠폰의 causal effect
1		0	0	0
2		0	0	0
3		1	1	0
4		1	1	0

- 실제 쿠폰의 효과(ATE)는 0

Selection Bias: Example

유저	쿠폰 지급 여부	쿠폰 없을 때 결제 여부	쿠폰 받았을 때 결제 여부	쿠폰의 causal effect
1	☑	0	0	0
2	⊖	0	0	0
3	☑	1	1	0
4	⊖	1	1	0

- 실제 쿠폰의 효과(ATE)는 0
- Selection bias 없는 쿠폰 지급(assignment)의 경우 측정된 ATE:
  (0 + 1)/2 - (0 + 1)/2 = 0

Selection Bias: Example

유저	쿠폰 지급 여부	쿠폰 없을 때 결제 여부	쿠폰 받았을 때 결제 여부	쿠폰의 causal effect
1	☑	0	0	0
2	☑	0	0	0
3	⊖	1	1	0
4	⊖	1	1	0

- 실제 쿠폰의 효과(ATE)는 0
- 쿠폰 지급 여부가 관심 결과(outcome)와 상관 관계가 있는 경우
  (예: “예측 결제 확률이 가장 낮은 유저에게만 쿠폰을 지급하자”)
  (0 + 0)/2 - (1 + 1)/2 = -1
- 실제로는 아무 효과가 없는 쿠폰이, 구매율을 낮춘다는 잘못된 결론!

Selection Bias: Example

유저	쿠폰 지급 여부	쿠폰 없을 때 결제 여부	쿠폰 받았을 때 결제 여부	쿠폰의 causal effect
1		0	0	0
2		0	0	0
3		1	1	0
4		1	1	0

- **Unbiased 쿠폰 지급이란? → 쿠폰 지급 여부와 관심 결과(outcome)의 상관관계가 없음, 즉:**
  - E[구폰 없을 때 결제 여부 | 쿠폰 지급 ✅] = E[쿠폰 없을 때 결제 여부 | 쿠폰 지급 ⛔]
  - E[구폰 받았을 때 결제 여부 | 쿠폰 지급 ✅] = E[쿠폰 받았을 때 결제 여부 | 쿠폰 지급 ⛔]
- **모순: 이를 판단하기 위해서는 potential outcome을 모두 알아야함**
- **Random assignment → 빈도주의적 입장에서 ATE의 기대치가 unbiased**

SUTVA: Example

- 택시앱에서 “무료승차권”의 효과를 측정하기 위해, random assignment로 유저 50%에게 “무료승차권” 지급 A/B test
- 실험 결과, “무료승차권” 받은 유저의 승차율이, 승차권을 받지 못한 유저보다 훨씬 높음
- 실제 무료승차권의 순수한 “효과”는 측정치보다 낮을 가능성이 큼. 왜?
- 한 유저 집단의 급격한 승차율 증가가 자연스럽게 다른 유저 집단의 승차율에도 영향을 미침 → SUTVA 위반

SUTVA: Example

- 이론적으로, n명의 실험 참가자가 있을 때, SUTVA가 위반되면, 각 참가자에 대한 potential outcome은 2개가 아니라 2^n으로 급속하게 증가!
  - 참가자 1, …, n-1의 assignment 여부에 따른 참가자 n의 각 outcome
- 현실적으로
  - 최대한 SUTVA를 위반하지 않는 범위 내에서 실험 설계
  - 위반하는 경우에 대해 실험 결과 해석에 주의
    (“효과의 최대치, 실제 효과는 실험 결과보다 작을 것”)

참고

- “Causal Inference” by Imbens and Rubin
- “Causality” by Judea Pearl

다음: 또 다른 상황

예산 제약 하에서의 이벤트 계획 및 의사결정

2부: ML 예측과 인과관계 분석

예산 제약 상황

데이터와 의사결정 | 정종빈

의사 결정 상황

경영진 日 | 유용한 결과 | 필요한 데이터/모형

의사 결정 상황

경영진 曰 | 유용한 결과 | 필요한 데이터/모형

경영진 曰:

“예산은 얼마나 되나?”

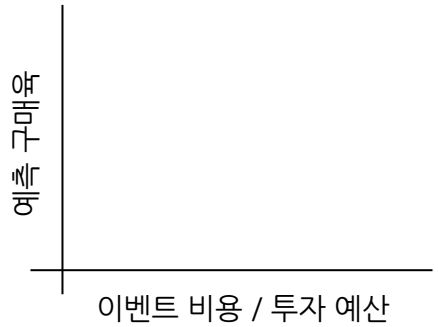
“한 5천만원 투자하면 구매율이 어느 정도 오르나?”

유용한 결과?

의사 결정 상황

경영진 日 | **유용한 결과** | 필요한 데이터/모형

유용한 결과:



필요한 데이터/모형?

의사 결정 상황

필요한 데이터/모형

- 이벤트 대상자 (현재 유저) 데이터 (`current_users.csv`)
- 쿠폰 지급 전/후 각 유저의 구매율 예측 모형
- 쿠폰 지급에 따른 비용 예측 모형
  - 고정 비용 제외 → “배송료 무료 쿠폰”의 비용은 실제 배송료 감당
  - 배송료 = f(basket, region)
  - 구매금액(basket) 예측 모형
  - 현재 유저 데이터의 미래 대표성에 대한 가정

Google Colab 실습

pandas, numpy, sklearn, xgboost, seaborn 활용

중요한 질문

- 전제/가정 파악 (예: “예산”의 정당성)
  - 왜 “5천만원”?
  - 비즈니스 목적/전략에 따라 합당한 이유가 있을 수 있음
  - 반면 합리적이지 않은 여러가지 이유도 있을 수 있음
  - DS의 중요한 역할 중 하나: 다양한 전제/가정에 의문을 제기하여 의사 결정 과정에 투명성을 더하는 것
- 목적/성과지표 구체화 (예: 왜 “구매율”?)
  - 성과지표(metric)의 미묘/복잡한 차이 파악
  - 외부에서 성과지표가 주어졌을 때는, 이에 대해 명확하게 정의/파악
  - 프로젝트의 궁극적 목적/성공의 척도에 집중

다음: ML 인과관계 분석

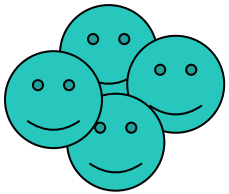
ML을 이용한 인과관계 분석의 한계와 주의점

2부: ML 예측과 인과관계 분석

ML 인과관계 분석

데이터와 의사결정 | 정종빈

예측 모형을 이용한 ATE 계산

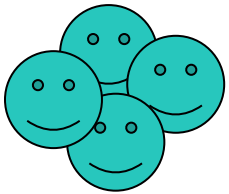


age: X1
gender: X2
...

coupon: True
checkout: 9%

쿠폰을
주지 않았다면 어떻게 됐을까?

예측 모형을 이용한 ATE 계산



age: X1
gender: X2
…

coupon: True
checkout: 9%

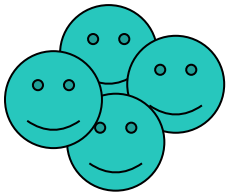
age: X1
gender: X2
…

coupon: False
checkout: 6%

쿠폰을 받지 **않았지만**,
다른 면에서 비슷한*
유저 집단을
찾아서 비교

* “비슷함”의 정확한 정의는, 사용하는 ML 예측 모형에 따라 결정 됨

예측 모형을 이용한 ATE 계산



age: X1
gender: X2
...

coupon: True
checkout: 9%

→ 쿠폰을 받지 않는다면?:

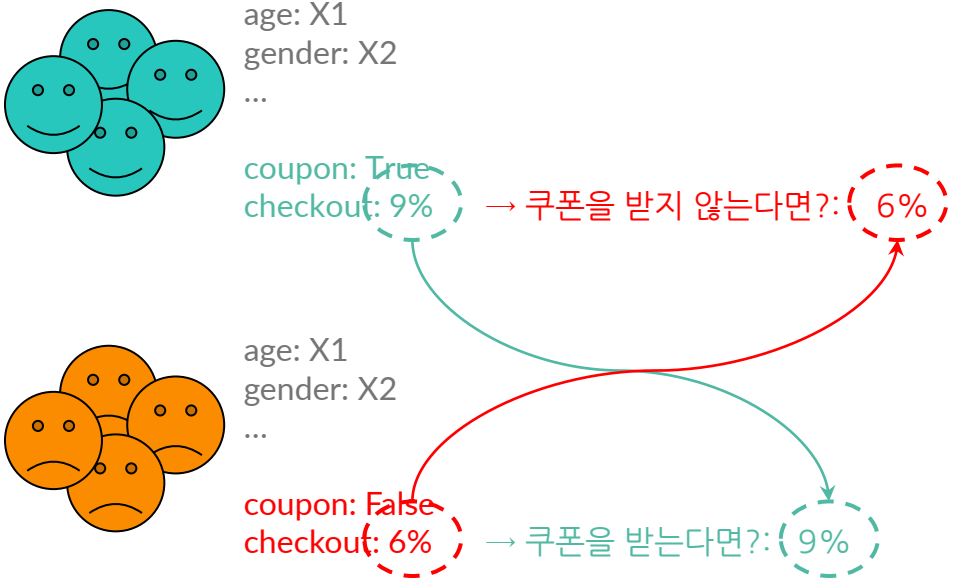
age: X1
gender: X2
...

coupon: False
checkout: 6%

→ 쿠폰을 받는다면?:

예측 모형을 이용한 ATE 계산

Assuming ignorability [response surface modeling]

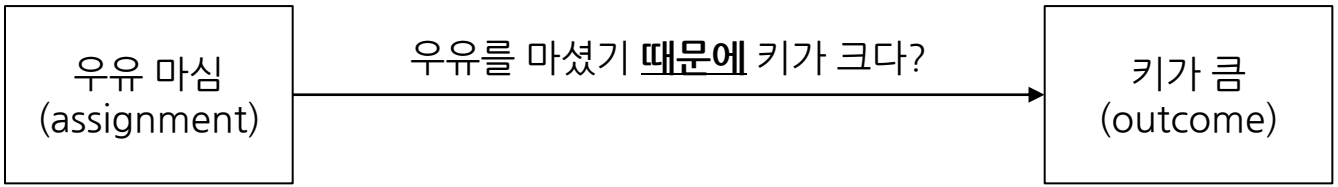


예측 모형을 이용한 ATE 계산

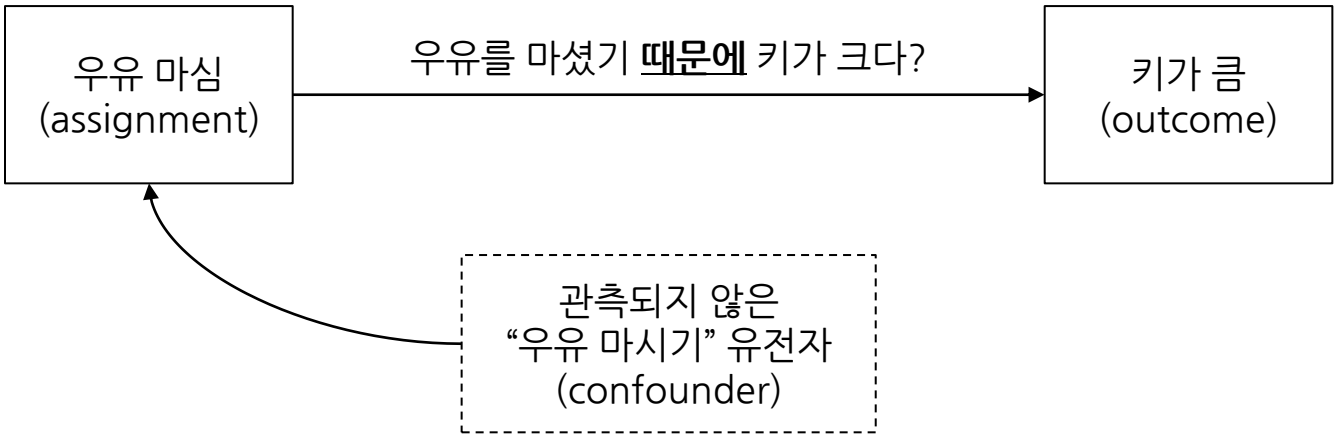
Assuming ignorability [response surface modeling]

- “비슷한” 유저는 [assignment, (potential) outcome]의 기대값이 같다
  - 위반 사례: 관찰하지 않은 “미지의 변수(confounder)”가
    - 쿠폰을 받는지 여부(assignment)와 상관관계가 있고
    - 관심 결과(outcome)와도 상관관계가 있을 경우
  - 두 가지 중 하나만 성립할 경우, 여전히 ignorability 만족
- 유저에 대한 정보가 많이 있을 수록 만족할 가능성이 큼
- Selection bias와 마찬가지로, 만족 여부 판단 불가능
- 예: “우유를 많이 마시는 사람이 키가 크다?”

Ignorability 위반의 예

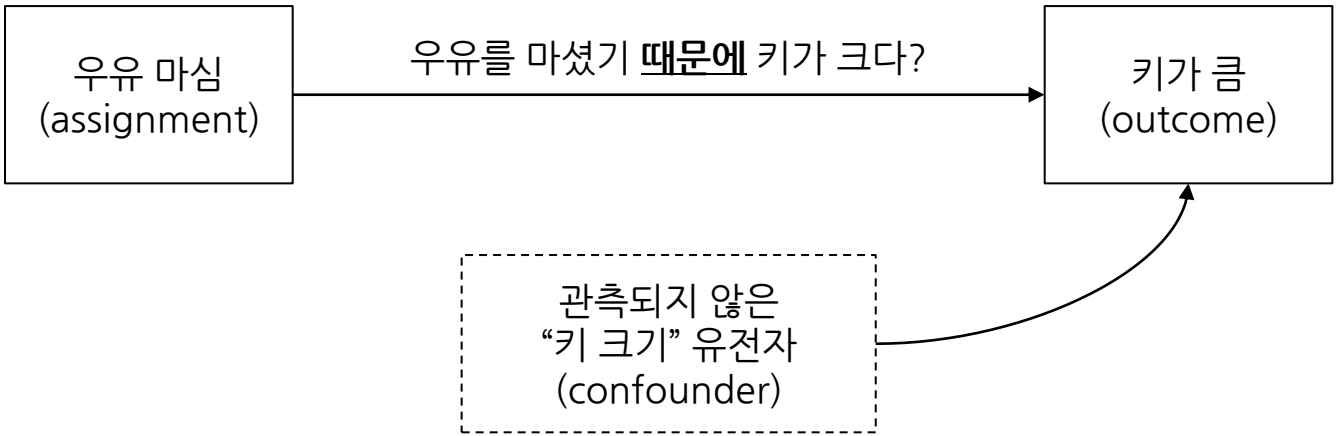


Ignorability 위반의 예



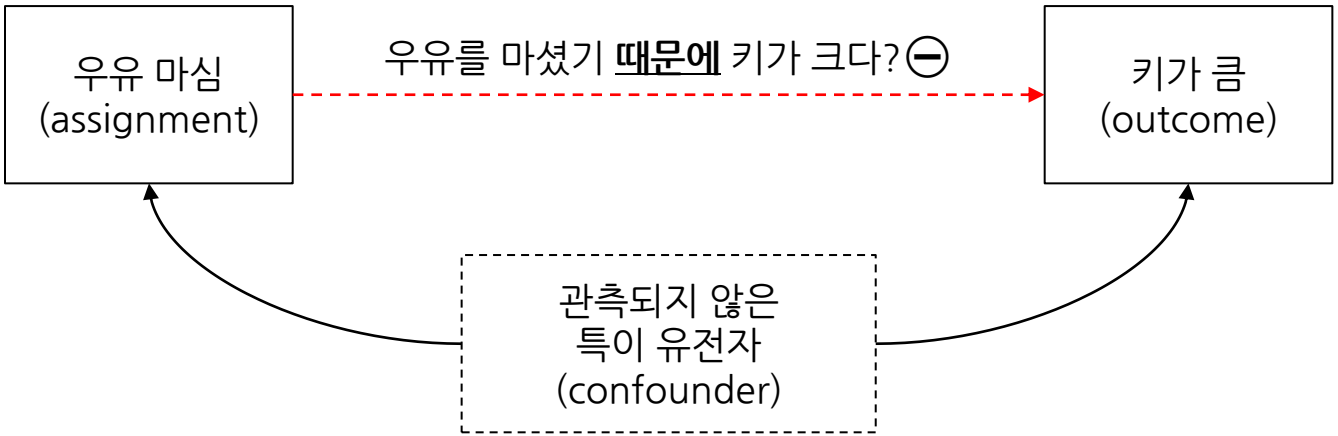
어떤 이들에게 있는 “우유를 더 마시게 하는 유전자”를 관측하진 못했어도,
그 유전자가 참가자들의 키와 상관관계를 갖지 않는다면 인과관계 분석 결과에는 영향이 없음

Ignorability 위반의 예



어떤 이들에게 있는 “키를 크게 하는 유전자”를 관측하진 못했어도,
그 유전자가 우유 섭취량과 상관 관계를 갖지 않는다면 인과관계 분석 결과에는 영향이 없음

Ignorability 위반의 예



우유 섭취량/키와 동시에 상관관계를 갖는 어떤 특이 유전자가 존재하는 경우
이에 대한 각 참가자의 정보를 관측하지 못했다면 ignorability 위반 → 인과관계 분석 결과 오류

추가 주의 사항

- Treatment assignment 이후에 관측 된 정보 사용 ⛔
  - 결제 금액/(쿠폰 적용 전) 배송료는 최종 결제여부와 상관관계 있음
  - 이를 feature로 포함할 경우, checkout에 대한 예측 성과 ⬆
  - 하지만 이는 배송료 무료 쿠폰 지급 이후 관찰한 항목이기 때문에, 이를 포함 할 경우 coupon에 대한 인과관계 분석 불가능
- 사용 예측 모형이 암시하는 바에 대해 주의
  - 예: 선형 모형을 사용할 경우 interaction의 부재가 시사하는 바?
  - 대부분 최대한 유연하면서 예측 성과가 우수한 모형(XGBoost)을 사용하면 크게 문제되지 않음
  - 반면, 문제에 대한 전문적 견해/이해가 있을 경우, 이를 반영할 수 있는 모형 사용이 도움이 될 수 있음

ML 인과관계 분석

- 아직 활발하게 연구 되고 있는 분야
- 딱히 “정답”이나 “공식”은 없음
  - 다양한 가정 및 데이터의 기원에 대한 이해 없이 “가져다 쓰기”는 잘못 된 결과에 이를 위험이 큼
- 다양한 가능성, 가정, 문제에 대한 이해를 바탕으로 판단
- 비현실적 가정은 피할 수 없음
  - 현실을 반영하는 복잡함 vs. 단순하지만 “적당”하고 유용함
- 목적은 “완벽함(verisimilitude)”이 아니라 “투명/명료함(clarity)”

다음: 마지막 상황

수익 극대화

2부: ML 예측과 인과관계 분석

수익/이익 극대화

데이터와 의사결정 | 정종빈

의사 결정 상황

경영진 日 | 유용한 결과 | 필요한 데이터/모형

의사 결정 상황

경영진 曰 | 유용한 결과 | 필요한 데이터/모형

경영진 曰:

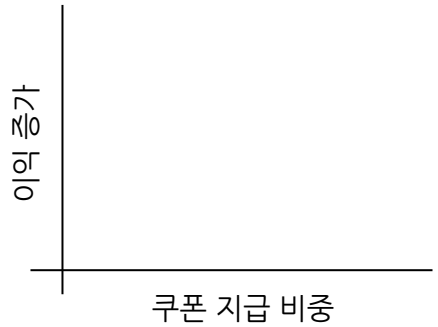
“상장도 했고 … 성장/구매율 보다는 이익에 초점 맞춰보자.”

유용한 결과?

의사 결정 상황

경영진 日 | 유용한 결과 | 필요한 데이터/모형

유용한 결과:



필요한 데이터/모형?

의사 결정 상황

필요한 데이터/모형

- 예산 제약 계산에 사용된 데이터/모형/가정 +
- 쿠폰 지급에 따른 이익 = 수익 - 비용 계산

Google Colab 실습

pandas, numpy, sklearn, xgboost, seaborn 활용

중요한 질문

- 배송료 무료 쿠폰 이벤트의 비용 = 배송료?
  - 유저에게 부과하는 배송료 ≠ 기업이 부담하는 배송료
  - 이외의 비용은 없는지?
- 예상 이익의 불확실성
  - 유저 기반 변화
  - Basket 예측 모형의 오차
  - Network effect (SUTVA 위반)

다음: 종강

데이터 사이언티스트로 행복하게 살기

DS로 행복하게 살기

종강

데이터와 의사결정 | 정종빈

DS의 삶의 근본 가치

Principles to live by (via Ron, Sharad, and Ramesh)

- 호기심과 늘 배우려는 자세
- 정밀/정확/명료함 (precision and clarity)
- 유쾌하고 합리적 의심
- 목적에 대한 초점과 자신감
- 윤리

호기심과 늘 배우려는 자세

- “내 일”에 대해 방어적이지 않도록 노력!
- “절대적 권위자” ⊖ → 늘 새로 배울 것 투성이

정밀/정확/명료함 (precision and clarity)

- 대충 얼버무리고 넘어가면 나중에 고생
- 정확하지 않은 것은 “모른다”고 인정
- 정확하지 않아도 괜찮은 것은 의도적/공개적으로 “이것은 정확하지 않은 것”을 분명하게 함
- 가장 중요한 것은 투명성/명료함 (clarity)

유쾌하고 합리적 의심

- “내가 한 건데 …”, “유명한 … 가 …”에 속지 말 것
- 새로운 문제/해법을 당면했을 때 한 발 물러서서 최대한 의심의 눈초리로 꼼꼼히 살펴 볼 것
- 직관적이지 않고 놀라운 일에 현혹되지 말고,
  직관이 생길 때까지 의심/깊이 있게 탐구

하지만 …

- 이 모든 과정에서 유쾌/즐겁게 임할 것!
- 만사에 불평/의심하는 사람으로 비추기 쉽상

목적에 대한 초점과 자신감

- 호기심, 정확도, 의심을 쫓다가 궁극적 목적을 잃기 쉬움
- 때로는 불만족스러운 상황에서의 도전이 필요
- “무모한” 도전이 아닌, 목적이 분명한 의도적 도전
- 불확실성에 대한 인정

윤리

- 생각지 못한 사이, 의도치 않게 많은 영향력을 갖기 쉬움
- 늘 본인의 “선”이 무엇인지 미리 고민
  - 돈을 벌기 위해 나는 어떤 일까지 할 것인가? 얼마까지?
  - 커리어를 위해 무엇까지 포기 할 수 있는/없는 가?
  - …
  - 닥쳤을 때는 늘 쫓기고 생각을 정리할 시간이 없어 실수하기 쉬움
- 본인의 (개인적인) 가치관을 일관적으로 지키기 위한 구체적인 노력이 항시 필요

감사합니다