Ⅲ 재현데이터

본 장에서는 기존 가명·익명처리 기법의 대안으로 재현데이터를 제시한다. 재현데이터는 정형데이터뿐 아니라 이미지, 텍스트 등의 다양한 비정형 데이터에도 적용되고 있다. 본 장에서 재현데이터의 역사와 특징에 대해 논하고 재현데이터의 통계학적 및 딥러닝 기반 생성 방법론을 설명한다. 더불어 현재 국내외 재현데이터의 활용사례도 살펴보겠다.

1. 주요 방법론

재현데이터는 실제로 관측된 데이터(Real Data)를 생성하는 모형 혹은 모집단이 존재한다고 가정하고, 통계적 방법이나 기계학습 방법 등을 이용해 추정된 모형에서 새롭게 생성한 모의 데이터(Simulated Data)이다. 개인정보 노출을 막는 효과적인 방법이며 민감정보를 활용할 수 있어 연구자들이 재현데이터를 활용하여 보다 세밀한 분석을 진행할 수 있다는 이점을 갖는다. 다양한 규제로 인해 데이터의 적극적 활용에 제한이 가해진 현 상황에서 재현데이터는 노출 제어의 효과적인 방법론으로 부상하고 있다.

재현데이터는 하버드대학교 통계학과 루빈 교수가 미국 정부 기관 프로젝트로 빈곤층에 대한 과소평가와 같은 문제들을 해결하는 연구 과정에서 처음 제시하였다.[13] 루빈 교수는 모집단에서 관측되지 않은 자료를 결측값으로 간주해 다중 대체법(Multiple imputation)을 적용했고 이를 반복적으로 샘플링하여 다수의 데이터셋을 생성했는데, 이렇게 만들어진 데이터셋을 재현데이터(Synthetic data)로 지칭하였다. 이후 재현데이터 관련 연구는 다양한 방향으로 확장되어 왔다.[14]

13) Rubin, D. B.(1993)
14) 유성준·박나리(2020)

가. 분류

재현데이터는 재현 범위와 방식에 따라 완전 재현데이터(Fully Synthetic Data)와 부분 재현데이터(Partially Synthetic Data)로 분류할 수 있다.

1) 완전 재현데이터(Fully Synthetic Data)

완전 재현데이터는 측정된 실제 데이터를 기반으로 모든 변수를 재현하여 가상으로 생성된 데이터를 의미한다. 정보보호 측면에서 가장 강력한 보안성을 가지며 제공되는 데이터가 실제 데이터를 포함하지 않으므로 민감정보가 노출되지 않는 구조이다. 루빈 교수가 최초로 정의한 개념으로 다중대체 기법을 기반으로 하고 있다. 완전 재현데이터는 보통 민감정보이거나 공개 불가능한 내용을 포함하고 있는 부분을 모두 결측치(Missing value)라고 간주하고 다중 대체(Multiple imputation)[15]를 적용한다. 다음으로, 완성된 모집단을 모형에 적합하여 무작위 추출(Random sampling)한 후 재현데이터를 생성하는 방식을 취한다. 관측치별로 정보 노출 방지의 안정성을 확보할 수 있고 올바른 모형을 사용하면 데이터 구조가 원본과 비슷하게 유지될 뿐만 아니라 마스킹과 같은 다른 가명·익명처리 기법보다 정보 손실이 적다는 장점이 있다.

2) 부분 재현데이터(Partially Synthetic Data)

부분 재현데이터는 공개하려는 변수 중 일부만을 선택하여 재현데이터로 대체한 데이터를 말하며 Little[16]이 최초로 제안했다고 알려져 있다. 완전 재현의 경우 생성 데이터의 차원이 과도하게 커져 과적합이 일어나거나 변수 중요도가 반영되지 않을 수 있는데, 부분 재현을 통해 이러한 이슈들을 해결할 수 있다. 부분 재현 시 대체되는 변수들은 보통 민감변수(Sensitive variable) 혹은 식별 변수(Identifiable variable)가 되지만 사용자 임의대로 또는 분석 목적에 따라 선택 가능하다. 일반적으로 노출위험이 높거나 공개 불가능한 정보들을 임의로 선택하여 그 변수에서만 값을 대체하기 때문에 정보 손실이 적고 데이터 구조가 이전과 비슷하게 유지될 수 있다.

---

15) Li, P., Stuart, E. A., and Allison, D. B.(2015)

16) Little, R. J.(1993)

완전 재현데이터와 또 다른 차이점은 생산 가능한 데이터셋의 수가 다르다는 것이다. 완전 재현데이터는 데이터를 무한정 생산할 수 있지만, 부분 재현데이터는 선택한 일부 변수만 채우기 때문에 기존 데이터와 동일한 크기의 데이터를 생산하게 된다.

또 다른 재현데이터의 범주로서 복합 재현데이터(Hybrid Synthetic Data)를 들 수 있다. 이는 완전 재현과 부분 재현 방법 둘 다 차용해서 생성하는 데이터를 묶어서 부르는 단어로, 사실 복합 재현데이터의 정의와 범위는 아직 정확히 정해지지 않고 있다. 다만, 복합 재현데이터는 일부 변수들의 값을 재현데이터로 생성한 후 남은 실제 데이터를 이용해 또 다른 변수들의 값을 다시 재현데이터로 생성하는 방법을 지칭하거나, 이미지 데이터에서 배경은 실제 이미지를 쓰고 사물은 재현된 합성 이미지를 사용해 새로운 이미지를 생성하는 방식을 일컫기도 한다.

나. 생성이론

재현데이터는 다음과 같은 개념적 단계를 통해 생성한다.[17]

(1) 데이터 스키마(Data schema) [18] 정의: 원데이터의 구조, 데이터 유형, 관계 및 데이터에 적용되는 제약 조건 또는 규칙을 정하고 도식화한다.

(2) 원데이터의 통계적 특성 확인: 원데이터가 가지는 분포, 상관관계 및 기타 통계적 패턴 등의 정보를 미리 확인한다. 이 정보는 실제 데이터를 분석하거나 데이터에 대한 사전 지식을 조사하여 기록해둘 수 있고 추후 재현데이터 생성 과정에서 활용될 수 있다.

(3) 적합한 재현데이터 생성 방법 선택: 재현데이터를 생성하는 알고리즘이 다양하므로 데이터 스키마 및 통계적 특성에 따라 생성 방법론 채택이 달라져야 한다. 따라서 어떤 방법을 사용하여 어느 유형의 재현데이터를 생성할지 결정하는 과정이 필요하다.

(4) 재현데이터 생성(Generation): 선택한 방법을 사용하여 재현데이터를 생성하고, 그 과정에서 원데이터의 통계적 특성과 일치하도록 보장하는 단계이다.

(5) 재현데이터의 품질 평가(Evaluation): 생성된 재현데이터를 원데이터와 비교·평가하

17) https://gretel.ai; https://limoss.london/how-does-sdc-work

18) 데이터베이스를 구성하는 데이터 개체(Entry), 속성(Attribute), 관계(Relationship) 및 데이터 조작 시 데이터값들이 갖는 제약 조건 등에 관해 전반적으로 정의함을 나타내는 용어임. 즉, 데이터베이스를 어떻게 설계할지에 대한 계획을 짜서 구조와 제약 조건을 정하는 것임

는 과정이다. 통계적 검정, 시각화 또는 원데이터와의 단변량 및 다변량 비교를 통해 평가할 수 있다. 품질이 기준에 미치지 못하거나 필요한 특성을 충분히 반영하고 있지 않다면, 데이터 스키마를 수정하거나 다른 생성 방법을 사용하여 위 과정을 반복할 수 있다. 동시에 민감한 개인정보의 노출위험이 충분히 제어되었는지도 같이 평가해야 한다.

(6) 재현데이터 배포 및 사용: 재현데이터의 품질이 검증되면, 머신러닝 학습 모델 훈련 및 테스트, 데이터 증강, 민감한 정보를 제외한 데이터 공유 등 다양한 용도로 사용될 수 있다.

위의 단계 (4)에서 재현데이터의 생성은 Python과 R 등의 다양한 소프트웨어 툴을 이용하여 구현할 수 있는데 이 중 오픈소스로 공개된 솔루션도 있고 상용화되어 내부적으로 어떤 모형·알고리즘을 사용하는지 알 수 없는 경우도 있다. 〈표 Ⅲ-1〉은 재현데이터 생성 소프트웨어들을 조사한 결과이다. 표에 포함되지 않은 상용 소프트웨어들도 많은데 2022년 기준으로 재현데이터 생성 서비스를 하는 크고 작은 업체는 대략 100여 개 이상으로 추산된다.[19]

〈표 Ⅲ-1〉 재현데이터 생성 관련 소프트웨어

소프트웨어	관련 홈페이지 URL	내용
synthpop	https://www.synthpop.org.uk/	분류 회귀모형을 사용하여 재현데이터에 대한 변수를 생성함. 정교한 샘플링 디자인을 필요로 하거나 가구 및 구성원 정보 등과 같은 계층 또는 클러스터 구조를 가진 데이터를 처리하는 기능은 없으나 편의성이 높음
sms	https://cran.r-project.org/web/packages/sms/index.html	주어진 영역 내 매크로 데이터로부터 마이크로 데이터를 시뮬레이션하는 기능을 제공함. 계층적 구조의 데이터 처리는 불가능하나, Simulated Annealing을 단순화하여 제한된 영역에 대한 설명을 최적화하는 기능이 존재함
simPop	https://cran.r-project.org/web/packages/simPop/index.html	주체가 가진 속성값에 따라 다르게 적용되는 정책의 거시적인 효과를 예측하기 위한 복잡한 구조의 데이터 재현에 매우 유용함. 가구와 가구 구성원 정보 등 계층적 구조 처리가 가능함. IPF와 SA를 사용한 통계량 조정, 로지스틱 회귀를 통한 모델링 기능을 제공함

19) 업체들의 목록은 다음의 링크를 참고함(https://elise-deux.medium.com/new-list-of-synthetic-data-vendors-2022-f06dbe91784)

〈표 Ⅲ-1〉 계속

소프트웨어	관련 홈페이지 URL	내용
PoPGen	https://www.mobilityanalytics.org/popgen.html	Arixona State University의 SimTRAVEL Research Initiative에서 개발되었으며, 상대적으로 전수 정확도가 높은 반복비율갱신(Iterative Proportional Updating) 알고리즘으로 전수 인구 데이터 생성이 가능함
TRANSIMS	https://sourceforge.net/projects/transims/	미국 Los Alamos National Laboratiry의 연구원이 개발한 운송 분석 시뮬레이션 시스템임. 인구조사 마이크로 데이터를 기반으로 재현데이터를 생성함
Synthia	https://synthia-dataset.net/	비영리 연구기관인 RTI에서 개발한 웹 기반 재현데이터 생성 프로그램으로, 사용자 정의 변수를 사용하여 사용자가 정의한 학습 영역에 대한 재현데이터를 생성함
SDV	https://sdv.dev/	datacebo에서 관리하는 재현데이터 생성 공개소스코드로, 개별, 관계형, 시계열 데이터에 대한 재현데이터를 생성함. 생성된 데이터에 대한 평가 및 시각화 모듈도 제공함. 소스코드 공유 사이트 github에서 1,500개 이상의 star를 받을 정도로 높은 대중적 인지도를 가지고 있음
DataSynthesizer	https://github.com/DataResponsibley/DataSynthesizer	차등보호(Differential privacy) 기술을 활용한 재현데이터 생성 오픈소스코드로, Django를 활용한 UI 앱도 제공함
Gretel	https://gretel.ai/	미국 샌디에이고에 위치한 재현데이터 생성 스타트업으로 LSTM, DGAN, Diffusion 등의 딥러닝 방법을 활용하여 관계형, 시계열, 비정형, 이미지 재현데이터를 생성하는 서비스를 제공함.

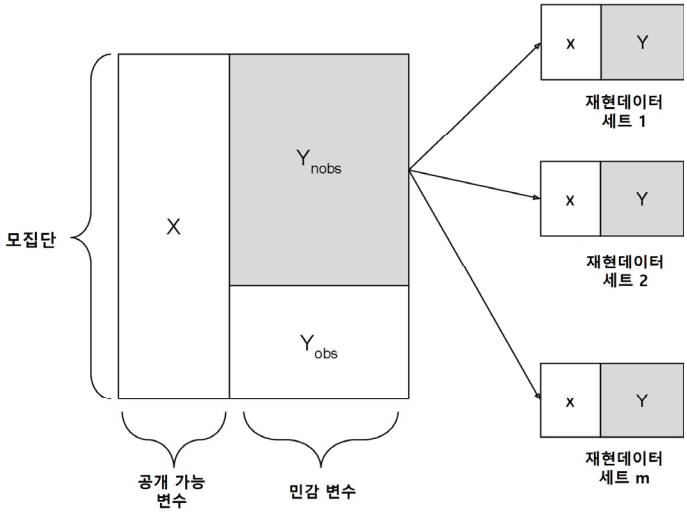
자료: 김승현(2020) 및 저자 조사

이제 대표적인 재현데이터 생성이론 몇 가지를 살펴보도록 하겠다. 재현데이터 생성 방법은 다양하게 발전되고 있고 아직 모든 경우에 잘 작동하는 표준화된 방법은 없다. 여기서는 통계적인 방식으로 재현데이터를 생성하는 이론을 주로 소개하고 추가로 최근 각광받고 있는 딥러닝 방법인 GAN(Generative Adversarial Network) 기반 생성 방법론도 간략하게 언급한다.

1) 모수적 다중대체 모형을 이용한 생성 방법론[20)]

통계학에서 대체(Imputation)란 결측된 값을 적절한 값으로 바꾸는 것을 뜻하며 경우에 따라 하나의 값으로 바꾸는 단순대체를 사용할 수도 있고, 대체되는 값의 불확실성을 감안해 여러 개의 값으로 제공하는 다중대체를 사용하기도 한다. 대체에 대한 통계적 내용은 방대해 여기서 모든 내용을 다루기 어렵기 때문에 여기서는 다중대체를 중심으로 설명하겠다. 모수적 다중대체는 사후예측분포(Posterior predictive distribution)를 추정하여 결측값을 생성하며, 이 과정을 반복하여 여러 개의 대체값을 만들어내는 방식을 취한다. 대체를 이용한 재현데이터의 생성 방법을 소개하기 위해 아래 〈그림 Ⅲ-1〉을 이용해 데이터 내부 자료 구조를 나누어 설명해 보겠다.

〈그림 Ⅲ-1〉 데이터 분할



자료: Dandekar, A., Zen, R. A., and Bressan, S.(2009)

그림의 왼쪽 X는 민감하지 않은 공개 가능한 변수이고, Y는 노출을 최소화해야 하는 민감변수들을 나타낸다. 추가로 Y는 Y_{obs}과 Y_{nobs}로 나뉘는데, 전자는 관측된 값들이고 후자인 Y_{nobs}는 수집되지 못한 결측치로 정의된다. 따라서 수집된 혹은 관측된 전체 데이터

20) 박민정(2020)

는 $D = \mathrm{X}, \mathrm{Y}_{\mathrm{obs}}$이다. 재현데이터는 사후예측분포인 $P(Y_{nobs}|X, Y_{obs}) = P(Y_{nobs}|D)$를 추정하고 이로부터 값을 추출하여 Y_{nobs}를 채우는 방식으로 생성된다. 이를 반복하면 다른 값들의 Y_{nobs}로 채울 수 있어 다중대체의 효과를 지닌다. 채워진 Y_{nobs}값들은 재현데이터인 Y_{syn}으로 이해할 수 있으므로 사후예측분포 $P(Y_{syn}|D)$는 재현데이터 생성기를 의미한다고도 할 수 있다.

위의 다중대체 방법은 맥락에 따라 몇 가지 다른 형태로 변형될 수 있다.

(1) 만약 위 그림에서 Y_{obs}의 공개가 불가하다면 생성기인 $P(Y_{syn}|D)$를 이용하여 Y_{obs}, Y_{nobs} 둘 다 대체된 재현데이터 (X, Y_{syn})를 생성할 수 있다. 이는 앞서 소개한 부분 재현데이터에 해당한다.

(2) 만약 위의 (1)에 추가로 X의 공개 역시 불가하다면 확장된 재현데이터 생성기인 $P(X_{syn}, Y_{syn}|D)$를 추정하고 이로부터 재현데이터 (X_{syn}, Y_{syn})를 생성할 수 있다. 이는 완전 재현데이터에 해당한다.

재현데이터 생성을 위한 대체에서 사후예측분포를 추정하는, 즉 생성기를 만드는 방법은 다양한데 기본적으로 다변량 데이터에 대해 분포추정을 할 수 있는 모든 모형이 사용 가능하다. 여기서는 모수적 방법에 국한하여 결합분포를 이용하는 방법과 각 주변분포에 대해 순차적으로 회귀모형을 적용하는 방법을 소개한다.

가) 결합모형을 이용한 분포 추정[21)]

생성하고자 하는 변수들이 벡터이고 이를 잘 설명할 수 있는 다변량 분포가 있다면 사후예측분포를 통해 재현변수들을 한 번에 벡터로 생성해 낼 수 있다. 만약 민감변수 벡터를 $Y_{nobs} = (Y_1, \cdots, Y_n)$이라고 하면 이 변수들의 분포를 동시에 고려한 결합분포(Joint distribution)를 추정해 이로부터 대체 값들을 벡터로 추출하는 것이다. 이는 수식

$$P(Y_{nobs}|D) = P(Y_{1,} Y_{2,}.., Y_n | X, Y_{obs}) = \int P(Y_{nobs}| X, Y_{obs}, \theta) P(\theta | X, Y_{obs}) d\theta$$

으로 표현할 수 있다. 여기서 $P(\theta | X, Y_{obs}) = P(\theta | D)$는 데이터가 주어졌을 때 θ의 사후

21) 김정연·박민정(2019); 박민정·김항준(2016)

분포이다. θ는 변수들의 추정된 모수, 즉 모집단의 특성을 나타내는데, 예를 들어 $Y_{nobs} = (Y_1, \cdots, Y_n)$가 다변량 정규 분포를 따른다고 하면 θ는 평균 벡터와 공분산 행렬이 된다. 이 사후분포는 베이지안 통계의 방법론을 따라 추정하는 것이 보통이다. 다음으로 $P(Y_{nobs} | \theta, X, Y_{obs})$는 데이터와 모수 θ가 주어졌을 때의 조건부 다변량 분포이다.

실제 위의 수식을 이용할 때는, 먼저 추정된 $P(\theta | X, Y_{obs})$로부터 하나의 θ'을 추출한 후 이를 조건부 분포인 $P(Y_{nobs} | \theta', X, Y_{obs})$에 넣고 $Y_{nobs} = Y_{syn}$을 생성하는 과정을 반복한다.

그러나 결합분포를 사용하는 이 같은 방식은 생성할 변수의 종류가 다양하고 차원 n이 커짐에 따라 결합확률분포를 추정하기가 어려워지고 차원의 저주(curse of dimensionality)[22]로 인해 정확성이 떨어지기 때문에 아래 설명할 조건부 모형을 이용한 순차회귀 방법이 보다 널리 사용된다.

나) 순차회귀를 이용한 분포추정

이 방법은 개별 주변변수에 대해 순차적으로 회귀모형 혹은 다른 지도학습모형을 적용하는 방법으로 변수들이 서로 다른 형태(이산형, 범주형, 연속형)를 가지거나 한꺼번에 다변량 분포로 추정하는 것이 어려울 때 사용할 수 있으며, 특히 변수들이 순차적인 관련성이 있을 때 잘 작동한다. 실제 적용할 때에도 조건부 확률분포를 변수 중요도 순서에 따라 추정하여 사용할 수 있고, 개별 변수의 사후예측분포를 순차적 혹은 연쇄적으로 추정하기 때문에 직관적이고 적용이 쉽다. 통계학에서 순차회귀 다중대체는 Sequential Regression Multiple Imputation(SRMI) 혹은 Multiple Imputation by Chained Equations(MICE)라고 불린다.

순차회귀를 예를 들어 살펴보기 위해 어떤 데이터 D가 주어졌을 때 세 변수의 결합확률분포 $P(W, Y, Z | D)$를 추정하는 것이 목표라고 하자. 만약 세 변수들 사이에 적절한 인과관계가 존재하거나 중요도 또는 상관관계가 있어 $Z \rightarrow Y \rightarrow W$의 순서대로 변수들을 정렬할 수 있다면, 조건부 확률의 성질을 이용해

22) 차원의 저주란 차원이 증가하면서 모형의 성능을 동일하게 유지하기 위해 필요한 학습데이터 수가 기하급수적으로 증가하는 현상을 말함

$P(W, Y, Z|D) = P(W|Y, Z, D)P(Y, Z|D) = P(W|Y, Z, D)P(Y|Z, D)P(Z|D)$

로 쓸 수 있고, 추가로 Y가 주어졌을 때 W와 Z가 독립이라면

$$P(W, Y, Z|D) = P(W|Y, D)P(Y|Z, D)P(Z|D)$$

와 같이 더욱 단순화할 수 있다.

여기서 모수적 순차회귀 다중대체를 위해 $P(Z|D)$, $P(Y|Z, D)$, $P(W|Y, Z, D)$의 분포를 순서대로 추정했는데, 이 분포들은 모두 일변량의 조건부 분포이므로 회귀모형을 포함한 적절한 지도학습모형을 이용하여 추정할 수 있다. 다시 말해 다변량 분포의 추정 문제를 쪼개어 다수의 일변량 분포 추정으로 전환한 셈이다. 추정된 조건부 분포들이 주어지면 개별 변수들을 동일한 $Z \rightarrow Y \rightarrow W$ 순서대로 하나씩 랜덤하게 추출할 수 있고, 이를 반복하면 원하는 수 만큼의 재현데이터를 생성할 수 있다.

원칙적으로 W, Y, Z의 순서는 인과관계에 기반하여 정해야 하지만 현실에서는 인과관계가 명확하지 않은 경우가 많기 때문에 상관성이 높은 순서로 정하기도 한다. 따라서 이 기법에서 순서의 결정에는 자의적인 면이 있다.

만약 변수 간의 순서가 존재하지 않는다면, 개별 변수에 대해 해당 변수를 제외한 나머지 변수들과 D를 설명변수, 즉, 조건부로 두고 분포추정을 할 수 있으며 이때 변수들의 순서는 무시하면 된다.

2) 비모수적 다중대체 모형을 이용한 생성 방법론[23)]

실제 재현데이터를 생성할 때 모수적으로 분포를 추정하는 것이 적절하지 않거나 어려운 경우가 많다. 혼합분포(Mixture distribution)를 활용하여 모수적 접근 방식의 한계를 어느 정도 극복할 수도 있으나 이러한 시도는 제한적이다.[24)] 이런 경우 비모수적인 방법을 이용할 수 있다. 비모수적 방법은 모집단에 대한 특정 분포 가정을 하지 않는 통계적 방법으로 여기서는 재현데이터 생성에 사용되는 몇 개의 기법들을 소개한다.

먼저 소개할 기법은 CART(Classification And Regression Tree)이다. CART는 주어진 여

23) 박민정(2020)

24) Chib, S.(1996)

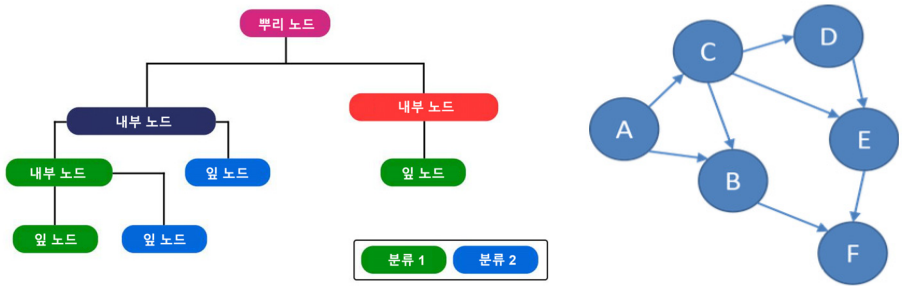
러 설명 변수에 기반하여 분할 규칙에 따라 의사결정나무를 만들고 반응 변수의 값을 예측하는 비모수적 방법이다. 각 분할은 뿌리노드(Root node)에서 시작해 각 변수별로 이진구조(Binary)로 쪼개는 방식을 순차적으로 적용해 진행되며 마지막에 말단노드(Terminal or leaf node)를 만든 후 종료된다. 분할의 기준은 보통 불순도(Impurity)[25]이다. CART는 연속형과 범주형 변수가 섞여 있는 데이터에 적합할 뿐만 아니라 분석 결과가 나무(Tree) 구조로 표현되어 이해가 쉽기 때문에 재현자료 생성에 자주 사용된다. 통계 소프트웨어 R의 패키지 synthpop 또한 method 옵션을 따로 지정하지 않으면 디폴트 옵션으로 CART 기법을 선택해 재현데이터를 생성한다. 보통 CART 알고리즘은 예측을 위해 개별 말단노드에서 관측된 반응 변수의 평균을 사용하지만 데이터 재현을 위해서는 비모수적으로 추정된 분포를 사용해 데이터를 생성하거나 무작위 복원추출법인 붓스트랩(Bootstrap)을 사용하는 것이 더 적절하다. synthpop에서는 베이지안 붓스트랩을 사용한다고 알려져 있다.

이외에도 CART를 확장한 Bagging이나 Random Forest, 그리고 서포트벡터머신(Support Vector Machine; SVM), 베이지안 네트워크(Bayesian Network) 등의 다른 비모수 방법론들도 재현데이터 생성에 사용될 수 있다.

베이지안 네트워크의 경우 개별 변수들을 노드로 두고 이들 간의 종속(인과)관계를 방향성이 있는 화살표로 연결한 네트워크 형태를 가지는 그래프 기반 모형이다. 원데이터로 학습된 베이지안 네트워크를 이용하면 다양한 쿼리에 대한 답을 얻을 수 있고, 각 노드에서 표본을 생성하는 과정을 통해 데이터 생성기로도 사용할 수 있다. 베이지안 네트워크는 Directed Acyclic Graph(DAG)라는 조건을 만족해야 하며, 개별 확률변수를 순차적으로 분할하는 CART와 달리 데이터 기저에 존재하는 다변량 분포를 종속관계에 따라 순차적으로 쪼갠 후 개별 조건부 확률분포를 범주형의 경우 전이행렬로, 연속형인 경우 회귀모형을 이용해 만드는 것이 보통이다. 〈그림 Ⅲ-2〉는 CART와 베이지안 네트워크의 예시를 보여준다.

25) 불순도는 다양한 범주(Factor)들의 개체들이 얼마나 포함되어 있는가를 의미함. 즉, 여러 가지의 클래스가 섞여 있는 정도를 말함. 회귀나무에서는 주로 지니계수(Gini Index)를 사용함

〈그림 Ⅲ-2〉 CART(좌) 및 베이지안 네트워크(우)의 예시



자료: https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/; Wei, J., Nie, Y., and Xie, W.(2020)

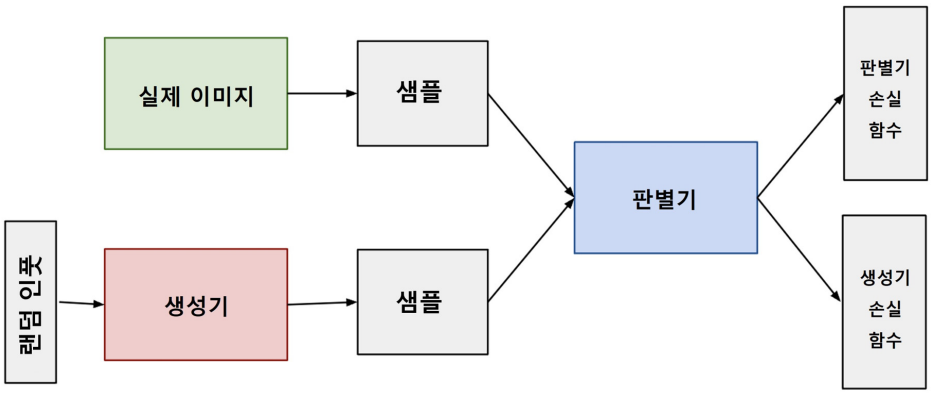
3) GAN 기반 생성 방법론

딥러닝 기술을 활용한 재현(합성)데이터는 컴퓨터 비전 등의 영역에서 AI학습용 데이터로 적극 활용되고 있다. 특히 이미지나 동영상 데이터에 대한 재현 기술이 많이 연구되고 왔으며 최근 정형데이터로의 확장도 이루어지고 있다. 본 보고서에서는 대표적인 딥러닝 기술인 GAN(Generative Adversarial Network)이 어떻게 재현데이터를 생성하는지 간략히 설명하려 한다.

회귀 및 분류 모형과 같은 많은 머신러닝 혹은 딥러닝 방법론들은 주어진 데이터를 바탕으로 특정 값을 도출하는 지도학습(Supervised learning)에 속한다. 지도학습모형을 훈련시키기 위해서는 모형을 통해 얻고 싶은 결과값이 포함된 데이터가 필요한데, Ian Goodfellow는 지도학습에 필요한 결과값 데이터를 모형이 자체적으로 만드는 생성모형(Generative model)인 GAN을 제안하였다.[26] 그가 모형을 설명할 때 사용한 비유를 인용하자면, GAN은 경찰과 위조지폐범 사이의 게임과 같다. 위조지폐범은 진짜 같은 화폐를 생성해 경찰을 속이려 하고, 경찰은 진짜 지폐와 가짜 지폐를 판별하려 한다. 이러한 경쟁적 학습이 지속되면 위조지폐범은 진짜와 매우 유사한 위조지폐를 만들 수 있게 되고, 경찰 또한 위조지폐를 판별하는 상당한 실력을 가지게 된다. 즉, 최종적으로 위조지폐범은 고품질의 위조지폐를 '생성'하게 된다. 이 비유에서 위조지폐범은 생성기(Generator)를, 경찰은 판별기(Discriminator)를 의미하며 이 두 모델이 서로 적대적 학습을 이어나가는 것이 GAN의 원리이다.

26) Goodfellow, I., Bengio, Y., and Courville, A.(2016)

〈그림 Ⅲ-3〉 GAN의 Generator와 Discriminator 관계



자료: Google developers(https://developers.google.com/machine-learning/gan/gan_structure?hl=ko)

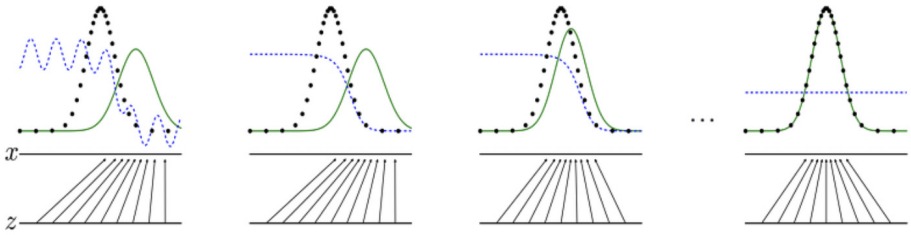
GAN은 〈그림 Ⅲ-3〉과 같이 데이터를 생성하는 생성기G와 생성된 데이터가 진짜인지 판별하는 판별기D를 학습시키는 과정을 반복한다. 이를 수식적으로 표현한 것이 아래의 손실 함수(Loss function)이다.

$$\min_{G}\max_{D} V(G,D) = E_{x \sim P_{data}(x)}[\log D(x)] + E_{z \sim P_z(z)}[\log(1 - D(G(z)))]$$

생성기는 데이터의 임의성(Randomization)을 주기 위해 정규분포 등에서 임의 생성한 값 z을 인자로 받아, 원데이터x와 유사한 데이터$G(z)$를 만들어 준다. 원데이터와 유사하게 만들도록 생성기를 학습하는 과정이 손실함수를 최소화($\min_G$)하는 과정이다. 원데이터와 만들어진 데이터$(x, G(z))$를 각각 판별기$D(\bullet)$에 넣어 판별하게 되는데, 판별력을 높이는 과정이 손실함수를 최대화($\max_D$)하는 과정에서 진행된다. 알고리즘의 목표는 진짜와 유사한 데이터를 만들어 내는 생성기이기에 최소화($\min_G$)를 최종적으로 진행하게 된다.

이렇게 손실함수를 이용하여 역전파(Backpropagation)의 방법을 통해 생성기와 학습기의 능력을 향상시키는데 Ian Goodfellow는 이 과정을 〈그림 Ⅲ-4〉를 통해 표현했다.

〈그림 Ⅲ-4〉 GAN 분포 학습 과정



자료: Goodfellow, Ian, et al.(2020)

〈그림 Ⅲ-4〉에서 굵은 점선은 원데이터의 확률 분포를 나타내고, 학습이 진행될수록 점차 이 선에 가까워지는 실선은 GAN의 생성기가 만든 확률 분포이다. 후반부로 갈수록 평행에 가까워지는 가는 점선은 판별기의 확률 분포를 나타낸다. 가장 왼쪽 그림이 학습 시작 시점에서의 상태이고 학습이 진행되면서 차례로 오른편의 상태로 변한다. 학습이 완료된 가장 오른쪽 상태가 되면 분류기가 생성된 것인지 원본인지 구분할 확률이 0.5에 수렴해, 생성모델이 실제 데이터와 거의 유사한 데이터를 만들어 냈음을 뜻한다.

이렇게 적대적 학습을 통해 얻게 된 생성모델 자체를 이용하여 재현데이터를 생성할 수 있다. 현재까지 GAN은 끊임없이 발전하여 이미지 데이터 분야에서는 DCGAN(Deep Convolutional GAN), medGAN(Choi et al. 2017), ehrGAN(Che et al. 2017) 등이 사용되고 있다. 뿐만 아니라 원하는 레이블에 대한 데이터를 생성하기 위해 학습데이터에 레이블을 포함하여 학습시키는 CGAN(Conditional-GAN)을 활용하여 정형데이터에서도 사용 가능한 GAN 기반 방법론에 대한 연구가 진행되고 있다.[27]

27) Xu, Lei, et al.(2019)

2. 보험데이터를 이용한 재현데이터의 예시

가. 데이터 소개

여기서는 실제 데이터를 이용해 재현데이터를 생성하고 생성된 데이터와 원본 데이터를 비교한다. 대표적인 비모수적 재현데이터 생성 알고리즘인 통계 소프트웨어 R의 synthpop 패키지와 그래프를 그리기 위한 tidyverse 패키지를 사용한다.

사용할 데이터는 뮌헨 공과대학교(TUM)가 실제 회사로부터 데이터를 얻어 공개 가능한 수준으로 조작한 보험데이터[28]로 15,000행과 11개의 열로 이루어져 있다. 이 데이터셋은 자동차 보험가입 여부를 포함해 성별, 직업, 혼인 상태, 신용불량 여부, 통화 시간 등 자동차 보험가입에 대한 정보를 수집하기 위해 연락을 받은(Cold call) 고객들의 다양한 정보를 포함하고 있다.[29] 원래의 데이터셋은 자동차 보험가입 여부를 예측하기 위해 훈련용 데이터(Training data)와 검증용 데이터(Test data)로 나누어져 있지만 본 보고서는 예측이나 분류의 문제를 다루는 것이 아닌 재현데이터 생성에 초점을 두고 있으므로 자동차 보험가입 여부가 라벨링되어 있는 훈련용 데이터만을 사용했다.

재현데이터 생성 전 데이터 전처리 과정은 다음과 같다. 먼저 이전 마케팅 콜에 대한 결과를 담고있는 'Outcome'변수의 결측값이 약 76%로 관측되어 변수를 삭제했다. 마케팅 콜 시작 시각과 종료 시각 정보가 각각 담겨있는 'Call Start'와 'Call End' 변수에서 가장 중요한 정보는 통화 시간이므로 후자 변수에서 전자 변수를 뺀 'Call_time' 변수를 연속형으로 생성하고 'Call Start'와 'Call End' 두 변수는 삭제했다. 통화 기기 종류를 나타내는 Communication 변수(Cellular와 Telephone 범주 존재)에서 N/A로 처리된 관측치들(22.5%)에 대해 'Others'라는 새로운 범주를 부여했다. 그 외 정보 동의를 하지 않거나 부정확한 정보를 포함한 관측치 169건(4%)에 대해 행을 삭제해 최종적으로 3,820건 관측치를 재현했다.

---

28) 데이터 출처(https://www.kaggle.com/datasets/kondla/carinsurance)

29) 전체 변수 정보는 〈표 Ⅲ-2〉를 참고하길 바람

〈표 Ⅲ-2〉 자동차 보험 Training Dataset Description

변수명	설명	예시
Id	고유 ID 번호	‘1’ … ‘5000’
Age	고객 나이	‘18’, ‘20’, … , ‘90’
Job	고객 직업	‘admin.’, ‘blue-collar’, 등
Marital	고객 혼인 상태	‘divorced’, ‘married’, ‘single’
Education	고객 학력 수준	‘primary’, ‘secondary’, 등
Default	신용불량자 여부	‘yes’ - 1, ‘no’ - 0
Balance	연간 평균 잔고(달러 기준)	-2119, 589, …
HHInsurance	가계 보험 보유 여부	‘yes’ - 1, ‘no’ - 0
CarLoan	자동차 대부금 보유 여부	‘yes’ - 1, ‘no’ - 0
Communication	마케팅콜 수단	‘cellular’, ‘telephone’, ‘N/A’
LastContactMonth	직전 마케팅콜 월	‘jan’, ‘feb’, 등
LastContactDay	직전 마케팅콜 일	‘1’ … ‘31’
CallStart	통화 시작 시각	12:43:15
CallEnd	통화 종료 시각	12:43:15
NoOfContacts	총 연락 횟수	‘0’. ‘1’. ‘2’. …
PrevAttempts	이전 마케팅콜 결과	‘failure’, ‘other’, ‘success’, ‘N/A’
CarInsurance	자동차 보험가입 결과	‘yes’ - 1, ‘no’ - 0

자료: https://www.kaggle.com/datasets/kondla/carinsurance

나. 재현데이터 생성

본 보고서는 재현데이터 생성을 간단하게 제시하고 유효한 데이터셋이 생성되었는지 확인하는 것이 목표이므로 재현될 변수들의 순서와 method를 디폴트 옵션으로 설정했다. 디폴트 옵션에서는 원데이터의 왼쪽에 있는 열부터 순차적으로 재현데이터를 생성하지만 필요하다면 패키지 내부 함수 syn의 옵션인 visit.sequence를 통해 변수들의 순서를 사용자가 지정할 수 있다. 생성 알고리즘의 선택은 옵션 ‘method =’으로 지정 가능한데 디폴트 옵션은 CART로서 범주형과 수치형 변수가 섞인 해당 데이터에 바로 적용 가능하다. CART 알고리즘에 변수들이 순차적으로 입력되면서 특정 조건을 만족하는지 여부에 따라 분류작업을 지속적으로 진행하며 지니 불순도(Gini Imputity)가 낮아지는 방향으로 구간을 정해 합리적인 관측치 집합을 만들어 낸다. 여기서는 최초분류변수로 Age를 선택하고 나머지는 CART 알고리즘을 따라 생성하였는데, 최초변수는 복원 랜덤 추출로 생성된다.

재현되는 데이터셋의 갯수는 syn 함수의 m으로 설정한다. 통계적 검증과 노출위험의 추정을 위해 다수의 재현데이터셋을 만들고자 한다면 m값을 조절하면 된다.[30] 본 예시에서는 m값을 1로 두고 한 세트의 재현데이터만를 생성해 원데이터와 직관적인 비교가 가능하게 하였다.

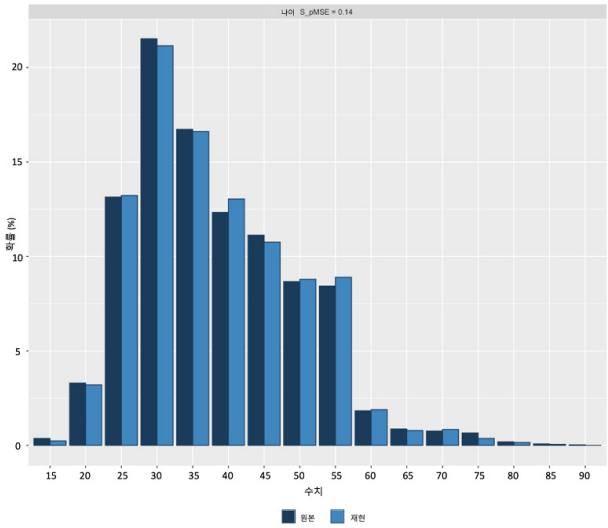
마지막으로 syn 함수를 이용해 생성한 재현데이터를 원데이터와 비교해 동일한 관측치가 있는지 확인한 결과, 총 9건이 원데이터와 중복되었는데 이는 전체 데이터의 0.23%이다. 따라서 이 9건의 관측치를 삭제하고 원데이터와 중복되지 않은 재현데이터 9건을 추가로 생성하여 원데이터와 동일한 크기인 3,820건을 최종 재현데이터로 채택했다.

다. 원데이터와 특성 비교

재현데이터에서 원데이터의 통계적 구조가 유지되는지 확인하기 위해 재현 전후 데이터의 단변량 및 다변량 비교 결과를 살펴보면 다음과 같다.

1) 단변량 비교

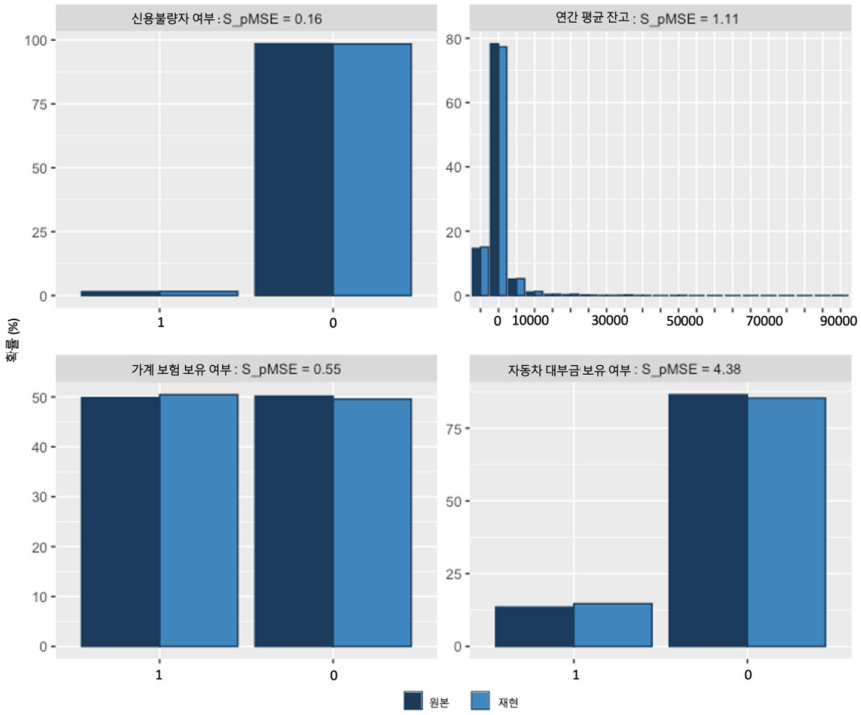
〈그림 Ⅲ-5〉 나이 변수 비교



30) Nowok, B., Raab, G. M., and Dibben, C.(2016)

〈그림 Ⅲ-5〉와 같이 랜덤 샘플링 기법으로 재현된 데이터는 원본과 그 데이터 분포가 유사하다. 실제로 이 두 분포의 차이를 확인해 보려 Kolmogorov-Smirnov Test[31]를 시행한 결과, 양측 검정 p-value가 약 0.99로 도출되어 재현이 잘 되었다고 볼 수 있다. 특히 원데이터에서는 최댓값이 95세로 노출이 일어날 확률이 높았으나 재현데이터에서는 최댓값이 86세로 나타나 원데이터나 외부의 데이터와 매칭을 어렵게 만들었다. 또 두 데이터에서 나이의 최솟값은 18세로 동일하다.

〈그림 Ⅲ-6〉 연속형 변수의 비교

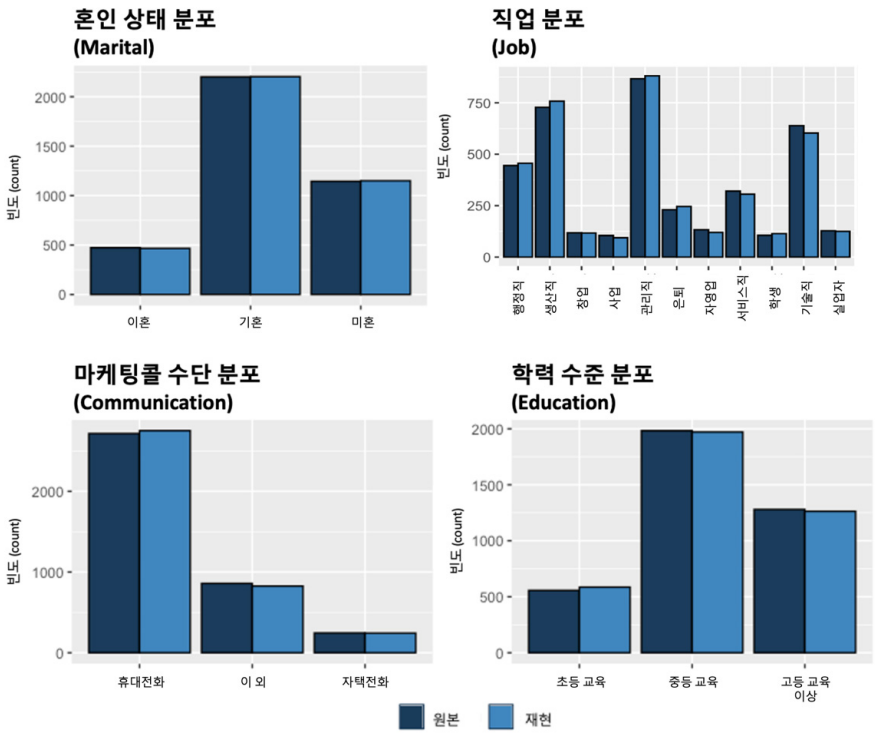


〈그림 Ⅲ-6〉은 신용불량자 여부(Default), 계좌 잔고(Balance), 가계 보험 보유 여부(HHInsurance), 자동차 대부금 보유 여부(CarLoan) 변수를 비교한 각각의 그래프로서 재현데이터와 원데이터의 분포가 유사함을 확인할 수 있다. 이를 수치적으로 보았을 때도

31) 데이터의 누적분포함수와 비교하고자 하는 분포의 누적분포함수 간의 최대 거리를 통계량으로 사용하는 가설검정 방법임. 귀무가설을 두 데이터의 분포가 동일한 것으로 설정하여 p-value가 α(주로 0.05, 0.01로 설정)을 으면 분포가 동일한 것으로 봄

마찬가지이다. 가계보험 보유(가입) 여부(HHInsurance) 변수의 경우, 보험가입(1)이 원데이터에서는 1,904건(49.8%), 재현데이터에서는 1,927건(50.4%)으로 나타났고 보험미가입(0)의 경우는 원데이터에서 1,916건(50.2%), 재현데이터에서 1,893건(49.6%)이 나타났다. 연속형 변수인 계좌 잔고(Balance)는 나이(Age) 변수와 같이 Kolmogorov-Smirnov Test를 통해 검정할 수 있는데 역시 p-value가 약 0.99로 두 데이터셋이 동일한 분포를 따른다는 결과를 보였다.

〈그림 Ⅲ-7〉 범주형 변수의 비교

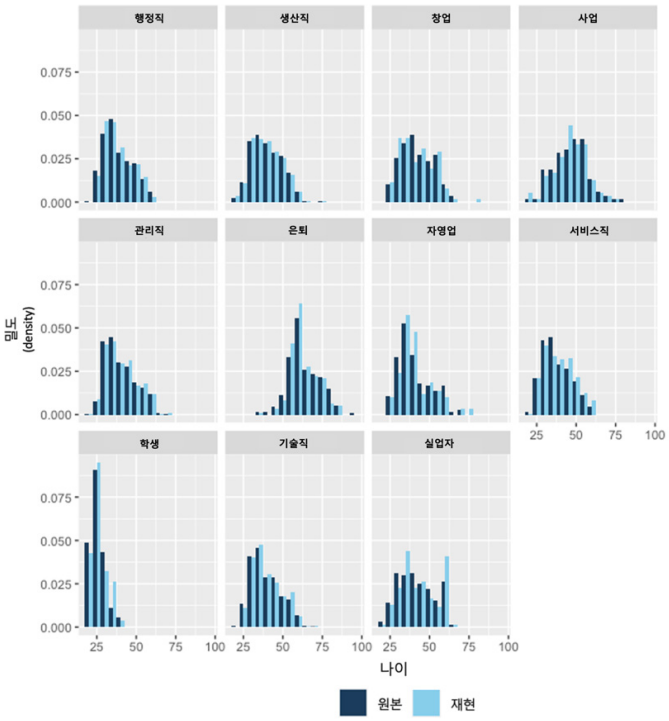


수치형으로 표현된 변수에 이어 〈그림 Ⅲ-7〉을 통해 범주형 변수 분포를 살펴보자. 혼인 상태(Marital), 직업군(Job), 마케팅콜 수단(Communication), 학력(Education)의 원데이터와 재현데이터를 비교한 막대그래프이다. 범주의 수가 10개 이상인 직업군(Job)에서도 원본 분포와 재현데이터 분포가 유사한 것을 확인할 수 있으며 특히 재현 순서가 비교적 앞 순서였던 혼인 상태(Marital)는 각 범주의 관측치 수가 원데이터와 99% 이상 동일했다.

2) 다변량 비교

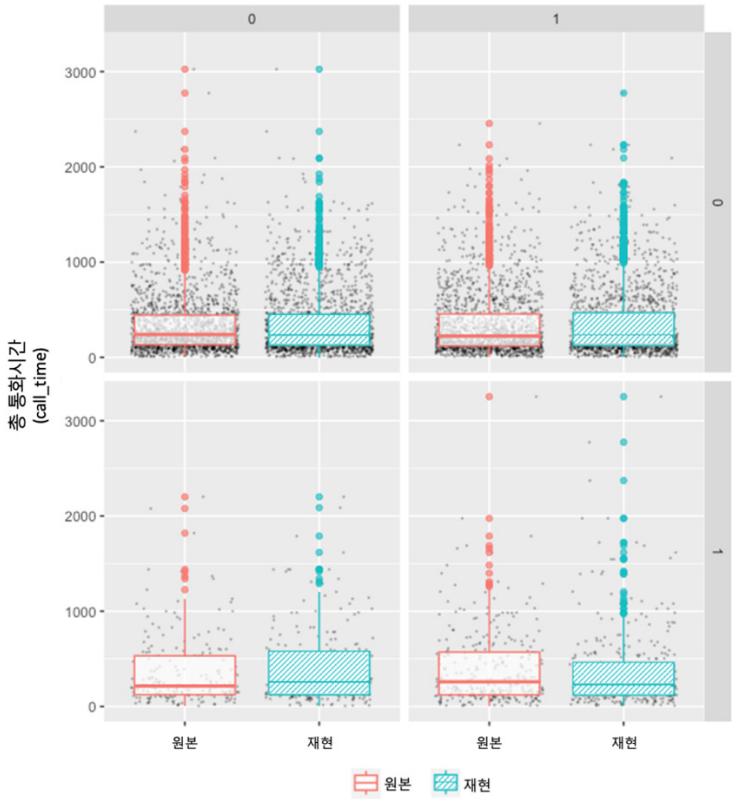
다변량 데이터에서 변수들의 상관관계는 매우 중요하다. 재현 전후 데이터가 유사한 상관관계를 유지하는지를 확인할 필요가 있지만, 모든 변수 조합을 고려하는 것은 지면상 어려우므로 여기서 예시로서 몇 가지 경우만 살펴보겠다.

〈그림 Ⅲ-8〉 직업군(Job)과 나이(Age) 이변량 분포 히스토그램



먼저 〈그림 Ⅲ-8〉은 이변량 분포를 히스토그램으로 나타낸 것이다. 연속형 변수인 나이(Age)를 가로 축으로 두고 범주 수가 많았던 직업군(Job)별로 각각의 범주에서 재현이 잘 되었는지를 보여준다. 예를 들어 학생 범주의 경우, 원데이터에서 20세 미만의 관측치는 12건이었고 재현데이터에서 20세 미만 관측치는 10건이었다. 그중 원데이터에서 18세로 관측된 행은 2건이었고 재현데이터에서 18세로 관측된 행 또한 2건으로 동일했다. 그 외 직업군 범주의 경우, 원데이터에서 20세 미만의 관측치가 발견되지 않았으며 재현데이터에서도 20세 미만 관측치가 없는 것으로 나타났다.

〈그림 Ⅲ-9〉 세 변수의 비교



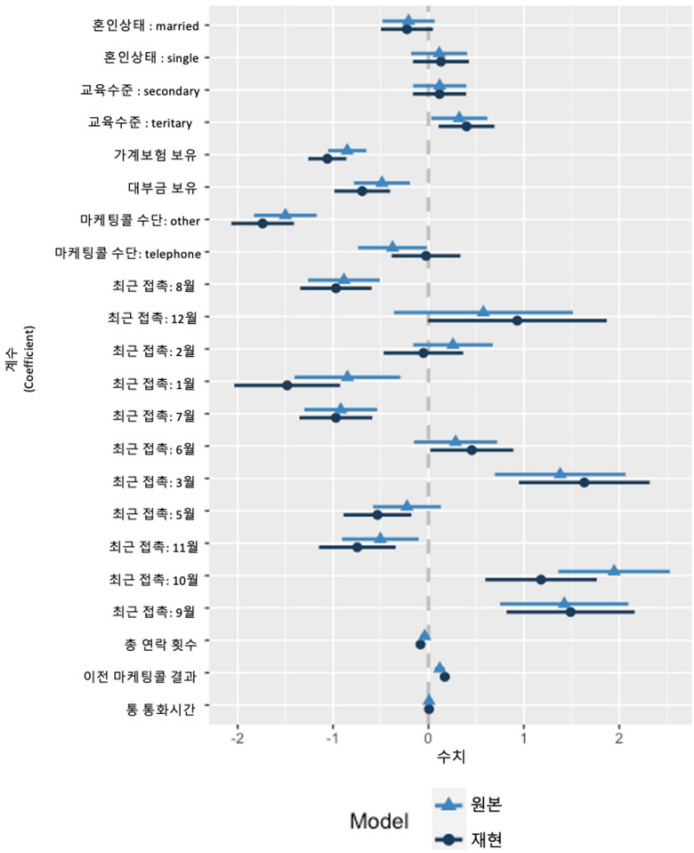
〈그림 Ⅲ-9〉는 마케팅콜 총 시간(call_time), 자동차 대부금 보유 여부(CarLoan), 가계 보험 보유 여부(HHInsurance)를 동시에 보여주는 박스 플롯이다. 원데이터 관측치 분포를 나타내는 눈금 없는 박스와 재현데이터 관측치 분포를 나타내는 눈금이 있는 박스가 각 범주마다 비슷한 범위 안에 놓여있다. 각 범주에 해당하는 데이터를 분할하여 총 4번의 Kolmogorov-Smirnov test를 시행한 결과, p-value가 모두 0.9에 가까운 수치를 기록하여 같은 분포를 따르고 있다는 결론을 도출했다.

3) 로지스틱회귀모형을 이용한 비교

엄밀한 의미에서 재현 전후 데이터를 비교하기 위해서는 가능한 모든 분석들을 시도하고 그 결과들을 비교해야 하겠지만 이는 현실적으로 불가능하다. 여기서는 대표적인 분석 모형인 회귀분석의 결과를 예시로 설명하겠다. 구체적으로, 변수를 다차원적으로 평가할 수 있는 로

지스틱회귀모형[32]을 적합해 두 데이터를 비교한다. 다중 회귀 모델에서 설명변수가 과도하게 많아지면 오히려 성능이 저하되기 때문에 본격적인 로지스틱회귀모형을 만들기 앞서 후진제거법(Backward elimination)을 이용해 설명력이 적은 변수를 순차적으로 제거하였다. 최종적으로 모델에 사용된 설명변수는 9가지로 혼인 상태(Marital), 교육 수준(Education), 자동차 대부금 보유 여부(CarLoan), 가계 보험 보유 여부(HHInsurance), 마케팅콜 연락 수단(Communication), 마케팅콜 통화 시간(Call_time), 총 연락 횟수(NoOfContacts), 이전 마케팅콜 결과(PrevAttempts), 직전 마케팅콜 월(LastContactMonth)이다. 종속변수로는 자동차 보험가입 여부(CarInsurance)를 설정하여 로지스틱회귀모형을 적합했다.

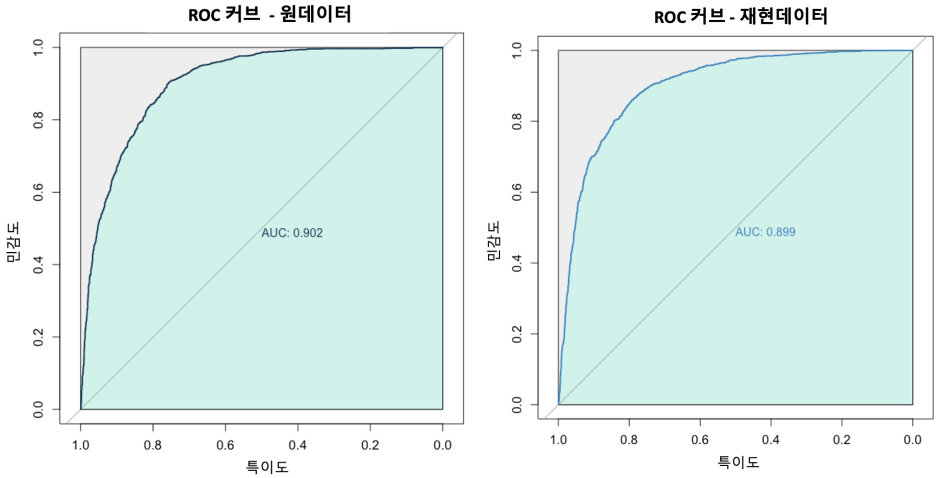
〈그림 Ⅲ-10〉 재현 전후 로지스틱회귀모형의 계수와 신뢰구간 비교



32) 종속변수(y)가 0 또는 1로 binary 변수일 때 사용하는 회귀분석 모델로 결과가 특정 분류로 나뉘기 때문에 일종의 분류 예측 모델임

〈그림 Ⅲ-11〉은 회귀분석 결과로 도출된 원데이터와 재현데이터 회귀계수와 신뢰구간을 비교한 결과이다. 원데이터와 재현데이터 범주형 변수 회귀계수의 신뢰구간이 Overlap되는 구간 평균은 약 0.71로 계산되었는데, 특히 범주형 변수의 회귀계수들의 Overlap 비율은 90% 이상의 정확성을 보였다.

〈그림 Ⅲ-11〉 재현 전후 데이터에 적합한 로지스틱회귀모형의 ROC 곡선 비교



적합된 두 개의 로지스틱회귀모형에서 도출된 ROC 곡선[33]은 〈그림 Ⅲ-11〉에서 보듯이 매우 유사하며, AUC[34]값 역시 원데이터는 0.902, 재현데이터는 0.899로 비슷하다. ROC 곡선을 도출할 때 보통 데이터를 훈련 집합과 검증 집합으로 나누는 것이 일반적이지만 여기서는 모형의 예측성능이 아니라 재현 전후 데이터의 유사성에 관심이 있으므로 검증 집합을 따로 두지 않고 전체 데이터를 훈련 집합으로 두고 진행하였다.

본 절에서는 재현 전후의 데이터를 다각도로 비교하였지만 여전히 제한적인 비교에 그친다는 점에서 재현데이터의 원데이터와의 유사성에 대해 일반적인 결론을 내리는 것은 어렵다. 사용자가 시도하는 쿼리나 분석의 종류와 범위는 매우 다양할 수 있어 이를 모두 조

33) ROC 곡선은 FPR(False Positive Rate)과 TPR(True Positive Rate)을 각각 x, y축으로 놓은 그래프로 임곗값(Threshold)을 바꿔가며 측정했을 때 FPR과 TPR의 변화를 나타낸 곡선임. 주로 이진 분류 모형의 성능평가에 사용되며 그래프가 좌상단에 근접할수록 좋은 성능을 보인다고 해석함

34) Area Under the ROC Curve의 약자로 ROC 곡선 아래의 면적을 뜻함. AUC 값은 ROC 곡선을 하나의 숫자로 요약한 값으로, 1(100%)에 가까울수록 모형의 성능이 좋다고 할 수 있음

사하는 것은 불가능하기 때문이다. 이런 이유로 특정 분석이 아니라 재현 전후 데이터를 분석 수준이 아니라 분포 수준에서 비교하는 것이 더 합당하다. 이는 재현데이터의 품질에 관한 중요한 대목으로 다음 장에서 좀 더 자세히 논하기로 한다.

3. 국내 및 해외 재현데이터 이용 사례

이 절에서는 금융과 관련된 국내외 몇 개의 사례들을 살펴보겠다. 이 외에도 다른 사례들이 있고 금융 외 다른 분야에서도 재현데이터의 사용 사례가 다수 있으나, 지면의 한계로 다 싣지 못함을 양해하기 바란다.[35)]

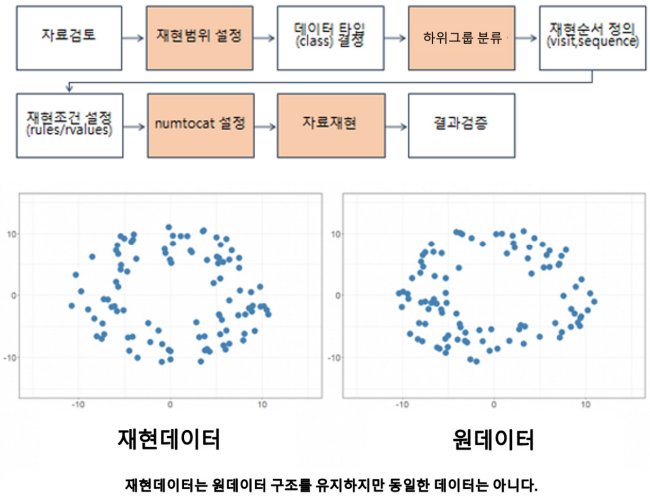
가. 국내 사례

1) 통계청 재현데이터 기초연구(2016~)

통계청은 마이크로데이터 공개에 따라 민감한 개인정보 노출위험을 줄이면서, 정보 손실을 최소화 하기 위한 방안으로 재현데이터 활용방안에 대한 탐색과 해외 사례 등에 대한 기초연구를 수행해 왔다. 구체적으로, 재현자료 생성을 위해 통계적 모형을 활용하는 방법론과 재현자료 노출위험 및 정보 손실 측정론에 대한 연구와 함께, 국내외 통계기관의 재현자료 생성 사례 및 통계데이터센터 DB에 대한 재현자료 시범 생성 결과 보고서를 발간한 바 있다.

35) 국내외 재현데이터 사용의 다른 사례들은 박민정(2020) 및 한국신용정보원(2020)를 참고하길 바람

〈그림 Ⅲ-12〉 통계청 K-통계시스템 구축 계획



자료: 통계청(2021b)

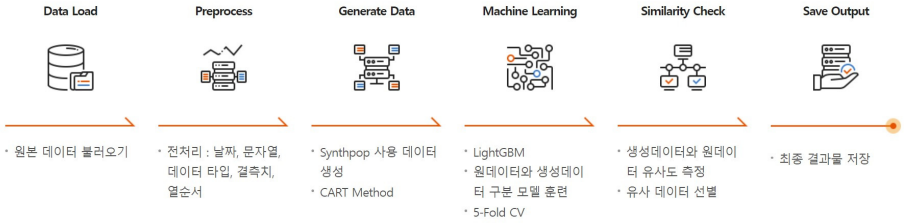
또한 〈그림 Ⅲ-12〉와 같이 2021년 공공데이터 활용을 증대하기 위한 통계청 K-통계 시스템 구축 계획을 발표했는데 새로운 데이터 보호 기술로 제안한 항목에 재현데이터가 포함되어 있는 것을 확인할 수 있다.

2) 코리아크레딧뷰로(KCB) 제주도 전입인구 특성 분석(2019)

코리아크레딧뷰로(이하, 'KCB'라 함)는 신용정보를 활용하여 개인신용평가를 시행하는 기업으로 신용평가 외에도 데이터의 제공과 활용을 도와주는 데이터 스토어를 운영하고 있다.

재현데이터와 관련해서, KCB는 2018년 이후 급증한 제주도 전출인구의 특성을 재현데이터로 만들어 분석·시각화하여 도내 인구 구조의 변화와 그 영향을 파악하기 위한 프로젝트를 진행한 바 있다. 재현데이터를 생성하는 과정에서 고차원으로 인한 데이터 부족 문제 해결을 위해 가능한 기준이 되는 변수 중 연속형은 축약된 명목형으로 바꿔 모델링하였고, 최종적으로 원데이터의 분포와 유사한 약 50만 건의 재현데이터를 생성해 제주도 전출인구 특성 분석을 진행하였다.

〈그림 Ⅲ-13〉 재현데이터 생성 순서



자료: KCB 데이터스토어(https://datastore.koreacb.com/support/utilizeCaseView8.do)

〈그림 Ⅲ-13〉은 이 프로젝트에서 재현데이터를 이용하여 성능평가를 했을 때의 진행 순서이다. Python에서 synthpop 패키지의 CART 방법론[36]을 적용하여 재현데이터를 생성했으며 머신러닝 모델 훈련 및 원데이터와 유사도 측정도 진행하였다. 그 결과 정확도와 AUC에서 각각 성능비 97.8%, 97.7%라는 우수한 평가수치를 보였다고 한다.

3) 신용정보원 부도예측을 위한 GAN 기반 재현데이터 생성 및 검증 보고서(2022)

신용정보원은 2022년 『부도 예측을 위한 인공지능 학습용 데이터 생성 및 검증 기법: GAN 기반 재현데이터를 중심으로』[37]라는 CIS 보고서를 발간했다. 인공지능 학습모형인 GAN을 적용해 원데이터의 통계적 특성을 유지한 재현데이터를 생성하고 그 결과를 평가한 내용을 담고 있으며, 생성된 재현데이터가 인공지능 학습데이터로 유용하게 활용될 수 있음을 시사했다.

내용을 간략히 요약하자면, 원데이터를 기반으로 한 재현데이터를 만들어 내기 위해 대출, 연체, 부도 정보를 포함한 실제 데이터를 준비하고 종속변수로는 부도 여부(Binary variable), 설명변수로는 신용공여 총 잔액, 원화 대출 총 기관 수, 연체율 등 신용정보를 사용했으며 Python 프로그래밍을 활용했다. 신용정보원 출처의 가명처리된 원데이터에서 적정 크기 표본 5만 행을 추출하고 이 샘플의 90%는 학습데이터로, 10%는 분류평가용 데이터로 분리해 사용했다. 학습데이터 중 부도 차주 레코드가 전체의 50%가 되도록 1차 재현데이터를 생성했고 이후 실제 데이터와 재현데이터 비율이 5:5가 되도록 구성한 새로

36) 추후 비모수적 생성 방법론에서 언급할 것임
37) 홍동숙(2022)

운 학습데이터를 생성했다고 한다. 그 이후 KS 통계량, AUC, 재현율을 평가지표로 사용하여 최종적으로 생성된 재현데이터를 평가한 결과, 재현데이터가 실제 데이터와 유사한 통계량을 보유하고 있음을 확인했고 데이터 불균형 문제에 따른 낮은 재현율이 개선되는 효과를 보여 실제 데이터를 대체할 수 있는 수준임을 보였다.

4) 국세청 재현데이터 도입 계획(2022)[38]

국세청 또한 2022년 국세 데이터 활용도를 높이기 위한 재현데이터 활용 계획을 밝혔다. 국세청의 데이터 센터는 방대한 양의 금융데이터를 보유하고 있으며 주요 정책 결정이 되는 소득자료를 포함하고 있다. 그러나 소득자료는 민감한 개인정보를 포함하고 있어 그동안의 국세청 데이터는 마스킹 등 전통적인 가명처리 기법을 적용하여 제공되어왔기 때문에 자료 훼손 또는 관계 왜곡으로 유용성이 저하되었다.

국세청은 이에 대한 대안으로 재현데이터를 선택하고 재현데이터 우선 구축대상은 활용빈도가 높은 종합소득세, 근로소득세 등 소득 분야라고 밝혔다. 추후 시범 구축된 재현데이터를 이용해 데이터 활용도 및 정보 노출위험을 테스트할 예정이다.[39]

나. 해외 사례

1) 미국 SIPP Synthetic Beta(SSB)(2013~2023년)

SIPP Synthetic Beta(이하, 'SBB'이라 함)는 가구 조사에서 비롯된 개별 관측치 수준의 마이크로데이터를 세금 및 사회보장 데이터와 통합한 결과물이다. 원데이터에는 Survey of Income and Program Participation 조사에 응한 응답자들의 설문기록이 담겨있으며 사회보장행정(SSA)/내부수입서비스(IRS) 양식 W-2 기록과 퇴직 및 장애 급여 수령에 대한 행정 기록 등 민감정보가 포함되어 있어 노출위험이 존재한다. 이러한 노출위험을 축소하기 위해 미국 인구조사국(Census Bureau)은 2013년부터 매년 누적된 데이터를 통합하여 새로운 버전의 부분 재현데이터를 가공하여 배포하고 있다.

---

38) 국세청(2021)

39) https://www.etnews.com/20220315000132

Census Bureau 소속 경제학자, 통계학자와 대학 연구원들이 협력하여 세금, 수입 등 개인정보가 담긴 데이터를 부분 재현했다고 알려져 있으며, 데이터 구조의 보존을 위해 변수 간 관계를 유지하도록 하고 각 관측치 기록을 대치하는 방식으로 연구를 진행했다. 총 9개의 SIPP 패널데이터(Panel data)가 1984년, 1990년, 1991년, 1992년, 1993년, 1996년, 2001년, 2004년, 2008년에 걸쳐 공개되었고 현재는 이 패널데이터 정보를 포함한 업데이트된 데이터가 주기적으로 업데이트되고 있다. 2018년에 업데이트된 버전 7.0에서는 누락되어 있던 잠재변수들도 재현데이터로 구현되어 채워졌고 데이터 내 논리적 불일치가 나타나는 부분들이 수정되었다. 600개 이상의 변수로 구성되어 있는 이 데이터는 주요 변수에 대한 통계적 특성이 원데이터와 매우 유사하다고 알려져 있고 가장 광범위하게 공개된 재현데이터로 평가받는다.

이 데이터는 공개 전에 노출위험에 대한 심사를 받았고 SSB의 기록과 외부 데이터를 연결해도 개인정보 파악이 불가능할 것이라는 판정을 받았다. SSB는 최종적으로 데이터 공개 검토 위원회와 IRS, SSA위원회의 승인을 받은 후 서버에 안전하게 등록되었다. 현재 SSB는 코넬 대학교의 가상 연구 데이터 센터(Virtual Research Data Center)에 있는 SDS(Synthetic Data Server)에 저장되어 있다. 연구자는 서버에서 무료계정을 만들고 키를 얻으면 SSB 데이터를 사용할 수 있다.

2) 영국 SynAE project(2018~2023년)[40]

이 프로젝트는 영국의 NHS(National Health Service) Digital이 제공하는 SUS(Secondary Uses Service) 데이터에서 추출한 Attendances&Emergency 데이터와 입원 환자 치료 데이터를 결합·가공하여 재현데이터로 구현한 시범사업으로서 현재까지도 추가된 정보가 업데이트되고 있다. 환자의 개인정보 노출을 막으면서도 데이터 공유를 확대하기 위한 취지로 시작된 이 프로젝트는 영국 정부의 NHS England 의무조항을 따르도록 수행되었으며 원데이터에 민감정보가 다수 포함되어 있어 ONS(Office for National Statistics)[41]에서 이 프로젝트를 검수하였다.

이 프로젝트에서 밝힌 데이터 가공 방법은 다음과 같다. 먼저 지역 인구통계학적 정보에

40) https://open-innovations.org/events/synae/

41) 영국에서 가장 규모가 큰 공식 통계의 독립 생산기관이자 공인된 국가 통계기관임

기반하여 지리적 정보를 제거하고 세분화된 연속형 변수를 밴드로 그룹화한다. 예를 들어, 연속형인 연령 변수를 '0~5', '5~10', '10~15'와 같이 5단위로 끊어 재코딩하는 것이다. 또 입원 일시, 시각과 같이 정확한 시간정보를 제거하고 이상치 정보는 추출해 마스킹 처리를 하였다. 추가로 고유한 값을 가진 관측치 및 일부 희귀 부분 집합도 제거하여 노출 위험을 줄였다.

이후 1차 가공된 데이터에 대해 R 패키지 'BNLearn'을 적용하여 재현데이터를 생성했다. 이는 베이지안 네트워크에 기반한 방식으로 계층 구조에서 일련의 조건부 확률을 기반으로 데이터 모델을 생성하는 것이다. 계층 구조의 상단(Top)의 네트워크 구조를 활용해 각 변수 분포에서 표본을 추출하면, 계층 구조 하부(Bottom)에 위치한 변수들의 표본 추출 가능 범위가 줄어들기 때문에 신빙성 있는 데이터를 만들어 낼 수 있다.

이 프로젝트에서 산출된 재현데이터는 매년 공개되고 있다. 건강·보건 관련 데이터 공개는 다양한 건강정보를 활용한 서비스 개발을 가능하게 하기에 보건의료 분야에서 재현데이터 생성에 대한 관심이 높은 편이다.

3) 스위스 보험회사 Die Mobiliar(2021년)

Die Mobiliar은 스위스의 보험회사로 재현데이터 솔루션을 제공하는 기업인 Statice와 협업을 통해 재현데이터를 이용한 고객이탈 예측모델을 활용하고 있다.[42] 애초에 Die Mobiliar는 보유한 실제 고객데이터를 기반으로 모델을 만드려고 했으나, 스위스 데이터 보호법의 규제로 인해 대신 Statice의 합성데이터 솔루션을 사내데이터에 적용해 재현데이터로 예측모델을 학습시켰다고 한다. 원데이터로 훈련한 모델과 재현데이터로 훈련한 모델을 비교했을 때, 재현데이터 훈련 모델이 원본의 95% 성능을 보인다고 알려져 있다. 이를 이용해 Die Mobiliar은 효과적인 데이터 익명화를 통해 소비자 개인정보를 보호하면서도 효과적인 마케팅 전략을 수립할 수 있었다. Die Mobiliar의 사례는 금융권에서의 재현데이터의 사용이 매우 큰 효과를 가져올 수 있음을 시사한다.

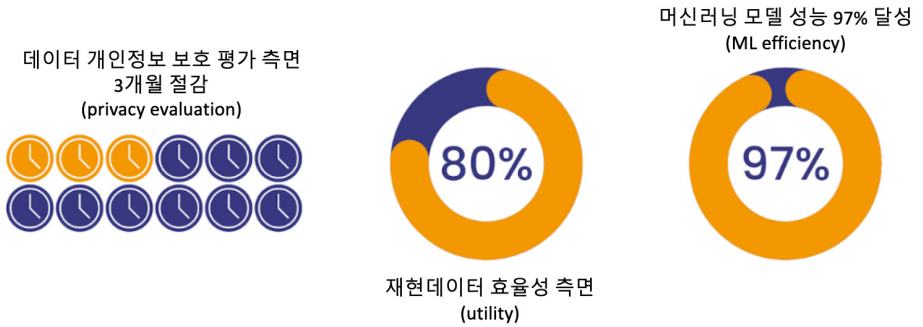
42) https://www.statice.ai/case-study/die-mobiliar

4) 독일 보험회사 Provinzial Rheinland(2021년)

재현데이터를 이용한 또 다른 해외 사례는 독일 보험회사 Provinzial Rheinland(이하, 'Provinzial'이라 함)에서 찾을 수 있다. Provinzial은 이미 소비자의 종류별 보험 수요를 예측하고 그에 맞는 제품을 추천하는 'Next best offer' 모델을 보유하고 있었으나, 이 마케팅 모델을 강화하기 위해 재현데이터를 활용하였다.

Provinzial은 데이터 가용성, 모델 사용, 개인정보보호 규정 세 가지 지표를 만족하는 재현데이터를 생성했고 성능을 테스트하는 과정을 진행하였고, 원데이터와 재현데이터를 비교한 결과 재현데이터의 80% 정도를 사용할 수 있음을 확인했으며 재현데이터로 학습시킨 모델의 성능이 원본 성능의 97% 이상을 충족했다고 한다. 이러한 결과는 원데이터와 재현데이터를 적절히 섞어 활용해 개인정보 노출 우려 없이 기존 머신러닝 모델을 발전시킨 것에 의의가 있다.

〈그림 Ⅲ-14〉 Provinzial사 모델 성능 결과



자료: Statice(https://www.statice.ai/case-study/provinzial-predictive-analytics-synthetic-insurance-data)

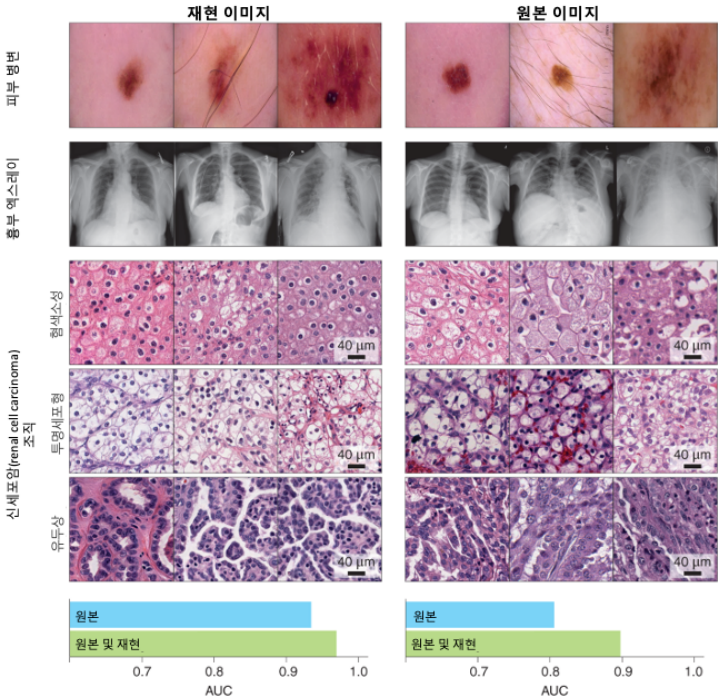
5) 해외 논문 사례[43)]

AI 분야에서 재현데이터를 이용한 이미지 관련 연구가 활발히 진행되고 있는데, 이 중 헬스케어 분야에서 재현데이터를 활용한 최근 논문 사례를 하나 소개한다.

43) Chen, R. J. et al.(2021)

이 논문은 의료학습용 이미지 데이터 부족의 해결책으로 재현데이터를 제시하고 있다. 예를 들어 특정 희귀질환 관련 의료 이미지를 분석하고 싶은 경우, 그 데이터 수가 현저히 부족하기 때문에 데이터 불균형으로 인한 편향 문제가 발생할 수 있다. 또한 병원마다 사용하는 의료장비가 다르기 때문에 데이터 표현에 이질성이 생기면 분석 정확도가 떨어질 수도 있다. 저자는 수집된 원래의 영상으로부터 대규모 합성데이터를 생성해 이 문제들을 해결할 수 있음을 주장하였다.

〈그림 Ⅲ-15〉 의료 이미지 재현데이터 예시



자료: Chen, R. J. et al.(2021)

〈그림 Ⅲ-15〉은 해당 논문에 제시된 그림이다. 피부 병변, 흉부 엑스레이 및 신장 세포 암의 세 가지 유형에 대해 우측에는 실제 이미지를, 좌측에는 GAN으로 생성된 재현데이터 이미지를 보여준다. 그림의 하단에는 재현데이터를 추가한 경우 AUC값이 증가해 예측모형의 성능이 더 좋아짐을 보여준다. 또한 해당 논문에서는 Visual Turing test[44]를 이용해

44) 자세한 내용은 Geman, D. et al.(2015)을 참고하길 바람

이미지 재현데이터 채택 및 평가 문제를 해결할 수 있을 것이라 제안하였다.

이 연구 사례는 의료 영상 분야에서의 재현데이터 사용이 개인정보보호 문제를 해결해 줄 뿐 아니라 학습데이터의 편향 문제도 해결해 사용하고자 하는 모형의 분석과 예측의 성능을 개선하는 데에도 도움을 줄 수 있음을 시사한다.

4. 재현데이터의 활용과 한계점

재현데이터는 알고리즘을 이용하기 때문에 제한 없이 빠르게 생성할 수 있으며 다음의 장점이 있다.

첫째, 재현데이터는 개인정보보호 등의 규제로부터 자유롭고 익명데이터로 분류되므로 사용자들은 이를 자유롭게 공유·전송·거래할 수 있다. 원데이터의 민감변수는 데이터 사용자들에게 공개되지 않지만, 재현데이터는 민감정보까지 포함하여 공개된다. 따라서 외부의 추가정보를 이용하더라도 민감정보 유출 가능성이 최소화되고 기존의 가명·익명처리 기법과 같이 사용할 수 있다. 이러한 특성으로 인해 재현데이터 사용자는 데이터의 모든 속성을 확인할 수 있고, 이를 이용하여 다양한 분석을 시행할 수 있게 된다.

특히 금융데이터의 경우 재현데이터의 생성기를 금융사의 내부망에 탑재해 사용할 수 있어 원데이터의 외부 유출에 대한 우려가 없다. 더 나아가 회사 내 부서별로 재현데이터를 독립적으로 생성한 후 부서별 재현데이터를 병합·결합해 전사적인 재현데이터를 구축할 수도 있다. 재현데이터가 규제로부터 자유로운 익명데이터임을 고려할 때, 이와 같은 방식은 개인정보가 담긴 금융데이터에 대해 엄격한 현행 규제 하에서 사내 수집된 다양한 빅데이터의 활용을 극대화할 수 있는 현실적인 대안이라고 하겠다.

둘째, 재현데이터는 모든 분야에서 더 많은 양의 고품질 학습데이터를 구축하는데 사용될 수 있다. 재현데이터의 생성을 통해 불균형 데이터의 문제를 해결할 수도 있고, 데이터 편향이 존재하는 경우에도 보정된 재현데이터를 사용해 분류기나 다른 지도학습모형의 성능 향상에 도움을 줄 수도 있다.[45)]

45) 데이터의 양과 질을 개선하는 작업을 데이터 증강이라고 하기도 함

그러나 재현데이터 사용 시 주의해야 하는 점도 존재한다. 원데이터와 완전히 같게 복제한 것이 아니기 때문에 원데이터와 비교할 때 오차가 존재할 수 밖에 없다.[46] 반대로 생각하면 재현데이터의 생성과정에서 확률은 낮지만 우연히 원자료와 비슷한 값이 나올 가능성이 있으므로 이를 제어하는 장치가 필요하기도 하다. 또한 원데이터의 성격에 따라 재현 알고리즘이 다를 수 있다. 예를 들어 횡단면 연구인지, 시계열 혹은 패널형태인지, 위계가 있는지, 극단값이 있는지에 따라 적합한 알고리즘이 다를 수 있다. 다양한 상황에서 적절한 재현 알고리즘을 구현하는 문제는 앞으로의 연구 문제로 남아있다. 마지막으로 선택된 모형과 알고리즘에 따라 재현데이터의 품질이 달라지므로, 주어진 재현데이터의 품질을 원데이터와 비교하고 평가할 수 있는 척도 또한 필요하다. 재현데이터의 효용과 노출위험을 측정하는 방법에 대해서는 본 보고서의 IV장에서 설명한다.

46) 여러 세트의 재현데이터를 생성해 제공하면 분석 오차의 크기 측정과 튜닝을 통해 오차 크기 조절이 가능함