

기계학습 운영화(MLOps) 실행 전략



조성익 || 한국전자통신연구원 책임연구원

유용식 || 한국전자통신연구원 실장

표철식 || 한국전자통신연구원 단장

데이터과학과 기계학습이 융합되는 DSML(Data Science & Machine Learning) 프로젝트에서 많은 기업이 실패를 경험하고 있으며, 데이터 의존적인 DSML 프로젝트 수명주기의 복잡성을 제어하지 못하는 것이 핵심 원인으로 거론되고 있다. 본 고에서는 주요한 실패 요인에 대한 분석을 토대로, 순방향과 역방향으로 복잡하게 얽히며 모니터링과 지속적 개선이 수반되어야 하는 DSML 프로젝트의 수명주기의 특성과, 이를 성공적으로 제어하여 기술 부채를 줄일 수 있도록 하는 기계학습 운영화(MLOps)의 개념을 살펴본다. DSML 프로젝트가 기계학습 운영화를 통해 실행되는 단계를 계획 수립, 데이터 준비, 분석 및 관리, 기계학습, 배포 및 생산화, 풀 스택 지원으로 구분하고, 각 단계에 포함되는 중요 프로세스를 정리한 프로세스 풍경(process landscape)을 제시한다. 기계학습 운영화 실행 전략에 활용할 수 있도록 각 프로세스의 특징 및 주요 기능을 상세히 정리하여, DSML 프로젝트를 통해 비즈니스 데이터 기반의 통찰을 얻고자 하는 분들에게 작은 도움을 제공하고자 한다.

I. 서론

신기술 영역에 대한 가트너의 하이프 사이클 분석[1]~[3]에서 인공지능의 위치를 보면, 2019년에 기대의 정점을 지나 환멸의 골짜기로 진입하면서 2020년에 새로운 몇 개의

* 본 내용은 조성익 책임연구원(☎ 042-860-1087, chosi@etri.re.kr)에게 문의하시기 바랍니다.

** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.

***이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 국가과학기술연구회 융합연구단 사업(No. CRC-15-05-ETRI)의 지원을 받아 수행된 연구임

세부 기술로 대체되는 흐름을 보여주고 있는데, 이는 인공지능에 대한 과대한 기대가 사라짐과 동시에 일부 초기 수용자들에 의해 제한적 범위에서 실제 비즈니스에의 적용이 시작된 것으로 해석할 수 있다. 여러 분석 보고서에 의하면, 인공지능 도입에 실패한 대부분 기업은 데이터 의존적 특성에 대한 이해 부족, 고품질 데이터 미확보, 생산 현장의 드리프트를 고려하지 못함 등의 경험을 한 것으로 나타난다. 즉, 인공지능은 알아서 문제를 해결하는 마술 램프가 아니라, 곤강(昆崗)의 자갈밭에 숨겨진 옥돌 원석을 찾듯이 고품질 데이터를 모으고, 원석을 깎아 구슬을 만들듯이 인공지능 실행 프로세스를 깔고 닦아 데이터 기반의 비즈니스 통찰을 얻는 기업만이 성과를 올릴 수 있다는 교훈을 제시하고 있다.

본 고에서는 기업의 경영진을 대상으로 하여 데이터과학 및 기계학습(Data Science & Machine Learning: DSML)이 융합되는 프로젝트 실행에 있어 도움이 되는 정보를 정리해 보고자 한다. 이를 위해 II장에서는 DSML 프로젝트의 실패 요인을 정리하고, III장에서는 DSML 프로젝트의 수명주기와 기계학습 운영화(Machine Learning Operationalizations/Operations: MLOps)의 개념을 정리하면서 DSML 프로젝트의 프로세스 풍경(process landscape)을 도입한다. IV장에서는 계획 수립, 데이터 준비, 분석 및 관리, 기계학습, 배포 및 생산화(productionization), 풀 스택 지원으로 구분되는 DSML 프로젝트의 실행 단계를 구성하는 프로세스의 특징 및 개념과 주요 기능을 정리하고, V장에서는 본 고의 결론을 제시한다.

II. DSML 프로젝트의 실패 요인

1. 인공지능에 대한 기대와 현실

코로나 19는 언택트라는 새로운 경험을 통해 사회 시스템, 라이프 스타일, 산업 지형도 등이 바뀌는 디지털 변혁(digital transformation)을 가속화시키는 전환점이 될 것으로 전망된다[4]. 디지털 변혁의 목표는 모든 비즈니스 영역에서 데이터와 기술을 통합한 디지털 비즈니스 솔루션을 이용하여 조직 내부의 문제 해결과 새로운 비즈니스 기회를 창출하고, 다양한 접점에서 고객 경험 증대와 진보된 고객 가치를 제공하기 위함에 있다. 디지털 변혁을 준비하는 조직은 인공지능을 토대로 비즈니스 데이터를 수집·준비·학습하여 비

즈니스 프로세스 최적화 및 운영 효율성을 높일 수 있는 실행 가능한 비즈니스 통찰을 얻을 것으로 기대하고 있다.

그러나 현실에서는 인공지능 기술의 도입을 시도했던 대부분의 글로벌 기업이 야망과 실행 사이에 큰 격차를 경험한 것으로 알려지고 있다. 여러 IT 컨설팅 기업의 보고서를 종합해 보면[5]-[9], 인공지능 프로젝트의 10~30% 정도가 제한적 성과를 내면서 이노베이터를 지나 얼리어답터로 이동하고 있는 것으로 나타난다. 하지만, 대부분 프로젝트는 실증 단계(PoC)에서 멈추거나 데이터 준비·통합에서 속도를 내지 못하며, 설사 이 단계를 통과해도 컨셉 드리프트(concept drift) 문제에 부딪히는 것으로 나타나고 있다. 이는 글로벌 기업도 인공지능 도입을 아직 본격화하고 있지 못하며, 여전히 잠재력을 시험하는 초기 단계에 머물고 있음을 의미한다.

2. DSML 프로젝트의 실패 요인[10]-[12]

DSML 프로젝트의 전형적인 실패 과정을 보면, 모델 개발에 몇 주가 걸리고 생산 적용은 더 많은 시간이 걸리지만, 생산 단계에서 발견된 심각한 문제로 전체 프로젝트를 재설계하거나 학습 모델을 폐기하는 경우가 자주 발생한다. 또한, 학습 모델의 데이터 규격이 생산 현장과 맞지 않아 배포되지 못하는 경우도 적지 않다고 알려져 있다. 이는 작은 규모의 프로젝트로 실증 단계까지 도달해서 도입 타당성을 검토하는 프로토타이핑을 거쳐, 조직의 전체 비즈니스 프로세스에서 데이터를 수집·가공하고 기계학습 모델을 생성·배포하고, 궁극적으로 제품·서비스에 사용할 수 있는 실행 가능한 기계학습 운영화 단계에 도달해서 비즈니스 가치를 창출하는 DSML 프로젝트의 생산 적용 과정에 다양한 장애요인이 있다는 것을 간과했기 때문이다.

대표적 장애 요인으로는 기계학습 운영화 전략·방법론에 대한 경영진과 엔지니어의 인식 문제, 조직에 존재하는 문화적·환경적 장애, 생산 현장의 부가 가치 생성 실패, 문화적 충돌, 시스템 사이의 충돌 등이 있고, 이외에도 전문 인력 채용의 어려움, DSML 솔루션의 생산 적용 과정에서의 소통 부족 등이 거론되고 있다.

DSML 프로젝트에 성공한 조직은 일관된 데이터 흐름과 더불어 기계학습 운영화를 이 해하는 협업 문화가 존재했다는 공통점이 있다. 그러나, DSML 플랫폼 도입으로 모든 문제가 해결된다고 믿는 경영진이나 기계학습 모델링이 프로젝트의 완성이라고 오해하는

엔지니어가 여전히 많다. 이런 조직에서는 프로젝트 실행에서 비즈니스 흐름, 기계학습 파이프라인, 생산 데이터 준비 등의 일관성을 유지하지 못하거나 데이터 의존성을 극복하지 못하는 경우가 많다. DSML 기술을 잘 활용하려면 정확한 문제 이해와 올바른 솔루션이 필요한데, 조직의 복잡한 문화(데이터 사일로, 문화적 갈등, 학술적 장벽 등)와 환경적 요인(학습과 생산의 분리, 데이터 준비의 방해 요인 등)이 문제의 해결을 막는 경우가 많다. DSML 프로젝트를 위한 문화·도구·인프라에 익숙하지 않은 조직은 생산 단계에서 그 모델의 부가가치가 떨어짐을 발견하거나, 모델 실행의 전제 조건(생산 데이터 구조, 모델 출력의 지연 시간 등)에 따른 프로젝트 실행의 복잡성이 너무 높아서 비즈니스 가치가 없음을 발견할 수도 있다.

데이터 보안을 위한 고도의 사일로 체계는 다양한 형식의 데이터를 통합하여 일관된 방식으로 활용하는 것을 방해하는 장벽으로 작용한다. 데이터의 준비를 위해 보통 가용 시간의 80% 정도를 소모하는데, 생산 데이터의 통합 접근 및 관리를 위한 일관된 인프라가 없으면, 데이터 엔지니어의 업무 과중 및 데이터 준비 부족으로 데이터 파이프라인이 중단될 가능성이 커지게 된다. 그 결과 생산 파이프라인의 결과를 활용하는 비즈니스 담당자, 생산 데이터의 드리프트에 맞춰 모델을 갱신하는 기계학습 엔지니어, 생산 데이터를 준비하는 데이터 엔지니어 등 여러 담당자의 의견 충돌과 협업 부족이 문화 충돌로 발전하면서 프로젝트 조직은 와해될 수도 있다.

기업의 다양한 엔터프라이즈 시스템은 Java, C/C++ 등으로 개발되어 프론트엔드와 백엔드가 이미 비즈니스 목적에 맞도록 최적화된 경우가 대부분이다. 반면, 기계학습 파이프라인과 모델 배포 아티팩트는 주로 Python, R 등의 언어로 개발되면서 언어와 인터페이스의 비호환성 문제로 인해 생산 및 활용 단계에서 엔터프라이즈 시스템과 기계학습 플랫폼의 충돌이 발생할 수 있다. 프로젝트의 가용성을 테스트하고 최종적 배포 승인에 이르는 데 상당한 시간(몇 달 또는 1년 이상)이 걸리는 경우가 많은 것이 현실인데, 호환성 충돌이 데이터 구조나 모델 성능에 영향을 미치면 원점으로 돌아가 프로젝트를 처음부터 다시 시작해야 할 수도 있다.

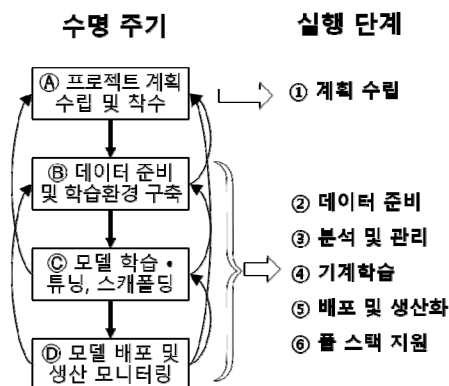
III. DSML 프로젝트의 수명주기와 기계학습 운영화

1. DSML 프로젝트의 수명주기

DSML 프로젝트의 실패를 줄이기 위해서는 프로젝트 실행 과정에서 데이터와 모델의 드리프트를 모니터링하면서 지속적인 개선을 도모하고, 프로젝트 실행 성과가 조직의 비즈니스 목표 달성을 위한 성과 지표에 미치는 영향을 분석하면서, DSML 프로젝트의 수명주기를 총괄 제어할 수 있어야 한다[13]-[16].

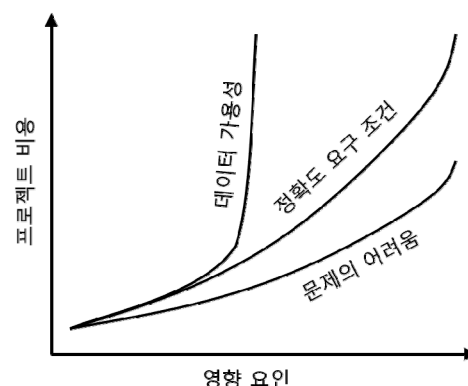
DSML 프로젝트의 수명주기는 [그림 1]에 나타난 것처럼, ① DSML 기술을 적용하여 비즈니스 가치 창출을 위한 통찰을 얻을 수 있도록 준비하는 프로젝트 계획 수립 및 착수, ② 기계학습에 활용할 수 있는 비즈니스 데이터를 수집·정제·관리하는 데이터과학 영역인 데이터 준비 및 학습 환경 구축, ③ 비즈니스 실행을 위한 데이터 의존적 모델을 개발하는 기계학습 영역인 모델 학습·튜닝 및 스캐폴딩, ④ 이를 생산 현장에 적용하여 솔루션 시스템을 구성하고 운영하도록 지원하는 기계학습 운영화 영역인 모델 배포 및 생산 모니터링으로 크게 구분할 수 있다[13],[14].

①에서는 비즈니스 목표 달성을 위한 성과 지표에 직결되는 메트릭을 규정하고, 생산 데이터와 주요 자산의 가용성을 분석하고 실행 계획을 설계하여 프로젝트에 착수한다. ②에서는 데이터 수집·정제, 특징 공학(feature engineering) 등의 데이터 준비와 더블



〈자료〉 [13]을 토대로 [14]에서 재구성

[그림 1] 수명 주기 및 실행 단계



〈자료〉 [13]을 토대로 [14]에서 재구성

[그림 2] 프로젝트 비용의 영향 요인

어 데이터 분석·관리의 고급 기능을 수행한다. ㉓에서는 모델의 학습·튜닝·평가를 수행하는 기계학습의 기본 기능과 더불어 자동 기계학습(AutoML)이나 모델 스캐폴딩(model scaffolding)과 같은 고급 기능을 수행한다. ㉔에서는 학습 모델을 생산 현장에 배포하여 생산화하면서 데이터·모델의 품질 관리 및 모니터링을 수행한다.

프로젝트가 ㉓→㉔→㉕→㉖의 일방적인 순방향 진행으로 완료되는 경우는 매우 드물고, 대부분 프로젝트에서는 [그림 1]에 나타난 것처럼 역방향 피드백이 발생한다.

역방향 피드백의 발생을 간략히 정리해 보면, ㉔→㉓ 피드백은 학습 데이터 또는 지상 진실(ground truths)이 부족하거나 데이터 및 학습 환경의 준비 시간이 부족할 때 발생할 수 있다. 데이터가 부족하면 공공 데이터 입수나 외부 데이터 구매 등의 전술을 사용하고, 시간이 부족하면 비즈니스 요구사항을 낮추거나 프로젝트 수행 기간을 늘리는 전술을 사용할 수 있다. ㉕→㉔ 피드백은 지상 진실의 오류가 발생하거나 모델 튜닝에 필요한 테스트 데이터가 부족할 때 발생할 수 있다. 지상 진실의 오류는 이를 다시 생성하거나 오류 데이터를 학습에서 제외하여 해결할 수 있고, 테스트 데이터 부족은 훈련·평가·테스트 데이터를 적절히 분배하여 해소할 수 있다.

㉕→㉓ 피드백은 학습 목표의 달성이 불가능하거나 학습 정확도와 비용 사이의 불균형으로 인해 발생할 수 있다. 중요한 프로젝트 영향 요인은 [그림 2]에 나타난 것처럼, 대처가 어려운 순서에 따라 데이터 가용성, 정확도 요구 조건, 문제의 어려움으로 구분할 수 있다. 데이터 가용성의 문제가 커질수록 비용은 빠르게 증가하며, 어느 한계에 이르면 비용을 추가 투입해도 성과 향상을 기대할 수 없게 된다. 이는 기계학습의 데이터 의존성을 나타내며, 저품질 데이터(학습 및 생산 데이터)로 인한 한계를 의미한다. 데이터 품질이 보장되는 경우, 학습 모델의 정확도 요구에 따라 모델의 복잡도가 높아지면서 학습·추론 시간의 증가로 인해 프로젝트 비용이 증가하게 된다. 예를 들면 현존 최고의 기술 수준을 넘어서는 학습 정확도를 요구하면 비용이 발생하게 된다. 데이터 품질이 보장되고 합리적인 정확도가 요구되는 경우, 기계학습에서 문제의 어려움을 해결하는 것은 덜 심각한 도전이 된다.

㉖→㉕ 피드백은 생산 모델(production model)의 출력(예측, 분류, 추론 등)이 너무 늦거나, 너무 무겁거나, 성능이 떨어지거나, 예상하지 못한 결과를 출력할 때 발생한다. 모델의 출력이 너무 늦거나 무거우면, 고성능 컴퓨팅 환경의 도입이나 경량 모델의 적용

(모델 스캐폴딩)으로 해결할 수 있다. 데이터에 과적합된 학습 모델은 생산 현장에서 예상하지 못한 결과를 출력할 수 있으며, 훈련·평가·테스트 데이터를 적절히 뒤섞거나 비중을 조절하고 재학습하여 해결을 도모할 수 있다.

㉔→㉕ 피드백은 생산 데이터(production data)의 특성이 학습 데이터와 다르거나 특징 공학의 적용 오류로 인해 예상하지 못한 결과를 출력할 때 발생한다. 예를 들면, 정규화 데이터로 학습한 모델에 비정규화 생산 데이터를 입력하거나, 학습과 생산 데이터의 특징 공학 로직이 다른 경우가 이에 해당한다. 학습 데이터 준비의 요구 조건을 생산 현장과 공유하는 특징 저장소를 도입하여 문제를 해결할 수 있다.

㉔→㉔ 피드백은 생산 모델을 평가하는 메트릭이 비즈니스 요구사항과 다르거나 컨셉 드리프트가 존재할 때 발생할 수 있으며, 문제의 원인을 제거한 다음 ㉔→㉔의 과정을 재실행해야 한다. 메트릭이 잘못되면 비즈니스 요구사항에 맞추어 프로젝트를 다시 계획해야 한다. 모델 출력을 감시하는 드리프트 트리거링을 생산 파이프라인에 추가하고, 드리프트가 발생하면 새로운 모델을 자동으로 학습하여 대체하는 기계학습 파이프라인을 연계 시키도록 프로젝트를 조정해야 한다.

2. 기계학습 운영화의 개념

[그림 1]에 나타낸 DSML 프로젝트 수명주기는 기계학습 운영화 개념과 밀접한 연관성을 가지고 있다. MLOps는 ML과 DevOps를 합친 용어이며, 대체로 “데이터과학 및 기계학습의 솔루션 개발과 DevOps 기반의 전주기 운영을 통합하여, 데이터 준비, 모델 학습, 모델 배포, 생산 적용 및 모니터링을 포함하는 전체 수명주기에서 안정적으로 서비스를 제공하면서도 신속하고 유연한 개발을 추구하는 문화·기술·인프라의 개념적 결합” 정도로 정의할 수 있다[14].

기계학습 운영화의 개념은 구글 엔지니어 D. Sculley가 기계학습의 숨겨진 기술 부채(hidden technical debt) 문제를 지적한 논문[17]에서 시작되었다. 이 논문은 기계학습 프로젝트에서 모델링(“ML coding”)은 아주 작은 범위에 불과하며, 비즈니스 문제를 기계학습 문제로 변환, 충분한 학습 데이터의 수집·정제·검증, 모델링 리소스 및 도구, 모델 배포 체계·인프라, 생산 현장의 드리프트 모니터링, 비즈니스 활용 과정 등의 단계가 훨씬 더 크고 복잡하다고 주장을 하였다. 모델링이 프로젝트의 전부라고 착각하면, 이를 제

외한 나머지가 숨겨진 기술 부채로 작용하여 기계학습 프로젝트를 실패로 이끌게 된다. 즉, 기계학습 모델의 일부 수정 불가능, 학습과 생산 데이터가 일치하지 않을 때의 예측 불가능성, 입력과 출력 사이의 블랙박스로 인한 해석 불가능성, 통상적인 SW 개발과 달리 엄격한 추상화 경계의 소실 등이 숨겨진 기술 부채로 작용하므로, 프로젝트를 제어할 수 있도록 기존과는 다른 운영 방식이 필요하다고 주장을 하였다.

3. DSML 프로젝트가 전개되는 풍경

[그림 1]의 DSML 프로젝트 수명주기를 기계학습 운영화를 위한 실행 단계로 표현하면, ① 계획 수립, ② 데이터 준비, ③ 분석 및 관리, ④ 기계학습, ⑤ 배포 및 생산화, ⑥ 풀 스택 지원으로 구분할 수 있다. [그림 3]은 DSML 프로젝트가 기계학습 운영화의 개념에서 프로세스와 파이프라인을 통해 실행되는 풍경을 나타낸 것이다[14].

[그림 3]은 MLOps 개념을 제시한 논문[17의 Fig.1] 및 풀 스택 심층학습 강의 자료[13]의 Lecture 4를 참조하고, DSML 최신 동향을 반영하여 필자가 작성한 것이다. 구체적으로, ① 계획 수립 단계는 요구 공학 방법론을 토대로 DSML 프로젝트의 고유 특성을 반영



중요 자산: 협업문화, 전문인력, 암묵지, 리소스(HW, SW), 데이터, 모델, 산출물 등
엔터프라이즈시스템: EIP, ERP, PMS, CRM, SCM, EAI, MES, BI, RPA, DBMS, EDW, KMS, GIS 등

〈자료〉 조성익, "기계학습 운영화(MLOps)의 프로세스 풍경", ETRI Technical Memo, 2020.12.

[그림 3] DSML 프로젝트의 프로세스 풍경

하였고, ② 데이터 준비에서 ⑥ 풀 스택 지원에 이르는 단계는 MLaaS(ML as a Service)와 DSML 올인원 플랫폼(all-in-one platform)에서 제공하는 주요 기능을 토대로 최신 기계학습 운영화 방법론 및 심층학습(deep learning) 연구 동향을 반영하였다. MLaaS는 Amazon AWS, Microsoft Azure, Google Cloud AI 등을 참조하였고, DSML 플랫폼은 Altair, Alteryx, Anaconda, BigML, Databricks, Dataiku, DataRobot, Domino, H2O.ai, KNIME, MathWorks, RapidMiner, Tableau, Tibco Software 등의 스타트업 제품과 한국전자통신연구원에서 개발한 BeeAI 인공지능 플랫폼[18],[19] 및 삼성 SDS의 Brightics AI 플랫폼[20]도 참조하였다.

DSML 프로젝트의 실행 파이프라인 그룹은 데이터과학, 기계학습, 생산화, 비즈니스의 네 종류로 크게 구분할 수 있다. ①에서 ⑤에 이르는 과정에서 순방향과 역방향의 파이프라인이 복잡하게 얽히며 반복되는 과정은 [그림 1]과 동일하며, DSML 프로젝트의 여러 프로세스가 중요 자산 및 엔터프라이즈 시스템에 연동되면서 실행되는 워크플로우를 ⑥의 수명주기 관리 프로세스가 제어하는 것이 기계학습 운영화의 핵심 개념이다.

IV. DSML 프로세스 풍경의 상세

[그림 3]에 나타낸 DSML 프로젝트의 프로세스 풍경이 지속적 통합(CI) 및 지속적 배포(CD)를 전제로 한다는 점은 DevOps와 유사하지만, 접근 방식 스케치, 데이터 품질, 거버넌스, 모델 배포, 생산 적용, 드리프트 추적(drift tracking) 등은 기계학습 운영화에서 특히 중요한 위치를 차지하고, 데이터 의존적이고 실험적인 특성으로 인해 컨셉 드리프트의 영향이 크고 역방향 피드백이 자주 발생한다는 점이 DevOps와 다르다고 할 수 있다.

이하에서 각 단계를 구성하는 프로세스의 개념·특징 및 핵심 기능을 설명한다.

1. 계획 수립 단계

비즈니스 목표 설정 프로세스에서는 주요 비즈니스 실행을 식별하고, 비즈니스 목표·경로를 분석하여 프로젝트 실행 범위와 달성 여부의 평가 항목을 정의하고, 비즈니스 문제를 기계학습 문제로 변환하여 해결하는 타당성을 검토해야 한다. 요구사항 정의 프로세스

에서는 산출물·제품·서비스 등의 고객 요구사항을 분석하여 명세화하고 검증하며, 요구사항 도출·식별·분석, 요구사항 검증 및 명세 작성 등을 수행한다. 프로젝트 설계 프로세스에서는 요구사항에 맞춘 개념 설계를 통해 프로젝트 실행 범위, 업무 분담, 수행 일정 등을 결정하고, 위험 및 구현 가능성 등을 계량하여 요구사항의 달성을 판단하는 기준선을 설정하고 자원·비용·제한 등을 추정한다.

전문 인력 배정 프로세스에서는 프로젝트 실행에 참여하는 전문가를 위한 다학제 커뮤니케이션 프로토콜을 확립하며, 학술적 장벽을 극복하고 협업 문화를 확립한다. 접근 방식 스케치 프로세스에서는 프로젝트 실행 인프라의 토대 위에서 기계학습 운영화 프로세스를 설계하여 작은 프로젝트로 구현하면서, 파이프라인 시험과 생산 데이터를 점검한다. 사용자 경험의 추적과 프로젝트의 유효성·신뢰성 등이 확인되면 조금씩 실행 규모를 늘리면서 프로젝트를 완성해 나갈 수 있도록 접근 방식을 스케치한다.

2. 데이터 준비 단계

데이터 수집 프로세스는 다양한 데이터 소스의 블렌딩·결합, 중앙집중화 및 표준화·일관화, 데이터 추적·관리, 규정 준수·규제, 트랜잭션 통합 등을 수행하여, 데이터 웨어하우스·마트나 데이터 레이크를 통해 통일되고 일관된 통합 접근을 제공한다. 데이터 분석 프로세스는 미리 준비된 다양한 기능을 이용하여 데이터 가공을 위해 결측값·이상값 처리, 샘플링·비닝·파티셔닝, 구조적 오류 수정, 정규화·표준화·일반화 등을 제공하고, 데이터 정제를 위해 다크 데이터 분석, 오류 모니터링, 중복·오염·오류·오타 제거, 데이터 검증·감사 등을 제공하고, 데이터 비식별화를 위해 익명화·가명화, 대체·삭제·범주화, 유사 식별자 억제 및 재식별 방지, 데이터·컨텍스트 리스크 측정 등을 수행한다.

특징 공학 프로세스는 도메인 지식이나 데이터 마이닝 기술을 토대로 특징을 추출·선택·필터링·변환·강조하고, 차원 감소, 심층 특징 합성 등을 수행하며, 원시 데이터에 내재된 특성을 더 잘 나타내는 구조로 변환·단순화하여, 고급 데이터 분석이나 기계학습 모델링의 성능을 개선할 수 있도록 한다. 데이터 탐색·시각화 프로세스는 다양한 데이터 계보를 정리·연결하면서, 데이터 탐색을 위해 고유치·빈도 계산, 파레토 분석, 상관 히트맵, 시계열·시공간 분석, 컨조인트 분석, 통계적 유의성 분석 등을 제공하고, 데이터

시각화를 위해 차트·테이블 시각화, 지리·그래프 시각화, 다차원 시각화, 네트워크 시각화, 리포트 생성 등을 수행하면서, 대시보드나 인포그래픽 등으로 사용자 상호작용을 제공한다. 데이터 품질 프로세스는 데이터의 완전성·고유성·적시성·유효성·무결성·정확성·일관성 등을 검증하고, 여러 품질 차원에서 데이터를 측정·평가·검증·보증·관리한다.

3. 데이터 분석 및 관리 단계

고급 데이터 접근 프로세스는 이기종 소스 데이터의 단일 구조 투영 및 일관성 있는 접근·관리를 지원하는 데이터 패브릭 환경을 제공하며, 고도의 대규모성·민첩성·속도성 지원과 데이터의 원활한 혼합·변환·통합 및 균질화·무결성 검증 등을 지원하는 데이터 매핑을 통해 데이터 활용 속도 및 능력의 증대를 제공한다. 고급 시각화 프로세스는 네트워크·연결성을 가지고 존재하는 데이터를 수집·분석·요약하고 탐색적 데이터 분석을 실행하면서, 패턴 추출, 경로·연결성 분석, 타임라인·네트워크 시각화, 계층적·다차원 시각화, 네트워크·지리 공간 시각화 등을 수행한다. 고급 데이터 분석 프로세스는 수학·통계 이론이나 시뮬레이션을 중심으로 정보 추출, 패턴 마이닝, 예측·처방 분석, 결정 분석, 추세 분석, 연관성·상관성 분석, 네트워크 및 클러스터 분석, 디지털 트윈 및 시뮬레이션 등을 실시하여 복잡한 이벤트 처리 및 미래 이벤트 예측, 통찰 요인 및 권장 사항의 도출을 수행한다.

데이터 통찰 프로세스는 고급 데이터 분석을 통해 인간 지능을 보완하는 통찰력을 생성하고, 개인화된 제품·서비스·콘텐츠 등을 추천하여 고객 경험 개선 및 비즈니스 가치 향상을 도모하며, 트렌드·수요 분석, 관련성·특성 분석, 비즈니스 모니터링, 개인화·초개인화 서비스 등을 수행한다. 증강 데이터 관리 프로세스는 데이터의 자동 수집·정제·구성, 관리 효율화, 생산성 향상 등을 제공하고, 데이터 품질·통합 등 관리 전반을 개선하여 비즈니스 의사 결정을 가속화하고, 자동 데이터 분석 및 데이터과학 민주화를 지원한다. 데이터 보안 프로세스는 데이터 접근 제어, 손실 방지, 규정 준수 등을 위한 물리적·논리적 보안을 제공하며, 데이터 위험 평가, 보안 모니터링, 데이터 보호 및 손실 방지, 백업·복구, 데이터 감사 등을 수행한다.

4. 기계학습 단계

데이터 라벨링 프로세스는 반자동·자동·합성 라벨링, 가이드 라벨링, 데이터 스키마 검증 등을 수행한다. 모델 학습 프로세스는 기계학습 및 심층학습 모델 매개변수의 최적화를 지원하고, 모델 튜닝 프로세스는 모델의 구조와 학습 환경을 제어하기 위한 초매개변수의 최적화를 수행한다. 모델 평가 프로세스는 모델의 성능·유효성·신뢰성 등을 평가하고 결함·오류 등을 검증한다. 자동 기계학습 프로세스는 모델 학습·튜닝의 자동화와 기계학습 민주화를 지원하고, 시민 데이터과학자를 지원하기 위해 휴리스틱 지식 제공, 워크플로우 추천, 셀프서비스 환경 지원 등을 수행한다.

모델 스캐폴딩 프로세스는 학습 모델이 다른 데이터나 컴퓨팅 환경에서도 잘 작동되도록 하는 모델 압축 및 전이 학습을 제공하고, 고유한 특성을 가진 데이터를 신규 생성하는 적대적 생성을 수행한다. 모델 압축은 규모가 크고 계산 능력이 뛰어난 심층학습 모델에서 가볍고 계산 속도가 빠른 구조의 모델로 변환을 지원한다. 전이 학습은 한 모델의 지식을 다른 모델로 이식하는 모델 구조 변경 또는 매개변수 최적화를 지원한다. 적대적 생성은 대립하는 두 개의 심층신경망(생성자와 구분자)을 적대적으로 학습시킨 후, 고품질 데이터를 생성하는 생성자의 특성을 이용하여 새로운 데이터를 생성하거나 기존 네트워크를 공격·방어하는 데이터를 생성할 수 있도록 하며, 생성적 적대 신경망(GAN) 및 적대적 공격·방어 등을 수행한다.

5. 배포 및 생산화 단계

컨테이너화(containerization) 프로세스는 리소스 독립 환경에서 기계학습 서비스의 구성·배포·관리·확장·네트워킹을 자동화하고, 이식 가능하며 유연하고 독립적인 실행 환경을 제공하여 기계학습 운영화의 생산성을 높일 수 있도록 지원한다.

거버넌스 프로세스는 데이터와 모델의 거버넌스로 구분된다. 데이터 거버넌스는 데이터의 계보·버전 관리, 데이터 정책·절차의 준수, 학습·생산 데이터의 일치성 유지, 데이터 투명성·책임성 보증 등을 통해 데이터 자산의 품질 관리와 데이터 관리의 효율성을 향상시킨다. 모델 거버넌스는 모델 아티팩트의 관리·추적, 모델의 편향·왜곡 검증, 적대적 견고성 검증, 모델 품질 인증, 모델 투명성·책임성 보증 등을 통해 데이터 의존적

특성에 대한 관리 체계를 제공하고, 윤리적 편견, 예측하지 못한 위험, 적대적 공격 등의 위험에 대비할 수 있는 통제 수단을 제공한다.

모델 배포 프로세스는 학습 모델의 계보 관리, 학습 아티팩트 관리, 모델 유효성 검사와 학습 이력, 실행 요구 조건, 신뢰도 정보 정리 등을 수행한다. 학습 모델을 배포 형식으로 패키징하여 모델 등록소에 등재하거나, 생산 데이터 준비 및 모델의 실행을 위한 모든 종속성을 캡슐화하여 특징 저장소에 등재하고 관리한다.

생산화 프로세스는 배포 모델이 생산 현장에서 제대로 동작하도록 이식성 검증, 특징 저장소 및 특징 메타데이터 관리, 모델·인프라 호환성 검증, 생산 리소스 검증, 모델 성능 실험 등을 통해 검증한다. 또한, 생산 현장의 일부에 릴리스하여 계약 테스트 및 실험 결과를 비교하면서 롤아웃·롤백 여부의 최종 결정을 지원하고, 장애 조치, 생산 이력 추적, 파이프라인 트리거 등을 지원한다.

드리프트 추적 프로세스는 인간-기계학습 상호작용을 토대로 생산 데이터 드리프트 감시, 실험 추적 및 모델 드리프트 모니터링 등을 지원하여 생산 데이터의 특성 변질이나 모델 붕괴를 추적·감시하고, 드리프트 파이프라인 트리거링 및 모델 재훈련·대체 등을 수행한다.

6. 풀 스택 지원 단계

협업 환경 프로세스는 프로젝트 참여자가 온라인·실시간·애자일 환경에서 지식 관리, 정보 공유, 협업, 채팅·미팅 등을 할 수 있는 기능을 제공한다. 전문가 개발 환경 프로세스는 통합 개발 및 노트북 환경, 코딩 환경, 기계학습 라이브러리·프레임워크 연동 환경 등을 제공하여 전문가의 솔루션 개발과 사용자 정의 구조·기능 확장을 지원하고, 온라인 사용자 포럼이나 오픈소스 커뮤니티 등을 통한 정보 공유 환경을 제공한다. 클라우드 연동 프로세스는 클라우드 호스팅, 클라우드 커넥터, 클라우드 스토리지 연결, 고성능 컴퓨팅(GPU, TPU 등) 서비스 지원 등을 통해 온프레미스 환경 또는 퍼블릭 클라우드 접근을 제공한다.

수명주기 관리 프로세스는 통합 메타데이터 저장소, 특징 등록소, 학습 아티팩트 등록소, 데이터·코드·모델 통합 추적, 지속적 통합·시험·배포, 생산 모니터링, 파이프라인 트리거, 파이프라인 오케스트레이션, 올인원 프로젝트 추적, 수명주기 통합 관리 등을 지

원한다. 또한, 프로젝트 수명주기 전반에서 비즈니스 프로세스의 무결성 보장을 위해 응집성·일관성을 가지고 기계학습 파이프라인 및 프로젝트 아티팩트를 자동화하여 구축·배포·관리·추적하며, DSML 프로젝트를 효율적으로 제어하고 통합 관리할 수 있도록 지원한다.

V. 결론

인공지능의 한 부분으로 DSML과 기계학습 운영화를 비즈니스 프로세스에 적용하려는 시도에도 불구하고 아직도 대부분의 글로벌 기업에서 가시적 성과를 올리지 못하고 있는 것이 현실이다. 데이터 준비·통합은 학습 데이터보다는 생산 데이터에 중점을 두어야 하며, 일단 드리프트가 발생하면 파이프라인의 부분 수정보다는 데이터 준비·모델링·배포에 이르는 전체 과정을 다시 점검할 수도 있는 것이 DSML 프로젝트에서 발생하는 데이터 의존성의 깊은 함의라고 할 수 있을 것이다. 드리프트 추적 및 트리거는 사람과 인공지능의 원활한 상호작용에 의존해야 하며, 프로젝트의 모든 파이프라인과 워크플로우는 사람이 감시·이해·대응할 수 있도록 구성해야 한다. DSML 프로젝트는 데이터 사일로의 해체뿐만 아니라, 소통을 가로막는 인적 사일로와 학술적 장벽의 해체를 통해 시작되어야 한다. DSML 프로젝트 실행을 위한 인프라도 중요하지만, 비즈니스 실행에 이르기까지 피드백이 반복되는 기계학습 운영화 과정을 극복할 수 있도록, 사람들 사이의 협업·소통 문화를 더 중요시하는 것이 기계학습 운영화 실행 전략의 핵심이라고 할 수 있을 것이다. DSML 프로젝트를 통해 비즈니스 데이터 기반의 통찰을 얻고자 하는 분들에게 본 고가 작은 도움이라도 줄 수 있기를 희망한다.

[참고문헌]

- [1] Gartner, “Gartner Top Strategic Predictions for 2019 and Beyond,” 2018. 10.
- [2] Gartner, “Top Trends on the Gartner Hype Cycle for Artificial Intelligence, 2019,” 2019. 9.
- [3] Gartner, “2 Megatrends Dominate the Gartner Hype Cycle for Artificial Intelligence, 2020,” 2020. 9.
- [4] Forbes, “COVID-19 Is a Before-and-After Moment in the Digital Transformation,” 2020. 3.

- [5] CIO, "How AI Can Help You Build a Picture of Your Digital Business," 2020. 8.
- [6] IDC, "AI Strategiesview 2020: Executive Summary," 2020. 4.
- [7] McKinsey, "Global Survey: The State of AI in 2020," 2020. 11.
- [8] RTInsights, "Forrester: Enterprise ML Development Maturing," 2020. 5.
- [9] UiPath, "C-Level Executives Want Employees to Have Automation and AI Skills: Survey," 2020. 11.
- [10] Analytics India, "Why Majority of Data Science Projects Never Make It to Production," 2020. 6.
- [11] Cognilytica, "ML Model Management and Operations 2020," CGR-MOM20, 2020. 3.
- [12] Forbes, "MLOps: What You Need to Know," 2020. 8.
- [13] J. Tobin, S. Karayev, and P. Abbeel, "Full Stack Deep Learning," Spring 2019 Full Stack Deep Learning Bootcamp 강의 자료, 2019. 3.
- [14] 조성익, "기계학습 운영화(MLOps)의 프로세스 풍경," ETRI Technical Memo, 2020. 12.
- [15] Sigmiod, "5 Challenges to Be Prepared for Before Scaling Machine Learning Models," 2020. 9.
- [16] Sigmiod, "5 Best Practices for Putting Machine Learning Models into Production," 2020, 10.
- [17] D. Sculley, et al. "Hidden Technical Debt in Machine Learning Systems," NIPS 2015.
- [18] 이연희, 표철식, "BeeAI: KSB 인공지능 서비스 플랫폼," OSIA S&TR Journal, 33, 2020. 5.
- [19] 이연희, 표철식, "AIoT를 위한 지능 서비스 소프트웨어 플랫폼 전략", 정보과학회지, 38(8), 2020. 8.
- [20] 삼성 SDS, <https://www.samsungsds.com/kr/ai/ai.html>