머신러닝 파이프라인

머신러닝 파이프라인 단계

송호연

목차

머신러닝 파이프라인 단계

학습목표

머신러닝 파이프라인의 이해

01. 머신러닝 파이프라인 단계에 대해 이해한다.

머신러닝 파이프라인의 각 단계에 대해 이해한다.

02. 데이터 수집, 버전관리, 데이터 검증의 특징을 이해한다.

머신러닝 기반의 소프트웨어가 갖고 있는 특징을 이해한다.

03. 모델 학습, 모델 분석, 모델 버전 관리에 대해 이해한다.

데이터의 품질을 체크하는 데이터 검증에 대해 이해한다.

04. 모델 배포, 피드백 루프 반복, 개인정보 보호에 대해 이해한다.

모델 학습에 따른 특징을 이해한다.

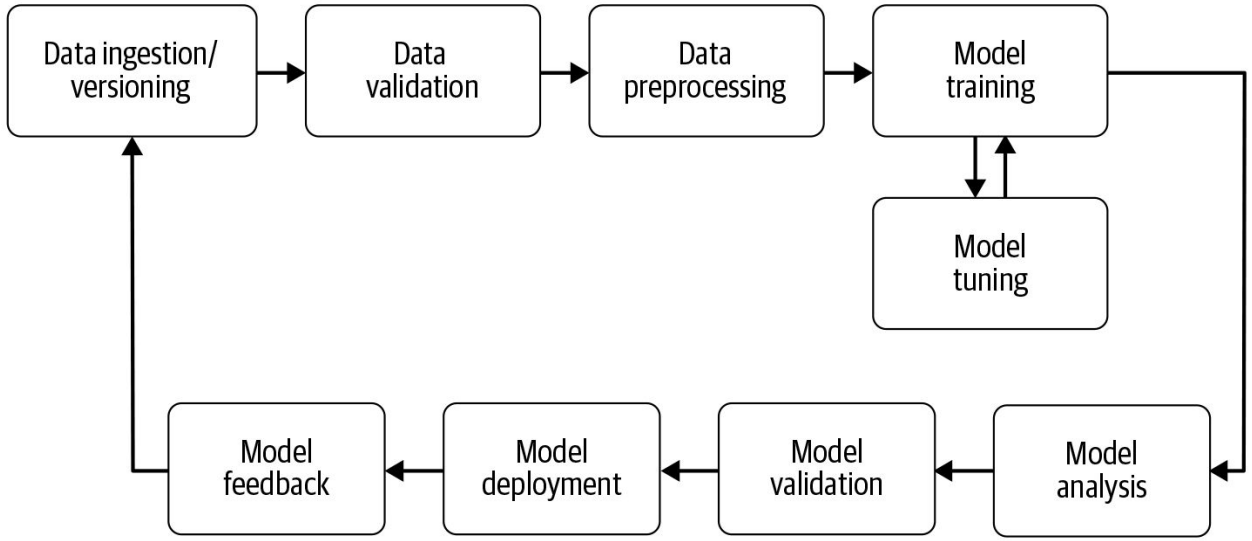
머신러닝 파이프라인 단계 개요

머신러닝 파이프라인 단계

머신러닝 파이프라인 개요

머신러닝 파이프라인 단계

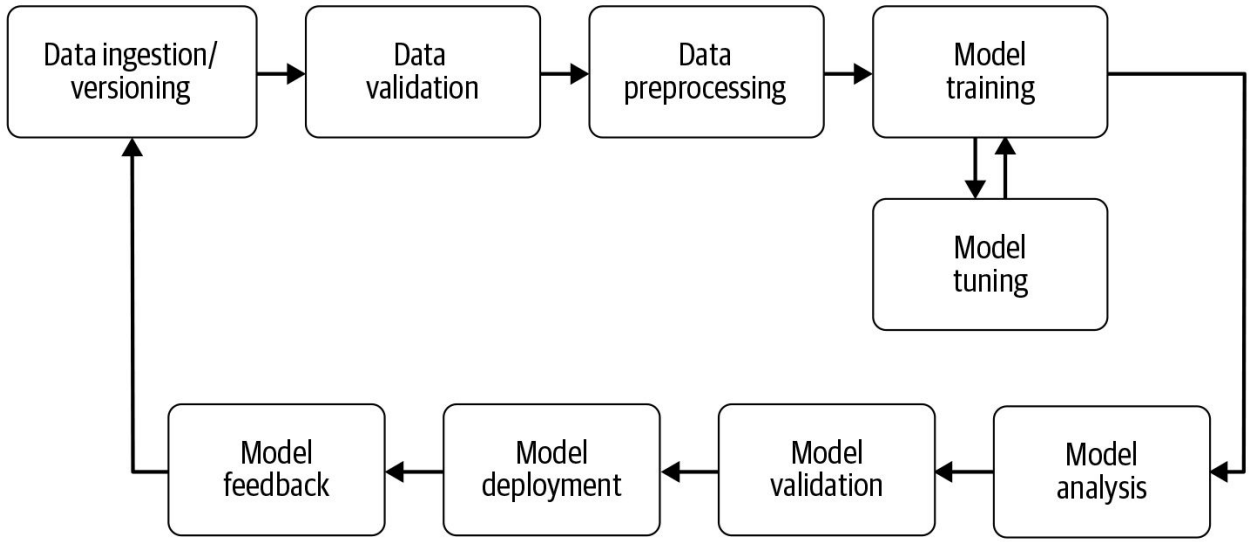
머신러닝 파이프라인은 새로운 학습 데이터를 수집하는 것으로 시작하여 새로 학습된 모델이 어떻게 작동하고 있는지에 대한 피드백을 받는 것으로 끝납니다. 이 피드백은 프로덕션 성능 메트릭 또는 제품 사용자의 피드백일 수 있습니다.



머신러닝 파이프라인 개요

머신러닝 파이프라인 단계

파이프라인에는 데이터 전처리, 모델 학습 및 모델 분석, 모델 배포 등 다양한 단계가 포함됩니다. 이러한 단계를 수동으로 수행하는 것은 번거롭고 오류가 발생하기 쉽다는 것을 상상할 수 있습니다.



머신러닝 파이프라인 개요

머신러닝 파이프라인 단계

데이터가 많을수록 일반적으로 모델이 개선됩니다. 이러한 지속적인 데이터 유입으로 인해 자동화가 핵심입니다. 실제 애플리케이션에서는 모델을 자주 재학습하려고 합니다. 그렇지 않으면 대부분의 경우 모형이 예측하는 새 데이터와 학습 데이터가 다르기 때문에 정확도가 저하됩니다.

머신러닝 파이프라인 개요

머신러닝 파이프라인 단계

재학습이 새로운 학습 데이터를 수동으로 검증하거나 업데이트된 모델을 분석해야 하는 수동 프로세스인 경우 데이터 과학자 또는 머신러닝 엔지니어는 전혀 다른 비즈니스 문제에 대한 새로운 모델을 개발할 시간이 없습니다.

데이터 수집, 버전관리, 데이터 검증

데이터 수집

버전 관리

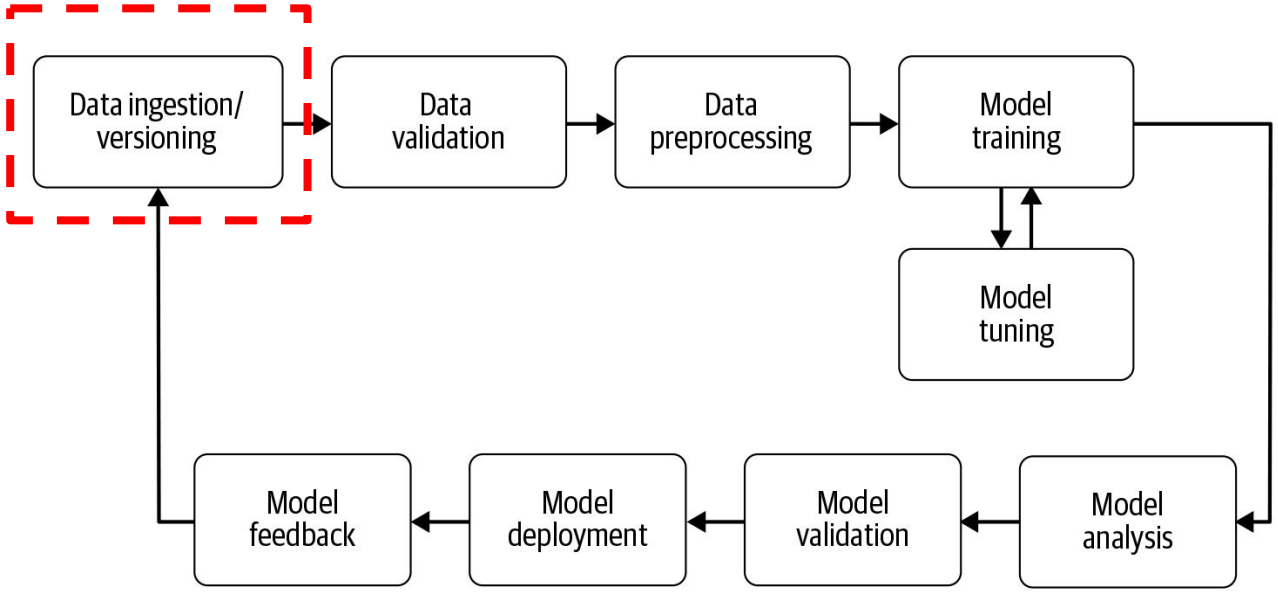
데이터 검증

02

데이터 수집, 버전관리, 데이터 검증

데이터 수집

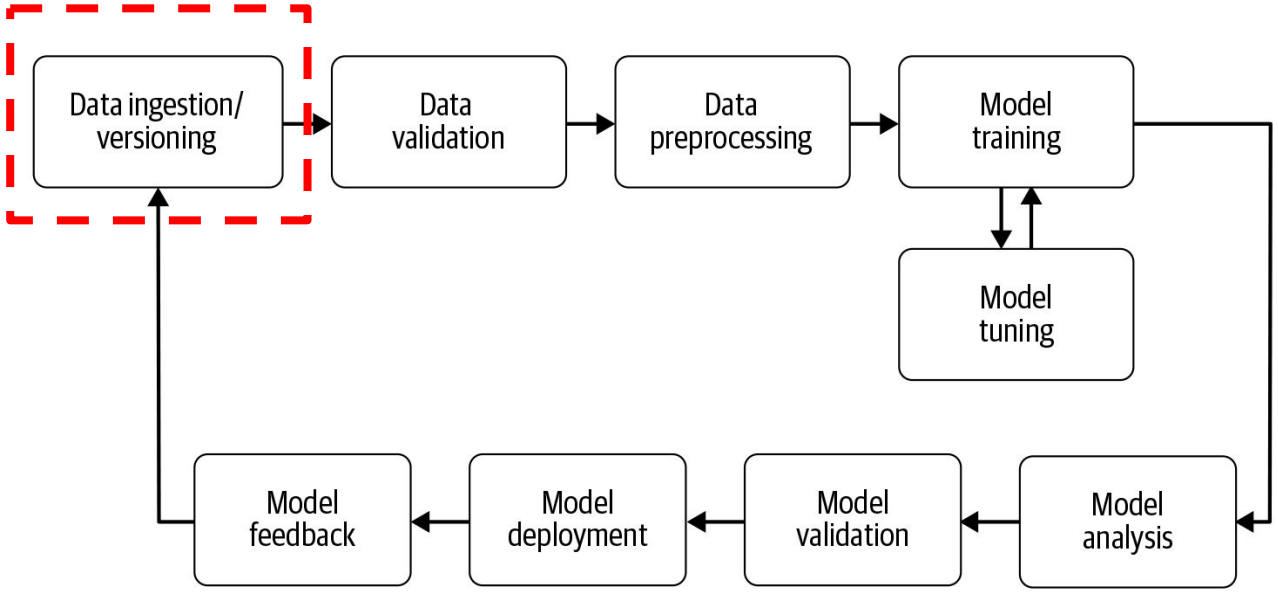
데이터 수집은 모든 머신러닝 파이프라인의 시작입니다. 이 파이프라인 단계에서는 다음 구성 요소가 소화할 수 있는 형식으로 데이터를 처리합니다. 데이터 수집 단계에서는 피쳐 엔지니어링을 수행하지 않습니다(데이터 유효성 검사 단계 후에 수행됨).



데이터 수집, 버전관리, 데이터 검증

데이터 버전관리

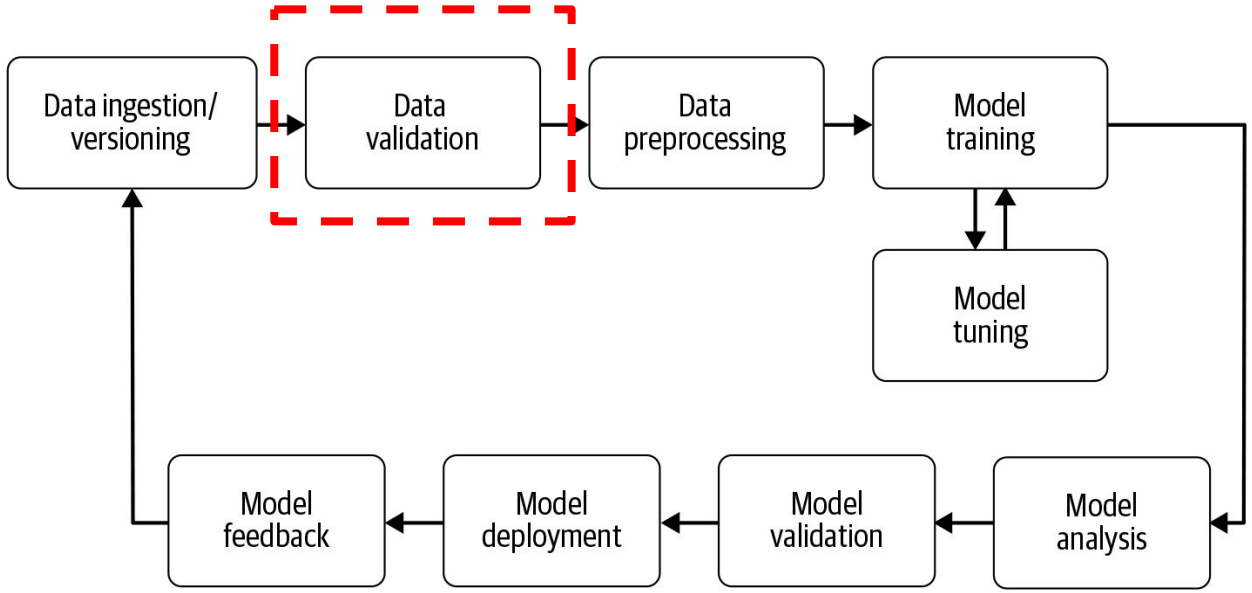
또한 들어오는 데이터를 버전 관리하여 데이터 스냅샷을 파이프라인의 끝에 있는 학습된 모델과 연결하는 것도 중요합니다.



데이터 수집, 버전관리, 데이터 검증

데이터 유효성 검사

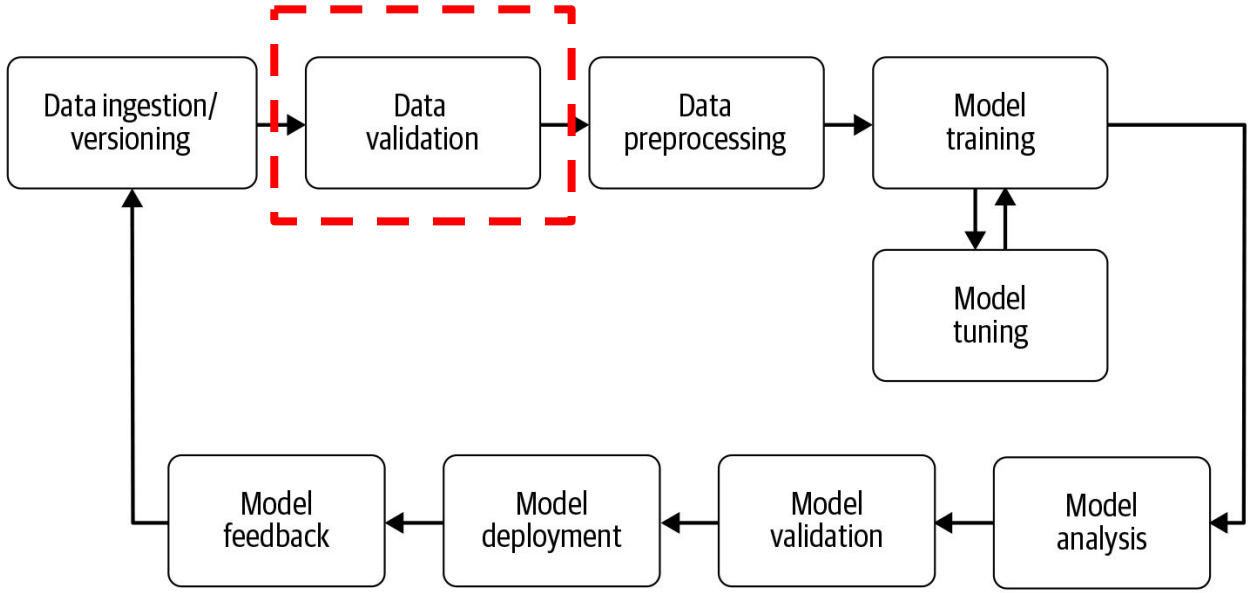
새 모델 버전을 학습하기 전에 새 데이터를 검증해야 합니다. 데이터 유효성 검사는 새 데이터의 통계가 예상대로인지 확인하는 데 초점을 맞춥니다(예: 범주의 범위, 범주 수 및 분포). 또한 이상 징후가 감지될 경우 데이터 과학자에게 경고합니다.



데이터 수집, 버전관리, 데이터 검증

데이터 유효성 검사

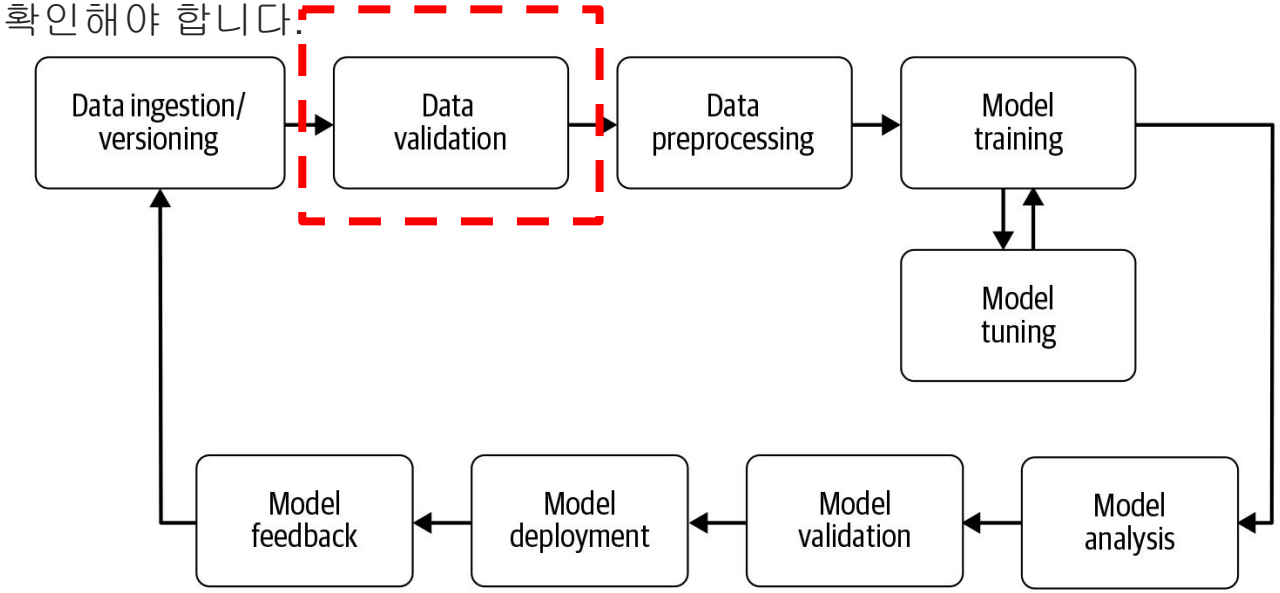
예를 들어, 이항 분류 모형을 학습하는 경우 학습 데이터에는 클래스 X 표본의 50%와 클래스 Y 표본의 50%가 포함될 수 있습니다. 데이터 유효성 검사 도구는 이러한 클래스 간의 분할이 변경될 경우 경고를 제공합니다. 여기서 새로 수집된 데이터는 두 클래스 간에 70/30으로 분할될 수 있습니다.



데이터 수집, 버전관리, 데이터 검증

데이터 유효성 검사

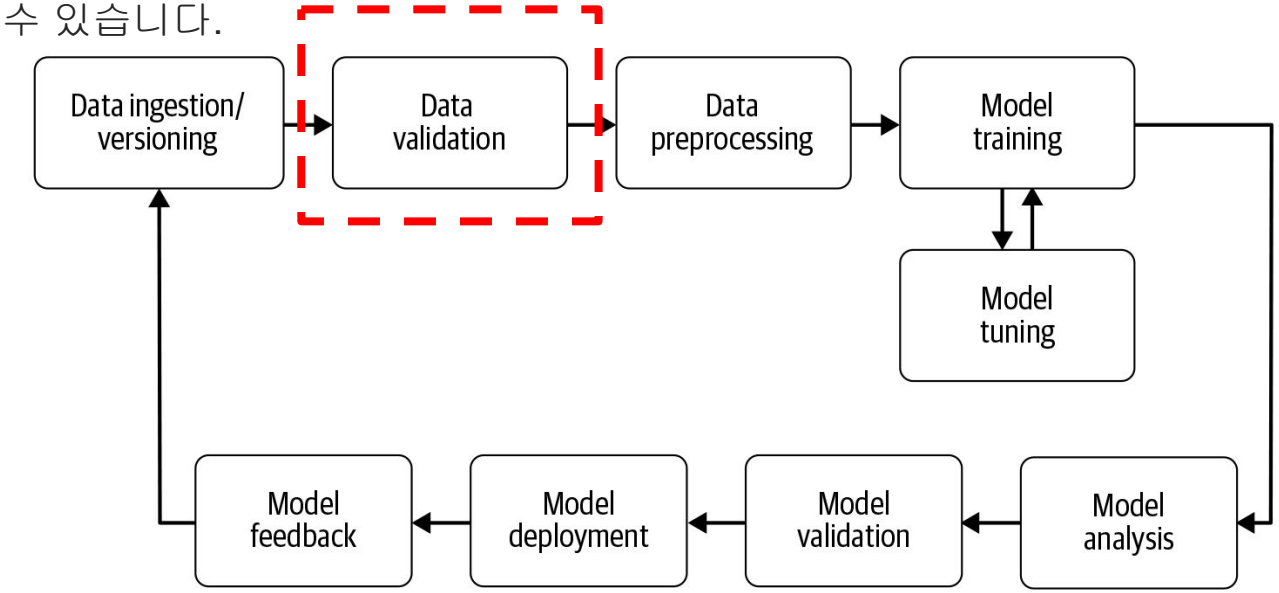
모형이 불균형한 훈련 집합으로 훈련되고 있는데 데이터 과학자가 모형의 손실 함수를 조정하지 않거나 X 또는 Y 범주를 과소 샘플링하지 않은 경우, 모형 예측은 지배적인 범주에 치우칠 수 있습니다. 또한 공통 데이터 검증 도구를 사용하여 다양한 데이터셋을 비교할 수 있습니다. 상위 레이블이 있는 데이터 집합이 있고 데이터 집합을 학습 및 검증 집합으로 분할하는 경우 두 데이터 집합 간에 레이블 분할이 거의 동일한지 확인해야 합니다.



데이터 수집, 버전관리, 데이터 검증

데이터 유효성 검사

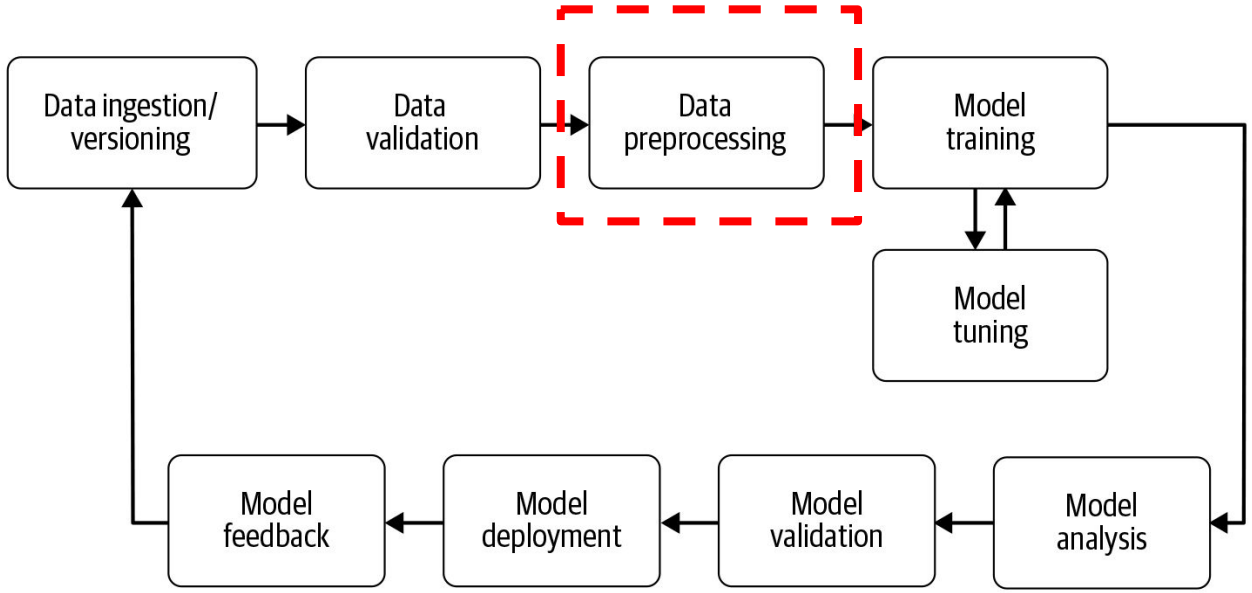
데이터 검증 도구를 사용하여 데이터셋을 비교하고 이상 징후를 강조할 수 있습니다. 검증에서 특이한 점이 발견되면 여기서 파이프라인을 중지하고 데이터 과학자에게 경고를 보낼 수 있습니다. 데이터 이동이 감지되면 데이터 과학자 또는 머신러닝 엔지니어는 개별 클래스의 샘플링을 변경하거나(예: 각 클래스에서 동일한 수의 예만 선택) 모델의 손실 피쳐을 변경하고 새 모델 빌드 파이프라인을 시작한 후 수명 주기를 다시 시작할 수 있습니다.



데이터 수집, 버전관리, 데이터 검증

데이터 전처리

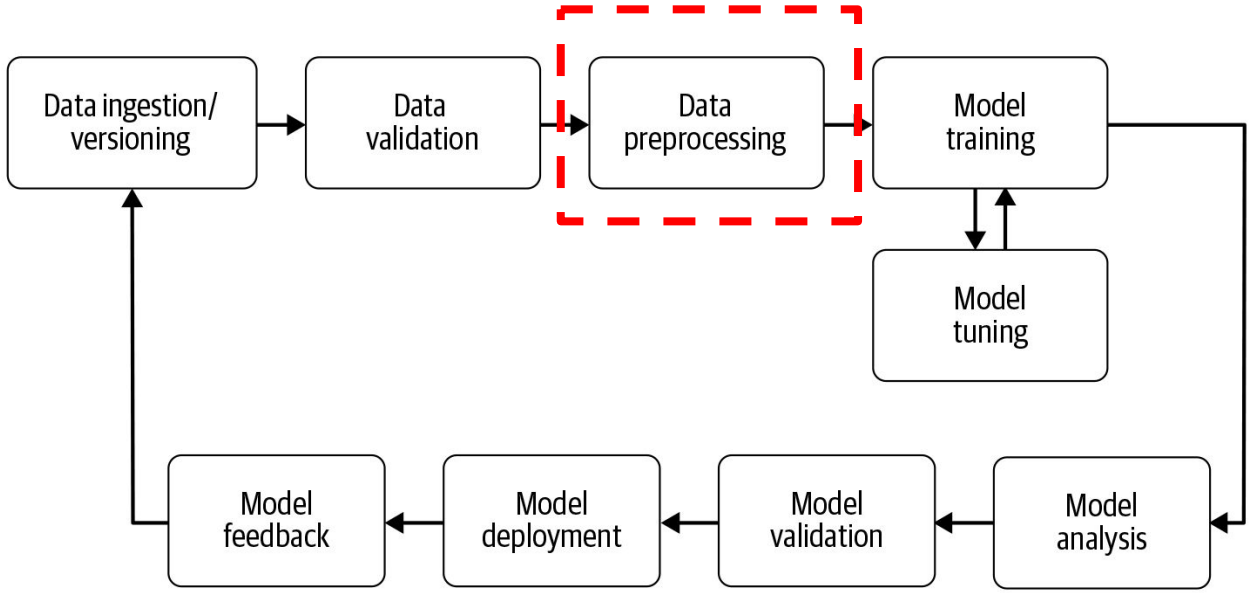
새로 수집한 데이터를 사용하고 머신러닝 모델을 직접 학습할 수 없을 가능성이 높습니다. 대부분의 경우 학습 실행에 사용하기 위해 데이터를 미리 처리해야 합니다. 레이블은 종종 하나 또는 다중 열 벡터 로 변환되어야 합니다. 모델 입력에도 동일하게 적용됩니다. 텍스트 데이터에서 모델을 학습하는 경우 텍스트의 문자를 인덱스로 변환하거나 텍스트 토큰을 워드 벡터로 변환할 수 있습니다.



데이터 수집, 버전관리, 데이터 검증

데이터 전처리

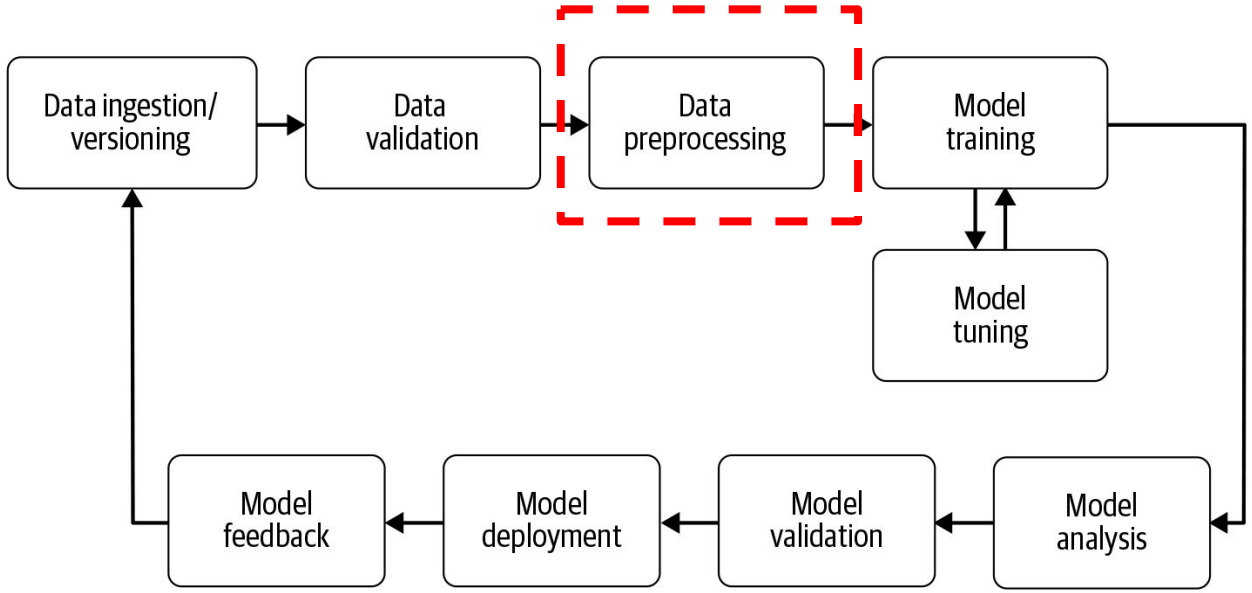
전처리가 모든 학습 에폭마다 실행될 필요는 없습니다. 모델 학습 전에 처리되기만 하면 되므로, 모델을 학습하기 전에 학습 라이프사이클 바로 전 단계에서 전처리를 실행하는 것이 가장 합리적입니다. 데이터 전처리 도구는 단순한 파이썬Python 스크립트부터 정교한 그래프 모델에 이르기까지 다양합니다.



데이터 수집, 버전관리, 데이터 검증

데이터 전처리

대부분의 데이터 과학자는 각자가 선호하는 도구로 피쳐를 처리하는 데 집중하지만, 전처리 단계를 수정하면 데이터셋에 영향을 주고 반대로 데이터셋이 전처리 단계에 영향을 줍니다. 즉, 처리 단계를 수정하는 경우 (예: 원핫벡터 변환에서 추가 라벨 허용), 이전 단계가 더 이상 유효하지 않게 되고 전체 파이프라인을 강제로 업데이트해야 합니다.



모델 학습, 모델 분석, 모델 버전 관리

모델 학습

모델 분석

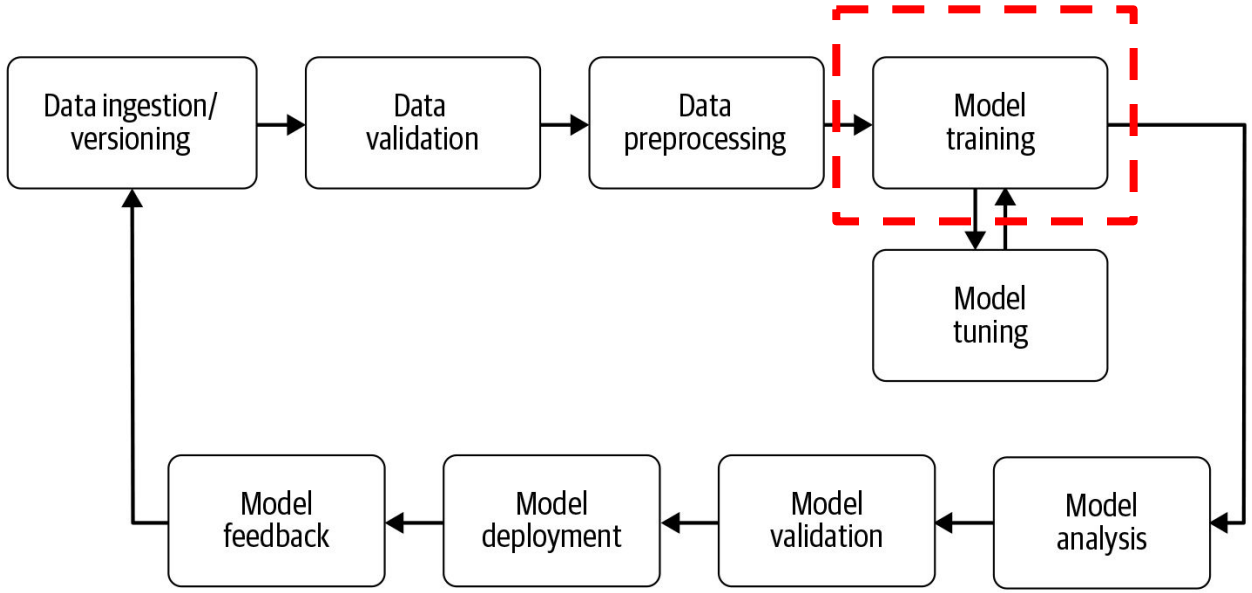
모델 버전 관리

03

모델 학습, 모델 분석, 모델 버전 관리

모델 학습

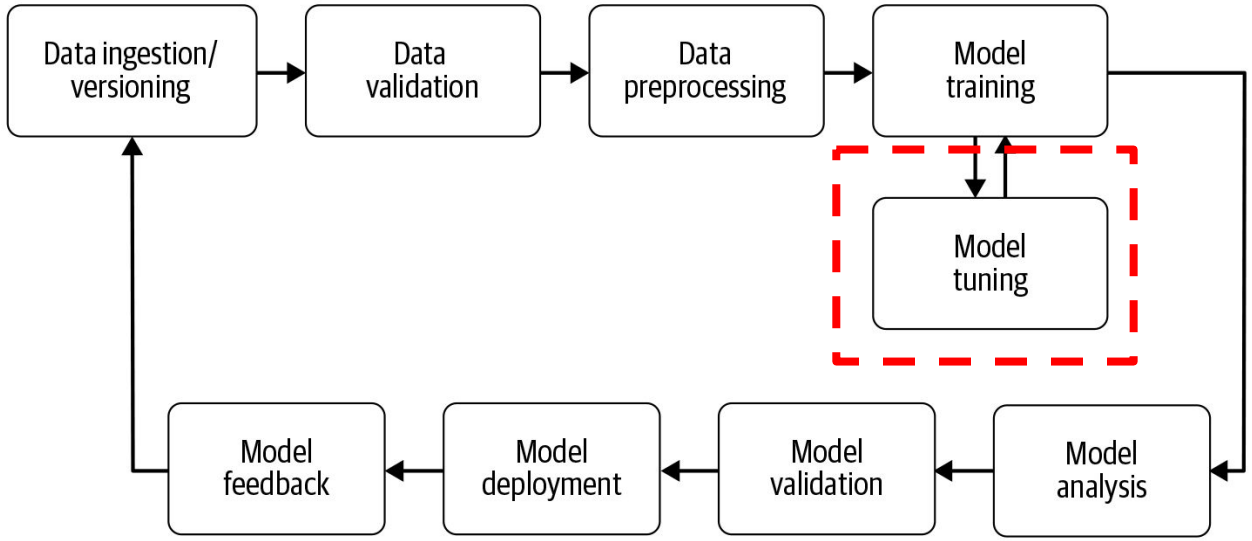
모델 학습 단계는 머신러닝 파이프라인의 핵심입니다. 이 단계에서는 가능한 가장 낮은 오차를 사용하여 입력을 수행하고 출력을 예측하는 모델을 학습합니다. 대규모 모델과 특히 대규모 학습 세트에서는 이 단계를 신속하게 관리하기가 어려워질 수 있습니다. 메모리는 일반적으로 우리의 계산에 한정된 자원이기 때문에, 모델 훈련의 효율적인 분포가 매우 중요합니다.



모델 학습, 모델 분석, 모델 버전 관리

모델 튜닝

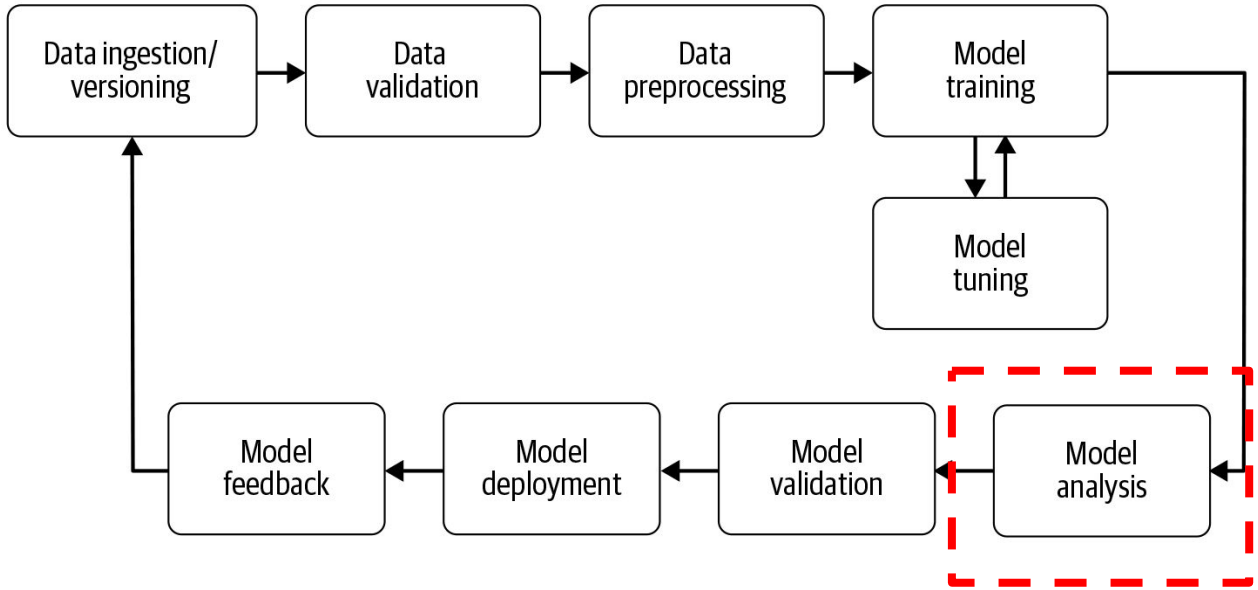
최근 모델 튜닝은 상당한 성능 개선과 경쟁 우위를 제공할 수 있기 때문에 많은 관심을 받고 있습니다. 머신러닝 프로젝트에 따라 머신러닝 파이프라인을 고려하기 전에 모델을 튜닝하도록 선택하거나 파이프라인의 일부로 튜닝하는 것이 좋습니다. 머신러닝 파이프라인 아키텍처는 확장 가능하기 때문에 다수의 모델을 병렬 또는 순차적으로 학습시킬 수 있습니다. 이를 통해 최종 모델에 적합한 모델 하이퍼 파라미터를 선택할 수 있습니다.



모델 학습, 모델 분석, 모델 버전 관리

모델 분석

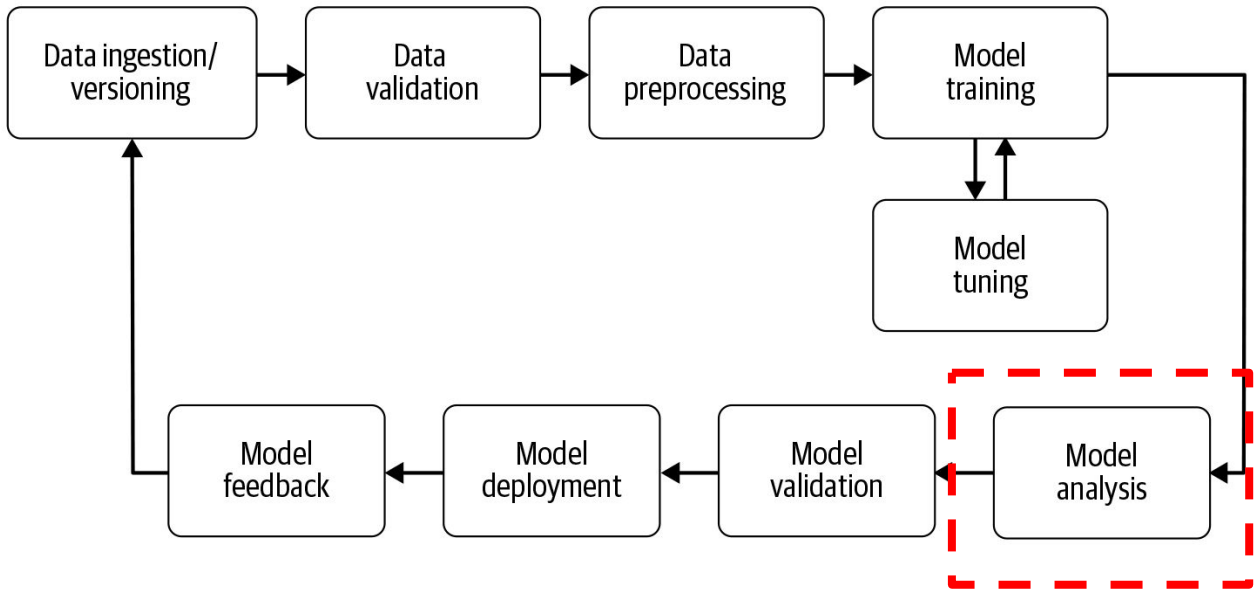
일반적으로 정확도 또는 손실을 사용하여 최적의 모형 모수 집합을 결정합니다. 그러나 모델의 최종 버전을 결정한 후에는 모델의 성능에 대해 보다 심층적으로 분석하는 것이 매우 유용합니다. 여기에는 정밀도, 리콜 및 AUC와 같은 다른 메트릭을 계산하거나 학습에 사용되는 검증 집합보다 더 큰 데이터 집합에서 성능을 계산하는 작업이 포함될 수 있습니다. 모델 분석을 심층적으로 하는 또 다른 이유는 모형의 예측이 공정하다는 것을 확인하기 위해서 입니다.



모델 학습, 모델 분석, 모델 버전 관리

모델 분석

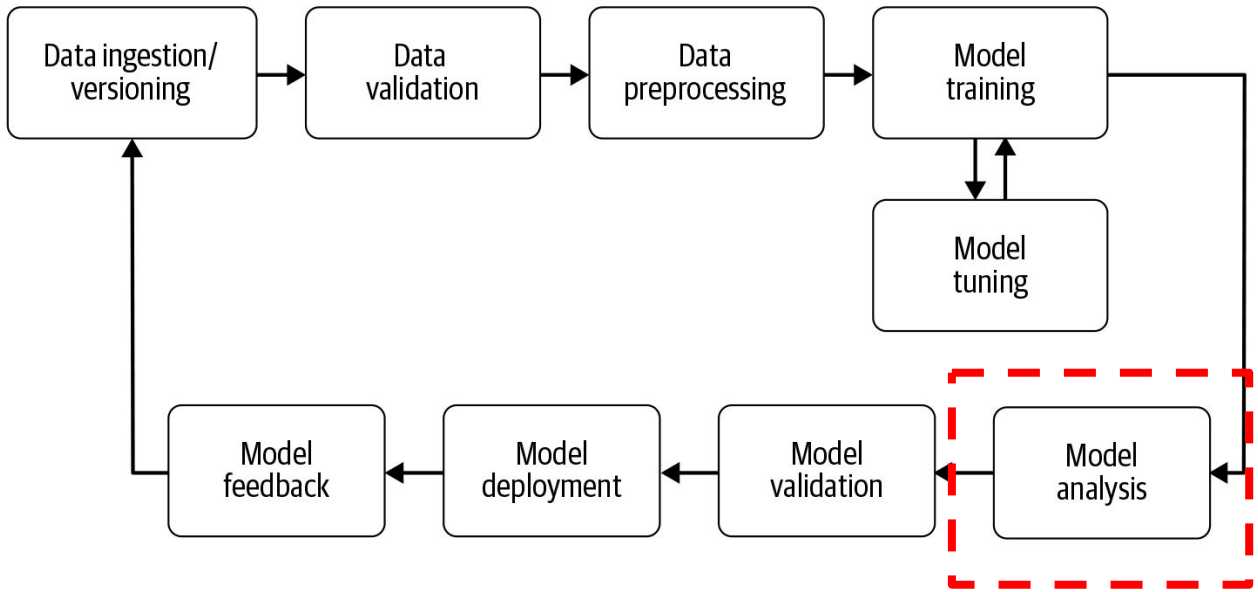
데이터 집합을 잘라내고 각 슬라이스의 성능을 계산하지 않는 한 모델이 여러 사용자 그룹에 대해 어떻게 작동하는지 알 수 없습니다. 또한 학습에 사용되는 피쳐에 대한 모델의 의존도를 조사하고, 단일 학습 예제의 피쳐을 변경할 경우 모델의 예측이 어떻게 변화하는지 살펴볼 수 있습니다.



모델 학습, 모델 분석, 모델 버전 관리

모델 분석

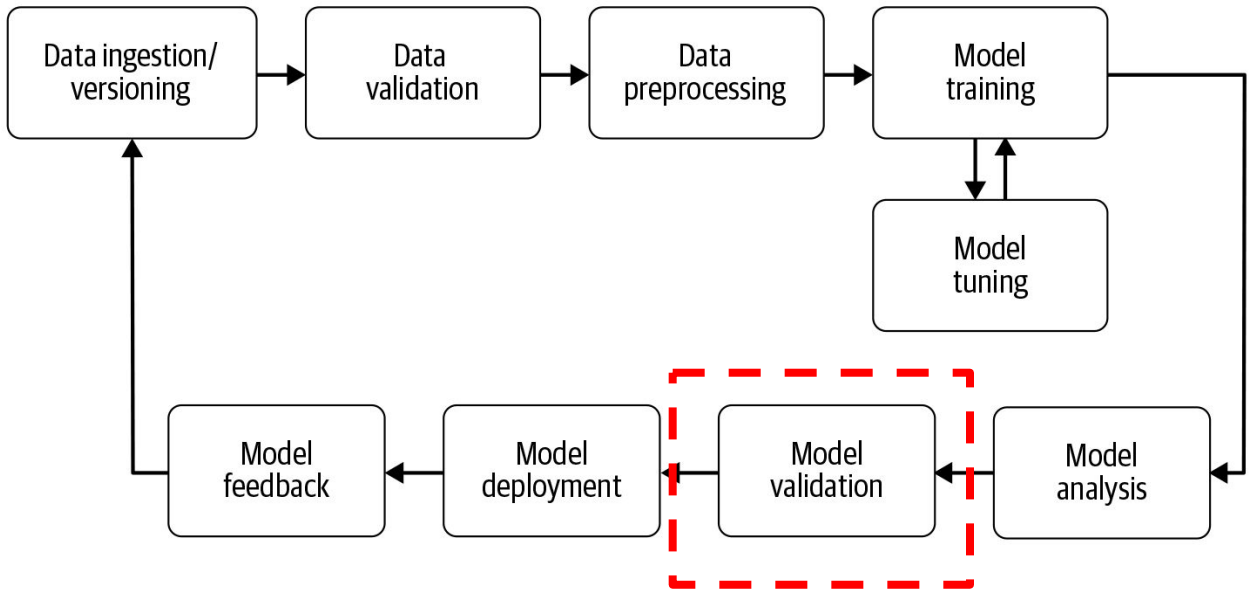
모델 조정 단계 및 최고의 성능을 갖춘 모델의 최종 선택과 마찬가지로, 이 워크플로우 단계에서는 데이터 과학자의 검토가 필요합니다. 다만, 최종적인 검토만으로 전체 분석을 자동화할 수 있는 방법을 시연해 보일 것입니다. 자동화를 통해 모델 분석이 다른 분석과 일관되고 비교 가능한 상태를 유지할 수 있습니다.



모델 학습, 모델 분석, 모델 버전 관리

모델 버전 관리

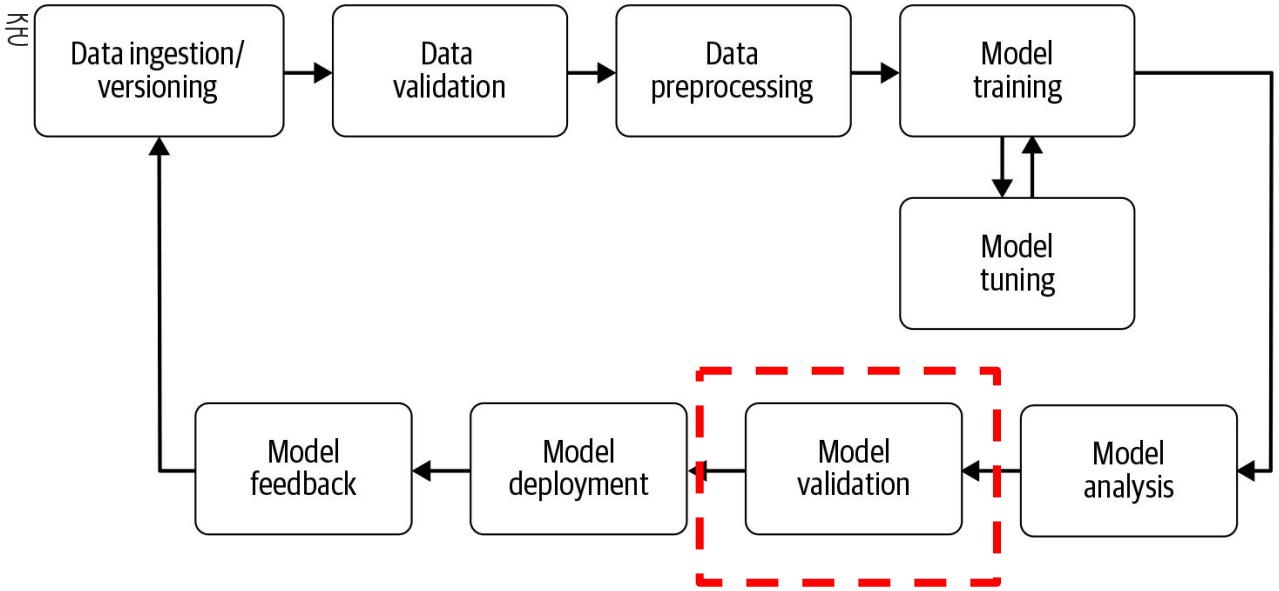
모델 버전 지정 및 검증 단계의 목적은 다음 버전으로 어떤 모델, 하이퍼 파라미터 세트 및 데이터셋이 선택되었는지 추적하는 것입니다.
소프트웨어 엔지니어링의 의미 버전 관리를 사용하려면 API에서 호환되지 않는 변경을 수행하거나 주요 피쳐을 추가할 때 메인 버전 번호를 늘려야 합니다. 그렇지 않으면 서브 버전 번호를 늘립니다.



모델 학습, 모델 분석, 모델 버전 관리

모델 버전 관리

모델 릴리스 관리에는 데이터세트라는 또 다른 파라미터가 있습니다. 학습 프로세스에 훨씬 더 많은 혹은 더 나은 데이터를 제공함으로써 단일 모델 매개변수 또는 아키텍처 설정을 변경하지 않고도 모델 성능의 상당한 차이를 달성할 수 있는 상황이 있습니다. 성능 향상이 주요 버전 업그레이드를 보장합니까? 이 질문에 대한 답변은 데이터 과학 팀마다 다를 수 있지만, 모든 입력을 새 모델 버전(하이퍼파라미터, 데이터셋, 아키텍처)으로 문서화하고 이 릴리스 단계의 일부로 추적하는 것이 중



모델 배포, 피드백 루프 반복, 개인 정보 보호

모델 배포

피드백 루프 반복

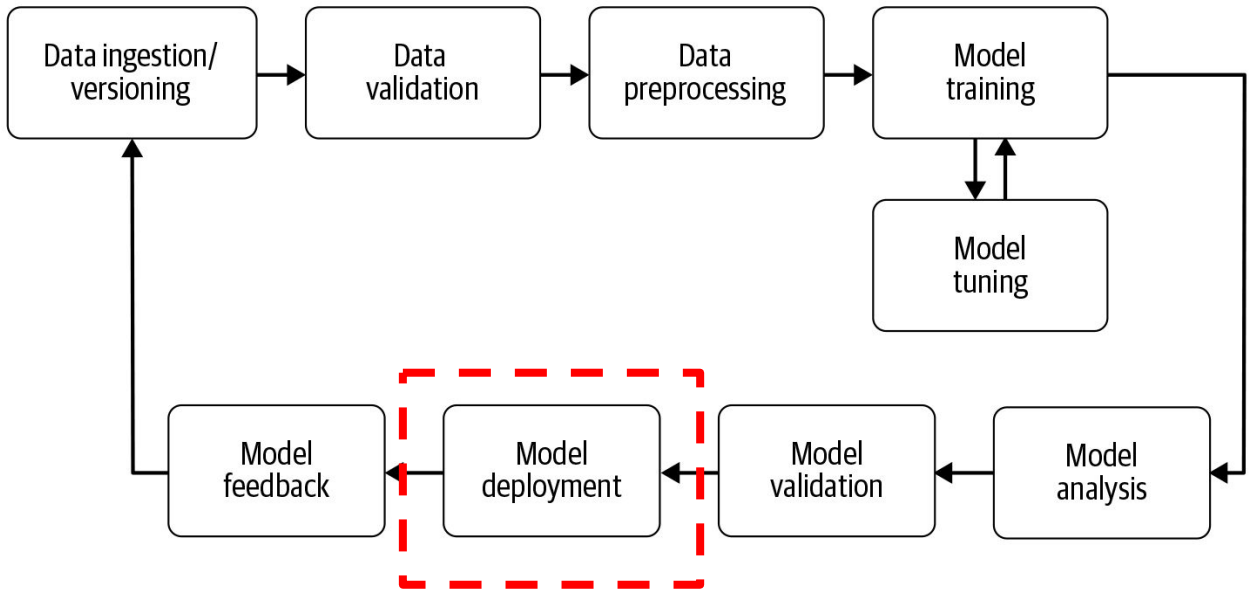
개인 정보 보호

04

모델 배포, 피드백 루프 반복, 개인 정보 보호

모델 배포

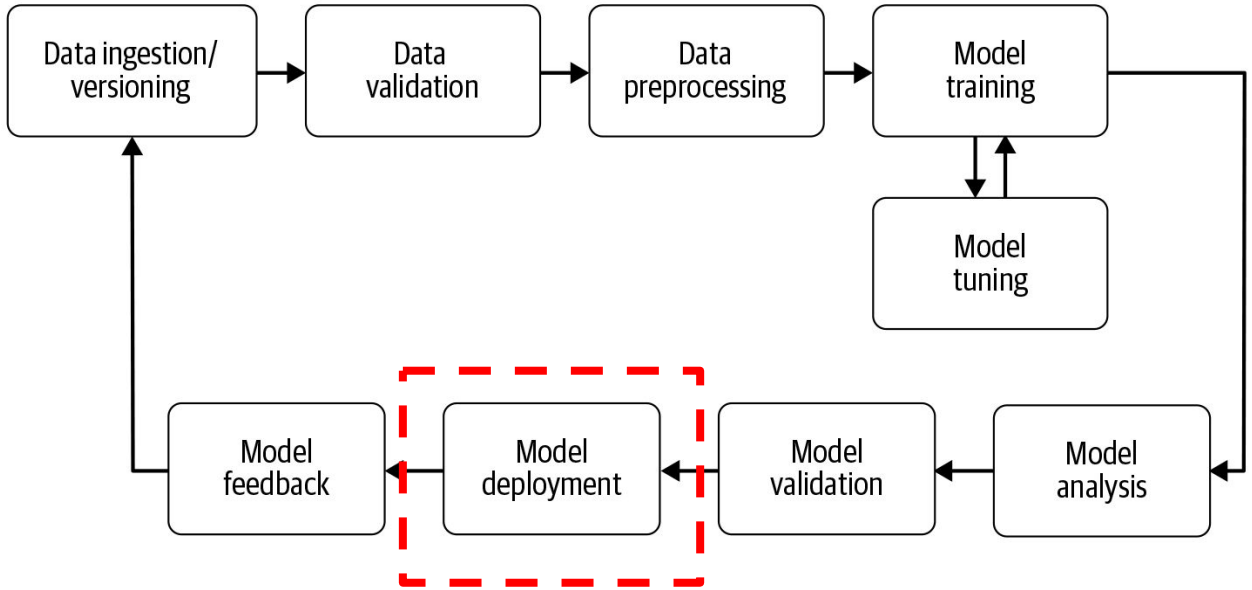
모델을 학습, 튜닝 및 분석한 뒤, 모델을 배포할 수 있습니다.
유감스럽게도 일회성 구현으로 구현된 모델이 너무 많다면,
모델 업데이트는 쉬운 프로세스가 아닙니다.



모델 배포, 피드백 루프 반복, 개인 정보 보호

모델 배포

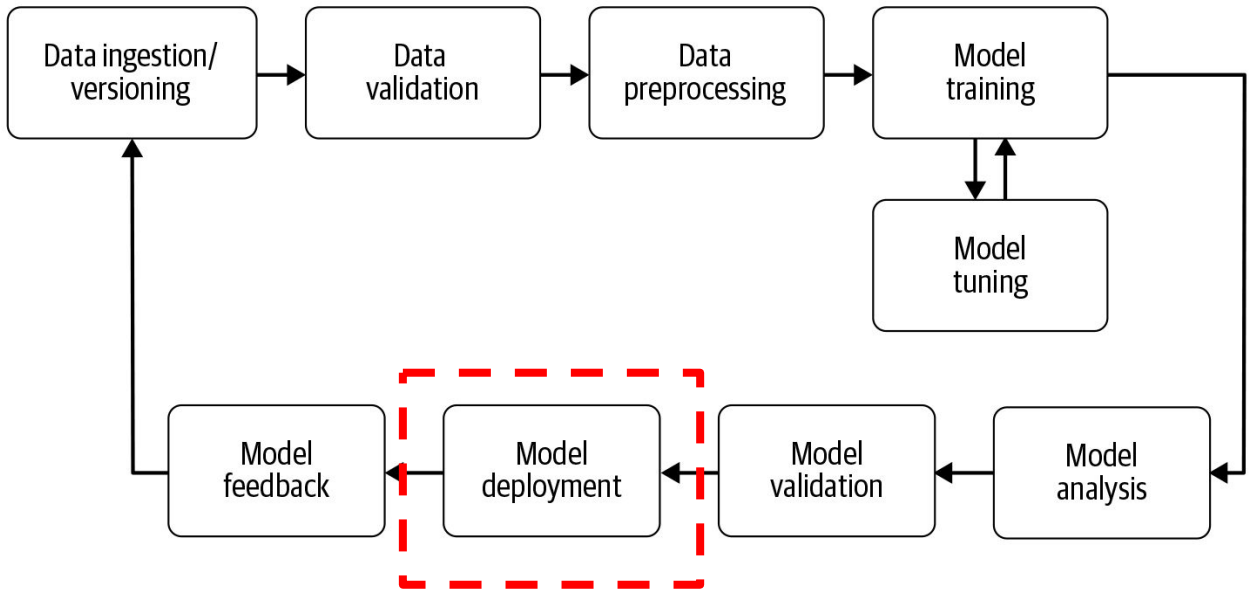
모던한 모델 서버를 사용하면 웹 서버 프로그램 코드를 작성하지 않고도 모델을 배포할 수 있습니다. REST(대표성 상태 전송) 또는 RPC(원격 프로시저 호출) 프로토콜과 같은 여러 API 인터페이스를 제공하여 동일한 모델의 여러 버전을 동시에 호스트할 수 있는 경우가 많습니다.



모델 배포, 피드백 루프 반복, 개인 정보 보호

모델 배포

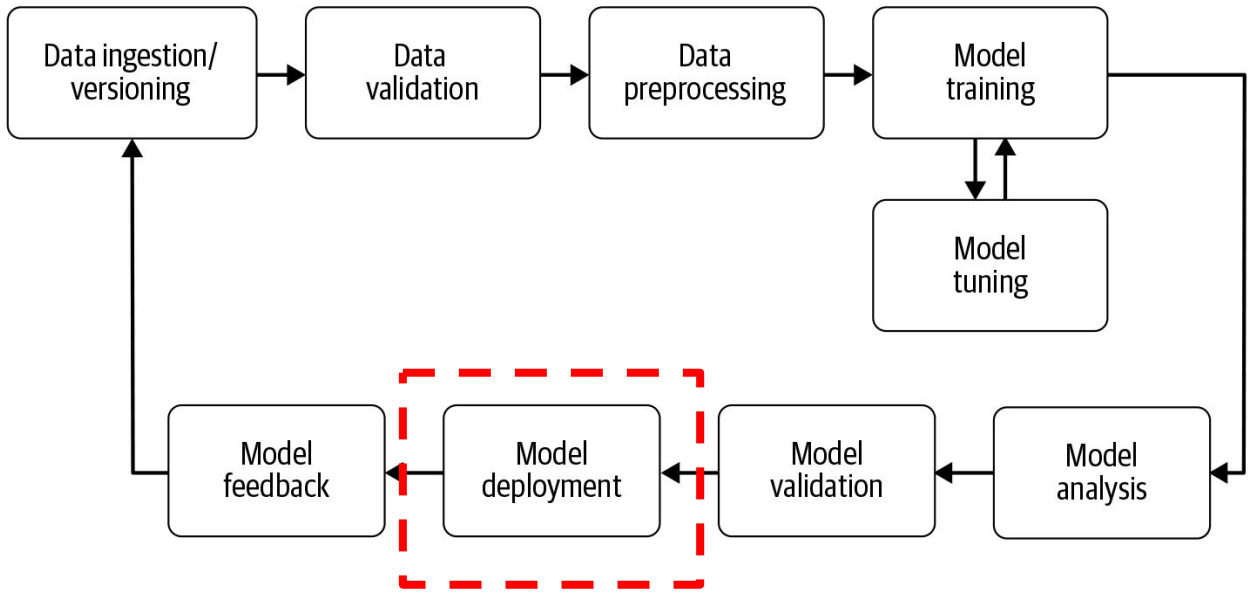
여러 버전을 동시에 호스팅한다면 모델 A/B 테스트를 실행하면서 모델 개선 사항에 대한 귀중한 피드백을 얻을 수 있습니다.



모델 배포, 피드백 루프 반복, 개인 정보 보호

모델 배포

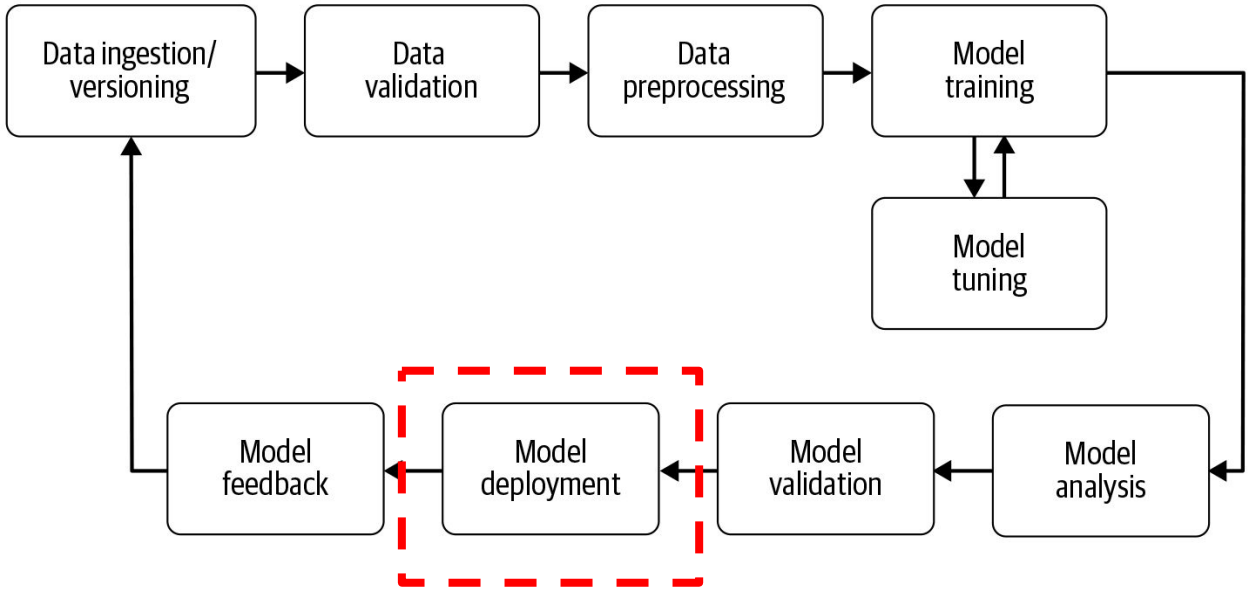
또한 모델 서버를 사용하면 애플리케이션을 다시 배포하지 않고도 모델 버전을 업데이트할 수 있으므로 애플리케이션 다운타임을 줄이고 애플리케이션 개발 팀과 머신러닝 팀 간의 커뮤니케이션을 줄일 수 있습니다.



모델 배포, 피드백 루프 반복, 개인 정보 보호

피드백 루프 반복

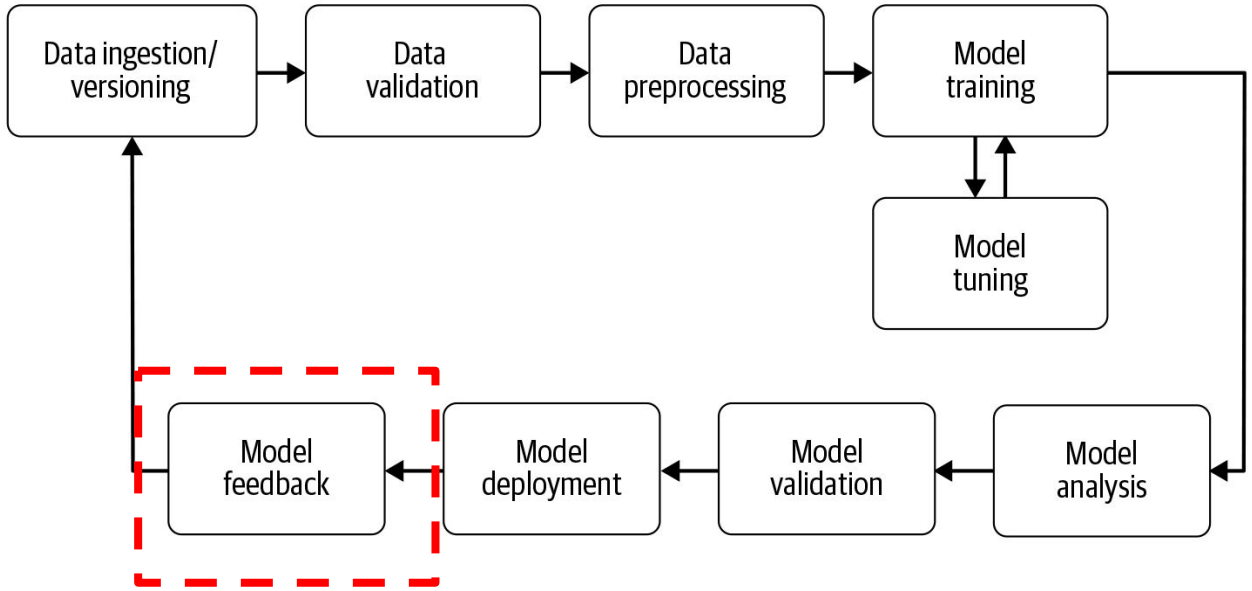
머신러닝 파이프라인의 마지막 단계인 피드백 루프 반복을 간과하기 쉽지만, 데이터 과학 프로젝트의 성공에는 매우 중요한 주제입니다.



모델 배포, 피드백 루프 반복, 개인 정보 보호

피드백 루프 반복

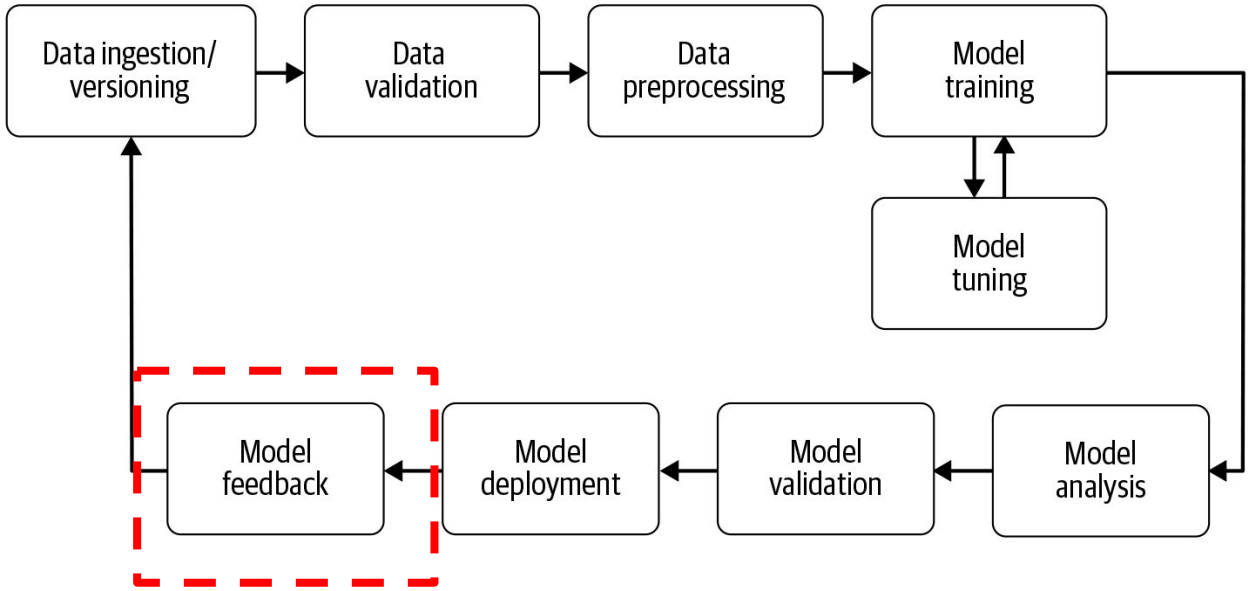
우리는 머신러닝 프로젝트에서 피드백 루프를 만들어야 합니다. 그래야 새로 배포된 모델의 효과와 성능을 측정할 수 있습니다. 이 단계에서는 모델의 성능에 대한 중요한 정보를 측정할 수 있습니다.



모델 배포, 피드백 루프 반복, 개인 정보 보호

피드백 루프 반복

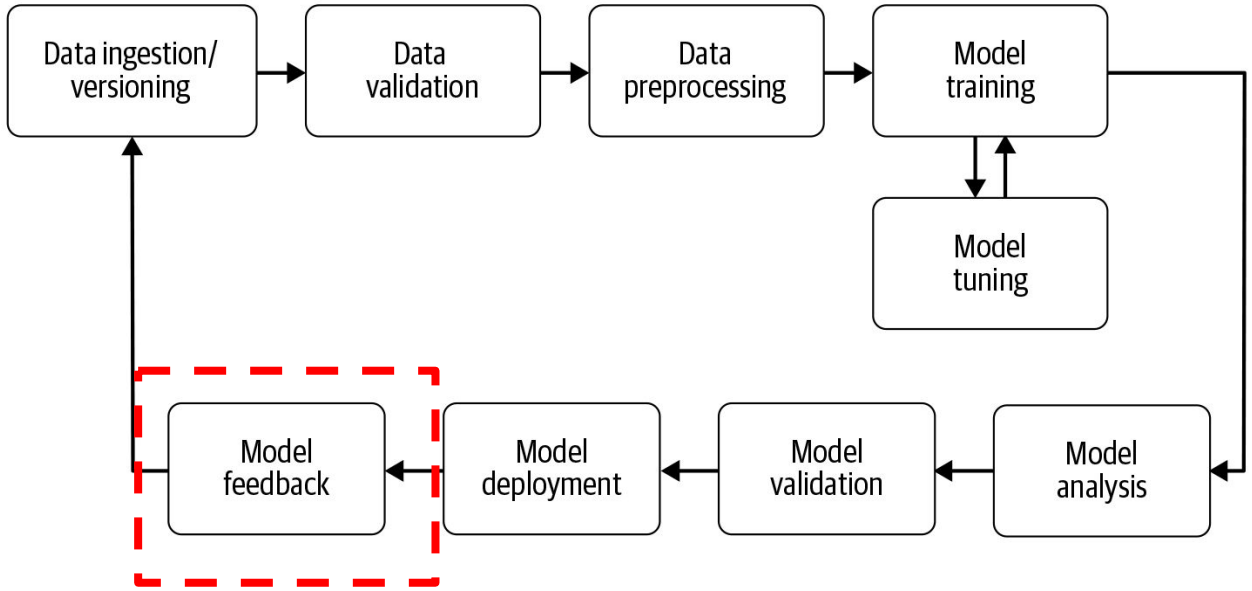
경우에 따라 데이터셋을 늘리고 모델을 업데이트하기 위해 새로운 학습 데이터를 수집할 수도 있습니다. 사람이 개입할 수도 있고, 자동화할 수도 있습니다. 13장에서 피드백 루프를 논의합니다.



모델 배포, 피드백 루프 반복, 개인 정보 보호

피드백 루프 반복

두 가지 수동 검토 단계(모델 분석 단계와 피드백 단계)를 제외하고 전체 파이프라인을 자동화할 수 있습니다. 데이터 과학자는 기존 모델을 업데이트하고 유지하는 것이 아니라 새로운 모델을 개발하는 데 집중할 수 있어야 합니다.



모델 배포, 피드백 루프 반복, 개인 정보 보호

개인 정보 보호

데이터 개인 정보 보호에 대한 고려사항은 표준 머신러닝 파이프라인 외부에 있습니다. 데이터 사용에 대한 소비자의 우려가 커지고, 개인 데이터의 사용을 제한하는 새로운 법률이 도입됨에 따라 향후 이러한 상황이 달라질 것으로 예상됩니다. 이로 인해 개인 정보 보호 기법이 머신러닝 파이프라인을 구축하기 위한 도구로 통합됩니다.

- 차등 개인 정보 보호Differential Privacy
- 연합 학습Federated Learning
- 암호화된 머신 러닝Encrypted Machine Learning

출처 : https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

짚어보기

머신러닝 파이프라인의 이해

01. 머신러닝 파이프라인 단계에 대해 이해한다.

머신러닝 파이프라인의 각 단계에 대해 이해한다.

02. 데이터 수집, 버전관리, 데이터 검증의 특징을 이해한다.

머신러닝 기반의 소프트웨어가 갖고 있는 특징을 이해한다.

03. 모델 학습, 모델 분석, 모델 버전 관리에 대해 이해한다.

데이터의 품질을 체크하는 데이터 검증에 대해 이해한다.

04. 모델 배포, 피드백 루프 반복, 개인정보 보호에 대해 이해한다.

모델 학습에 따른 특징을 이해한다.

머신러닝 파이프라인

머신러닝 파이프라인 단계

송호연

감사합니다.

THANKS FOR WATCHING