본 게시물은 W&B가 제공한 MLOps의 Model Development에 대한 내용을 기술한다. {: .notice}

1. Introduce

모델을 개발하여 실제 업무에 적용하기 위해선 아래 두가지 사항을 고려해야 한다.

- 모델이 실제 운영 환경에서 올바르게 동작할 수 있는 가능성 검증
- 정확도, 오차율 등의 비지니스 효과를 측정

이러한 작업은 보통 단순한 baseline(default model)을 구축하고 반복적인 모델링(튜닝, feature set 설정 등)을 수행하고, 그 결과를 평가를 통해 검증한다. 이번 포스팅에서는 이처럼 Model Developement를 위해 수행하는 과정과 이를 효과적으로 수행할 수 있는 방법에 대해 서술하도록 할 것이다.

2. Modeling & Evaluation

2.1 Understanding the business context

Machin Learning 프로젝트의 첫 번째 단계는 비즈니스 맥락을 이해하는 것이다. 이는 비즈니스 목표를 설정하기 위한 모델 성능 평가 지표(=metric)을 정하는 것과 매우 연관이 있으며, 다양한 실험에서 좋은 모델을 선정하기 위함이다. 비즈니스를 이해하고 metric을 선정하였다면 고차원의 모델링을 수행하기 전, 간단한 baseline을 구축하고 전문가와 함께 검토하는 것이 유용하다.

2.2 Feedback from experts

전문가의 피드백은 실험을 효과적으로 할 수 있는 첫 단추이다. 본인의 도메인이 아닌 새로운 프로젝트에 참여하는 데이터 분석가는 현업과의 인터뷰를 통해 목표를 분명히 하며 인사이트를 얻을 수 있다. 혹은 반대로 도메인 지식은 있으나 데이터 분석 역량이 부족하다면 전문 분석가에게 metric 및 baseline을 검토 요청하여 효과적으로 분석을 설계할 수 있다. 따라서 각 이해 관계자들의 협업을 가능하게 하는 분석 툴을 수립하는 것도 model development에서 중요한 역할을 한다고 볼 수 있다.

2.3 Tracking model

앞선 과정들이 준비되었다면 모델을 점진적으로 고차원으로 개발하고 평가하는 과정을 수행해야한다. 해당 과정을 수행할 시 다양한 모델을 구축할 것이며 각 모델의 파라미터들을 조정하며 많은 실험 결과를 산출하게 된다. 이때 만약 실험의 결과를 기록하거나 추적할 수 없다면 최적의 결과를 찾기까지 많은 시간이 소모될 것이다. 반면 실험 결과를 추적할 수 있다면, 주요하게 작용한 feature, parameter 등을 고려해 최적의 결과로 수렴하는 과정이 수월해지며, 협업시에도 빠른 의사결정을 할 수 있을 것이다. 따라서 다양한 가설을 통해 구축한 모델을 실험하고 점진적으로 모델을 개선시키기 위해서는 tracking이 필요하다.

3. Development Tools

효과적인 모델 개발 및 평가를 위한 필수 조건 중 하나는 엔지니어의 개발 환경을 구축하는 것이다. 아래의 기능들은 반복 가능하고 신뢰성 있는 개발 환경을 구축하는데 필요한 것들이다.

3.1 Dependency Management

대부분의 Machine Learning 프로젝트에서 모델 및 파이프라인 구축은 python으로 개발된다. 따라서 python의 다양한 라이브러리들을 관리할 수 있는 의존성 관리 도구인 Poetry나 Conda가 필요하며, Docker, virtualenv 또는 pipenv는 이러한 의존성 관리를 일정한 환경에서 수행할 수 있도록 보장한다. 따라서 협업을 위해서는 몇 가지 표준화된 환경과 의존성 하에서 모델 구축 및 실험을 하는 것이 유리하다.

3.2 Sandbox Environments

많은 기업들이 컴퓨팅 요구사항에 맞게 cloud 혹은 on-prem의 분석 환경을 제공한다. 그러나 의존성과 환경이 서로 호환되지 않는 경우가 종종 있기에 개발자들은 본인이 원하는 환경(=sandbox environment)을 구축하는 능력이 필요하다. 개발 환경과 제공되는 환경 간의 불일치를 최소화하는 것이 신속한 모델 개발을 가능하게 하는 데 중요하기 때문이다.

4. Decision Optimization

Machine Learning은 운영 의사 결정을 최적화 하는데 사용되곤 한다. 하지만 모델은 비즈니스의 의사 결정이 아니라 단순 예측치만을 제공할 뿐이다. 따라서 모델을 사용하여 의사 결정을 내리는 정책을 세워야 한다.

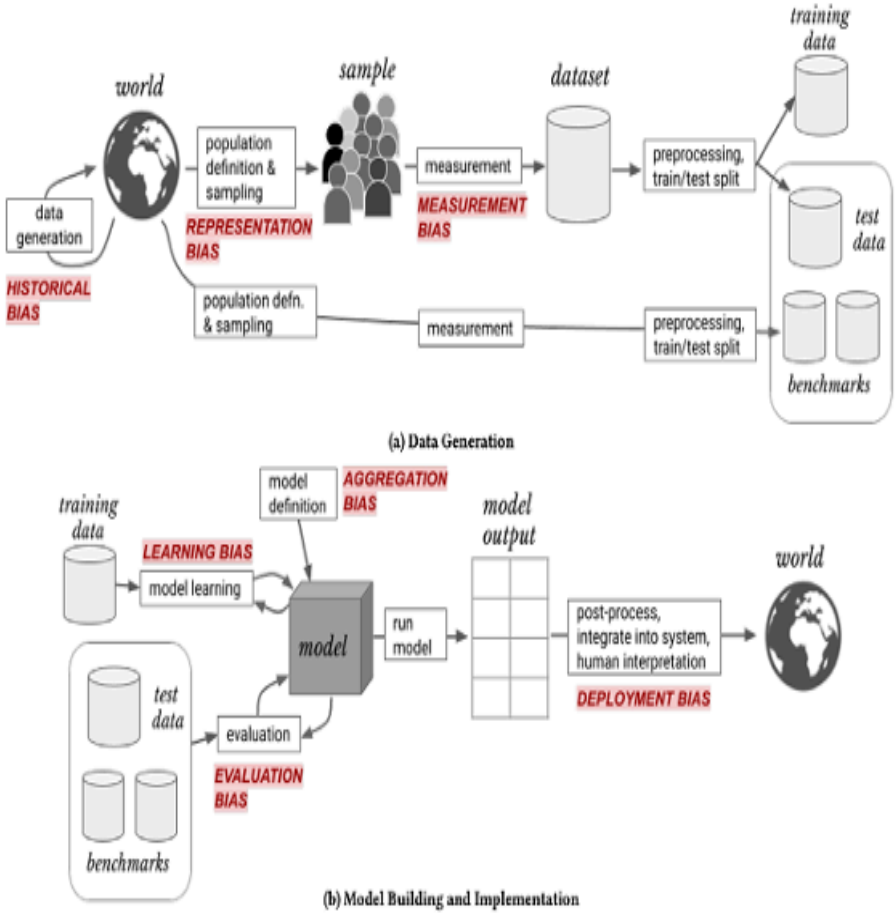
📍 **Example**

> 의료 분야에서 양성(=질병이 있는 경우)과 음성(=질병이 없는 경우)을 판별하는 분류 모델에서 임계값을 낮게 설정한다면 많은 케이스가 양성(=True)으로 분류된다. 이로 인해 False Positive(=거짓 양성)의 수가 증가하고, 이는 의료업에서 문제가 될 수 있다. 예를 들어 거짓 양성으로 치료를 시작하는 경우 비용이 많이 들 수 있으며, 환자에게 불필요한 불안감을 줄 수 있기 때문이다.
>
> 반면 임계값을 높게 설정한다면 거짓 양성의 수는 감소하지만 False Negative(=거짓 음성)의 수가 증가한다. 이는 거짓 음성이 실제로 질병을 갖고 있는 놓칠 수 있는 위험을 야기한다.

위의 예시로 알 수 있듯, 모델 예측을 올바르게 사용하기 위한 적절한 의사결정 정책이 없다면 비즈니스 효과를 오히려 부정적으로 이끌 수 있다. "Tuning The Model For Business Value" 이 블로그의 글은 비즈니스 가치를 극대화 하는 방법을 보여주는 실제적인 예시를 보여주며 의사 결정 최적화를 쉽게 이룰 수 있는 Tracking Tool을 소개한다.

5. Governance

Machine Learning에서 발생할 수 있는 bias들의 원인은 매우 다양하다. 모델을 개선하기 위해선 bias의 각 유형이 무엇인지, 근본 원인이 무엇인지, 그리고 이를 완화하기 위한 프로세스를 어떻게 구현할 것인지 이해하는 것이 중요하다. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle" 이 논문에서는 이를 7가지 일반적인 bias의 유형 구분하였으며 아래는 그에 대한 도표이다.



- **Historical bias**: 데이터가 생성되는 원본 과정에서 발생하는 것으로, 데이터 자체가 잘못 생성되는 경우를 의미한다.
- **Representation bias**: 데이터 샘플링이 정의되는 방식에 따라 발생하는 것으로, 어떤 데이터를 타겟으로 수집할지에 따라 누락되는 정보가 생기는 경우를 의미한다.
- **Measurement bias**: 데이터 샘플링 전략이 실행되는 방식에 대한 것으로, 데이터를 수집하고 처리하는 과정에서 품질이 저하되는 경우 발생할 수 있다.
- **Aggregation bias**: 모델이 정의되는 방식에 따른 것으로 데이터를 요약하거나 집계하는 방법에 따라 모델이 영향을 받을 수 있다.
- **Lerning bias**: 모델이 훈련되는 방식에 따른 것으로 훈련 데이터에 따라 모델이 치우친 결과른 내는 것이다.
- **Evaluation bias**: 모델이 검증되는 방식에 대한 것으로 평가시 지표나 데이터의 구성에 따라 성능이 달라질 수 있다.
- **Deployment bias**: ML system의 최종 구현 및 사용 방식에 대한 것으로 실제 환경에서 특정 그룹이 불공평하게 영향을 받을 수 있다.