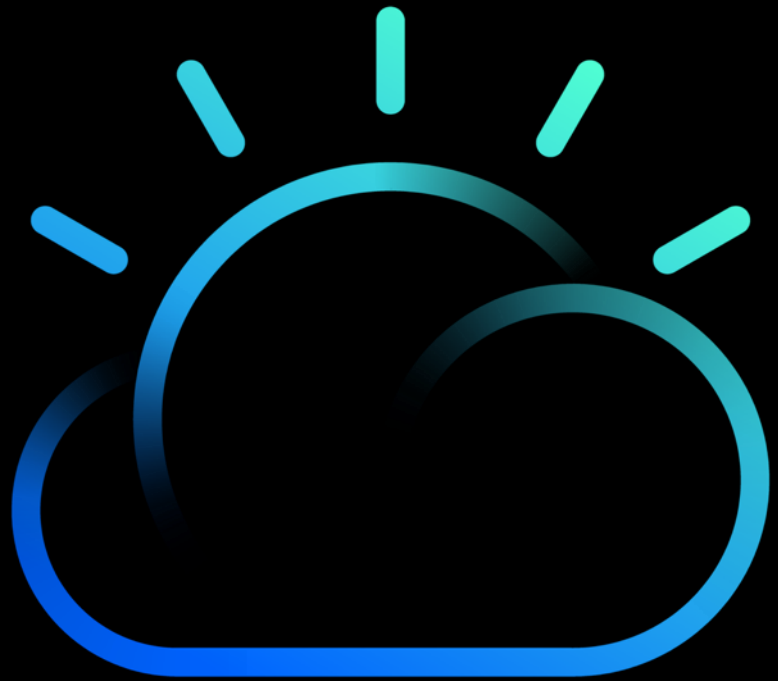
IBM Watson을 활용한

신뢰가능한 AI 모니터링 및 효율적 개선 방안 제시



2021. 4. 23.

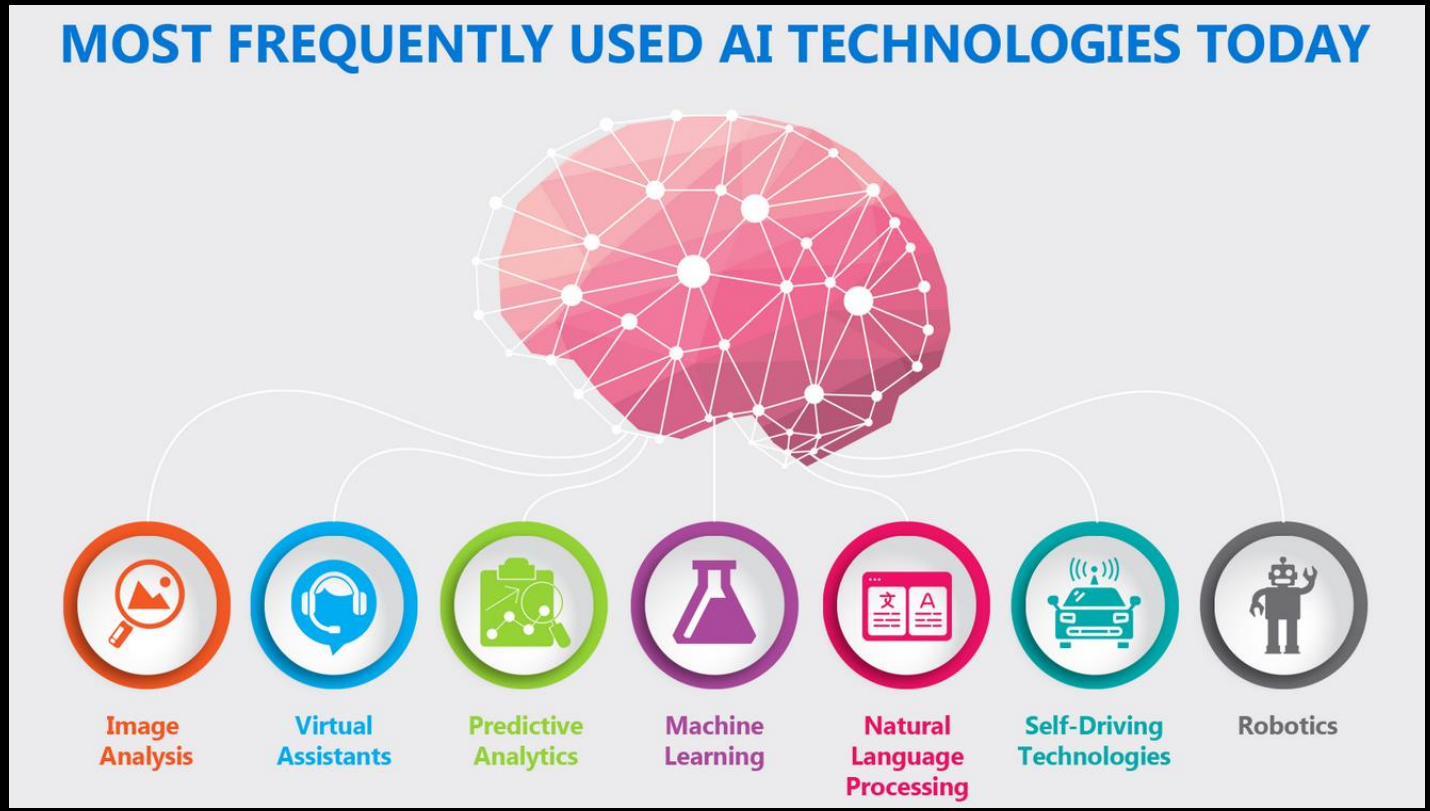
김지관 부장
Data&AI Technical Specialist
IBM Technology Sales, Korea

IBM Cloud

Contents

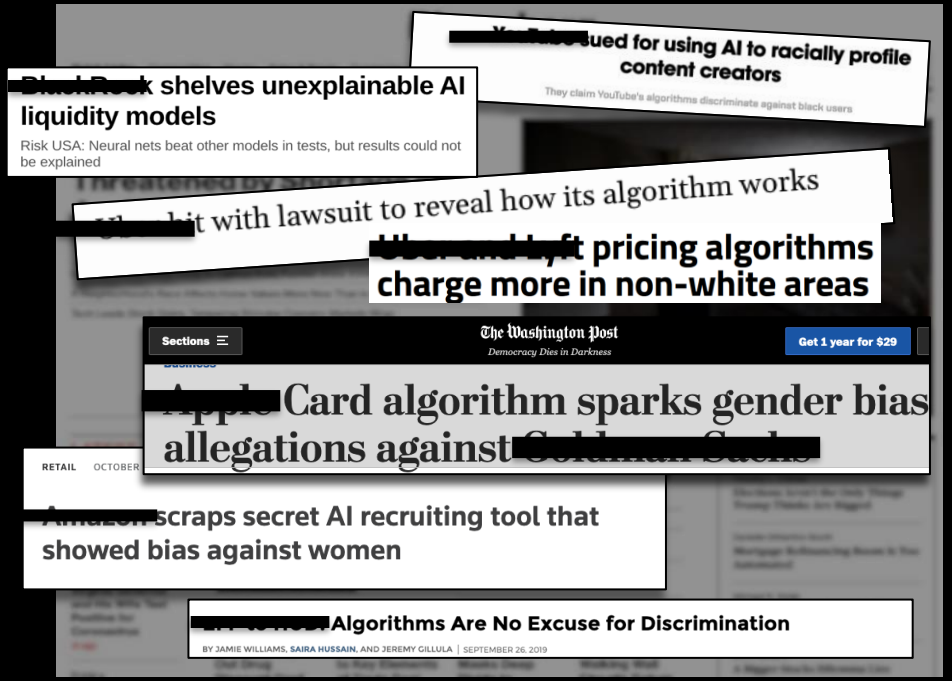
1. AI 신뢰에 대한 문제 인식, 어디까지 왔나?
2. 신뢰 가능한 AI를 위한 핵심 키워드는?
3. 신뢰 가능한 AI를 위한 IBM의 방안 – Watson OpenScale
4. 개방적이면서도 통제 가능한, 그리고 자동화된 AI Lifecycle을 지원하는 IBM Data AI 플랫폼

AI 기술은 이미 산업 분야와 상관 없이 우리 생활 속에 깊게 자리잡고 있습니다.



Source : https://medium.com/@Brian.johnson_62680/artificial-intelligence-ai-top-use-cases-and-technologies-used-today-3c22e1a63e78

AI 기술이 우리 삶에 빠르게 스며든 속도 만큼 많은 문제점들이 빠르게 발생하고 있습니다.
이는 특정 개인, 회사의 문제가 아닌, AI 기술을 활용하는 우리 모두의 문제일 수 있습니다.



1. *알고리즘의 투명성 문제*

   - *알고리즘이 도출한 결과에 대하여 설명/해석이 가능한가?*

2. *알고리즘의 편향(bias) 문제*

   - *알고리즘이 도출한 결과가 특정 그룹/계층을 차별하지 않는가?*

국제사회에서는 이미 AI의 신뢰성을 담보하기 위한 원칙을 세워나가고 있습니다.

OECD BETTER POLICIES FOR BETTER LIVES 60 YEARS

Artificial intelligence > OECD Principles on AI

What are the OECD Principles on AI?

The OECD Principles on Artificial Intelligence promote artificial intelligence (AI) that is innovative and trustworthy and that respects human rights and democratic values. They were adopted in May 2019 by OECD member countries when they approved the **OECD Council Recommendation on Artificial Intelligence**. The OECD AI Principles are the first such principles signed up to by governments. Beyond OECD members, other countries including Argentina, Brazil, Costa Rica, Malta, Peru, Romania and Ukraine have already adhered to the AI Principles, with further adherents welcomed.

The OECD AI Principles set standards for AI that are practical and flexible enough to stand the test of time in a rapidly evolving field. They complement existing OECD standards in areas such as privacy, digital security risk management and responsible business conduct.

In June 2019, the **G20 adopted human-centred AI Principles** that draw from the OECD AI Principles.

The OECD AI Principles

The Recommendation identifies five complementary values-based principles for the responsible stewardship of trustworthy AI:

- AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.
- AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.
- There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.
- AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.
- Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

신뢰가능한 AI 구현을 위한 원칙

- ✓ 포용성장, 지속가능 발전, 복지 증진
- ✓ 인간 중심 가치 지향, 공정성
- ✓ 투명성 및 설명 가능성 확보
- ✓ 견고성, 보안, 안전성 확보
- ✓ 상기 원칙에 따른 책임 완수

Source : https://www.oecd.org/going-digital/ai/principles/

Contents

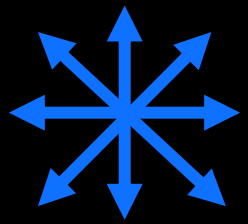
신뢰 가능한 AI를 위한 핵심 키워드와 핵심 지표

“Fair”

1. Fairness (공정성)

AI 모델의 Bias(편향)은 상시 모니터링 및 개선 되어야 함

“Robust”



2. Quality (품질)

AI 모델은 전체 라이프사이클 측면에서 그 품질이 담보되어야 함

3. Drift (추세)

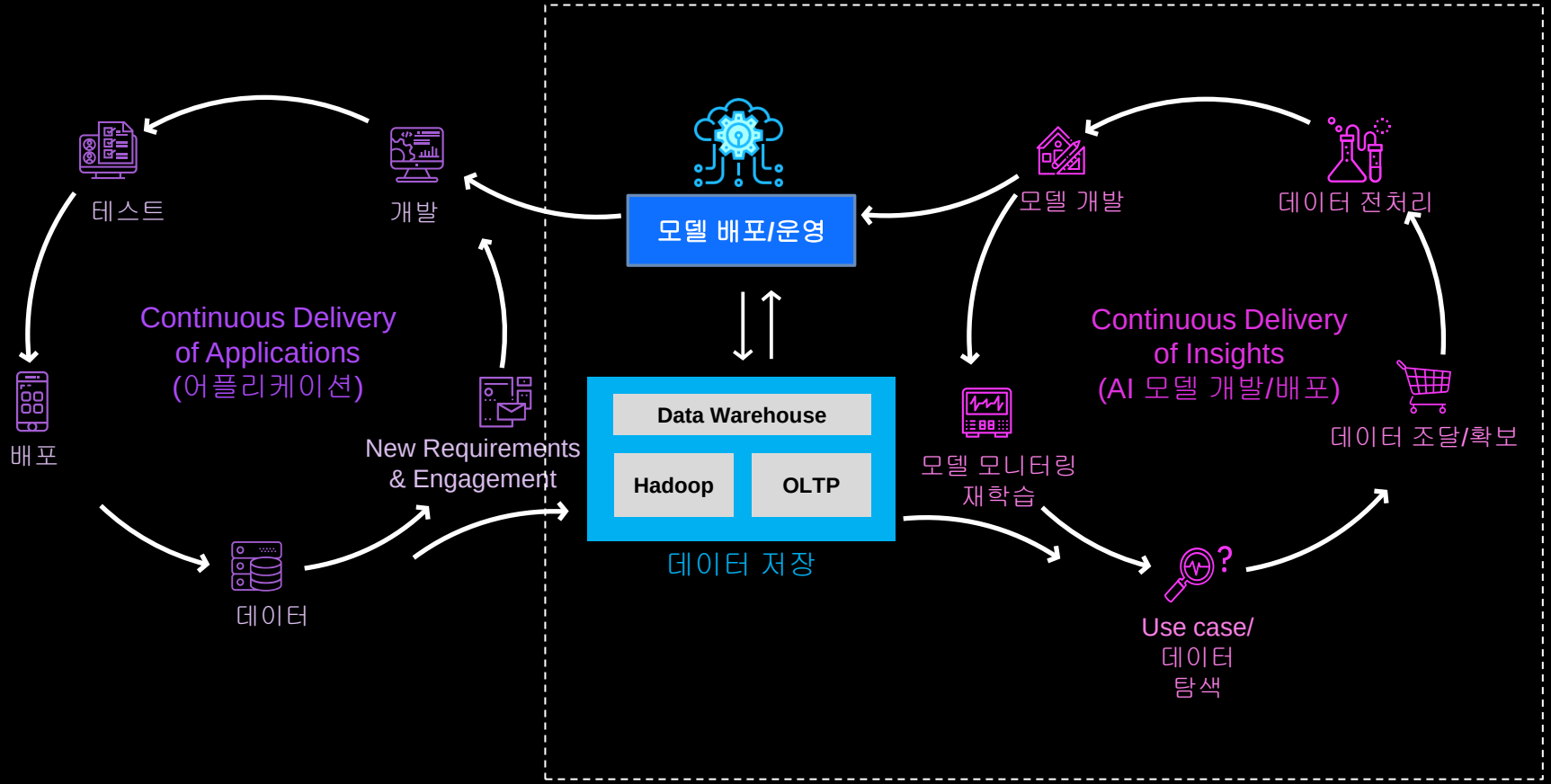
AI 모델의 품질 문제는 최대한 빠르게 인지되고 개선되어야 함

“Transparent”

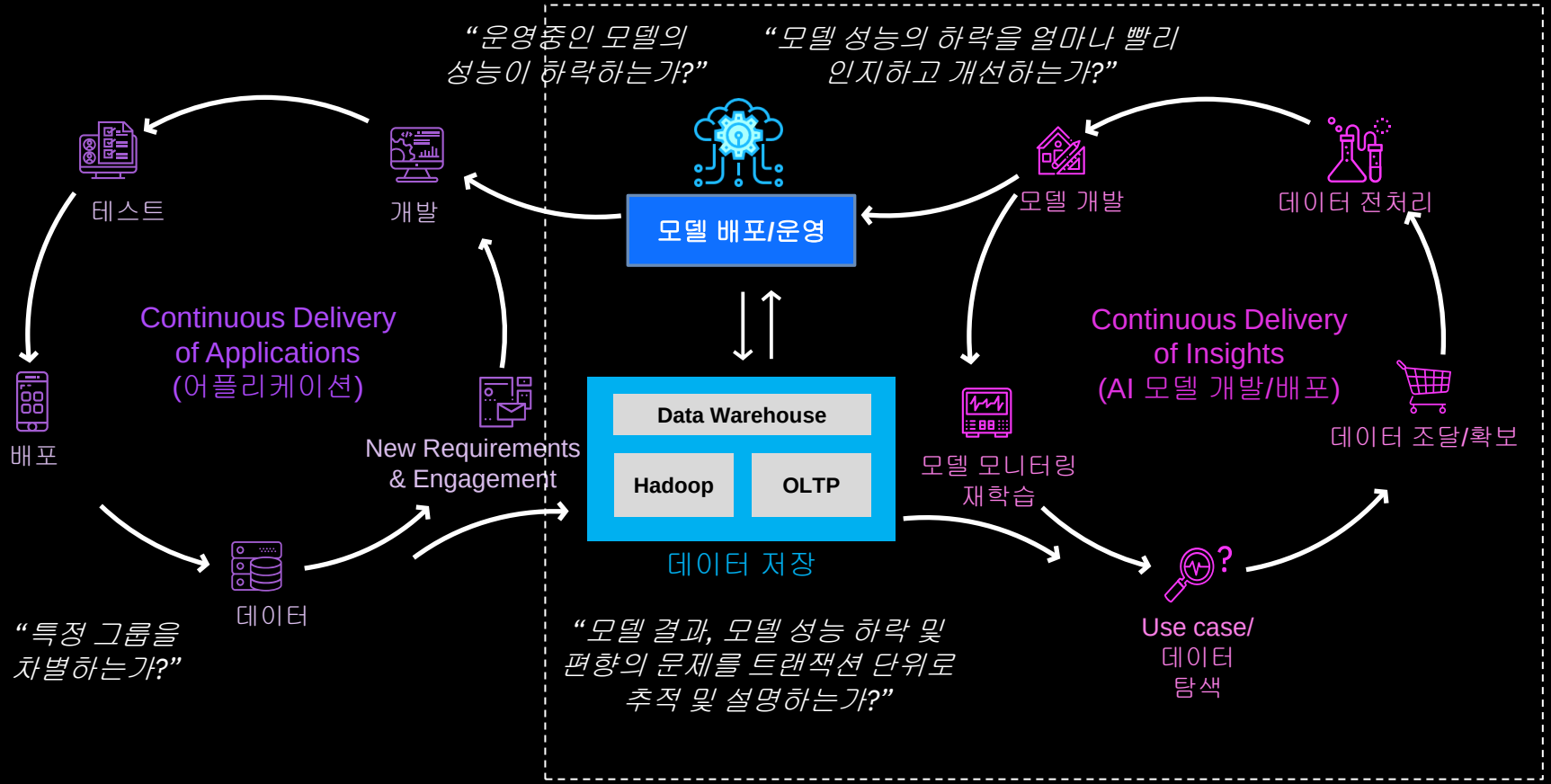
4. Explainability (설명)

AI 모델은 개별 트랜젝션 단위로 추적되고 설명되어야 함

AI의 라이프사이클과 신뢰가능한 AI를 위한 고려사항



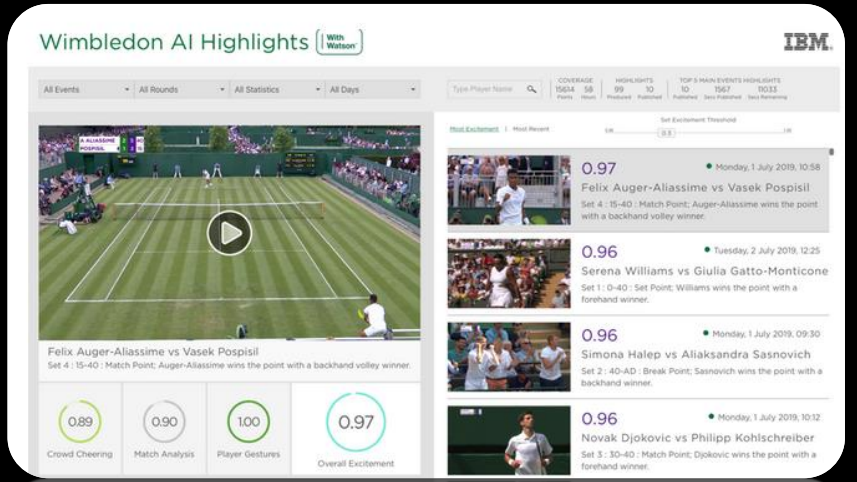
AI의 라이프사이클과 신뢰가능한 AI를 위한 고려사항



Contents

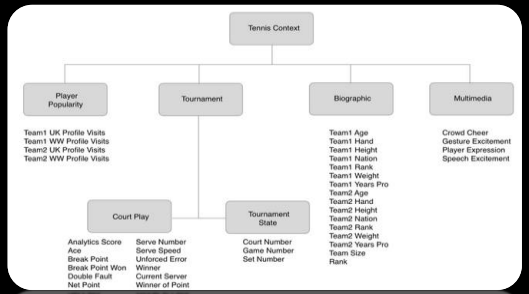
사례 - 윔블던

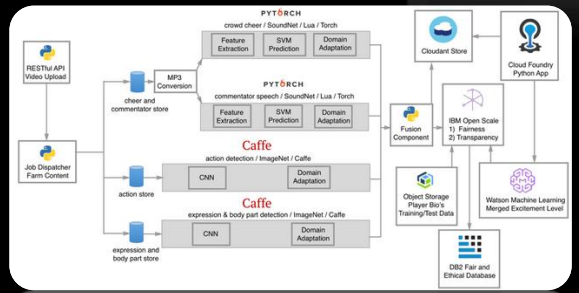
AI가 공정하고 빠르게 하이라이트 장면을 자동 선정

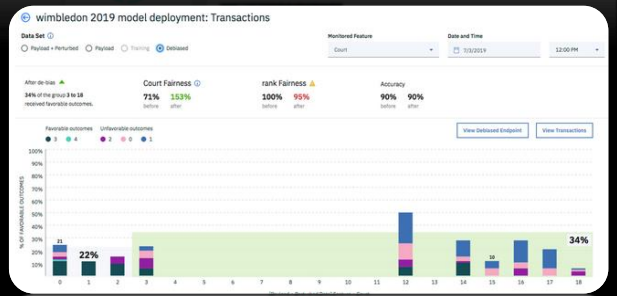


Business challenge & Solution

- 윔블던은 675개 이상의 경기와 147,000개 이상의 테니스 포인트를 가지고 있음
- 편집자 또는 팬들이 토너먼트에서 가장 좋은 장면을 보기가 어려움
- 윔블던은 IBM의 디지털 및 AI 기능을 사용하여 하이라이트에 신속하게 액세스하여 팬들에게 최고의 콘텐츠를 2분 이내 제공







Source : https://www.ibm.com/blogs/journey-to-ai/2019/07/how-ai-picks-the-highlights-from-wimbledon-fairly-and-fast-2/

사례 - Nature Conservancy(비영리 글로벌 환경 단체)

AI가 수질 환경 문제를 모니터링하고 문제 원인 분석

Business challenge & Solution

- Nature Conservancy는 미국 버지니아 주 알링턴에 본사를 둔 글로벌 환경 단체(미국 전역 및 전세계 79개국 활동)
- 타나 강은 1,000KM의 케냐에서 가장 긴 강으로 이 지역 수백만 주민 식수와 농경지 관개에 매우 중요
- 수질 모니터링 및 문제 조기 식별을 위한 AI 적용
  - 수질, 수위, 온도 등 센서에 의한 데이터 자동 수집
  - 실시간 시각화 분석, 모니터링 및 AI 해석에 의한 문제 원인 파악 및 대응책 마련

Tackle Climate Change | Protect Land & Water | Provide Food & Water Sustainably | Build Healthy Cities

Source : https://www.ibm.com/blogs/journey-to-ai/2019/07/how-ai-picks-the-highlights-from-wimbledon-fairly-and-fast-2/

사례 - KPMG

Watson OpenScale을 통한 책임 있는 AI 관리

Business challenge:

KPMG는 AI 도입을 저해하는 신뢰 원칙들을 식별

- Integrity(무결성): 라이프사이클 전체에 걸쳐 데이터 품질을 어떻게 보장합니까?
- Fairness(공정성) : 그룹, 개인 또는 데이터 속성에 대한 편견을 어떻게 극복할 수 있을까요??
- Explainability(설명력) : AI가 어떻게 결론에 도달했는지에 대한 결정을 비즈니스 관점에서 어떻게 설명할 수 있습니까??

Solution :

KPMG는 IBM Watson OpenScale을 도입함으로써,

- 인지되지 못하는 AI의 편향에 대하여 보다 잘 이해
- 규제 변화에 보다 능동적으로 대응
- AI 도입함에 있어 필요한 신뢰 확보

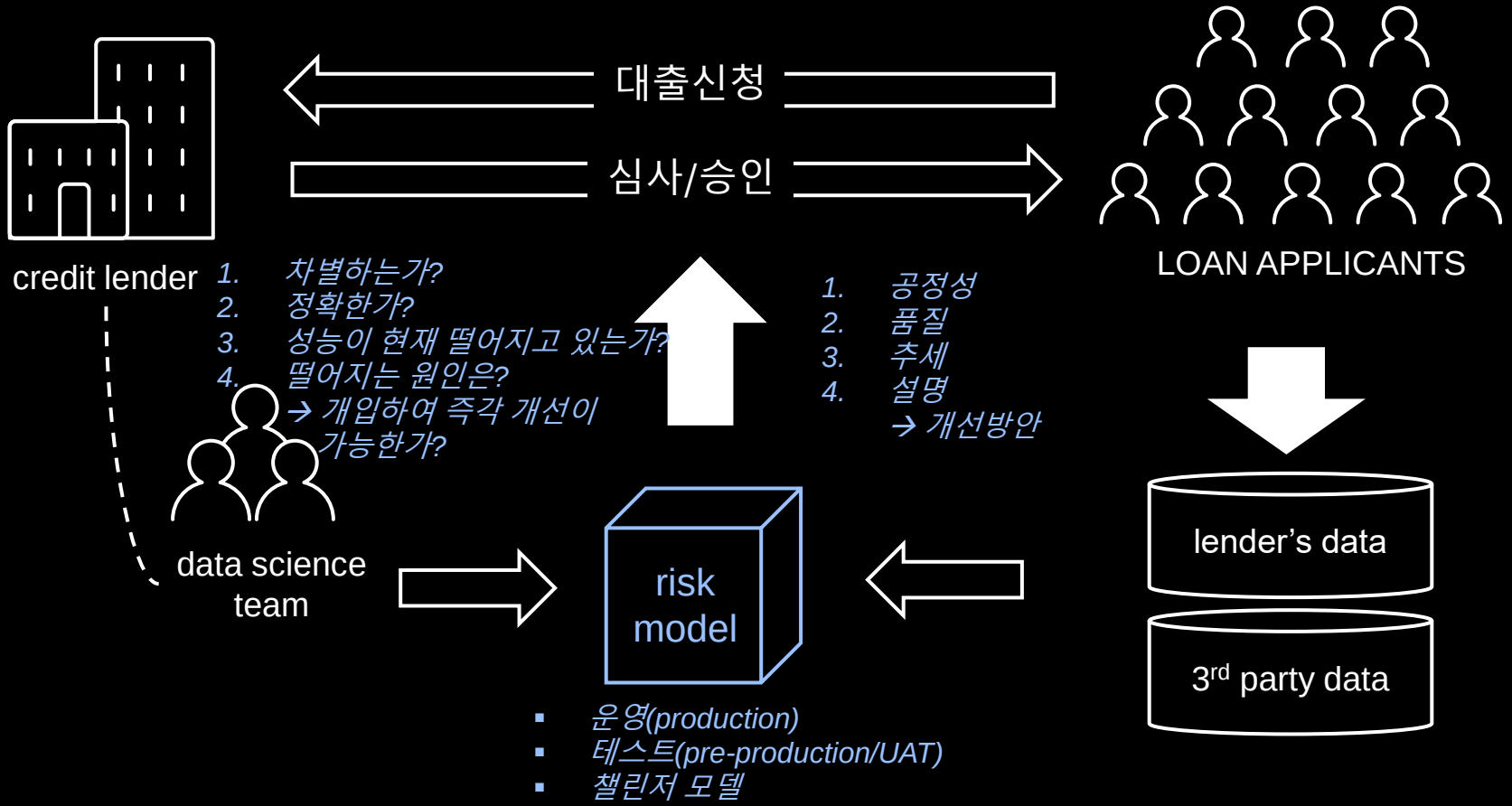
"Watson OpenScale은
AI 모델이 어떤 데이터 속성을 사용했는지, 어떻게,
왜 결정하게 됐는지 등을 명확히 밝힘으로써,
고객에게 비즈니스 용어의 투명성을 부여하는
시장에서 상용화된 독보적 기술"

— Kelly Combs, Director,
Emerging Technology Risk at KPMG

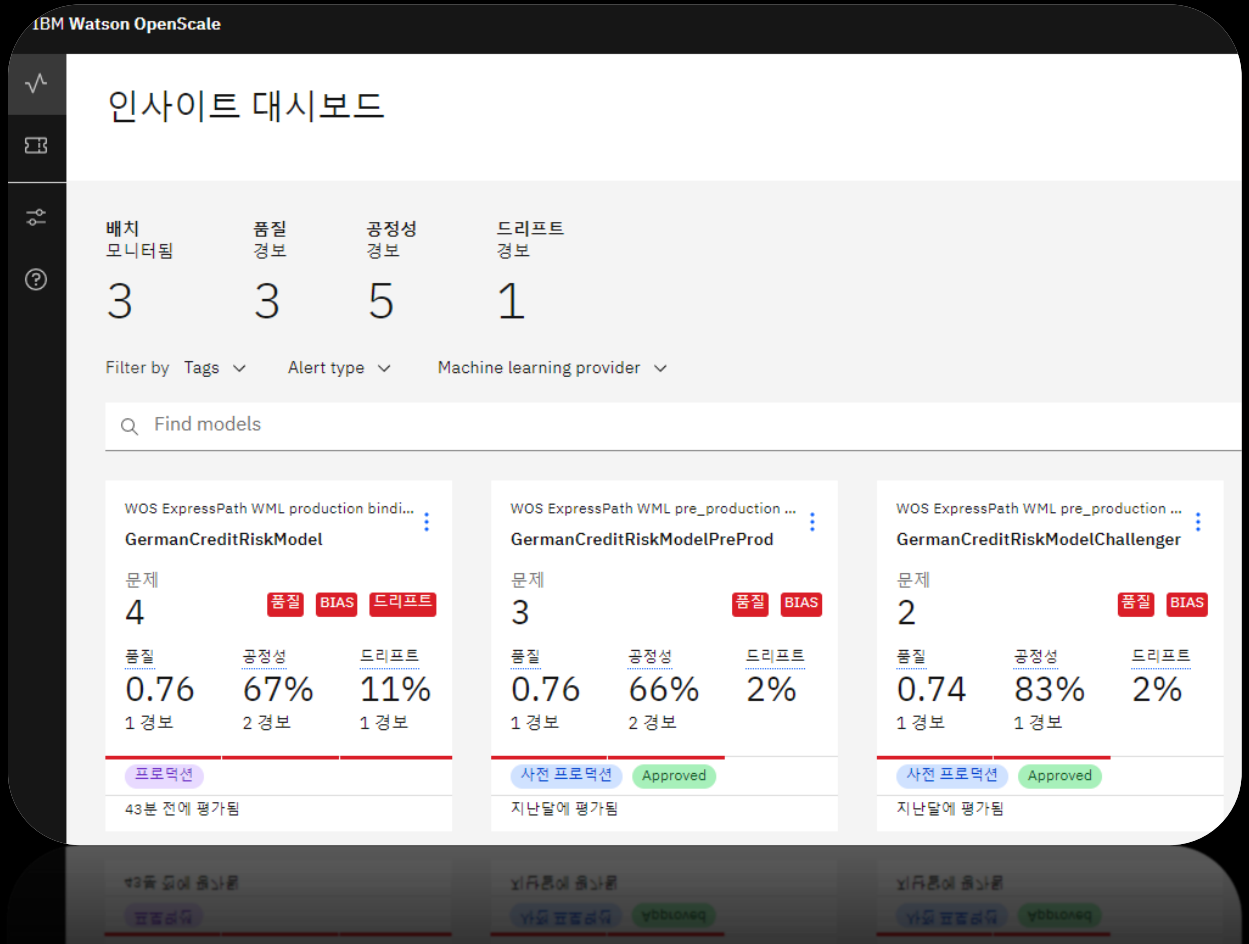
Source : https://www.youtube.com/watch?app=desktop&v=aEmzG02Mk0U&feature=youtu.be/

IBM Watson OpenScale을 활용한 가상 시나리오 : Credit Risk Model

공정하고 정확하고 투명한 리스크 판정 모델을 통해 기업 위험 손실을 최소화하기 위함

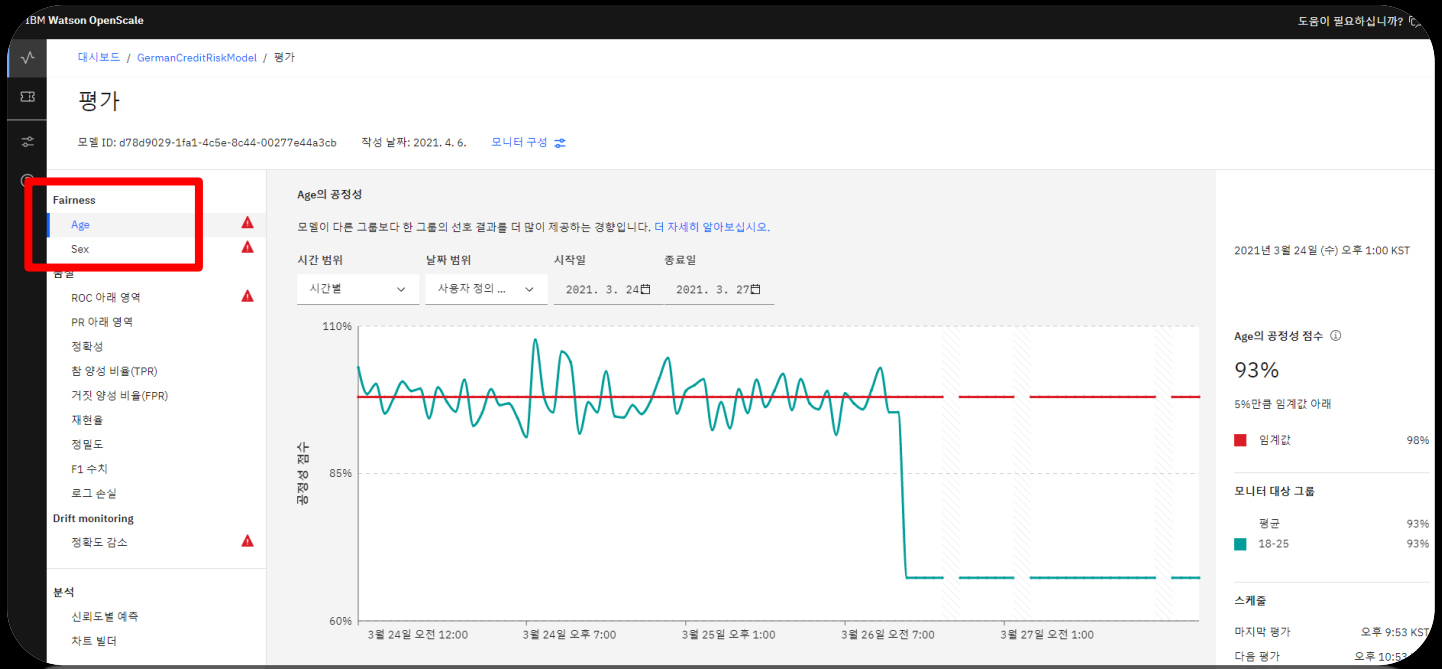


Watson OpenScale 시나리오 > 인사이트 대시보드(전체 모델 현황 관리)



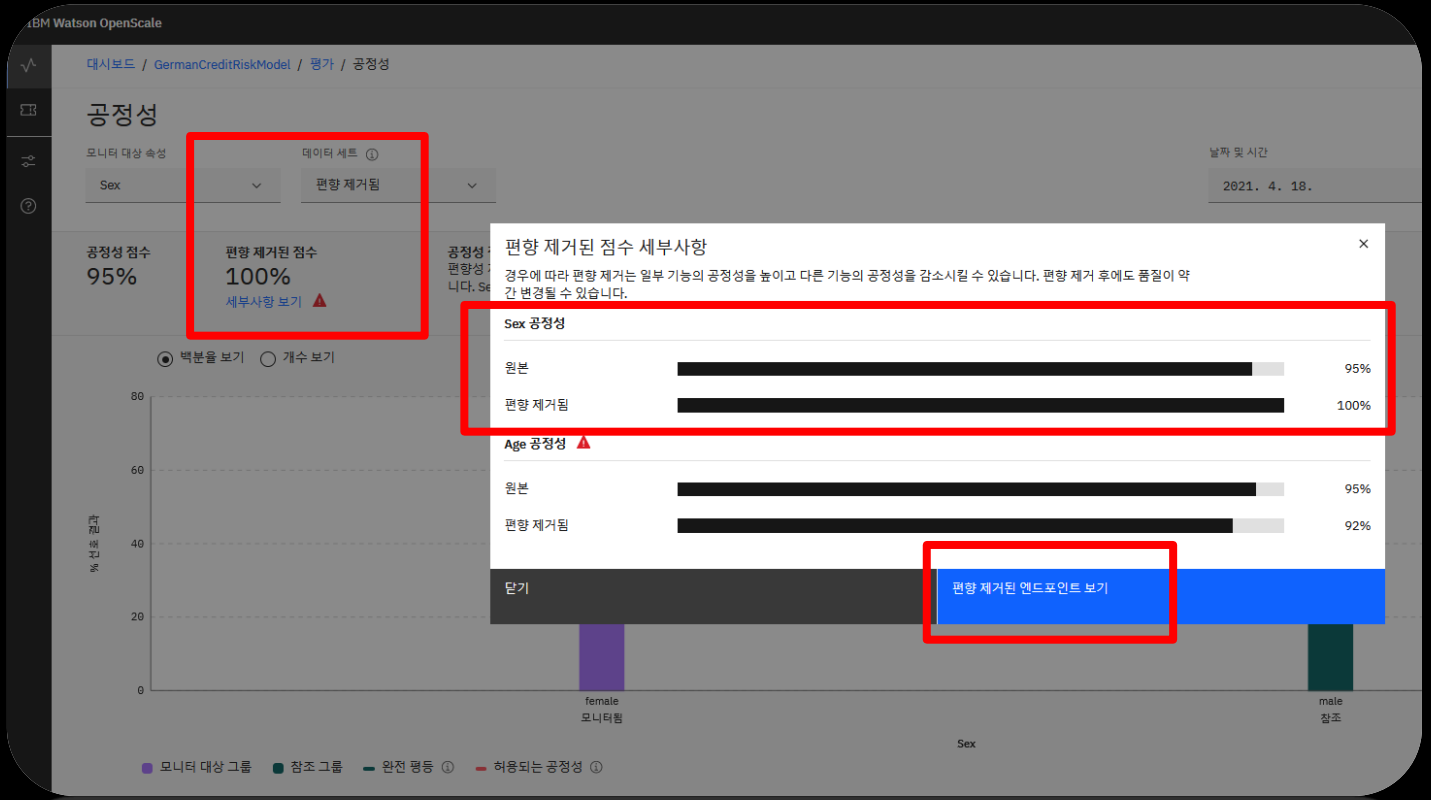
- ✓ 기업 내 관리가 필요한 모든 머신러닝 모델에 대한 현황 파악 및 관리
- ✓ 실 운영중 모델 뿐만 아니라 테스트 모델, 챌린저 모델 등도 동시 관리
- ✓ 본 시나리오에서는 리스크 판정 모델 1개 use case에 대하여 아래 3개 모델을 모니터링
  . 운영(production)
  . 테스트(pre-production/UAT)
  . 챌린저 모델

Watson OpenScale 시나리오 > 1. Fairness(공정성)



- ✓ 실 운영중인 모델에 대하여 bias(편향) 여부를 모니터링 해야하는 변수를 지정 (예 : 성별, 연령대)
- ✓ 시간 단위 별 bias를 모니터링하며, 특정 임계치 이하 시 경보

Watson OpenScale 시나리오 > 1. Fairness(공정성)



- ✓ 지정된 변수별로 참조 그룹 대비 모니터링 대상 그룹의 bias를 탐지
- ✓ 편향이 제거된 데이터 셋을 자동 생성하고 모델에 적용하여 즉각 개선 및 적용 가능 (업무 담당자/분석가 선택 사항임)
- ✓ 모델 공정성을 지속적으로 담보

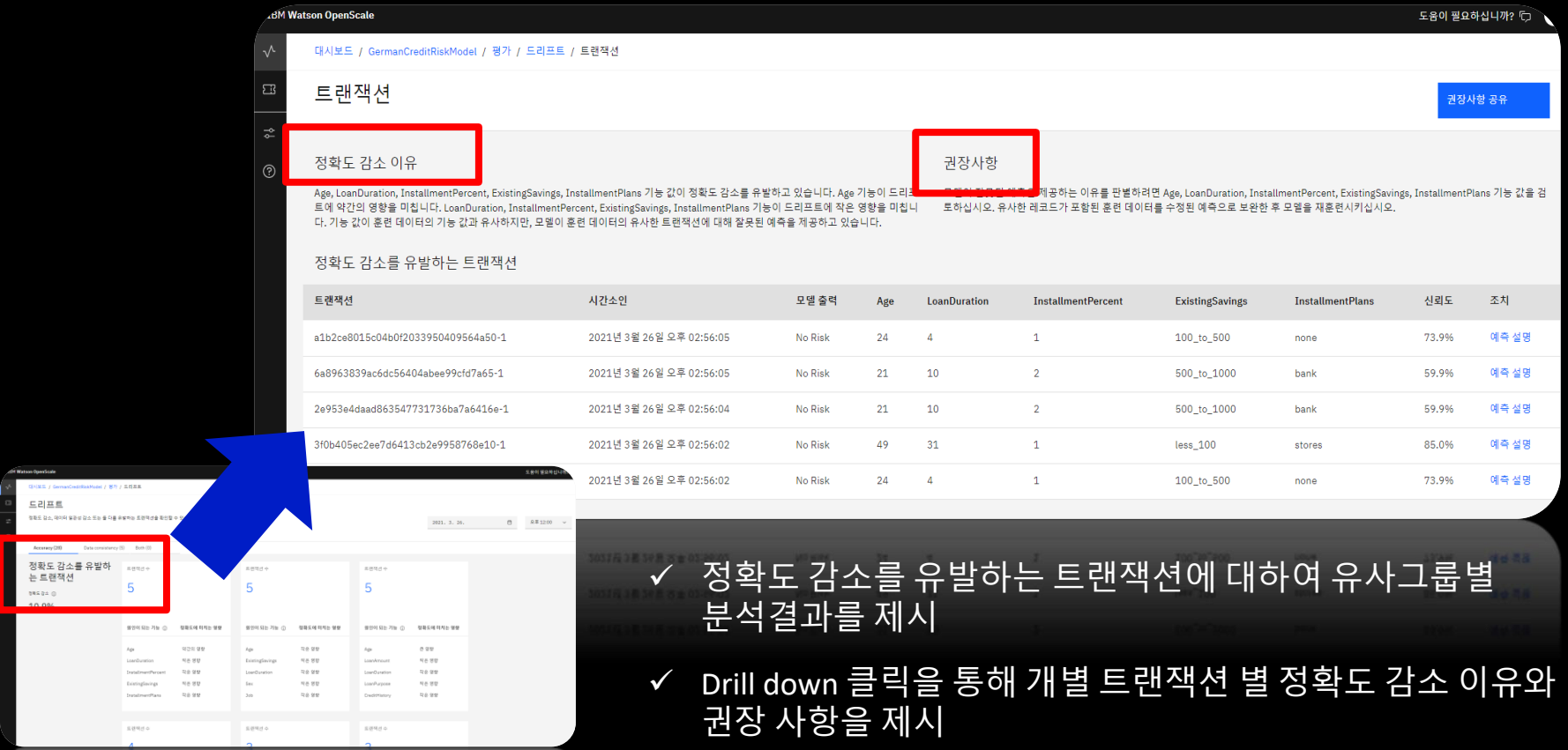
Watson OpenScale 시나리오 > 2. 품질 & 3. Drift(추세)

- ✓ 모델 품질의 하락 추세를 감지하기 위하여 아래 두가지 관점을 모니터링함

. 모델 정확도 감소 유발 트랜잭션 탐지

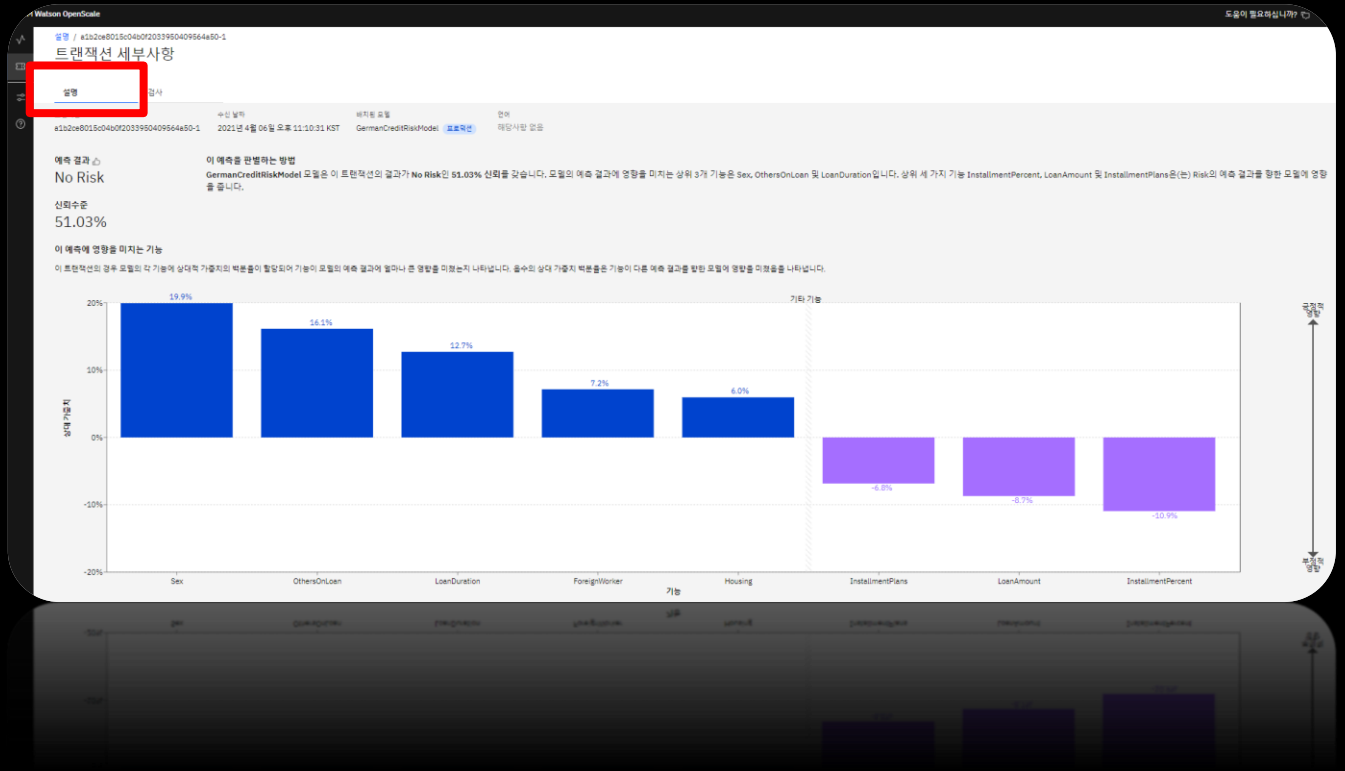
. 데이터 일관성 감소 트랜잭션 탐지

Watson OpenScale 시나리오 > 3. Drift(추세)



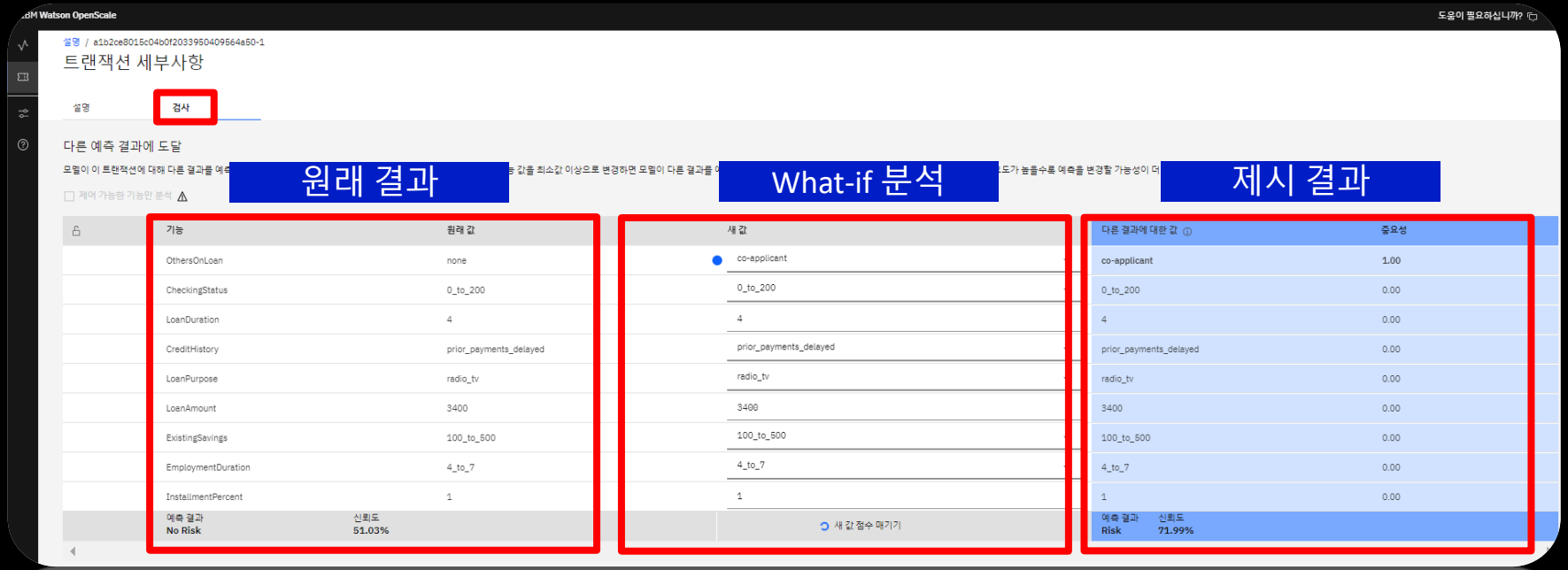
- ✓ 정확도 감소를 유발하는 트랜잭션에 대하여 유사그룹별 분석결과를 제시
- ✓ Drill down 클릭을 통해 개별 트랜잭션 별 정확도 감소 이유와 권장 사항을 제시

Watson OpenScale 시나리오 > 4. Explainability(설명)



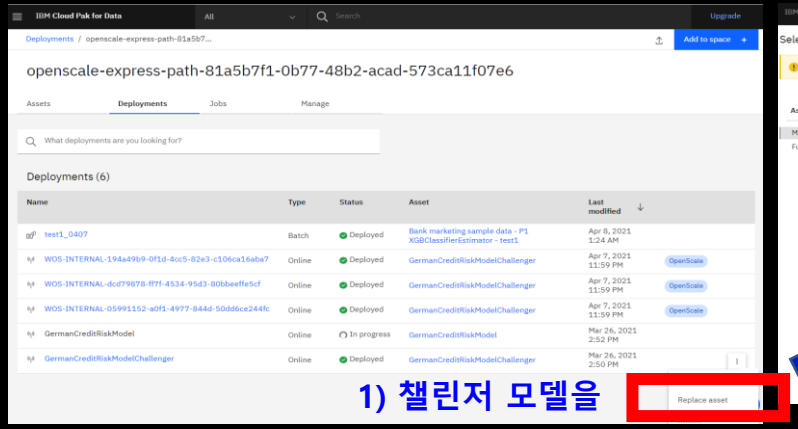
- ✓ 내장된 AI가 모든 개별 트랜잭션을 분석하며, drill down을 통해 모델 결과를 쉽고 직관적으로 설명
- ✓ 실 운영중인 모델에 input되는 개별 트랜잭션에 대하여 주요 변수별 중요도 및 입력값을 설명함으로써 컴플라이언스 규제 준수 및 고객 관리 업무를 지원

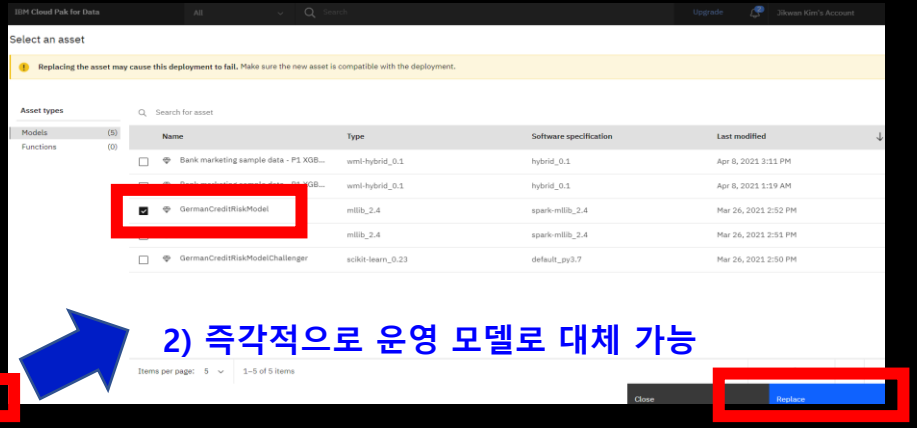
Watson OpenScale 시나리오 > 4. Explainability(설명)

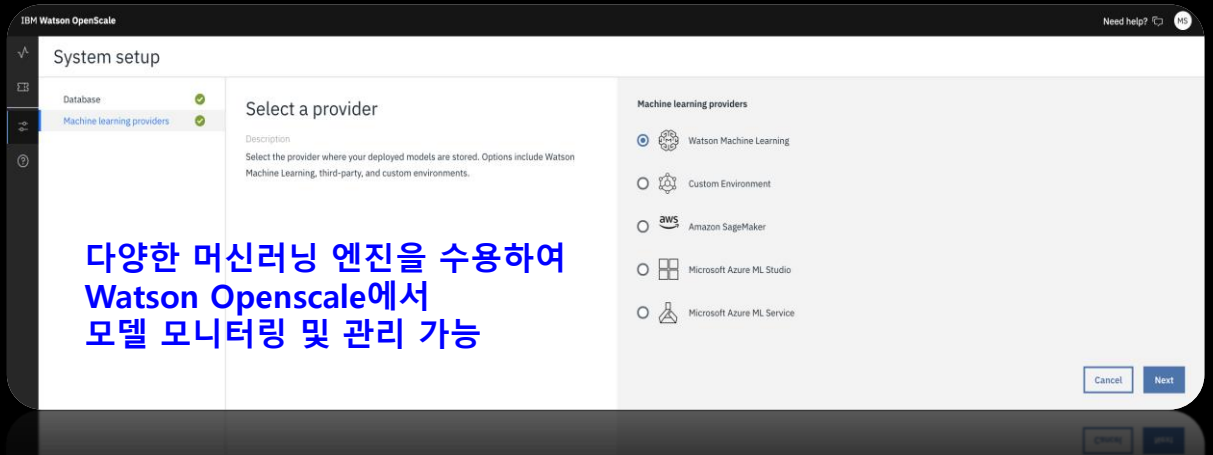


- ✓ 특정 트랜잭션에 대하여 영향력이 큰 변수가 변할 때의 모델 결과 값이 얼마나, 어떻게 달라지는지 자체 분석 결과 제시
- ✓ 다른 input 값을 대입할 때 모델 행태 및 결과가 어떻게 달라지는지 what-if 방식의 분석을 지원

Watson OpenScale 시나리오 > 기타 참고 기능







IBM은 오랜 기간 오픈소스를 통해 Trusted AI에 기여해 왔습니다.

Did anyone tamper with it?

ROBUSTNESS

Adversarial Robustness 360

↳ (ART)

github.com/IBM/adversarial-robustness-toolbox

art-demo.mybluemix.net

Is it fair?

FAIRNESS

AI Fairness 360

↳ (AIF360)

github.com/IBM/AIF360

aif360.mybluemix.net

Is it easy to understand?

EXPLAINABILITY

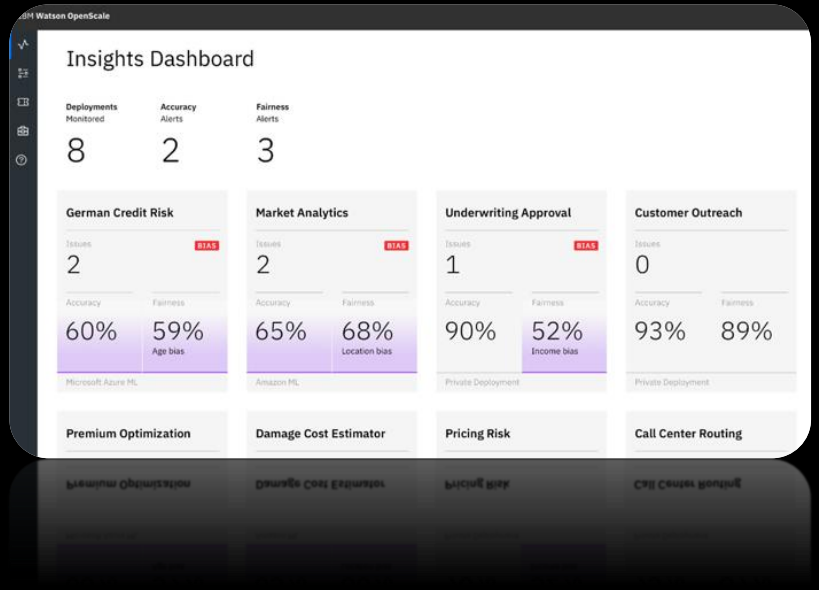
AI Explainability 360

↳ (AIX360)

github.com/IBM/AIX360

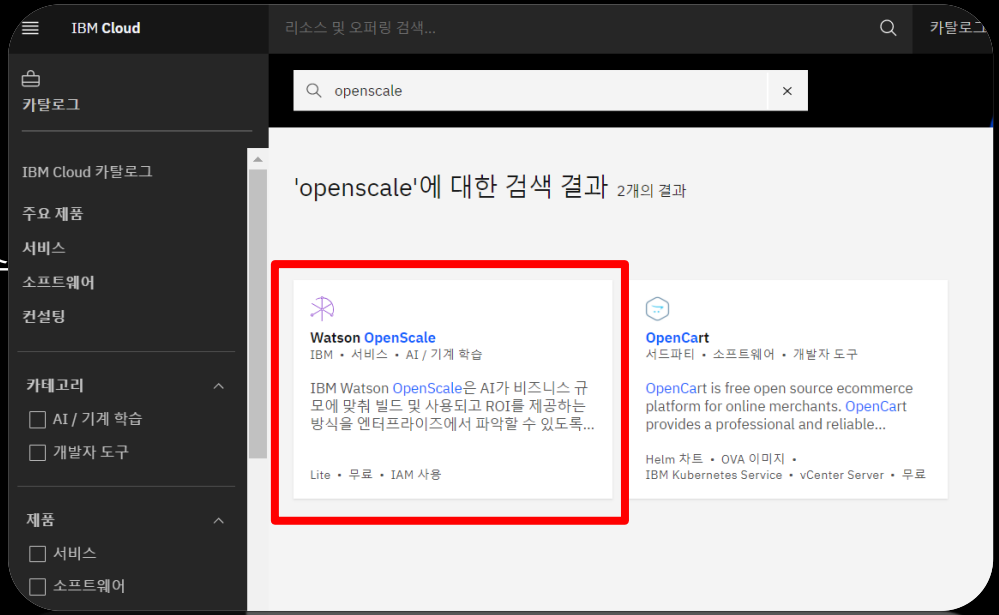
aix360.mybluemix.net

Watson OpenScale



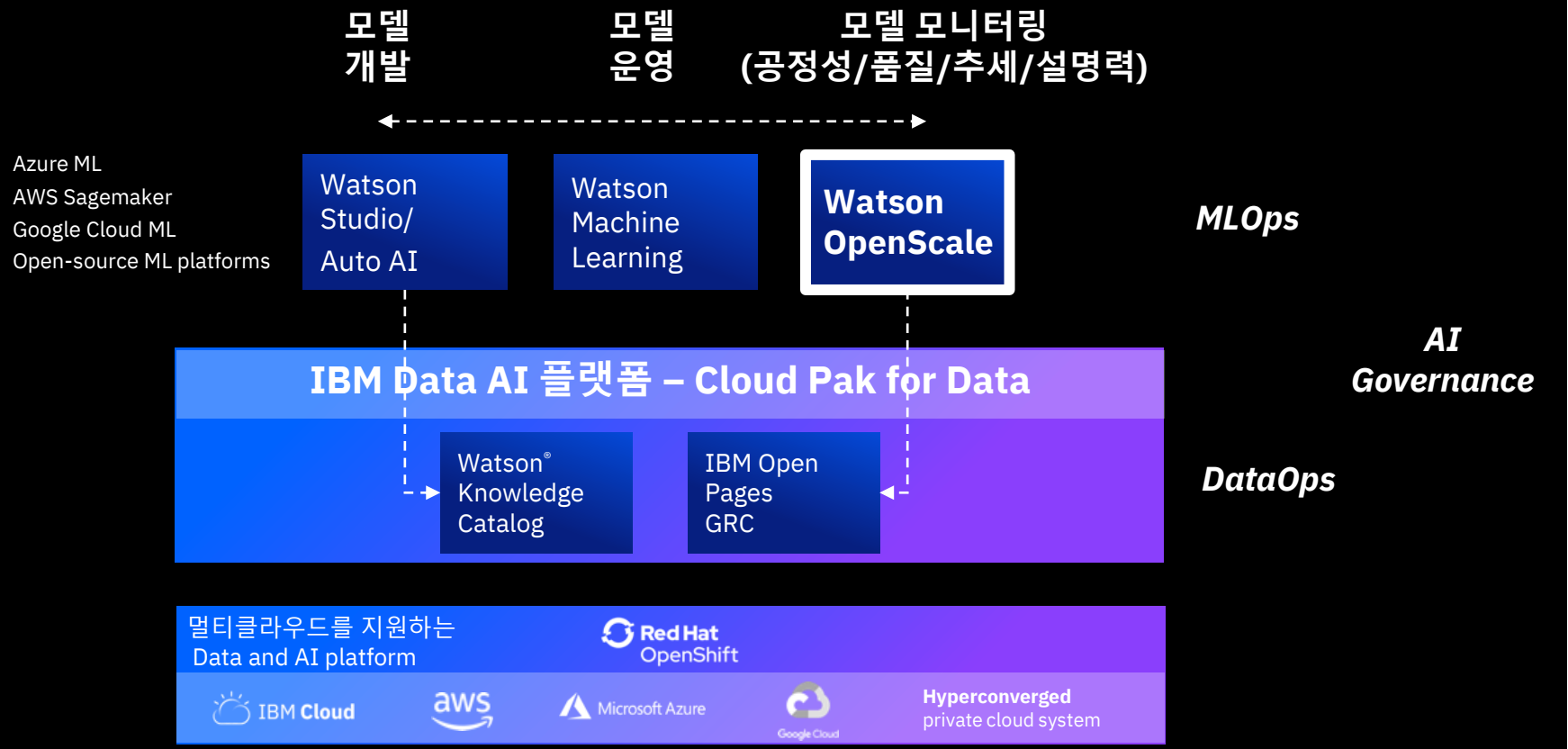
Try it out for free

- IBM Cloud에서 Watson OpenScale Lite 버전 무료 사용이 가능합니다.(일부 기능 및 워크로드 상 제약)
- 제공되는 샘플 데이터와 매뉴얼을 통해 기능을 살펴보고, 보유하고 계신 모델을 직접 활용해 보실 수 있습니다.
- https://cloud.ibm.com/catalog/services/watson-openscale

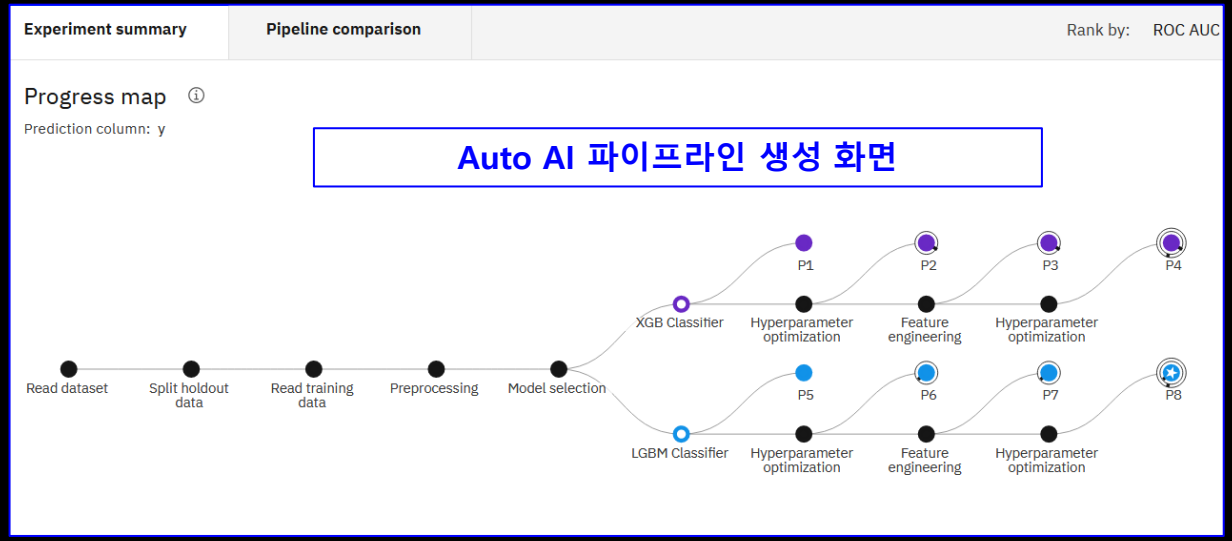


Contents

개방적이면서도 통제 가능한, 그리고 자동화된 IBM Cloud Pak for Data 는 신뢰 가능한 AI 라이프 싸이클을 지원합니다.



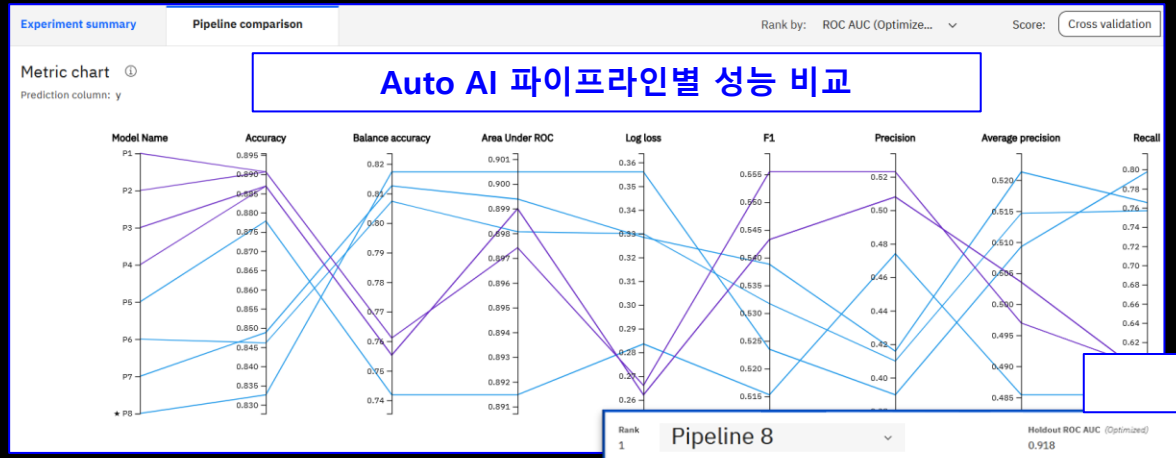
Watson Studio 내의 Auto AI 는 자동으로 데이터 전처리를 진행하고 다양한 알고리즘을 모델링에 적용하여 최적의 알고리즘을 선택합니다.
이후, 하이퍼파라미터 최적화, Feature engineering을 순차적으로 자동 수행 하고 모델링하면서 더욱 정교한 모델을 산출합니다.



왜 Auto AI 인가?

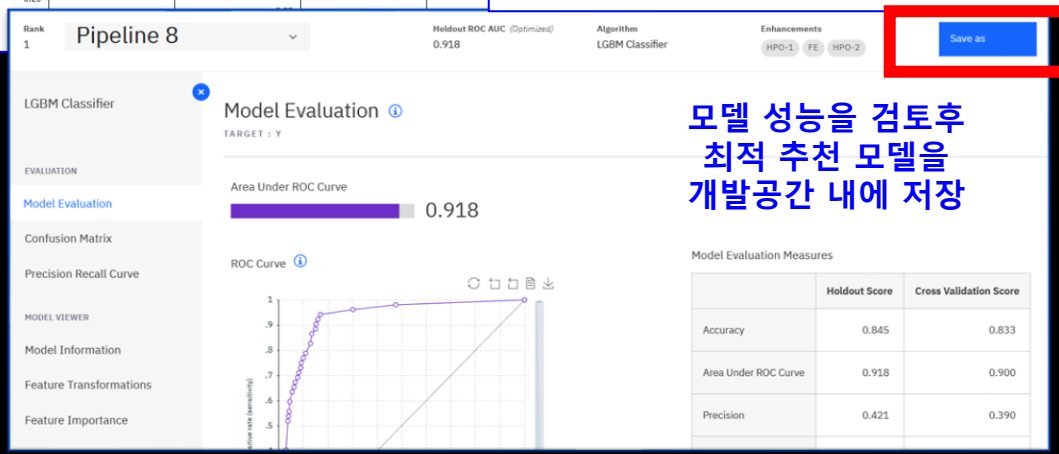
- 일반 분석가 육성 및 역량 강화(No code)
- 전문 데이터사이언티스트 생산성 향상
  (업무 관련 아이디어 및 변수 생성에 집중!)
- 모델 정교화/최적화를 지속적 자동 수행
  (신뢰 가능 AI에 있어서의 필수 요소!)

Watson Studio(모델 개발 공간) 내의 Auto AI가 자동으로 만든 여러개의 모델 결과를 검토하고 추천 받은 최적 모델을 선택/저장 후 즉시 배포할 수 있습니다.



- ✓ 모델 평가를 위한 다양한 성능지표 값을 산출하여 비교표를 제시
- ✓ Training 데이터/Test 데이터별 선택하여 각각의 성능 지표값을 비교

특정 모델에 대한 세부 성능 검토



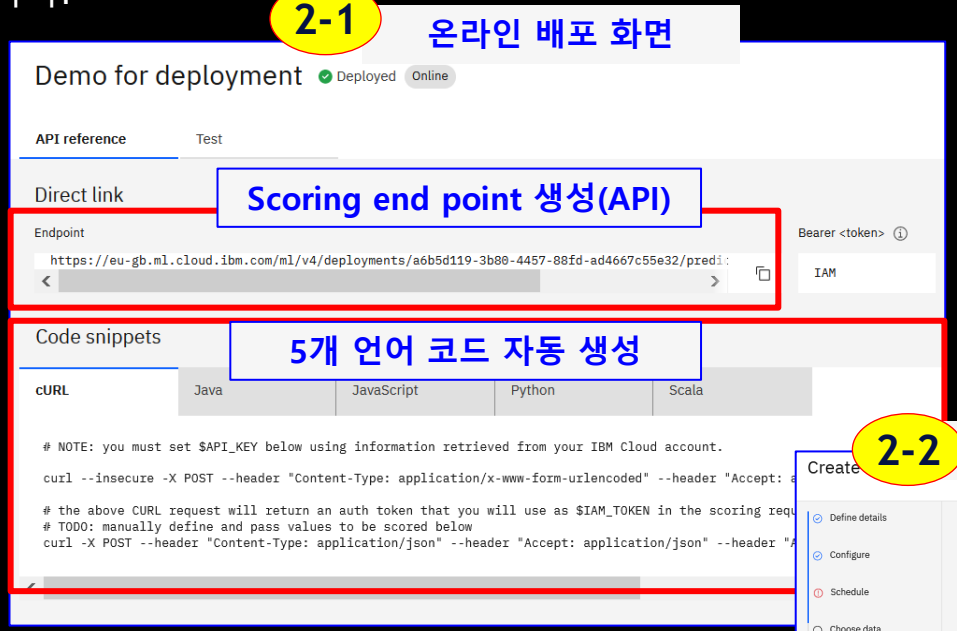
모델 성능을 검토후 최적 추천 모델을 개발공간 내에 저장

"좋은 품질의 모델을 쉽고 빠르게 만들고 적용하는 일은 신뢰 가능한 AI를 구현하는데 필수 요소입니다."

운영공간인 Watson Machine Learning(기계학습엔진)에 모델이 이동된 후, 온라인 또는 배치로 모델을 즉시, 손쉽게 배포/운영화 할 수 있습니다.

1 온라인 또는 배치 형태로 운영방법 선택

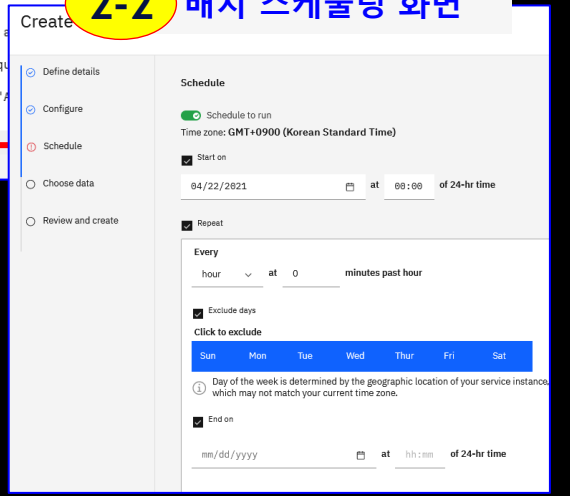
2-1 온라인 배포 화면



Scoring end point 생성(API)

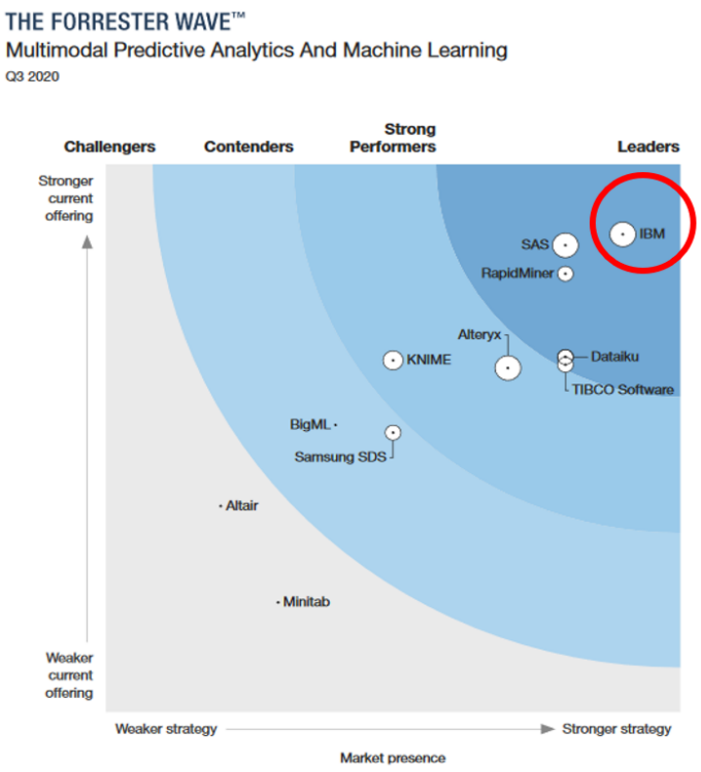
5개 언어 코드 자동 생성

2-2 배치 스케줄링 화면

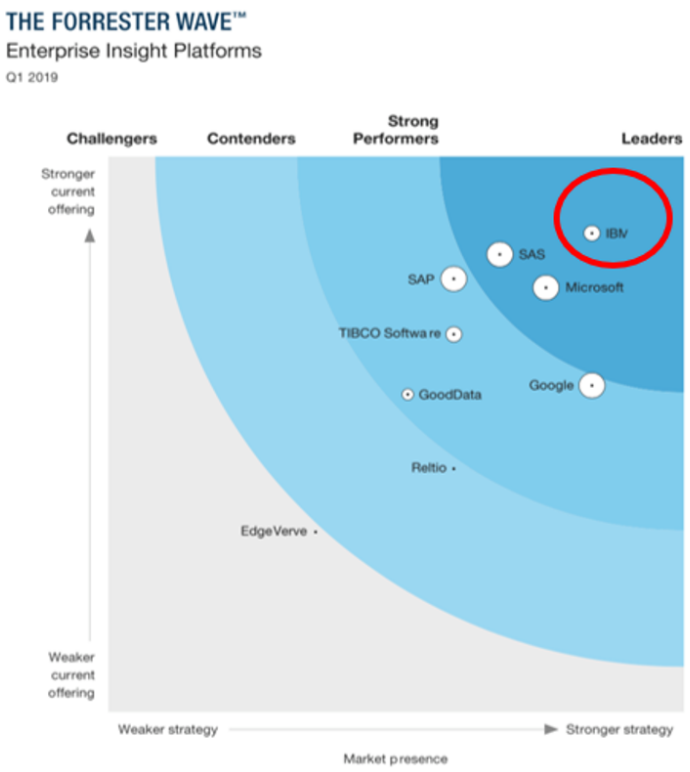


최신 외부기관 평가-1(The Forrester Wave)

The Forrester Wave Q3 2020
Multimodal Predictive Analytics And Machine Learning



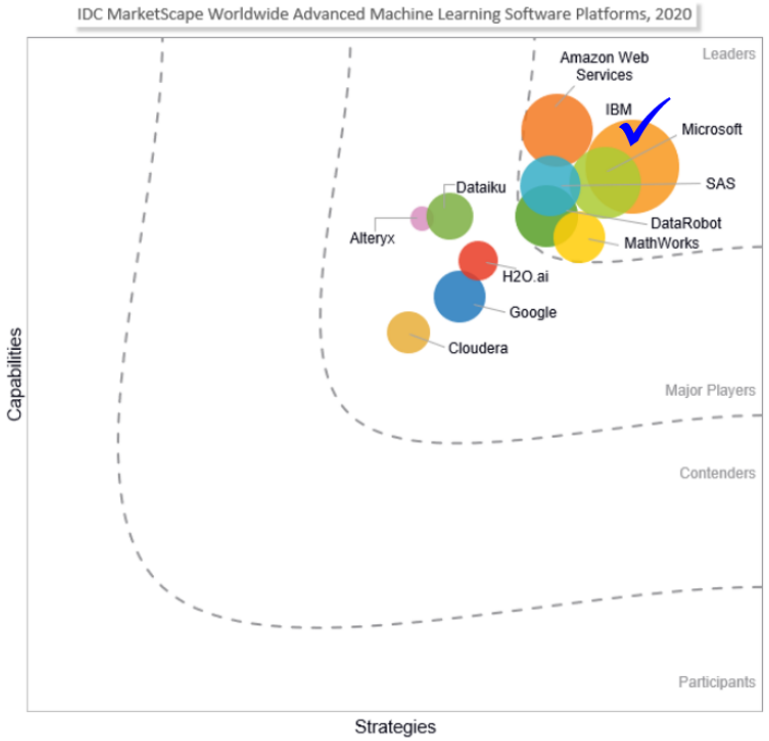
The Forrester Wave Q1 2019
Enterprise Insight Platform



최신 외부기관 평가-2(Gartner Magic Quadrant 2021/ IDC 2020)

Gartner 2021 Magic Quadrant for Data Science and Machine Learning Platforms

DC MarketScape: Worldwide Advanced Machine Learning Software Platforms 2020 Vendor Assessment




IBM