

3. 데이터 기반 제품 개선

머신 러닝을 기반으로 한 제품 개선이 무엇인지 알아보자

Contents

1. 데이터 기반 제품 개선(Product Science)이란?
2. 머신 러닝이란?
3. ML 모델 개발시 고려할 점
4. MLOps란?
5. 머신 러닝 사용시 고려할 점
6. 실습: 지표 정의하고 차트 만들어보기



데이터 기반 제품 개선이란?

머신 러닝 기술을 사용해 제품/서비스의 기능을
개선하는 것에 대해 살펴보자

◆ 데이터 과학자의 역할

❖ 머신러닝의 형태로 사용자들의 경험을 개선

- 문제에 맞춰 가설을 세우고 데이터를 수집한 후에 예측 모델을 만들고 이를 테스트
 - 장시간이 필요하지만 이를 짧은 사이클로 단순하게 시작해서 고도화하는 것이 좋음
- 테스트는 가능하면 **A/B** 테스트를 수행하는 것이 더 좋음

❖ 데이터 과학자에게 필요한 스킬셋

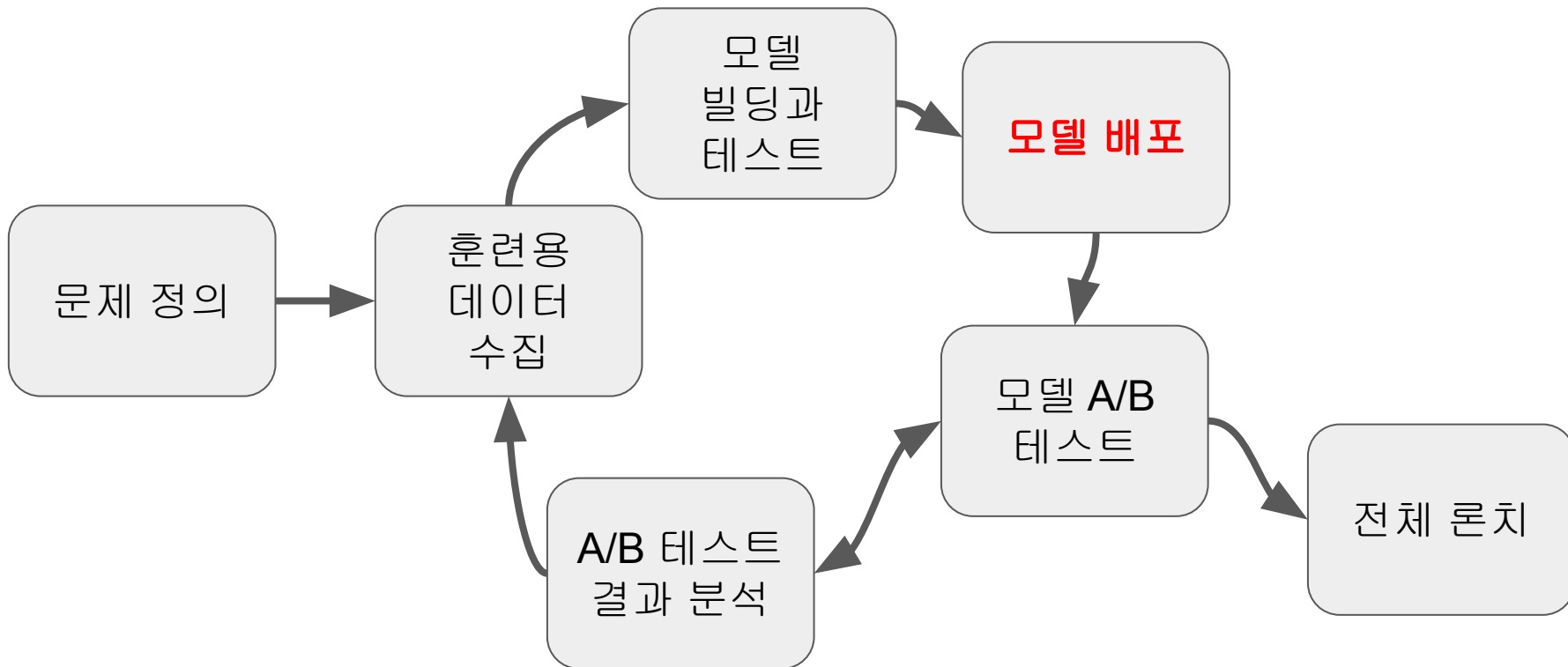
- 머신러닝/인공지능에 대한 깊은 지식과 경험
- 코딩 능력 (파이썬과 **SQL**)
- 통계 지식, 수학 지식
- 끈기와 열정. 박사 학위가 도움이 되는 이유 중의 하나

◆ 훌륭한 데이터 과학자란?

- ❖ 열정과 끈기?
- ❖ 다양한 경험?
- ❖ 코딩 능력?
- ❖ 현실적인 접근 방법?
 - 애자일 기반의 모델링
 - 딥러닝이 모든 문제의 해답은 아님을 명심
- ❖ 과학적인 접근 방법?
 - 지표기반 접근
 - 내가 만드는 모델이 목표는 무엇이고 그걸 어떻게 측정할 것인가?

제일 중요한 것은 모델링을
위한 데이터의 존재 여부

모델 개발 전체 과정 (Life-Cycle)



◆ 머신 러닝 모델링 예 - 개인화된 추천 엔진

- ❖ 유데미: 규칙 기반에서 머신 러닝 기반으로 전환
- ❖ 머신 러닝 전에는 마케터들이 규칙 기반으로 추천: **AB** 테스트가 중요해짐

◆ 머신 러닝 모델링 예 - 사기 결제 감지

❖ 훈련 데이터를 수집하는 두 가지 방법:

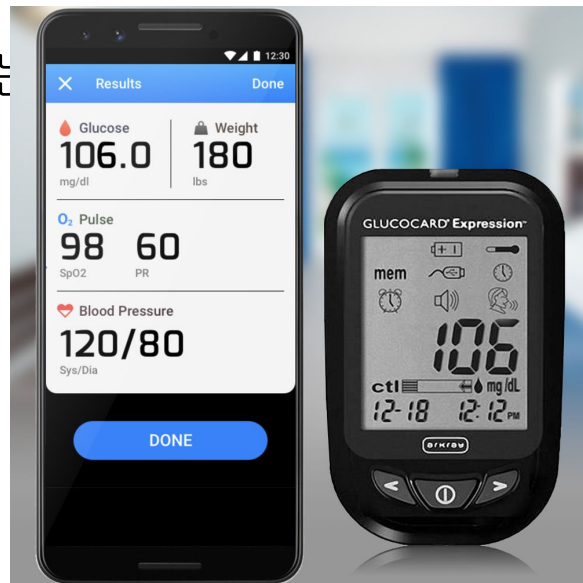
- 실제 사례 수집(신용카드 회사와 협업)
- 이상값 탐지를 실행하고 사람에게 검토 요청(휴먼 인더 루프)

❖ 몇 가지 패턴이 나타남

- 가격이 \$\$\$\$인 신규 코스가 며칠 만에 정가로 판매됨 -> 의심스러움
- 다른 강력한 지표는 무엇일까?
 - "머신 러닝의 편향성" 또는 "머신 러닝 윤리"의 중요성

◆ 머신 러닝 모델링 예 - 환자 이상 징후 예측

- ❖ 원격 환자 모니터링 (Remote Patient Monitoring)에서 많이 사용됨
- ❖ 환자의 다양한 측정 데이터를 기반으로 환자의 상태가 혹시라도 치료를 필요로 하는지 예측
- ❖ 목표는 환자의 병원 입원이나 응급실 방문을 막는



◆ 머신 러닝 모델링 예 - 농업용 자율 트랙터

❖ 존디어는 ML을 사용하여 자율 트랙터 개발

- 밭을 탐색하고 사람보다 더 효율적으로 심기 및 수확과 같은 작업을 수행



의료 이미지 (Medical Imaging) 분석

- 로봇 방사선 기술자의 대두:
 - 딥러닝 알고리즘이 MRI와 엑스레이 이미지 분석에서 사람을 능가하기 시작
 - 하지만 잘못된 진단의 경우 누구 책임인가?
- VoxelMorph라는 오픈소스 프레임웍은 딥러닝을 이용해 몇 초만에 MRI 분석
 - 사람이 하는 경우 적어도 2시간이 걸림
 - 캐글에 데모 모델 존재: <https://www.kaggle.com/kmader/voxelmorph-demo>
- 초음파 사진 기반의 심장병 진단 기술
 - Caption Health는 초음파 사진 기반의 심장병 진단 기술 개발
 - 인공 지능 기반의 이미징 기술로 FDA 승인도 받음
 - 기존의 엑스레이 기반의 CT Scan과 비교하면 안전성과 비용과 시간이 있어 엄청난 잇점 존재



머신 러닝이란?

머신 러닝이 무엇이고 어떤 종류들이 있는지 알아보도록
하자

◆ 머신 러닝(Machine Learning)의 정의

❖ Machine Learning:

- ‘A field of study that gives computers the ability to learn without being explicitly programmed’ (Arthur Samuel)

❖ “배움이 가능한 기계의 개발”

- 결국 데이터의 패턴을 보고 흉내내는 방식 (imitation)
- 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야
- 딥 러닝(신경망의 다른 이름)은 이 중의 일부
 - 비전, 자연언어처리 (텍스트/오디오)등에 적용되고 있음
- AI는 머신러닝을 포괄하는 개념

◆ 머신 러닝 모델이란?

❖ 머신 러닝을 통해서 최종적으로 만드는 것이 머신 러닝 모델

- 특정 방식의 예측을 해주는 블랙박스
 - 선택한 머신 러닝 알고리즘에 따라 내부가 달라짐
 - 디버깅은 쉽지 않음
- 입력 데이터를 주면 그를 기반으로 예측
 - 정확히 이야기하자면 **Supervised ML** (지도기계학습)

❖ 모델 트레이닝/빌딩

- 이런 머신 러닝 모델을 만드는 것을 지칭

◆ 머신 러닝의 종류

❖ 지도 기계 학습 (Supervised Machine Learning)

- 명시적 예제 (트레이닝셋)을 통해 학습: 정답이 존재
- 크게 두 종류가 존재
 - 분류 지도 학습 (Classification): 이진 분류(Binary)와 다중 분류 (Multi-class)
 - 회귀 지도 학습 (Regression)

❖ 비지도 기계 학습 (Unsupervised Machine Learning)

- 클러스터링 혹은 뉴스 그룹핑처럼 주어진 데이터를 몇 개의 그룹으로 분리
- GPT 같은 언어 모델의 훈련도 여기에 속함

❖ 강화 학습 (Reinforcement Learning)

- 알파고 혹은 자율주행

◆ 지도 기계 학습 예제: 타이타닉 승객 생존 여부 예측

❖ 이진 분류 문제 (Binary Classification)

❖ 탑승 승객별로 승객 정보와 최종 생존 여부가 트레이닝셋으로 제공됨

- 최종 생존 여부처럼 모델이 예측해야하는 필드를 레이블/타겟이라고 부름
- 기존 필드로부터 새로운 필드를 뽑아내는 것이 일반적: Feature Engineering

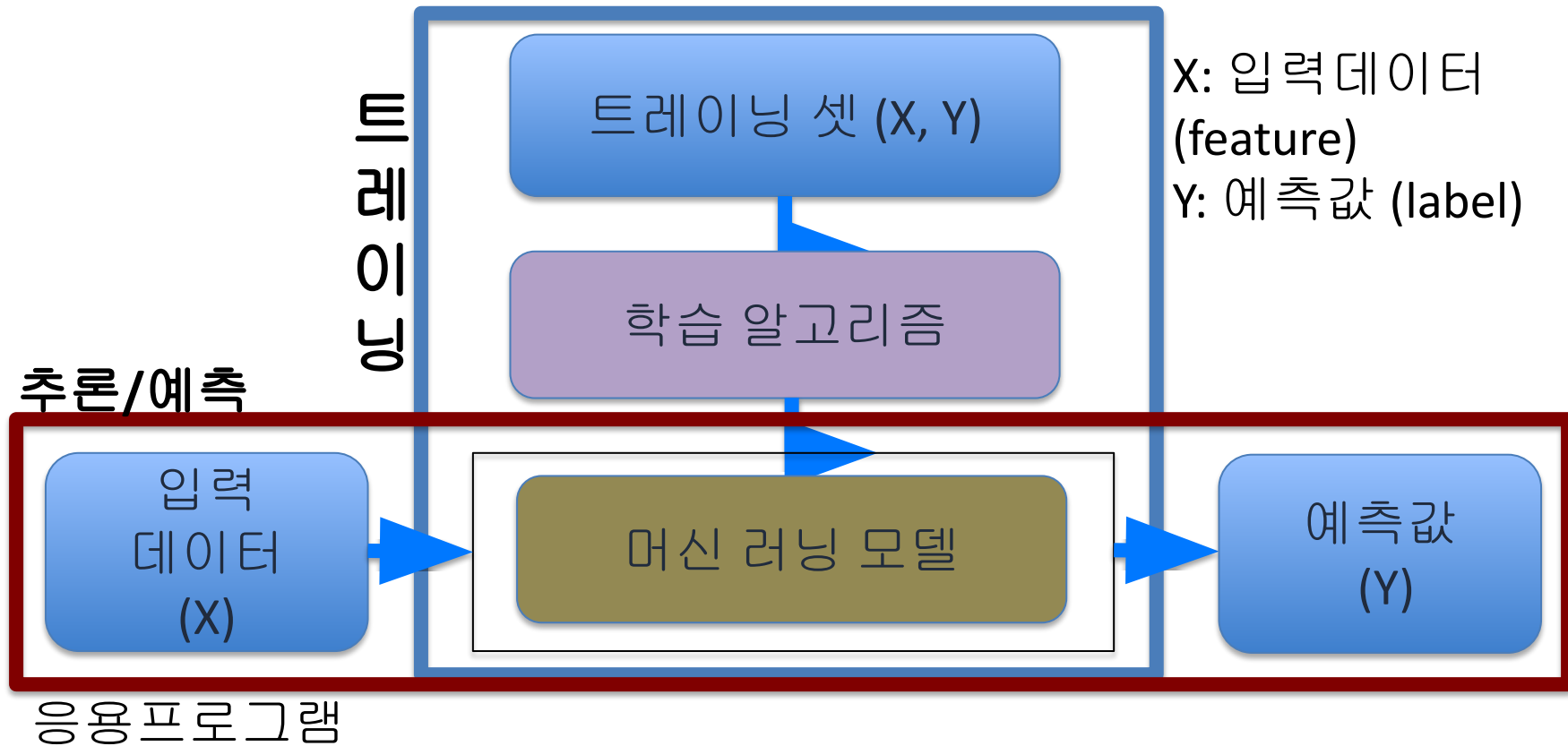
survived,pclass,name,sex,age,sibsp,parch,ticket,fare,cabin,embarked

1,2,"Hewlett, Mrs. (Mary D Kingcome) ",female,55,0,0,248706,16,,S

0,3,"Rice, Master. Eugene",male,2,4,1,382652,29.125,,Q

...

◆ 지도 기계학습



❖ 지도 학습 예: 스팸 웹 페이지 분류기

"클라우드 컴퓨팅"(Cloud Computing)
이란 집적·공유된 정보통신기기,
정보통신설비, 소프트웨어 등
정보통신자원을 이용자의 요구나 수요
변화에 따라 정보통신망을 통하여
신축적으로 이용할 수 있도록 하는
정보처리체계를 말한다(클라우드컴퓨팅
발전 및 이용자 보호에 관한 법률 제2조
제1호).

A

클라우드 컴퓨팅, 온라인 도박,
간편즉시 대출, 정보통신기기,
신용불량 대출, 온라인 카지노,
교통사고 상해 변호사, 주식투자,
부동산투자, 중고차매매, 해외여행,
저가항공편, ..

B

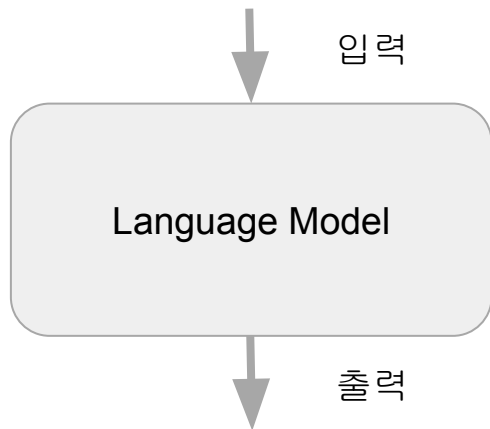
어느 쪽이 정상 웹 페이지일까?

어떤 특징을 뽑아내면 정상인지 스팸인지 결정하는데 도움이 될까?

◆ 비지도 학습 예: Language Model

- ❖ 문장의 일부를 보고 비어있는 단어를 확률적으로 맞추는 모델
- ❖ 훈련은 위키피디아에 있는 자연스러운 문장들을 대상으로 진행

Seoul is the capital of ()



Seoul is the capital of (Korea)
(South Korea)
(Republic of Korea)

(OpenAI) (transitioned) (from)
(non-profit) (to) (for-profit)

["OpenAI transitioned from",
"non-profit"]
["transitioned from non-profit", "to"]
["from non-profit to", "for-profit"]

위의 경우 context window가 4가 됨:

- 3개의 토큰을 보고 1개의 토큰 예측을 훈련
- Context window의 크기가 결국 모델의 메모리를 결정



ML 모델 개발시 고려할 점

현업에서 모델 개발시 알아야할 점들은?

◆ 모델 개발시 데이터 과학자들의 일반적인 생각

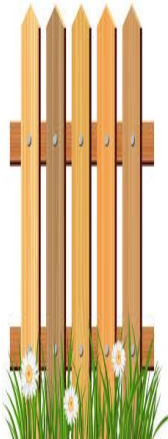
- ❖ 데이터 과학자: 아주 좋은 머신러닝 모델을 만들고 말겠어!
- ❖ 엔지니어: 모델 만들고 나서 다음 스텝은 뭐야?
- ❖ 데이터 과학자: ???

```
23 rf_model = RandomForestClassifier(  
24     n_estimators=1,  
25     criterion='gini',  
26     max_depth=7,  
27     min_samples_split=2,  
28     min_samples_leaf=5,  
29     min_weight_fraction_leaf=0.0,  
30     max_features='auto',  
31     max_leaf_nodes=None,  
32     bootstrap=True,  
33     oob_score=False,  
34     n_jobs=16,  
35     random_state=None,  
36     verbose=0,  
37     warm_start=False,  
38     class_weight=None).fit(X_train, y_train)
```



데이터 과학자

ML Model

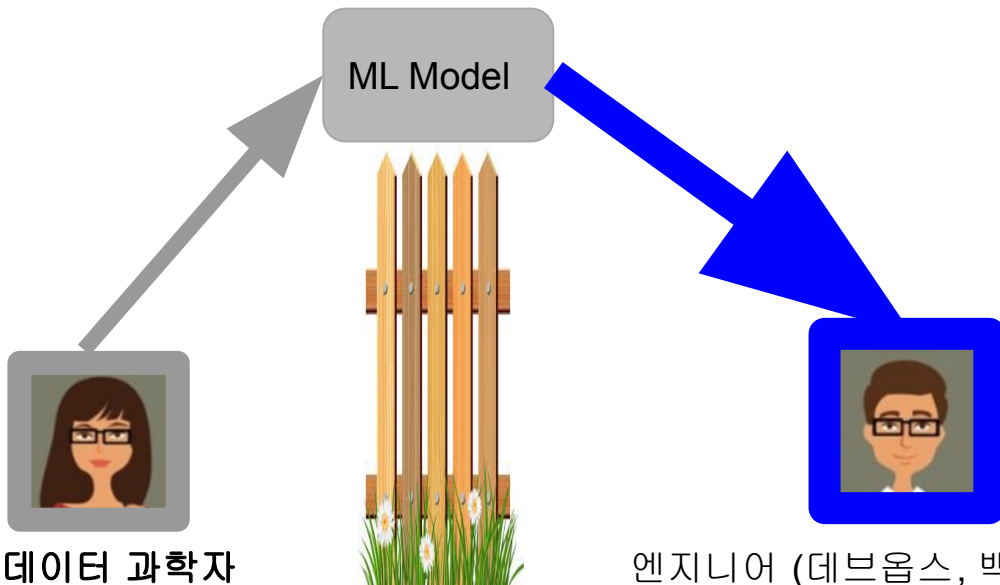


엔지니어 (데브옵스, 백엔드, ...)

◆ 모델 개발시 엔지니어들의 일반적인 생각

- ❖ 엔지니어: 머신러닝 모델을 받긴 했는데 어떻게 배포하지?
(시간이 지난 후)
- ❖ 데이터 과학자: 모델 잘 론치되었어?
- ❖ 엔지니어: 어? 응

```
23 rf_model = RandomForestClassifier(  
24     n_estimators=1,  
25     criterion='gini',  
26     max_depth=7,  
27     min_samples_split=2,  
28     min_samples_leaf=5,  
29     min_weight_fraction_leaf=0.0,  
30     max_features='auto',  
31     max_leaf_nodes=None,  
32     bootstrap=True,  
33     oob_score=False,  
34     n_jobs=16,  
35     random_state=None,  
36     verbose=0,  
37     warm_start=False,  
38     class_weight=None).fit(X_train, y_train)
```

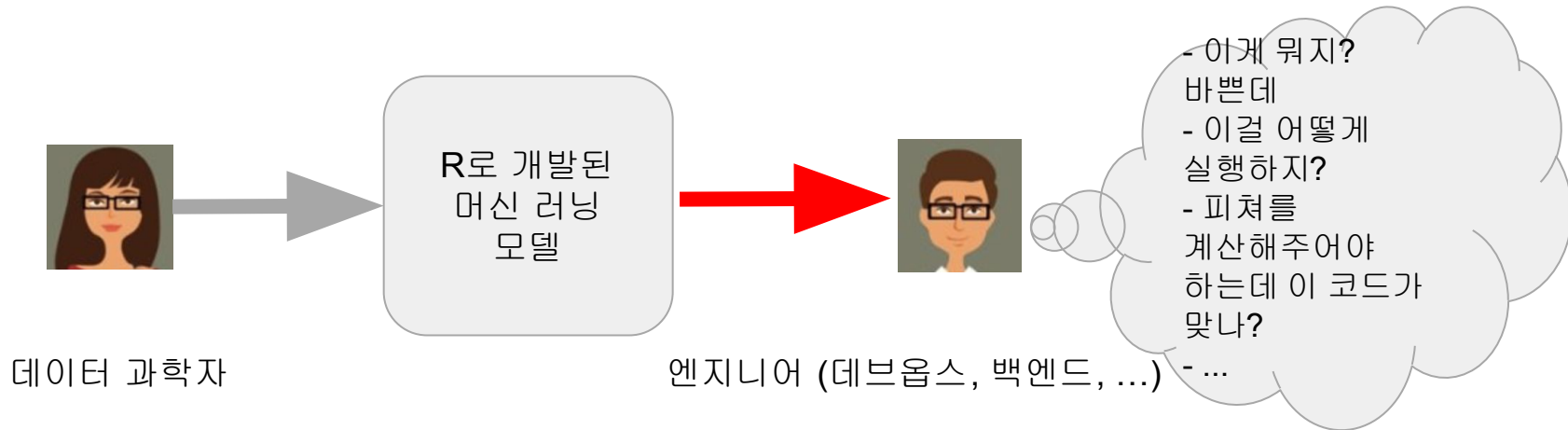


데이터 과학자

엔지니어 (데브옵스, 백엔드, ...)

◆ 마찰이 생기는 지점 - 개발된 모델의 이양 관련

- ❖ 많은 수의 데이터 과학자들은 R을 비롯한 다양한 툴로 모델 개발
- ❖ 하지만 실제 프로덕션 환경은 다양한 모델들을 지원하지 못함
 - 개발/검증된 모델의 프로덕션 환경 론치시 시간이 걸리고 오류 가능성이 존재
 - 심한 경우 모델 관련 개발을 다시 해야함 (피쳐 계산과 모델 실행 관련)



◆ 모델 개발시 꼭 기억할 포인트 (1)

❖ 누군가 모델 개발부터 최종 론치까지 책임질 사람이 필요

- 모델 개발은 시작일 뿐이고 성공적인 프로덕션 론치가 최종적인 목표
- 이 일에 참여하는 사람들이 같이 크레딧을 받아야 협업이 더 쉬워짐
 - 최종 론치하는 엔지니어들과 소통하는 것이 중요

❖ 모델 개발 초기부터 개발/론치 과정을 구체화하고 소통

- 모델 개발시 모델을 어떻게 검증할 것인지?
- 모델을 어떤 형태로 엔지니어들에게 넘길 것인지?
 - 피쳐 계산을 어떻게 하는지? 모델 자체는 어떤 포맷인지?
- 모델을 프로덕션에서 **A/B** 테스트할 것인지?
 - 한다면 최종 성공판단 지표가 무엇인지?

◆ 모델 개발시 꼭 기억할 포인트 (2)

❖ 개발된 모델이 바로 프로덕션에 론치가능한 프로세스/프레임웍이 필요

- 예를 들어 **R**로 개발된 모델은 바로 프로덕션 론치가 불가능
- 트위터: 데이터 과학자들에게 특정 파이썬 라이브러리로 모델개발 정책화
 - 툴을 하나로 통일하면 제반 개발과 론치 관련 프레임웍의 개발이 쉬워짐
- **머신러닝 전반 개발/배포 프레임웍**의 등장
 - 머신러닝 모델 개발, 검증, 배포를 하나의 프레임웍에서 수행
 - **AWS SageMaker**가 대표적인 프레임웍
 - 검증된 모델을 버튼 클릭 하나로 **API** 형태로 론치 가능!
 - **AutoPilot**이란 **AutoML** 기능도 제공
 - **Google Cloud**와 **Azure**도 비슷한 프레임웍 지원
 - 우버/리프트/넷플릭스 등의 **IT** 기업도 자체 머신러닝 개발/배포 프레임웍을 개발

◆ 모델 개발시 꼭 기억할 포인트 (3)

❖ 첫 모델 론치는 시작일 뿐

- 론치가 아닌 운영을 통해 점진적인 개선을 이뤄내는 것이 중요!
- 데이터 과학자의 경우 모델 개발하고 끝이 아니라는 점 명심!

❖ 결국 피드백 루프가 필요

- 운영에서 생기는 데이터를 가지고 개선점 찾기
 - 검색이라면 CTR(Click Through Rate)을 모니터링하고 모든 데이터를 기록
- 주기적으로 모델을 재빌딩하고 배포
 - Continuous Model Update and Monitoring
 - 이로 인해 탄생한 직군이 MLOps

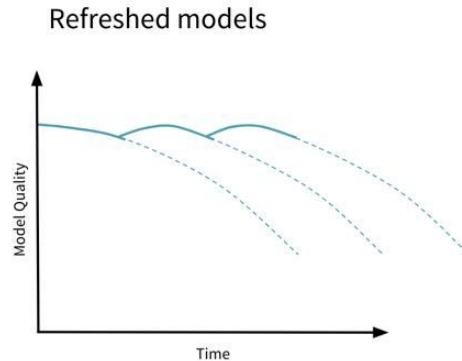
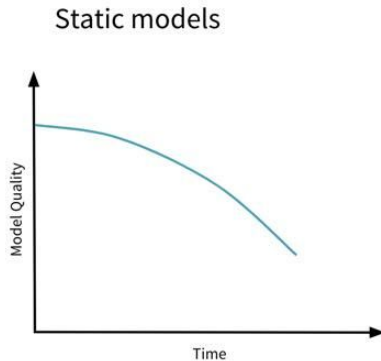
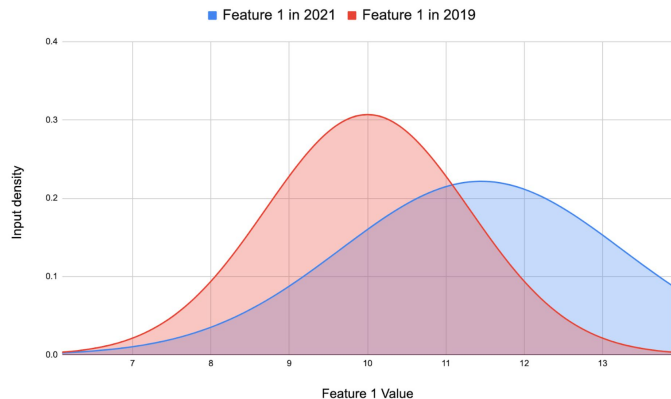


MLOps란?

모델 빌드, 배포, 모니터링 전체 프로세스를 자동화하는
직군!

Data Drift로 인한 모델 성능 저하

- ML 모델에서 가장 중요한 것은 훈련 데이터
- 시간이 지나면서 훈련에 사용한 데이터와 실제 환경의 데이터가 다르게 변화함
 - 이를 **Data drift**라고 부르며 이를 모니터링하는 것이 중요
- 주기적으로 ML 모델을 다시 학습시키는 것이 중요

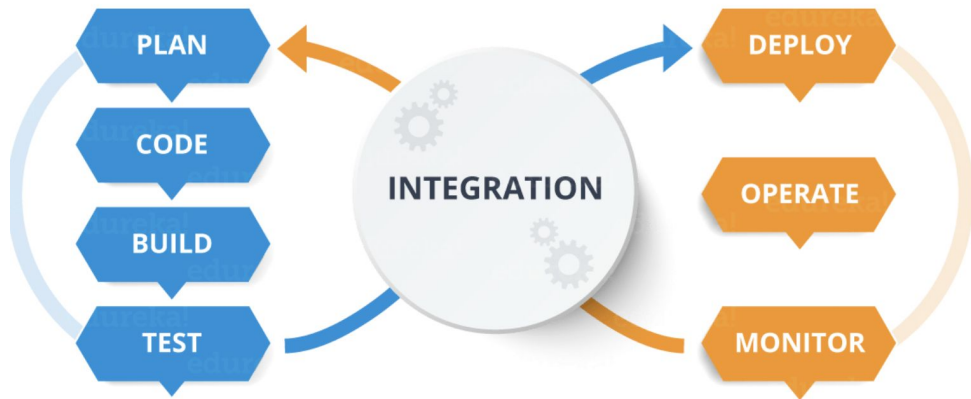


MLOps vs. DevOps

- DevOps가 하는 일은?
 - Deliver software faster and more reliably in automated fashion
 - 개발자가 만든 코드를 시스템에 반영하는 프로세스 (CI/CD)
 - 시스템이 제대로 동작하는지 모니터링 그리고 이슈 감지시 **escalation** 프로세스 수행
 - On-call 프로세스
- MLOps가 하는 일은?
 - Deliver ML models faster and more reliably in automated fashion
 - 앞의 DevOps가 하는 일과 동일. 차이점은 개발자 코드가 아니라 ML 모델이 대상이 된다는 점
 - 모델을 계속적으로 빌딩하고 배포하고 성능을 모니터링
 - ML모델 빌딩과 프로덕션 배포를 자동화할 수 있을까? 지속적인 모델 빌딩(CT)과 배포!
 - 모델 서빙 환경과 모델의 성능 저하를 모니터링하고 필요시 **escalation** 프로세스 진행
 - Latency의 중요성
 - Data drift 측정

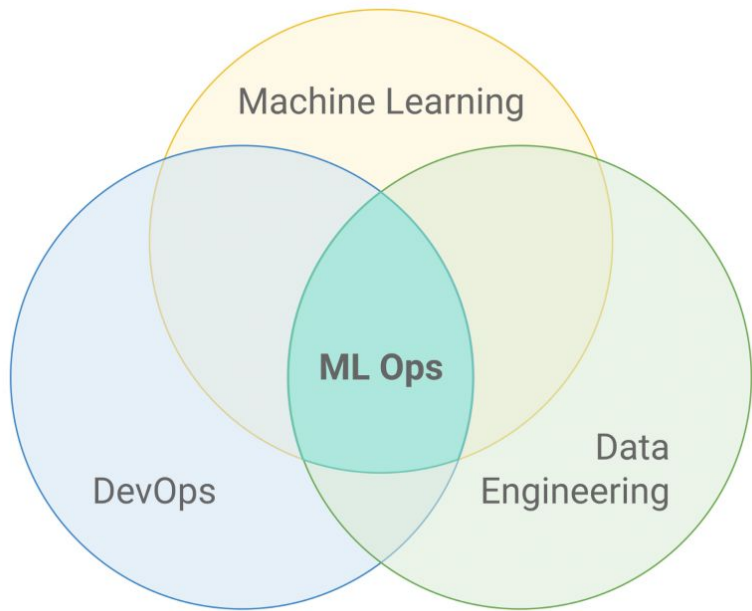
CI & CD

- CI (Continuous Integration)
 - Developers frequently merge code changes into a central repo
 - Building and testing are automated
- CD (Continuous Delivery or Deployment)
 - Passing builds (packages) are deployed directly to the production environment



MLOps 엔지니어가 알아야하는 기술

- 데이터 엔지니어가 알아야 하는 기술
 - 파이썬/스칼라/자바
 - 데이터 파이프라인과 데이터 웨어하우스
- DevOps 엔지니어가 알아야 하는 기술
 - CI/CD, 서비스 모니터링, ...
 - 컨테이너 기술 (K8S, 도커)
 - 클라우드 (AWS, GCP, Azure)
 - Infrastructure As Code (Configuration As Code)
- 머신러닝 관련 경험/지식
 - 머신러닝 모델 빌딩과 배포
 - ML 모델 빌딩 프레임웍 경험
 - SageMaker, Kubeflow, MLflow



<https://builtin.com/machine-learning/mlops>

머신 러닝 사용시 고려할 점

머신 러닝(혹시 AI)을 제대로 사용하는 것은 쉽지 않다.
어떤 고려할 점들이 있는지 알아보자

◆ 데이터 윤리와 주의할 점, MLOps

❖ 데이터로부터 패턴을 찾아 학습

- 데이터의 품질과 크기가 중요
- 데이터로 인한 왜곡 (**bias**) 발생 가능
 - AI 윤리
- 내부동작 설명 가능 여부도 중요
 - ML Explainability
- 데이터 권리도 중요한 문제

데이터 기반 AI는 완벽한가? (1)

- 트레이닝셋의 품질은 어떤가?
 - 데이터의 양도 중요하지만 품질도 중요: Garbage In Garbage Out
 - 미국 EMR(Electronic Medical Record)이 아주 좋은 예
- AI 도입시 가능한 문제들을 어떻게 해결할 것인가?
 - 왜 어떤 결과가 나왔는지 설명이 가능한가?
 - 알고리즘 자체에 인종이나 특정 편향성이 있지는 않은가?
- 많은 시도와 실패가 필요 -> 혁신을 만들어낼 생태계와 법률이 필요

데이터 기반 AI는 완벽한가? (2)

- EU의 관련 법규는 많은 시사점을 줌: Trustworthy AI
 - 감독 (human agency and oversight)
 - 견고성과 안전성 (robustness and safety)
 - 개인 정보 보호 및 데이터 거버넌스 (privacy and data governance)
 - 투명성 (Transparency)
 - 다양성과 비차별성과 공정성 (Diversity, nondiscrimination and fairness)
 - 사회/환경 친화적 (Societal and environmental well-being)
 - 문제 발생시 책임 소재 (Accountability)

잘못된 개인정보 보존으로 인한 페널티

- **HIPAA (Health Insurance Portability and Accountability Act)**
 - 개인 의료 정보 보호를 목적으로 하며 1996년부터 효력 발표
 - 다음과 같은 전자 의료 정보를 보호하려는 목적: ePHI (electronic Protected Health Information)
 - 개인을 식별할 수 있는 정보로 대략 18개가 존재
 - 이름, 주소, 생년월일, 전화번호, 이메일 주소, 주민등록번호, 라이선스 번호, IP 주소 등등
 - MRN (Medical Record Number), 계좌 정보, 바이오메트릭 정보 (지문 등)
- **GDPR/CCPA**
 - 각각 유럽연합과 미국 캘리포니아 주의 온라인 상에서 개인정보 보호에 관한 법률
 - 데이터 암호화
 - 예를 들면 데이터 저장시 암호화, 데이터 송수신시 암호화 (암호화 프로토콜 사용)

집단 이기주의: 의료분야 예

- 한국에서는 왜 비디오 진료가 안 될까?
 - 1999년에 이미 서울대 병원과 분당 KT가 원격 진료 연결 시범 사업을 했음
 - 코로나로 한시 허용된 원격의료, 의사 반발에 또 표류: 의료법이 여전히 개정되지 못함
 - 미국은 50개주 모두 일정 부분 원격 진료 허용 (Telehealth Parity Law)
 - 원격진료는 과연 의사들에게 나쁜 영향을 줄까?
- AI 발전에 영향받는 분야의 교육 방향에 대한 시사점
 - AI 시대에 의사의 역할은 무엇인가? 진료시간 확대와 공감 능력을 더 중요시?
 - 기존 교육 시스템의 점검 뿐만 아니라 재교육 필요성 증대
 - 일이 없어진다고 보다는 바뀐다는 점이 강조되어야 함
- 세상의 변화를 거스르기 보다는 새로운 역할을 찾는 것이 더 건강하지 않을까?

AI의 발전과 미래 직업의 변화: 예) 의사의 역할

- AI는 의사를 대체하기 보다는 의사의 효율성과 진단/치료의 정확성을 높이는 보조적 역할
 - 현재 의사는 다른 잡무로 인해 환자와 충분한 시간을 보내지 못함
 - 아무리 경험이 많은 의사라 해도 실수를 할 수 있고 의사마다 굉장히 다른 진료결과를 냄
 - AI는 진단 절차를 체계적으로 만들고 작업을 빠르고 정확하게 하는데 사용가능
 - 일종의 의사결정트리 (Decision Tree)
- 중단기적으로 의사의 역할에 대해 재고가 필요
 - 그에 따라 교육 시스템도 변경이 필요
 - 데이터 관련 교육 (Data Literacy)이 절대적으로 필요
 - 환자와의 진료/대화 (공감)에 더 많은 시간을 쏟기
 - Compassionomics(책 제목)에 따르면 공감을 더 잘하는 의사에게 진료를 받은 환자가 더 좋은 의학적 결과가 보였고 공감을 더 잘하는 의사들이 일을 더 재미있게 하고 번아웃이 덜 되었다고 함

미래의 의사 모습은 어떨까?

- 현대 비행기의 기장 역할이 좋은 예
 - 현대 비행기 조종사는 비행 소프트웨어가 보여주는 각종 정보를 대시보드를 통해 제공받음
 - 조종사들은 소프트웨어가 주는 정보를 따라하는데 거부감이 없음
 - 또한 조종사들은 매번 비행마다 안전을 보장하기 위해 반드시 체크해야하는 리스트가 존재
- 미래의 의사도 비슷하지 않을까?
 - 인공지능 기반의 각종 진단과 치료 정보를 제공받고 그걸 기반으로 의료 서비스를 제공
 - 이를 통해 효율적이고 오진이 적은 의료 서비스 제공
 - 의사들의 진료전 체크리스트
 - 병원에서 발생하는 많은 이차감염은 의사/간호사들의 비위생적인 행동으로 발생
 - 예를 들면 수술전에 손을 씻지 않음
 - 이렇게 간단하지만 필수적인 행동들을 체크리스트로 관리하고 시행



실습: 머신 러닝 모델 만들어보기

Simple ML for Sheets을 사용해보자

◆ Simple ML for Sheets

- ❖ 구글 스프레드시트의 무료 확장판
- ❖ 시트 상의 데이터를 훈련 데이터로 간단한 모델을 만들 수 있음

The screenshot shows the 'Simple ML for Sheets' interface. A Google Sheet with penguin data is open. The data includes columns for island, bill length, bill depth, flipper length, body mass, sex, year, and species. The 'species' column is highlighted in blue. A green box labeled '1. Select a task' points to the 'Predict species' option in the sidebar. Another green box labeled '2. Click on Predict' points to the 'Predict' button at the bottom of the sidebar. A third green box labeled '3. And get the predictions' points to the predicted species values in the 'Pred:Conf:species' column.

1. Select a task

3. And get the predictions

2. Click on Predict

What do you want to do?

Predict missing values

Predict the values of the empty cells of a column.

Learn how to use this task

Column with empty cells

species

Source columns

Advanced options

Results

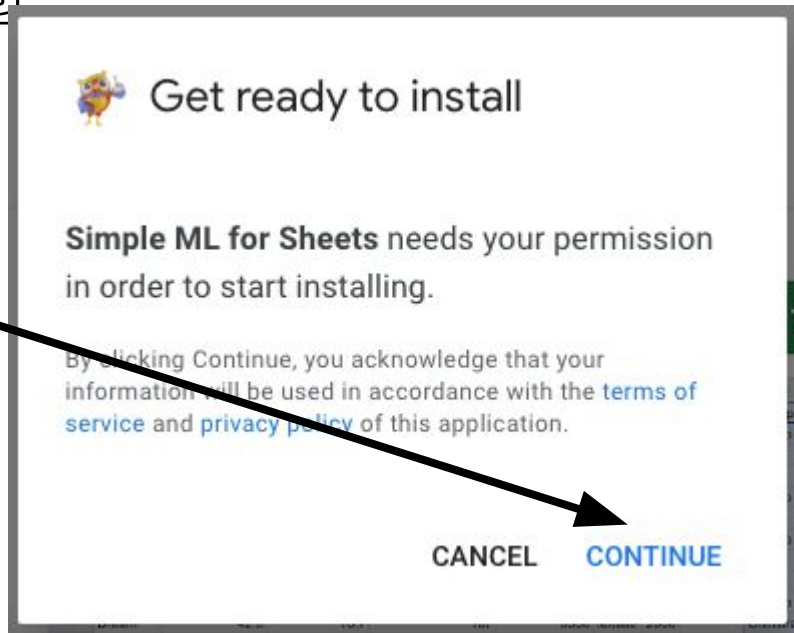
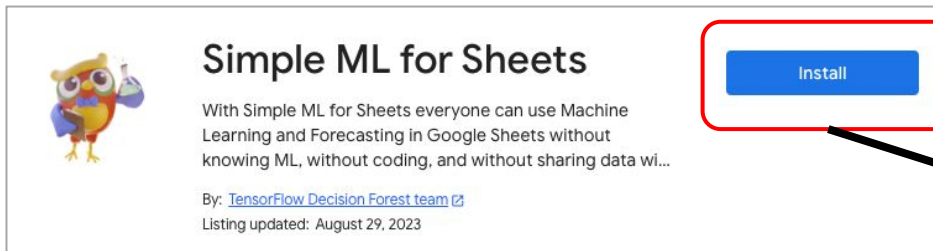
Predict

Send feedback

	A	B	C	D	E	F	G	H	I	J
	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year	species	Pred:species	Pred:Conf:species
1	Biscoe	47.8	15	215	5650	male	2007	Gentoo		99.23 %
2	Dream	40.2	17.1	193	3400	female	2009	Adelle		99.06 %
3	Dream	36	17.8	195	3450	female	2009	Adelle		99.25 %
4	Biscoe	49.8	15.9	229	5950	male	2009	Gentoo		99.23 %
5	Biscoe	38.6	17.2	199	3750	female	2009	Adelle		99.25 %
6								Gentoo		99.21 %
7								Adelle		99.25 %
8								Adelle		99.25 %
9								Gentoo		99.23 %
10								Chinstrap		93.60 %
11	Dream	42.5	16.7	187	3350	female	2008	Adelle		99.25 %
12	Torgersen	34.1	18.1	193	3475		2007	Adelle		99.26 %
13	Dream	37.5	18.5	199	4475	male	2009	Adelle		99.21 %
14	Dream	36.4	17	195	3325	female	2007	Chinstrap		99.25 %
15	Dream	45.7	17.3	193	3600	female	2009	Chinstrap		99.22 %
16	Dream	51.9	19.5	209	3950	male	2009	Chinstrap		99.26 %
17	Biscoe	46.2	14.5	209	4800	female	2007	Gentoo		99.26 %
18	Dream	42.5	17.3	187	3350	female	2009	Chinstrap		91.22 %
19	Torgersen	34.6	21.1	198	4400	male	2007	Adelle		
20	Biscoe	45.2	15.8	215	5300	male	2008	Gentoo		
21	Biscoe	45.3	13.8	208	4200	female	2008	Gentoo		
22	Dream	35.7	18	202	3550	female	2008	Adelle		
23	Torgersen	45.8	18.9	197	4150	male	2008	Adelle		
24	Dream	50.3	20	197	3300	male	2007	Chinstrap		
25	Biscoe	37.7	16	183	3075					
26	Dream	40.2	20.1	200	3975					
27	Dream	38.1	18.6	190	3700					
28	Biscoe	36.4	17.1	184	2850					
29	Dream	50.8	19	210	4100	male	2009	Chinstrap		
30	Dream	46.6	17.8	193	3800	female	2007	Chinstrap		

◆ Simple ML for Sheets 설치

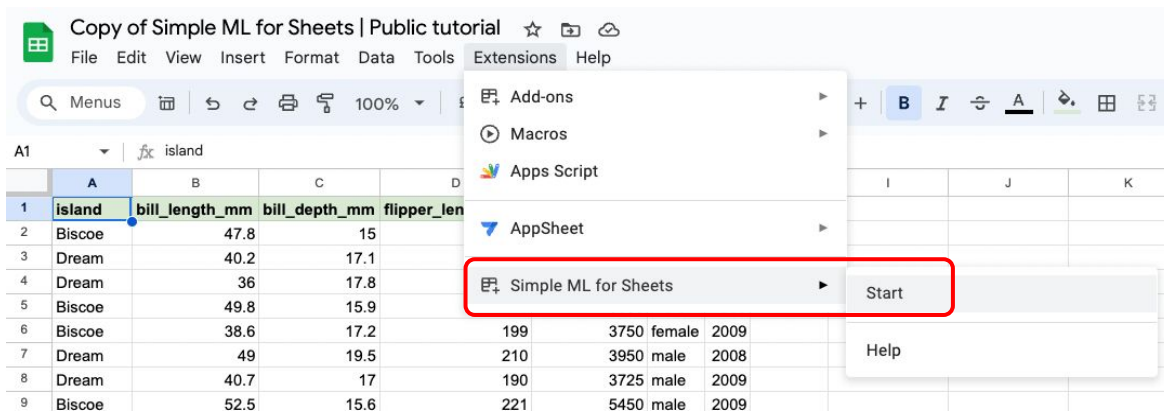
- ❖ 위 링크를 눌러 페이지로 이동
- ❖ 여기서 우측 상단의 **Install** 버튼 클릭



◆ Simple ML for Sheets 실습 (1)

❖ 예제 시트를 복사

❖ Extensions 메뉴에서 Simple ML for Sheets 선택



◆ Simple ML for Sheets 실습 (2)

❖ Simple ML for Sheets 메뉴바에서 “Predict missing values” 선택

	A	B	C	D	E	F	G	H
1	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year	species
25	Dream	37.5	18.5	199	4475	male	2009	
26	Dream	36.4	17	195	3325	female	2007	
27	Dream	45.7	17.3	193	3600	female	2009	
28	Dream	51.9	19.5	206	3950	male	2009	
29	Biscoe	46.2	14.5	209	4800	female	2007	
30	Dream	42.5	17.3	187	3350	female	2009	
31	Torgersen	34.6	21.1	198	4400	male	2007	Adelie
32	Biscoe	45.2	15.8	215	5000	male	2008	Gentoo
33	Biscoe	45.3	13.8	215	5000	male	2008	Gentoo

1

Simple ML for Sheets

← Predict missing values

Find the most likely values of empty cells.
[Documentation](#)

Column with empty cells

species

> Source columns ⓘ

> Advanced options


> Results


Predict


3

Simple ML for Sheets

What do you want to do?

 **Predict missing values**
Find the most likely values of empty cells.

 **Spot abnormal values**
Find values that look strange and what value would be expected instead.

 **Forecast future values**
Predict future data based on past data.
For example, predict future sales from past ones.

> Advanced tasks

2

◆ Simple ML for Sheets 실습 (3)

❖ 예측 결과 확인

A	B	C	D	E	F	G	H	I	J
island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year	species	Pred:species	Pred:Conf.species
Biscoe	47.8	15	215	5650	male	2007		Gentoo	99.23 %
Dream	40.2	17.1	193	3400	female	2009		Adelie	99.06 %
Dream	36	17.8	195	3450	female	2009		Adelie	99.25 %
Biscoe	49.8	15.9	229	5950	male	2009		Gentoo	99.23 %
Biscoe	38.6	17.2	199	3750	female	2009		Adelie	99.25 %
Dream	49	19.5	210	3950	male	2008		Chinstrap	99.20 %
Dream	40.7	17	190	3725	male	2009		Adelie	99.10 %
Biscoe	52.5	15.6	221	5450	male	2009		Gentoo	99.23 %
Biscoe	46.2	14.4	214	4650		2008		Gentoo	99.23 %
Torgersen	40.2	17	176	3450	female	2009		Adelie	99.14 %
Biscoe	46.5	14.5	213	4400	female	2007		Gentoo	99.25 %
Biscoe	49.5	16.2	229	5800	male	2008		Gentoo	99.23 %
Torgersen	36.2	16.1	187	3550	female	2008		Adelie	98.69 %
Biscoe	41.3	21.1	195	4400	male	2008		Adelie	99.18 %
Biscoe	45.1	14.5	207	5050	female	2007		Gentoo	99.26 %
Biscoe	47.5	15	218	4950	female	2009		Gentoo	99.25 %
Biscoe	49.1	15	228	5500	male	2009		Gentoo	99.23 %



속제

이번 강의 속제를 알아보자

◆ 3장 숙제

- ❖ 앞서 스프레드시트 기반 **ML** 실행해보고 스크린샷을 슬랙 **DM**으로 제출할 것
- ❖ 아래 퀴즈 풀어볼 것
 - <https://forms.gle/5gcsWvPwWW1ENXTY8>



Q & A

이번 강의에 질문이 있으면 알려주세요!