INNOVATE

AI/ML EDITION

24 February 2022

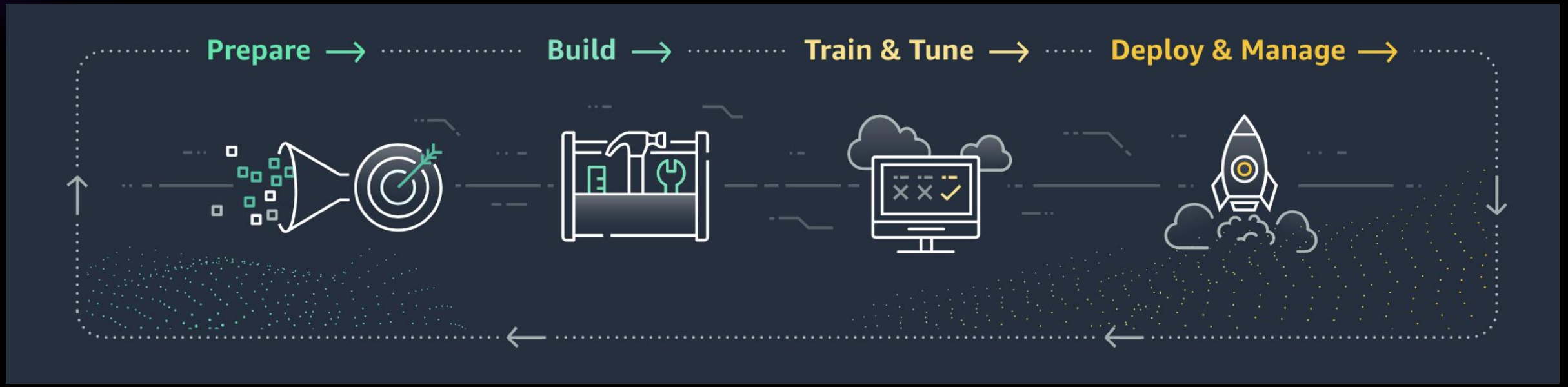
ML 데이터 준비 및 ML Workflow 프로토타이핑 배워보기

문곤수
AI/ML 전문 솔루션즈 아키텍트
AWS

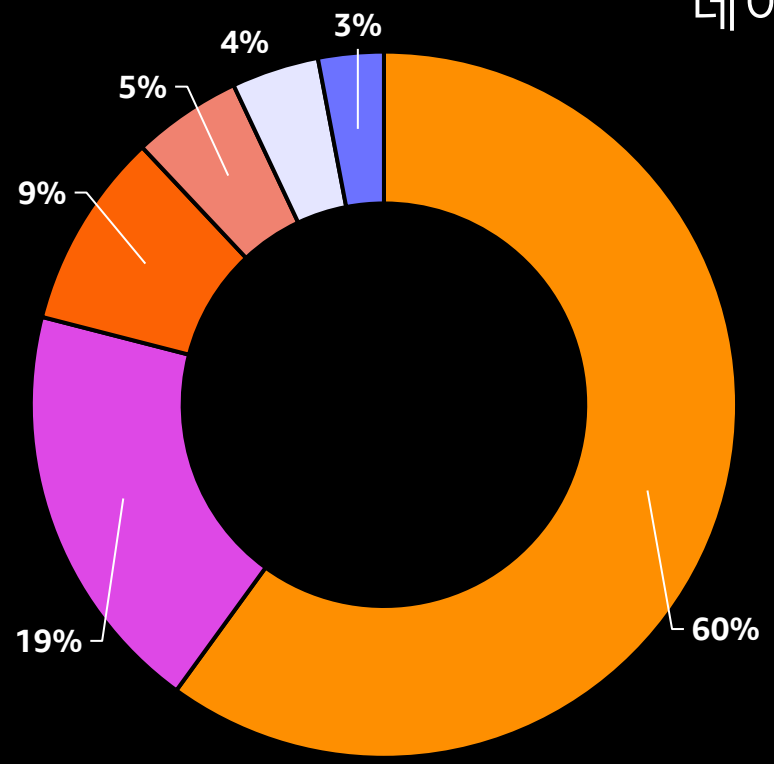
Agenda

- 머신 러닝 개요 및 데이터 준비의 중요성
- 예시1: Tabular (CSV 포맷) 데이터 준비 및 ML Workflow 프로토타이핑
- 예시 2: 이미지 데이터 준비 및 ML Workflow 프로토타이핑

머신 러닝 프로세스



80%의 시간이 데이터 준비에 소요됨



데이터 과학자가 가장 많은 시간을 보내는 작업

- Cleaning and organizing data
- Collecting data sets
- Mining data for patterns
- Other
- Refining algorithms
- Building training sets

Source: Forbes survey of 80 data scientists, March 2016

“The model and the code for many applications are basically a solved problem,” says Ng. “Now that the models have advanced to a certain point, we got to make the data work as well.”

"많은 응용 프로그램의 모델과 코드는 기본적으로 해결된 문제입니다. 이제 모델이 특정 지점까지 발전 했으므로 데이터도 작동하도록 해야 합니다."

“If 80 percent of our work is data preparation, then ensuring data quality is the important work of a machine learning team.”

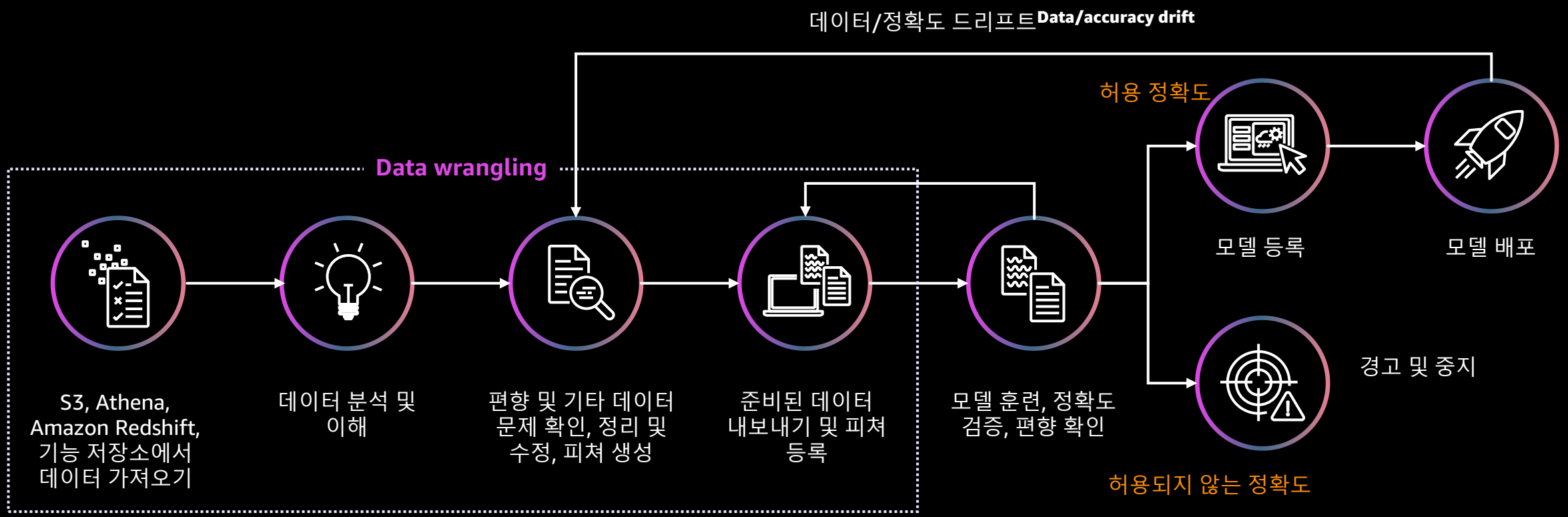
"우리 작업의 **80%**가 데이터 준비라면 데이터 품질을 보장하는 것은 머신 러닝 팀의 중요한 작업입니다."

Andrew Ng

Founder & CEO – Landing AI

Andrew Ng Launches A Campaign For Data-Centric AI (Dec 6, 2021)
https://www.forbes.com/sites/gilpress/2021/06/16/andrew-ng-launches-a-campaign-for-data-centric-ai/?sh=2dee9f3574f5
A Chat with Andrew on MLOps: From Model-centric to Data-centric AI (Mar 2021)
https://www.youtube.com/watch?v=06-AZXmwHjo

데이터 준비는 종단 간 **ML Workflow** 의 중요한 부분



AWS 머신러닝 스택

가장 깊이있고 폭넓은 머신러닝 역량 제공

AI 서비스

개발자를 위한 API 서비스

SPECIALIZED

비즈니스 프로세스
- Amazon Personalize
- Amazon Forecast
- Amazon Fraud Detector
- Amazon Lookout for Metrics

검색
- Amazon Kendra

코드 + DevOps
- Amazon CodeGuru
- Amazon DevOps Guru

제조
- Amazon Monitron
- Amazon Lookout for Equipment
- Amazon Lookout for Vision

헬스케어
- Amazon HealthLake
- Amazon Comprehend Medical
- Amazon Transcribe Medical

CORE

텍스트 & 문서
- Amazon Translate
- Amazon Comprehend
- Amazon Textract

챗봇
- Amazon Lex

음성
- Amazon Polly
- Amazon Transcribe
- Amazon Transcribe Call Analytics

비전
- Amazon Rekognition
- AWS Panorama

AMAZON SAGEMAKER

데이터 과학자를 위한 완전 관리 서비스

데이터 레이블링

SAGEMAKER CANVAS

비즈니스 분석가를 위한 코드프리 머신 러닝

SAGEMAKER STUDIO LAB

머신 러닝 학습

SAGEMAKER STUDIO IDE

데이터 준비 | 피쳐 저장 | 노트북 개발 | 모델 훈련 | 파라메터 튜닝 | 프로덕션 배포 | 관리& 모니터링

CI/CD

엣지 디바이스 관리

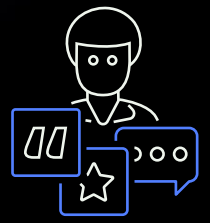
ML 프레임워크 & 인프라

데이터 과학자를 위한 셀프서비스

TensorFlow, PyTorch, Apache MXNet, **Hugging Face** | Deep learning AMIs & containers | CPUs | GPUs | AWS Inferentia | AWS Trainium | Habana Gaudi | Elastic inference | FPGA

Amazon SageMaker Notebooks

빠른 시작 공유 가능한 노트북

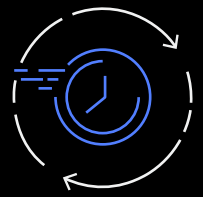


SSO(Single Sign-On)를 통한 손쉬운 액세스

몇 초 만에 노트북에 액세스

완전 관리형 및 보안

관리자는 액세스 및 권한을 관리합니다.

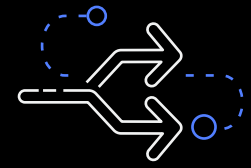


빠른 시작

컴퓨팅 리소스 준비 없이 노트북 시작

손쉬운 협업

클릭 한 번으로 노트북 공유



탄력적 리소스

컴퓨팅 리소스를 쉽게 늘리고 줄일 수 있음

Agenda

- 머신 러닝 개요 및 데이터 준비의 중요성
- 예시1: Tabular (CSV 포맷) 데이터 준비 및 ML Workflow 프로토타이핑
- 예시 2: 이미지 데이터 준비 및 ML Workflow 프로토타이핑

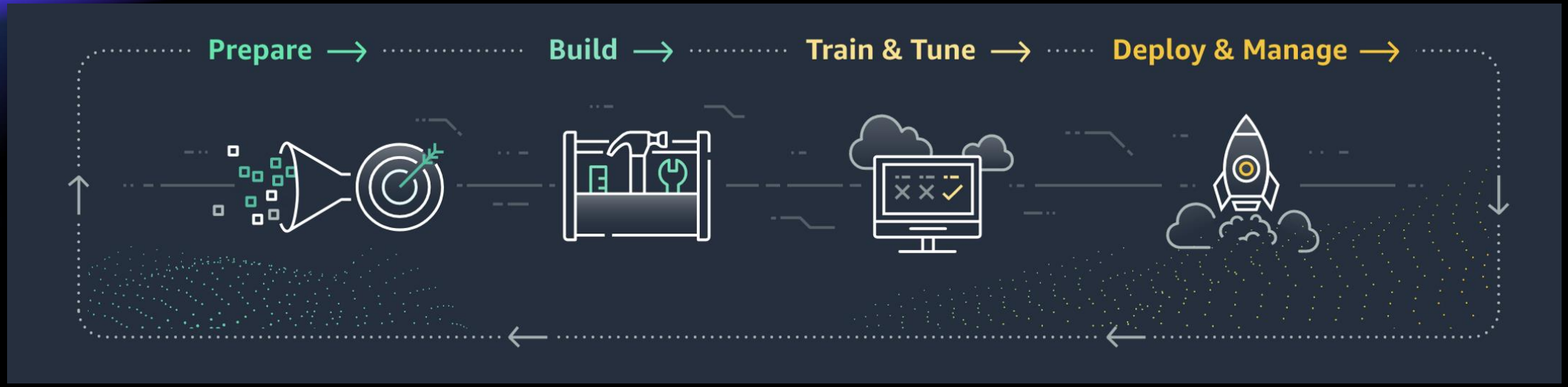
예시1: **Tabular (CSV** 포맷) 데이터 준비 및 고객 이탈 분류 프로토타이핑

Code: https://bit.ly/ml-data-prep

	Churn?	State	Account Length	Area Code	Phone	Int'l Plan	VMail Plan	VMail Message	Day Mins	Day Calls	Day Charge
0	1	PA	163	806	403-2562	no	yes	300	8.162204	3	7.579174
1	0	SC	15	836	158-8416	yes	no	0	10.018993	4	4.226289
2	0	MO	131	777	896-6253	no	yes	300	4.708490	3	4.768160
3	0	WY	75	878	817-5729	yes	yes	700	1.268734	3	2.567642
4	1	WY	146	878	450-4942	yes	no	0	2.696177	3	5.908916

이동 통신 가입자의 “고객 이탈” 유무를 분류하는 문제 임.

“고객 이탈 분류” 프로토타이핑 개요



Tabular 데이터 수집 및 준비

- 이동통신 고객이탈 데이터

데이터 탐색

- 히스토그램
- 상관관계 분석

데이터 전처리

피쳐 선택

알고리즘 선택

- XGBoost
- AutoGluon

모델 훈련 코드 작성

모델 훈련

- SageMaker XGBoost
- AutoGluon 앙상블
- 모델 설명 (SHAP)

모델 배포 및 추론

- Local Inference
- 모델 평가 (AUC-ROC)

(1) 데이터 수집 및 준비

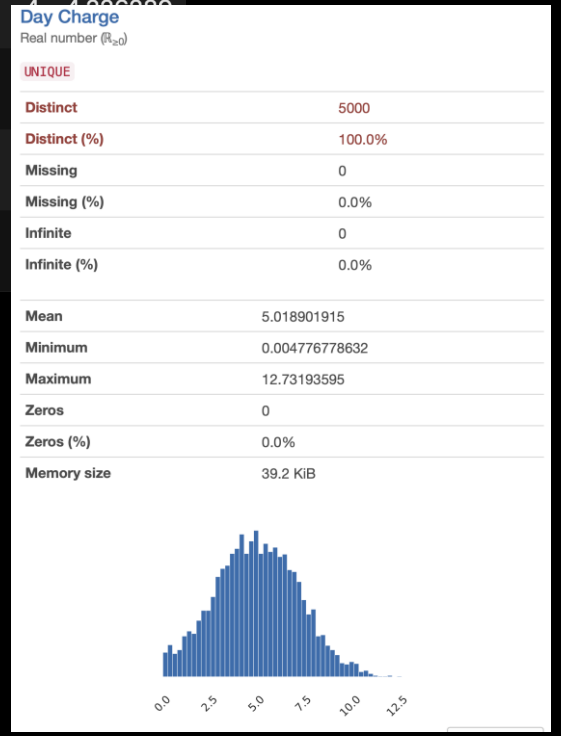
"Pandas-Profiling 패키지" 이용하여 데이터 탐색

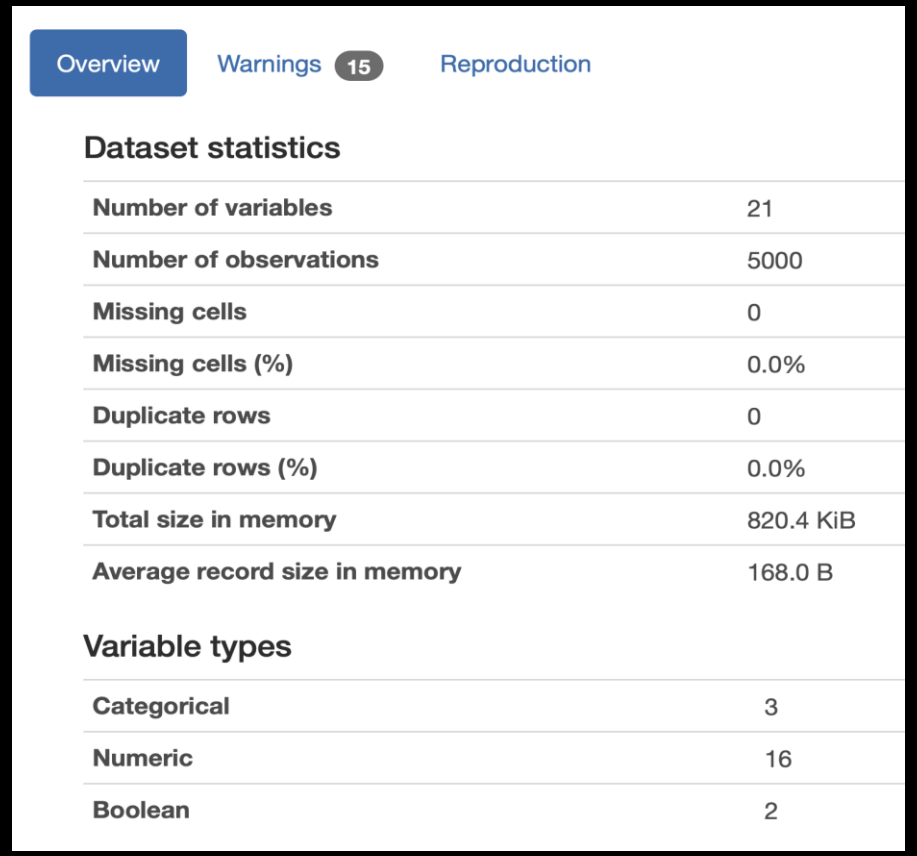
Tabular 데이터 수집 및 준비 → 데이터 탐색

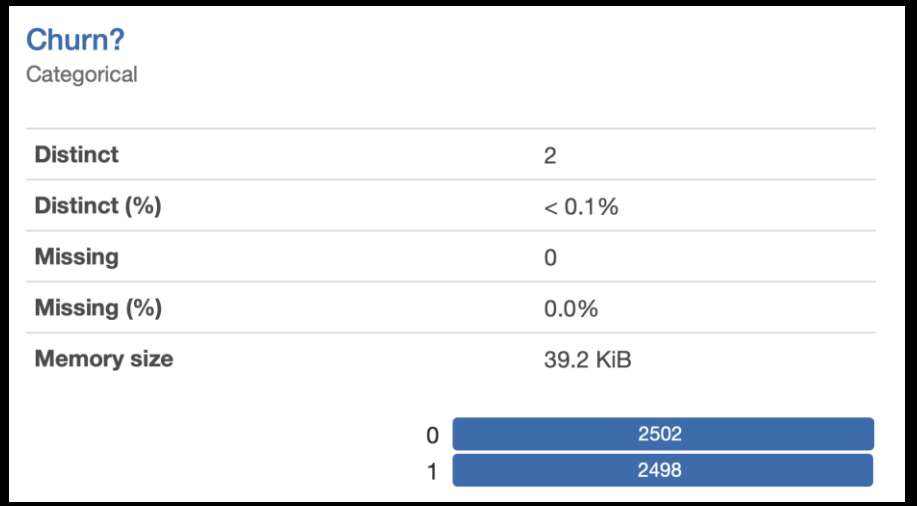
- 이동통신 고객이탈 데이터

- 히스토그램
- 상관관계 분석

	Churn?	State	Account Length	Area Code	Phone	Int'l Plan	VMail Plan	VMail Message	Day Mins	Day Calls	Day Charge
0	1	PA	163	806	403-2562	no	yes	300	8.162204	3	7.579174
1	0	SC	15	836	158-8416	yes	no	0	10.018993		
2	0	MO	131	777	896-6253	no	yes	300	4.708490		
3	0	WY	75	878	817-5729	yes	yes	700	1.268734		
4	1	WY	146	878	450-4942	yes	no	0	2.696177		







(2) 모델 훈련 준비

데이터 전처리 → 피쳐 선택 → 알고리즘 선택 → 모델 훈련 코드 작성

데이터 프로파일링 인사이트

Overview | Warnings 15 | Reproduction

Warnings

`State` has a high cardinality: 51 distinct values	High cardinality
`Phone` has a high cardinality: 4999 distinct values	High cardinality
`Phone` is uniformly distributed	Uniform
`Day Mins` has unique values	Unique
`Day Charge` has unique values	Unique
`Eve Mins` has unique values	Unique
`Eve Charge` has unique values	Unique
`Night Mins` has unique values	Unique
`Night Charge` has unique values	Unique
`Intl Mins` has unique values	Unique
`Intl Charge` has unique values	Unique
`VMail Message` has 2549 (51.0%) zeros	Zeros
`Day Calls` has 170 (3.4%) zeros	Zeros
`Eve Calls` has 800 (16.0%) zeros	Zeros
`Night Calls` has 73 (1.5%) zeros	Zeros

상관계수 분석

	Churn?	Account Length	Area Code	VMail Message	Day Mins	Day Calls	Day Charge	Eve Mins	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls	Intl Charge	CustServ Calls
Churn?	nan	0.00	0.01	0.19	0.58	0.00	0.44	0.43	0.22	0.46	0.02	0.51	0.57	0.00	0.01	0.11	0.02
Account Length	nan	nan	0.04	0.01	0.02	0.01	0.01	0.00	0.03	0.01	0.02	0.00	0.03	0.02	0.00	0.03	0.04
Area Code	nan	nan	nan	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.02	0.01	0.01	0.01
VMail Message	nan	nan	nan	nan	0.14	0.00	0.18	0.10	0.10	0.03	0.06	0.14	0.16	0.02	0.13	0.01	0.07
Day Mins	nan	nan	nan	nan	nan	0.09	0.67	0.48	0.18	0.77	0.19	0.45	0.57	0.00	0.24	0.24	0.20
Day Calls	nan	nan	nan	nan	nan	nan	0.22	0.03	0.19	0.05	0.09	0.08	0.05	0.02	0.05	0.12	0.07

Tabular Data (CSV 데이터) 피쳐 선택 기본 가이드

https://github.com/gonsoomoon-ml/Self-Study-On-SageMaker/blob/main/data_preparation/Feature_Selection_Guide.md

피쳐 제거

- State:
  이유: "High Cardinality"
- Area Code, Phone
  이유: "High Cardinality"
- Eve Charge
  이유: Day Mins와 높은 상관관계

(3) XGBoost 모델 훈련 및 모델 설명

모델 훈련 → 모델 설명

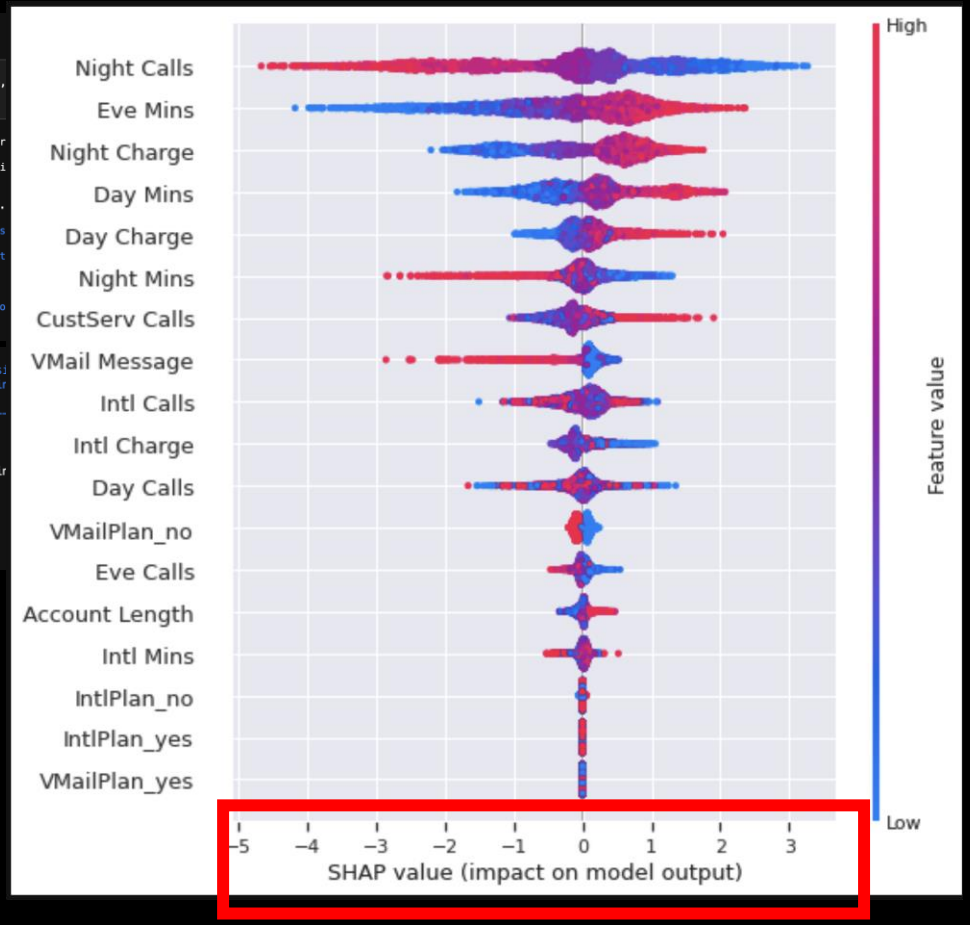
SageMaker XGBoost

피쳐 중요도 (SHAP 분석)

SageMaker XGBoost 모델 훈련

```
%%time
xgb_estimator.fit({'training':training, 'validation':validation},
                  wait=True,
                  )
2022-01-06 01:02:56 Starting - Starting the training job...
2022-01-06 01:03:00 Starting - Launching requested ML instancesPr
.....
2022-01-06 01:04:11 Starting - Preparing the instances for traini
2022-01-06 01:06:26 Downloading - Downloading input data
2022-01-06 01:06:26 Training - Downloading the training image...
2022-01-06 01:06:51 Training - Training image download completed.
emaker_xgboost_container.training
INFO:sagemaker-containers:No GPUs detected (normal if no gpus ins
INFO:sagemaker_xgboost_container.training:Invoking user training
INFO:sagemaker-containers:Module train_xgb does not provide a set
Generating setup.py
INFO:sagemaker-containers:Generating setup.cfg
INFO:sagemaker-containers:Generating MANIFEST.in
INFO:sagemaker-containers:Installing module with the following co
/miniconda3/bin/python3 -m pip install .
Processing /opt/ml/code
  SM_HP_LABEL_COLUMN=Churn?
  PYTHONPATH=/miniconda3/bin:/:/miniconda3/lib/python3.6/si
  /lib/python3.6:/miniconda3/lib/python3.6/lib-dynload:/mir
  Invoking script with the following command:
  /miniconda3/bin/python3 -m train_xgb --eval-metric auc --
  XGBoost 0.90
  AUC  0.938

2022-01-06 01:07:11 Uploading - Uploading generated trair
2022-01-06 01:07:11 Completed - Training job completed
Training seconds: 69
Billable seconds: 69
CPU times: user 621 ms, sys: 52.3 ms, total: 673 ms
Wall time: 4min 45s
```



피쳐 중요도 상세 분석

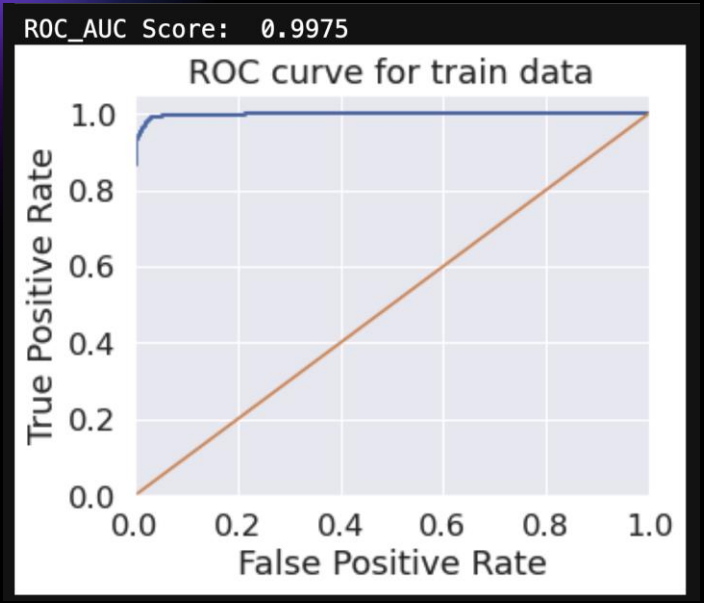
- SHAP Value 0 을 중심으로 양의 방향이면 고객 이탈 영향 (레이블 값이 1)
  - Night Calls 가 작고
  - Eve Mins 가 많고
  - Night Charge 가 많고
  - Day Mins 가 많고
  - Day Charge 가 많고
  - Night Mins 가 적고
  - Customer Calls 가 많고
  - VMail Messages 숫자가 적습니다.

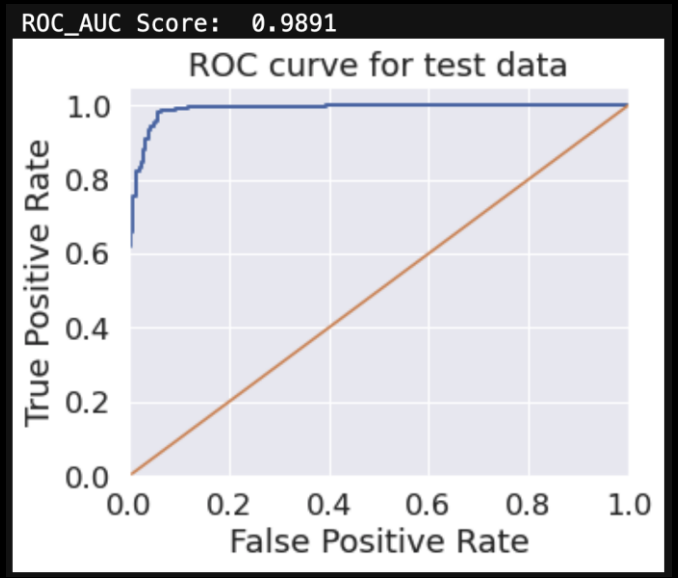
- SHAP Value 0 을 중심으로 음의 방향이면 고객 유지 영향 (레이블 값이 0)
  - Night Calls 가 많고
  - Eve Mins 가 적고
  - Night Charge 가 적고
  - Day Mins 가 적고
  - Day Charge 가 적고
  - Night Mins 가 많고
  - Customer Calls 가 적고
  - VMail Messages 가 많습니다.

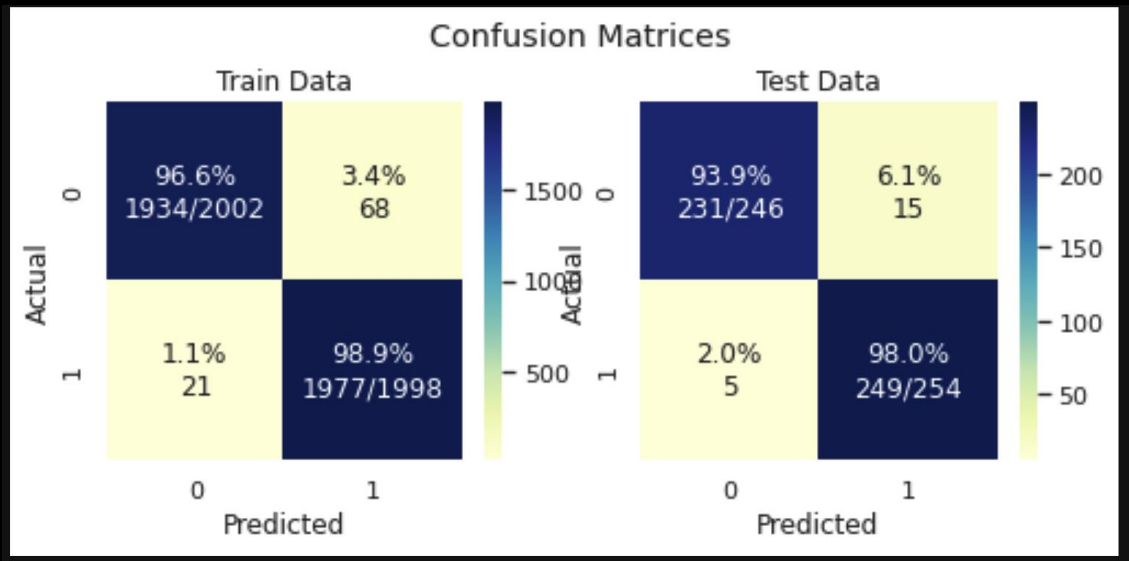
(4) XGBoost 모델 추론 및 평가

모델 배포 및 추론

- Local Inference







훈련 및 테스트 데이터 추론 분석

- threshold=0.5 (레이블의 Fraud 확률)을 기준으로 0, 1을 구분
- False Postive (FP)
  - 훈련: 3.4%
  - 테스트: 6.1%
- False Negative (FN)
  - 훈련: 1.1%
  - 테스트: 2.0%
- 위의 훈련 과 테스트의 "차이"는 과적합의 정도를 나타냄.

(5) AutoGluon 훈련, 추론 및 평가 준비

모델 훈련 → 모델 배포 및 추론

- AutoGluon 앙상블
- 추론 평가

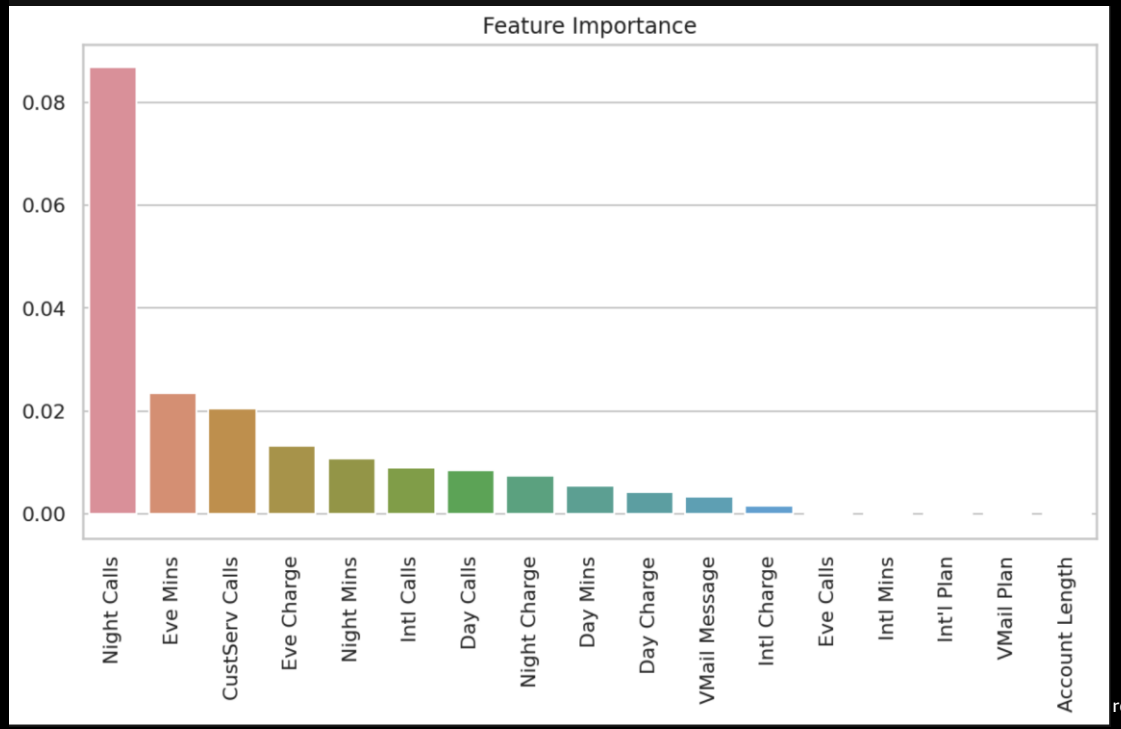
```
Fitting model: LightGBMXT ...
'verbose_eval' argument is deprecated and will be removed in
rgument instead.
        0.9916   = Validation score   (roc_auc)
        1.78s    = Training   runtime
        0.02s    = Validation runtime
Fitting model: LightGBM ...
'verbose_eval' argument is deprecated and will be removed in
rgument instead.
        0.9916   = Validation score   (roc_auc)
        0.99s    = Training   runtime
        0.01s    = Validation runtime
Fitting model: CatBoost ...
        0.9925   = Validation score   (roc_auc)
        4.35s    = Training   runtime
        0.0s     = Validation runtime
Fitting model: ExtraTreesGini ...
        0.9864   = Validation score   (roc_auc)
        0.81s    = Training   runtime
        0.1s     = Validation runtime
Fitting model: ExtraTreesEntr ...
        0.9849   = Validation score   (roc_auc)
        0.81s    = Training   runtime
        0.1s     = Validation runtime
Fitting model: XGBoost ...
        0.9916   = Validation score   (roc_auc)
        1.4s     = Training   runtime
        0.01s    = Validation runtime
Fitting model: LightGBMLarge ...
'verbose_eval' argument is deprecated and will be removed in
rgument instead.
        0.9879   = Validation score   (roc_auc)
        1.8s     = Training   runtime
Fitting model: WeightedEnsemble_L2 ...
        0.9933   = Validation score   (roc_auc)
        0.0s     = Validation runtime
AutoGluon training complete, total runtime = 13.87s ...
TabularPredictor saved. To load, use: predictor = TabularPre
```

최종 앙상블 모델

리더 보드 생성

```
predictor.leaderboard(test_data,  extra_info=False, silent=True)
```

	model	score_test	score_val	pred_time_test	pred_time_val	fit_time
0	WeightedEnsemble_L2	0.993790	0.993326	0.100449	0.022293	7.499869
1	XGBoost	0.993342	0.991582	0.074782	0.009923	1.399325
2	LightGBM	0.993214	0.991630	0.018969	0.008403	0.988137
3	CatBoost	0.993070	0.992526	0.003241	0.002819	4.349078
4	LightGBMXT	0.991534	0.991614	0.031216	0.015588	1.777600
5	LightGBMLarge	0.991518	0.987901	0.022254	0.006897	1.801324
6	ExtraTreesEntr	0.990918	0.984932	0.126840	0.103847	0.813927
7	ExtraTreesGini	0.990822	0.986357	0.128747	0.103795	0.814664



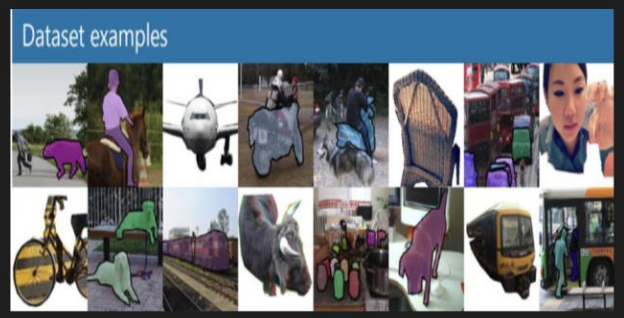
Agenda

- 머신 러닝 개요 및 데이터 준비의 중요성
- 예시1: Tabular (CSV 포맷) 데이터 준비 및 ML Workflow 프로토타이핑
- 예시 2: 이미지 데이터 준비 및 ML Workflow 프로토타이핑

예시 2: 이미지 데이터 준비 및 “이미지 분류” 프로토타이핑

Code: https://bit.ly/ml-data-prep

COCO Dataset



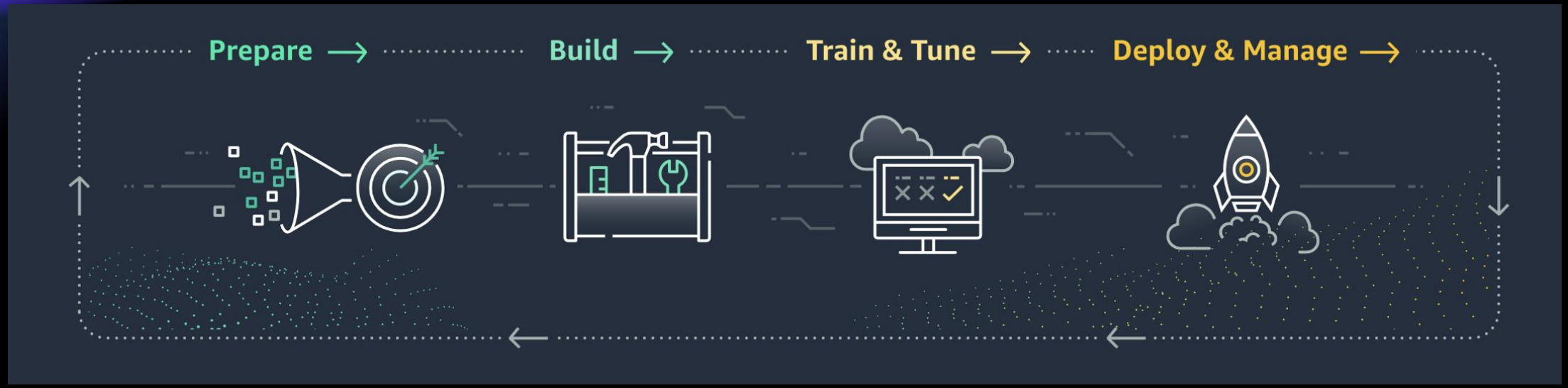
Cifar-10 Dataset

10개의 동물 카테고리 + 1 개의 동물 카테고리

→ 11개 의 동물 카테고리에 해당하는 동물 이미지 준비

동물 이미지가 주어지면 동물 이미지 분류를
11개 카테고리안에서 분류하는 문제 임.

“이미지 분류” 프로토타이핑 개요



이미지 수집 및 준비

- COCO
- CIFAR 10

프레임워크 선택

- TensorFlow
- Pytorch
- SageMaker 내장 알고리즘

이미지 전처리

- Resize
- Scale

모델 훈련 코드 작성

모델 훈련

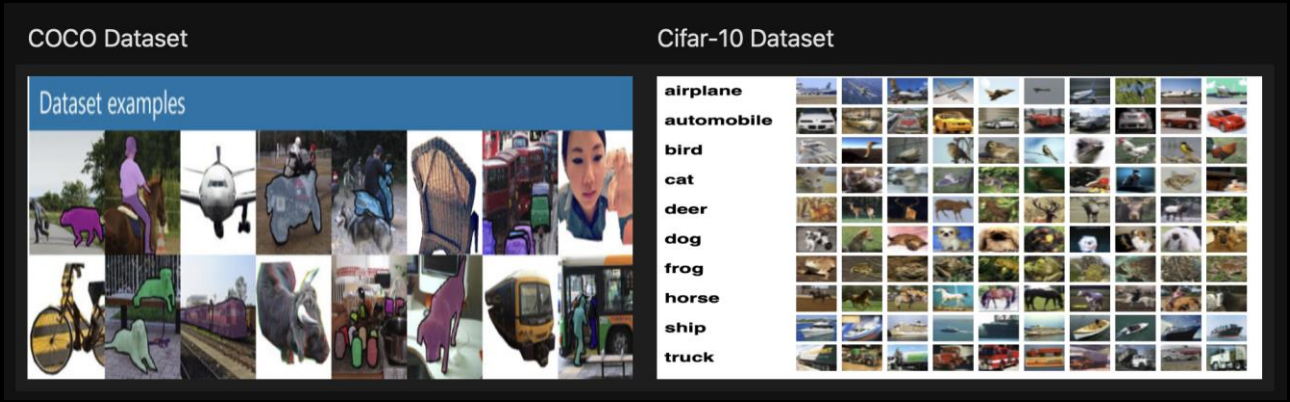
- Image 증강
- Transfer Learning

모델 배포 및 추론

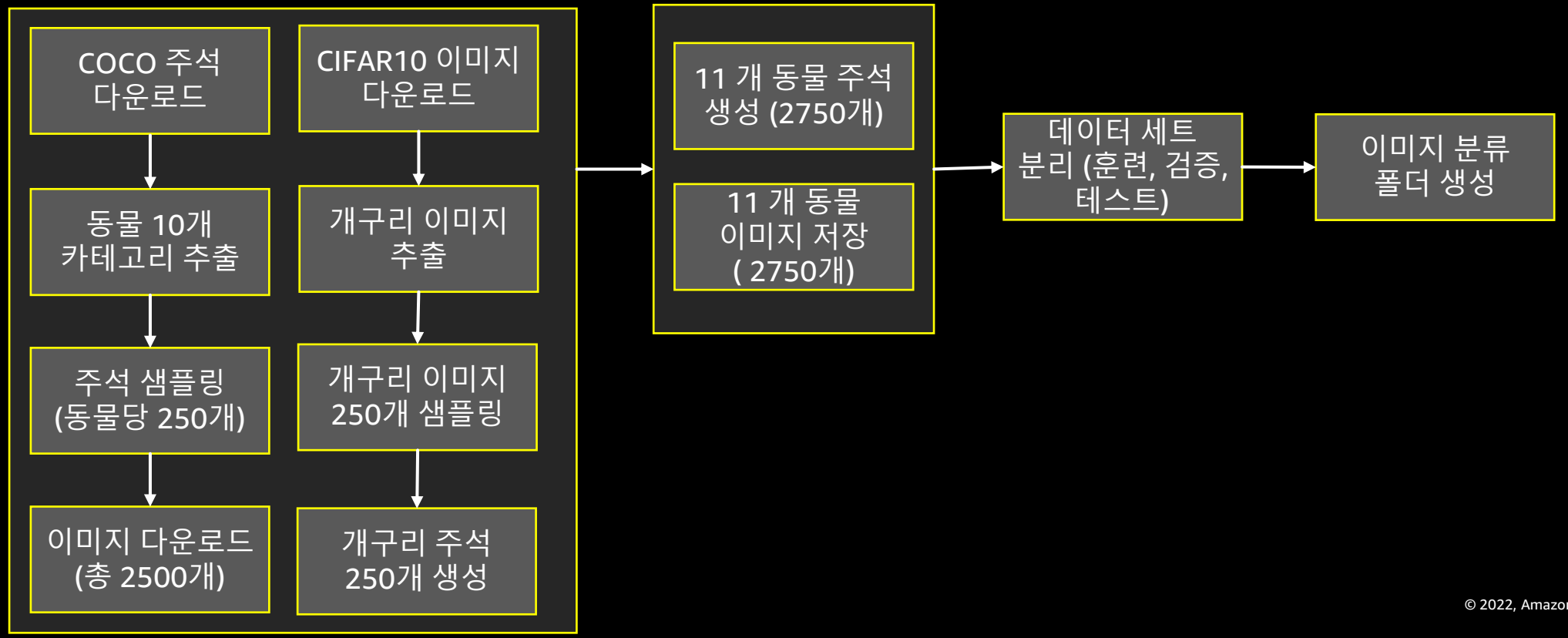
- Local Inference

(1) 이미지 수집 및 준비

사용한 이미지 데이터 세트 :



주석 (Annotation) : 이미지의 메타 정보로서 이미지 이름, 이미지 경로, 이미지 카테고리, 바운딩 박스, 세그먼테이션 정보



표준 이미지 분류 폴더: TensorFlow, PyTorch, MXNet 프로임워크가 모두 수용 가능한 폴더 및 파일 구조

```
+-- train
|   +-- class_A
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|   +-- class_B
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|
+-- val
|   +-- class_A
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|   +-- class_B
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|
+-- test
|   +-- class_A
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|   +-- class_B
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
```

(2) 텐서 플로우 : 이미지 전처리 및 모델 훈련

표준 이미지 분류 폴더:

```
+-- train
|   +-- class_A
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|   +-- class_B
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|
+-- val
|   +-- class_A
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|   +-- class_B
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|
+-- test
|   +-- class_A
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|   +-- class_B
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
```

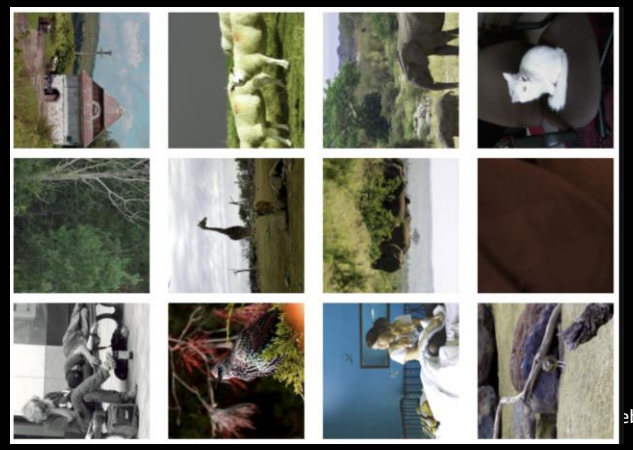
리사이즈 예시: (424, 640) → (224, 224)

추론 예시

Tensorflow → Resize → TFRecord 파일 생성

train.py 생성

모델 훈련 (Image 증강) → 모델 추론



aws

(3) 파이토치 : 이미지 전처리 및 모델 훈련

표준 이미지 분류 폴더:

```
+-- train
|   +-- class_A
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|   +-- class_B
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|
+-- val
|   +-- class_A
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|   +-- class_B
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|
+-- test
|   +-- class_A
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|   +-- class_B
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
```

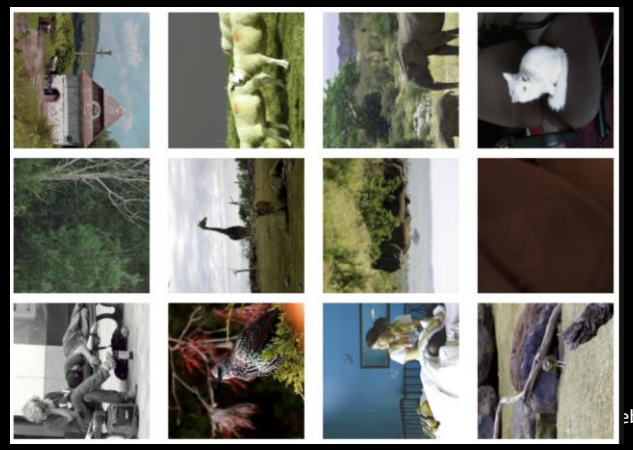
리사이즈 예시: (424, 640) → (224, 224)

추론 예시

PyTorch → Resize → Resized Image 저장

train.py 생성

모델 훈련 (Image 증강) → 모델 추론



aws

(4) SageMaker 내장 알고리즘 : 이미지 전처리 및 모델 훈련

표준 이미지 분류 폴더:

```
+-- train
|   +-- class_A
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|   +-- class_B
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|
+-- val
|   +-- class_A
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|   +-- class_B
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|
+-- test
|   +-- class_A
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
|   +-- class_B
|       +-- filename.jpg
|       +-- filename.jpg
|       +-- filename.jpg
```

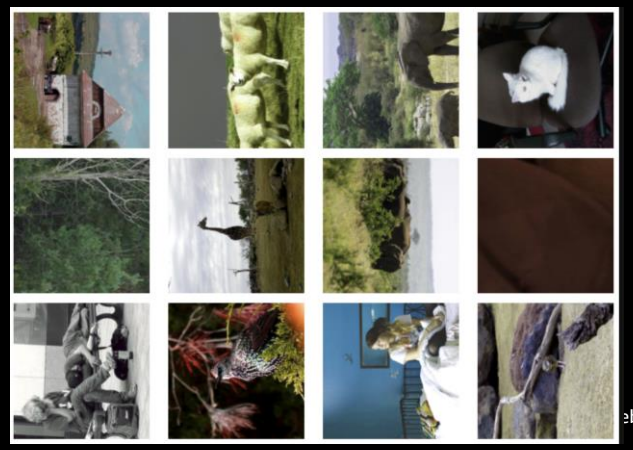
리사이즈 예시: (424, 640) → (224, 224)

추론 예시

MXNet → Resize → RecordIO 파일 생성 → 모델 훈련 → 모델 추론

훈련 코드 필요 없음

Image 증강



정리 및 요약

예시1: **Tabular (CSV** 포맷) 데이터 준비 및 고객 이탈 분류 프로토타이핑

Code: https://bit.ly/ml-data-prep

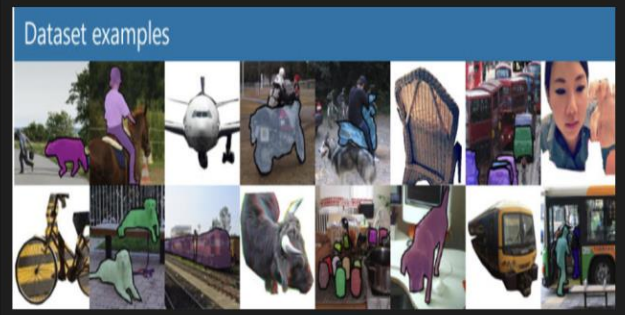
	Churn?	State	Account Length	Area Code	Phone	Int'l Plan	VMail Plan	VMail Message	Day Mins	Day Calls	Day Charge
0	1	PA	163	806	403-2562	no	yes	300	8.162204	3	7.579174
1	0	SC	15	836	158-8416	yes	no	0	10.018993	4	4.226289
2	0	MO	131	777	896-6253	no	yes	300	4.708490	3	4.768160
3	0	WY	75	878	817-5729	yes	yes	700	1.268734	3	2.567642
4	1	WY	146	878	450-4942	yes	no	0	2.696177	3	5.908916

이동 통신 가입자의 “고객 이탈” 유무를 분류하는 문제 임.

예시 **2:** 이미지 데이터 준비 및 이미지 분류 프로토타이핑

Code: https://bit.ly/ml-data-prep

COCO Dataset



Cifar-10 Dataset

10개의 동물 카테고리 + 1 개의 동물 카테고리

→ 11개 의 동물 카테고리에 해당하는 동물 이미지 준비

동물 이미지가 주어지면 동물 이미지 분류를 11개 카테고리안에서 분류하는 문제 임.

aws

데모

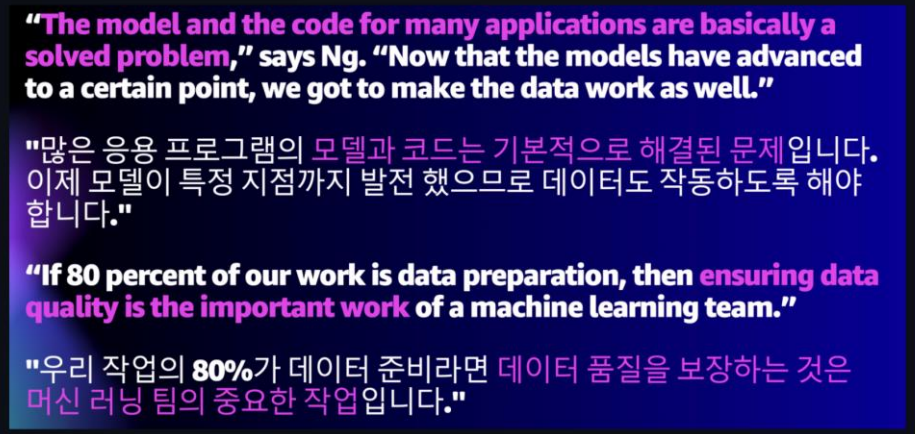
Code: https://bit.ly/ml-data-prep

ML 데이터 준비 및 ML Workflow 프로토 타이핑

1. 워크샵 배경

1.1 Andrew Ng 의 "데이터 준비" 의 중요성에 대한 의견

ML Workflow 를 개발하기 위해서는 "ML 데이터 준비" (데이터 수집, 정제, 탐색, 분석, 이해 및 정리) 를 하는 과정이 약 80% 정도를 차지 한다고 합니다. Andrew Ng 는 "From Model-Centric To Data-Centric" 으로 바꾸어야 한다고 합니다. 이유는 많은 ML 알고리즘 및 코드는 많이 발전하였고, 이미 검증이 되었다고 합니다. 하지만 "데이터 준비" 는 많이 과소평가 되고, "낮은 데이터 품질" 로 인해서 ML Workflow의 개발 속도의 저하 및 Production 시에 낮은 모델 추론 성능이 나온다고 합니다.



소스:

- Andrew Ng Launches A Campaign For Data-Centric AI (Dec 6, 2021)
  - https://www.forbes.com/sites/gilpress/2021/06/16/andrew-ng-launches-a-campaign-for-data-centric-ai/?sh=2dee9f3574f5
- A Chat with Andrew on MLOps: From Model-centric to Data-centric AI (Mar 2021)
  - https://www.youtube.com/watch?v=06-AZXmwHjo
- Data Prep Still Dominates Data Scientists' Time, Survey Finds, July 2020
  - https://www.datanami.com/2020/07/06/data-prep-still-dominates-data-scientists-time-survey-finds/
- Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, Mar 2016
  - https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=6a0651e26f63

2. 워크샵 개요

gonsoomoon-ml / ml-data-prep-workshop Public

Unwatch 1 Fork 0 Star 0

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 0 tags Go to file Add file Code

gonsoomoon-ml update cf link 4149fae 17 hours ago 14 commits

image-classificaton	edit comment	2 days ago
img	add images	4 days ago
setup	update cf link	17 hours ago
tabular	edit comments	5 days ago
LICENSE	Initial commit	12 days ago
README.md	edit recommended nb type	2 days ago

README.md

ML 데이터 준비 및 ML Workflow 프로토 타이핑

1. 워크샵 배경

1.1 Andrew Ng 의 "데이터 준비" 의 중요성에 대한 의견

ML Workflow 를 개발하기 위해서는 "ML 데이터 준비" (데이터 수집, 정제, 탐색, 분석, 이해 및 정리) 를 하는 과정이 약 80% 정도

About

No description, website, or topics provided.

Readme
MIT License
0 stars
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

Jupyter Notebook 99.9%
Python 0.1%

참고 자료

공통
- 메인 코드: ML 데이터 준비 및 ML Workflow 프로토타이핑
  - https://bit.ly/ml-data-prep

- AWS SageMaker Examples / Pred_Data
  - https://github.com/aws/amazon-sagemaker-examples/tree/master/prep_data

Tabular (CSV 형식의 데이터)
- Customer churn prediction with SageMaker XGBoost
  - https://github.com/mullue/churn-pred-xgboost

- Pandas Profiling
  - https://pandas-profiling.github.io/pandas-profiling/docs/master/rtd/

- Tablur Data (CSV 데이터) 피쳐 선택 기본 가이드
  - https://github.com/gonsoomoon-ml/Self-Study-On-SageMaker/blob/main/data_preparation/Feature_Selection_Guide.md

- AutoGluon Quick Start
  - https://github.com/mullue/autogluon

Image
- COCO Dataset
  - https://cocodataset.org

- The CIFAR-10 dataset
  - https://www.cs.toronto.edu/~kriz/cifar.html

- SageMaker built-in Image Classification Algorithm
  - https://docs.aws.amazon.com/sagemaker/latest/dg/image-classification.html

- TensorFlow Datasets 공식 페이지
  - https://www.tensorflow.org/datasets/overview

AI & ML 리소스 허브

AWS가 제공하는 AI 및 ML에 관한 다양한 자료들을 통해 더욱 심층적으로 학습해보세요!

- 기계 학습 여정 가이드
- 기계 학습의 7가지 주요 사용 사례
- 데이터, 분석 및 기계 학습을 위한 전략 플레이북
- 올바른 클라우드 서비스 및 인프라를 통한 기계 학습 혁신 가속화 전략 가이드
- 기계 학습에 적합한 컴퓨팅 인프라 선택 가이드
- 컨택트 센터의 서비스 개선 및 비용 절감 방법

- \+ 외의 다양한 동영상 학습 자료 및 기술 학습 자료!

https://bit.ly/3yUk0Kx

리소스 허브 방문하기

AWS Innovate - AI/ML 특집에 참석해주셔서 대단히 감사합니다.

저희가 준비한 강연, 어떻게 보셨나요?
더 나은 세미나를 위하여 **설문을 꼭 작성해 주세요!**


aws-korea-marketing@amazon.com


twitter.com/AWSKorea


facebook.com/amazonwebservices.ko


youtube.com/user/AWSKorea


linkedin.com/company/amazon-web-services


twitch.tv/aws

Thank you!