

9. 세이지메이커 스튜디오 I

3강. 세이지메이커 스튜디오 모델 훈련 실습

학습목표

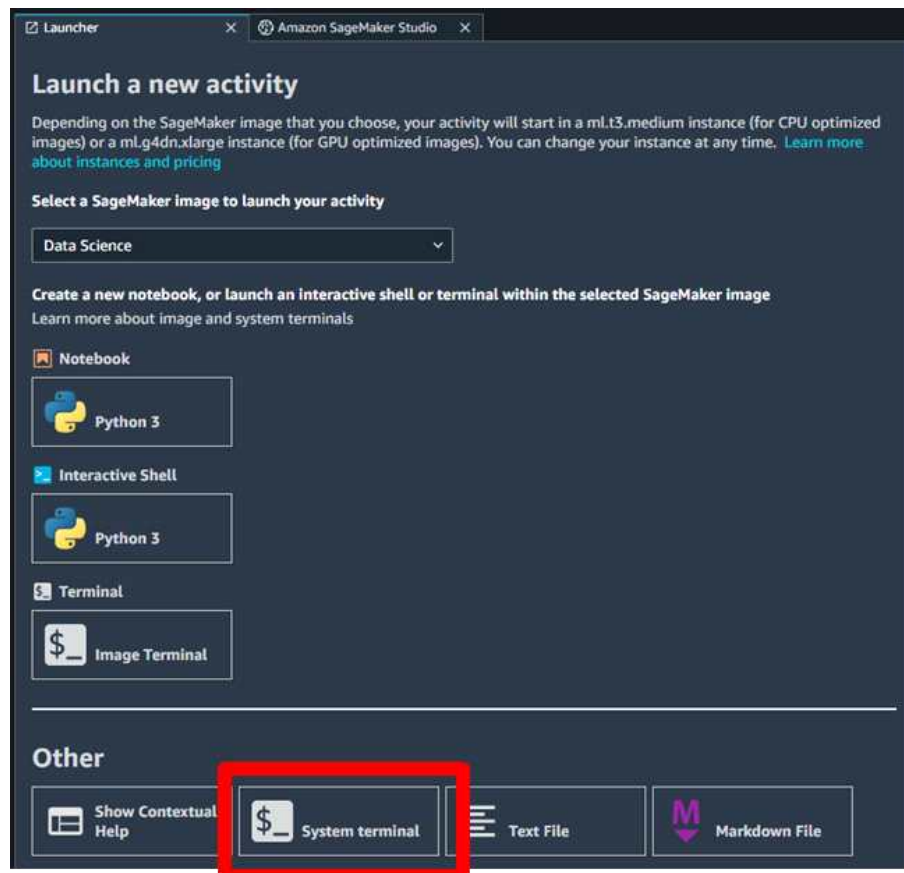
- 세이지메이커 스튜디오에서 데이터 준비 및 구축과 모델 훈련을 실행할 수 있다.

학습내용

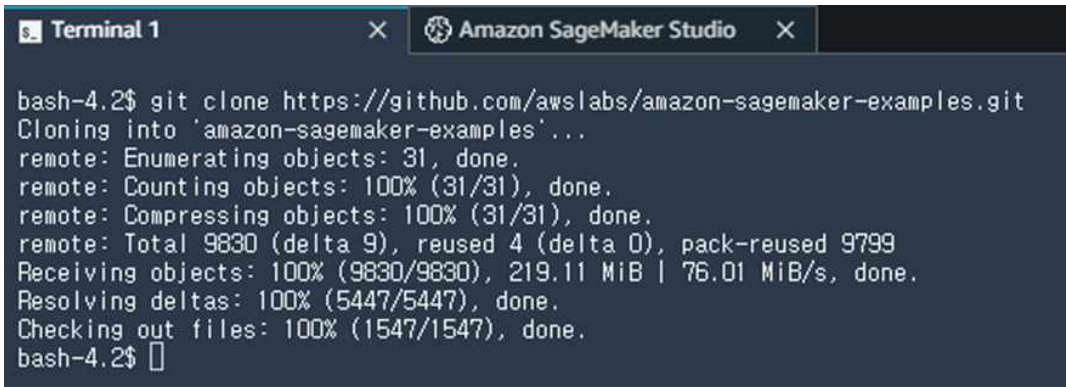
- 데이터 준비 및 구축 실습
- 모델 훈련 실습

1. 데이터 준비 및 구축 실습

- 스튜디오 실행 후 [Launcher] 탭에서 system terminal 클릭



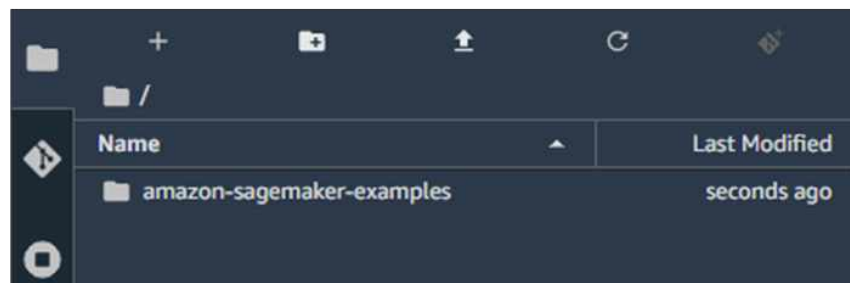
- `git clone https://github.com/aws-labs/amazon-sagemaker-examples.git`
명령어 입력
- ✓ `amazon-sagemaker-examples` 폴더 생성 완료



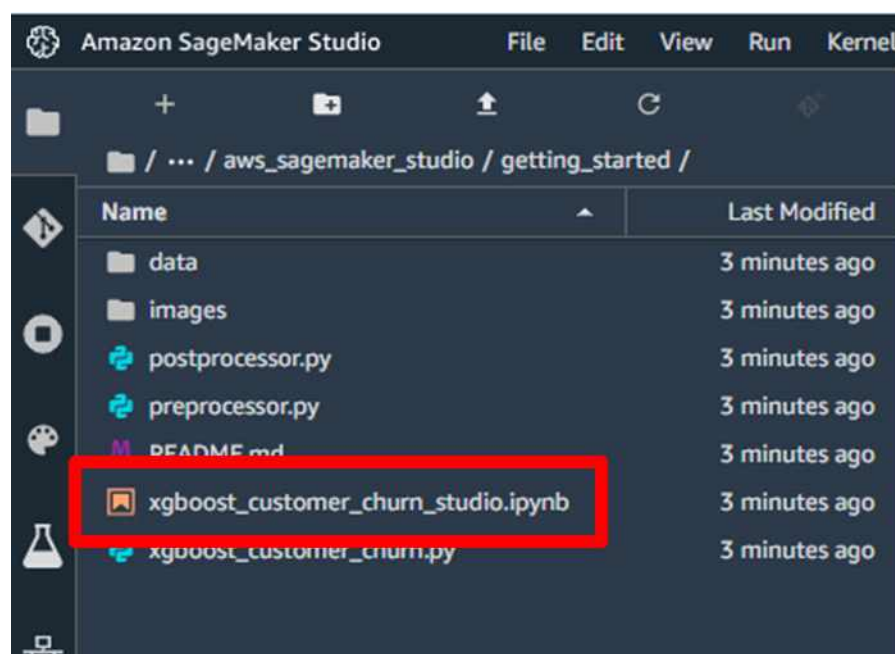
```

Terminal 1
Amazon SageMaker Studio

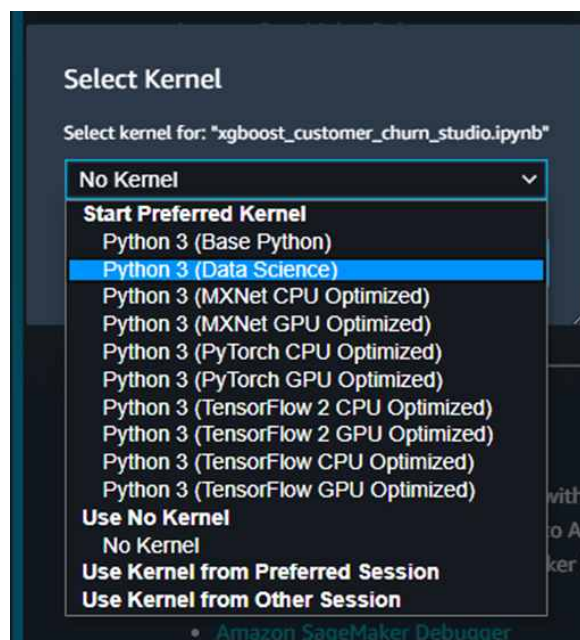
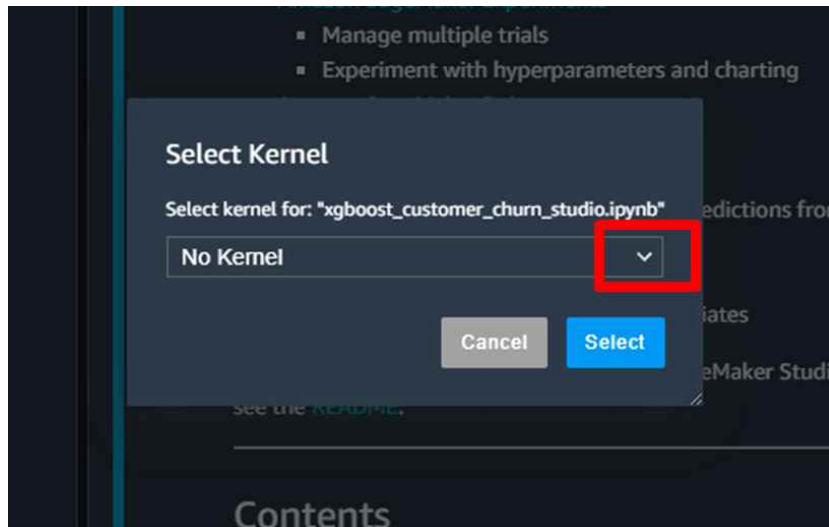
bash-4.2$ git clone https://github.com/aws-labs/amazon-sagemaker-examples.git
Cloning into 'amazon-sagemaker-examples'...
remote: Enumerating objects: 31, done.
remote: Counting objects: 100% (31/31), done.
remote: Compressing objects: 100% (31/31), done.
remote: Total 9830 (delta 9), reused 4 (delta 0), pack-reused 9799
Receiving objects: 100% (9830/9830), 219.11 MiB | 76.01 MiB/s, done.
Resolving deltas: 100% (5447/5447), done.
Checking out files: 100% (1547/1547), done.
bash-4.2$
  
```



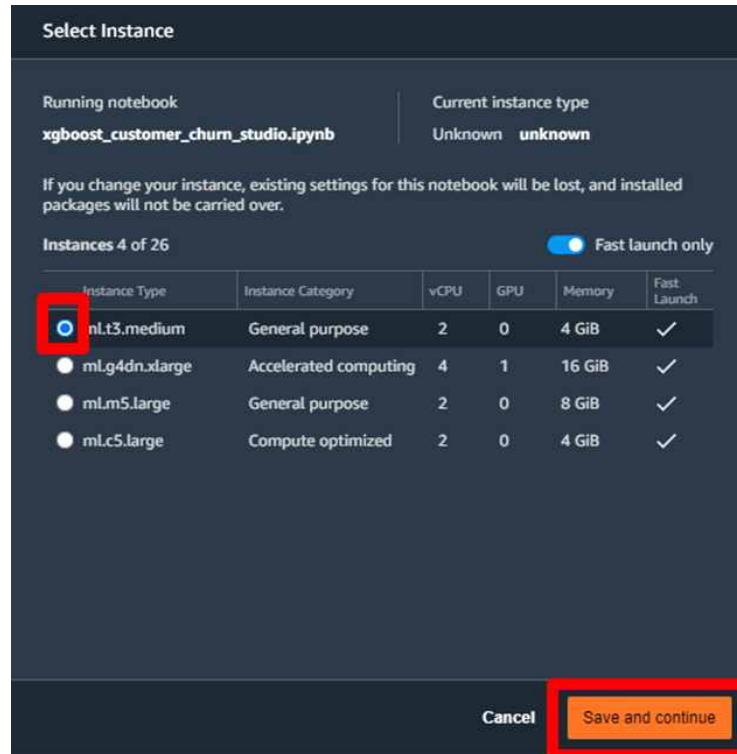
- 파일브라우저 아이콘 선택 → 준비된 파일 예제 코드 열기
- ✓ `amazon-sagemaker-examples/aws_sagemaker_studio_getting_started/xgboost_customer_churn_studio.ipynb`



- 커널 선택 - python3(data science)



- 인스턴스 선택 - python3(data science)
 - ✓ ml.t3.medium 선택 후 [save and continue] 클릭



2. 모델 훈련 실습

- 필요한 python 라이브러리를 가져옴

```
[ ]: import sys
      !{sys.executable} m pip install sagemaker -U
      !{sys.executable} m pip install sagemaker experiments

[ ]: import pandas as pd
      import numpy as np
      import matplotlib.pyplot as plt
      import io
      import os
      import sys
      import time
      import json
      from IPython.display import display
      from time import strftime, gmtime
      import boto3
      import re

      import sagemaker
      from sagemaker import get_execution_role
      from sagemaker.predictor import csv_serializer
      from sagemaker.debugger import rule_configs, Rule, DebuggerHookConfig
      from sagemaker.model_monitor import DataCaptureConfig, DatasetFormat, DefaultModelI
      from sagemaker.s3 import S3Uploader, S3Downloader

      from smexperiments.experiment import Experiment
      from smexperiments.trial import Trial
      from smexperiments.trial_component import TrialComponent
      from smexperiments.tracker import Tracker

[ ]: sess = boto3.Session()
      sm = sess.client('sagemaker')
      role = sagemaker.get_execution_role()
```

- pandas를 사용하여 헤더가 있는 교육 데이터 버전의 일부 데이터를 로드

```
# Set the path we can find the data files that go with this notebook
%cd /root/amazon-sagemaker-examples/aws_sagemaker_studio/getting_started
local_data_path = '../data/training-dataset-with-header.csv'
data = pd.read_csv(local_data_path)
pd.set_option('display.max_columns', 500) # Make sure we can see all of the columns
pd.set_option('display.max_rows', 10) # Keep the output on one page
data
```

- 훈련을 위해 파일을 S3에 업로드 하지만 먼저 데이터에 대한 S3 버킷이 없는 경우 이를 생성

```
account_id = sess.client('sts', region_name=sess.region_name).get_caller_identity().get('AccountID')
bucket = 'sagemaker-studio-{}-{}'.format(sess.region_name, account_id)
prefix = 'xgboost-churn'

try:
    if sess.region_name == "us-east-1":
        sess.client('s3').create_bucket(Bucket=bucket)
    else:
        sess.client('s3').create_bucket(Bucket=bucket,
                                         CreateBucketConfiguration={'LocationConstraint': sess.region_name})
except Exception as e:
    print("Looks like you already have a bucket of this name. That's good. Uploading files.")

# Return the URLs of the uploaded file, so they can be reviewed or used elsewhere
s3url = S3Uploader.upload('data/train.csv', 's3://{}{}/{}'.format(bucket, prefix, 'train.csv'))
print(s3url)
s3url = S3Uploader.upload('data/validation.csv', 's3://{}{}/{}'.format(bucket, prefix, 'validation.csv'))
print(s3url)
```

Amazon S3 > sagemaker-studio-us-east-1-402345010737 > xgboost-churn

sagemaker-studio-us-east-1-402345010737

개요

🔍 검색하려면 접두사를 입력하고 Enter 키를 누릅니다. 지우려면 Esc 키를 누릅니다.

📁 업로드 + 폴더 만들기 다운로드 작업 ▼ 미국 동부(버지니아 북부) ↻

<input type="checkbox"/> 이름 ▼	마지막 수정 ▼	크기 ▼	스토리지 클래스 ▼
<input type="checkbox"/> 📁 train	--	--	--
<input type="checkbox"/> 📁 validation	--	--	--

보기 1 대상 2

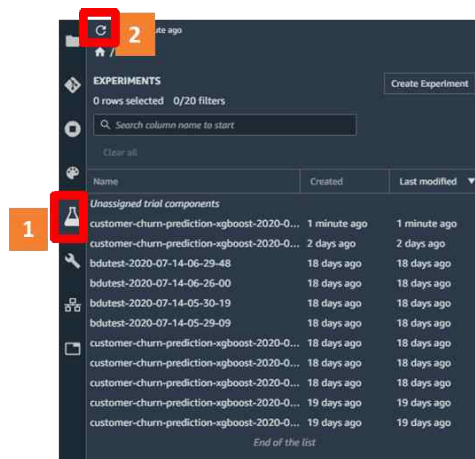
- XGBoost 라이브러리를 사용하여 방금 업로드 한 데이터에서 그라디언트 강화 의사 결정 트리라고하는 모델 클래스를 학습 진행
- XGBoost를 사용하고 있으므로 먼저 XGBoost 알고리즘 컨테이너의 위치를 지정

```
from sagemaker.amazon.amazon_estimator import get_image_uri
docker_image_name = get_image_uri(boto3.Session().region_name, 'xgboost', repo_version='1.0.0')
```

- S3의 파일에 대한 포인터로 사용할 수 있는 s3_inputs을 생성

```
s3_input_train = sagemaker.s3_input(s3_data='s3://{}/train/'.format(bucket, prefix))
s3_input_validation = sagemaker.s3_input(s3_data='s3://{}/validation/'.format(bucket, prefix))
```

- 실험 생성
 - ✓ 왼쪽 사이드바에서 실험 목록 아이콘 클릭
 - ✓ 실험 목록 새로고침 클릭
 - ✓ 실험 목록에서 실험 확인



	Name	Created	Experiment type	Tags	Last modified
3	customer-churn-prediction-xgboost-2020-08-01-08-02-22	4 minutes ago			4 minutes ago
	customer-churn-prediction-xgboost-2020-08-01-08-02-22	4 minutes ago			4 minutes ago
	bdutest-2020-07-14-06-29-48	18 days ago			18 days ago
	bdutest-2020-07-14-06-26-00	18 days ago			18 days ago
	bdutest-2020-07-14-05-30-19	18 days ago			18 days ago
	bdutest-2020-07-14-05-29-09	18 days ago			18 days ago
	customer-churn-prediction-xgboost-2020-07-14-05-28-13	18 days ago			18 days ago
	customer-churn-prediction-xgboost-2020-07-14-05-11-54	18 days ago			18 days ago
	customer-churn-prediction-xgboost-2020-07-14-05-10-58	18 days ago			18 days ago
	customer-churn-prediction-xgboost-2020-07-13-17-51-28	19 days ago			19 days ago
	customer-churn-prediction-xgboost-2020-07-13-13-57-36	19 days ago			19 days ago

- 모델 훈련 코드 실행
 - ✓ XGBoost 하이퍼 매개 변수를 지정

```
hyperparams = {"max_depth":5,
               "subsample":0.8,
               "num_round":600,
               "eta":0.2,
               "gamma":4,
               "min_child_weight":6,
               "silent":0,
               "objective":'binary:logistic'}
```





- ✓ XGBoost 알고리즘을 사용하여 모델 훈련

```
trial = Trial.create(trial_name="algorithm-mode-trial-{}".format(strftime("%Y-%m-%d-%H-%M-%S", gmtime())),
                  experiment_name=customer_churn_experiment.experiment_name,
                  sagemaker_boto_client=boto3.client('sagemaker'))

xgb = sagemaker.estimator.Estimator(image_name=docker_image_name,
                                   role=role,
                                   hyperparameters=hyperparams,
                                   train_instance_count=1,
                                   train_instance_type='ml.m4.xlarge',
                                   output_path='s3://{}x/output'.format(bucket, prefix),
                                   base_job_name="demo-xgboost-customer-churn",
                                   sagemaker_session=sess)

xgb.fit({'train': s3_input_train,
        'validation': s3_input_validation},
       experiment_config={
           "ExperimentName": customer_churn_experiment.experiment_name,
           "TrialName": trial.trial_name,
           "TrialComponentDisplayName": "Training",
       })
```

- 훈련 결과 확인
 - ✓ 실험 아이콘 클릭 -> 실험명 더블 클릭

	Name	Last modified
	Unassigned trial components	
	customer-churn-prediction-...	20 minutes ago
	customer-churn-prediction-...	2 days ago
	bductest-2020-07-14-06-29...	18 days ago
	bductest-2020-07-14-06-26...	18 days ago
	bductest-2020-07-14-05-30...	18 days ago
	bductest-2020-07-14-05-29...	18 days ago

✓ 시도 더블 클릭

TRIALS
0 rows selected 0/20 filters

Search column name to start

Clear all

Name	Created	Trial type	Tags	Last modified
algorithm-m...	28 minutes ago			25 minutes ago
algorithm-m...	38 minutes ago			38 minutes ago

End of the list

less than 20 seconds ago

/ customer-churn-prediction-xgboost-2020-08-01-08-02-22 / algorithm-mode-trial-2020-08-01-08-26-45 /

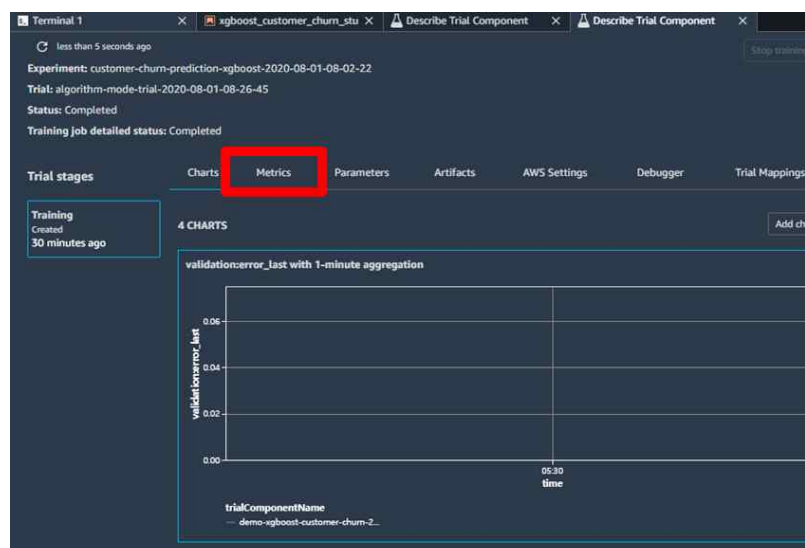
TRIAL COMPONENTS
0 rows selected 0/20 filters

Search column name to start

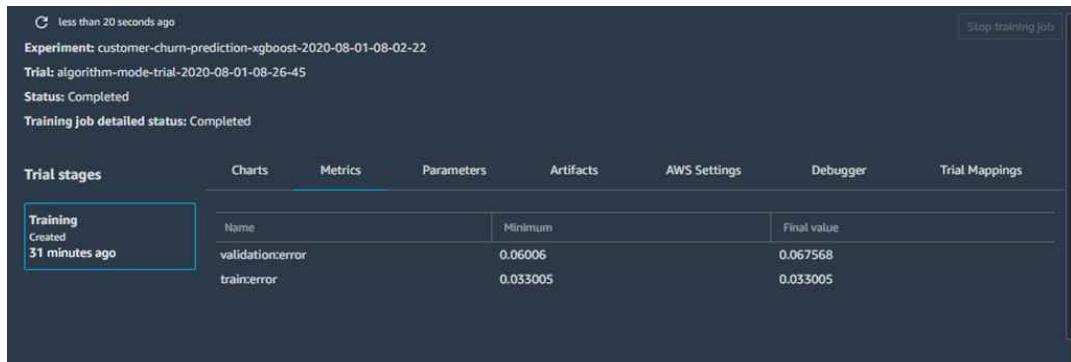
Clear all

Name	Created	Last modified
Training	29 minutes ago	25 minutes ago

End of the list



✓ 훈련 작업의 진행 상황을 확인



- 여러 시도 생성 및 비교

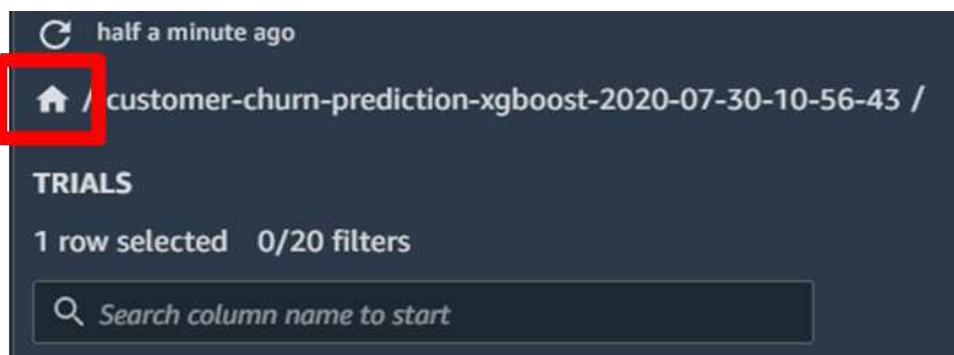
```
min_child_weights = [1, 2, 4, 8, 10]

for weight in min_child_weights:
    hyperparams["min_child_weight"] = weight
    trial = Trial.create(trial_name="algorithm-node-trial-{}-weight-{}".format(strftime("%Y-%m-%d-%H-%M-%S", gmtime()), weight),
        experiment_name=customer_churn_experiment.experiment_name,
        sagemaker_boto_client=boto3.client("sagemaker"))

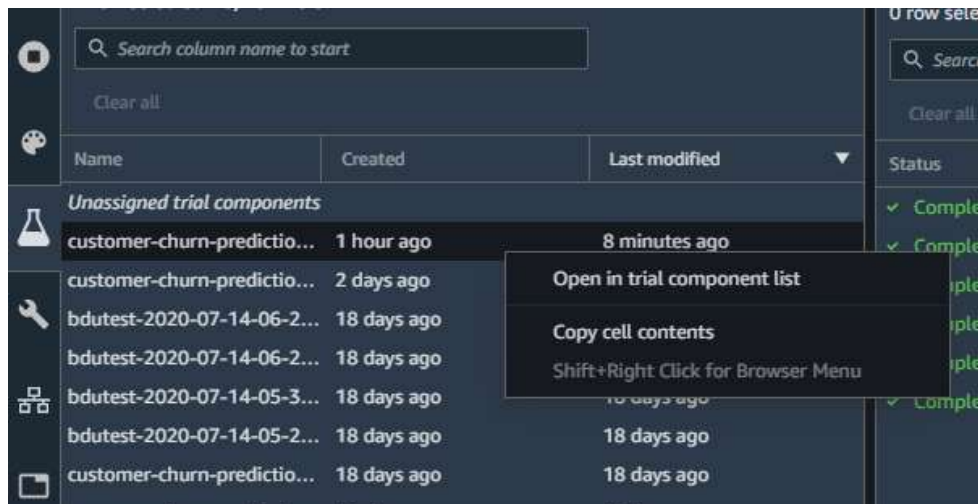
    t_xgb = sagemaker.estimator.Estimator(image_name=docker_image_name,
        role=role,
        hyperparameters=hyperparams,
        train_instance_count=1,
        train_instance_type='ml.m4.xlarge',
        output_path='s3://{}/{}/output'.format(bucket, prefix),
        base_job_name="demo-xgboost-customer-churn",
        sagemaker_session=sess)

    t_xgb.fit({'train': s3_input_train,
        'validation': s3_input_validation},
        wait=False,
        experiment_config={
            "experimentName": customer_churn_experiment.experiment_name,
            "trialName": trial.trial_name,
            "TrialComponentDisplayName": "Training",
        })
```

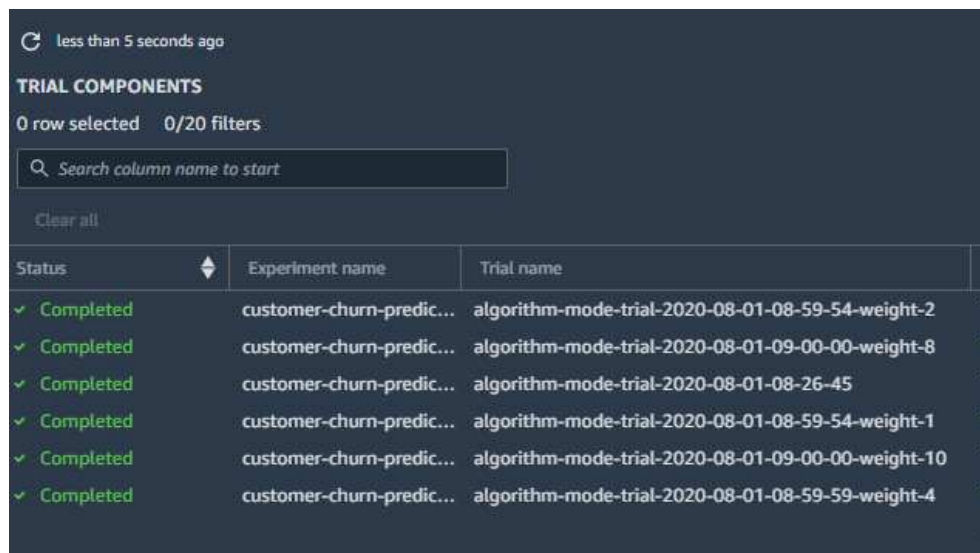
- ✓ Studio에서 진행 상황을 추적하고 결과를 보려면 시도 구성 요소 위에 있는 홈 아이콘 선택



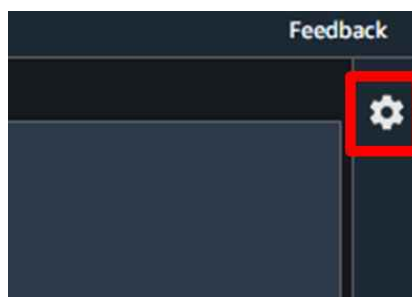
- ✓ 새로고침 클릭 후 실험명 오른쪽 클릭 후 [Open in trial component list] 클릭



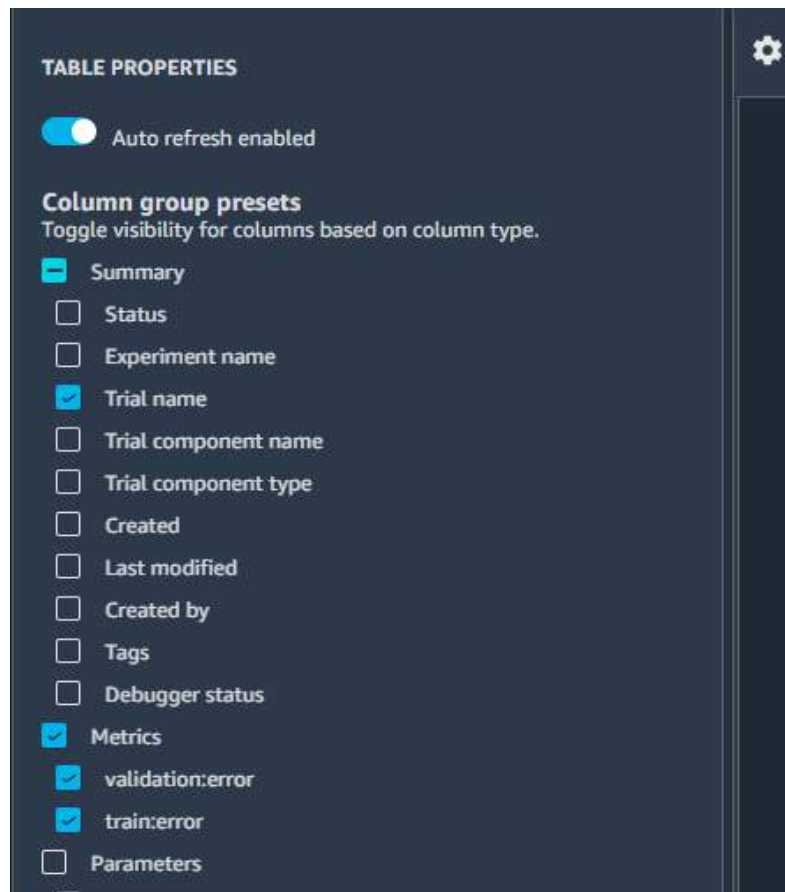
- ✓ 시도 구성 목록



- ✓ 상단 모소리에 설정 아이콘 클릭



- ✓ Summary 에서 Trial name 클릭
- ✓ Metrics 전체 클릭



- ✓ min_child_weight 하이퍼파라미터 값에 대한 validation:error, train:error 비교 가능

less than a minute ago

TRIAL COMPONENTS

0 row selected 0/20 filters

Search column name to start

Clear all

Trial name	validation:error	train:error
algorithm-mode-trial-2020-08-01-08-59-54-weight-2	0.064565	0.025289
algorithm-mode-trial-2020-08-01-09-00-00-weight-8	0.06606599688529968	0.02100300043821335
algorithm-mode-trial-2020-08-01-08-26-45	0.067568	0.033005
algorithm-mode-trial-2020-08-01-08-59-54-weight-1	0.069069	0.04072
algorithm-mode-trial-2020-08-01-08-59-59-weight-4	0.070571	0.046721
algorithm-mode-trial-2020-08-01-09-00-00-weight-10	0.066066	0.029147

End of the list

※ 나도 전문가다

- AI 서비스 사용자에게 따른 개발 환경 비교

	로컬 컴퓨터 (1+)	EC 2 (10+)	EC 2 클러스터 (100+)	SageMaker (10000+)
인프라 설정	DIY(GPU 필요)	Auto(EC2 API/CLI)	Auto(Cloud Formation)	Managed
ML SW 설정	DIY(Tensorflow 설치)	Auto(DL AMI)	Auto	Managed
모델 개발	DIY(Jupyter Notebook)	DIY	DIY	Managed(built-in Model)
모델 훈련	DIY(장시간 소요)	Auto(EC2 API/CLI)	Auto	Managed
모델 배포	DIY(직접 평가)	N/A	DIY(Model Server)	Managed
CI/CD 통합	N/A	N/A	DIY(Jenkins, Puppet)	Managed (SDK, Step Functions)
고 가용성	N/A	N/A	Auto(ELB, Auto Scaling)	Managed (On-demand training, Auto-scaling Deployment)
비용 최적화	N/A	N/A	DIY(Spot, RI, 자동화)	Managed (On-demand/Spot training, Auto-scaling Deployment)
보안	N/A	Auto(IAM, VPC, KMS)	Auto	Managed

평가하기

1. 세이지메이커 스튜디오에서 모델 훈련 작업에 대한 정보를 확인하기 위하여 선택해야 할 아이콘을 고르시오.



- 정답 : ③번

해설 : 세이지메이커 실험 목록 아이콘은 데이터 사전 처리 작업 및 훈련 작업에 대한 Charts, Metrics, Parameters, Artifacts, AWS Setting, Debugger, Trial Mappings에 대한 자세한 정보를 확인할 수 있습니다.

2. 시도(Trial) 세부 정보 중 하이퍼파라미터 및 인스턴스 정보를 확인할 수 있는 메뉴(탭)는?

① Chart

② Metrics

③ Parameters

④ Trial Mappings

- 정답 : ③번

해설 : Parameters 메뉴(탭)에서 훈련된 모델의 하이퍼파라미터 및 인스턴스의 정보를 확인할 수 있습니다.

학습정리

1. 데이터 준비 및 구축 실습

- Studio에 로그인
- 예제 소스 다운로드
- 소스파일 열기
- 커널 및 인스턴스 선택 : Python3, ml.t3.medium

2. 모델 훈련 실습

- 모델 훈련 준비
- 실험 생성
- 모델 훈련
- 훈련 결과 확인
- 여러 시도(Trial) 생성 및 비교