

## 10. 세이지메이커 스튜디오 II

### 2강. 세이지메이커 스튜디오 모델 훈련

#### 학습목표

- 세이지메이커 스튜디오에서 훈련된 모델 배포 및 최상의 모델 배포 단계를 설명할 수 있다.
- 세이지메이커 스튜디오에서 리소스 정리 단계를 설명할 수 있다.

#### 학습내용

- 모델 배포
- 최상의 모델 배포
- 리소스 정리

#### 1. 모델 배포

- 모델 배포 단계

- ① 모델 호스팅
- ② 배포된 모델 호출
- ③ 데이터 캡처

##### ✓ 모델 호스팅

- 모델을 모니터링하기 위해 엔드포인트로 전송되는 데이터를 캡처하는 구성도 추가
- 훈련된 모델을 호스팅 된 엔드포인트에 배포하기 위하여 엔드포인트 이름 설정

```
data_capture_prefix = '{} /datacapture'.format(prefix)

endpoint_name = "demo-xgboost-customer-churn-" + strftime("%Y-%m-%d-%H-%M-%S", gmti
print("EndpointName = {}".format(endpoint_name))

EndpointName = demo-xgboost-customer-churn-2020-08-01-17-57-02
```

- 훈련된 모델을 호스팅 된 엔드포인트에 배포하기 위하여 엔드포인트 인스턴스 설정

- 이 후 코드 실행 시 배포 작업 동작

```
xgb_predictor = xgb.deploy(initial_instance_count=1,
                           instance_type='ml.m4.xlarge',
                           endpoint_name=endpoint_name,
                           data_capture_config=DataCaptureConfig(enable_capture=True,
                                                                    sampling_percentage=100,
                                                                    destination_s3_uri=
                                                                    )
                           )
```

- 배포 후 엔드포인트 및 엔드포인트 구성

엔드포인트 구성			
<input type="text" value="엔드포인트 구성 검색"/>			
이름	ARN	생성 시간	
demo-xgboost-customer-churn-2020-08-01-17-57-02	arn:aws:sagemaker:us-east-1:402345010737:endpoint-config/demo-xgboost-customer-churn-2020-08-01-17-57-02	Aug 01, 2020 17:57 UTC	
엔드포인트			
<input type="text" value="엔드포인트 검색"/>			
이름	ARN	생성 시간	상태
demo-xgboost-customer-churn-2020-08-01-17-57-02	arn:aws:sagemaker:us-east-1:402345010737:endpoint/demo-xgboost-customer-churn-2020-08-01-17-57-02	Aug 01, 2020 17:57 UTC	Creating
			최종 업데이트 날짜
			Aug 01, 2020 17:57 UTC

#### ✓ 배포된 모델 호출

- 호스트 엔드포인트가 실행 중이므로 http POST 요청을 작성하여 모델에서 실시간 예측 수행가능
- test\_data NumPy 배열을 엔드 포인트 뒤의 모델로 전달하기 위해 시리얼라이저 및 디시리얼라이저를 설정

```
xgb_predictor.content_type = 'text/csv'
xgb_predictor.serializer = csv_serializer
xgb_predictor.deserializer = None
```

- XGBoost 엔드 포인트를 호출하여 테스트 데이터 세트를 반복하고 예측을 수집

```
print("Sending test traffic to the endpoint {}. \nPlease wait for a minute...".format(endpoint_name))

with open('data/test_sample.csv', 'r') as f:
    for row in f:
        payload = row.rstrip('\n')
        response = xgb_predictor.predict(data=payload)
        time.sleep(0.5)

Sending test traffic to the endpoint demo-xgboost-customer-churn-2020-08-01-09-59-29.
Please wait for a minute...
```

✓ 데이터 캡처

- 데이터를 엔드 포인트로 전송하여 실시간 예측을 할 때 모니터링 목적으로 해당 데이터를 캡처해야 함

```
from time import sleep

current_endpoint_capture_prefix = '{}/{}'.format(data_capture_prefix, endpoint_name)
for _ in range(12): # wait up to a minute to see captures in S3
    capture_files = S3Downloader.list("s3://{}{}".format(bucket, current_endpoint_capture_prefix))
    if capture_files:
        break
    sleep(5)

print("Found Data Capture Files:")
print(capture_files)
```

- 캡처된 모든 데이터는 세이지메이커 특정 json 라인 형식 파일에 저장됨
- s3에 저장된 캡처 된 데이터 확인

```
capture_file = S3Downloader.read_file(capture_files[-1])

print("====Single Data Capture====")
print(json.dumps(json.loads(capture_file.split('\n')[0]), indent=2)[:2000])
```

- json 타입의 캡처 된 데이터

```
=====Single Data Capture=====
{
  "captureData": {
    "endpointInput": {
      "observedContentType": "text/csv",
      "mode": "INPUT",
      "data": "186,0.1,137.8,97,187.7,118,146.4,85,8.7,6,1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,0.10,0.11,0.12,0.13,0.14,0.15,0.16,0.17,1.1,0.18,0.19,0.20,0.21,0.22,0.23,0.24,0.25,0.26,0.27,0.28,0.29,0.30,0.31,0.32,0.33,0.34,0.35,0.36,0.37,0.38,0.39,0.40,0.41,0.42,0.43,0.44,0.45,0.46,0.47,0.48,0.49,0.50,0.51,0.52,0.53,1.2,1.3,0.54,1.4,0.55",
      "encoding": "CSV"
    },
    "endpointOutput": {
      "observedContentType": "text/csv; charset=utf-8",
      "mode": "OUTPUT",
      "data": "0.014719205908477306",
      "encoding": "CSV"
    }
  },
  "eventMetadata": {
    "eventId": "7ae2cbc6-8a04-4228-92fe-f785fc1fb09b",
    "inferenceTime": "2020-08-01T10:09:04Z"
  },
  "eventVersion": "0"
}
```

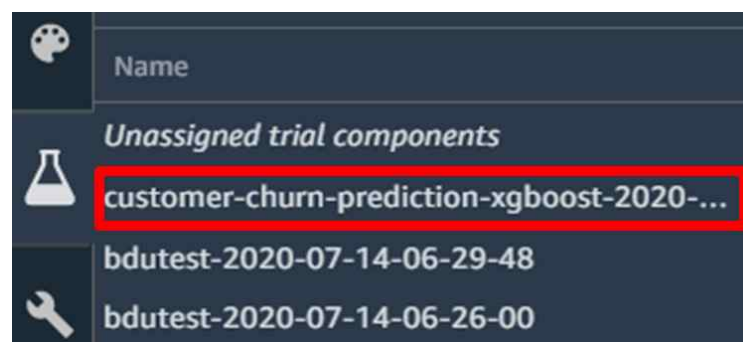
## 2. 최상의 모델 배포

- 최상의 모델 배포 단계

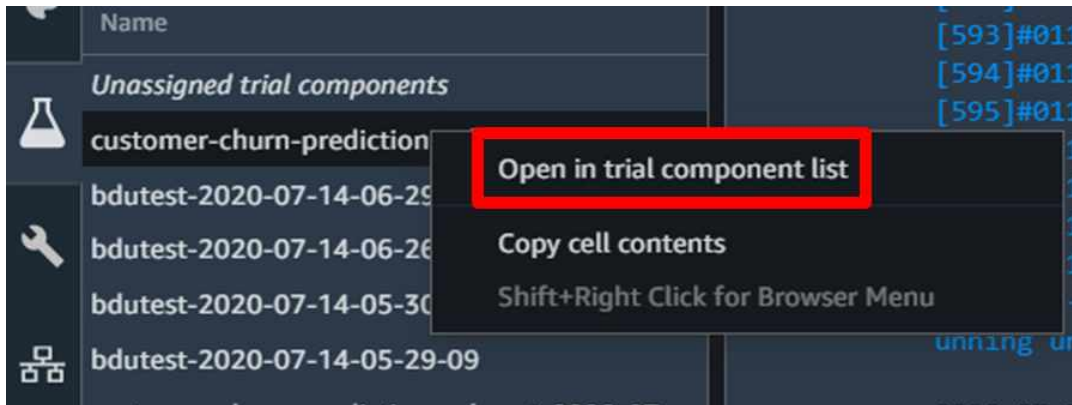
- ① 시도 구성 요소 목록 진입
- ② 배포할 모델 선택
- ③ 엔드포인트 설정
- ④ 모델 배포

- ✓ 시도 구성 요소 목록 진입

- 실험 목록에서 홈 아이콘 클릭
- 시도명을 오른쪽 버튼으로 클릭
- customer-churn-prediction-xgboost-...

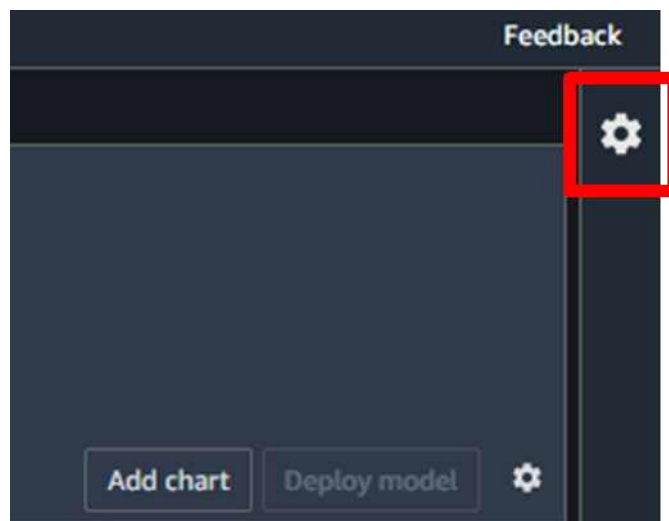


- 시도 구성 요소 목록 열기 선택

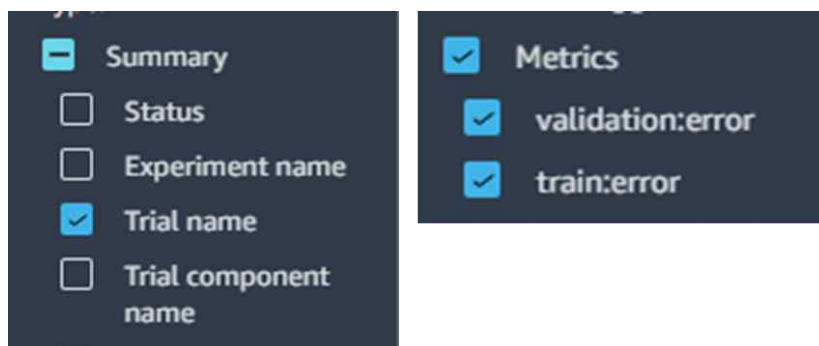


- ✓ 배포할 모델 선택

- 오른쪽 상단의 환경 설정 아이콘을 선택



- Summary 모두 해체 후 Trial name 체크
- Metrics 모두 체크



- 시도별 validation:error, train:error 모니터링 가능

Clear all

Trial name	validation:error ▼	train:error ▲
algorithm-mode-trial-2020-07-30-12-49-46-weight-10	0.075075	0.042006
algorithm-mode-trial-2020-07-30-12-19-44	0.072072	0.033862
algorithm-mode-trial-2020-07-30-12-52-56	0.072072	0.036434
framework-mode-trial-2020-07-30-13-53-02	0.069069	0.03129
algorithm-mode-trial-2020-07-30-12-49-44-weight-8	0.06606599688529968	0.02100300043821335
algorithm-mode-trial-2020-07-30-12-49-38-weight-2	0.064565	0.024432
algorithm-mode-trial-2020-07-30-12-49-38-weight-1	0.06006	0.030862
algorithm-mode-trial-2020-07-30-12-49-40-weight-4	0.06006	0.032576

End of the list

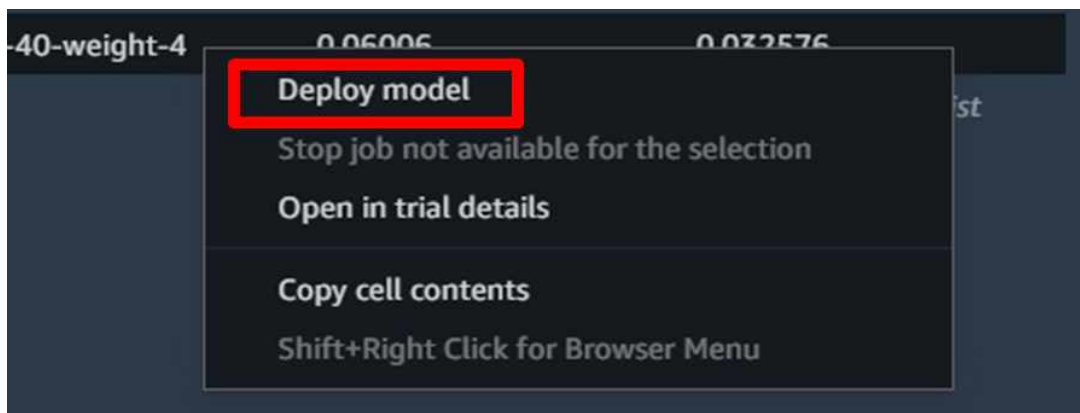
- validation:error 헤더를 선택하여 시도를 정렬

Clear all

Trial name	validation:error ▲	train:error ▲
framework-mode-trial-2020-08-01-09-27-41	0.072072	0.03129
algorithm-mode-trial-2020-08-01-09-00-00-weight-8	0.06606599688529968	0.02100300043821335
algorithm-mode-trial-2020-08-01-09-00-00-weight-10	0.066066	0.029147
algorithm-mode-trial-2020-08-01-08-59-59-weight-4	0.070571	0.046721
algorithm-mode-trial-2020-08-01-08-59-54-weight-1	0.069069	0.04072
algorithm-mode-trial-2020-08-01-08-59-54-weight-2	0.064565	0.025289
algorithm-mode-trial-2020-08-01-08-26-45	0.067568	0.033005

End of

- validation:error가 가장 낮은 시도를 마우스 오른쪽 버튼으로 클릭하고 모델 배포를 선택



**Deploy model**

**REQUIRED SETTINGS**

Endpoint name  
  
 Maximum of 63 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in an AWS Region.

Instance type      Instance count  
     

Data capture  
 SageMaker Studio will save prediction requests and responses from the endpoint to an Amazon S3 location specified below

☐ Save prediction requests  
☐ Save prediction responses

**ADVANCED SETTINGS - Optional** ▼

**Deploy model**

✓ 엔드포인트 설정

- 모델 배포에서 모델을 호스팅할 엔드포인트의 세부 정보 입력
  - 엔드포인트 이름
  - 인스턴스 타입 및 개수
  - 데이터 캡처
  - 고급설정-옵션

✓ 모델 배포

- 모델 배포에서 모델을 호스팅할 엔드포인트의 세부 정보 입력

**Deploy model**

**REQUIRED SETTINGS**

Endpoint name  
  
 Maximum of 63 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in an AWS Region.


Instance type      Instance count  
     

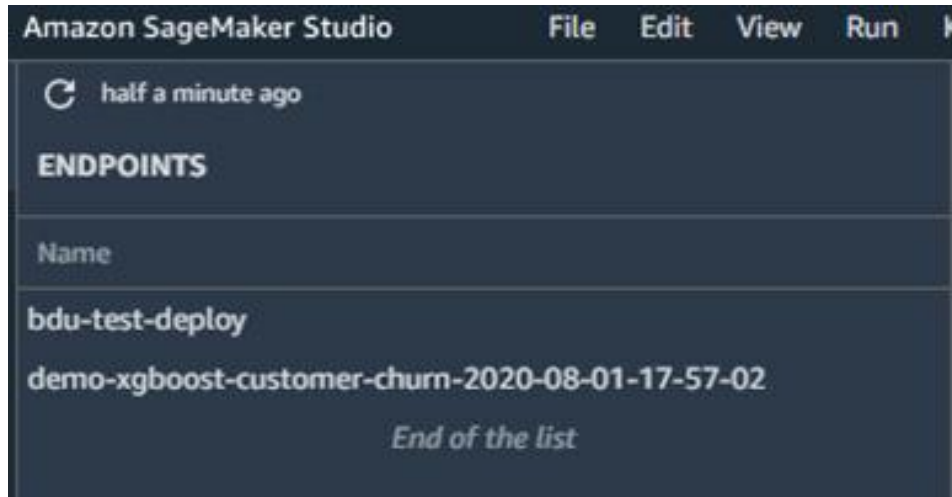
Data capture  
 SageMaker Studio will save prediction requests and responses from the endpoint to an Amazon S3 location specified below

☐ Save prediction requests  
☐ Save prediction responses

**Deploy model**



- 스튜디오 왼쪽 탭에서 엔드포인트 리스트 아이콘(  ) 클릭



- 스튜디오 콘솔 대시보드에서 엔드포인트 탭 클릭

엔드포인트					
<input type="text" value="엔드포인트 검색"/>		<input type="button" value="엔드포인트 업데이트"/>	<input type="button" value="작업"/>	<input type="button" value="엔드포인트 생성"/>	
이름	ARN	생성 시간	상태	최종 업데이트 날짜	
<a href="#">bdu-test-deploy</a>	arn:aws:sagemaker:us-east-1:402345010737:endpoint/bdu-test-deploy	Aug 01, 2020 18:32 UTC	Creating	Aug 01, 2020 18:32 UTC	
<a href="#">demo-xgboost-customer-churn-2020-08-01-17-57-02</a>	arn:aws:sagemaker:us-east-1:402345010737:endpoint/demo-xgboost-customer-churn-2020-08-01-17-57-02	Aug 01, 2020 17:57 UTC	InService	Aug 01, 2020 18:05 UTC	

- 스튜디오 콘솔 대시보드에서 엔드포인트 구성 탭 클릭

엔드포인트 구성			
<input type="text" value="엔드포인트 구성 검색"/>		<input type="button" value="엔드포인트에 적용"/>	<input type="button" value="복제"/>
		<input type="button" value="작업"/>	<input type="button" value="엔드포인트 구성 생성"/>
이름	ARN	생성 시간	
<a href="#">bdu-test-deploy-endpointConfig-1596306732110</a>	arn:aws:sagemaker:us-east-1:402345010737:endpoint-config/bdu-test-deploy-endpointconfig-1596306732110	Aug 01, 2020 18:32 UTC	
<a href="#">demo-xgboost-customer-churn-2020-08-01-17-57-02</a>	arn:aws:sagemaker:us-east-1:402345010737:endpoint-config/demo-xgboost-customer-churn-2020-08-01-17-57-02	Aug 01, 2020 17:57 UTC	



### 3. 리소스 정리

- 최상의 모델 배포 단계
  - 추가적인 비용이 발생하지 않으려면 생성된 리소스를 정리해야함

```
try:
    sess.delete_monitoring_schedule(mon_schedule_name)
except:
    pass
while True:
    try:
        print("Waiting for schedule to be deleted")
        sess.describe_monitoring_schedule(mon_schedule_name)
        sleep(15)
    except:
        print("Schedule deleted")
        break

sess.delete_endpoint(xgb_predictor.endpoint)

def cleanup(experiment):
    """Clean up everything in the given experiment"""
    for trial_summary in experiment.list_trials():
        trial = Trial.load(trial_name=trial_summary.trial_name)

        for trial_comp_summary in trial.list_trial_components():
            trial_step = TrialComponent.load(trial_component_name=trial_comp_summary.trial_component_name)
            print("Starting to delete TrialComponent... " + trial_step.trial_component_name)
            sm.disassociate_trial_component(TrialComponentName=trial_step.trial_component_name, TrialName=trial_name)
            trial_step.delete()
            time.sleep(1)

        trial.delete()

    experiment.delete()

cleanup(customer_churn_experiment)
```

- aws 콘솔에서 모델, 엔드포인트 구성, 엔드포인트 등 재확인



- S3 콘솔 접속 후 버킷 삭제



**평가하기**

1. 모델 배포 시 입력해야 하는 엔드포인트의 세부 정보로 옳지 않은 것은?

- ① 엔드포인트 이름
- ② 데이터 캡처
- ③ 인스턴스 타입
- ④ 리전

- 정답 : ④ 번

해설 : 모델 배포 시 입력해야 하는 엔드포인트의 세부 정보로는 엔드포인트 이름, 인스턴스 타입/개수, 데이터 캡처, 고급설정이 있습니다.

2. 모델 배포 작업 완료 후 추가적인 비용이 발생하지 않도록 세이지메이커 콘솔에서 정리해야 할 리소스로 옳지 않은 것은?

- ① 엔드포인트
- ② 엔드포인트 구성
- ③ 훈련 작업
- ④ 모델

- 정답 : ③ 번

해설 : 세이지메이커 콘솔에서 훈련작업은 모델 훈련 과정 정보(작업설정, 사용 알고리즘, 입력 데이터, 파라미터, 모니터링 등)을 확인할 수 있으며, 별도의 제거 기능은 없습니다.

### 학습정리

#### 1. 모델 배포

- 모델 호스팅
- 배포된 모델 호출
- 데이터 캡처

#### 2. 최상의 모델 배포

- 시도 구성 요소 목록 진입
- 배포할 모델 선택
- 엔드포인트 선택
- 모델 배포

#### 3. 리소스 정리

- 소스 코드로 리소스 정리
- 콘솔에서 리소스 정리