<http://bit.ly/sagemaker11>



Amazon SageMaker  
Hands-on Guidebook

6th November 2018

강지양 SA / 김태현 SA

Table of Contents

[Lab 개요 3](#_Toc527349486)

[목표 3](#_Toc527349487)

[준비 조건 3](#_Toc527349488)

[LAB 실습 가이드 4](#_Toc527349489)

[Module 1: Notebook Instance생성하기 5](#_Toc527349490)

[Module 2: 실습용 코드 다운 받기 10](#_Toc527349491)

[Module 3: Factorization Machine을 이용한 영화 추천 서비스 Notebook 12](#_Toc527349492)

[Module 4: TensorFlow 를 활용한 분산 학습 Notebook 14](#_Toc527349493)

[Module 5: TensorFlow를 활용한 분산 학습의 자동 모델 튜닝 Notebook 15](#_Toc527349494)

[Module 6: 자동 모델 튜닝 결과 분석 Notebook 18](#_Toc527349495)

[서비스 종료 가이드 19](#_Toc527349496)

# Lab 개요

Amazon SageMaker는 데이터 사이언티스트와 개발자들이 쉽고 빠르게 구성, 학습하고 어떤 규모로든 기계 학습된 모델을 배포할 수 있도록 해주는 관리형 서비스 입니다. 이 워크샵을 통해 SageMaker notebook instance를 생성하고 샘플 Jupyter notebook을 실습하면서 SageMaker의 일부 기능을 알아보도록 합니다.

# 목표

* SageMaker에 내장된 학습 기능을 사용하여 모델 훈련 Job을 생성 합니다.
* SageMaker의 endpoint 기능을 사용하여 생성된 모델이 예측에 사용될 수 있도록 endpoint를 생성합니다.
* 머신 러닝이 정형 데이터(e.g. CSV 파일)와 비정형 데이터(e.g. 이미지)에 모두 적용 될수 있음을 확인 합니다.

# 준비 조건

* AWS 계정: AWS IAM, S3, SageMaker 자원을 생성할 수 있는 권한이 필요합니다.
* AWS Region: SageMaker는 지원되는 region은 <https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/> 에서 확인하실 수 있습니다. 이번 실습은Seoul (ap-northeast-2) region에서 실행 합니다.
* Browser: 최신 버전의 Chrome, Firefox를 사용하세요.

**※ 주의 사항:** Notebook 안의 Cell에서 코드 실행후 결과 값이 나오는 데는 수 초가 걸립니다. 훈련 Job을 실행하는 경우 수 분이 걸릴 수도 있습니다. 실습 완료 후에는 아래 가이드에 따라 생성된 자원을 꼭 종료/삭제해 주세요.

# LAB 실습 가이드

실습은 총 6개 모듈로 구성되어 있습니다. 1번 완료후 2번을 순서대로 진행하셔야 합니다. 3번, 4번, 5번 모듈은 원하는 순서대로 진행하실 수 있습니다. 6번 모듈은 반드시 5번 모듈 완료 후에 실행하시기 바랍니다.

1. Notebook Instance 생성하기
2. 실습용 코드 다운받기
3. Factorization Machine을 이용한 영화 추천 서비스 Notebook
4. TensorFlow를 활용한 분산 학습 Notebook
5. TensorFlow를 활용한 분산 학습의 자동 모델 튜닝 Notebook
6. 자동 모델 튜닝 결과 분석 Notebook

## Module 1: Notebook Instance생성하기

**1. S3 Bucket생성하기**

SageMaker는 S3를 데이터와 모델 저장소로 사용합니다. 여기서는 해당 목적으로 S3 Bucket을 생성합니다.

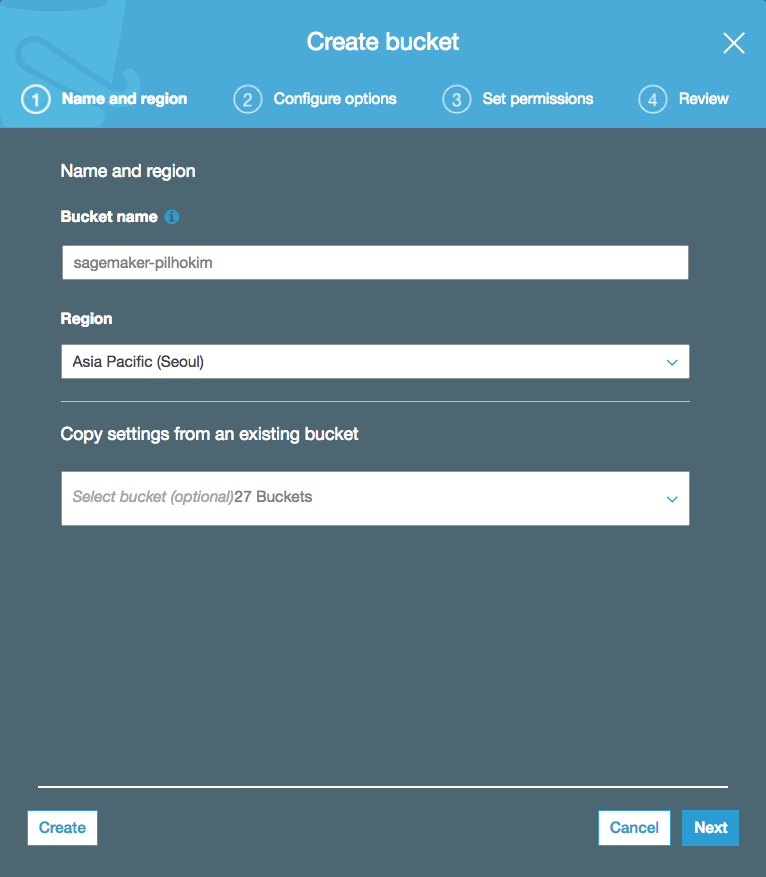
1) AWS 관리 콘솔 (<https://console.aws.amazon.com/>)에 Sign in 합니다.

2) AWS Services 리스트에서 S3 로 이동합니다.

3) “+ Create Bucket” 버튼을 선택합니다.

4) 아래 내용 설정 후 화면 왼쪽 아래 **Create** 클릭합니다.

* Bucket name: sagemaker-{userid} [반드시 고유한 값 설정]
* Region : Asia Pacific (Seoul)



**2. Notebook instance 생성**

1) AWS관리 콘솔에서 오른쪽 상단에서 Seoul Region선택 후 AWS Services 리스트에서 Amazon SageMaker 서비스를 선택합니다.



2) 새로운 Notebook instance를 생성하기 위해 왼쪽 패널 메뉴 중 **Notebook Instances** 선택 후 오른쪽 상단의 **Create notebook instance** 버튼을 클릭 합니다.



3) Notebook instance 이름으로 **[First Name]-[Last Name]-workshop**으로 넣은 뒤 **ml.m4.xlarge** 인스턴스 타입을 선택 합니다.



4) IAM role은 **Create a new role**을 선택하고, 생성된 팝업 **창에서는 S3 buckets you specify – optional** 밑에 **Specific S3 Bucket**을 선택 합니다. 그리고 텍스트 필드에 위에서 만든 S3 bucket 이름(예: sagemaker-xxxxx)을 선택 합니다. 이후 **Create role**을 클릭합니다.



5) 다시 Create Notebook instance 페이지로 돌아온 뒤 **Create notebook** **instance**를 클릭합니다.

**3. Notebook Instance 접근하기**

1) 서버 상태가 **InService** 로 바뀔 때까지 기다립니다. 보통 5분정도의 시간이 소요 됩니다.



2) Open을 클릭하면 방금 생성한 notebook instance의 Jupyter 홈페이지로 이동하게 됩니다.



## Module 2: 실습용 코드 다운 받기

SageMaker의 Jupyter 노트북도 Linux 기반의 서버입니다. Jupyter 노트북에서 서버의 Terminal을 바로 실행하는 기능을 제공하고 있습니다. Figure 1와 같이 Terminal을 선택합니다.



Figure 1. SageMaker 노트북 서버에 접속하기 위한 Terminal 실행 화면.

Figure 2과 같이 터미널이 실행되면 아래의 명령어들을 입력해서 실행합니다.

cd SageMaker/

git clone <https://github.com/aws-samples/aws-ai-ml-workshop-kr>

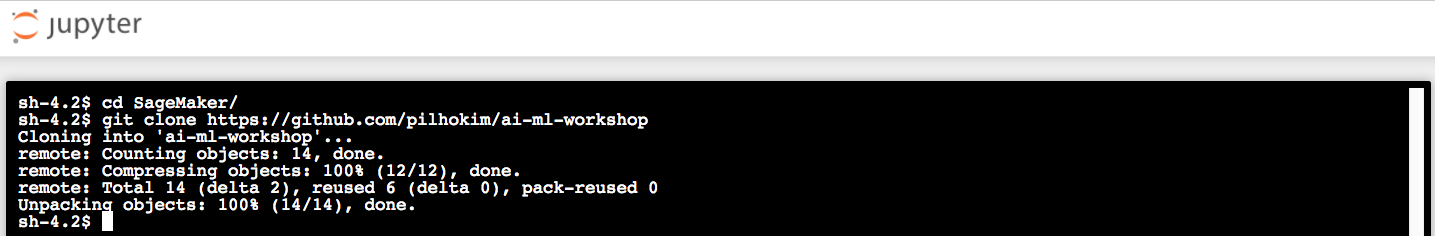


Figure 2. GitHub에서 실습 코드 다운 받기.

코드를 다운 받고 난 후 Jupyter 노트북을 갱신 하면 (오른쪽 상단의 Refresh 아이콘을 클릭하세요) 새롭게 다운 받은 코드 폴더가 보입니다 (Figure 3).

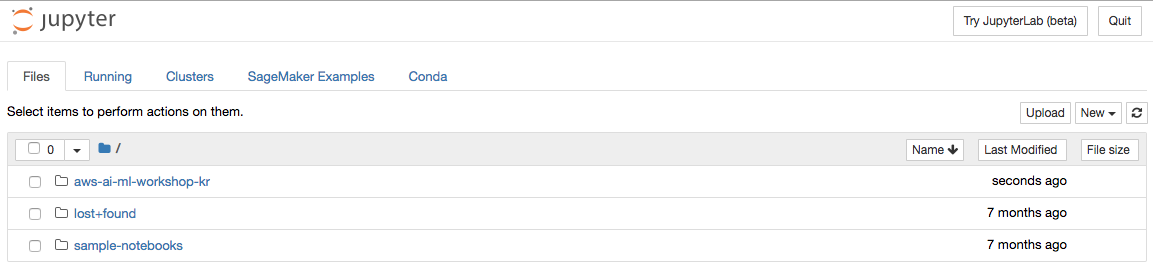


Figure 3. 새로 다운 받은 폴더 생성 확인 화면.

폴더를 클릭해 aws-ai-ml-workshop-kr/src/release/dev-day 폴더를 찾습니다 (Figure 4). 이 폴더에서 다음 단계의 실습을 진행합니다.

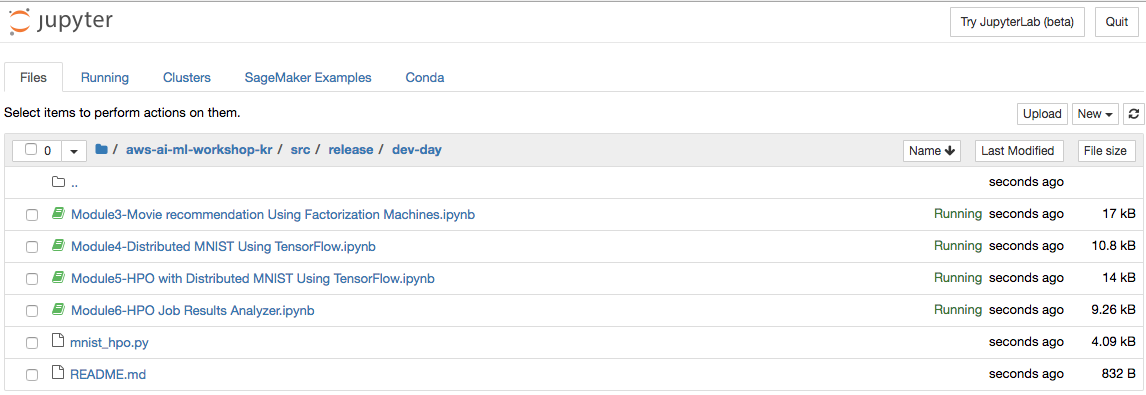


Figure 4. 실습 폴더 확인 화면.

## Module 3: Factorization Machine을 이용한 영화 추천 서비스 Notebook

이 모듈에서는 Jupyter notebook 예제를 통해 어떻게 아마존이 제공하는 알고리즘을 SageMaker에서 사용할 수 있는지 알아 봅니다. 특히 SageMaker버전의 Factorization Machines 알고리즘을 사용하게 되는데, Factorization Machines는 2010년에 소개된 Supervised Machine Learning 알고리즘으로 행렬 인수분해(Matrix Factorization)과 SVM(Support Vector Machine) 개념을 결합한 알고리즘입니다. Factorization Machines 알고리즘은 분류 또는 회귀에 사용할 수 있으며 선형 회귀와 같은 알고리즘보다는 고차원 희소 데이터 셋을 처리하는데 훨씬 효율적인 알고리즘 입니다. Amazon SageMaker의 Factorization Machines 알고리즘은 확장성이 높고 배포된 인스턴스에 걸쳐 교육이 가능하고 희소 및 밀집 데이터 세트에 대해 CPU 인스턴스를 사용한 교육 및 추론을 권장합니다.

이 모듈에서는 Factorization Machines에 대한 소개와 이를 이용한 영화 추천 서비스를 만드는 과정입니다. 본 실습을 통해:

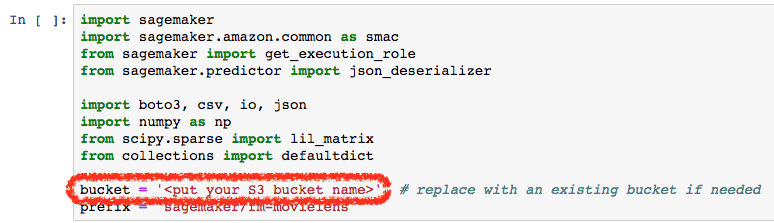
* Factorization Machines 알고리즘을 위한 데이터 준비 과정
* Factorization Machines모델 훈련 및 배포

을 다루게 됩니다.

실습을 위해서는 현재 설치되어 있는 SageMaker의 Jupyter 노트북의 예제들 중 아래의 디렉토리에 위한 Jupyter 노트북을 실행하시면 됩니다.

**/aws-ai-ml-workshop-kr/src/release/dev-day/Module3-Movie recommendation Using Factorization Machines.ipynb**

모듈 실행중 아래 코드를 만나면 모듈 1, 2에서 사용한 s3 버킷의 이름(예: sagemaker-xxxxx)을 넣고 실행합니다.



※ 이 모델을 훈련하는데는 약 10분에서 15 분이 소요됩니다.

## Module 4: TensorFlow 를 활용한 분산 학습 Notebook

이 모듈에서는[MNIST Database](http://yann.lecun.com/exdb/mnist/) 에서 손으로 쓴 숫자의 이미지 데이터를 활용하여 SageMaker에서 어떻게 분산 학습을 실행하는지 배우게 됩니다. 이 모듈은 [TensorFlow MNIST Example](https://github.com/tensorflow/models/tree/master/official/mnist) 에 기반한 Convolutional Neural Network model을 활용합니다.

이 모듈에서는 데이터 전처리 작업과 모델 학습은 물론, SageMaker의 hosted endpoint생성과 훈련된 모델을 endpoint에 실제로 적용 하기 위해 어떻게 Jupyter notebook과 SageMaker Python SDK를 사용하는지를 차례로 실습할 수 있습니다.

생성된 모델은 실제로 사용자가 그려 넣은 숫자가 무엇인지 예측합니다. 이 예제에서는 TensorFlow를 사용해 자신의 코드를 가져와 실행하는 것 뿐만 아니라, SageMaker에서 모델 학습을 위해 여러 대의 인스턴스 클러스터를 얼마나 쉽게 생성할 수 있는지 보여 줍니다.

실습을 위해서는 현재 설치되어 있는 SageMaker의 Jupyter 노트북의 예제들 중 아래의 디렉토리에 위한 Jupyter 노트북을 실행하시면 됩니다.

**/aws-ai-ml-workshop-kr/src/release/dev-day/Module4-Distributed MNIST Using TensorFlow.ipynb**

※ 이 모델을 훈련하는데는 약 15분에서 20분 정도가 소요됩니다.

## Module 5: TensorFlow를 활용한 분산 학습의 자동 모델 튜닝 Notebook

이 모듈에서는 앞 모듈의 MNIST 이미지 분류 예제를 기반으로 SageMaker의 자동 모델 튜닝 기능을 실습합니다. 이 기능은 기계 학습 알고리즘의 최적의 하이퍼파라미터(Hyperparameter) 값을 베이지안 최적화 기법을 통해 찾아줍니다.

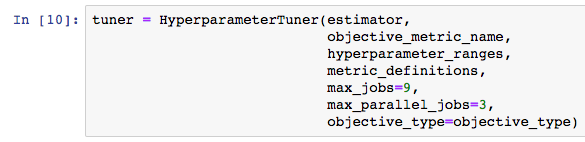
실습을 위해서 현재 설치되어 있는 SageMaker의 Jupyter 노트북의 예제들 중 아래의 디렉토리에 위한 Jupyter 노트북을 실행하시면 됩니다.

**/aws-ai-ml-workshop-kr/src/release/dev-day/Module5-HPO with Distributed MNIST Using TensorFlow.ipynb**

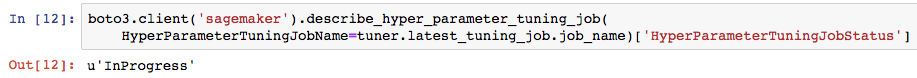
이 모듈에서는 MNIST 이미지 분류 예제의 하이퍼파라미터 중에서 learning rate 값을 자동으로 튜닝하며, 효과적인 탐색을 위해 최대값과 최소값을 아래 그림과 같이 설정합니다.



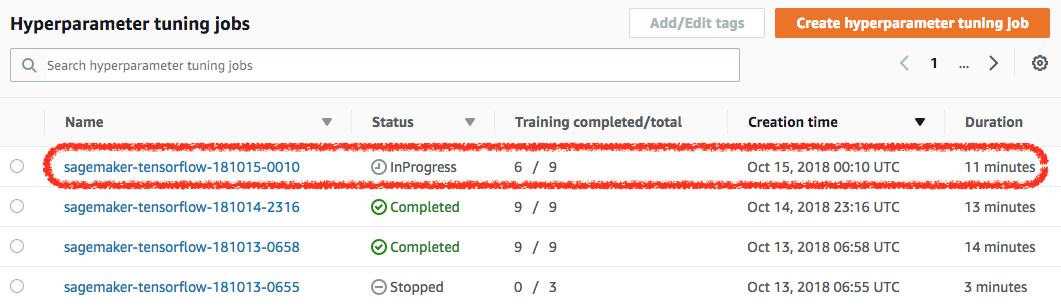
베이지안 최적화 기법은 하이퍼파라미터를 변경하면서 미리 지정된 숫자만큼 실험을 반복하는 특징이 있습니다. 이번 모듈에서는 병렬로 3개의 학습을 3번, 즉 총 9번의 실험을 시도하도록 아래와 같이 설정합니다.



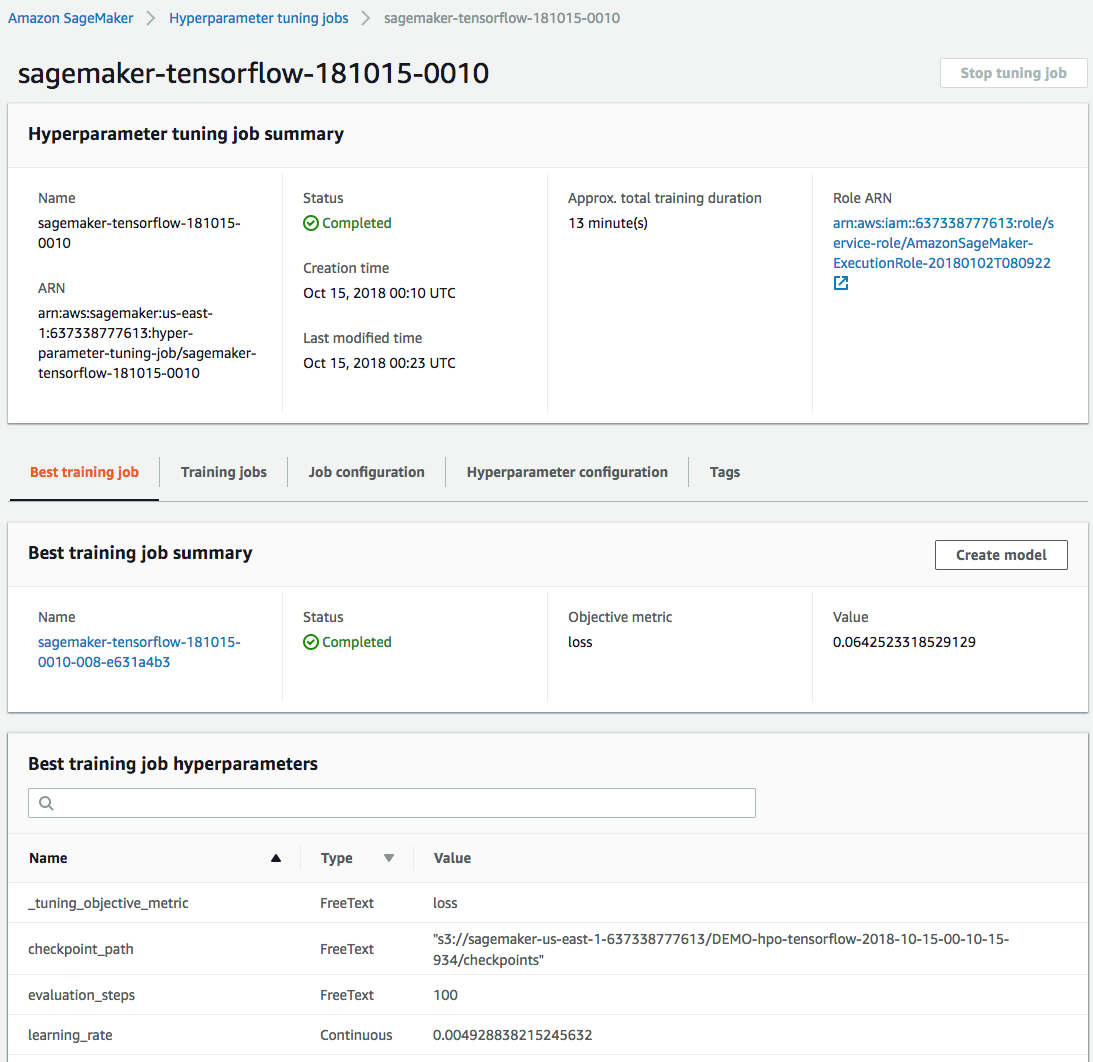
하이퍼파라미터 튜닝 작업은 아래와 같은 코드로 실행하며, 실행하면 각 하이퍼파라미터 값에 대한 개별 학습이 백그라운드에서 시작됩니다.



이 때, SageMaker의 콘솔에서 새로운 **하이퍼파라미터 튜닝 작업** (**Hyperparameter tuning jobs**)이 생성된 것을 확인할 수 있습니다. 다음 모듈을 위해 이 작업의 이름을 메모해 놓습니다.



실험이 모두 끝나면 하이퍼파라미터 튜닝 작업의 이름을 클릭해 튜닝 결과를 확인합니다. 아래 그림에서는 learning\_rate가 0.004928838215245632가 최적의 값이며 이때의 loss 값은 0.0642523318529129인 것을 확인할 수 있습니다.



※ 이 모델을 훈련하는데는 약 20분이 소요됩니다.

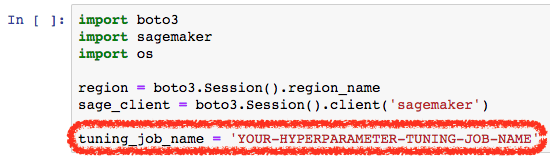
## Module 6: 자동 모델 튜닝 결과 분석 Notebook

이 모듈에서는 앞에서 실행한 하이퍼파라미터 튜닝 작업의 결과를 해석하는 과정을 실습합니다. BokehJS와 pandas 라이브러리를 사용해 튜닝 결과를 Jupyter 노트북에서 테이블과 그래프 형태로 시각화해볼 수 있습니다.

실습을 위해서 현재 설치되어 있는 SageMaker의 Jupyter 노트북의 예제들 중 아래의 디렉토리에 위한 Jupyter 노트북을 실행하시면 됩니다.

**/aws-ai-ml-workshop-kr/src/release/dev-day/Module6-HPO Job Results Analyzer.ipynb**

모듈의 첫 부분에서 아래의 코드를 만나면 앞 모듈에서 실행된 **하이퍼파라미터 튜닝 작업** (**Hyperparameter tuning jobs**)의 이름을 따옴표 안에 넣습니다.



실행 결과로 나오는 두 개의 그래프에서, 탐색된 하이퍼파라미터 값의 변화에 따른 loss 함수 값의 변화를 해석해 보시기 바랍니다.

※ 이 모듈의 실습에는 약 5분이 소요됩니다.

# 서비스 종료 가이드

워크 샵 이후 발생 되는 비용을 방지하기 위해서 아래의 단계에 따라 모두 종료/삭제 해 주세요. 비용이 발생하더라도 실습하신 Internet-facing App을 유지하고 싶으신 경우에는 아래의 Notebook instance의 경우만 처리하시면 됩니다.

* Notebook instance:

1) 만약 향후 사용을 위해 인스턴스를 저장하고 싶다면 **stop**을 하시면 됩니다. 이 경우 스토리지 비용은 발생합니다. 향후 다시 재가동 하시려면 Start button을 클릭하면 됩니다.

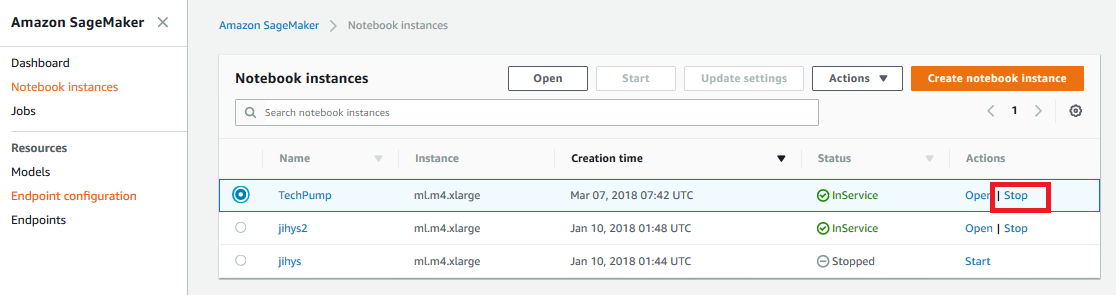


Figure 4. SageMaker 노트북 인스턴스 중단 화면.

2) 삭제를 할 경우는 **stop** 되어 있는 해당 notebook instance를 선택하고 **Action**  Dropdown 메뉴에서 **Delete** 선택 하시면 됩니다.

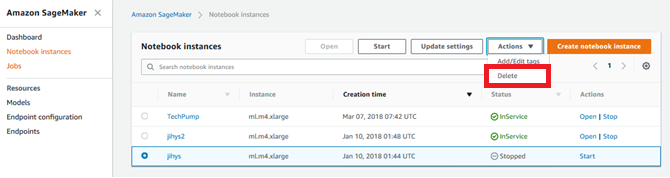


Figure 5. SageMaker 노트북 인스턴스 삭제 화면.

* SageMaker Endpoints:

훈련된 모델을 실제 예측 업무를 위해 배포된 한대 이상으로 구성된 클러스터입니다. Notebook안에서 명령어로 삭제하거나 SageMaker console에서 삭제 하실 수 있습니다. 삭제 하시기 위해서는 왼쪽 패널의 Endpoints를 선택 하신 후 해당 endpoints들 옆에 radio button을 클릭 하신 후 Action Dropdown 메뉴에서 Delete 선택 하시면 됩니다.

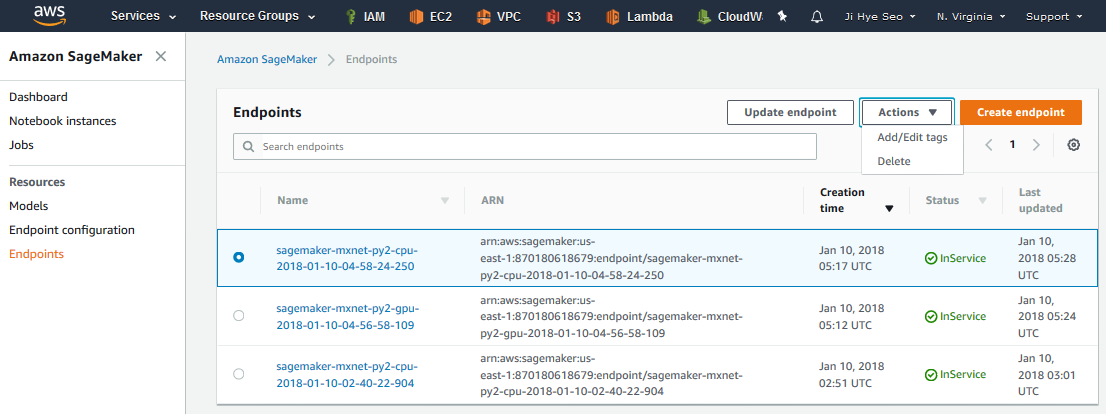


Figure 6. SageMaker Endpoint 삭제 화면.

* Lambda instance: 생성하신 Lambda instance를 삭제합니다.

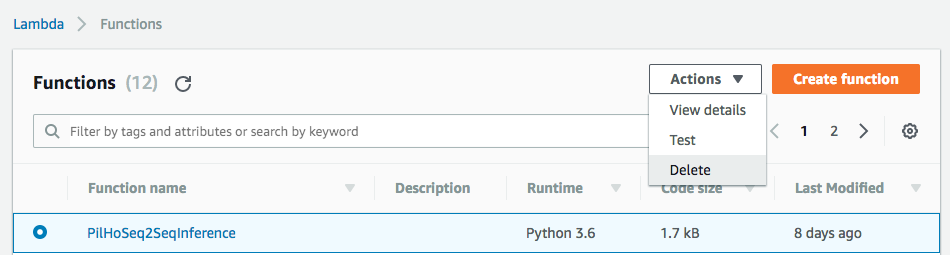


Figure 7. Lambda 인스턴스 삭제 화면.

* Amazon API Gateway instance: 생성하신 Gateway instance를 삭제합니다.

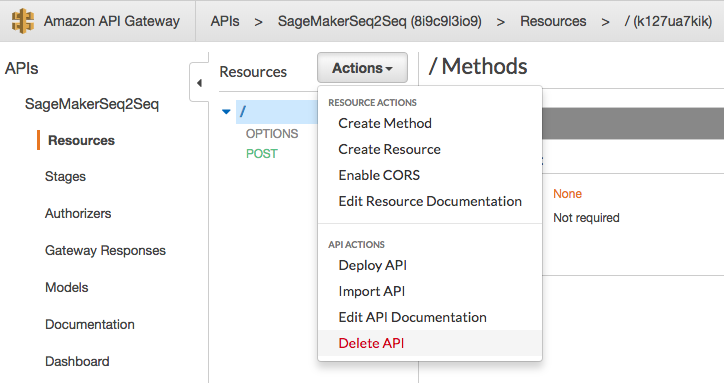


Figure 8. API Gateway 삭제 화면.

* Amazon S3 buckets: 생성하신 S3 Bucket (SageMaker용, Public Internet용)들을 모두 삭제합니다.

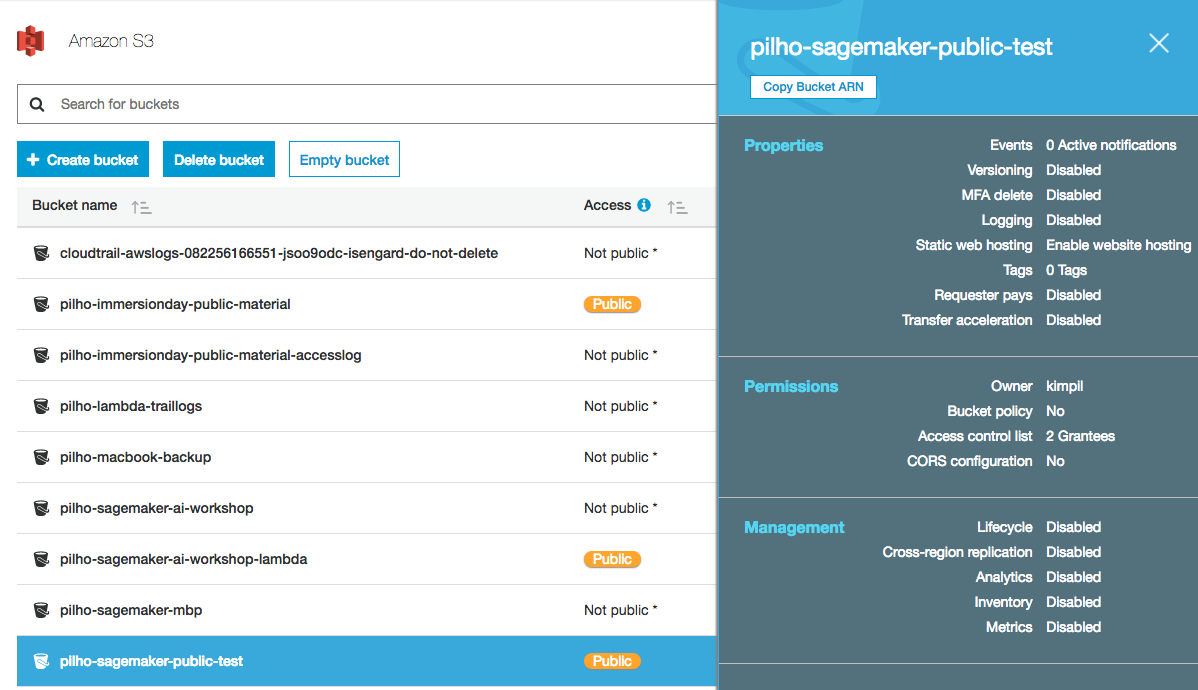


Figure 9. S3 버킷 삭제 화면.

이상으로 본 핸즈온 세션의 모든 과정을 마무리 하셨습니다. 수고하셨습니다.