

Training by Amazon
SageMaker console

Objectives

- Using Sklearn in Amazon SageMaker
- Starting a training job through Console using XGBoost built-in algorithm
- Using SageMaker Python SDK to train a linear learner algorithm
- Using SageMaker built-in frameworks to train a model

1) Using Sklearn in Amazon SageMaker

Using AWS Academy in Virginia

Run the Day06-01-Sklearn notebook

- In this notebook we **use sklearn library** to run a **linear regression** on the Jupyter instance. Please note that we **do not** use Amazon SageMaker training APIs in this notebook, everything happens inside this Jupyter notebook
- Upload the **IceCreamData.csv** to S3 bucket

2) Starting a training job
through **Console** using XGBoost
built-in algorithm

Create a folder structure in S3

- Create a folder in your day-06 bucket called console
- Add three folders in that called “training”, “validation” and “output”

<input type="checkbox"/>	Name ▲	Type
<input type="checkbox"/>	📁 output/	Folder
<input type="checkbox"/>	📁 training/	Folder
<input type="checkbox"/>	📁 validation/	Folder

- Add two files called “ console-train” and “console-validation” to the respective training and validation folders

A little about the data

- The data is about university admission

	GRE_Score	TOEFL_Score	University_Rating	SOP	LOR	CGPA	Research	Chance_of_Admission
0	337	118	4	4.5	4.5	9.65	1	0.92
1	324	107	4	4.0	4.5	8.87	1	0.76
2	316	104	3	3.0	3.5	8.00	1	0.72
3	322	110	3	3.5	2.5	8.67	1	0.80
4	314	103	2	2.0	3.0	8.21	0	0.65
5	330	115	5	4.5	3.0	9.34	1	0.90

- I have cleaned the data to focus on the training activity but just to let you know the admission column that is the label column has been moved to the first column as that is the SageMaker xgboost algorithm requirement

A quick look into the training data set

- Just showing the first 23 lines
- We have 600 rows

1	0.9, 332.0, 118.0, 2.0, 4.5, 3.5, 9.36, 1.0
2	0.7, 308.0, 110.0, 4.0, 3.5, 3.0, 8.6, 0.0
3	0.93, 328.0, 116.0, 4.0, 5.0, 3.5, 9.6, 1.0
4	0.71, 309.0, 105.0, 5.0, 3.5, 3.5, 8.56, 0.0
5	0.77, 312.0, 109.0, 3.0, 3.0, 3.0, 8.69, 0.0
6	0.72, 316.0, 105.0, 3.0, 3.0, 3.5, 8.73, 0.0
7	0.46, 308.0, 110.0, 3.0, 3.5, 3.0, 8.0, 1.0
8	0.79, 313.0, 109.0, 3.0, 4.0, 3.5, 9.0, 0.0
9	0.7, 324.0, 111.0, 3.0, 2.5, 1.5, 8.79, 1.0
10	0.8, 327.0, 113.0, 3.0, 3.5, 3.0, 8.66, 1.0
11	0.64, 316.0, 102.0, 3.0, 2.0, 3.0, 7.4, 0.0
12	0.8, 317.0, 110.0, 3.0, 4.0, 4.5, 9.11, 1.0
13	0.59, 300.0, 101.0, 3.0, 3.5, 2.5, 7.88, 0.0
14	0.54, 299.0, 96.0, 2.0, 1.5, 2.0, 7.86, 0.0
15	0.65, 308.0, 104.0, 2.0, 2.5, 3.0, 8.07, 0.0
16	0.71, 309.0, 105.0, 5.0, 3.5, 3.5, 8.56, 0.0
17	0.44, 301.0, 97.0, 2.0, 3.0, 3.0, 7.88, 1.0
18	0.67, 318.0, 110.0, 1.0, 2.5, 3.5, 8.54, 1.0
19	0.56, 313.0, 94.0, 2.0, 2.5, 1.5, 8.13, 0.0
20	0.73, 323.0, 107.0, 3.0, 3.5, 3.5, 8.55, 1.0
21	0.59, 299.0, 100.0, 1.0, 1.5, 2.0, 7.89, 0.0
22	0.78, 311.0, 107.0, 4.0, 4.5, 4.5, 9.0, 1.0
23	0.86, 331.0, 120.0, 3.0, 4.0, 4.0, 8.96, 1.0

Configure a training job

- Select SageMaker built-in algorithm

[Amazon SageMaker](#) > [Training jobs](#) > Create training job

Create training job

When you create a training job, Amazon SageMaker sets up the distributed compute cluster, performs the training, and deletes the cluster when training has completed. The resulting model artifacts are stored in the location you specified when you created the training job. [Learn more](#)

Job settings

Job name

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

IAM role

Amazon SageMaker requires permissions to call other services on your behalf. Choose a role or let us create a role that has the [AmazonSageMakerFullAccess](#) IAM policy attached.

Algorithm options

Use an Amazon SageMaker built-in algorithm, your own algorithm, or a third-party algorithm from AWS Marketplace.

▼ Algorithm source

- ☒ Amazon SageMaker built-in algorithm [Learn more](#)
- ☐ Your own algorithm resource
- ☐ Your own algorithm container in ECR [Learn more](#)
- ☐ An algorithm subscription from AWS Marketplace

Select XGboost algorithm

▼ Choose an algorithm

Tabular - XGBoost : v1.3 ▼

Container

The registry path where the training image is stored in Amazon ECR. [Learn more](#)

683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:1.3-1

Input mode

You can provide your training data as a file or pipe.

File ▼

Metrics

The algorithm you selected will publish the following metrics to CloudWatch metrics

Select the server that you want to train

- This is a server that will be started by SageMaker and will be shut down automatically when the training is completed.

Resource configuration

Instance type	Instance count	Additional storage volume per instance (GB)
ml.m4.xlarge ▼	1	1

Encryption key - *optional*
Encrypt your data. Choose an existing KMS key or enter a key's ARN.

No Custom Encryption ▼

Stopping condition

Maximum runtime

24 hours ▼

Hyperparameters

- Set the two hyperparameter values as shown below:
 - **Objective: binary:logistic**
 - **num_round: 2**
- You need to define two channels for data: one for training data and another for validation data

Training channel

- Set the training Channel as shown

Channels

▼ train

Remove

Channel name

train

Input mode - optional

▼

Content type - optional

text/csv

Choose one of the formats below

• libsvm

• csv

Compression type

None ▼

Record wrapper

None ▼

Data source

☒ S3

☐ File system

S3 data type

S3Prefix ▼

S3 data distribution type

FullyReplicated ▼

S3 location

https://day-06-mk.s3.amazonaws.com/console/training/console-train.csv

Add channel

Validation channel

- Add a new validation channel

▼ validation

Remove

Channel name

validation

Input mode - *optional*

File

Content type - *optional*

text/csv

Choose one of the formats below

- libsvm
- csv

Compression type

None

Record wrapper

None

Data source

☒ S3

☐ File system

S3 data type

S3Prefix

S3 data distribution type

FullyReplicated

S3 location

https://day-06-mk.s3.amazonaws.com/console/validation/console-validation.csv

Add channel

Where the model artifact will be stored

- Set the output folder





Output data configuration

S3 output path






Encryption key - optional
If you want Amazon SageMaker to encrypt the output of your training job using your own AWS KMS encryption key instead of the default S3 service key, provide its ID or ARN.

Start the training job

- Review the training job

Training jobs Info							Actions 	Create training job
<input type="text" value="Search training jobs"/>						< 1 > 		
	Name	Creation time	Duration	Job status	Warm pool status			
<input type="radio"/>	xgboost-mk	Nov 16, 2022 02:42 UTC	4 minutes	 Completed	-			

- See the result of training job in the output folder

output/	
Objects	Properties
Objects (1) <small>Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 in</small>	
  Copy S3 URI  Copy URL  Download	
<input type="text" value="Find objects by prefix"/>	
<input type="checkbox"/>	Name ▲ Type
<input type="checkbox"/>	 model.tar.gz gz

3) Using SageMaker Python SDK
to train a linear learner
algorithm

Start the Day 06-02-Linear Learner notebook

- Upload the **experience_and_salary.csv** to Notebook instance
- Use **conda_python3** kernel
- Run the **Day 06-02-Linear-Learner** notebook

4) Using SageMaker built-in frameworks to train a model

Run Day-03-Script mode notebook

- Upload the **script.py** in to the instance

Assignment

- Amazon SageMaker has built-in algorithms. They are listed in this URL: <https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html>
- Do not use KNN algorithm, Other than KNN, select one of them that you are comfortable with to train a model
- Use an appropriate data set from <https://www.openml.org/> or www.kaggle.com
- You need to show a completed training job and the model artifact in the S3 bucket
- You need to upload the **selected data set and the notebook file** that you use to train the model to BB.
- **NOTE:** Please make sure you do not start a **Hosting job**. If you do so, make sure you terminate the hosting endpoint as it will consume your credit.