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 Amazom 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”

Name	Type
output/	Folder
training/	Folder
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

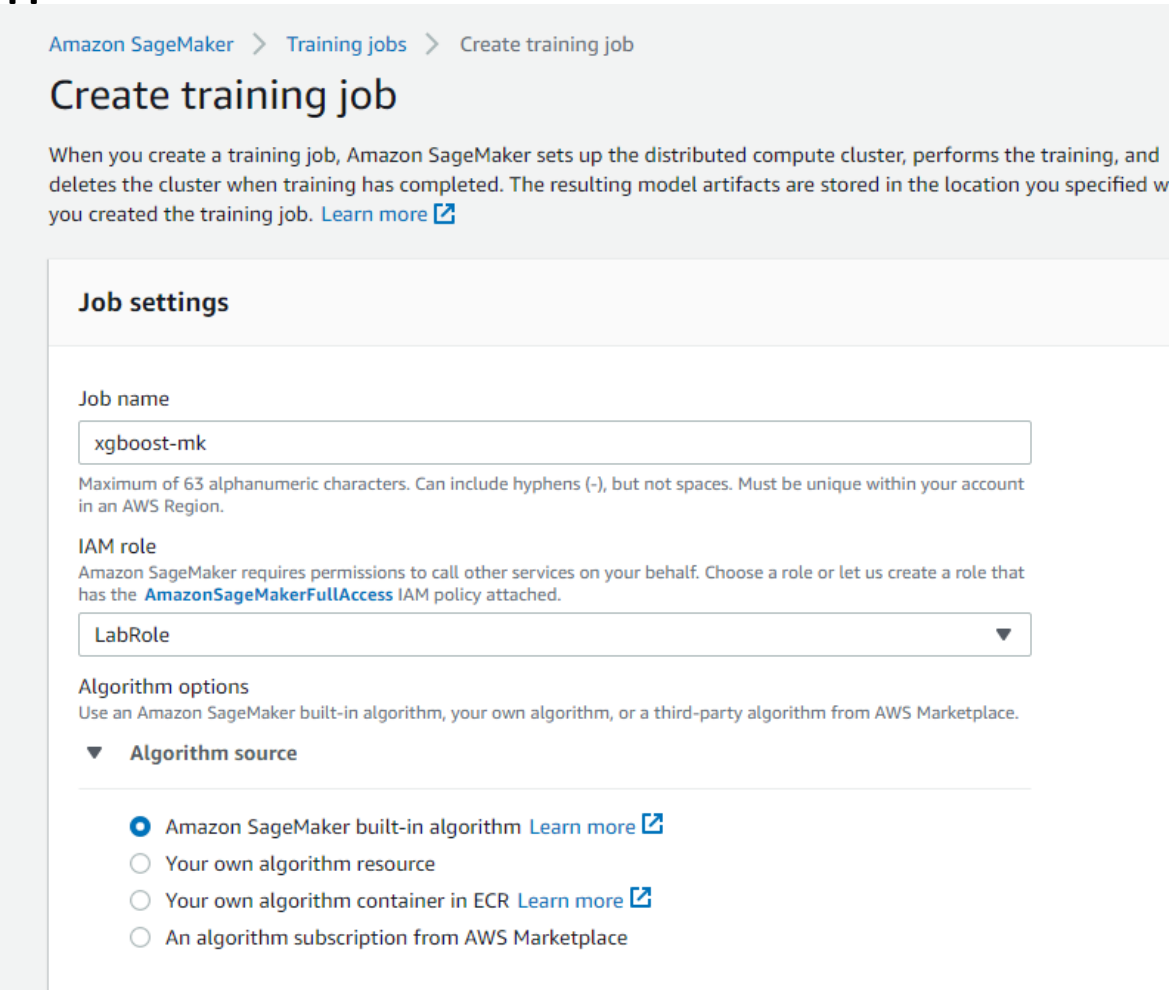
A quick look into the training data set

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

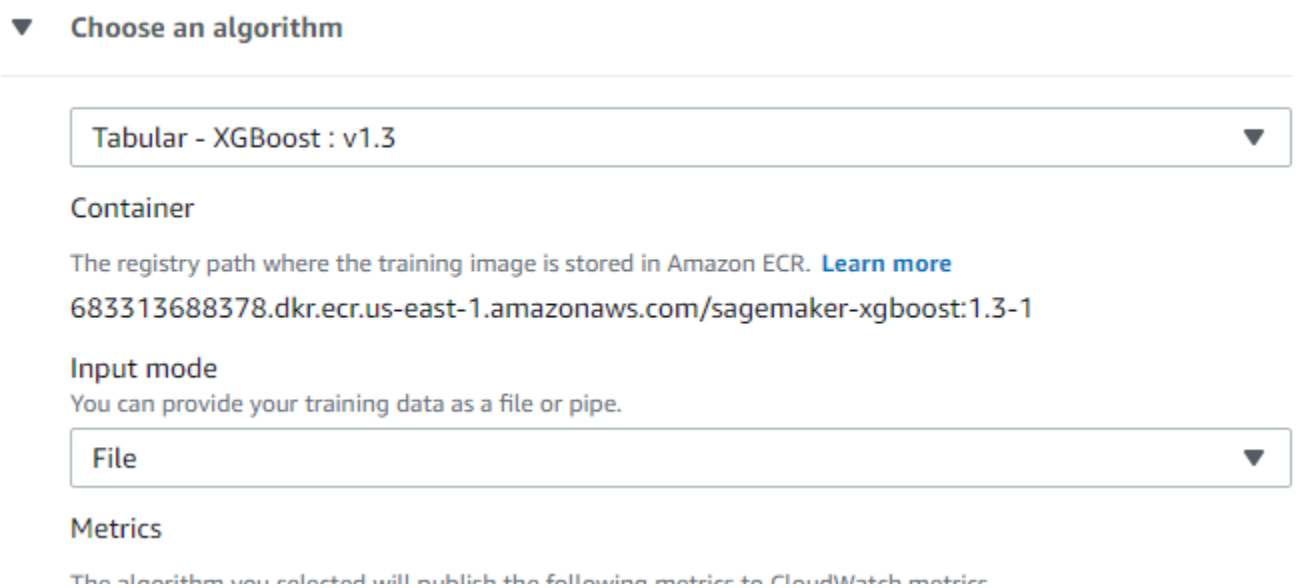
```
0.9,332.0,118.0,2.0,4.5,3.5,9.36,1.0
0.7,308.0,110.0,4.0,3.5,3.0,8.6,0.0
0.93,328.0,116.0,4.0,5.0,3.5,9.6,1.0
0.71,309.0,105.0,5.0,3.5,3.5,8.56,0.0
0.77,312.0,109.0,3.0,3.0,3.0,8.69,0.0
0.72,316.0,105.0,3.0,3.0,3.5,8.73,0.0
0.46,308.0,110.0,3.0,3.5,3.0,8.0,1.0
0.79,313.0,109.0,3.0,4.0,3.5,9.0,0.0
0.7,324.0,111.0,3.0,2.5,1.5,8.79,1.0
0.8,327.0,113.0,3.0,3.5,3.0,8.66,1.0
0.64,316.0,102.0,3.0,2.0,3.0,7.4,0.0
0.8,317.0,110.0,3.0,4.0,4.5,9.11,1.0
0.59,300.0,101.0,3.0,3.5,2.5,7.88,0.0
0.54,299.0,96.0,2.0,1.5,2.0,7.86,0.0
0.65,308.0,104.0,2.0,2.5,3.0,8.07,0.0
0.71,309.0,105.0,5.0,3.5,3.5,8.56,0.0
0.44,301.0,97.0,2.0,3.0,3.0,7.88,1.0
0.67,318.0,110.0,1.0,2.5,3.5,8.54,1.0
0.56,313.0,94.0,2.0,2.5,1.5,8.13,0.0
0.73,323.0,107.0,3.0,3.5,3.5,8.55,1.0
0.59,299.0,100.0,1.0,1.5,2.0,7.89,0.0
0.78,311.0,107.0,4.0,4.5,4.5,9.0,1.0
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

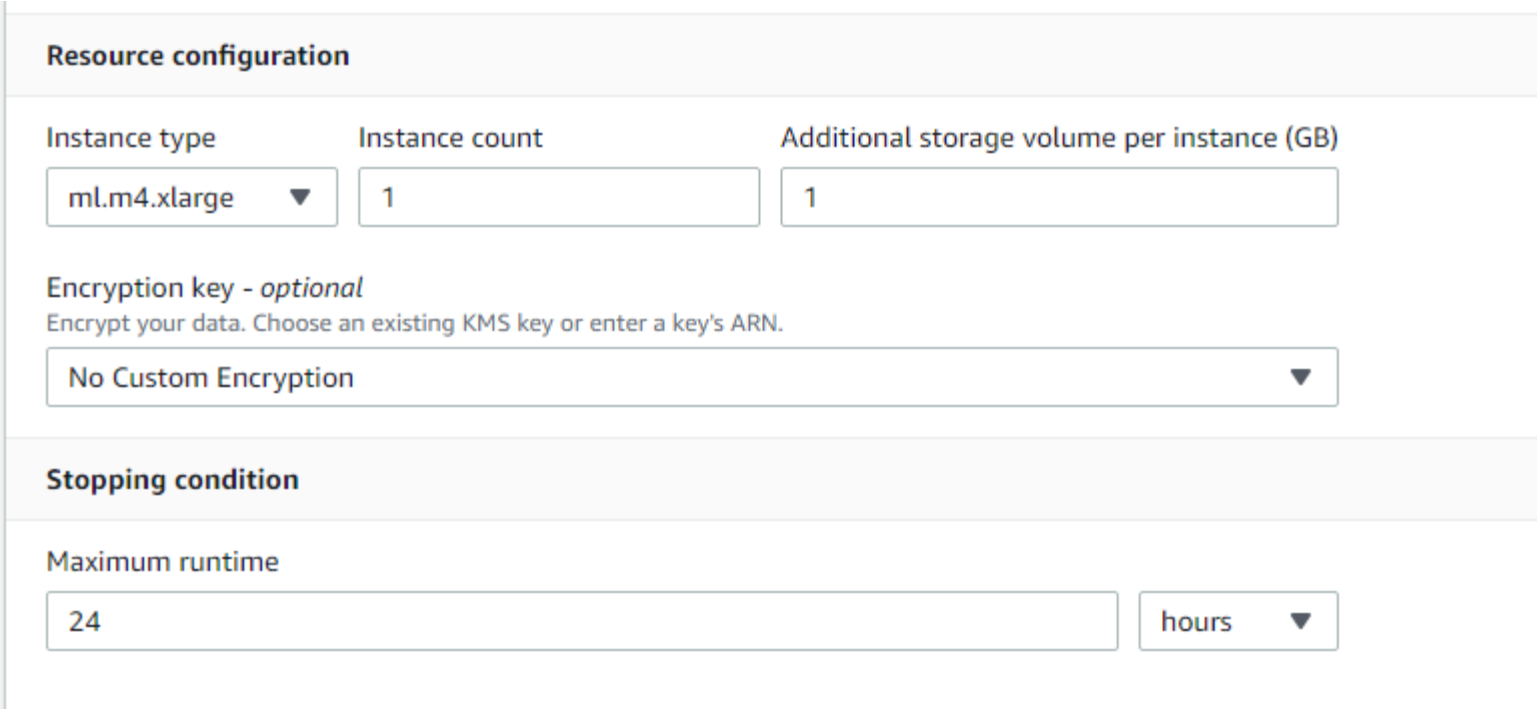


Select XGboost algorithm



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.



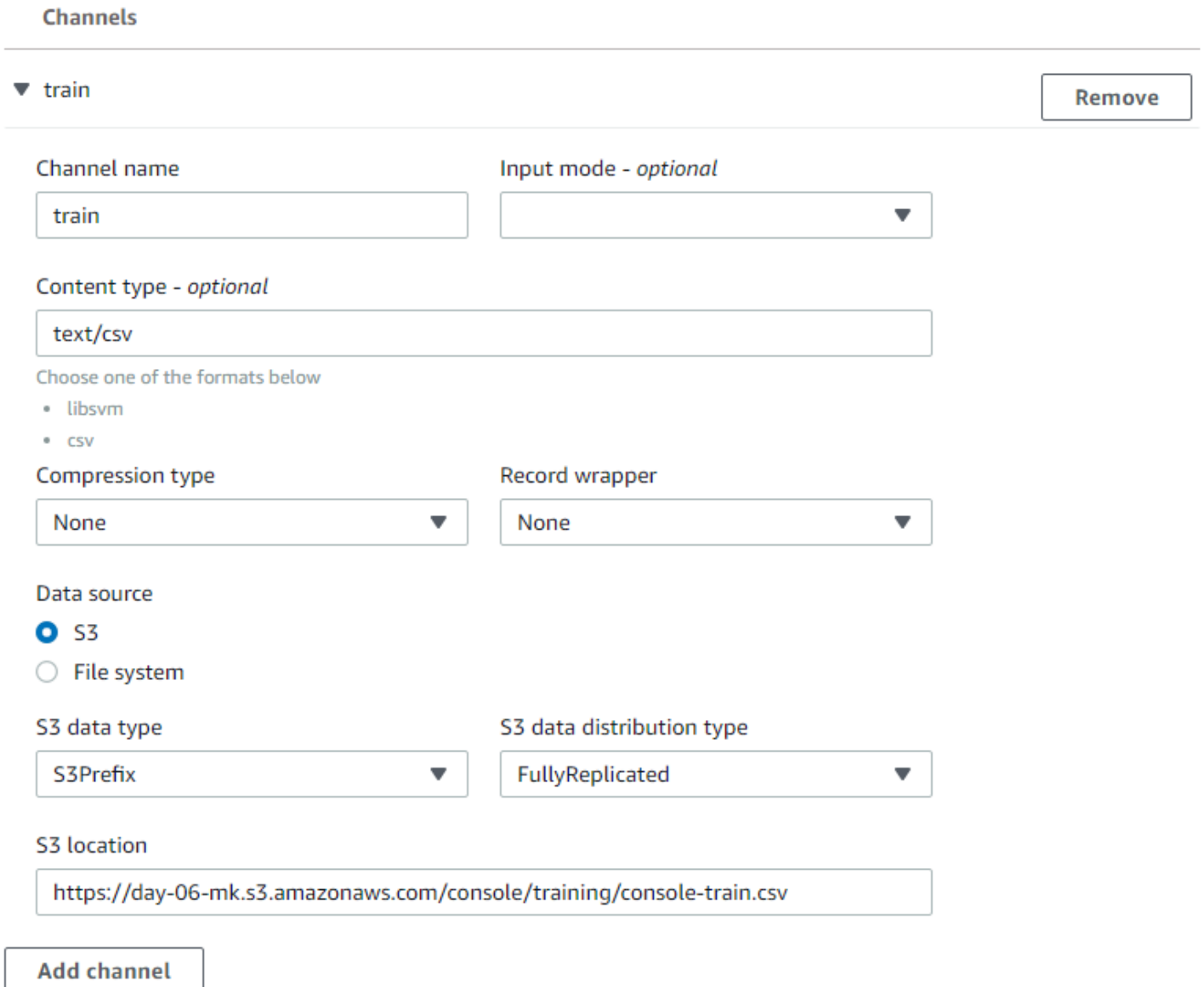
Hyperparameters

- 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

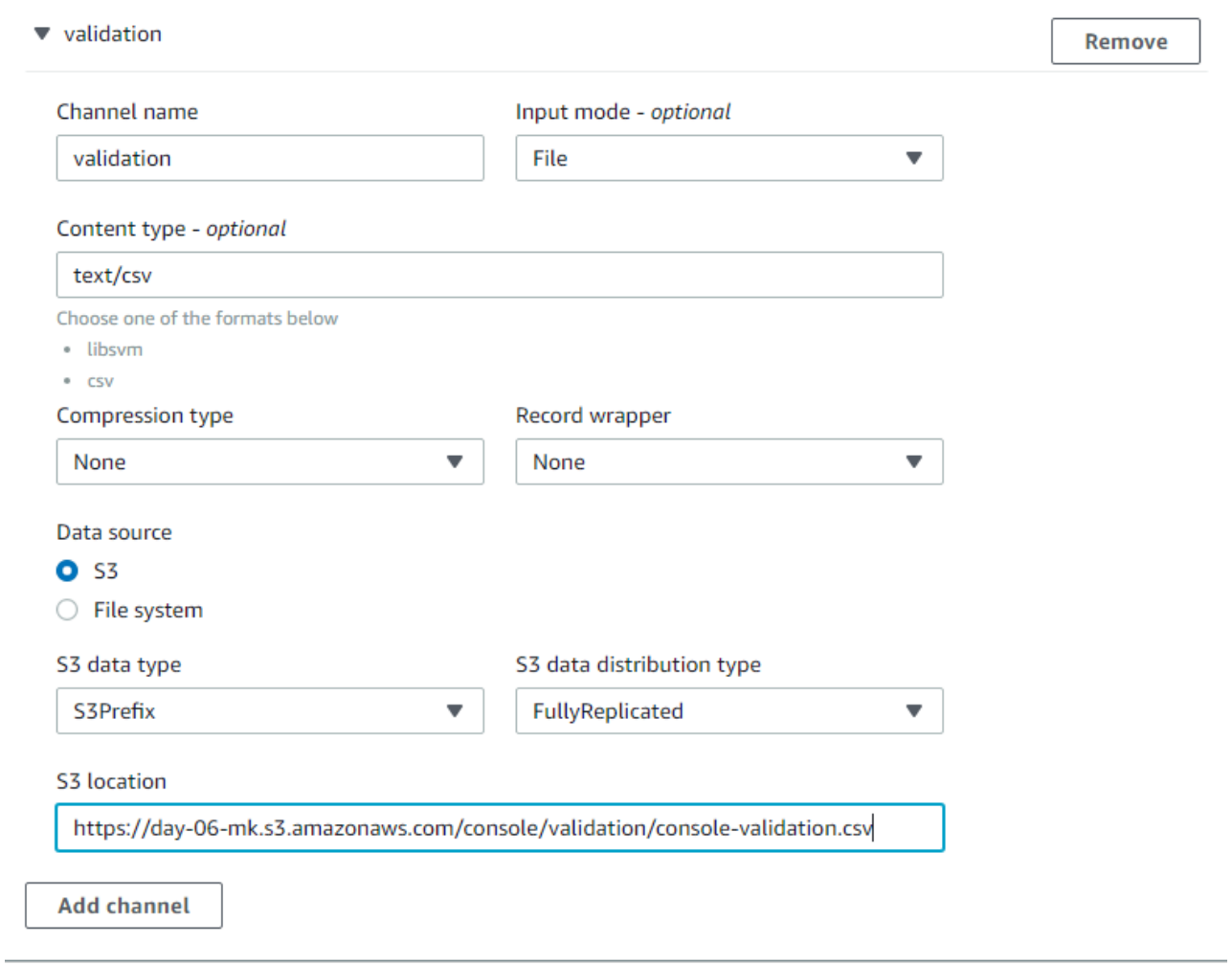
Training channel

- Set the training Channel as shown



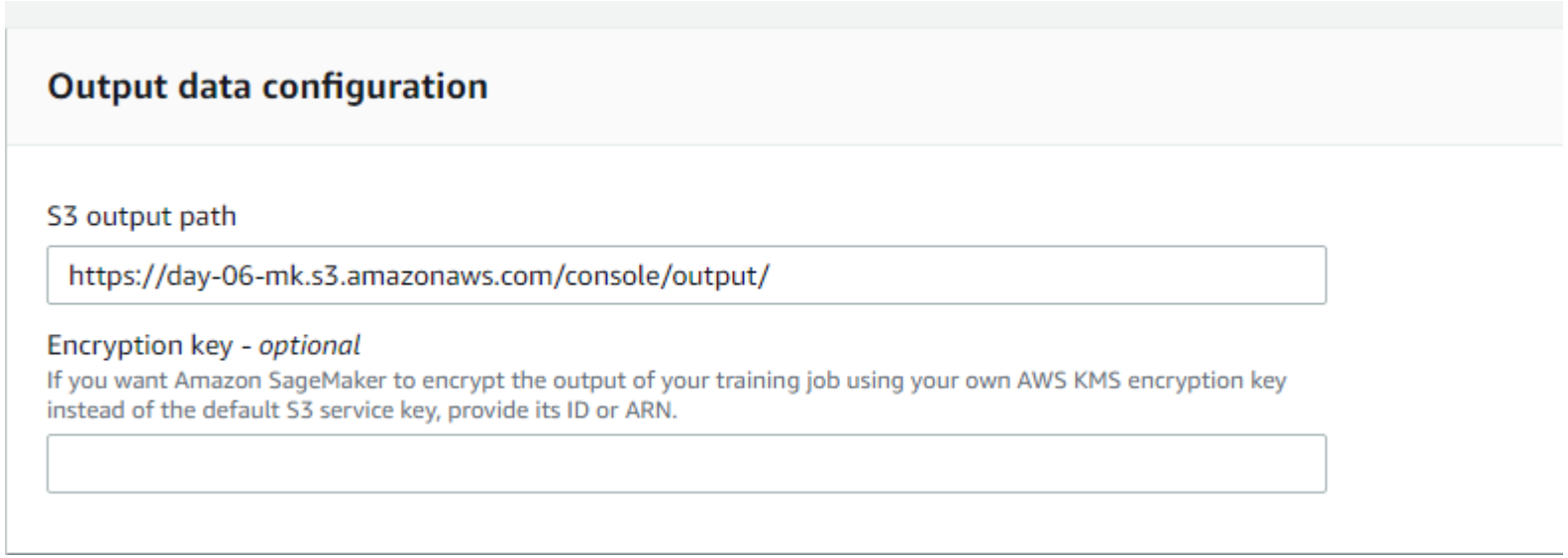
Validation channel

- Add a new validation channel



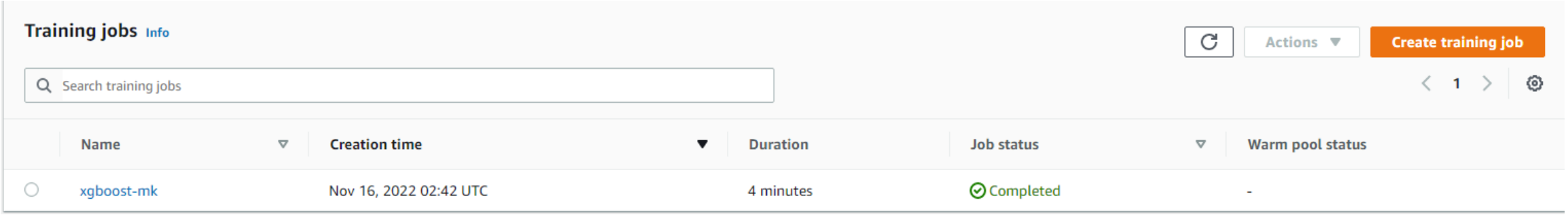
Where the model artifact will be stored

- Set the output folder

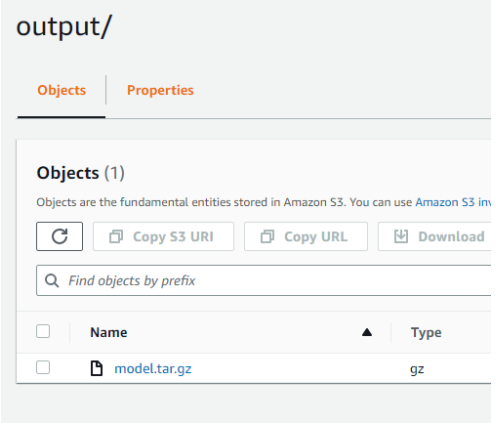


Start the training job

- Review the training job



- See the result of training job in the output folder



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.