Module

Going back to the session

Go into module 4 folder and run :

None

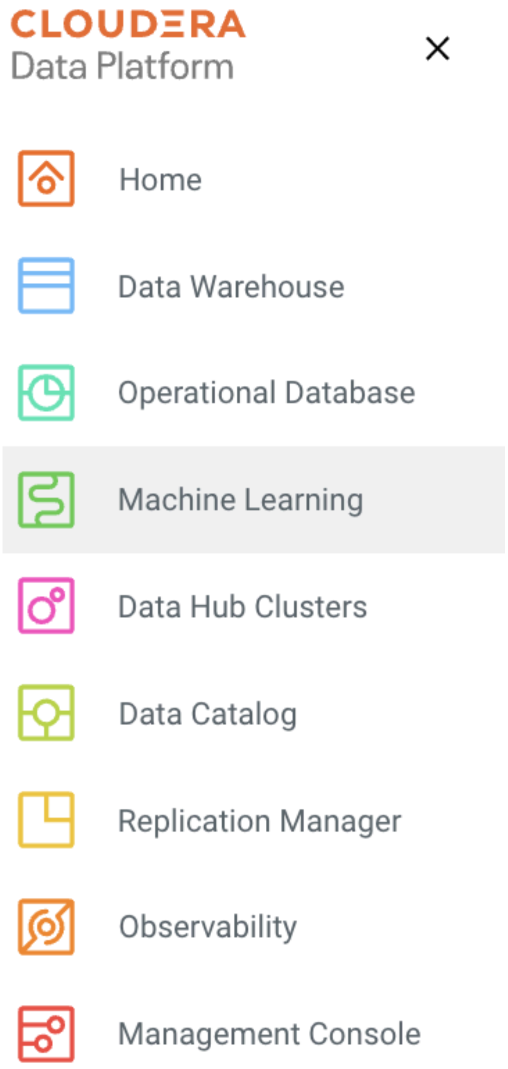
```
python create_onnx_model.py
```

What This Script Does

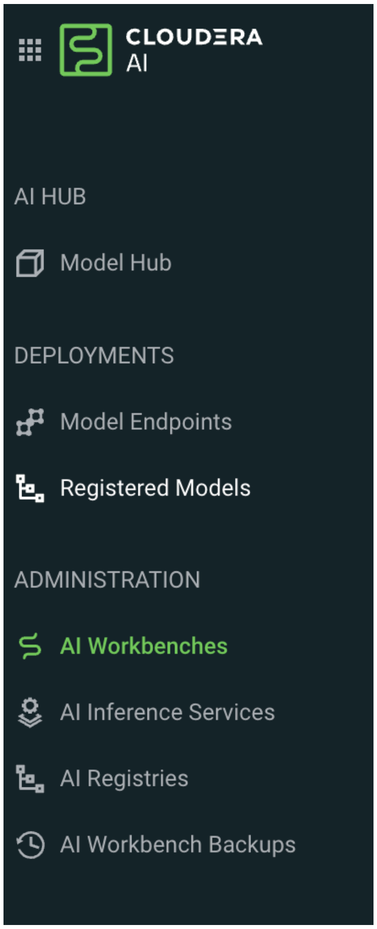
- Loads original bank marketing data plus newly labeled customer data
- Combines datasets and trains a fresh RandomForest model with improved performance
- Converts the sklearn model to ONNX format for optimized inference serving
- Registers the ONNX model in MLflow's Model Registry with proper versioning and metadata
- Makes the model ready for deployment to Cloudera AI Inference Service
- Outputs JSON file with performance metrics (F1 score, accuracy) and MLflow registration details

Ensure the script runs successfully.

Then go to the main CAI page

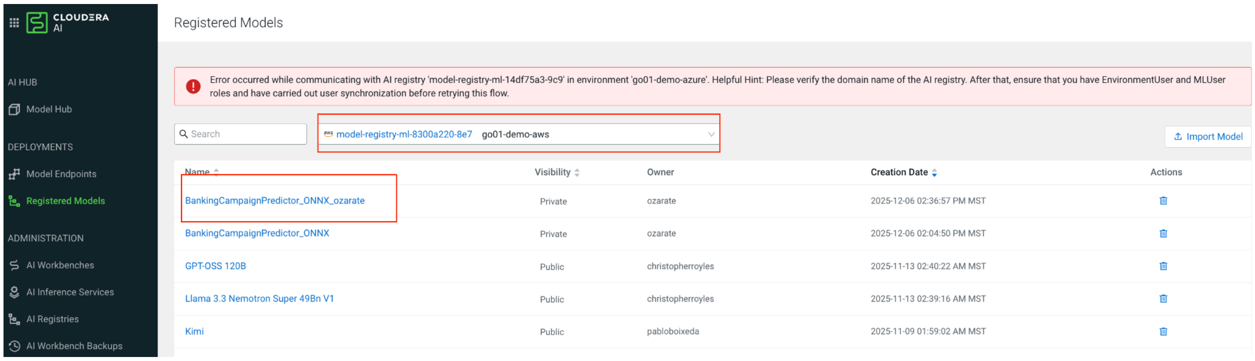


Check to make sure you model got registered (select registered models):



Once there,

1. pick the right model registry
2. Look for your model (identified w/ you username)



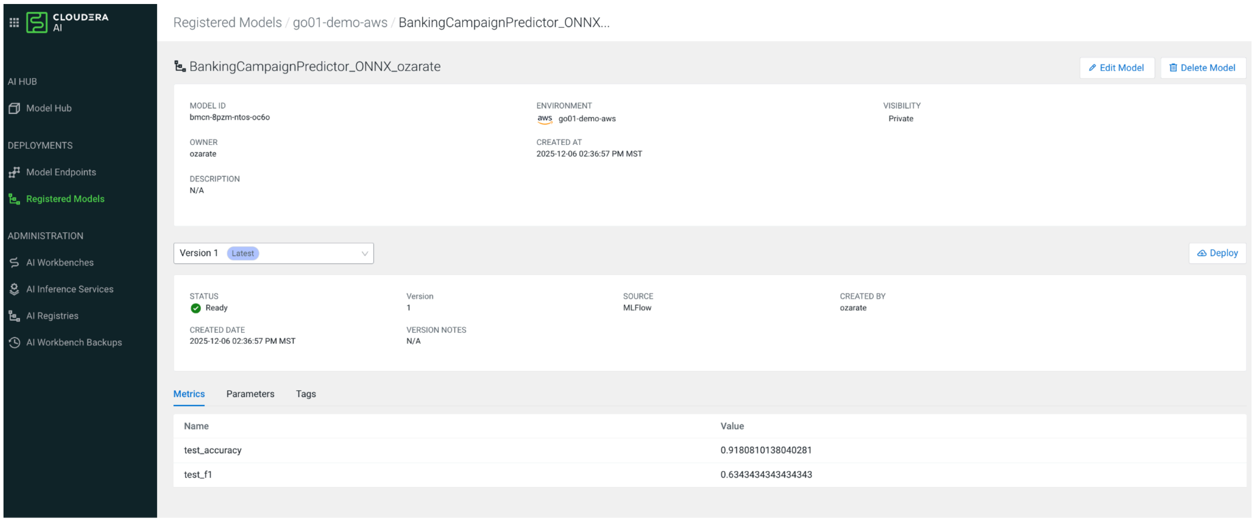
Registered Models

Error occurred while communicating with AI registry 'model-registry-ml-14df75a3-9c9' in environment 'go01-demo-azure'. Helpful Hint: Please verify the domain name of the AI registry. After that, ensure that you have EnvironmentUser and MLUser roles and have carried out user synchronization before retrying this flow.

model-registry-ml-8300a220-8e7 go01-demo-aws

Name	Visibility	Owner	Creation Date	Actions
BankingCampaignPredictor_ONNX_ozarate	Private	ozarate	2025-12-06 02:36:57 PM MST	🗑
BankingCampaignPredictor_ONNX	Private	ozarate	2025-12-06 02:04:50 PM MST	🗑
GPT-OSS 120B	Public	christopherroyles	2025-11-13 02:40:22 AM MST	🗑
Llama 3.3 Nemotron Super 49Bn V1	Public	christopherroyles	2025-11-13 02:39:16 AM MST	🗑
Kimi	Public	pabloboixeda	2025-11-09 01:59:02 AM MST	🗑

Click on your model and you can see some important model information such as model id, metrics, parameters, model version and more.



Registered Models / go01-demo-aws / BankingCampaignPredictor_ONNX...

BankingCampaignPredictor_ONNX_ozarate

MODEL ID: bmcn-8pzm-ntos-oc6o

ENVIRONMENT: aws go01-demo-aws

VISIBILITY: Private

OWNER: ozarate

CREATED AT: 2025-12-06 02:36:57 PM MST

DESCRIPTION: N/A

Version 1 Latest

STATUS: Ready

Version: 1

SOURCE: MLFlow

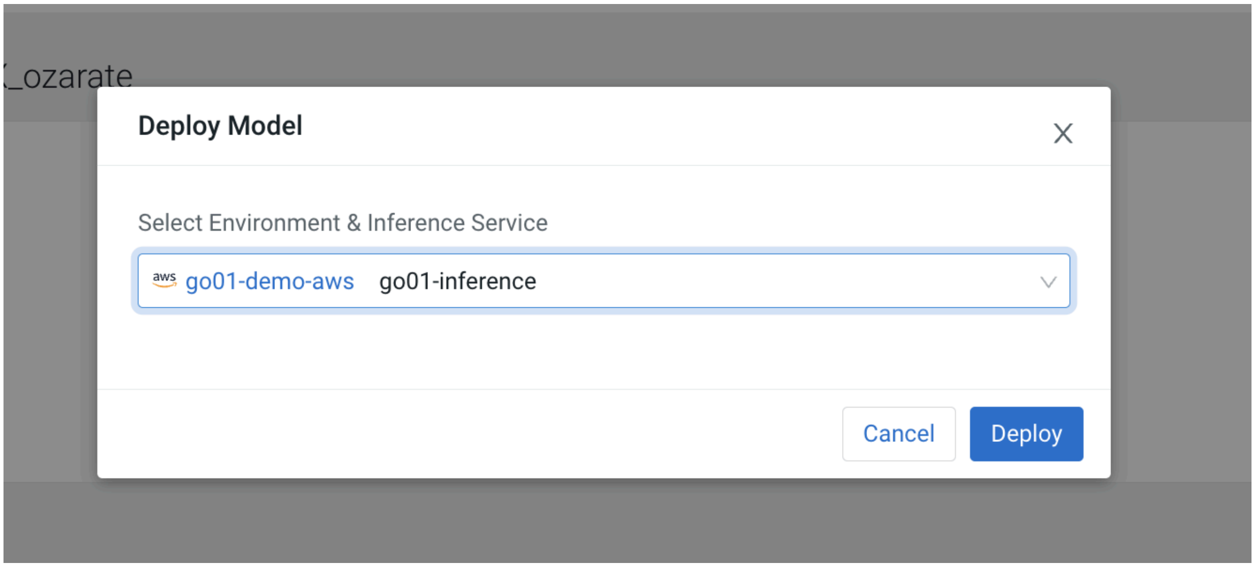
CREATED BY: ozarate

CREATED DATE: 2025-12-06 02:36:57 PM MST

VERSION NOTES: N/A

Name	Value
test_accuracy	0.9180810138040281
test_f1	0.6343434343434343

Go ahead and click on the deploy button. And select the environment and inference services.

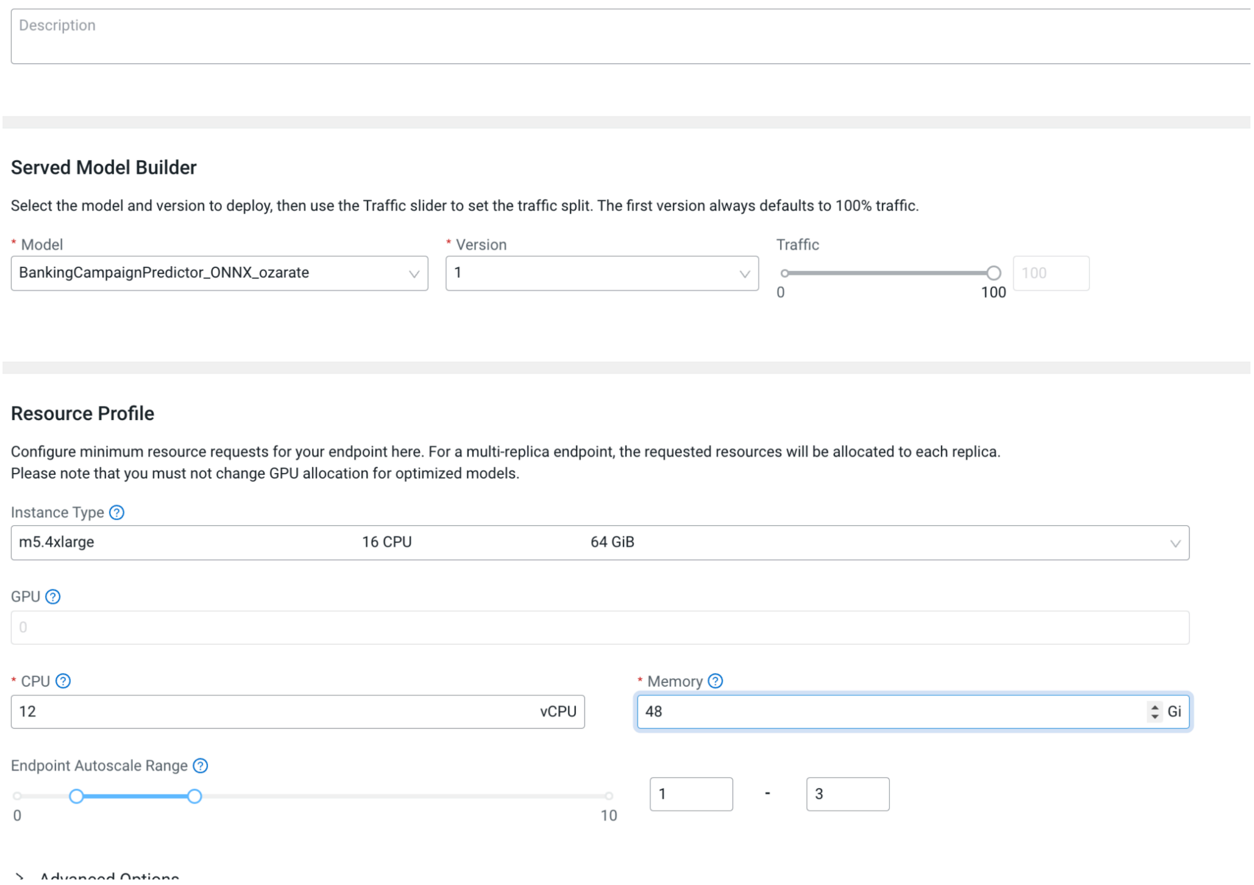


Deploy Model

Select Environment & Inference Service

aws go01-demo-aws go01-inference

Fill in the following information - name “bank-marketing-campaign-your user name” and pick the instance type show below. Select 12 vcpus and 48 GB of ram for memory :



Description

Served Model Builder

Select the model and version to deploy, then use the Traffic slider to set the traffic split. The first version always defaults to 100% traffic.

* Model: BankingCampaignPredictor_ONNX_ozarate

* Version: 1

Traffic: 100

Resource Profile

Configure minimum resource requests for your endpoint here. For a multi-replica endpoint, the requested resources will be allocated to each replica.
Please note that you must not change GPU allocation for optimized models.

Instance Type: m5.4xlarge 16 CPU 64 GiB

GPU: 0

* CPU: 12 vCPU

* Memory: 48 Gi

Endpoint Autoscale Range: 1 - 3