

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



Home



Data Warehouse



Operational Database



Machine Learning



Data Hub Clusters



Data Catalog



Replication Manager

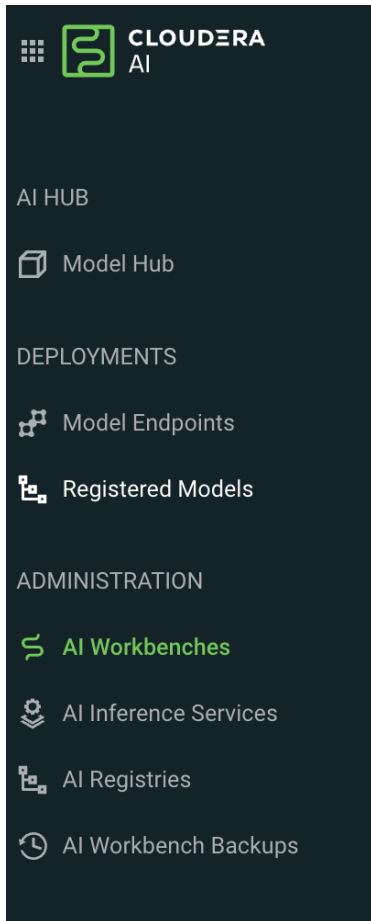


Observability



Management Console

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)

A screenshot of the 'Registered Models' page in CloudEra AI. The page shows a table of registered models with columns: Name, Visibility, Owner, Creation Date, and Actions. The first two rows are highlighted with red boxes. The first row is 'BankingCampaignPredictor_ONNX_ozarate' and the second row is 'BankingCampaignPredictor_ONNX'. A red box also highlights the search bar and the registry dropdown menu. A red error message is visible at the top of the page.

Registered Models

Error occurred while communicating with AI registry 'model-registry-ml-14df75a3-9c9' in environment 'go01-demo-aws'. 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.

Search model-registry-ml-8300a220-8e7 go01-demo-aws [Import Model](#)

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

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

The screenshot shows the Cloudera AI console interface. On the left is a dark sidebar with navigation links: AI HUB, Model Hub, DEPLOYMENTS, Model Endpoints, Registered Models (highlighted), ADMINISTRATION, AI Workbenches, AI Inference Services, AI Registries, and AI Workbench Backups. The main content area is titled 'Registered Models / go01-demo-aws / BankingCampaignPredictor_ONNX...'. It displays details for the model 'BankingCampaignPredictor_ONNX_ozarate'. At the top right are 'Edit Model' and 'Delete Model' buttons. The model details include: MODEL ID (bmcm-8pzm-ritos-ocfo), ENVIRONMENT (aws go01-demo-aws), VISIBILITY (Private), OWNER (ozarate), CREATED AT (2025-12-06 02:36:57 PM MST), and DESCRIPTION (N/A). Below this is a 'Version 1' dropdown with a 'Latest' button and a 'Deploy' button. A table shows the model's status as 'Ready' (with a green checkmark), Version 1, Source as 'MLFlow', and Created by 'ozarate'. The 'CREATED DATE' is '2025-12-06 02:36:57 PM MST' and 'VERSION NOTES' are 'N/A'. At the bottom, there are tabs for 'Metrics', 'Parameters', and 'Tags'. The 'Metrics' tab is active, showing a table with columns 'Name' and 'Value':

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.

The screenshot shows a 'Deploy Model' dialog box with a close button (X) in the top right corner. The dialog has a title bar 'Deploy Model' and a subtitle 'Select Environment & Inference Service'. Below the subtitle is a dropdown menu showing 'aws go01-demo-aws go01-inference'. At the bottom right of the dialog are two buttons: 'Cancel' and 'Deploy'.

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

0

100

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

m5.4xlarge16 CPU64 GiB

GPU

0

* CPU

12vCPU

* Memory

48Gi

Endpoint Autoscale Range

0

10

1

-

3

Advanced Options