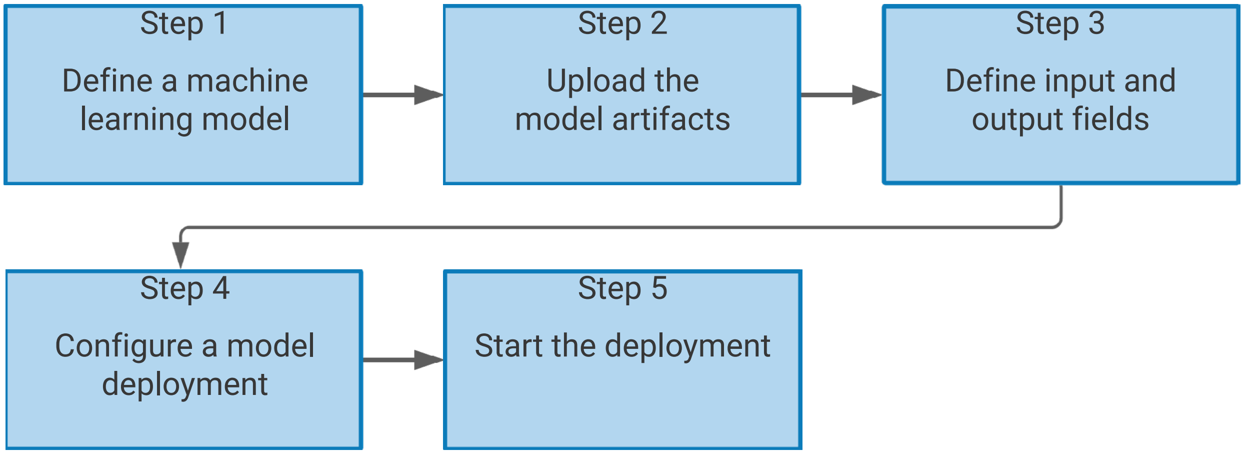
Model Serve Quick Start Guide

Model Serve provides a fully-managed platform to deploy and manage your data science and artificial intelligence models. You upload your machine learning model, and Model Serve deploys the model in a cloud environment. You can then make API requests to a URL endpoint to get predictions from the model.

Use this quick start guide to register and deploy a machine learning model:



Step 1
Define a machine learning model

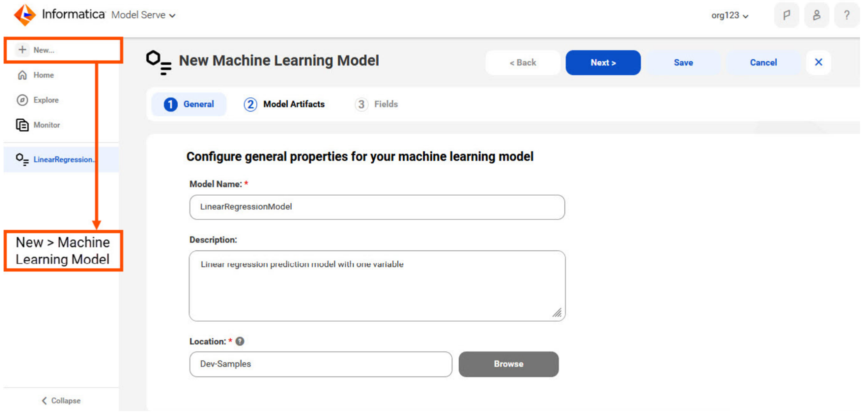
Step 2
Upload the model artifacts

Step 3
Define input and output fields

Step 4
Configure a model deployment

Step 5
Start the deployment

Define a Machine Learning Model



New > Machine Learning Model

Configure general properties for your machine learning model

Model Name: * LinearRegressionModel

Description: Linear regression prediction model with one variable

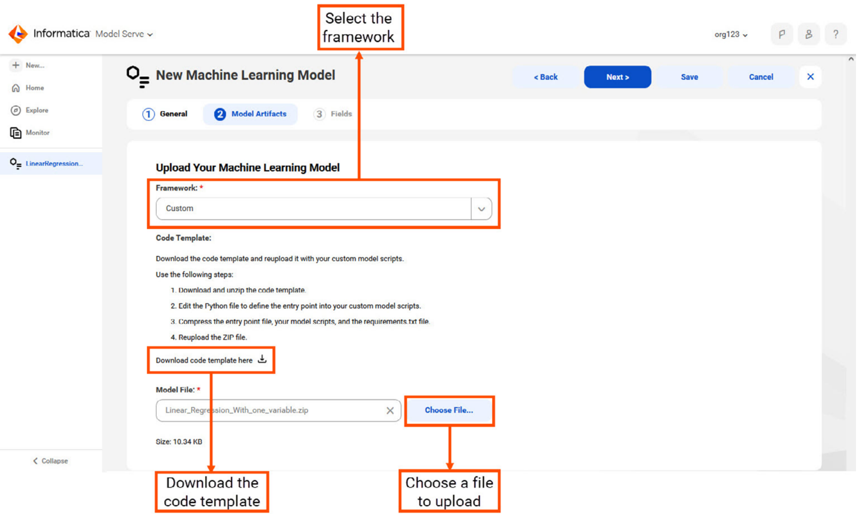
Location: * Dev-Samples

1. In the navigation bar, select **New > Machine Learning Model** to create a machine learning model.

2. Configure the name, description, and location of the model.

Learn more about defining a machine learning model.

Upload the Model Artifacts



New Machine Learning Model

Upload Your Machine Learning Model

Framework: * Custom

Code Template:

Download the code template and reupload it with your custom model scripts.

Use the following steps:

1. Download and unzip the code template.
2. Edit the Python file to define the entry point into your custom model scripts.
3. Compress the entry point file, your model scripts, and the requirements.txt file.
4. Reupload the ZIP file.

Download code template here

Model File: * Linear_Regression_With_one_variable.zip

Size: 10.34 KB

Select the framework

Download the code template

Choose a file to upload

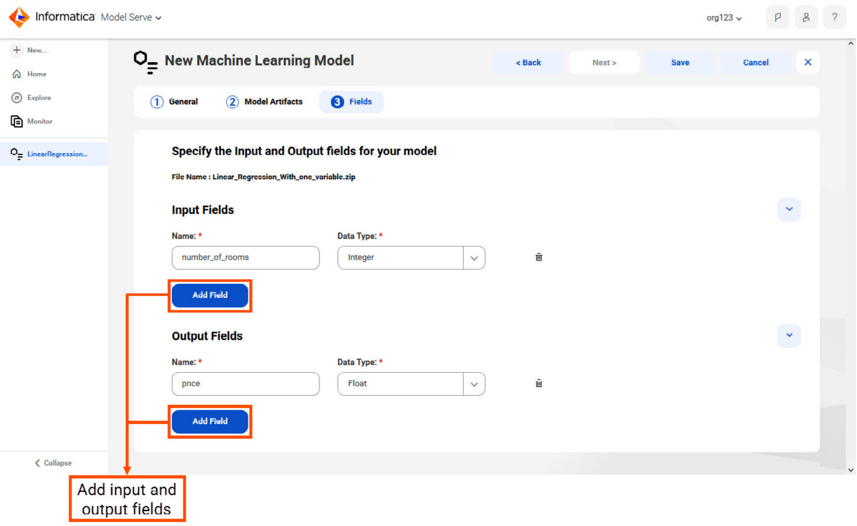
1. Select the machine learning framework that you used to build the model.

2. If you're using a custom model, download and edit the code template to define the entry point to your custom files.

3. Upload a ZIP file containing the model artifacts, including the code template, if applicable.

[Learn more about uploading the model artifacts.](#)

Define Input and Output Fields



New Machine Learning Model

Specify the Input and Output fields for your model

File Name : Linear_Regression_With_one_variable.zip

Input Fields

Name: * number_of_rooms

Data Type: * Integer

Output Fields

Name: * price

Data Type: * Float

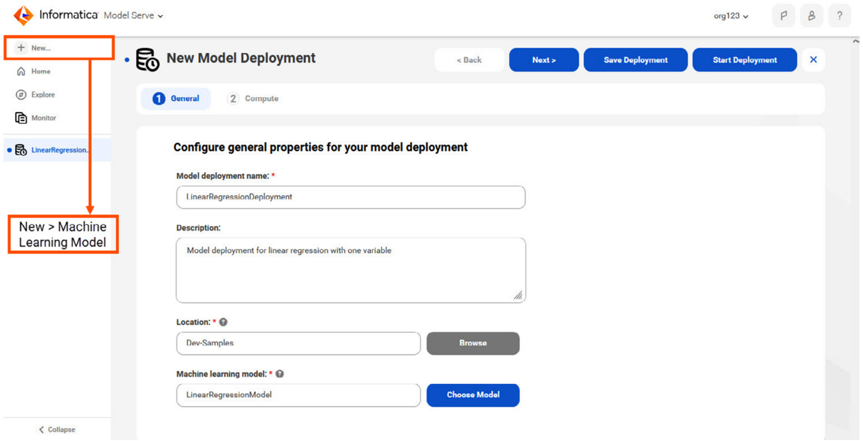
Add input and output fields

Define input and output fields. Configure the name and data type of each input that the model expects and output that the model returns.

Then save the machine learning model.

[Learn more about defining input and output fields.](#)

Configure a Model Deployment



Configure general properties for your model deployment

Model deployment name: * LinearRegressionDeployment

Description: Model deployment for linear regression with one variable

Location: * Dev-Samples

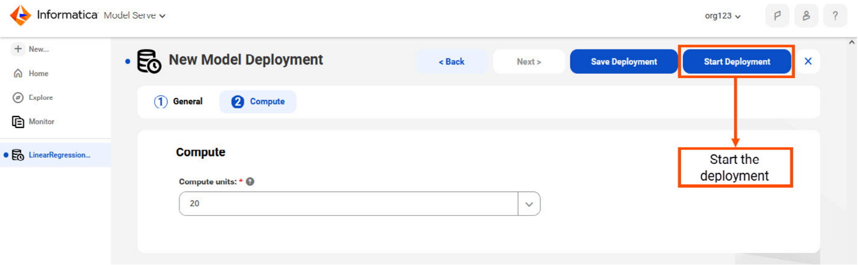
Machine learning model: * LinearRegressionModel

New > Machine Learning Model

1. In the navigation bar, select **New > Model Deployment** to create a model deployment.
2. On the General tab, configure the name, description, and location of the deployment. Also, select the model that you'll deploy.
3. On the Compute tab, configure the maximum number of compute units that the model deployment can use.

[Learn more about configuring a model deployment.]

Start the Deployment



Compute

Compute units: * 20

Start the deployment

After you save a valid model deployment, you can start the deployment. Model Serve provisions the cloud resources and deploys your model.

[Learn more about starting a model deployment.]

Next Steps

1. Send a sample request to [test that the model is generating predictions as you expect].
2. In your own application, add an API call that sends requests to the URL endpoint of the deployed model.
3. [Monitor the model deployment] and stop it when it's not in use.