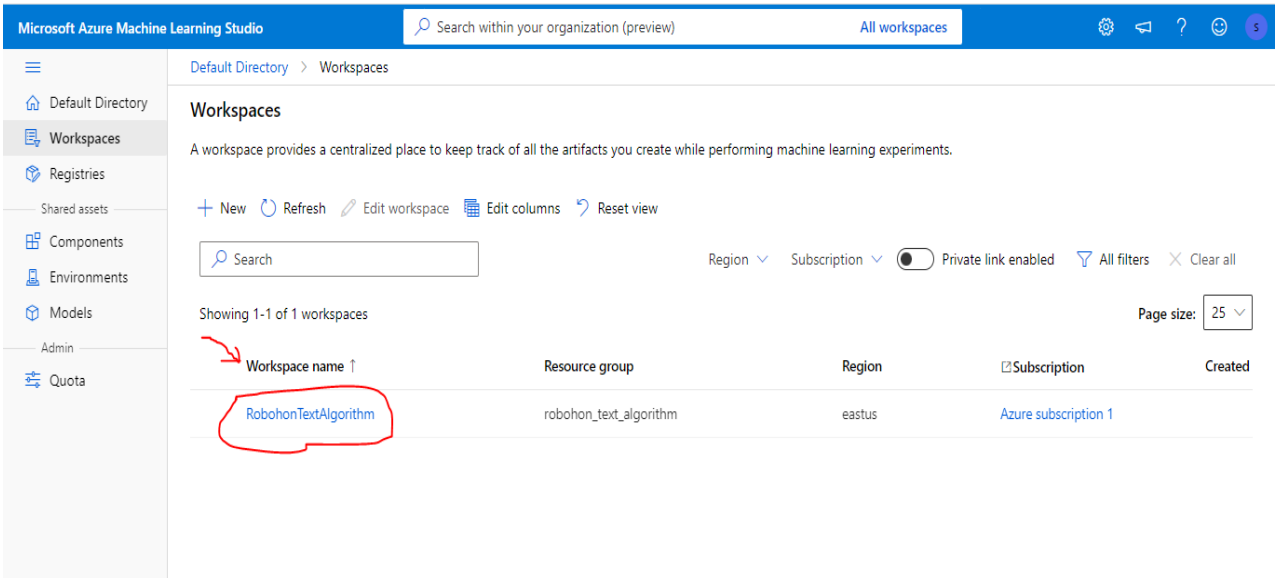
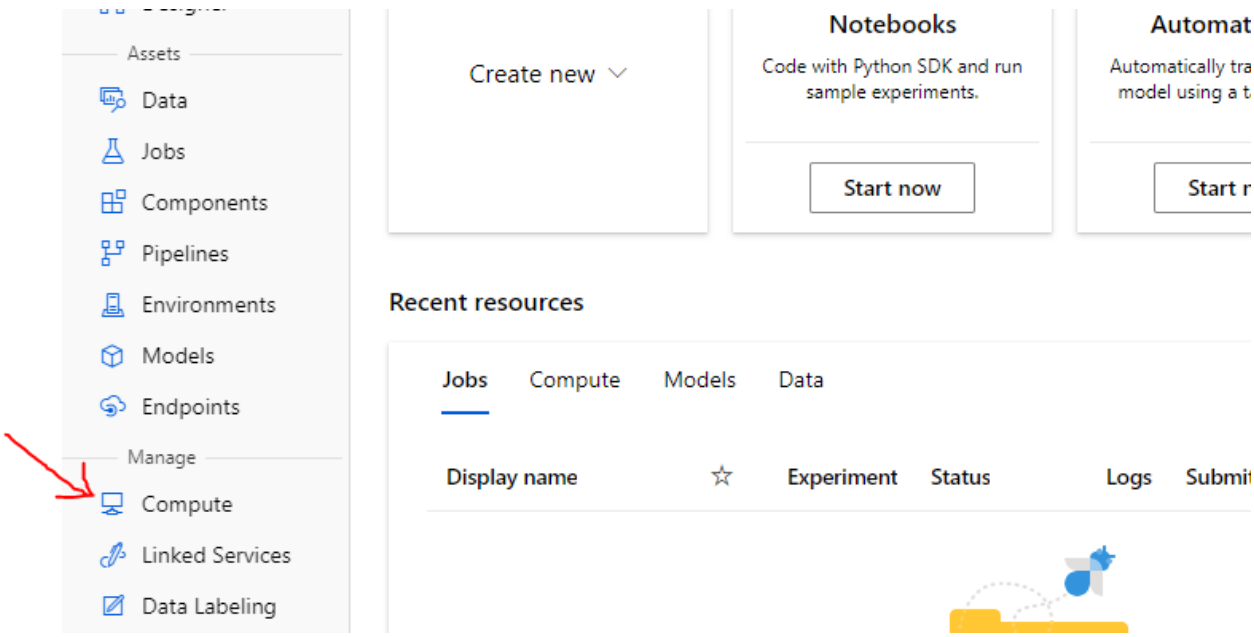
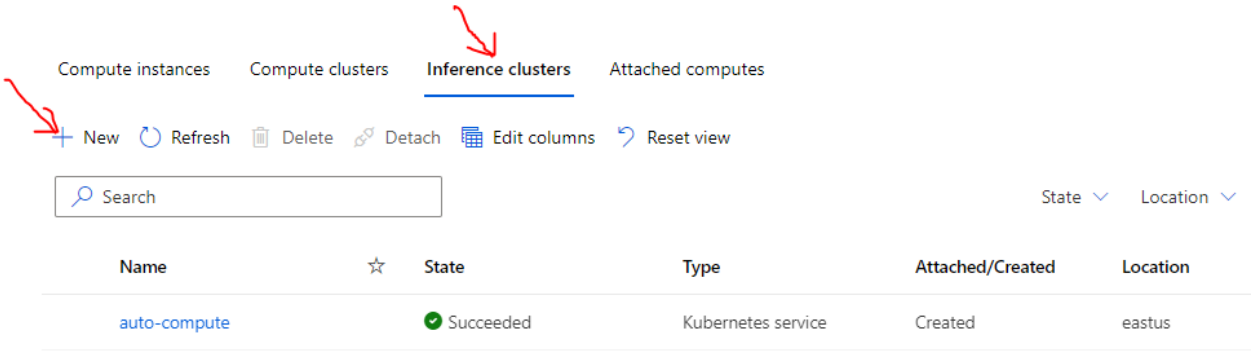
Step 1: Create a workspace in Azure Machine learning, wait till the workspace is created and then click on it.



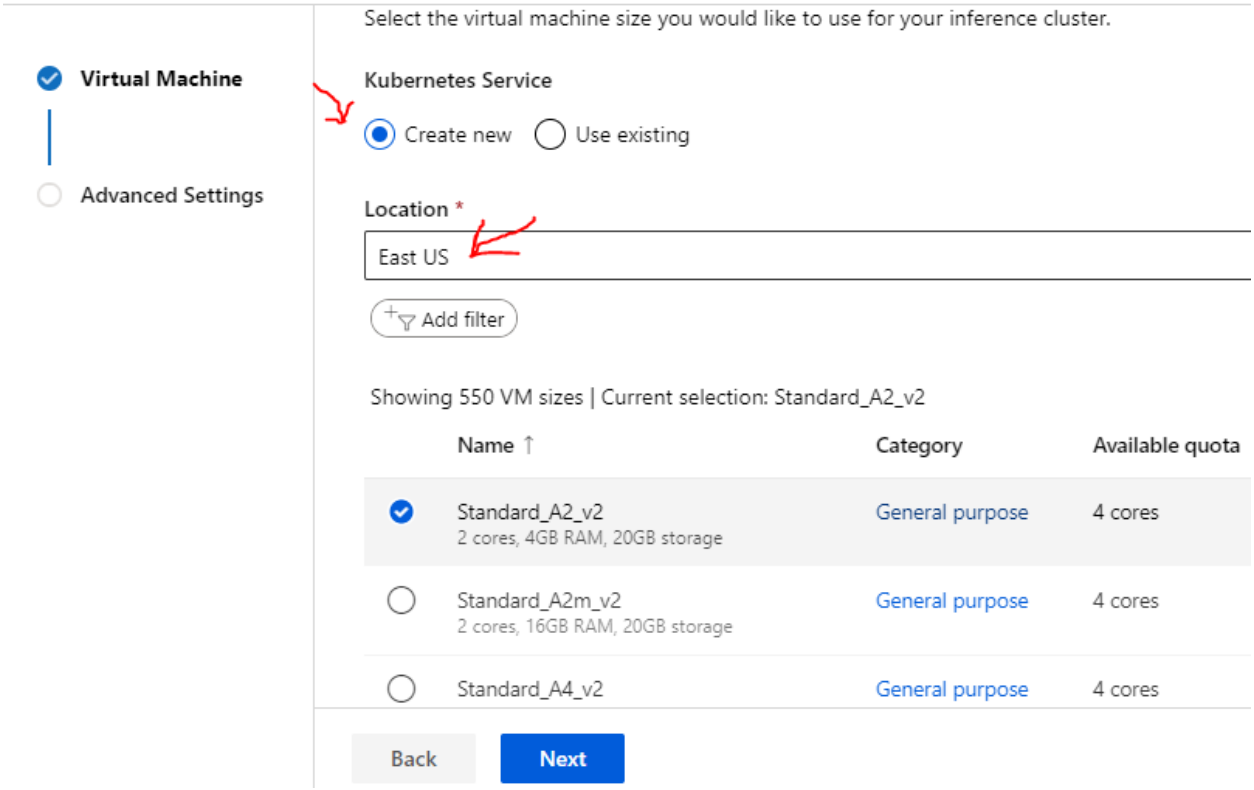
Step 2: Inside that workspace, click on the “Compute” icon.



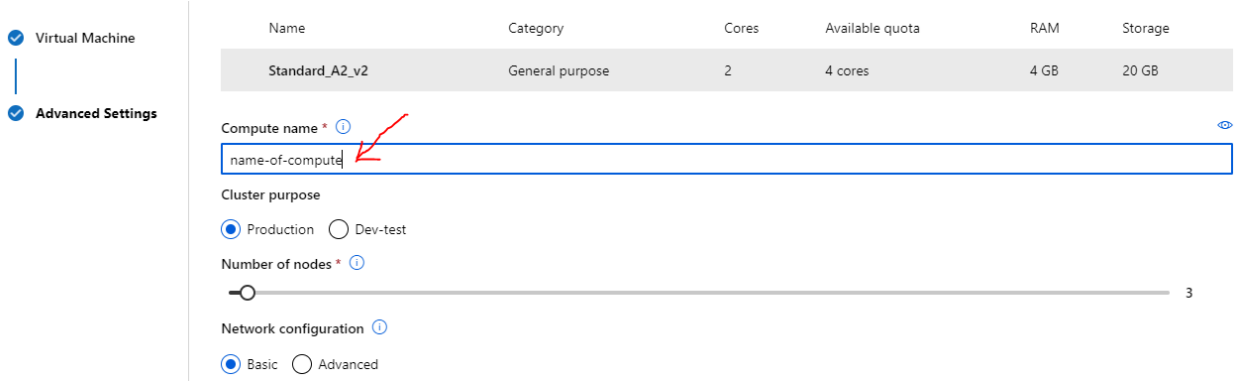
Step 3: Click on “inference clusters” and then click on “New”.



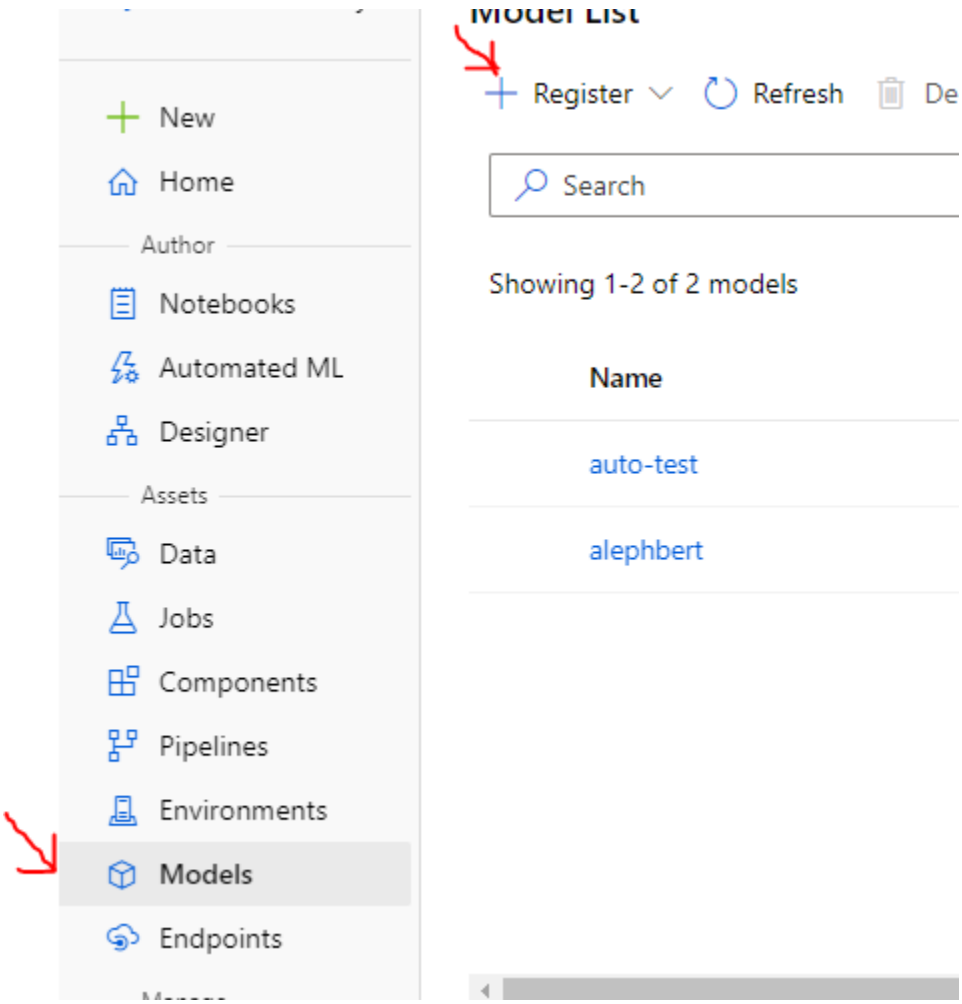
Step 4: Click on “Create new” and then fill the respective fields and click on next button



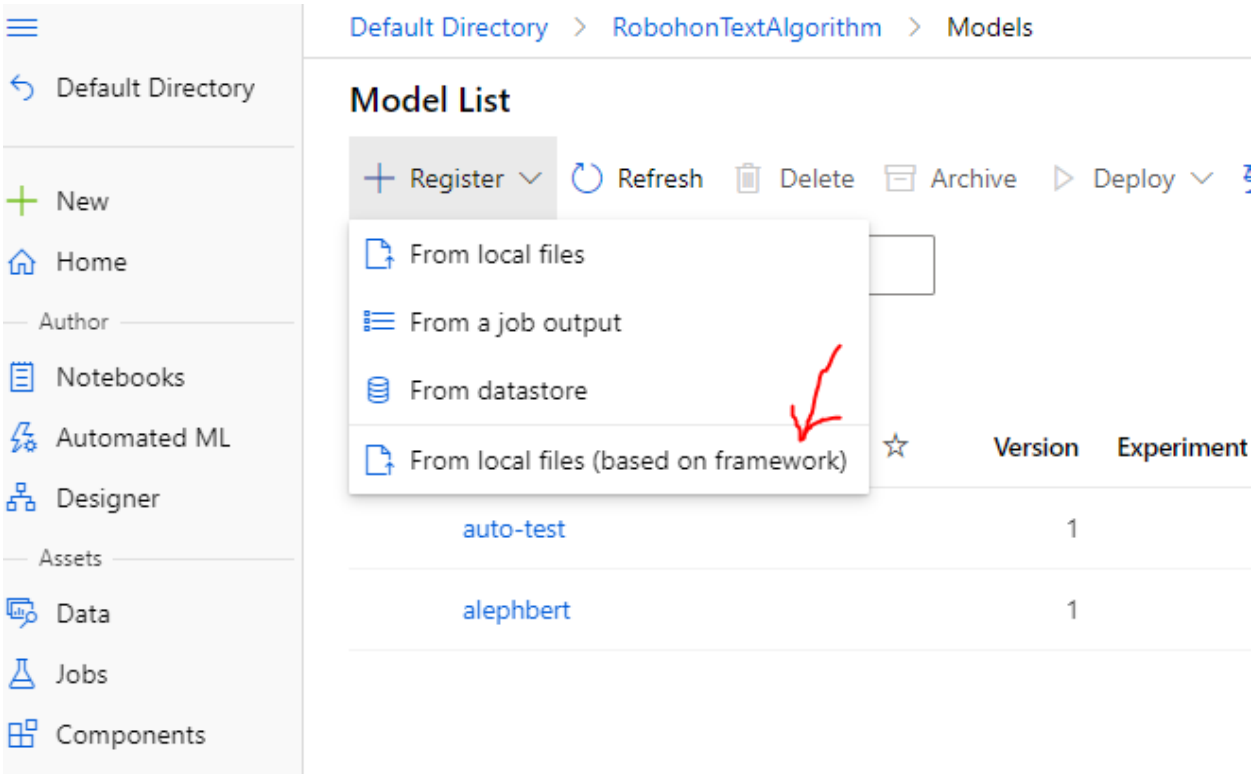
Step 5: Just fill these fields as well and then click on create. This will create the inference cluster.



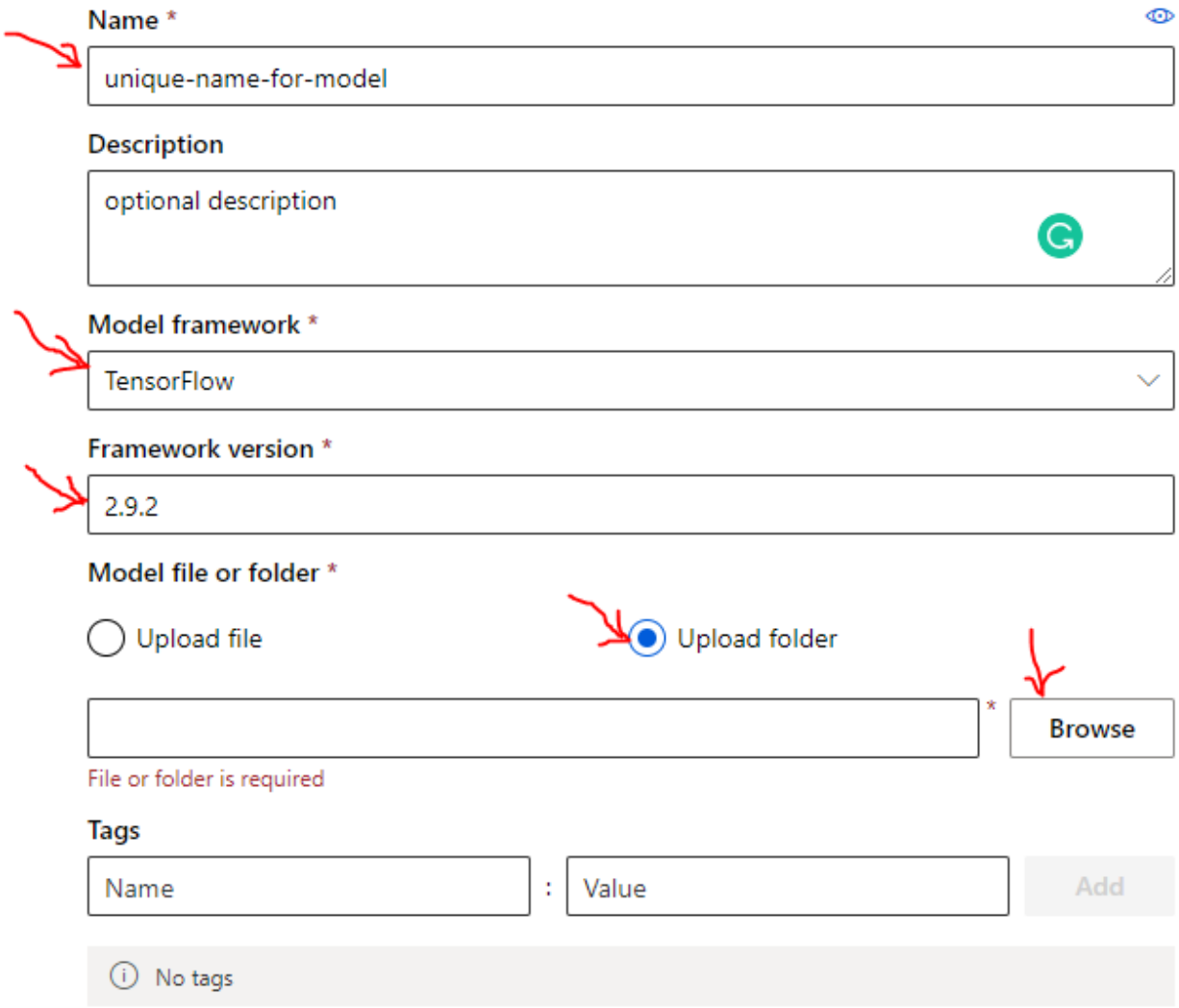
Step 6: After that, click on the “Model” and then “Register” to register the trained model.



Step 7: Click the option directed by the red arrow.

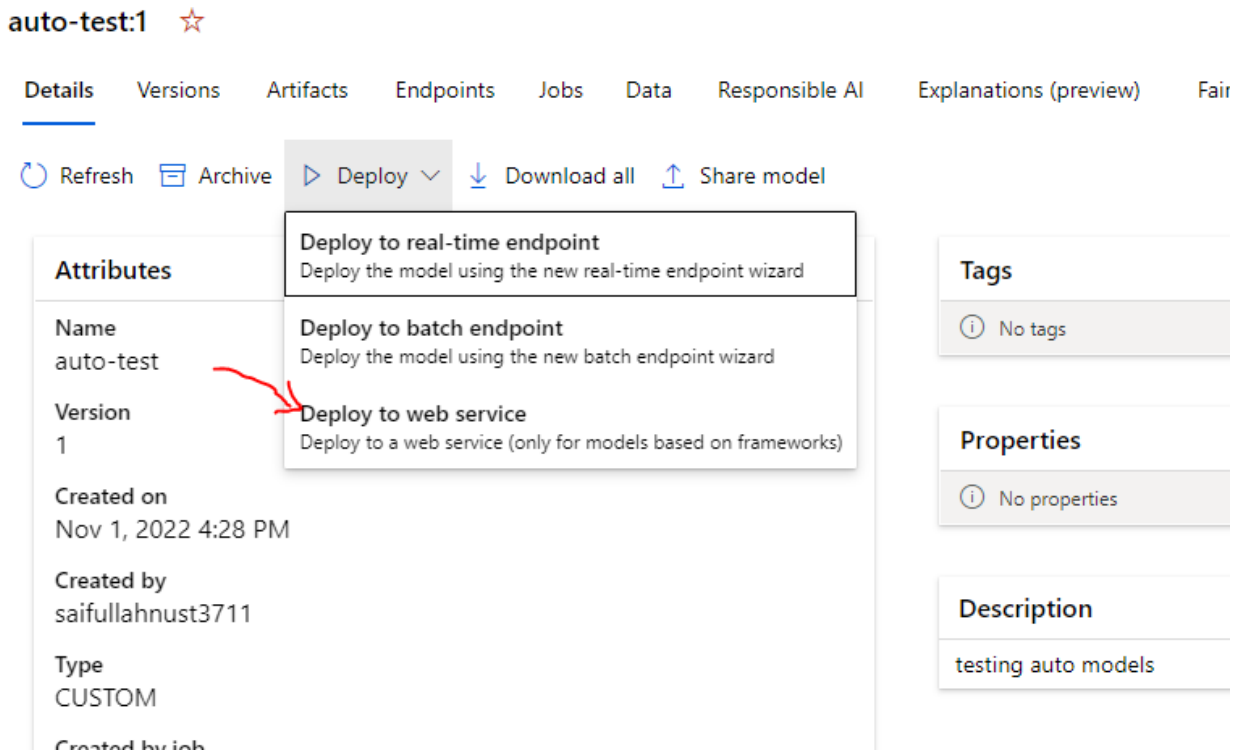


Step 8: Fill the respective fields as directed

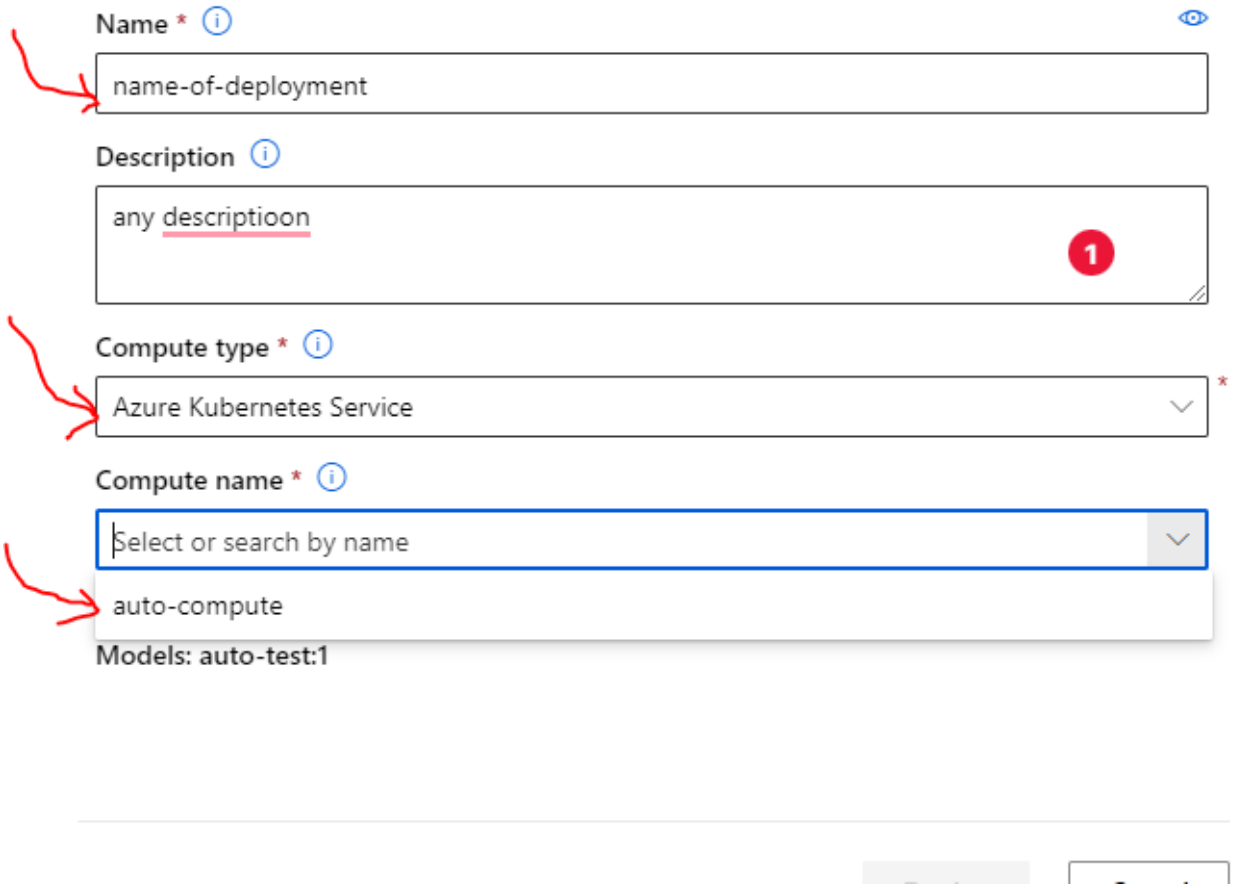


Step 9: When the model is registered successfully. This will take some time (depends on the model size)

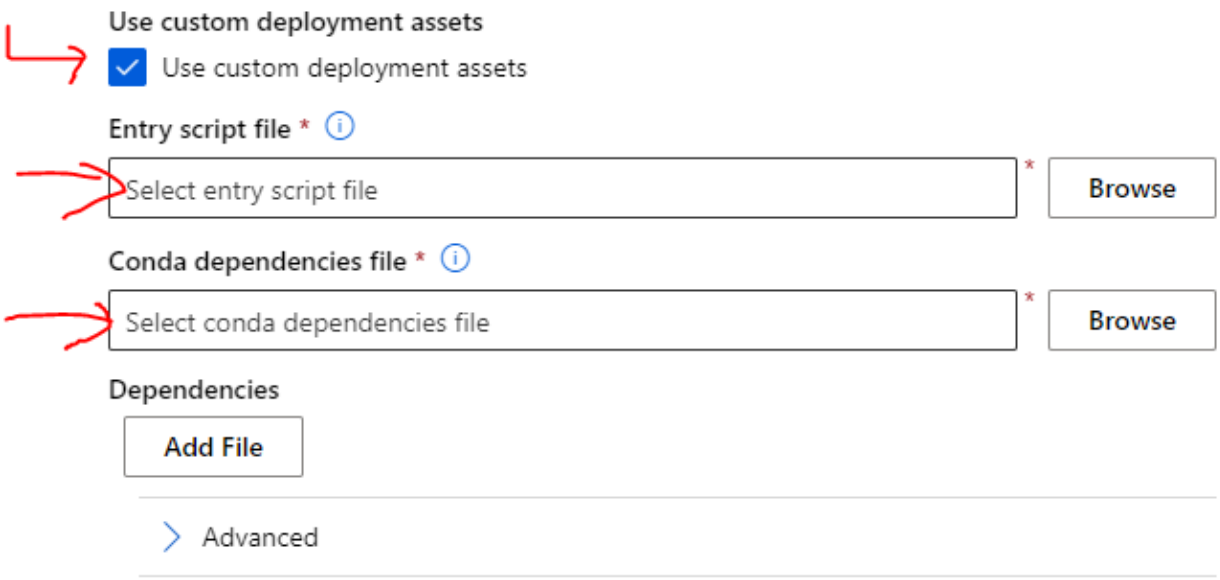
Step 10: Click on Deploy and then Deploy to Web Services as directed by the arrow.



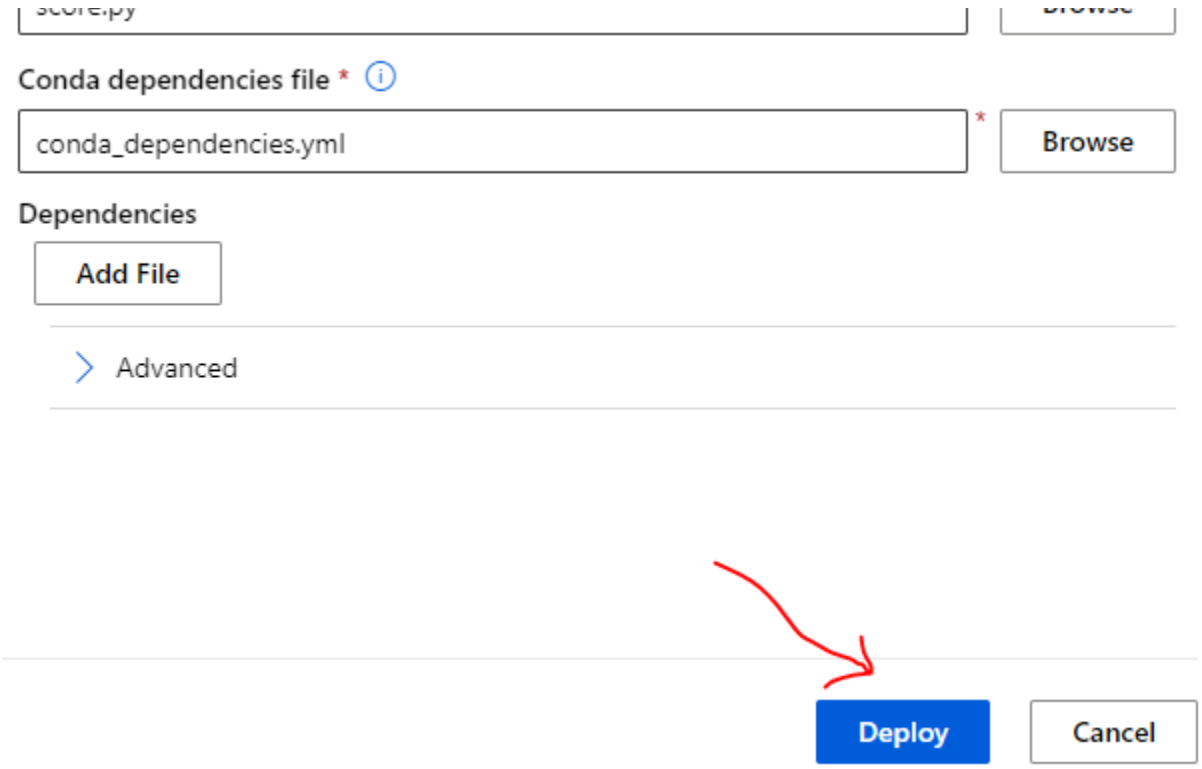
Step 11: Fill the respective fields for the model deployment for real-time inference.



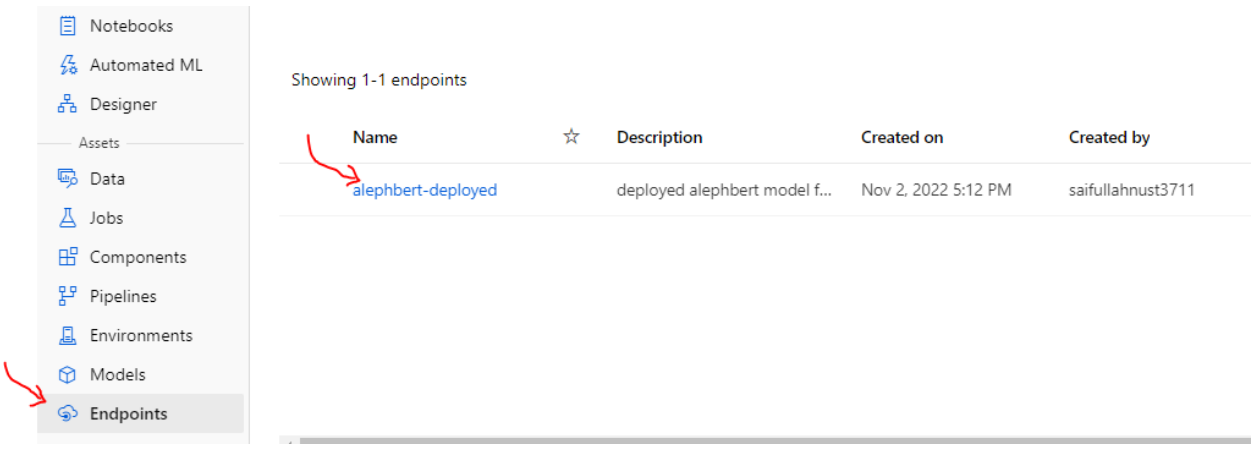
Step 12: Here insert the model scoring script script.py and conda_dependencies.yml. Both of these are available on the Git Hub.



Step 13: Click on the Deploy button. Wait till the model is deployed.



Step 14: Check the endpoints for the deployed model endpoints. Click on the deployed model.



Step 15: URL is here which can be used for real-time inferencing.