Machine Learning Implementations and Operations Quiz 1

Model deployed in a container in an EC2 instance

A docker container that contains the end user's chosen model and necessary scripts must be prepared prior to deploying a model in an EC2 instance. Afterwards, an EC2 instance needs to be launched with the appropriate instance type based on the end user's discretion or the project's requirements. Once the instance has been launched, the end user needs to connect to it via SSH—for example—and install docker if it wasn't included in the AMI. Afterwards, the docker container along with the necessary files need to be transferred into the EC2 instance and the security group needs to be configured based on the specifications such as the ephemeral ports. Finally, the endpoint may be accessed via the EC2 instance's public IPv4 address or DNS along with the specified port.

Built-in algorithm + SageMaker endpoint

This method is the most straightforward among the deployment choices since everything is packaged inside SageMaker itself. You simply need to instantiate an estimator object with your chosen container that houses the built-in algorithm and use the .deploy method to create an inference endpoint. Unlike the other deployment processes, everything regarding the deployment is abstracted from the end user.