Congratulations! You passed!

Grade received 96.43% **To pass** 80% or higher

Model deployment and integration

Practice

Model integration and monitoring

Lecture Notes (Optional)

Practice

Practice Quiz: Week 2
7 questions

Week 2

Practice Quiz • 30 min

Submit your assignment

Try again

0.75 / 1 point

Receive grade

To Pass 80% or higher

Your grade
96.43%

We keep your highest score

View Feedback

Like **Dislike** **Report an issue**

1. To get the most value out of your machine learning models, you need to deploy them into production to start making practical decisions quickly. Which of the following are the most important factors to consider when choosing a deployment option?

   - [x] Cost

     Correct
     Correct! Cost is definitely something to consider when choosing a deployment option depending on your use case (ie. real-time, batch, edge, etc)

   - [x] Latency

     Correct
     Correct! The response time of the deployed model is very essential to the choice of deployment.

   - [x] Network connectivity

     Correct
     That's right! This is important because some deployment options require access to a network or internet.

   - [x] Model Accuracy

     This should not be selected
     Not quite! The accuracy of the model is not vital to choosing a deployment option. Please revisit the lecture "Model Deployment Overview'' in Week 1.

2. As a machine learning engineer on the marketing team, you want to serve advertisements to a customer based on their current session activity. For a better user experience, you want to serve the ads as quickly as possible. What is the most ideal cloud deployment option for a model of this type?

   1 / 1 point

   - ( ) Edge deployment
   - (•) Real-time deployment
   - ( ) Batch inference deployment
   - ( ) All of the above

   Correct
   Correct! A real-time deployment option is ideal for this use case.

3. You would like to deploy a model that detects pneumonia from x-ray images of patients in a rural area. Internet connection is a major challenge here and can slow down the process.

   1 / 1 point

   True/False: Edge deployment and batch inference deployment are both great options for deploying this model.

   - ( ) True
   - (•) False

   Correct
   Correct! Edge deployment works well in cases where there is limited or no network connectivity while the batch inference pipeline requires some form of network connectivity to work properly.

4. When selecting a strategy to deploy a new or updated model, you want to minimize risk, and reduce downtime. Additionally, we need to measure and compare performance relative to the original model. Which of the following statements correctly describes the canary deployment strategy.

   1 / 1 point

   - ( ) All users are shifted from the older version to the new, replacement version all at once.
   - (•) A small percentage of users are initially exposed to the new version of the model while the majority of users continue using the original version of the model.
   - ( ) A new version of the model is deployed into production alongside a live model, and traffic is sent to both models simultaneously.
   - ( ) None of the above

   Correct
   Correct! This is called the "canary deployment" strategy, when sets of users are incrementally exposed to the new model version.

5. In both A/B testing and canary deployment strategies, you split your traffic to compare different model versions. What is the difference between these 2 strategies?

   1 / 1 point

   - ( ) Canary deployment is more dynamic and uses reinforcement learning to dynamically shift traffic to the winning model versions.
   - ( ) A/B testing deployment is more dynamic and uses reinforcement learning to dynamically shift traffic to the winning model versions.
   - (•) A/B testing targets larger groups and typically runs for longer periods of time.
   - ( ) Canary deployment targets larger groups and typically runs for longer periods of time.

   Correct
   Correct! A/B testing needs to run over longer periods of time to gather performance data that are statistically significant.

6. Amazon SageMaker provides a feature that allows you to host a sequential chain of models behind a single endpoint. What is this feature called?

   1 / 1 point

   - ( ) Autoscale
   - ( ) SageMaker CloudWatch
   - ( ) Multi-model Endpoints