Google Discussions

Exam Professional Machine Learning Engineer All Questions

View all questions & answers for the Professional Machine Learning Engineer exam

Go to Exam

EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 234 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 234
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a retail company that is using a regression model built with BigQuery ML to predict product sales. This model is being used to serve online predictions. Recently you developed a new version of the model that uses a different architecture (custom model). Initial analysis revealed that both models are performing as expected. You want to deploy the new version of the model to production and monitor the performance over the next two months. You need to minimize the impact to the existing and future model users. How should you deploy the model?

A. Import the new model to the same Vertex AI Model Registry as a different version of the existing model. Deploy the new model to the same Vertex AI endpoint as the existing model, and use traffic splitting to route 95% of production traffic to the BigQuery ML model and 5% of production traffic to the new model.

B. Import the new model to the same Vertex AI Model Registry as the existing model. Deploy the models to one Vertex AI endpoint. Route 95% of production traffic to the BigQuery ML model and 5% of production traffic to the new model.

C. Import the new model to the same Vertex AI Model Registry as the existing model. Deploy each model to a separate Vertex AI endpoint.

D. Deploy the new model to a separate Vertex AI endpoint. Create a Cloud Run service that routes the prediction requests to the corresponding endpoints based on the input feature values.

Show Suggested Answer

by pikachu007 at *Jan. 13, 2024, 8:07 a.m.*

Comments

Type your comment...

Submit

bfdf9c8 2 months, 3 weeks ago

Selected Answer: B

I'm considering two options, A and B. Both deploy to the same endpoint and divide traffic in a similar way. However, option B is more appropriate because it generates a new model rather than just creating a new version of the existing model.

upvoted 1 times

fitri001 6 months, 1 week ago

Selected Answer: A

Minimal Disruption: Deploying the new model to the same endpoint avoids changes for existing users. Traffic splitting ensures a gradual rollout, minimizing any potential impact on production.
Performance Monitoring: By routing a small percentage of traffic (5%) to the new model, you can monitor its performance in a controlled environment for the next two months. Metrics like prediction accuracy and latency can be compared with the BigQuery ML model.
Versioning in Model Registry: Storing both models in the same Vertex AI Model Registry with clear versioning allows easy tracking and management.

upvoted 3 times

fitri001 6 months, 1 week ago

why not others option?
B. Deploying Models to One Endpoint without Traffic Splitting: This approach doesn't allow for controlled rollout and could abruptly switch all traffic to the new model, potentially causing disruptions.
C. Deploying Models to Separate Endpoints: This requires users to update their prediction pipelines to interact with the new endpoint, introducing unnecessary complexity and potential delays.
D. Cloud Run Service with Feature-Based Routing: While Cloud Run can route traffic, feature-based routing might be more complex to implement for sales prediction and might not be necessary with traffic splitting.

upvoted 1 times

pinimichele01 6 months, 2 weeks ago

Selected Answer: A

https://cloud.google.com/vertex-ai/docs/general/deployment#models-endpoint

upvoted 2 times

Yan_X 8 months, 3 weeks ago

Selected Answer: A

A, no need to separate endpoint.

upvoted 4 times

BlehMaks 9 months, 1 week ago

Selected Answer: C

as i understand we need to minimize the impact to the model users, so if we take a part of the traffic from the old model users, we will effect them. As for me we should deploy models to separated endpoints and duplicate the traffic

upvoted 2 times

pikachu007 9 months, 1 week ago

Selected Answer: A

B. Doesn't Specify Traffic Splitting: Deploying models to a single endpoint without explicit traffic splitting might lead to unpredictable model selection behavior, hindering controlled evaluation.
C. Separate Endpoints: While isolating models, it introduces complexity in managing multiple endpoints and routing logic, increasing operational overhead.
D. Cloud Run Routing: Adds complexity by requiring a separate service to manage routing, potentially increasing latency and maintenance overhead compared to Vertex AI's built-in traffic splitting.

upvoted 2 times

Start Learning for free

Social Media

Facebook , Twitter

YouTube , Reddit

Pinterest

We are the biggest and most updated IT certification exam material website.

Using our own resources, we strive to strengthen the IT professionals community for free.

© 2024 ExamTopics

ExamTopics doesn't offer Real Microsoft Exam Questions. ExamTopics doesn't offer Real Amazon Exam Questions. ExamTopics Materials do not contain actual questions and answers from Cisco's Certification Exams.

CFA Institute does not endorse, promote or warrant the accuracy or quality of ExamTopics. CFA® and Chartered Financial Analyst® are registered trademarks owned by CFA Institute.