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 266 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 266

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for an organization that operates a streaming music service. You have a custom production model that is serving a “next song” recommendation based on a user's recent listening history. Your model is deployed on a Vertex AI endpoint. You recently retrained the same model by using fresh data. The model received positive test results offline. You now want to test the new model in production while minimizing complexity. What should you do?

A. Create a new Vertex AI endpoint for the new model and deploy the new model to that new endpoint. Build a service to randomly send 5% of production traffic to the new endpoint. Monitor end-user metrics such as listening time. If end-user metrics improve between models over time, gradually increase the percentage of production traffic sent to the new endpoint.

B. Capture incoming prediction requests in BigQuery. Create an experiment in Vertex AI Experiments. Run batch predictions for both models using the captured data. Use the user’s selected song to compare the models performance side by side. If the new model’s performance metrics are better than the previous model, deploy the new model to production.

C. Deploy the new model to the existing Vertex AI endpoint. Use traffic splitting to send 5% of production traffic to the new model. Monitor end-user metrics, such as listening time. If end-user metrics improve between models over time, gradually increase the percentage of production traffic sent to the new model.

D. Configure a model monitoring job for the existing Vertex AI endpoint. Configure the monitoring job to detect prediction drift and set a threshold for alerts. Update the model on the endpoint from the previous model to the new model. If you receive an alert of prediction drift, revert to the previous model.

Show Suggested Answer

by pikachu007 at *Jan. 13, 2024, 4:12 p.m.*

Comments

Type your comment...

Submit

fitri001 6 months, 3 weeks ago

Selected Answer: C

For Simplicity: If speed and simplicity are your top priorities, deploying to the existing endpoint with caution (close monitoring during deployment) can work.--> choose C
For Safety and Control: If minimizing risk and having better control over the testing process are more important, creating a new endpoint is the better option. This is generally the recommended approach for most production deployments. --> choose A

upvoted 2 times

daidai75 9 months, 2 weeks ago

Selected Answer: C

Here's why the option C is preferable:
Minimized complexity:
Leverages existing endpoint: No need to create and manage a new endpoint, reducing setup and maintenance overhead.
Traffic splitting readily available: Vertex AI provides built-in traffic splitting functionality, simplifying traffic distribution.
Efficient testing and monitoring:
Direct comparison: Sending a percentage of traffic to the new model allows for direct comparison with the current model's performance on real user data.
Gradual rollout: Starting with a small percentage mitigates potential risks and allows for gradual transition based on observed improvements.
End-user metric monitoring: Focusing on metrics like listening time directly reflects user engagement and preference for the new recommendations.

upvoted 3 times

b1a8fae 9 months, 2 weeks ago

Selected Answer: C

Traffic splitting is a feature of Vertex AI that allows you to distribute the prediction requests among multiple models or model versions within the same endpoint. You can specify the percentage of traffic that each model or model version receives, and change it at any time. Traffic splitting can help you test the new model in production without creating a new endpoint or a separate service. You can deploy the new model to the existing Vertex AI endpoint, and use traffic splitting to send 5% of production traffic to the new model. You can monitor the end-user metrics, such as listening time, to compare the performance of the new model and the previous model. If the end-user metrics improve between models over time, you can gradually increase the percentage of production traffic sent to the new model. This solution can help you test the new model in production while minimizing complexity and cost.

upvoted 2 times

pikachu007 9 months, 3 weeks ago

Selected Answer: C

Option A: Building a separate service adds unnecessary complexity and requires managing two endpoints.
Option B: Batch predictions in Vertex AI Experiments might not reflect real-time user behavior and don't directly affect the production environment.
Option D: Model monitoring alerts for prediction drift might be triggered by natural variations in user behavior instead of genuine performance issues and could lead to unnecessary model rollbacks.

upvoted 2 times

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.

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.