



Home / Dashboard / My Courses / AWS Certified Machine Learning Specialty / Machine Learning Implementation and Operations / Report

← Back to the **Course**



Level: Advanced

AWS Certified Machine Learning Specialty

Machine Learning Implementation and Operations

Completed on **Mon, 27 Jun 2022**



1st
Attempt



7/12
Marks Obtained



58.33%
Your Score



0h 21m 40s
Time Taken



FAIL
Result

Domain wise Quiz Performance Report



Join us on **Slack community**

No.	Domain	Total Question	Correct	Incorrect	Unattempted
1	ML Implementation and Operations	12	7	5	0
Total	All Domains	12	7	5	0

Review the Answers

Filter By **All Questions**

Question 1

Incorrect

Domain: ML Implementation and Operations

You are a machine learning specialist at a large bank where you are rolling out a new marketing campaign. The campaign is to determine whether a given customer will enroll in a new term deposit at the bank. You have decided to use the SageMaker Autopilot service to produce the best machine learning pipeline for your enrollment predictions. Your data set, which consists of 17 attributes and

hundreds of millions of rows, is stored in the libsvm format on one of your S3 buckets. You plan to use SageMaker Autopilot on your dataset to get the most accurate machine learning pipeline by exploring several potential options, or candidate models. However, when you attempt to run AutoPilot using your Autopilot job, you get an error stating:

Could not complete the data builder processing job. The AutoML Job cannot continue.

What might be the cause of your Autopilot job failure?

- A. Your data source is too large, 17 attributes x hundreds of millions of rows. You need to break your data source into subset files and run your Autopilot job over the subsets of data.
- B. The format of your data source is not supported by the Autopilot service. right
- C. Your data source is in a tabular format, which needs to be transformed using a machine learning tool such as sklearn or pandas.
- ☒ D. Your data source has a header row in the first file in your set of source data files. You need to remove the header row. wrong

Explanation:

Answer: B

Option A is incorrect. The AutoPilot service, as it is built upon the SageMaker service, has no limit to the size of your source data.

Option B is CORRECT. SageMaker AutoPilot only supports tabular data sets in the CSV format. Your source data is in the libsvm format. This will cause your Autopilot job to fail with the given error.

Option C is incorrect. SageMaker AutoPilot supports tabular data sets in the CSV format. If your data source is in a tabular format that cannot be the source of your error. Your source data is in the libsvm format, which will cause the given error.

Option D is incorrect. The AutoPilot service supports tabular data where either all files have a header row or the first file of the dataset when sorted in alphabetical/lexical order, has a header row. Therefore, your data source having a header row cannot be the source of your error.

Reference:

Please see the **Amazon Sagemaker developer guide** titled [Samples: Explore modeling with Amazon SageMaker Autopilot](#).

Please refer to the GitHub repository titled **Direct Marketing with Amazon SageMaker Autopilot**.

Ask our Experts

 [View Queries](#)



Question 2

Correct

Domain: ML Implementation and Operations

As a machine learning specialist for a credit card company, you build a model to extract information from scanned bank application documents. These forms generate datasets containing customer Personally Identifiable Information (PII) such as credit card numbers. The resulting datasets need to be highly accurate; your company has a very low tolerance for inaccurate bank application data.

How can you ensure the PII data remains encrypted and the credit card information is secure while ensuring the highest level of quality from the scanned forms?

- A. Build a custom encryption algorithm to encrypt the data. Then store the data in your SageMaker instance in your private subnet in your VPC. Use the SageMaker DeepAR algorithm to randomize the credit card numbers. Finally, use SageMaker GroundTruth to provide a human review of the image scan data.
- B. Use an IAM policy to encrypt the data on your S3 bucket. Then use Kinesis Data Analytics to obfuscate the credit card numbers. Finally, use SageMaker AutoPilot to provide a human review of the image scan data.
- C. Use a SageMaker custom image to encrypt the data when it is loaded into the SageMaker instance in your private subnet in your VPC. Use the SageMaker principal component analysis built-in algorithm to obfuscate the credit card numbers.
- ☒ D. Use AWS KMS to encrypt the data on S3 and in your SageMaker environment. Then obfuscate the credit card numbers from the customer data using AWS Comprehend. Finally, use SageMaker Augmented AI to provide a human review of the image scan data. right

Explanation:

Answer: D

Option A is incorrect. Building a custom encryption algorithm will be time-consuming and will most likely produce a less secure encryption technique than using AWS KMS. Also, the SageMaker DeepAR built-in algorithm is used for forecasting, not randomizing credit card numbers. Finally, the GroundTruth service is used to label datasets, not for verifying datasets.

Option B is incorrect. An IAM policy cannot be used to encrypt your data. It is used to build policies to control access to your data on S3 and in SageMaker. You could use Kinesis Data Analytics to obfuscate your data. But it would be a very inefficient way of achieving this goal. Also, the question doesn't state that the scanned data is streamed to your S3 bucket, so a Kinesis solution is not a good choice. Finally, SageMaker AutoPilot is a service that allows you to automate your SageMaker pipeline. It does not provide a human review capability.

Option C is incorrect. A SageMaker custom image, a component of the SageMaker Studio service, will not give you the capability to encrypt your data. Also, the PCA built-in algorithm is used for dimensionality reduction. It would not be a good choice for data obfuscation.

Option D is CORRECT. AWS KMS is the best choice provided for encrypting your data on your S3 bucket and in your SageMaker environment. Also, AWS Comprehend now provides a very efficient

way to obfuscate your credit card data. Lastly, SageMaker Augmented AI can be used to provide a human review step in the data engineering step of your modeling workflow.

Reference:

Please see the [AWS SageMaker developer guide](#) titled **Using Amazon Augmented AI for Human Review**.

Please refer to the [Amazon SageMaker GroundTruth overview page](#).

Please review the [AWS Machine Learning blog](#) titled **Detecting and redacting PII using Amazon Comprehend**.

Ask our Experts

 [View Queries](#)

**Question 3**

Correct

Domain: ML Implementation and Operations

You are a machine learning specialist at a government agency that has gathered election data that they plan to use to predict election outcomes. Using the agile methodology, a minimum viable product for the model idea has been built using a relative data sample. Now your agency is ready to create a machine learning model using SageMaker Studio. The election data you plan to use for training data is stored in RDS SQL Server.

Which option gives you the highest performing and most efficient method of loading the data into your SageMaker Studio environment?

- ☐ A. Load the election data into your SageMaker Studio jupyter notebook using a direct connection to the SQL Server database from the code in your notebook cells.
- ☒ B. Use AWS Data Pipeline to load the election data from your SQL Server instance to one of your S3 buckets. Load the data into your SageMaker Studio jupyter notebook from the S3 bucket. right
- ☐ C. Use AWS Data Pipeline to load the election data from your SQL Server instance to a set of DynamoDB tables. Then connect to your DynamoDB tables from the code in your jupyter notebook cells to load the data into your SageMaker Studio environment.
- ☐ D. Use AWS DMS to load the election data from your SQL Server instance to an ElastiCache cluster. Then connect to your Elasticache cluster from the code in your jupyter notebook cells to load the data into your SageMaker Studio environment.

Explanation:

Answer: B

Option A is incorrect. SageMaker requires that your model artifacts, including your training data, be stored in S3. This option describes using a direct connection to SQL Server in RDS to load the data.

Option B is CORRECT. This option correctly describes moving the training data from SQL Server to your S3 bucket. Then loading the data into your jupyter notebook from the S3 bucket.

Option C is incorrect. SageMaker requires that your model artifacts, including your training data, be stored in S3. This option describes using AWS Data Pipeline to load the data into DynamoDB then loading the data into your jupyter notebook from DynamoDB.

Option D is incorrect. SageMaker requires that your model artifacts, including your training data, be stored in S3. This option describes using AWS DMS to load the data into Elasticache then loading the data into your jupyter notebook from Elasticache.

Reference:

Please see the **Amazon SageMaker developer guide** titled **Get Started with Amazon SageMaker Notebook Instances and SDKs**.

Please refer to the **GitHub repository** titled **Getting Started with Amazon SageMaker Studio**.

Please review the **GitHub repository** titled **Amazon SageMaker Studio Walkthrough**.

Ask our Experts

 [View Queries](#)



Question 4

Correct

Domain: ML Implementation and Operations

You are an engineering manager for a healthcare conglomerate that is expanding its web application suite of online patient tools and services. You have just started a new web user experience lab where you are responsible for creating a SageMaker environment for your analyst and engineering staff. Your corporation's data security policies require that communication within your web lab machine learning environment not traverse the internet.

How can you provide the SageMaker service to your user experience labs without allowing internet access for your lab SageMaker notebook instances?

- A. Use a NAT gateway in the public subnet of your corporate VPC to enable the SageMaker service.
- B. Use a NAT gateway in the private subnet of your corporate VPC to enable the SageMaker service.
- ☒ C. Use a SageMaker VPC interface endpoint in your corporate VPC to enable the SageMaker service. right

D. Use VPC peering between the VPC hosting the SageMaker service and your corporate VPC to enable the SageMaker service.

Explanation:

Answer: C

Option A is incorrect. You cannot enable the SageMaker service without a connection between your VPC and the Amazon SageMaker service, which runs in an Amazon managed service account. Using a NAT Gateway in your public subnet within your corporate VPC will send your traffic over the public internet.

Option B is incorrect. NAT Gateways are instantiated in your public subnet, not your private subnet. Also, using a NAT Gateway in your public subnet within your corporate VPC will send your traffic over the public internet.

Option C is CORRECT. Using a VPC interface endpoint to enable communication between your VPC and the SageMaker managed service account allows your traffic to flow entirely and securely within the AWS network.

Option D is incorrect. The VPC hosting the SageMaker service is an Amazon managed service account. VPC Peering is used for client account peering, not for communication with an Amazon managed service account.

Reference:

Please see the [AWS Machine Learning blog](#) titled [Understanding Amazon SageMaker notebook instance networking configurations and advanced routing options](#).

Please refer to the [Amazon SageMaker developer guide](#) titled [Connect to SageMaker Through a VPC Interface Endpoint](#).

Ask our Experts

 [View Queries](#)



Question 5

Incorrect

Domain: ML Implementation and Operations

You are on a machine learning team working for a retail website. Your team is responsible for the shopping cart leg of the “complete purchase” journey. Your team is building an up-sell recommendation model to be used to show “customers who purchased this item also purchased these other items” recommendations on the shopping cart page. Your model will use a dataset containing customer credit card information and other Personally Identifiable Information (PII). The PII data needs to be protected to meet the Payment Card Industry Data Security Standard (PCI DSS) requirements.

What is the most efficient way to obfuscate the PII data, including the credit card information?

- ☐ A. Use KMS to encrypt the PII data in transit and at rest. wrong
- ☐ B. Tokenize the PII data. right
- ☐ C. Use AWS Shield Advanced for all website traffic.
- ☐ D. Use GuardDuty for website traffic.

Explanation:

Answer: B

Option A is incorrect. Using KMS and encrypting your data in transit and at rest are more complex and costly than using tokenization on the specific PII data, including the credit card data.

Option B is CORRECT. You can use tokenization instead of encryption when you only need to protect specific highly sensitive data for regulatory compliance requirements, such as PCI DSS.

Option C is incorrect. You can use AWS Shield to protect your retail website from distributed denial of service (DDoS) attacks, and Shield Advanced gives you a DDoS response team. However, Shield and Shield advanced won't protect your PII credit card data from being exposed. You need to either encrypt your data or tokenize your customer's PII data.

Option D is incorrect. You can use GuardDuty to systematically monitor network traffic to detect anomalies in your website users' behavior by using machine learning. However, GuardDuty won't protect your PII credit card data from being exposed. You need to either encrypt your data or tokenize your customer's PII data.

Reference:

Please see the [AWS Compute Blog](#) titled **Building a serverless tokenization solution to mask sensitive data**.

Please see the [AWS Big Data blog](#) titled **Best practices for securing sensitive data in AWS data stores**.

Please see the [Wikipedia page](#) titled **Payment Card Industry Data Security Standard**.

Please refer to the [AWS GuardDuty overview page](#).

Please see the [AWS Shield overview page](#).

Ask our Experts

[+ View Queries](#)



Question 6

Incorrect

Domain: ML Implementation and Operations

You are a machine learning specialist at a medical research institute where you are using deep learning techniques to analyze brain scan images. Your team has built and trained your deep learning model using one of the pre-built SageMaker container images for TensorFlow. You are now attempting to deploy it for inferences in production.

You have used the TensorFlow 2.3.0 framework using Horovod on GPU servers using python version 3.7. Your training jobs run well. But when you deploy to your container image for serving inferences, your container fails. Why is your inference container failing?

- ☐ A. You cannot use TensorFlow in a SageMaker inference container image. wrong
- ☐ B. The TensorFlow SageMaker inference container image only runs on CPU servers.
- ☐ C. You need to remove the Horovod operations from your inference container. right
- ☐ D. You should use the DeepAR SageMaker built-in algorithm.

Explanation:

Answer: C

Option A is incorrect. You can use TensorFlow in a SageMaker inference container. You just can't use the container with Horovod for inference.

Option B is incorrect. You can use TensorFlow in a SageMaker inference container on GPUs or CPUs. You just can't use the container with Horovod for inference.

Option C is CORRECT. When running inference on a SageMaker inference container for TensorFlow that was trained with Horovod, you need to remove the references to Horovod before deploying for inference.

Option D is incorrect. The DeepAR SageMaker built-in algorithm is an algorithm for forecasting one-dimensional time series using recurrent neural networks (RNN). You would not use this algorithm for image analysis.

Reference:

Please see the [Amazon SageMaker developer guide](#) titled **Using Docker containers with SageMaker**.

Please see the [Amazon SageMaker developer guide](#) titled **Prebuilt SageMaker Docker Images for TensorFlow, MXNet, Chainer, and PyTorch**.

Please see the [GitHub repository](#) titled **Available Deep Learning Containers Images**.

Please refer to the [GitHub repository](#) overview page titled **Horovod**.

Please see the [GitHub repository](#) titled **Horovod Inference**.

Please see the [Amazon SageMaker developer guide](#) titled **DeepAR Forecasting Algorithm**.

[Ask our Experts](#)

[+ View Queries](#)**Question 7**

Incorrect

Domain: ML Implementation and Operations

You are a machine learning specialist at a financial services firm. Your team has been tasked with developing a forecasting model to predict the price movement of the S&P 500 Emini futures contract on the CME trading exchange. You have historical data from the past year as your training data. You are also considering using trading data from the closely related NASDAQ 100 Emini futures contract as part of your training data.

Which AWS machine learning service and algorithm should you use for your model that gives your team the most expeditious result with the least amount of administrative overhead?

- ☐ A. SageMaker using the DeepAR Forecasting algorithm wrong
- ☐ B. Amazon Forecast using the Prophet algorithm
- ☐ C. Amazon Forecast using the NTPS algorithm
- ☐ D. Amazon Forecast using the ARIMA algorithm
- ☒ E. Amazon Forecast using the DeepAR+ algorithm right

Explanation:**Answer: E**

Option A is incorrect. While the SageMaker DeepAR Forecasting algorithm would be a good choice for this type of forecasting, it requires much more administrative effort than using Amazon Forecast.

Option B is incorrect. The Amazon Forecast Prophet algorithm is best suited for time series with strong seasonal effects and several seasons of historical data. Stock index futures data is not seasonal in nature.

Option C is incorrect. The Amazon Forecast NTPS algorithm is best suited for sparse or intermittent time series. Stock index futures data is not sparse or intermittent in nature.

Option D is incorrect. The Amazon Forecast ARIMA algorithm is best suited for simple datasets with under 100 time series. Stock index futures data is complex with many features and millions of observations across many time series.

Option E is CORRECT. The Amazon Forecast DeepAR+ algorithm works best with large datasets containing hundreds of feature time series. It also works with forward-looking related time series. Additionally, Amazon Forecast requires much less administrative overhead than SageMaker. So this is the best option for the given scenario.

Reference:

Please see the **Amazon Forecast developer guide** titled **Getting Started (Console)**.

Please see the **Amazon Forecast developer guide** titled **Choosing an Amazon Forecast Algorithm**.

Please see the **Amazon SageMaker developer guide** titled **Use Amazon SageMaker Built-in Algorithms**.

Ask our Experts

 [View Queries](#)



Question 8

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a software company that sells an app that is used to identify plants in the wild. Users can take a picture of a plant, maybe poison ivy, and your app will identify the plant using an image classification algorithm. You have built your classification algorithm based model. You need to make sure your model produces inferences quickly because your users expect a classification within less than 3 seconds. How should you optimize your model for performance?

- ☐ A. Use accuracy for your evaluation metric
- ☐ B. Use f1 for your evaluation
- ☒ C. Use accuracy for your evaluation metric and runtime as your satisficing metric right
- ☐ D. Use accuracy as your satisficing metric

Explanation:

Correct Answer: C

Option A is incorrect. Using accuracy as an evaluation metric alone will not limit your model's runtime. You may get very accurate image classification but your runtime may be too slow for your user's expectations.

Option B is incorrect. Using f1 as an evaluation metric alone will not limit your model's runtime. You may get very accurate image classification but your runtime may be too slow for your user's expectations.

Option C is correct. Using accuracy as your model evaluation metric while also limiting your classifier runtime to your satisficing metric will give you the most accurate classification within the acceptable runtime.

Option D is incorrect. You need to use accuracy as your evaluation metric, not as your satisfying metric.

Reference:

Please see the book Machine Learning Yearning by Andrew Ng, chapter 9: **9 Optimizing and satisficing metrics** (<https://github.com/ajaymache/machine-learning-yearning>), and the Structuring Your Machine Learning Projects article titled **How to choose your Satisficing and optimizing metric?** (<https://medium.com/structuring-your-machine-learning-projects/satisficing-and-optimizing-metric-24372e0a73c>)

Ask our Experts

 View Queries



Question 9

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a bank in their fraud detection department. You and your machine learning team have been assigned the task of creating a fraud detection model that can be used in an automated way. When a transaction is attempted on one of your bank's issued credit cards, the endpoint of your model will be called and it should return an inference result flagging the transaction as fraudulent or not. Which Amazon machine learning services should you use to build your solution?

- ☐ A. Lambda to receive HTTPS REST API inference requests and make inference requests to model endpoint, deploy an XGBoost fraud detection model using SageMaker batch transform
- ☒ B. API Gateway to receive HTTPS REST API inference requests, Lambda to process inference requests from API Gateway and make inference requests to model endpoint, deploy an XGBoost fraud detection model using SageMaker hosting service right
- ☐ C. API Gateway to receive HTTPS REST API inference requests and make inference requests to model endpoint, deploy an XGBoost fraud detection model using SageMaker hosting service
- ☐ D. Lambda to receive HTTPS REST API inference requests and make inference requests to model endpoint, deploy an XGBoost fraud detection model using SageMaker hosting service

Explanation:

Correct Answer: B

Option A is incorrect. Lambda cannot receive HTTPS REST API inference requests without an API service (frequently API Gateway) to receive the request. Also, SageMaker batch transform is used to process batches of inference requests, not single inference requests.

Option B is correct. API Gateway can receive REST API calls and invoke a Lambda function. The Lambda function can invoke your endpoint one transaction at a time. A SageMaker hosting service endpoint can process transactions one at a time.

Option C is incorrect. API Gateway can receive REST API calls but it cannot invoke your inference endpoint without some processing logic (frequently implemented in a Lambda function).

Option D is incorrect. Lambda cannot receive HTTPS REST API inference requests without an API service (frequently API Gateway) to receive the request.

Reference:

Please see the AWS Machine Learning Blog titled **Simplify machine learning with XGBoost and Amazon SageMaker** (<https://aws.amazon.com/blogs/machine-learning/simplify-machine-learning-with-xgboost-and-amazon-sagemaker/>), the Fraud Detection Using Machine Learning page titled **Architecture Overview** (<https://docs.aws.amazon.com/solutions/latest/fraud-detection-using-machine-learning/architecture.html>), and the Amazon SageMaker developer guide titled **Deploy a Model in Amazon SageMaker** (<https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-deployment.html#how-it-works-hosting>)

Ask our Experts

 [View Queries](#)



Question 10

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a company that produces a ride share service. The service mobile app is the primary way for users to make requests to your service. As part of the company's machine learning team, you are responsible for the model that optimizes driver placement to increase profit. You have built a model based on the k-Means clustering algorithm and you have deployed it to a hosted service endpoint on SageMaker. Your mobile app service runs in your company's VPC on AWS. How can you secure your call from the ride share service to your model endpoint so that your service-to-endpoint traffic does not go over the internet?

- ☐ A. Route all traffic from the service destined for the model endpoint through the Internet Gateway associated with the VPC.
- ☐ B. Route all traffic from the service destined for the model endpoint through a NAT Gateway running in your VPC.
- ☐ C. Route all traffic from the service destined for the model endpoint through Direct Connect.
- ☒ D. Route all traffic from the service destined for the model endpoint through PrivateLink right

Explanation:

Correct Answer: D

Option A is incorrect. Routing your requests through the Internet Gateway associated with your VPC would send the traffic over the internet.

Option B is incorrect. Using a NAT Gateway would add unnecessary infrastructure components and bottlenecks to your solution.

Option C is incorrect. Using Direct Connect to connect your VPC based service code to the SageMaker endpoint service would also add unnecessary infrastructure components and bottlenecks to your solution.

Option D is correct. The PrivateLink service allows you to make requests to your endpoint without traversing the internet.

Reference:

Please see the Towards Data Science article titled **How I used Clustering to analyze ride-sharing data from Uber** (<https://towardsdatascience.com/how-does-uber-use-clustering-43b21e3e6b7d>), and the AWS Machine Learning blog titled **Secure prediction calls in Amazon SageMaker with AWS PrivateLink** (<https://aws.amazon.com/blogs/machine-learning/secure-prediction-calls-in-amazon-sagemaker-with-aws-privatelink/>)

Ask our Experts

 [View Queries](#)



Question 11

Incorrect

Domain: ML Implementation and Operations

You work as a machine learning specialist for a financial services firm. Your machine learning team is responsible for creating model endpoints that allow for A/B testing of new modeling ideas for your user experience teams. The user experience teams use hypothesis testing to decide when a user experience change works based on the stated hypothesis. For example, a hypothesis is posited that if a certain change is made to your model that uses clickstream data to decide when and which product ads to insert you'll see a 10% clickthrough rate improvement. How can you implement the most efficient solution that allows you to send most inference traffic to the existing model and a small subset of the inference traffic to your proposed new model?

- ☐ A. Deploy the new model to a new endpoint and use a Lambda function to send 5% of the inference requests to the new model endpoint, while sending the remaining 95% to the existing model endpoint.
- ☐ B. Deploy the new model to the same endpoint as the existing model using an Endpoint Configuration. Then send 5% of the inference requests to the new model and send the remaining 95% to the existing model using the same endpoint. right
- ☒ C. Deploy the new model to the same endpoint as the existing model using an Elastic Inference. Then send 5% of the inference requests to the new model and send the remaining 95% to the existing model using the same endpoint. wrong

D. Deploy the new model to the same endpoint as the existing model using a Batch Transform. Then send 5% of the inference requests to the new model and send the remaining 95% to the existing model using the same endpoint.

Explanation:

Correct Answer: B

Option A is incorrect. This approach is far less efficient than using an Endpoint Configuration to define two variants available at the same endpoint.

Option B is correct. You can define an Endpoint Configuration that allows for the endpoint to serve the primary endpoint and the production variant endpoint. This is the most efficient option.

Option C is incorrect. Elastic Interface is used to improve performance for inference requests, not for defining production variants.

Option D is incorrect. Batch Transform endpoints are used to allow for multiple inference requests to be processed in batch, not for defining production variants.

Reference:

Please see the Amazon SageMaker developer guide titled **Deploy Models for Inference** (<https://docs.aws.amazon.com/sagemaker/latest/dg/deploy-model.html>), and the Amazon SageMaker developer guide titled **Deploy a Model in Amazon SageMaker** (<https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-deployment.html#how-it-works-hosting>), and the Amazon SageMaker developer guide titled **Use Amazon SageMaker Elastic Inference (EI)** (<https://docs.aws.amazon.com/sagemaker/latest/dg/ei.html>)

Ask our Experts

 View Queries



No queries found.

Question 12

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for an online retail business. Your business has spikes in demand that are predictable, such as holiday shopping. But, it also has unpredictable spikes in traffic. Your website uses the machine learning models your team has created to place personalized advertisement ads throughout your customer's online purchase journey. How can you take advantage of SageMakers features to make sure your model endpoints scale along with the spikes and drops in traffic to your website?

A. Use an Elastic Inference to improve the performance of your real-time inference requests.

B. Deploy a model variant using an Endpoint Configuration that will allow you to have multiple variants of your model running at one endpoint.

☒ C. Use SageMaker to configure your production variant so that it uses autoscaling to scale out and in based on your demand. right

D. Use SageMaker to configure your Elastic Inference so that it uses autoscaling to scale out and in based on your demand.

Explanation:

Correct Answer: C

Option A is incorrect. Elastic Inference will improve the performance of your SageMaker endpoint, but it will not provide the horizontal scaling that auto scaling provides.

Option B is incorrect. Having multiple variants to handle your inference requests will help but this configuration will not scale in and out based on your demand with you manually changing the Endpoint Configuration.

Option C is correct. Using a production variant and application autoscaling is the most efficient way to make your model scale to your traffic demands.

Option D is incorrect. Elastic Inference alone will not provide an auto scaling configuration.

Reference:

Please see the Amazon SageMaker developer guide titled **Deploy a Model in Amazon SageMaker** (<https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-deployment.html#how-it-works-hosting>), and the Amazon SageMaker developer guide titled **Use Amazon SageMaker Elastic Inference (EI)** (<https://docs.aws.amazon.com/sagemaker/latest/dg/ei.html>), the Amazon SageMaker developer guide titled **Configure model autoscaling with the console** (<https://docs.aws.amazon.com/sagemaker/latest/dg/endpoint-auto-scaling-add-console.html>), and the Amazon SageMaker developer guide titled **Define a scaling policy** (<https://docs.aws.amazon.com/sagemaker/latest/dg/endpoint-auto-scaling-add-code-define.html>)

Ask our Experts

 [View Queries](#)



Finish Review

Certification

Company

[Cloud Certification](#)

[Java Certification](#)

[PM Certification](#)

[Big Data Certification](#)

[Become Our Instructor](#)

[Support](#)

[Discussions](#)


[Blog](#)

[Business](#)

Support

[Contact Us](#)

[Help Topics](#)

 **Join us on Slack!**

Join our open **Slack community** and
get your queries answered instantly!
Our experts are online to answer your
questions!