



**Amazon**

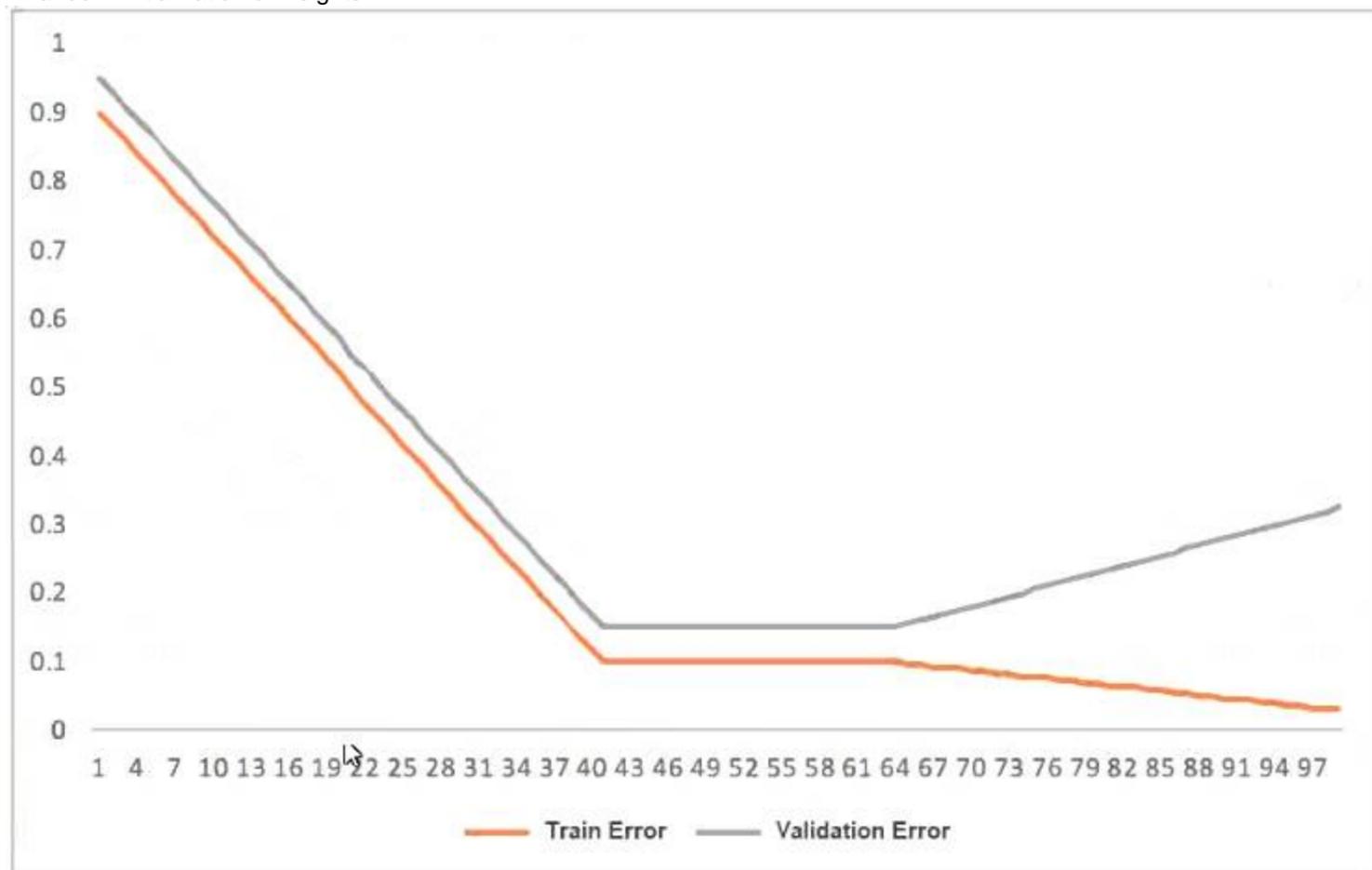
## **Exam Questions AWS-Certified-Machine-Learning-Specialty**

AWS Certified Machine Learning - Specialty

**NEW QUESTION 1**

This graph shows the training and validation loss against the epochs for a neural network The network being trained is as follows

- Two dense layers one output neuron
- 100 neurons in each layer
- 100 epochs
- Random initialization of weights



Which technique can be used to improve model performance in terms of accuracy in the validation set?

- A. Early stopping
- B. Random initialization of weights with appropriate seed
- C. Increasing the number of epochs
- D. Adding another layer with the 100 neurons

**Answer: D**

**NEW QUESTION 2**

A machine learning (ML) specialist is using Amazon SageMaker hyperparameter optimization (HPO) to improve a model's accuracy. The learning rate parameter is specified in the following HPO configuration:

```
{
  "Name": "learning_rate",
  "MaxValue" : "0.0001",
  "MinValue": "0.1"
}
```

During the results analysis, the ML specialist determines that most of the training jobs had a learning rate between 0.01 and 0.1. The best result had a learning rate of less than 0.01. Training jobs need to run regularly over a changing dataset. The ML specialist needs to find a tuning mechanism that uses different learning rates more evenly from the provided range between MinValue and MaxValue.

Which solution provides the MOST accurate result?

- A. Modify the HPO configuration as follows: C:\Users\Admin\Desktop\Data\Odt data\Untitled.jpgSelect the most accurate hyperparameter configuration form this HPO job.

```
{
  "Name": "learning_rate",
  "MaxValue" : "0.0001",
  "MinValue": "0.1"
  "ScalingType": "ReverseLogarithmic"
}
```

- B. Run three different HPO jobs that use different learning rates form the following intervals for MinValue and MaxValue while using the same number of training jobs for each HPO job:[0.01, 0.1][0.001, 0.01][0.0001, 0.001]Select the most accurate hyperparameter configuration form these three HPO jobs.
- C. Modify the HPO configuration as follows: C:\Users\Admin\Desktop\Data\Odt data\Untitled.jpg

```
{
  "Name": "learning_rate",
  "MaxValue" : "0.0001",
  "MinValue": "0.1"
  "ScalingType": "Logarithmic"
}
```

Select the most accurate hyperparameter configuration form this training job.

D. Run three different HPO jobs that use different learning rates form the following intervals for MinValue and MaxValu

E. Divide the number of training jobs for each HPO job by three:[0.01, 0.1][0.001, 0.01][0.0001, 0.001]Select the most accurate hyperparameter configuration form these three HPO jobs.

**Answer: C**

### NEW QUESTION 3

A company will use Amazon SageMaker to train and host a machine learning (ML) model for a marketing campaign. The majority of data is sensitive customer data. The data must be encrypted at rest. The company wants AWS to maintain the root of trust for the master keys and wants encryption key usage to be logged. Which implementation will meet these requirements?

A. Use encryption keys that are stored in AWS Cloud HSM to encrypt the ML data volumes, and to encryptthe model artifacts and data in Amazon S3.

B. Use SageMaker built-in transient keys to encrypt the ML data volume

C. Enable default encryption for new Amazon Elastic Block Store (Amazon EBS) volumes.

D. Use customer managed keys in AWS Key Management Service (AWS KMS) to encrypt the ML data volumes, and to encrypt the model artifacts and data in Amazon S3.

E. Use AWS Security Token Service (AWS STS) to create temporary tokens to encrypt the ML storage volumes, and to encrypt the model artifacts and data in Amazon S3.

**Answer: C**

### NEW QUESTION 4

A manufacturer of car engines collects data from cars as they are being driven The data collected includes timestamp, engine temperature, rotations per minute (RPM), and other sensor readings The company wants to predict when an engine is going to have a problem so it can notify drivers in advance to get engine maintenance The engine data is loaded into a data lake for training

Which is the MOST suitable predictive model that can be deployed into production'?

A. Add labels over time to indicate which engine faults occur at what time in the future to turn this into a supervised learning problem Use a recurrent neural network (RNN) to train the model to recognize when an engine might need maintenance for a certain fault.

B. This data requires an unsupervised learning algorithm Use Amazon SageMaker k-means to cluster the data

C. Add labels over time to indicate which engine faults occur at what time in the future to turn this into a supervised learning problem Use a convolutional neural network (CNN) to train the model to recognize when an engine might need maintenance for a certain fault.

D. This data is already formulated as a time series Use Amazon SageMaker seq2seq to model the time series.

**Answer: B**

### NEW QUESTION 5

A bank wants to launch a low-rate credit promotion. The bank is located in a town that recently experienced economic hardship. Only some of the bank's customers were affected by the crisis, so the bank's credit team must identify which customers to target with the promotion. However, the credit team wants to make sure that loyal customers' full credit history is considered when the decision is made.

The bank's data science team developed a model that classifies account transactions and understands credit eligibility. The data science team used the XGBoost algorithm to train the model. The team used 7 years of bank transaction historical data for training and hyperparameter tuning over the course of several days. The accuracy of the model is sufficient, but the credit team is struggling to explain accurately why the model denies credit to some customers. The credit team has almost no skill in data science.

What should the data science team do to address this issue in the MOST operationally efficient manner?

A. Use Amazon SageMaker Studio to rebuild the mode

B. Create a notebook that uses the XGBoost training container to perform model trainin

C. Deploy the model at an endpoint

D. Enable Amazon SageMaker Model Monitor to store inference

E. Use the inferences to create Shapley values that help explain model behavio

F. Create a chart that shows features and SHapley Additive explanation (SHAP) values to explain to the credit team how the features affect the model outcomes.

G. Use Amazon SageMaker Studio to rebuild the mode

H. Create a notebook that uses the XGBoost training container to perform model trainin

I. Activate Amazon SageMaker Debugger, and configure it to calculate and collect Shapley value

J. Create a chart that shows features and SHapley Additive explanation (SHAP) values to explain to the credit team how the features affect the model outcomes.

K. Create an Amazon SageMaker notebook instanc

L. Use the notebook instance and the XGBoost library to locally retrain the mode

M. Use the plot\_importance() method in the Python XGBoost interface to create a feature importance char

N. Use that chart to explain to the credit team how the features affect the model outcomes.

O. Use Amazon SageMaker Studio to rebuild the mode

P. Create a notebook that uses the XGBoost training container to perform model trainin

Q. Deploy the model at an endpoint

R. Use Amazon SageMakerProcessing to post-analyze the model and create a feature importance explainability chart automatically for the credit team.

**Answer: C**

### NEW QUESTION 6

An e-commerce company needs a customized training model to classify images of its shirts and pants products. The company needs a proof of concept in 2 to 3 days with good accuracy. Which compute choice should the Machine Learning Specialist select to train and achieve good accuracy on the model quickly?

- A. m5.4xlarge (general purpose)
- B. r5.2xlarge (memory optimized)
- C. p3.2xlarge (GPU accelerated computing)
- D. p3.8xlarge (GPU accelerated computing)

**Answer: C**

#### NEW QUESTION 7

A Machine Learning team runs its own training algorithm on Amazon SageMaker. The training algorithm requires external assets. The team needs to submit both its own algorithm code and algorithm-specific parameters to Amazon SageMaker.

What combination of services should the team use to build a custom algorithm in Amazon SageMaker? (Choose two.)

- A. AWS Secrets Manager
- B. AWS CodeStar
- C. Amazon ECR
- D. Amazon ECS
- E. Amazon S3

**Answer: CE**

#### NEW QUESTION 8

A company needs to quickly make sense of a large amount of data and gain insight from it. The data is in different formats, the schemas change frequently, and new data sources are added regularly. The company wants to use AWS services to explore multiple data sources, suggest schemas, and enrich and transform the data. The solution should require the least possible coding effort for the data flows and the least possible infrastructure management.

Which combination of AWS services will meet these requirements?

- A. Amazon EMR for data discovery, enrichment, and transformation; Amazon Athena for querying and analyzing the results in Amazon S3 using standard SQL; Amazon QuickSight for reporting and getting insights
- B. Amazon Kinesis Data Analytics for data ingestion; Amazon EMR for data discovery, enrichment, and transformation; Amazon Redshift for querying and analyzing the results in Amazon S3
- C. AWS Glue for data discovery, enrichment, and transformation; Amazon Athena for querying and analyzing the results in Amazon S3 using standard SQL; Amazon QuickSight for reporting and getting insights
- D. AWS Data Pipeline for data transfer; AWS Step Functions for orchestrating AWS Lambda jobs for data discovery, enrichment, and transformation; Amazon Athena for querying and analyzing the results in Amazon S3 using standard SQL; Amazon QuickSight for reporting and getting insights

**Answer: A**

#### NEW QUESTION 9

A financial services company wants to adopt Amazon SageMaker as its default data science environment. The company's data scientists run machine learning (ML) models on confidential financial data. The company is worried about data egress and wants an ML engineer to secure the environment.

Which mechanisms can the ML engineer use to control data egress from SageMaker? (Choose three.)

- A. Connect to SageMaker by using a VPC interface endpoint powered by AWS PrivateLink.
- B. Use SCPs to restrict access to SageMaker.
- C. Disable root access on the SageMaker notebook instances.
- D. Enable network isolation for training jobs and models.
- E. Restrict notebook presigned URLs to specific IPs used by the company.
- F. Protect data with encryption at rest and in transit.
- G. Use AWS Key Management Service (AWS KMS) to manage encryption keys.

**Answer: BDE**

#### Explanation:

<https://aws.amazon.com/blogs/machine-learning/millennium-management-secure-machine-learning-using-amaz>

#### NEW QUESTION 10

A Data Scientist needs to migrate an existing on-premises ETL process to the cloud. The current process runs at regular time intervals and uses PySpark to combine and format multiple large data sources into a single consolidated output for downstream processing.

The Data Scientist has been given the following requirements for the cloud solution:

- \* Combine multiple data sources
- \* Reuse existing PySpark logic
- \* Run the solution on the existing schedule
- \* Minimize the number of servers that will need to be managed

Which architecture should the Data Scientist use to build this solution?

- A. Write the raw data to Amazon S3. Schedule an AWS Lambda function to submit a Spark step to a persistent Amazon EMR cluster based on the existing schedule. Use the existing PySpark logic to run the ETL job on the EMR cluster. Output the results to a "processed" location in Amazon S3 that is accessible for downstream use.
- B. Write the raw data to Amazon S3. Create an AWS Glue ETL job to perform the ETL processing against the input data. Write the ETL job in PySpark to leverage the existing logic. Create a new AWS Glue trigger to trigger the ETL job based on the existing schedule. Configure the output target of the ETL job to write to a "processed" location in Amazon S3 that is accessible for downstream use.
- C. Write the raw data to Amazon S3. Schedule an AWS Lambda function to run on the existing schedule and process the input data from Amazon S3. Write the Lambda logic in Python and implement the existing PySpark logic to perform the ETL process. Have the Lambda function output the results to a "processed" location in Amazon S3 that is accessible for downstream use.
- D. Use Amazon Kinesis Data Analytics to stream the input data and perform realtime SQL queries against the stream to carry out the required transformations within the stream. Deliver the output results to a "processed" location in Amazon S3 that is accessible for downstream use.

**Answer:** A

#### NEW QUESTION 10

A data scientist is developing a pipeline to ingest streaming web traffic data. The data scientist needs to implement a process to identify unusual web traffic patterns as part of the pipeline. The patterns will be used downstream for alerting and incident response. The data scientist has access to unlabeled historic data to use, if needed.

The solution needs to do the following:

- Calculate an anomaly score for each web traffic entry.
- Adapt unusual event identification to changing web patterns over time. Which approach should the data scientist implement to meet these requirements?

- A. Use historic web traffic data to train an anomaly detection model using the Amazon SageMaker Random Cut Forest (RCF) built-in mode
- B. Use an Amazon Kinesis Data Stream to process the incoming webtrafficdat
- C. Attach a preprocessing AWS Lambda function to perform data enrichment by calling the RCF modelto calculate the anomaly score for each record.
- D. Use historic web traffic data to train an anomaly detection model using the Amazon SageMaker built-inXGBoost mode
- E. Use an Amazon Kinesis Data Stream to process the incoming web traffic dat
- F. Attach apreprocessing AWS Lambda function to perform data enrichment by calling the XGBoost model to calculate the anomaly score for each record.
- G. Collect the streaming data using Amazon Kinesis Data Firehos
- H. Map the delivery stream as an inputsource for Amazon Kinesis Data Analytic
- I. Write a SQL query to run in real time against the streaming datawith the k-Nearest Neighbors (kNN) SQL extension to calculate anomaly scores for each record using a tumbling window.
- J. Collect the streaming data using Amazon Kinesis Data Firehos
- K. Map the delivery stream as an inputsource for Amazon Kinesis Data Analytic
- L. Write a SQL query to run in real time against the streaming datawith the Amazon Random Cut Forest (RCF) SQL extension to calculate anomaly scores for each record using a sliding window.

**Answer:** D

#### NEW QUESTION 14

A retail company wants to combine its customer orders with the product description data from its product catalog. The structure and format of the records in each dataset is different. A data analyst tried to use a spreadsheet to combine the datasets, but the effort resulted in duplicate records and records that were not properly combined. The company needs a solution that it can use to combine similar records from the two datasets and remove any duplicates. Which solution will meet these requirements?

- A. Use an AWS Lambda function to process the dat
- B. Use two arrays to compare equal strings in the fields from the two datasets and remove any duplicates.
- C. Create AWS Glue crawlers for reading and populating the AWS Glue Data Catalo
- D. Call the AWS Glue SearchTables API operation to perform a fuzzy-matching search on the two datasets, and cleanse the data accordingly.
- E. Create AWS Glue crawlers for reading and populating the AWS Glue Data Catalo
- F. Use the FindMatches transform to cleanse the data.
- G. Create an AWS Lake Formation custom transfor
- H. Run a transformation for matching products from the Lake Formation console to cleanse the data automatically.

**Answer:** D

#### NEW QUESTION 17

A manufacturing company has structured and unstructured data stored in an Amazon S3 bucket A Machine Learning Specialist wants to use SQL to run queries on this data. Which solution requires the LEAST effort to be able to query this data?

- A. Use AWS Data Pipeline to transform the data and Amazon RDS to run queries.
- B. Use AWS Glue to catalogue the data and Amazon Athena to run queries
- C. Use AWS Batch to run ETL on the data and Amazon Aurora to run the quenes
- D. Use AWS Lambda to transform the data and Amazon Kinesis Data Analytics to run queries

**Answer:** D

#### NEW QUESTION 19

Amazon Connect has recently been tolled out across a company as a contact call center The solution has been configured to store voice call recordings on Amazon S3

The content of the voice calls are being analyzed for the incidents being discussed by the call operators Amazon Transcribe is being used to convert the audio to text, and the output is stored on Amazon S3

Which approach will provide the information required for further analysis?

- A. Use Amazon Comprehend with the transcribed files to build the key topics
- B. Use Amazon Translate with the transcribed files to train and build a model for the key topics
- C. Use the AWS Deep Learning AMI with Gluon Semantic Segmentation on the transcribed files to train and build a model for the key topics
- D. Use the Amazon SageMaker k-Nearest-Neighbors (kNN) algorithm on the transcribed files to generate a word embeddings dictionary for the key topics

**Answer:** B

#### NEW QUESTION 24

A company supplies wholesale clothing to thousands of retail stores. A data scientist must create a model that predicts the daily sales volume for each item for each store. The data scientist discovers that more than half of the stores have been in business for less than 6 months. Sales data is highly consistent from week to week. Daily data from the database has been aggregated weekly, and weeks with no sales are omitted from the current dataset. Five years (100 MB) of sales data is available in Amazon S3.

Which factors will adversely impact the performance of the forecast model to be developed, and which actions should the data scientist take to mitigate them? (Choose two.)

- A. Detecting seasonality for the majority of stores will be an issue
- B. Request categorical data to relate new stores with similar stores that have more historical data.
- C. The sales data does not have enough variance
- D. Request external sales data from other industries to improve the model's ability to generalize.
- E. Sales data is aggregated by week
- F. Request daily sales data from the source database to enable building a daily model.
- G. The sales data is missing zero entries for item sale
- H. Request that item sales data from the source database include zero entries to enable building the model.
- I. Only 100 MB of sales data is available in Amazon S3. Request 10 years of sales data, which would provide 200 MB of training data for the model.

**Answer:** AB

#### NEW QUESTION 28

A data engineer at a bank is evaluating a new tabular dataset that includes customer data. The data engineer will use the customer data to create a new model to predict customer behavior. After creating a correlation matrix for the variables, the data engineer notices that many of the 100 features are highly correlated with each other.

Which steps should the data engineer take to address this issue? (Choose two.)

- A. Use a linear-based algorithm to train the model.
- B. Apply principal component analysis (PCA).
- C. Remove a portion of highly correlated features from the dataset.
- D. Apply min-max feature scaling to the dataset.
- E. Apply one-hot encoding category-based variables.

**Answer:** BD

#### NEW QUESTION 30

A Data Scientist is developing a machine learning model to classify whether a financial transaction is fraudulent. The labeled data available for training consists of 100,000 non-fraudulent observations and 1,000 fraudulent observations.

The Data Scientist applies the XGBoost algorithm to the data, resulting in the following confusion matrix when the trained model is applied to a previously unseen validation dataset. The accuracy of the model is 99.1%, but the Data Scientist needs to reduce the number of false negatives.

Predicted	0	1
Actual 0	99,966	34
Actual 1	877	123

Which combination of steps should the Data Scientist take to reduce the number of false negative predictions by the model? (Choose two.)

- A. Change the XGBoost eval\_metric parameter to optimize based on Root Mean Square Error (RMSE).
- B. Increase the XGBoost scale\_pos\_weight parameter to adjust the balance of positive and negative weights.
- C. Increase the XGBoost max\_depth parameter because the model is currently underfitting the data.
- D. Change the XGBoost eval\_metric parameter to optimize based on Area Under the ROC Curve (AUC).
- E. Decrease the XGBoost max\_depth parameter because the model is currently overfitting the data.

**Answer:** BD

#### NEW QUESTION 35

A machine learning specialist is developing a regression model to predict rental rates from rental listings. A variable named Wall\_Color represents the most prominent exterior wall color of the property. The following is the sample data, excluding all other variables:

Property_ID	Wall_Color
1000	Red
1001	White
1002	Green

The specialist chose a model that needs numerical input data.

Which feature engineering approaches should the specialist use to allow the regression model to learn from the Wall\_Color data? (Choose two.)

- A. Apply integer transformation and set Red = 1, White = 5, and Green = 10.
- B. Add new columns that store one-hot representation of colors.
- C. Replace the color name string by its length.
- D. Create three columns to encode the color in RGB format.
- E. Replace each color name by its training set frequency.

**Answer:** AD

#### NEW QUESTION 36

A company is building a demand forecasting model based on machine learning (ML). In the development stage, an ML specialist uses an Amazon SageMaker notebook to perform feature engineering during work hours that consumes low amounts of CPU and memory resources. A data engineer uses the same notebook to perform data preprocessing once a day on average that requires very high memory and completes in only 2 hours. The data preprocessing is not configured to use GPU. All the processes are running well on an ml.m5.4xlarge notebook instance.

The company receives an AWS Budgets alert that the billing for this month exceeds the allocated budget. Which solution will result in the MOST cost savings?

- A. Change the notebook instance type to a memory optimized instance with the same vCPU number as the ml.m5.4xlarge instance has
- B. Stop the notebook when it is not in use
- C. Run both data preprocessing and feature engineering development on that instance.
- D. Keep the notebook instance type and size the same
- E. Stop the notebook when it is not in use
- F. Run data preprocessing on a P3 instance type with the same memory as the ml.m5.4xlarge instance by using Amazon SageMaker Processing.
- G. Change the notebook instance type to a smaller general purpose instance

- H. Stop the notebook when it is not in use
- I. Run data preprocessing on an ml.r5 instance with the same memory size as the ml.m5.4xlarge instance by using Amazon SageMaker Processing.
- J. Change the notebook instance type to a smaller general purpose instance
- K. Stop the notebook when it is not in use
- L. Run data preprocessing on an R5 instance with the same memory size as the ml.m5.4xlarge instance by using the Reserved Instance option.

**Answer: B**

#### NEW QUESTION 41

A Data Science team within a large company uses Amazon SageMaker notebooks to access data stored in Amazon S3 buckets. The IT Security team is concerned that internet-enabled notebook instances create a security vulnerability where malicious code running on the instances could compromise data privacy. The company mandates that all instances stay within a secured VPC with no internet access, and data communication traffic must stay within the AWS network. How should the Data Science team configure the notebook instance placement to meet these requirements?

- A. Associate the Amazon SageMaker notebook with a private subnet in a VPC
- B. Place the Amazon SageMaker endpoint and S3 buckets within the same VPC.
- C. Associate the Amazon SageMaker notebook with a private subnet in a VPC
- D. Use IAM policies to grant access to Amazon S3 and Amazon SageMaker.
- E. Associate the Amazon SageMaker notebook with a private subnet in a VPC
- F. Ensure the VPC has S3 VPC endpoints and Amazon SageMaker VPC endpoints attached to it.
- G. Associate the Amazon SageMaker notebook with a private subnet in a VPC
- H. Ensure the VPC has a NAT gateway and an associated security group allowing only outbound connections to Amazon S3 and Amazon SageMaker

**Answer: D**

#### NEW QUESTION 46

A company is launching a new product and needs to build a mechanism to monitor comments about the company and its new product on social media. The company needs to be able to evaluate the sentiment expressed in social media posts, and visualize trends and configure alarms based on various thresholds. The company needs to implement this solution quickly, and wants to minimize the infrastructure and data science resources needed to evaluate the messages. The company already has a solution in place to collect posts and store them within an Amazon S3 bucket. What services should the data science team use to deliver this solution?

- A. Train a model in Amazon SageMaker by using the BlazingText algorithm to detect sentiment in the corpus of social media post
- B. Expose an endpoint that can be called by AWS Lambda
- C. Trigger a Lambda function when posts are added to the S3 bucket to invoke the endpoint and record the sentiment in an Amazon DynamoDB table and in a custom Amazon CloudWatch metric
- D. Use CloudWatch alarms to notify analysts of trends.
- E. Train a model in Amazon SageMaker by using the semantic segmentation algorithm to model the semantic content in the corpus of social media post
- F. Expose an endpoint that can be called by AWS Lambda
- G. Trigger a Lambda function when objects are added to the S3 bucket to invoke the endpoint and record the sentiment in an Amazon DynamoDB table
- H. Schedule a second Lambda function to query recently added records and send an Amazon Simple Notification Service (Amazon SNS) notification to notify analysts of trends.
- I. Trigger an AWS Lambda function when social media posts are added to the S3 bucket
- J. Call Amazon Comprehend for each post to capture the sentiment in the message and record the sentiment in an Amazon DynamoDB table
- K. Schedule a second Lambda function to query recently added records and send an Amazon Simple Notification Service (Amazon SNS) notification to notify analysts of trends.
- L. Trigger an AWS Lambda function when social media posts are added to the S3 bucket
- M. Call Amazon Comprehend for each post to capture the sentiment in the message and record the sentiment in a custom Amazon CloudWatch metric and in S3. Use CloudWatch alarms to notify analysts of trends.

**Answer: A**

#### NEW QUESTION 50

A company wants to create a data repository in the AWS Cloud for machine learning (ML) projects. The company wants to use AWS to perform complete ML lifecycles and wants to use Amazon S3 for the data storage. All of the company's data currently resides on premises and is 40 TB in size. The company wants a solution that can transfer and automatically update data between the on-premises object storage and Amazon S3. The solution must support encryption, scheduling, monitoring, and data integrity validation. Which solution meets these requirements?

- A. Use the S3 sync command to compare the source S3 bucket and the destination S3 bucket
- B. Determine which source files do not exist in the destination S3 bucket and which source files were modified.
- C. Use AWS Transfer for FTPS to transfer the files from the on-premises storage to Amazon S3.
- D. Use AWS DataSync to make an initial copy of the entire dataset
- E. Schedule subsequent incremental transfers of changing data until the final cutover from on premises to AWS.
- F. Use S3 Batch Operations to pull data periodically from the on-premises storage
- G. Enable S3 Versioning on the S3 bucket to protect against accidental overwrites.

**Answer: C**

#### Explanation:

Configure DataSync to make an initial copy of your entire dataset, and schedule subsequent incremental transfers of changing data until the final cut-over from on-premises to AWS.

#### NEW QUESTION 52

A data scientist is working on a public sector project for an urban traffic system. While studying the traffic patterns, it is clear to the data scientist that the traffic behavior at each light is correlated, subject to a small stochastic error term. The data scientist must model the traffic behavior to analyze the traffic patterns and reduce congestion.

How will the data scientist MOST effectively model the problem?

- A. The data scientist should obtain a correlated equilibrium policy by formulating this problem as a multi-agent reinforcement learning problem.
- B. The data scientist should obtain the optimal equilibrium policy by formulating this problem as a single-agent reinforcement learning problem.
- C. Rather than finding an equilibrium policy, the data scientist should obtain accurate predictors of traffic flow by using historical data through a supervised learning approach.
- D. Rather than finding an equilibrium policy, the data scientist should obtain accurate predictors of traffic flow by using unlabeled simulated data representing the new traffic patterns in the city and applying an unsupervised learning approach.

**Answer:** D

#### NEW QUESTION 57

A credit card company wants to build a credit scoring model to help predict whether a new credit card applicant will default on a credit card payment. The company has collected data from a large number of sources with thousands of raw attributes. Early experiments to train a classification model revealed that many attributes are highly correlated, the large number of features slows down the training speed significantly, and that there are some overfitting issues. The Data Scientist on this project would like to speed up the model training time without losing a lot of information from the original dataset. Which feature engineering technique should the Data Scientist use to meet the objectives?

- A. Run self-correlation on all features and remove highly correlated features
- B. Normalize all numerical values to be between 0 and 1
- C. Use an autoencoder or principal component analysis (PCA) to replace original features with new features
- D. Cluster raw data using k-means and use sample data from each cluster to build a new dataset

**Answer:** B

#### NEW QUESTION 61

A company is running an Amazon SageMaker training job that will access data stored in its Amazon S3 bucket. A compliance policy requires that the data never be transmitted across the internet. How should the company set up the job?

- A. Launch the notebook instances in a public subnet and access the data through the public S3 endpoint
- B. Launch the notebook instances in a private subnet and access the data through a NAT gateway
- C. Launch the notebook instances in a public subnet and access the data through a NAT gateway
- D. Launch the notebook instances in a private subnet and access the data through an S3 VPC endpoint.

**Answer:** D

#### NEW QUESTION 64

A Data Engineer needs to build a model using a dataset containing customer credit card information. How can the Data Engineer ensure the data remains encrypted and the credit card information is secure?

- A. Use a custom encryption algorithm to encrypt the data and store the data on an Amazon SageMaker instance in a VP
- B. Use the SageMaker DeepAR algorithm to randomize the credit card numbers.
- C. Use an IAM policy to encrypt the data on the Amazon S3 bucket and Amazon Kinesis to automatically discard credit card numbers and insert fake credit card numbers.
- D. Use an Amazon SageMaker launch configuration to encrypt the data once it is copied to the SageMaker instance in a VP
- E. Use the SageMaker principal component analysis (PCA) algorithm to reduce the length of the credit card numbers.
- F. Use AWS KMS to encrypt the data on Amazon S3 and Amazon SageMaker, and redact the credit card numbers from the customer data with AWS Glue.

**Answer:** D

#### NEW QUESTION 68

A monitoring service generates 1 TB of scale metrics record data every minute. A Research team performs queries on this data using Amazon Athena. The queries run slowly due to the large volume of data, and the team requires better performance. How should the records be stored in Amazon S3 to improve query performance?

- A. CSV files
- B. Parquet files
- C. Compressed JSON
- D. RecordIO

**Answer:** D

#### NEW QUESTION 71

A Mobile Network Operator is building an analytics platform to analyze and optimize a company's operations using Amazon Athena and Amazon S3. The source systems send data in CSV format in real time. The Data Engineering team wants to transform the data to the Apache Parquet format before storing it on Amazon S3. Which solution takes the LEAST effort to implement?

- A. Ingest .CSV data using Apache Kafka Streams on Amazon EC2 instances and use Kafka Connect S3 to serialize data as Parquet
- B. Ingest .CSV data from Amazon Kinesis Data Streams and use Amazon Glue to convert data into Parquet.
- C. Ingest .CSV data using Apache Spark Structured Streaming in an Amazon EMR cluster and use Apache Spark to convert data into Parquet.
- D. Ingest .CSV data from Amazon Kinesis Data Streams and use Amazon Kinesis Data Firehose to convert data into Parquet.

**Answer:** B

#### Explanation:

<https://medium.com/search/convert-csv-json-files-to-apache-parquet-using-aws-glue-a760d177b45f> <https://github.com/ecloudvalley/Building-a-Data-Lake-with-AWS-Glue-and-Amazon-S3>

### NEW QUESTION 73

A financial services company is building a robust serverless data lake on Amazon S3. The data lake should be flexible and meet the following requirements:

- \* Support querying old and new data on Amazon S3 through Amazon Athena and Amazon Redshift Spectrum.
- \* Support event-driven ETL pipelines.
- \* Provide a quick and easy way to understand metadata. Which approach meets these requirements?

- A. Use an AWS Glue crawler to crawl S3 data, an AWS Lambda function to trigger an AWS Glue ETL job, and an AWS Glue Data catalog to search and discover metadata.
- B. Use an AWS Glue crawler to crawl S3 data, an AWS Lambda function to trigger an AWS Batch job, and an external Apache Hive metastore to search and discover metadata.
- C. Use an AWS Glue crawler to crawl S3 data, an Amazon CloudWatch alarm to trigger an AWS Batch job, and an AWS Glue Data Catalog to search and discover metadata.
- D. Use an AWS Glue crawler to crawl S3 data, an Amazon CloudWatch alarm to trigger an AWS Glue ETL job, and an external Apache Hive metastore to search and discover metadata.

**Answer: A**

### NEW QUESTION 76

A Machine Learning Specialist works for a credit card processing company and needs to predict which transactions may be fraudulent in near-real time.

Specifically, the Specialist must train a model that returns the probability that a given transaction may be fraudulent.

How should the Specialist frame this business problem?

- A. Streaming classification
- B. Binary classification
- C. Multi-category classification
- D. Regression classification

**Answer: C**

### NEW QUESTION 81

A company uses camera images of the tops of items displayed on store shelves to determine which items were removed and which ones still remain. After several hours of data labeling, the company has a total of 1,000 hand-labeled images covering 10 distinct items. The training results were poor.

Which machine learning approach fulfills the company's long-term needs?

- A. Convert the images to grayscale and retrain the model
- B. Reduce the number of distinct items from 10 to 2, build the model, and iterate
- C. Attach different colored labels to each item, take the images again, and build the model
- D. Augment training data for each item using image variants like inversions and translations, build the model, and iterate.

**Answer: A**

### NEW QUESTION 82

A health care company is planning to use neural networks to classify their X-ray images into normal and abnormal classes. The labeled data is divided into a training set of 1,000 images and a test set of 200 images. The initial training of a neural network model with 50 hidden layers yielded 99% accuracy on the training set, but only 55% accuracy on the test set.

What changes should the Specialist consider to solve this issue? (Choose three.)

- A. Choose a higher number of layers
- B. Choose a lower number of layers
- C. Choose a smaller learning rate
- D. Enable dropout
- E. Include all the images from the test set in the training set
- F. Enable early stopping

**Answer: ADE**

### NEW QUESTION 84

A Machine Learning Specialist needs to be able to ingest streaming data and store it in Apache Parquet files for exploration and analysis. Which of the following services would both ingest and store this data in the correct format?

- A. AWS DMS
- B. Amazon Kinesis Data Streams
- C. Amazon Kinesis Data Firehose
- D. Amazon Kinesis Data Analytics

**Answer: C**

### NEW QUESTION 87

A company has set up and deployed its machine learning (ML) model into production with an endpoint using Amazon SageMaker hosting services. The ML team has configured automatic scaling for its SageMaker instances to support workload changes. During testing, the team notices that additional instances are being launched before the new instances are ready. This behavior needs to change as soon as possible.

How can the ML team solve this issue?

- A. Decrease the cooldown period for the scale-in activity
- B. Increase the configured maximum capacity of instances.
- C. Replace the current endpoint with a multi-model endpoint using SageMaker.
- D. Set up Amazon API Gateway and AWS Lambda to trigger the SageMaker inference endpoint.
- E. Increase the cooldown period for the scale-out activity.

**Answer:** A

**NEW QUESTION 89**

A company offers an online shopping service to its customers. The company wants to enhance the site's security by requesting additional information when customers access the site from locations that are different from their normal location. The company wants to update the process to call a machine learning (ML) model to determine when additional information should be requested. The company has several terabytes of data from its existing ecommerce web servers containing the source IP addresses for each request made to the web server. For authenticated requests, the records also contain the login name of the requesting user. Which approach should an ML specialist take to implement the new security feature in the web application?

- A. Use Amazon SageMaker Ground Truth to label each record as either a successful or failed access attempt
- B. Use Amazon SageMaker to train a binary classification model using the factorization machines (FM) algorithm.
- C. Use Amazon SageMaker to train a model using the IP Insights algorithm
- D. Schedule updates and retraining of the model using new log data nightly.
- E. Use Amazon SageMaker Ground Truth to label each record as either a successful or failed access attempt
- F. Use Amazon SageMaker to train a binary classification model using the IP Insights algorithm.
- G. Use Amazon SageMaker to train a model using the Object2Vec algorithm
- H. Schedule updates and retraining of the model using new log data nightly.

**Answer:** C

**NEW QUESTION 90**

A large consumer goods manufacturer has the following products on sale

- 34 different toothpaste variants
- 48 different toothbrush variants
- 43 different mouthwash variants

The entire sales history of all these products is available in Amazon S3 Currently, the company is using custom-built autoregressive integrated moving average (ARIMA) models to forecast demand for these products The company wants to predict the demand for a new product that will soon be launched Which solution should a Machine Learning Specialist apply?

- A. Train a custom ARIMA model to forecast demand for the new product.
- B. Train an Amazon SageMaker DeepAR algorithm to forecast demand for the new product
- C. Train an Amazon SageMaker k-means clustering algorithm to forecast demand for the new product.
- D. Train a custom XGBoost model to forecast demand for the new product

**Answer:** B

**Explanation:**

The Amazon SageMaker DeepAR forecasting algorithm is a supervised learning algorithm for forecasting scalar (one-dimensional) time series using recurrent neural networks (RNN). Classical forecasting methods, such as autoregressive integrated moving average (ARIMA) or exponential smoothing (ETS), fit a single model to each individual time series. They then use that model to extrapolate the time series into the future.

**NEW QUESTION 94**

A Machine Learning Specialist receives customer data for an online shopping website. The data includes demographics, past visits, and locality information. The Specialist must develop a machine learning approach to identify the customer shopping patterns, preferences and trends to enhance the website for better service and smart recommendations. Which solution should the Specialist recommend?

- A. Latent Dirichlet Allocation (LDA) for the given collection of discrete data to identify patterns in the customer database.
- B. A neural network with a minimum of three layers and random initial weights to identify patterns in the customer database
- C. Collaborative filtering based on user interactions and correlations to identify patterns in the customer database
- D. Random Cut Forest (RCF) over random subsamples to identify patterns in the customer database

**Answer:** C

**NEW QUESTION 96**

Which of the following metrics should a Machine Learning Specialist generally use to compare/evaluate machine learning classification models against each other?

- A. Recall
- B. Misclassification rate
- C. Mean absolute percentage error (MAPE)
- D. Area Under the ROC Curve (AUC)

**Answer:** D

**NEW QUESTION 101**

A data scientist wants to use Amazon Forecast to build a forecasting model for inventory demand for a retail company. The company has provided a dataset of historic inventory demand for its products as a .csv file stored in an Amazon S3 bucket. The table below shows a sample of the dataset.

timestamp	item_id	demand	category	lead_time
2019-12-14	uni_000736	120	hardware	90
2020-01-31	uni_003429	98	hardware	30
2020-03-04	uni_000211	234	accessories	10

How should the data scientist transform the data?

- A. Use ETL jobs in AWS Glue to separate the dataset into a target time series dataset and an item metadata dataset
- B. Upload both datasets as .csv files to Amazon S3.
- C. Use a Jupyter notebook in Amazon SageMaker to separate the dataset into a related time series dataset and an item metadata dataset
- D. Upload both datasets as tables in Amazon Aurora.
- E. Use AWS Batch jobs to separate the dataset into a target time series dataset, a related time series dataset, and an item metadata dataset
- F. Upload them directly to Forecast from a local machine.
- G. Use a Jupyter notebook in Amazon SageMaker to transform the data into the optimized protobuf recordIO format
- H. Upload the dataset in this format to Amazon S3.

**Answer:** A

**Explanation:**

<https://docs.aws.amazon.com/forecast/latest/dg/dataset-import-guidelines-troubleshooting.html>

**NEW QUESTION 105**

A Machine Learning Specialist kicks off a hyperparameter tuning job for a tree-based ensemble model using Amazon SageMaker with Area Under the ROC Curve (AUC) as the objective metric. This workflow will eventually be deployed in a pipeline that retrains and tunes hyperparameters each night to model click-through on data that goes stale every 24 hours.

With the goal of decreasing the amount of time it takes to train these models, and ultimately to decrease costs, the Specialist wants to reconfigure the input hyperparameter range(s).

Which visualization will accomplish this?

- A. A histogram showing whether the most important input feature is Gaussian.
- B. A scatter plot with points colored by target variable that uses t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the large number of input variables in an easier-to-read dimension.
- C. A scatter plot showing the performance of the objective metric over each training iteration.
- D. A scatter plot showing the correlation between maximum tree depth and the objective metric.

**Answer:** D

**NEW QUESTION 108**

A logistics company needs a forecast model to predict next month's inventory requirements for a single item in 10 warehouses. A machine learning specialist uses Amazon Forecast to develop a forecast model from 3 years of monthly data. There is no missing data. The specialist selects the DeepAR+ algorithm to train a predictor. The predictor means absolute percentage error (MAPE) is much larger than the MAPE produced by the current human forecasters.

Which changes to the CreatePredictor API call could improve the MAPE? (Choose two.)

- A. Set PerformAutoML to true.
- B. Set ForecastHorizon to 4.
- C. Set ForecastFrequency to W for weekly.
- D. Set PerformHPO to true.
- E. Set FeaturizationMethodName to filling.

**Answer:** CD

**NEW QUESTION 111**

A Machine Learning Specialist is applying a linear least squares regression model to a dataset with 1,000 records and 50 features. Prior to training, the ML Specialist notices that two features are perfectly linearly dependent.

Why could this be an issue for the linear least squares regression model?

- A. It could cause the backpropagation algorithm to fail during training.
- B. It could create a singular matrix during optimization which fails to define a unique solution.
- C. It could modify the loss function during optimization causing it to fail during training.
- D. It could introduce non-linear dependencies within the data which could invalidate the linear assumptions of the model.

**Answer:** C

**NEW QUESTION 114**

A data scientist needs to identify fraudulent user accounts for a company's ecommerce platform. The company wants the ability to determine if a newly created account is associated with a previously known fraudulent user. The data scientist is using AWS Glue to cleanse the company's application logs during ingestion.

Which strategy will allow the data scientist to identify fraudulent accounts?

- A. Execute the built-in FindDuplicates Amazon Athena query.
- B. Create a FindMatches machine learning transform in AWS Glue.
- C. Create an AWS Glue crawler to infer duplicate accounts in the source data.
- D. Search for duplicate accounts in the AWS Glue Data Catalog.

**Answer:** B

**NEW QUESTION 117**

A data scientist is using the Amazon SageMaker Neural Topic Model (NTM) algorithm to build a model that recommends tags from blog posts. The raw blog post data is stored in an Amazon S3 bucket in JSON format. During model evaluation, the data scientist discovered that the model recommends certain stopwords such as "a," "an," and "the" as tags to certain blog posts, along with a few rare words that are present only in certain blog entries. After a few iterations of tag review with the content team, the data scientist notices that the rare words are unusual but feasible. The data scientist also must ensure that the tag recommendations of the generated model do not include the stopwords.

What should the data scientist do to meet these requirements?

- A. Use the Amazon Comprehend entity recognition API operation.
- B. Remove the detected words from the blog post data.

- C. Replace the blog post data source in the S3 bucket.
- D. Run the SageMaker built-in principal component analysis (PCA) algorithm with the blog post data from the S3 bucket as the data source.
- E. Replace the blog post data in the S3 bucket with the results of the training job.
- F. Use the SageMaker built-in Object Detection algorithm instead of the NTM algorithm for the training job to process the blog post data.
- G. Remove the stopwords from the blog post data by using the Count Vectorizer function in the scikit-learn library.
- H. Replace the blog post data in the S3 bucket with the results of the vectorizer.

**Answer:** D

#### NEW QUESTION 121

A manufacturer is operating a large number of factories with a complex supply chain relationship where unexpected downtime of a machine can cause production to stop at several factories. A data scientist wants to analyze sensor data from the factories to identify equipment in need of preemptive maintenance and then dispatch a service team to prevent unplanned downtime. The sensor readings from a single machine can include up to 200 data points including temperatures, voltages, vibrations, RPMs, and pressure readings.

To collect this sensor data, the manufacturer deployed Wi-Fi and LANs across the factories. Even though many factory locations do not have reliable or high-speed internet connectivity, the manufacturer would like to maintain near-real-time inference capabilities.

Which deployment architecture for the model will address these business requirements?

- A. Deploy the model in Amazon SageMaker
- B. Run sensor data through this model to predict which machines need maintenance.
- C. Deploy the model on AWS IoT Greengrass in each factory
- D. Run sensor data through this model to infer which machines need maintenance.
- E. Deploy the model to an Amazon SageMaker batch transformation job
- F. Generate inferences in a daily batch report to identify machines that need maintenance.
- G. Deploy the model in Amazon SageMaker and use an IoT rule to write data to an Amazon DynamoDB table. Consume a DynamoDB stream from the table with an AWS Lambda function to invoke the endpoint.

**Answer:** B

#### Explanation:

<https://aws.amazon.com/blogs/iot/industrial-iot-from-condition-based-monitoring-to-predictive-quality-to-digitize/> <https://aws.amazon.com/blogs/iot/using-aws-iot-for-predictive-maintenance/>

#### NEW QUESTION 123

A Machine Learning Specialist is required to build a supervised image-recognition model to identify a cat. The ML Specialist performs some tests and records the following results for a neural network-based image classifier:

Total number of images available = 1,000 Test set images = 100 (constant test set)

The ML Specialist notices that, in over 75% of the misclassified images, the cats were held upside down by their owners.

Which techniques can be used by the ML Specialist to improve this specific test error?

- A. Increase the training data by adding variation in rotation for training images.
- B. Increase the number of epochs for model training.
- C. Increase the number of layers for the neural network.
- D. Increase the dropout rate for the second-to-last layer.

**Answer:** A

#### NEW QUESTION 125

A Machine Learning Specialist is planning to create a long-running Amazon EMR cluster. The EMR cluster will have 1 master node, 10 core nodes, and 20 task nodes. To save on costs, the Specialist will use Spot Instances in the EMR cluster.

Which nodes should the Specialist launch on Spot Instances?

- A. Master node
- B. Any of the core nodes
- C. Any of the task nodes
- D. Both core and task nodes

**Answer:** A

#### NEW QUESTION 129

A city wants to monitor its air quality to address the consequences of air pollution. A Machine Learning Specialist needs to forecast the air quality in parts per million of contaminants for the next 2 days in the city. As this is a prototype, only daily data from the last year is available.

Which model is MOST likely to provide the best results in Amazon SageMaker?

- A. Use the Amazon SageMaker k-Nearest-Neighbors (kNN) algorithm on the single time series consisting of the full year of data with a predictor\_type of regressor.
- B. Use Amazon SageMaker Random Cut Forest (RCF) on the single time series consisting of the full year of data.
- C. Use the Amazon SageMaker Linear Learner algorithm on the single time series consisting of the full year of data with a predictor\_type of regressor.
- D. Use the Amazon SageMaker Linear Learner algorithm on the single time series consisting of the full year of data with a predictor\_type of classifier.

**Answer:** C

#### NEW QUESTION 131

.....

## Thank You for Trying Our Product

### We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questions and Answers in PDF Format

### AWS-Certified-Machine-Learning-Specialty Practice Exam Features:

- \* AWS-Certified-Machine-Learning-Specialty Questions and Answers Updated Frequently
- \* AWS-Certified-Machine-Learning-Specialty Practice Questions Verified by Expert Senior Certified Staff
- \* AWS-Certified-Machine-Learning-Specialty Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- \* AWS-Certified-Machine-Learning-Specialty Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

**100% Actual & Verified — Instant Download, Please Click**  
[Order The AWS-Certified-Machine-Learning-Specialty Practice Test Here](#)