



Amazon

Exam Questions AWS-Certified-Machine-Learning-Specialty

AWS Certified Machine Learning - Specialty

NEW QUESTION 1

A Data Scientist needs to create a serverless ingestion and analytics solution for high-velocity, real-time streaming data. The ingestion process must buffer and convert incoming records from JSON to a query-optimized, columnar format without data loss. The output datastore must be highly available, and Analysts must be able to run SQL queries against the data and connect to existing business intelligence dashboards. Which solution should the Data Scientist build to satisfy the requirements?

- A. Create a schema in the AWS Glue Data Catalog of the incoming data format.
- B. Use an Amazon Kinesis Data Firehose delivery stream to stream the data and transform the data to Apache Parquet or ORC format using the AWS Glue Data Catalog before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.
- C. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and writes the data to a processed data location in Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.
- D. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and inserts it into an Amazon RDS PostgreSQL database.
- E. Have the Analysts query and run dashboards from the RDS database.
- F. Use Amazon Kinesis Data Analytics to ingest the streaming data and perform real-time SQL queries to convert the records to Apache Parquet before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.

Answer: A

NEW QUESTION 2

A Machine Learning Specialist observes several performance problems with the training portion of a machine learning solution on Amazon SageMaker. The solution uses a large training dataset 2 TB in size and is using the SageMaker k-means algorithm. The observed issues include the unacceptable length of time it takes before the training job launches and poor I/O throughput while training the model. What should the Specialist do to address the performance issues with the current solution?

- A. Use the SageMaker batch transform feature.
- B. Compress the training data into Apache Parquet format.
- C. Ensure that the input mode for the training job is set to Pipe.
- D. Copy the training dataset to an Amazon EFS volume mounted on the SageMaker instance.

Answer: B

NEW QUESTION 3

A company is building a new version of a recommendation engine. Machine learning (ML) specialists need to keep adding new data from users to improve personalized recommendations. The ML specialists gather data from the users' interactions on the platform and from sources such as external websites and social media. The pipeline cleans, transforms, enriches, and compresses terabytes of data daily, and this data is stored in Amazon S3. A set of Python scripts was coded to do the job and is stored in a large Amazon EC2 instance. The whole process takes more than 20 hours to finish, with each script taking at least an hour. The company wants to move the scripts out of Amazon EC2 into a more managed solution that will eliminate the need to maintain servers. Which approach will address all of these requirements with the LEAST development effort?

- A. Load the data into an Amazon Redshift cluster.
- B. Execute the pipeline by using SQS.
- C. Store the results in Amazon S3.
- D. Load the data into Amazon DynamoDB.
- E. Convert the scripts to an AWS Lambda function.
- F. Execute the pipeline by triggering Lambda execution.
- G. Store the results in Amazon S3.
- H. Create an AWS Glue job.
- I. Convert the scripts to PySpark.
- J. Execute the pipeline.
- K. Store the results in Amazon S3.
- L. Create a set of individual AWS Lambda functions to execute each of the scripts.
- M. Build a step function by using the AWS Step Functions Data Science SDK.
- N. Store the results in Amazon S3.

Answer: B

NEW QUESTION 4

A Marketing Manager at a pet insurance company plans to launch a targeted marketing campaign on social media to acquire new customers. Currently, the company has the following data in Amazon Aurora:

- Profiles for all past and existing customers
- Profiles for all past and existing insured pets
- Policy-level information
- Premiums received
- Claims paid

What steps should be taken to implement a machine learning model to identify potential new customers on social media?

- A. Use regression on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media.
- B. Use clustering on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media.
- C. Use a recommendation engine on customer profile data to understand key characteristics of consumer segments.
- D. Find similar profiles on social media.
- E. Use a decision tree classifier engine on customer profile data to understand key characteristics of consumer segments.
- F. Find similar profiles on social media.

Answer: C

NEW QUESTION 5

A company ingests machine learning (ML) data from web advertising clicks into an Amazon S3 data lake. Click data is added to an Amazon Kinesis data stream by using the Kinesis Producer Library (KPL). The data is loaded into the S3 data lake from the data stream by using an Amazon Kinesis Data Firehose delivery stream. As the data volume increases, an ML specialist notices that the rate of data ingested into Amazon S3 is relatively constant. There also is an increasing backlog of data for Kinesis Data Streams and Kinesis Data Firehose to ingest. Which next step is MOST likely to improve the data ingestion rate into Amazon S3?

- A. Increase the number of S3 prefixes for the delivery stream to write to.
- B. Decrease the retention period for the data stream.
- C. Increase the number of shards for the data stream.
- D. Add more consumers using the Kinesis Client Library (KCL).

Answer: C

NEW QUESTION 6

An office security agency conducted a successful pilot using 100 cameras installed at key locations within the main office. Images from the cameras were uploaded to Amazon S3 and tagged using Amazon Rekognition, and the results were stored in Amazon ES. The agency is now looking to expand the pilot into a full production system using thousands of video cameras in its office locations globally. The goal is to identify activities performed by non-employees in real time. Which solution should the agency consider?

- A. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream.
- B. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection of known employees, and alert when non-employees are detected.
- C. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream.
- D. On each stream, use Amazon Rekognition Image to detect faces from a collection of known employees and alert when non-employees are detected.
- E. Install AWS DeepLens cameras and use the DeepLens_Kinesis_Video module to stream video to Amazon Kinesis Video Streams for each camera.
- F. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection on each stream, and alert when nonemployees are detected.
- G. Install AWS DeepLens cameras and use the DeepLens_Kinesis_Video module to stream video to Amazon Kinesis Video Streams for each camera.
- H. On each stream, run an AWS Lambda function to capture image fragments and then call Amazon Rekognition Image to detect faces from a collection of known employees, and alert when non-employees are detected.

Answer: C

NEW QUESTION 7

An agency collects census information within a country to determine healthcare and social program needs by province and city. The census form collects responses for approximately 500 questions from each citizen. Which combination of algorithms would provide the appropriate insights? (Select TWO)

- A. The factorization machines (FM) algorithm
- B. The Latent Dirichlet Allocation (LDA) algorithm
- C. The principal component analysis (PCA) algorithm
- D. The k-means algorithm
- E. The Random Cut Forest (RCF) algorithm

Answer: CD

Explanation:

The PCA and K-means algorithms are useful in collection of data using census form.

NEW QUESTION 8

A medical imaging company wants to train a computer vision model to detect areas of concern on patients' CT scans. The company has a large collection of unlabeled CT scans that are linked to each patient and stored in an Amazon S3 bucket. The scans must be accessible to authorized users only. A machine learning engineer needs to build a labeling pipeline. Which set of steps should the engineer take to build the labeling pipeline with the LEAST effort?

- A. Create a workforce with AWS Identity and Access Management (IAM). Build a labeling tool on Amazon EC2. Queue images for labeling by using Amazon Simple Queue Service (Amazon SQS). Write the labeling instructions.
- B. Create an Amazon Mechanical Turk workforce and manifest file.
- C. Create a labeling job by using the built-in image classification task type in Amazon SageMaker Ground Truth.
- D. Write the labeling instructions.
- E. Create a private workforce and manifest file.
- F. Create a labeling job by using the built-in bounding box task type in Amazon SageMaker Ground Truth.
- G. Write the labeling instructions.
- H. Create a workforce with Amazon Cognito.
- I. Build a labeling web application with AWS Amplify.
- J. Build a labeling workflow backend using AWS Lambda.
- K. Write the labeling instructions.

Answer: C

Explanation:

<https://docs.aws.amazon.com/sagemaker/latest/dg/sms-workforce-private.html>

NEW QUESTION 9

A web-based company wants to improve its conversion rate on its landing page. Using a large historical dataset of customer visits, the company has repeatedly

trained a multi-class deep learning network algorithm on Amazon SageMaker However there is an overfitting problem training data shows 90% accuracy in predictions, while test data shows 70% accuracy only

The company needs to boost the generalization of its model before deploying it into production to maximize conversions of visits to purchases Which action is recommended to provide the HIGHEST accuracy model for the company's test and validation data?

- A. Increase the randomization of training data in the mini-batches used in training.
- B. Allocate a higher proportion of the overall data to the training dataset
- C. Apply L1 or L2 regularization and dropouts to the training.
- D. Reduce the number of layers and units (or neurons) from the deep learning network.

Answer: C

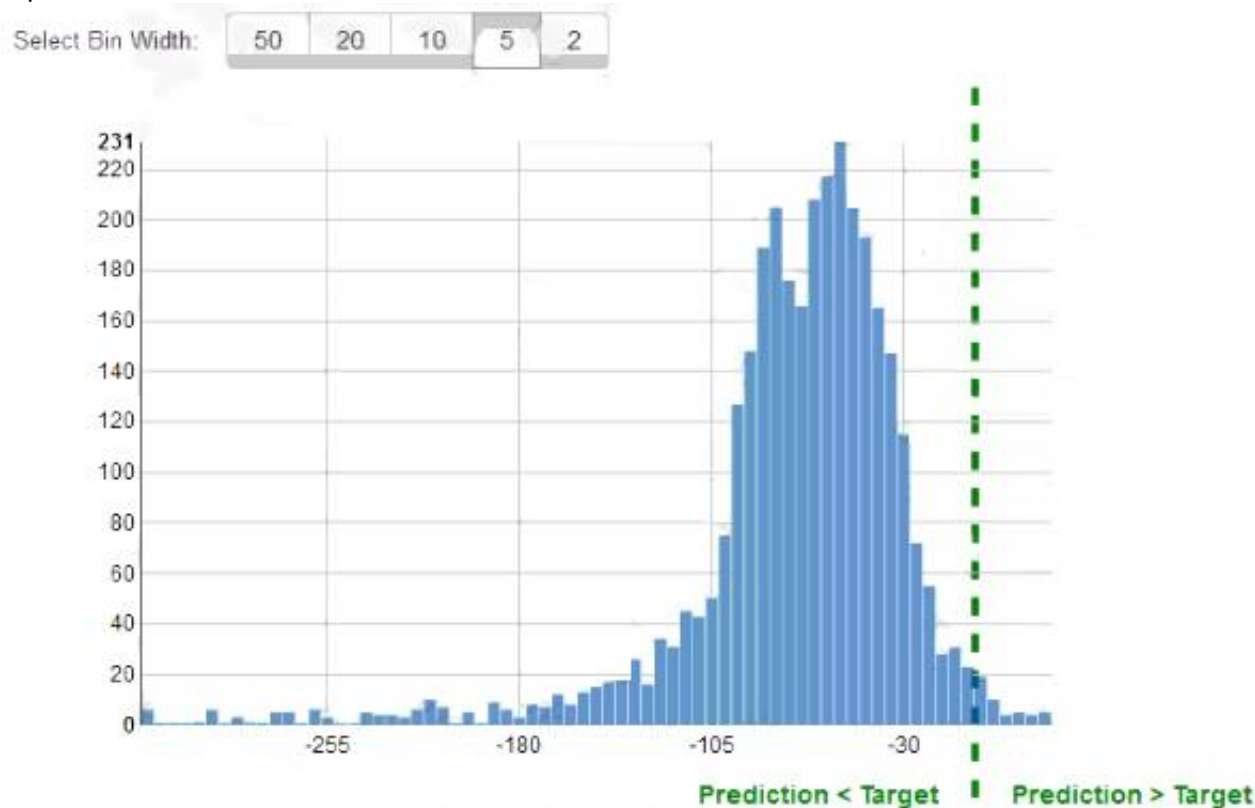
Explanation:

If this is a ComputerVision problem augmentation can help and we may consider A an option. However in analyzing customer historic data, there is no easy way to increase randomization in training. If you go deep into modelling and coding. When you build model with tensorflow/pytorch, most of the time the trainloader is already sampling in data in random manner (with shuffle enable). What we usually do to reduce overfitting is by adding dropout.

<https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html>

NEW QUESTION 10

While reviewing the histogram for residuals on regression evaluation data a Machine Learning Specialist notices that the residuals do not form a zero-centered bell shape as shown What does this mean?



- A. The model might have prediction errors over a range of target values.
- B. The dataset cannot be accurately represented using the regression model
- C. There are too many variables in the model
- D. The model is predicting its target values perfectly.

Answer: D

NEW QUESTION 10

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 12

A large mobile network operating company is building a machine learning model to predict customers who are likely to unsubscribe from the service. The company plans to offer an incentive for these customers as the cost of churn is far greater than the cost of the incentive.

The model produces the following confusion matrix after evaluating on a test dataset of 100 customers: Based on the model evaluation results, why is this a viable model for production?

| n = 100 | PREDICTED CHURN | |
|------------------|-----------------|----|
| | Yes | No |
| ACTUAL Churn Yes | 10 | 4 |
| Actual No | 10 | 76 |

A. The model is 86% accurate and the cost incurred by the company as a result of false negatives is less than the false positives.

B. The precision of the model is 86%, which is less than the accuracy of the model.

C. The model is 86% accurate and the cost incurred by the company as a result of false positives is less than the false negatives.

D. The precision of the model is 86%, which is greater than the accuracy of the model.

Answer: A

NEW QUESTION 13

A Machine Learning Specialist is using Apache Spark for pre-processing training data As part of the Spark pipeline, the Specialist wants to use Amazon SageMaker for training a model and hosting it Which of the following would the Specialist do to integrate the Spark application with SageMaker? (Select THREE)

A. Download the AWS SDK for the Spark environment

B. Install the SageMaker Spark library in the Spark environment.

C. Use the appropriate estimator from the SageMaker Spark Library to train a model.

D. Compress the training data into a ZIP file and upload it to a pre-defined Amazon S3 bucket.

E. Use the sageMakerMode

F. transform method to get inferences from the model hosted in SageMaker

G. Convert the DataFrame object to a CSV file, and use the CSV file as input for obtaining inferences from SageMaker.

Answer: DEF

NEW QUESTION 15

A Machine Learning Specialist is preparing data for training on Amazon SageMaker The Specialist is transformed into a numpy .array, which appears to be negatively affecting the speed of the training

What should the Specialist do to optimize the data for training on SageMaker'?

A. Use the SageMaker batch transform feature to transform the training data into a DataFrame

B. Use AWS Glue to compress the data into the Apache Parquet format

C. Transform the dataset into the Recordio protobuf format

D. Use the SageMaker hyperparameter optimization feature to automatically optimize the data

Answer: C

NEW QUESTION 18

The Chief Editor for a product catalog wants the Research and Development team to build a machine learning system that can be used to detect whether or not individuals in a collection of images are wearing the company's retail brand The team has a set of training data

Which machine learning algorithm should the researchers use that BEST meets their requirements?

A. Latent Dirichlet Allocation (LDA)

B. Recurrent neural network (RNN)

C. K-means

D. Convolutional neural network (CNN)

Answer: C

NEW QUESTION 19

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 24

A Machine Learning Specialist deployed a model that provides product recommendations on a company's website. Initially, the model was performing very well and resulted in customers buying more products on average. However, within the past few months, the Specialist has noticed that the effect of product recommendations has diminished and customers are starting to return to their original habits of spending less. The Specialist is unsure of what happened, as the model has not changed from its initial deployment over a year ago. Which method should the Specialist try to improve model performance?

- A. The model needs to be completely re-engineered because it is unable to handle product inventory changes.
- B. The model's hyperparameters should be periodically updated to prevent drift.
- C. The model should be periodically retrained from scratch using the original data while adding a regularization term to handle product inventory changes.
- D. The model should be periodically retrained using the original training data plus new data as product inventory changes.

Answer: D

NEW QUESTION 29

A company wants to use automatic speech recognition (ASR) to transcribe messages that are less than 60 seconds long from a voicemail-style application. The company requires the correct identification of 200 unique product names, some of which have unique spellings or pronunciations. The company has 4,000 words of Amazon SageMaker Ground Truth voicemail transcripts it can use to customize the chosen ASR model. The company needs to ensure that everyone can update their customizations multiple times each hour. Which approach will maximize transcription accuracy during the development phase?

- A. Use a voice-driven Amazon Lex bot to perform the ASR customization.
- B. Create customer slots within the bot that specifically identify each of the required product names.
- C. Use the Amazon Lex synonym mechanism to provide additional variations of each product name as mis-transcriptions are identified in development.
- D. Use Amazon Transcribe to perform the ASR customization.
- E. Analyze the word confidence scores in the transcript, and automatically create or update a custom vocabulary file with any word that has a confidence score below an acceptable threshold value.
- F. Use this updated custom vocabulary file in all future transcription tasks.
- G. Create a custom vocabulary file containing each product name with phonetic pronunciations, and use it with Amazon Transcribe to perform the ASR customization.
- H. Analyze the transcripts and manually update the custom vocabulary file to include updated or additional entries for those names that are not being correctly identified.
- I. Use the audio transcripts to create a training dataset and build an Amazon Transcribe custom language model.
- J. Analyze the transcripts and update the training dataset with a manually corrected version of transcripts where product names are not being transcribed correctly.
- K. Create an updated custom language model.

Answer: A

NEW QUESTION 34

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 37

A machine learning (ML) specialist wants to secure calls to the Amazon SageMaker Service API. The specialist has configured Amazon VPC with a VPC interface endpoint for the Amazon SageMaker Service API and is attempting to secure traffic from specific sets of instances and IAM users. The VPC is configured with a

single public subnet.

Which combination of steps should the ML specialist take to secure the traffic? (Choose two.)

- A. Add a VPC endpoint policy to allow access to the IAM users.
- B. Modify the users' IAM policy to allow access to Amazon SageMaker Service API calls only.
- C. Modify the security group on the endpoint network interface to restrict access to the instances.
- D. Modify the ACL on the endpoint network interface to restrict access to the instances.
- E. Add a SageMaker Runtime VPC endpoint interface to the VPC.

Answer: AC

NEW QUESTION 38

A Machine Learning Specialist wants to bring a custom algorithm to Amazon SageMaker. The Specialist implements the algorithm in a Docker container supported by Amazon SageMaker.

How should the Specialist package the Docker container so that Amazon SageMaker can launch the training correctly?

- A. Modify the bash_profile file in the container and add a bash command to start the training program
- B. Use CMD config in the Dockerfile to add the training program as a CMD of the image
- C. Configure the training program as an ENTRYPOINT named train
- D. Copy the training program to directory /opt/ml/train

Answer: B

NEW QUESTION 42

An ecommerce company is automating the categorization of its products based on images. A data scientist has trained a computer vision model using the Amazon SageMaker image classification algorithm. The images for each product are classified according to specific product lines. The accuracy of the model is too low when categorizing new products. All of the product images have the same dimensions and are stored within an Amazon S3 bucket. The company wants to improve the model so it can be used for new products as soon as possible.

Which steps would improve the accuracy of the solution? (Choose three.)

- A. Use the SageMaker semantic segmentation algorithm to train a new model to achieve improved accuracy.
- B. Use the Amazon Rekognition DetectLabels API to classify the products in the dataset.
- C. Augment the images in the dataset
- D. Use open source libraries to crop, resize, flip, rotate, and adjust the brightness and contrast of the images.
- E. Use a SageMaker notebook to implement the normalization of pixels and scaling of the image
- F. Store the new dataset in Amazon S3.
- G. Use Amazon Rekognition Custom Labels to train a new model.
- H. Check whether there are class imbalances in the product categories, and apply oversampling or undersampling as require
- I. Store the new dataset in Amazon S3.

Answer: BCE

NEW QUESTION 43

A company is using Amazon Polly to translate plaintext documents to speech for automated company announcements. However, company acronyms are being mispronounced in the current documents. How should a Machine Learning Specialist address this issue for future documents?

- A. Convert current documents to SSML with pronunciation tags
- B. Create an appropriate pronunciation lexicon.
- C. Output speech marks to guide in pronunciation
- D. Use Amazon Lex to preprocess the text files for pronunciation

Answer: A

NEW QUESTION 48

A Machine Learning Specialist is building a convolutional neural network (CNN) that will classify 10 types of animals. The Specialist has built a series of layers in a neural network that will take an input image of an animal, pass it through a series of convolutional and pooling layers, and then finally pass it through a dense and fully connected layer with 10 nodes. The Specialist would like to get an output from the neural network that is a probability distribution of how likely it is that the input image belongs to each of the 10 classes.

Which function will produce the desired output?

- A. Dropout
- B. Smooth L1 loss
- C. Softmax
- D. Rectified linear units (ReLU)

Answer: C

NEW QUESTION 53

A real-estate company is launching a new product that predicts the prices of new houses. The historical data for the properties and prices is stored in .csv format in an Amazon S3 bucket. The data has a header, some categorical fields, and some missing values. The company's data scientists have used Python with a common open-source library to fill the missing values with zeros. The data scientists have dropped all of the categorical fields and have trained a model by using the open-source linear regression algorithm with the default parameters.

The accuracy of the predictions with the current model is below 50%. The company wants to improve the model performance and launch the new product as soon as possible.

Which solution will meet these requirements with the LEAST operational overhead?

- A. Create a service-linked role for Amazon Elastic Container Service (Amazon ECS) with access to the S3 bucket
- B. Create an ECS cluster that is based on an AWS Deep Learning Containers image

- C. Write the code to perform the feature engineerin
- D. Train a logistic regression model for predicting the price, pointing to the bucket with the datase
- E. Wait for the training job to complet
- F. Perform the inferences.
- G. Create an Amazon SageMaker notebook with a new IAM role that is associated with the noteboo
- H. Pull the dataset from the S3 bucke
- I. Explore different combinations of feature engineering transformations, regression algorithms, and hyperparameter
- J. Compare all the results in the notebook, and deploy the most accurate configuration in an endpoint for predictions.
- K. Create an IAM role with access to Amazon S3, Amazon SageMaker, and AWS Lambd
- L. Create a training job with the SageMaker built-in XGBoost model pointing to the bucket with the datase
- M. Specify the price as the target featur
- N. Wait for the job to complet
- O. Load the model artifact to a Lambda function for inference on prices of new houses.
- P. Create an IAM role for Amazon SageMaker with access to the S3 bucke
- Q. Create a SageMaker AutoML job with SageMaker Autopilot pointing to the bucket with the datase
- R. Specify the price as the target attribut
- S. Wait for the job to complet
- T. Deploy the best model for predictions.

Answer: A

NEW QUESTION 55

Example Corp has an annual sale event from October to December. The company has sequential sales data from the past 15 years and wants to use Amazon ML to predict the sales for this year's upcoming event. Which method should Example Corp use to split the data into a training dataset and evaluation dataset?

- A. Pre-split the data before uploading to Amazon S3
- B. Have Amazon ML split the data randomly.
- C. Have Amazon ML split the data sequentially.
- D. Perform custom cross-validation on the data

Answer: C

NEW QUESTION 57

A Machine Learning Specialist is developing a daily ETL workflow containing multiple ETL jobs The workflow consists of the following processes

- * Start the workflow as soon as data is uploaded to Amazon S3
- * When all the datasets are available in Amazon S3, start an ETL job to join the uploaded datasets with multiple terabyte-sized datasets already stored in Amazon S3
- * Store the results of joining datasets in Amazon S3
- * If one of the jobs fails, send a notification to the Administrator Which configuration will meet these requirements?

- A. Use AWS Lambda to trigger an AWS Step Functions workflow to wait for dataset uploads to complete in Amazon S3. Use AWS Glue to join the datasets Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure
- B. Develop the ETL workflow using AWS Lambda to start an Amazon SageMaker notebook instance Use a lifecycle configuration script to join the datasets and persist the results in Amazon S3 Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure
- C. Develop the ETL workflow using AWS Batch to trigger the start of ETL jobs when data is uploaded to Amazon S3 Use AWS Glue to join the datasets in Amazon S3 Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure
- D. Use AWS Lambda to chain other Lambda functions to read and join the datasets in Amazon S3 as soon as the data is uploaded to Amazon S3 Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure

Answer: A

NEW QUESTION 60

A machine learning (ML) specialist needs to extract embedding vectors from a text series. The goal is to provide a ready-to-ingest feature space for a data scientist to develop downstream ML predictive models. The text consists of curated sentences in English. Many sentences use similar words but in different contexts. There are questions and answers among the sentences, and the embedding space must differentiate between them.

Which options can produce the required embedding vectors that capture word context and sequential QA information? (Choose two.)

- A. Amazon SageMaker seq2seq algorithm
- B. Amazon SageMaker BlazingText algorithm in Skip-gram mode
- C. Amazon SageMaker Object2Vec algorithm
- D. Amazon SageMaker BlazingText algorithm in continuous bag-of-words (CBOW) mode
- E. Combination of the Amazon SageMaker BlazingText algorithm in Batch Skip-gram mode with a custom recurrent neural network (RNN)

Answer: AC

NEW QUESTION 63

A Machine Learning Specialist is training a model to identify the make and model of vehicles in images The Specialist wants to use transfer learning and an existing model trained on images of general objects The Specialist collated a large custom dataset of pictures containing different vehicle makes and models

- A. Initialize the model with random weights in all layers including the last fully connected layer
- B. Initialize the model with pre-trained weights in all layers and replace the last fully connected layer.
- C. Initialize the model with random weights in all layers and replace the last fully connected layer
- D. Initialize the model with pre-trained weights in all layers including the last fully connected layer

Answer: D

NEW QUESTION 66

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 68

A media company with a very large archive of unlabeled images, text, audio, and video footage wishes to index its assets to allow rapid identification of relevant content by the Research team. The company wants to use machine learning to accelerate the efforts of its in-house researchers who have limited machine learning expertise.

Which is the FASTEST route to index the assets?

- A. Use Amazon Rekognition, Amazon Comprehend, and Amazon Transcribe to tag data into distinct categories/classes.
- B. Create a set of Amazon Mechanical Turk Human Intelligence Tasks to label all footage.
- C. Use Amazon Transcribe to convert speech to tex
- D. Use the Amazon SageMaker Neural Topic Model (NTM) and Object Detection algorithms to tag data into distinct categories/classes.
- E. Use the AWS Deep Learning AMI and Amazon EC2 GPU instances to create custom models for audio transcription and topic modeling, and use object detection to tag data into distinct categories/classes.

Answer: A

NEW QUESTION 69

An Amazon SageMaker notebook instance is launched into Amazon VPC The SageMaker notebook references data contained in an Amazon S3 bucket in another account The bucket is encrypted using SSE-KMS The instance returns an access denied error when trying to access data in Amazon S3.

Which of the following are required to access the bucket and avoid the access denied error? (Select THREE)

- A. An AWS KMS key policy that allows access to the customer master key (CMK)
- B. A SageMaker notebook security group that allows access to Amazon S3
- C. An IAM role that allows access to the specific S3 bucket
- D. A permissive S3 bucket policy
- E. An S3 bucket owner that matches the notebook owner
- F. A SageMaker notebook subnet ACL that allow traffic to Amazon S3.

Answer: ACF

NEW QUESTION 74

A company uses a long short-term memory (LSTM) model to evaluate the risk factors of a particular energy sector. The model reviews multi-page text documents to analyze each sentence of the text and categorize it as either a potential risk or no risk. The model is not performing well, even though the Data Scientist has experimented with many different network structures and tuned the corresponding hyperparameters.

Which approach will provide the MAXIMUM performance boost?

- A. Initialize the words by term frequency-inverse document frequency (TF-IDF) vectors pretrained on a large collection of news articles related to the energy sector.
- B. Use gated recurrent units (GRUs) instead of LSTM and run the training process until the validation loss stops decreasing.
- C. Reduce the learning rate and run the training process until the training loss stops decreasing.
- D. Initialize the words by word2vec embeddings pretrained on a large collection of news articles related to the energy sector.

Answer: C

NEW QUESTION 77

A Data Scientist wants to gain real-time insights into a data stream of GZIP files. Which solution would allow the use of SQL to query the stream with the LEAST latency?

- A. Amazon Kinesis Data Analytics with an AWS Lambda function to transform the data.
- B. AWS Glue with a custom ETL script to transform the data.
- C. An Amazon Kinesis Client Library to transform the data and save it to an Amazon ES cluster.
- D. Amazon Kinesis Data Firehose to transform the data and put it into an Amazon S3 bucket.

Answer: A

NEW QUESTION 78

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 81

A technology startup is using complex deep neural networks and GPU compute to recommend the company's products to its existing customers based upon each customer's habits and interactions. The solution currently pulls each dataset from an Amazon S3 bucket before loading the data into a TensorFlow model pulled from the company's Git repository that runs locally. This job then runs for several hours while continually outputting its progress to the same S3 bucket. The job can be paused, restarted, and continued at any time in the event of a failure, and is run from a central queue.

Senior managers are concerned about the complexity of the solution's resource management and the costs involved in repeating the process regularly. They ask for the workload to be automated so it runs once a week, starting Monday and completing by the close of business Friday.

Which architecture should be used to scale the solution at the lowest cost?

- A. Implement the solution using AWS Deep Learning Containers and run the container as a job using AWS Batch on a GPU-compatible Spot Instance
- B. Implement the solution using a low-cost GPU-compatible Amazon EC2 instance and use the AWS Instance Scheduler to schedule the task
- C. Implement the solution using AWS Deep Learning Containers, run the workload using AWS Fargate running on Spot Instances, and then schedule the task using the built-in task scheduler
- D. Implement the solution using Amazon ECS running on Spot Instances and schedule the task using the ECS service scheduler

Answer: C

NEW QUESTION 82

A library is developing an automatic book-borrowing system that uses Amazon Rekognition. Images of library members' faces are stored in an Amazon S3 bucket. When members borrow books, the Amazon Rekognition CompareFaces API operation compares real faces against the stored faces in Amazon S3.

The library needs to improve security by making sure that images are encrypted at rest. Also, when the images are used with Amazon Rekognition, they need to be encrypted in transit. The library also must ensure that the images are not used to improve Amazon Rekognition as a service.

How should a machine learning specialist architect the solution to satisfy these requirements?

- A. Enable server-side encryption on the S3 bucket
- B. Submit an AWS Support ticket to opt out of allowing images to be used for improving the service, and follow the process provided by AWS Support.
- C. Switch to using an Amazon Rekognition collection to store the image
- D. Use the IndexFaces and SearchFacesByImage API operations instead of the CompareFaces API operation.
- E. Switch to using the AWS GovCloud (US) Region for Amazon S3 to store images and for Amazon Rekognition to compare face
- F. Set up a VPN connection and only call the Amazon Rekognition API operations through the VPN.
- G. Enable client-side encryption on the S3 bucket
- H. Set up a VPN connection and only call the Amazon Rekognition API operations through the VPN.

Answer: B

NEW QUESTION 85

A Machine Learning Specialist is attempting to build a linear regression model.

Given the displayed residual plot only, what is the MOST likely problem with the model?

- A. Linear regression is inappropriate
- B. The residuals do not have constant variance.
- C. Linear regression is inappropriate
- D. The underlying data has outliers.
- E. Linear regression is appropriate
- F. The residuals have a zero mean.
- G. Linear regression is appropriate
- H. The residuals have constant variance.

Answer: D

NEW QUESTION 87

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 88

A company is building a line-counting application for use in a quick-service restaurant. The company wants to use video cameras pointed at the line of customers at a given register to measure how many people are in line and deliver notifications to managers if the line grows too long. The restaurant locations have limited bandwidth for connections to external services and cannot accommodate multiple video streams without impacting other operations. Which solution should a machine learning specialist implement to meet these requirements?

- A. Install cameras compatible with Amazon Kinesis Video Streams to stream the data to AWS over the restaurant's existing internet connectio
- B. Write an AWS Lambda function to take an image and send it to Amazon Rekognition to count the number of faces in the imag
- C. Send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.
- D. Deploy AWS DeepLens cameras in the restaurant to capture vide
- E. Enable Amazon Rekognition on the AWS DeepLens device, and use it to trigger a local AWS Lambda function when a person is recognize
- F. Use the Lambda function to send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.
- G. Build a custom model in Amazon SageMaker to recognize the number of people in an imag
- H. Install cameras compatible with Amazon Kinesis Video Streams in the restauran
- I. Write an AWS Lambda function to take an imag
- J. Use the SageMaker endpoint to call the model to count peopl
- K. Send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.
- L. Build a custom model in Amazon SageMaker to recognize the number of people in an imag
- M. Deploy AWS DeepLens cameras in the restaurant
- N. Deploy the model to the camera
- O. Deploy an AWS Lambda function to the cameras to use the model to count people and send an Amazon Simple Notification Service (Amazon SNS) notification if the line is too long.

Answer: A

NEW QUESTION 93

A retail chain has been ingesting purchasing records from its network of 20,000 stores to Amazon S3 using Amazon Kinesis Data Firehose To support training an improved machine learning model, training records will require new but simple transformations, and some attributes will be combined The model needs lo be retrained daily
Given the large number of stores and the legacy data ingestion, which change will require the LEAST amount of development effort?

- A. Require that the stores to switch to capturing their data locally on AWS Storage Gateway for loading into Amazon S3 then use AWS Glue to do the transformation
- B. Deploy an Amazon EMR cluster running Apache Spark with the transformation logic, and have the cluster run each day on the accumulating records in Amazon S3, outputting new/transformed records to Amazon S3
- C. Spin up a fleet of Amazon EC2 instances with the transformation logic, have them transform the data records accumulating on Amazon S3, and output the transformed records to Amazon S3.
- D. Insert an Amazon Kinesis Data Analytics stream downstream of the Kinesis Data Firehose stream that transforms raw record attributes into simple transformed values using SQL.

Answer: D

NEW QUESTION 96

An aircraft engine manufacturing company is measuring 200 performance metrics in a time-series. Engineers want to detect critical manufacturing defects in near-real time during testing. All of the data needs to be stored for offline analysis.
What approach would be the MOST effective to perform near-real time defect detection?

- A. Use AWS IoT Analytics for ingestion, storage, and further analysi
- B. Use Jupyter notebooks from within AWS IoT Analytics to carry out analysis for anomalies.
- C. Use Amazon S3 for ingestion, storage, and further analysi
- D. Use an Amazon EMR cluster to carry out Apache Spark ML k-means clustering to determine anomalies.
- E. Use Amazon S3 for ingestion, storage, and further analysi
- F. Use the Amazon SageMaker Random Cut Forest (RCF) algorithm to determine anomalies.
- G. Use Amazon Kinesis Data Firehose for ingestion and Amazon Kinesis Data Analytics Random Cut Forest (RCF) to perform anomaly detectio
- H. Use Kinesis Data Firehose to store data in Amazon S3 for further analysis.

Answer: B

NEW QUESTION 101

An ecommerce company sends a weekly email newsletter to all of its customers. Management has hired a team of writers to create additional targeted content. A data scientist needs to identify five customer segments based on age, income, and location. The customers' current segmentation is unknown. The data scientist previously built an XGBoost model to predict the likelihood of a customer responding to an email based on age, income, and location.
Why does the XGBoost model NOT meet the current requirements, and how can this be fixed?

- A. The XGBoost model provides a true/false binary output
- B. Apply principal component analysis (PCA) with five feature dimensions to predict a segment.
- C. The XGBoost model provides a true/false binary output
- D. Increase the number of classes the XGBoost model predicts to five classes to predict a segment.
- E. The XGBoost model is a supervised machine learning algorithm
- F. Train a k-Nearest-Neighbors (kNN) model with K = 5 on the same dataset to predict a segment.
- G. The XGBoost model is a supervised machine learning algorithm
- H. Train a k-means model with K = 5 on the same dataset to predict a segment.

Answer: C

NEW QUESTION 102

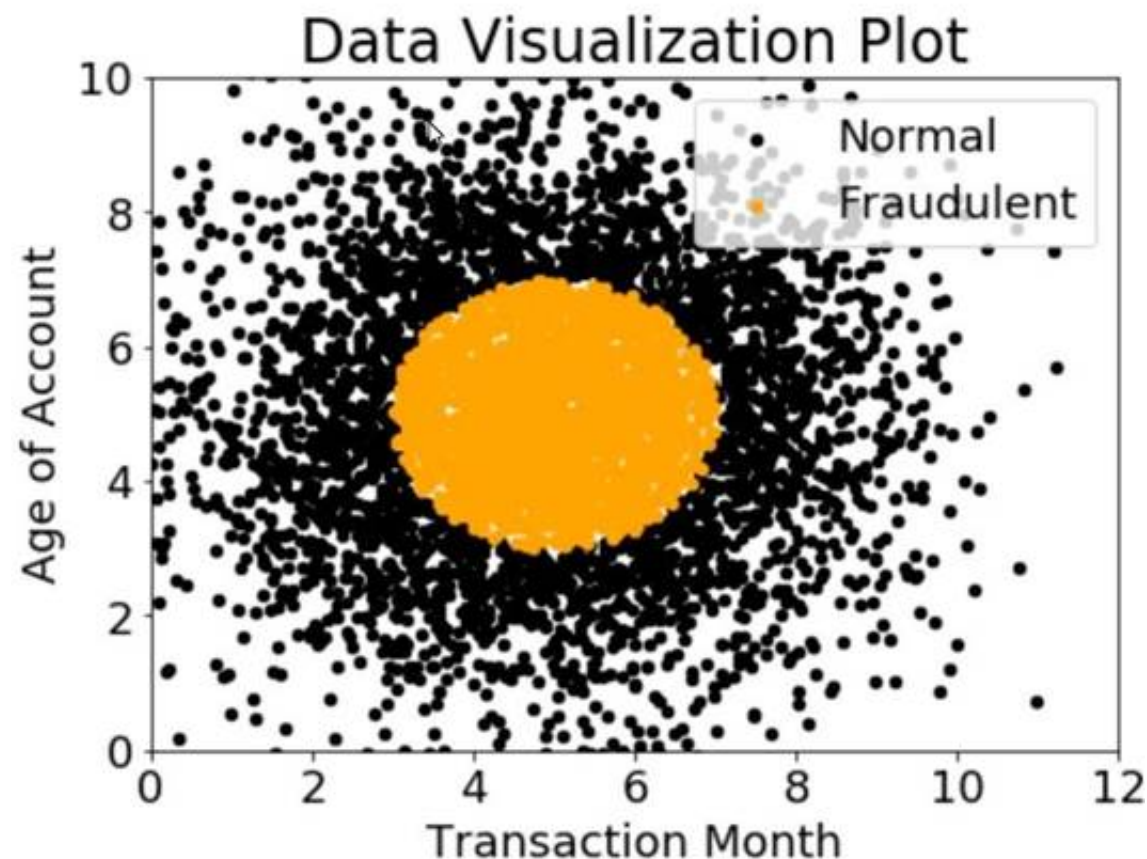
A credit card company wants to build a credit scoring model to help predict whether a new credit card applicant will default on a credit card payment. The company has collected data from a large number of sources with thousands of raw attributes. Early experiments to train a classification model revealed 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 107

A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a Machine Learning Specialist would like to build a binary classifier based on two features: age of account and transaction month. The class distribution for these features is illustrated in the figure provided.



Based on this information, which model would have the HIGHEST recall with respect to the fraudulent class?

- A. Decision tree
- B. Linear support vector machine (SVM)
- C. Naive Bayesian classifier
- D. Single Perceptron with sigmoidal activation function

Answer: C

NEW QUESTION 109

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 113

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 117

A Machine Learning Specialist is working with a large company to leverage machine learning within its products. The company wants to group its customers into categories based on which customers will and will not churn within the next 6 months. The company has labeled the data available to the Specialist. Which machine learning model type should the Specialist use to accomplish this task?

- A. Linear regression
- B. Classification
- C. Clustering
- D. Reinforcement learning

Answer: B

Explanation:

The goal of classification is to determine to which class or category a data point (customer in our case) belongs to. For classification problems, data scientists would use historical data with predefined target variables AKA labels (churner/non-churner) – answers that need to be predicted – to train an algorithm. With classification, businesses can answer the following questions:

- Will this customer churn or not?
- Will a customer renew their subscription?
- Will a user downgrade a pricing plan?
- Are there any signs of unusual customer behavior?

NEW QUESTION 122

A Machine Learning Specialist is building a prediction model for a large number of features using linear models, such as linear regression and logistic regression. During exploratory data analysis the Specialist observes that many features are highly correlated with each other. This may make the model unstable. What should be done to reduce the impact of having such a large number of features?

- A. Perform one-hot encoding on highly correlated features
- B. Use matrix multiplication on highly correlated features.
- C. Create a new feature space using principal component analysis (PCA)
- D. Apply the Pearson correlation coefficient

Answer: B

NEW QUESTION 124

A company is observing low accuracy while training on the default built-in image classification algorithm in Amazon SageMaker. The Data Science team wants to use an Inception neural network architecture instead of a ResNet architecture. Which of the following will accomplish this? (Select TWO.)

- A. Customize the built-in image classification algorithm to use Inception and use this for model training.
- B. Create a support case with the SageMaker team to change the default image classification algorithm to Inception.
- C. Bundle a Docker container with TensorFlow Estimator loaded with an Inception network and use this for model training.
- D. Use custom code in Amazon SageMaker with TensorFlow Estimator to load the model with an Inception network and use this for model training.
- E. Download and apt-get install the inception network code into an Amazon EC2 instance and use this instance as a Jupyter notebook in Amazon SageMaker.

Answer: AD

NEW QUESTION 128

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 132

A Data Scientist is building a model to predict customer churn using a dataset of 100 continuous numerical features. The Marketing team has not provided any insight about which features are relevant for churn prediction. The Marketing team wants to interpret the model and see the direct impact of relevant features on the model outcome. While training a logistic regression model, the Data Scientist observes that there is a wide gap between the training and validation set accuracy.

Which methods can the Data Scientist use to improve the model performance and satisfy the Marketing team's needs? (Choose two.)

- A. Add L1 regularization to the classifier
- B. Add features to the dataset
- C. Perform recursive feature elimination
- D. Perform t-distributed stochastic neighbor embedding (t-SNE)
- E. Perform linear discriminant analysis

Answer: BE

NEW QUESTION 135

When submitting Amazon SageMaker training jobs using one of the built-in algorithms, which common parameters MUST be specified? (Select THREE.)

- A. The training channel identifying the location of training data on an Amazon S3 bucket.
- B. The validation channel identifying the location of validation data on an Amazon S3 bucket.
- C. The IAM role that Amazon SageMaker can assume to perform tasks on behalf of the users.
- D. Hyperparameters in a JSON array as documented for the algorithm used.
- E. The Amazon EC2 instance class specifying whether training will be run using CPU or GPU.
- F. The output path specifying where on an Amazon S3 bucket the trained model will persist.

Answer: CEF

NEW QUESTION 140

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 143

A telecommunications company is developing a mobile app for its customers. The company is using an Amazon SageMaker hosted endpoint for machine learning model inferences.

Developers want to introduce a new version of the model for a limited number of users who subscribed to a preview feature of the app. After the new version of the model is tested as a preview, developers will evaluate its accuracy. If a new version of the model has better accuracy, developers need to be able to gradually release the new version for all users over a fixed period of time.

How can the company implement the testing model with the LEAST amount of operational overhead?

- A. Update the ProductionVariant data type with the new version of the model by using the CreateEndpointConfig operation with the InitialVariantWeight parameter set to 0. Specify the TargetVariant parameter for InvokeEndpoint calls for users who subscribed to the preview feature.
- B. When the new version of the model is ready for release, gradually increase InitialVariantWeight until all users have the updated version.
- C. Configure two SageMaker hosted endpoints that serve the different versions of the model.
- D. Create an Application Load Balancer (ALB) to route traffic to both endpoints based on the TargetVariant query string parameter.
- E. Reconfigure the app to send the TargetVariant query string parameter for users who subscribed to the preview feature.
- F. When the new version of the model is ready for release, change the ALB's routing algorithm to weighted until all users have the updated version.
- G. Update the DesiredWeightsAndCapacity data type with the new version of the model by using the UpdateEndpointWeightsAndCapacities operation with the DesiredWeight parameter set to 0. Specify the TargetVariant parameter for InvokeEndpoint calls for users who subscribed to the preview feature.
- H. When the new version of the model is ready for release, gradually increase DesiredWeight until all users have the updated version.
- I. Configure two SageMaker hosted endpoints that serve the different versions of the model.
- J. Create an Amazon Route 53 record that is configured with a simple routing policy and that points to the current version of the model.
- K. Configure the mobile app to use the endpoint URL for users who subscribed to the preview feature and to use the Route 53 record for other users.
- L. When the new version of the model is ready for release, add a new model version endpoint to Route 53, and switch the policy to weighted until all users have the updated version.

Answer: D

NEW QUESTION 148

A company has raw user and transaction data stored in Amazon S3, a MySQL database, and Amazon Redshift. A Data Scientist needs to perform an analysis by joining the three datasets from Amazon S3, MySQL, and Amazon Redshift, and then calculating the average of a few selected columns from the joined data. Which AWS service should the Data Scientist use?

- A. Amazon Athena
- B. Amazon Redshift Spectrum
- C. AWS Glue
- D. Amazon QuickSight

Answer: A

NEW QUESTION 149

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 152

A company has video feeds and images of a subway train station. The company wants to create a deep learning model that will alert the station manager if any passenger crosses the yellow safety line when there is no train in the station. The alert will be based on the video feeds. The company wants the model to detect the yellow line, the passengers who cross the yellow line, and the trains in the video feeds. This task requires labeling. The video data must remain confidential. A data scientist creates a bounding box to label the sample data and uses an object detection model. However, the object detection model cannot clearly demarcate the yellow line, the passengers who cross the yellow line, and the trains. Which labeling approach will help the company improve this model?

- A. Use Amazon Rekognition Custom Labels to label the dataset and create a custom Amazon Rekognition object detection mode
- B. Create a private workforce
- C. Use Amazon Augmented AI (Amazon A2I) to review the low-confidence predictions and retrain the custom Amazon Rekognition model.
- D. Use an Amazon SageMaker Ground Truth object detection labeling tas
- E. Use Amazon Mechanical Turk as the labeling workforce.
- F. Use Amazon Rekognition Custom Labels to label the dataset and create a custom Amazon Rekognition object detection mode
- G. Create a workforce with a third-party AWS Marketplace vendo
- H. Use Amazon Augmented AI (Amazon A2I) to review the low-confidence predictions and retrain the custom Amazon Rekognition model.
- I. Use an Amazon SageMaker Ground Truth semantic segmentation labeling tas
- J. Use a private workforce as the labeling workforce.

Answer: B

NEW QUESTION 155

A Machine Learning Specialist is designing a system for improving sales for a company. The objective is to use the large amount of information the company has on users' behavior and product preferences to predict which products users would like based on the users' similarity to other users. What should the Specialist do to meet this objective?

- A. Build a content-based filtering recommendation engine with Apache Spark ML on Amazon EMR.
- B. Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR.
- C. Build a model-based filtering recommendation engine with Apache Spark ML on Amazon EMR.
- D. Build a combinative filtering recommendation engine with Apache Spark ML on Amazon EMR.

Answer: B

Explanation:

Many developers want to implement the famous Amazon model that was used to power the "People who bought this also bought these items" feature on Amazon.com. This model is based on a method called Collaborative Filtering. It takes items such as movies, books, and products that were rated highly by a set of users and recommending them to other users who also gave them high ratings. This method works well in domains where explicit ratings or implicit user actions can be gathered and analyzed.

NEW QUESTION 160

A machine learning specialist is developing a proof of concept for government users whose primary concern is security. The specialist is using Amazon SageMaker to train a convolutional neural network (CNN) model for a photo classifier application. The specialist wants to protect the data so that it cannot be accessed and transferred to a remote host by malicious code accidentally installed on the training container. Which action will provide the MOST secure protection?

- A. Remove Amazon S3 access permissions from the SageMaker execution role.
- B. Encrypt the weights of the CNN model.
- C. Encrypt the training and validation dataset.
- D. Enable network isolation for training jobs.

Answer: D

NEW QUESTION 162

A Machine Learning Specialist is building a supervised model that will evaluate customers' satisfaction with their mobile phone service based on recent usage The model's output should infer whether or not a customer is likely to switch to a competitor in the next 30 days Which of the following modeling techniques should the Specialist use1?

- A. Time-series prediction
- B. Anomaly detection
- C. Binary classification
- D. Regression

Answer: D

NEW QUESTION 163

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 has been asked to reduce the number of false negatives.

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

Which combination of steps should the Data Scientist take to reduce the number of false positive predictions by the model? (Select TWO.)

- A. Change the XGBoost eval_metric parameter to optimize based on rmse instead of error.

- 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 evaljnetric parameter to optimize based on AUC instead of error.
- E. Decrease the XGBoost max_depth parameter because the model is currently overfitting the data.

Answer: DE

NEW QUESTION 168

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 (-Distributed Stochastic Neighbor Embedding (I-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 169

An interactive online dictionary wants to add a widget that displays words used in similar contexts. A Machine Learning Specialist is asked to provide word features for the downstream nearest neighbor model powering the widget.

What should the Specialist do to meet these requirements?

- A. Create one-hot word encoding vectors.
- B. Produce a set of synonyms for every word using Amazon Mechanical Turk.
- C. Create word embedding factors that store edit distance with every other word.
- D. Download word embedding's pre-trained on a large corpus.

Answer: D

NEW QUESTION 170

A Data Scientist is training a multilayer perception (MLP) on a dataset with multiple classes. The target class of interest is unique compared to the other classes within the dataset, but it does not achieve an acceptable recall metric. The Data Scientist has already tried varying the number and size of the MLP's hidden layers, which has not significantly improved the results. A solution to improve recall must be implemented as quickly as possible.

Which techniques should be used to meet these requirements?

- A. Gather more data using Amazon Mechanical Turk and then retrain.
- B. Train an anomaly detection model instead of an MLP.
- C. Train an XGBoost model instead of an MLP.
- D. Add class weights to the MLP's loss function and then retrain.

Answer: C

NEW QUESTION 174

A Machine Learning Specialist wants to determine the appropriate SageMakerVariant Invocations Per Instance setting for an endpoint automatic scaling configuration. The Specialist has performed a load test on a single instance and determined that peak requests per second (RPS) without service degradation is about 20 RPS. As this is the first deployment, the Specialist intends to set the invocation safety factor to 0.5.

Based on the stated parameters and given that the invocations per instance setting is measured on a per-minute basis, what should the Specialist set as the sageMakervariantinvocationsPerinstance setting?

- A. 10
- B. 30
- C. 600
- D. 2,400

Answer: C

NEW QUESTION 179

A Machine Learning Specialist is building a model to predict future employment rates based on a wide range of economic factors. While exploring the data, the Specialist notices that the magnitude of the input features vary greatly. The Specialist does not want variables with a larger magnitude to dominate the model. What should the Specialist do to prepare the data for model training?

- A. Apply quantile binning to group the data into categorical bins to keep any relationships in the data by replacing the magnitude with distribution.
- B. Apply the Cartesian product transformation to create new combinations of fields that are independent of the magnitude.
- C. Apply normalization to ensure each field will have a mean of 0 and a variance of 1 to remove any significant magnitude.
- D. Apply the orthogonal sparse Diagram (OSB) transformation to apply a fixed-size sliding window to generate new features of a similar magnitude.

Answer: C

NEW QUESTION 184

A retail company is selling products through a global online marketplace. The company wants to use machine learning (ML) to analyze customer feedback and identify specific areas for improvement. A developer has built a tool that collects customer reviews from the online marketplace and stores them in an Amazon S3 bucket.

bucket. This process yields a dataset of 40 reviews. A data scientist building the ML models must identify additional sources of data to increase the size of the dataset.

Which data sources should the data scientist use to augment the dataset of reviews? (Choose three.)

- A. Emails exchanged by customers and the company's customer service agents
- B. Social media posts containing the name of the company or its products
- C. A publicly available collection of news articles
- D. A publicly available collection of customer reviews
- E. Product sales revenue figures for the company
- F. Instruction manuals for the company's products

Answer: BDF

NEW QUESTION 185

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 187

A Machine Learning Specialist must build out a process to query a dataset on Amazon S3 using Amazon Athena. The dataset contains more than 800,000 records stored as plaintext CSV files. Each record contains 200 columns and is approximately 1.5 MB in size. Most queries will span 5 to 10 columns only. How should the Machine Learning Specialist transform the dataset to minimize query runtime?

- A. Convert the records to Apache Parquet format
- B. Convert the records to JSON format
- C. Convert the records to GZIP CSV format
- D. Convert the records to XML format

Answer: A

Explanation:

Using compressions will reduce the amount of data scanned by Amazon Athena, and also reduce your S3 bucket storage. It's a Win-Win for your AWS bill. Supported formats: GZIP, LZO, SNAPPY (Parquet) and ZLIB.

NEW QUESTION 190

A Machine Learning Specialist is designing a scalable data storage solution for Amazon SageMaker. There is an existing TensorFlow-based model implemented as a train.py script that relies on static training data that is currently stored as TFRecords.

Which method of providing training data to Amazon SageMaker would meet the business requirements with the LEAST development overhead?

- A. Use Amazon SageMaker script mode and use train.py unchanged
- B. Point the Amazon SageMaker training invocation to the local path of the data without reformatting the training data.
- C. Use Amazon SageMaker script mode and use train.py unchanged
- D. Put the TFRecord data into an Amazon S3 bucket
- E. Point the Amazon SageMaker training invocation to the S3 bucket without reformatting the training data.
- F. Rewrite the train.py script to add a section that converts TFRecords to protobuf and ingests the protobuf data instead of TFRecords.
- G. Prepare the data in the format accepted by Amazon SageMaker
- H. Use AWS Glue or AWS Lambda to reformat and store the data in an Amazon S3 bucket.

Answer: B

Explanation:

<https://github.com/aws-samples/amazon-sagemaker-script-mode/blob/master/tf-horovod-inference-pipeline/train>

NEW QUESTION 193

A Data Science team is designing a dataset repository where it will store a large amount of training data commonly used in its machine learning models. As Data Scientists may create an arbitrary number of new datasets every day, the solution has to scale automatically and be cost-effective. Also, it must be possible to explore the data using SQL.

Which storage scheme is MOST adapted to this scenario?

- A. Store datasets as files in Amazon S3.
- B. Store datasets as files in an Amazon EBS volume attached to an Amazon EC2 instance.
- C. Store datasets as tables in a multi-node Amazon Redshift cluster.
- D. Store datasets as global tables in Amazon DynamoDB.

Answer: A

NEW QUESTION 198

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 dat
- 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 NMF 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 201

A Machine Learning Specialist is packaging a custom ResNet model into a Docker container so the company can leverage Amazon SageMaker for training. The Specialist is using Amazon EC2 P3 instances to train the model and needs to properly configure the Docker container to leverage the NVIDIA GPUs. What does the Specialist need to do?

- A. Bundle the NVIDIA drivers with the Docker image.
- B. Build the Docker container to be NVIDIA-Docker compatible.
- C. Organize the Docker container's file structure to execute on GPU instances.
- D. Set the GPU flag in the Amazon SageMaker CreateTrainingJob request body

Answer: B

NEW QUESTION 202

A Machine Learning Specialist is developing recommendation engine for a photography blog. Given a picture, the recommendation engine should show a picture that captures similar objects. The Specialist would like to create a numerical representation feature to perform nearest-neighbor searches. What actions would allow the Specialist to get relevant numerical representations?

- A. Reduce image resolution and use reduced resolution pixel values as features
- B. Use Amazon Mechanical Turk to label image content and create a one-hot representation indicating the presence of specific labels
- C. Run images through a neural network pre-trained on ImageNet, and collect the feature vectors from the penultimate layer
- D. Average colors by channel to obtain three-dimensional representations of images.

Answer: A

NEW QUESTION 205

A gaming company has launched an online game where people can start playing for free but they need to pay if they choose to use certain features. The company needs to build an automated system to predict whether or not a new user will become a paid user within 1 year. The company has gathered a labeled dataset from 1 million users.

The training dataset consists of 1,000 positive samples (from users who ended up paying within 1 year) and 999.1 negative samples (from users who did not use any paid features). Each data sample consists of 200 features including user age, device, location, and play patterns.

Using this dataset for training, the Data Science team trained a random forest model that converged with over 99% accuracy on the training set. However, the prediction results on a test dataset were not satisfactory.

Which of the following approaches should the Data Science team take to mitigate this issue? (Select TWO.)

- A. Add more deep trees to the random forest to enable the model to learn more features.
- B. Indicate a copy of the samples in the test database in the training dataset.
- C. Generate more positive samples by duplicating the positive samples and adding a small amount of noise to the duplicated data.
- D. Change the cost function so that false negatives have a higher impact on the cost value than false positives.
- E. Change the cost function so that false positives have a higher impact on the cost value than false negatives.

Answer: CD

NEW QUESTION 209

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 211

.....

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](#)