

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

AWS Certified Machine Learning - Specialty

<https://www.2passeasy.com/dumps/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

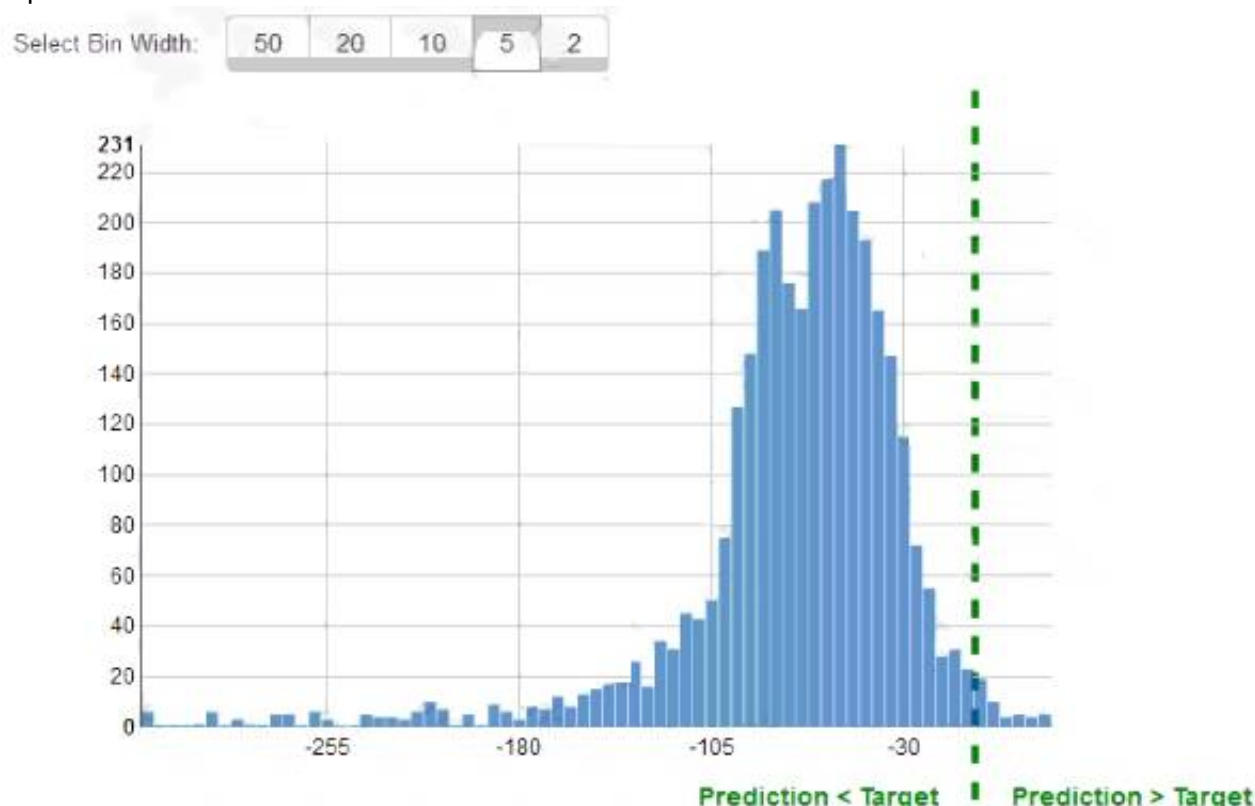
A Machine Learning Specialist works for a credit card processing company and needs to predict which transactions may be fraudulent in near-real time. Specifically, the Specialist must train a model that returns the probability that a given transaction may be fraudulent. How should the Specialist frame this business problem?

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

Answer: A

### NEW QUESTION 4

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 5

A Machine Learning Specialist is creating a new natural language processing application that processes a dataset comprised of 1 million sentences. The aim is to then run Word2Vec to generate embeddings of the sentences and enable different types of predictions.

Here is an example from the dataset:

"The quck BROWN FOX jumps over the lazy dog."

Which of the following are the operations the Specialist needs to perform to correctly sanitize and prepare the data in a repeatable manner? (Select THREE)

- A. Perform part-of-speech tagging and keep the action verb and the nouns only
- B. Normalize all words by making the sentence lowercase
- C. Remove stop words using an English stopwords dictionary.
- D. Correct the typography on "quck" to "quick."
- E. One-hot encode all words in the sentence
- F. Tokenize the sentence into words.

**Answer:** BCF

#### NEW QUESTION 6

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 7

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 8

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 model
- B. Create a notebook that uses the XGBoost training container to perform model training
- 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 behavior
- 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 model
- H. Create a notebook that uses the XGBoost training container to perform model training
- I. Activate Amazon SageMaker Debugger, and configure it to calculate and collect Shapley values
- 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 instance
- L. Use the notebook instance and the XGBoost library to locally retrain the model
- M. Use the `plot_importance()` method in the Python XGBoost interface to create a feature importance chart
- 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 model
- P. Create a notebook that uses the XGBoost training container to perform model training
- Q. Deploy the model at an endpoint
- R. Use Amazon SageMaker Processing to post-analyze the model and create a feature importance explainability chart automatically for the credit team.

**Answer:** C

#### NEW QUESTION 9

A company is setting up an Amazon SageMaker environment. The corporate data security policy does not allow communication over the internet. How can the company enable the Amazon SageMaker service without enabling direct internet access to Amazon SageMaker notebook instances?

- A. Create a NAT gateway within the corporate VPC.
- B. Route Amazon SageMaker traffic through an on-premises network.
- C. Create Amazon SageMaker VPC interface endpoints within the corporate VPC.
- D. Create VPC peering with Amazon VPC hosting Amazon SageMaker.

**Answer:** A

#### NEW QUESTION 10

An Machine Learning Specialist discover the following statistics while experimenting on a model.

Experiment 1  
Baseline model  
Train error = 5%  
Test error = 16%

Experiment 2  
The Specialist added more layers and neurons to the model and received the following results:  
Train error = 5.2%  
Test error = 15.7%

Experiment 3  
The Specialist reverted back to the original number of neurons from Experiment 1 and implemented regularization in the neural network, which yielded the following results:  
Train error = 4.7%  
Test error = 9.5%

What can the Specialist from the experiments?

- A. The model In Experiment 1 had a high variance error that was reduced in Experiment 3 by regularization Experiment 2 shows that there is minimal bias error in Experiment 1
- B. The model in Experiment 1 had a high bias error that was reduced in Experiment 3 by regularization Experiment 2 shows that there is minimal variance error in Experiment 1
- C. The model in Experiment 1 had a high bias error and a high variance error that were reduced in Experiment 3 by regularization Experiment 2 shows that high bias cannot be reduced by increasing layers and neurons in the model
- D. The model in Experiment 1 had a high random noise error that was reduced in Experiment 3 by regularization Experiment 2 shows that random noise cannot be reduced by increasing layers and neurons in the model

**Answer:** C

#### NEW QUESTION 10

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 11

A data science team is planning to build a natural language processing (NLP) application. The application's text preprocessing stage will include part-of-speech tagging and key phrase extraction. The preprocessed text will be input to a custom classification algorithm that the data science team has already written and trained using Apache MXNet.

Which solution can the team build MOST quickly to meet these requirements?

- A. Use Amazon Comprehend for the part-of-speech tagging, key phrase extraction, and classification tasks.
- B. Use an NLP library in Amazon SageMaker for the part-of-speech tagging
- C. Use Amazon Comprehend for the key phrase extraction
- D. Use AWS Deep Learning Containers with Amazon SageMaker to build the custom classifier.
- E. Use Amazon Comprehend for the part-of-speech tagging and key phrase extraction task
- F. Use Amazon SageMaker built-in Latent Dirichlet Allocation (LDA) algorithm to build the custom classifier.
- G. Use Amazon Comprehend for the part-of-speech tagging and key phrase extraction task
- H. Use AWS Deep Learning Containers with Amazon SageMaker to build the custom classifier.

**Answer:** B

#### NEW QUESTION 12

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 Create TrainingJob request body

**Answer:** A

#### NEW QUESTION 16

A global financial company is using machine learning to automate its loan approval process. The company has a dataset of customer information. The dataset contains some categorical fields, such as customer location by city and housing status. The dataset also includes financial fields in different units, such as account balances in US dollars and monthly interest in US cents.

The company's data scientists are using a gradient boosting regression model to infer the credit score for each customer. The model has a training accuracy of 99% and a testing accuracy of 75%. The data scientists want to improve the model's testing accuracy.

Which process will improve the testing accuracy the MOST?

- A. Use a one-hot encoder for the categorical fields in the dataset
- B. Perform standardization on the financial fields in the dataset
- C. Apply L1 regularization to the data.
- D. Use tokenization of the categorical fields in the dataset
- E. Perform binning on the financial fields in the dataset
- F. Remove the outliers in the data by using the z-score.
- G. Use a label encoder for the categorical fields in the dataset
- H. Perform L1 regularization on the financial fields in the dataset
- I. Apply L2 regularization to the data.
- J. Use a logarithm transformation on the categorical fields in the dataset
- K. Perform binning on the financial fields in the dataset
- L. Use imputation to populate missing values in the dataset.

**Answer:** B

#### NEW QUESTION 20

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 22

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 26

A large JSON dataset for a project has been uploaded to a private Amazon S3 bucket. The Machine Learning Specialist wants to securely access and explore the data from an Amazon SageMaker notebook instance. A new VPC was created and assigned to the Specialist.

How can the privacy and integrity of the data stored in Amazon S3 be maintained while granting access to the Specialist for analysis?

- A. Launch the SageMaker notebook instance within the VPC with SageMaker-provided internet access enabled. Use an S3 ACL to open read privileges to the everyone group.
- B. Launch the SageMaker notebook instance within the VPC and create an S3 VPC endpoint for the notebook to access the data. Copy the JSON dataset from Amazon S3 into the ML storage volume on the SageMaker notebook instance and work against the local dataset.
- C. Launch the SageMaker notebook instance within the VPC and create an S3 VPC endpoint for the notebook to access the data. Define a custom S3 bucket policy to only allow requests from your VPC to access the S3 bucket.
- D. Launch the SageMaker notebook instance within the VPC with SageMaker-provided internet access enabled.
- E. Generate an S3 pre-signed URL for access to data in the bucket.

**Answer:** B

#### NEW QUESTION 30

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 33

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 37

A machine learning (ML) specialist wants to create a data preparation job that uses a PySpark script with complex window aggregation operations to create data for training and testing. The ML specialist needs to evaluate the impact of the number of features and the sample count on model performance.

Which approach should the ML specialist use to determine the ideal data transformations for the model?

A. Add an Amazon SageMaker Debugger hook to the script to capture key metric

B. Run the script as an AWS Glue job.

C. Add an Amazon SageMaker Experiments tracker to the script to capture key metric

D. Run the script as an AWS Glue job.

E. Add an Amazon SageMaker Debugger hook to the script to capture key parameter

F. Run the script as a SageMaker processing job.

G. Add an Amazon SageMaker Experiments tracker to the script to capture key parameter

H. Run the script as a SageMaker processing job.

**Answer:** B

### NEW QUESTION 40

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 41

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 44

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 47

A Data Scientist is developing a binary classifier to predict whether a patient has a particular disease on a series of test results. The Data Scientist has data on 400 patients randomly selected from the population. The disease is seen in 3% of the population. Which cross-validation strategy should the Data Scientist adopt?

- A. A k-fold cross-validation strategy with k=5
- B. A stratified k-fold cross-validation strategy with k=5
- C. A k-fold cross-validation strategy with k=5 and 3 repeats
- D. An 80/20 stratified split between training and validation

**Answer: B**

#### NEW QUESTION 50

A data scientist is training a text classification model by using the Amazon SageMaker built-in BlazingText algorithm. There are 5 classes in the dataset, with 300 samples for category A, 292 samples for category B, 240 samples for category C, 258 samples for category D, and 310 samples for category E. The data scientist shuffles the data and splits off 10% for testing. After training the model, the data scientist generates confusion matrices for the training and test sets.

Training data confusion matrix

		Predicted class					
		A	B	C	D	E	Total
True class	A	270	0	0	0	0	270
	B	1	260	0	0	2	263
	C	0	0	111	100	5	216
	D	4	3	132	92	1	232
	E	0	0	2	3	274	279
	Total	275	263	245	195	282	1260

Test data confusion matrix

		Predicted class					
		A	B	C	D	E	Total
True class	A	9	1	0	0	0	10
	B	2	25	0	2	0	29
	C	10	2	11	10	1	34
	D	1	0	12	14	0	27
	E	9	1	4	1	25	40
	Total	31	29	27	27	26	140

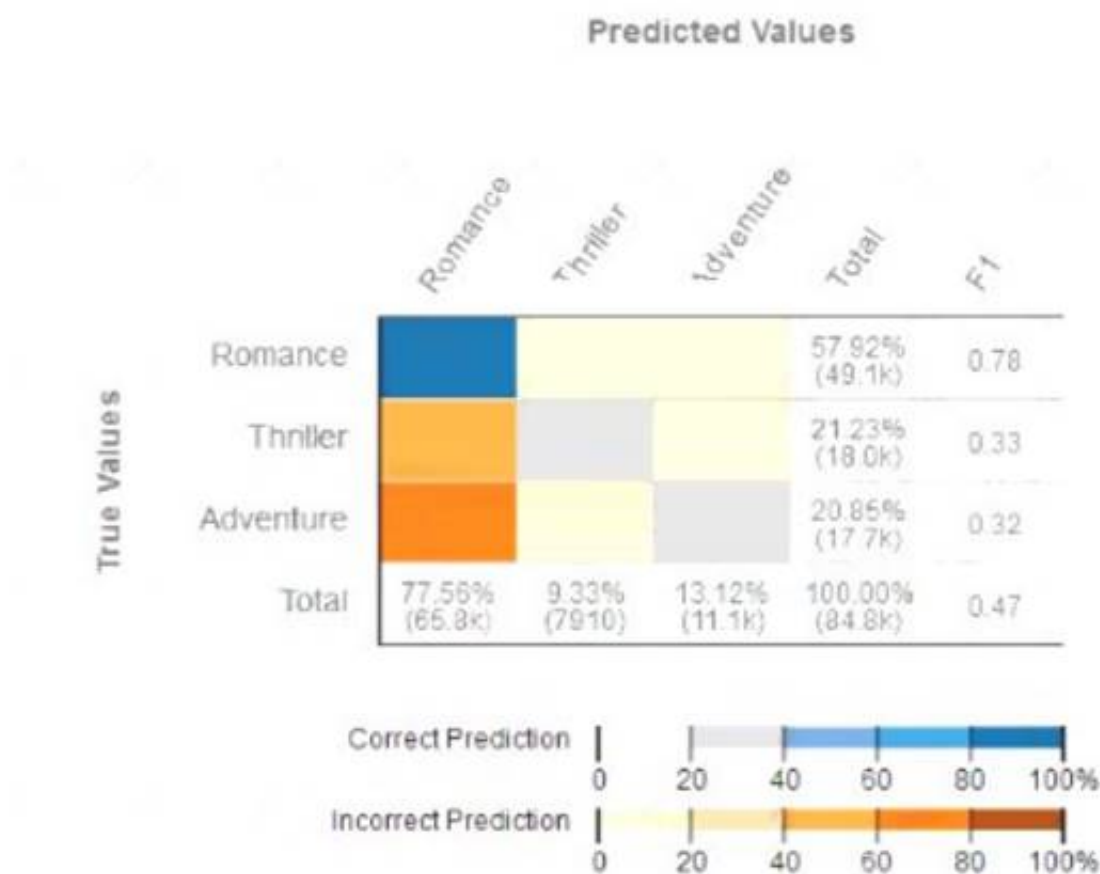
What could the data scientist conclude from these results?

- A. Classes C and D are too similar.
- B. The dataset is too small for holdout cross-validation.
- C. The data distribution is skewed.
- D. The model is overfitting for classes B and E.

**Answer: B**

#### NEW QUESTION 51

Given the following confusion matrix for a movie classification model, what is the true class frequency for Romance and the predicted class frequency for Adventure?



- A. The true class frequency for Romance is 77.56% and the predicted class frequency for Adventure is 20.85%  
 B. The true class frequency for Romance is 57.92% and the predicted class frequency for Adventure is 13.12%  
 C. The true class frequency for Romance is 0.78 and the predicted class frequency for Adventure is (0.47 - 0.32).  
 D. The true class frequency for Romance is 77.56% \* 0.78 and the predicted class frequency for Adventure is 20.85% \* 0.32

**Answer: B**

**Explanation:**

<https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-model-insights.html>

**NEW QUESTION 53**

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 57**

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 62

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 restauran
- 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 67

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 72

A Data Scientist needs to analyze employment data. The dataset contains approximately 10 million observations on people across 10 different features. During the preliminary analysis, the Data Scientist notices that income and age distributions are not normal. While income levels shows a right skew as expected, with fewer individuals having a higher income, the age distribution also show a right skew, with fewer older individuals participating in the workforce.

Which feature transformations can the Data Scientist apply to fix the incorrectly skewed data? (Choose two.)

- A. Cross-validation
- B. Numerical value binning
- C. High-degree polynomial transformation
- D. Logarithmic transformation
- E. One hot encoding

**Answer:** AB

#### NEW QUESTION 76

A Machine Learning Specialist previously trained a logistic regression model using scikit-learn on a local machine, and the Specialist now wants to deploy it to production for inference only.

What steps should be taken to ensure Amazon SageMaker can host a model that was trained locally?

- A. Build the Docker image with the inference cod
- B. Tag the Docker image with the registry hostname and upload it to Amazon ECR.
- C. Serialize the trained model so the format is compressed for deploymen
- D. Tag the Docker image with the registry hostname and upload it to Amazon S3.
- E. Serialize the trained model so the format is compressed for deploymen
- F. Build the image and upload it to Docker Hub.
- G. Build the Docker image with the inference cod
- H. Configure Docker Hub and upload the image to AmazonECR.

**Answer:** D

#### NEW QUESTION 79

A manufacturing company uses machine learning (ML) models to detect quality issues. The models use images that are taken of the company's product at the end of each production step. The company has thousands of machines at the production site that generate one image per second on average.

The company ran a successful pilot with a single manufacturing machine. For the pilot, ML specialists used an industrial PC that ran AWS IoT Greengrass with a long-running AWS Lambda function that uploaded the images to Amazon S3. The uploaded images invoked a Lambda function that was written in Python to perform inference by using an Amazon SageMaker endpoint that ran a custom model. The inference results were forwarded back to a web service that was hosted at the production site to prevent faulty products from being shipped.

The company scaled the solution out to all manufacturing machines by installing similarly configured industrial PCs on each production machine. However, latency

for predictions increased beyond acceptable limits. Analysis shows that the internet connection is at its capacity limit. How can the company resolve this issue MOST cost-effectively?

- A. Set up a 10 Gbps AWS Direct Connect connection between the production site and the nearest AWS Region
- B. Use the Direct Connect connection to upload the image
- C. Increase the size of the instances and the number of instances that are used by the SageMaker endpoint.
- D. Extend the long-running Lambda function that runs on AWS IoT Greengrass to compress the images and upload the compressed files to Amazon S3. Decompress the files by using a separate Lambda function that invokes the existing Lambda function to run the inference pipeline.
- E. Use auto scaling for SageMaker
- F. Set up an AWS Direct Connect connection between the production site and the nearest AWS Region
- G. Use the Direct Connect connection to upload the images.
- H. Deploy the Lambda function and the ML models onto the AWS IoT Greengrass core that is running on the industrial PCs that are installed on each machine
- I. Extend the long-running Lambda function that runs on AWS IoT Greengrass to invoke the Lambda function with the captured images and run the inference on the edge component that forwards the results directly to the web service.

**Answer:** D

#### NEW QUESTION 82

IT leadership wants to transition a company's existing machine learning data storage environment to AWS as a temporary ad hoc solution. The company currently uses a custom software process that heavily leverages SQL as a query language and exclusively stores generated CSV documents for machine learning. The ideal state for the company would be a solution that allows it to continue to use the current workforce of SQL experts. The solution must also support the storage of CSV and JSON files, and be able to query over semi-structured data. The following are high priorities for the company:

- Solution simplicity
- Fast development time
- Low cost
- High flexibility

What technologies meet the company's requirements?

- A. Amazon S3 and Amazon Athena
- B. Amazon Redshift and AWS Glue
- C. Amazon DynamoDB and DynamoDB Accelerator (DAX)
- D. Amazon RDS and Amazon ES

**Answer:** B

#### NEW QUESTION 86

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 89

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 90

A retail company is using Amazon Personalize to provide personalized product recommendations for its customers during a marketing campaign. The company sees a significant increase in sales of recommended items to existing customers immediately after deploying a new solution version, but these sales decrease a short time after deployment. Only historical data from before the marketing campaign is available for training. How should a data scientist adjust the solution?

- A. Use the event tracker in Amazon Personalize to include real-time user interactions.
- B. Add user metadata and use the HRNN-Metadata recipe in Amazon Personalize.
- C. Implement a new solution using the built-in factorization machines (FM) algorithm in Amazon SageMaker.
- D. Add event type and event value fields to the interactions dataset in Amazon Personalize.

**Answer:** A

#### NEW QUESTION 91

A manufacturing company has a large set of labeled historical sales data. The manufacturer would like to predict how many units of a particular part should be produced each quarter. Which machine learning approach should be used to solve this problem?

- A. Logistic regression
- B. Random Cut Forest (RCF)
- C. Principal component analysis (PCA)
- D. Linear regression

**Answer:** D

#### NEW QUESTION 95

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 99

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 102

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 107

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 112

A machine learning specialist works for a fruit processing company and needs to build a system that categorizes apples into three types. The specialist has collected a dataset that contains 150 images for each type of apple and applied transfer learning on a neural network that was pretrained on ImageNet with this dataset.

The company requires at least 85% accuracy to make use of the model.

After an exhaustive grid search, the optimal hyperparameters produced the following: 68% accuracy on the training set, 67% accuracy on the validation set.

What can the machine learning specialist do to improve the system's accuracy?



- A. Upload the model to an Amazon SageMaker notebook instance and use the Amazon SageMaker HPO feature to optimize the model's hyperparameters.
- B. Add more data to the training set and retrain the model using transfer learning to reduce the bias.
- C. Use a neural network model with more layers that are pretrained on ImageNet and apply transfer learning to increase the variance.
- D. Train a new model using the current neural network architecture.

**Answer:** B

#### NEW QUESTION 113

A machine learning specialist stores IoT soil sensor data in Amazon DynamoDB table and stores weather event data as JSON files in Amazon S3. The dataset in DynamoDB is 10 GB in size and the dataset in Amazon S3 is 5 GB in size. The specialist wants to train a model on this data to help predict soil moisture levels as a function of weather events using Amazon SageMaker.

Which solution will accomplish the necessary transformation to train the Amazon SageMaker model with the LEAST amount of administrative overhead?

- A. Launch an Amazon EMR cluster
- B. Create an Apache Hive external table for the DynamoDB table and S3 data
- C. Join the Hive tables and write the results out to Amazon S3.
- D. Crawl the data using AWS Glue crawler
- E. Write an AWS Glue ETL job that merges the two tables and writes the output to an Amazon Redshift cluster.
- F. Enable Amazon DynamoDB Streams on the sensor table
- G. Write an AWS Lambda function that consumes the stream and appends the results to the existing weather files in Amazon S3.
- H. Crawl the data using AWS Glue crawler
- I. Write an AWS Glue ETL job that merges the two tables and writes the output in CSV format to Amazon S3.

**Answer:** C

#### NEW QUESTION 116

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 model
- 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 task
- 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 model
- G. Create a workforce with a third-party AWS Marketplace vendor
- 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 task
- J. Use a private workforce as the labeling workforce.

**Answer:** B

#### NEW QUESTION 118

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 122

A manufacturing company wants to use machine learning (ML) to automate quality control in its facilities. The facilities are in remote locations and have limited internet connectivity. The company has 20 of training data that consists of labeled images of defective product parts. The training data is in the corporate on-premises data center.

The company will use this data to train a model for real-time defect detection in new parts as the parts move on a conveyor belt in the facilities. The company needs a solution that minimizes costs for compute infrastructure and that maximizes the scalability of resources for training. The solution also must facilitate the



company's use of an ML model in the low-connectivity environments.  
Which solution will meet these requirements?

- A. Move the training data to an Amazon S3 bucket
- B. Train and evaluate the model by using Amazon SageMaker
- C. Optimize the model by using SageMaker Ne
- D. Deploy the model on a SageMaker hosting services endpoint.
- E. Train and evaluate the model on premise
- F. Upload the model to an Amazon S3 bucket
- G. Deploy the model on an Amazon SageMaker hosting services endpoint.
- H. Move the training data to an Amazon S3 bucket
- I. Train and evaluate the model by using Amazon SageMaker
- J. Optimize the model by using SageMaker Ne
- K. Set up an edge device in the manufacturing facilities with AWS IoT Greengrass
- L. Deploy the model on the edge device.
- M. Train the model on premise
- N. Upload the model to an Amazon S3 bucket
- O. Set up an edge device in the manufacturing facilities with AWS IoT Greengrass
- P. Deploy the model on the edge device.

**Answer:** A

#### NEW QUESTION 126

A machine learning specialist needs to analyze comments on a news website with users across the globe. The specialist must find the most discussed topics in the comments that are in either English or Spanish.  
What steps could be used to accomplish this task? (Choose two.)

- A. Use an Amazon SageMaker BlazingText algorithm to find the topics independently from language. Proceed with the analysis.
- B. Use an Amazon SageMaker seq2seq algorithm to translate from Spanish to English, if necessary
- C. Use a SageMaker Latent Dirichlet Allocation (LDA) algorithm to find the topics.
- D. Use Amazon Translate to translate from Spanish to English, if necessary
- E. Use Amazon Comprehend topic modeling to find the topics.
- F. Use Amazon Translate to translate from Spanish to English, if necessary
- G. Use Amazon Lex to extract topics from the content.
- H. Use Amazon Translate to translate from Spanish to English, if necessary
- I. Use Amazon SageMaker Neural Topic Model (NTM) to find the topics.

**Answer:** B

#### NEW QUESTION 129

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 131

A Machine Learning Specialist has built a model using Amazon SageMaker built-in algorithms and is not getting expected accurate results. The Specialist wants to use hyperparameter optimization to increase the model's accuracy.  
Which method is the MOST repeatable and requires the LEAST amount of effort to achieve this?

- A. Launch multiple training jobs in parallel with different hyperparameters
- B. Create an AWS Step Functions workflow that monitors the accuracy in Amazon CloudWatch Logs and relaunches the training job with a defined list of hyperparameters
- C. Create a hyperparameter tuning job and set the accuracy as an objective metric.
- D. Create a random walk in the parameter space to iterate through a range of values that should be used for each individual hyperparameter

**Answer:** B

#### NEW QUESTION 136

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 141**

A company wants to predict the sale prices of houses based on available historical sales data. The target variable in the company's dataset is the sale price. The features include parameters such as the lot size, living area measurements, non-living area measurements, number of bedrooms, number of bathrooms, year built, and postal code. The company wants to use multi-variable linear regression to predict house sale prices. Which step should a machine learning specialist take to remove features that are irrelevant for the analysis and reduce the model's complexity?

- A. Plot a histogram of the features and compute their standard deviation
- B. Remove features with high variance.
- C. Plot a histogram of the features and compute their standard deviation
- D. Remove features with low variance.
- E. Build a heatmap showing the correlation of the dataset against itself
- F. Remove features with low mutual correlation scores.
- G. Run a correlation check of all features against the target variable
- H. Remove features with low target variable correlation scores.

**Answer:** D

**NEW QUESTION 144**

.....

## THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual AWS-Certified-Machine-Learning-Specialty Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the AWS-Certified-Machine-Learning-Specialty Product From:

<https://www.2passeasy.com/dumps/AWS-Certified-Machine-Learning-Specialty/>

## Money Back Guarantee

### **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