

**MLS-C01**

**AWS Certified Machine Learning - Specialty**

## Exam A

### QUESTION 1

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:

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

Based on the model evaluation results, why is this a viable model for production?

- 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.

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

### QUESTION 2

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

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

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.

Reference: <https://aws.amazon.com/blogs/big-data/building-a-recommendation-engine-with-spark-ml-on-amazon-emr-using-zeppelin/>

### QUESTION 3

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.

**Correct Answer: C**

**Section: (none)**

**Explanation**

**Explanation/Reference:**

#### QUESTION 4

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`.

**Correct Answer: C**

**Section: (none)**

**Explanation**

**Explanation/Reference:**

Reference: <https://aws.amazon.com/blogs/machine-learning/build-a-model-to-predict-the-impact-of-weather-on-urban-air-quality-using-amazon-sagemaker/?ref=Welcome.AI>

#### QUESTION 5

A Data Engineer needs to build a model using a dataset containing customer credit card information

How can the Data Engineer ensure the data remains encrypted and the credit card information is secure?

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

numbers from the customer data with AWS Glue.

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://docs.aws.amazon.com/sagemaker/latest/dg/pca.html>

#### QUESTION 6

A Machine Learning Specialist is using an Amazon SageMaker notebook instance in a private subnet of a corporate VPC. The ML Specialist has important data stored on the Amazon SageMaker notebook instance's Amazon EBS volume, and needs to take a snapshot of that EBS volume. However, the ML Specialist cannot find the Amazon SageMaker notebook instance's EBS volume or Amazon EC2 instance within the VPC.

Why is the ML Specialist not seeing the instance visible in the VPC?

- A. Amazon SageMaker notebook instances are based on the EC2 instances within the customer account, but they run outside of VPCs.
- B. Amazon SageMaker notebook instances are based on the Amazon ECS service within customer accounts.
- C. Amazon SageMaker notebook instances are based on EC2 instances running within AWS service accounts.
- D. Amazon SageMaker notebook instances are based on AWS ECS instances running within AWS service accounts.

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://docs.aws.amazon.com/sagemaker/latest/dg/gs-setup-working-env.html>

#### QUESTION 7

A Machine Learning Specialist is building a model that will perform time series forecasting using Amazon SageMaker. The Specialist has finished training the model and is now planning to perform load testing on the endpoint so they can configure Auto Scaling for the model variant.

Which approach will allow the Specialist to review the latency, memory utilization, and CPU utilization during the load test?

- A. Review SageMaker logs that have been written to Amazon S3 by leveraging Amazon Athena and Amazon QuickSight to visualize logs as they are being produced.
- B. Generate an Amazon CloudWatch dashboard to create a single view for the latency, memory utilization, and CPU utilization metrics that are outputted by Amazon SageMaker.
- C. Build custom Amazon CloudWatch Logs and then leverage Amazon ES and Kibana to query and visualize the log data as it is generated by Amazon SageMaker.
- D. Send Amazon CloudWatch Logs that were generated by Amazon SageMaker to Amazon ES and use Kibana to query and visualize the log data

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-cloudwatch.html>

#### QUESTION 8

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 queries.
- D. Use AWS Lambda to transform the data and Amazon Kinesis Data Analytics to run queries.

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 9

A Machine Learning Specialist is developing a custom video recommendation model for an application. The dataset used to train this model is very large with millions of data points and is hosted in an Amazon S3 bucket. The Specialist wants to avoid loading all of this data onto an Amazon SageMaker notebook instance because it would take hours to move and will exceed the attached 5 GB Amazon EBS volume on the notebook instance.

Which approach allows the Specialist to use all the data to train the model?

- A. Load a smaller subset of the data into the SageMaker notebook and train locally. Confirm that the training code is executing and the model parameters seem reasonable. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.
- B. Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to the instance. Train on a small amount of the data to verify the training code and hyperparameters. Go back to Amazon SageMaker and train using the full dataset
- C. Use AWS Glue to train a model using a small subset of the data to confirm that the data will be compatible with Amazon SageMaker. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.
- D. Load a smaller subset of the data into the SageMaker notebook and train locally. Confirm that the training code is executing and the model parameters seem reasonable. Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to train the full dataset.

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 10

A Machine Learning Specialist has completed a proof of concept for a company using a small data sample, and now the Specialist is ready to implement an end-to-end solution in AWS using Amazon SageMaker. The historical training data is stored in Amazon RDS.

Which approach should the Specialist use for training a model using that data?

- A. Write a direct connection to the SQL database within the notebook and pull data in
- B. Push the data from Microsoft SQL Server to Amazon S3 using an AWS Data Pipeline and provide the S3 location within the notebook.
- C. Move the data to Amazon DynamoDB and set up a connection to DynamoDB within the notebook to pull data in.
- D. Move the data to Amazon ElastiCache using AWS DMS and set up a connection within the notebook to pull data in for fast access.

**Correct Answer:** B

**Section:** (none)

## Explanation

### Explanation/Reference:

#### QUESTION 11

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

Which solution should the Specialist recommend?

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

**Correct Answer: C**

**Section: (none)**

### Explanation

### Explanation/Reference:

#### QUESTION 12

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

**Correct Answer: B**

**Section: (none)**

### Explanation

### Explanation/Reference:

Explanation:

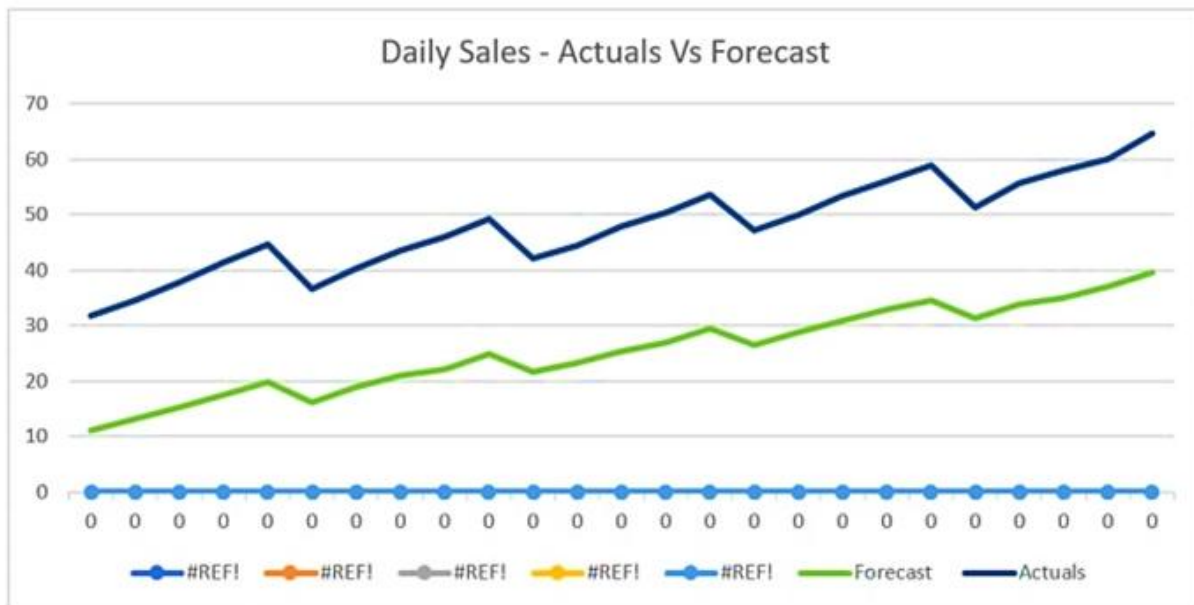
The goal of classification is to determine to which class or category a data point (customer in our case) belongs to. For classification problems, data scientists would use historical data with predefined target variables AKA labels (churner/non-churner) – answers that need to be predicted – to train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

Reference: <https://www.kdnuggets.com/2019/05/churn-prediction-machine-learning.html>

#### QUESTION 13

The displayed graph is from a forecasting model for testing a time series.



Considering the graph only, which conclusion should a Machine Learning Specialist make about the behavior of the model?

- A. The model predicts both the trend and the seasonality well
- B. The model predicts the trend well, but not the seasonality.
- C. The model predicts the seasonality well, but not the trend.
- D. The model does not predict the trend or the seasonality well.

**Correct Answer:** D

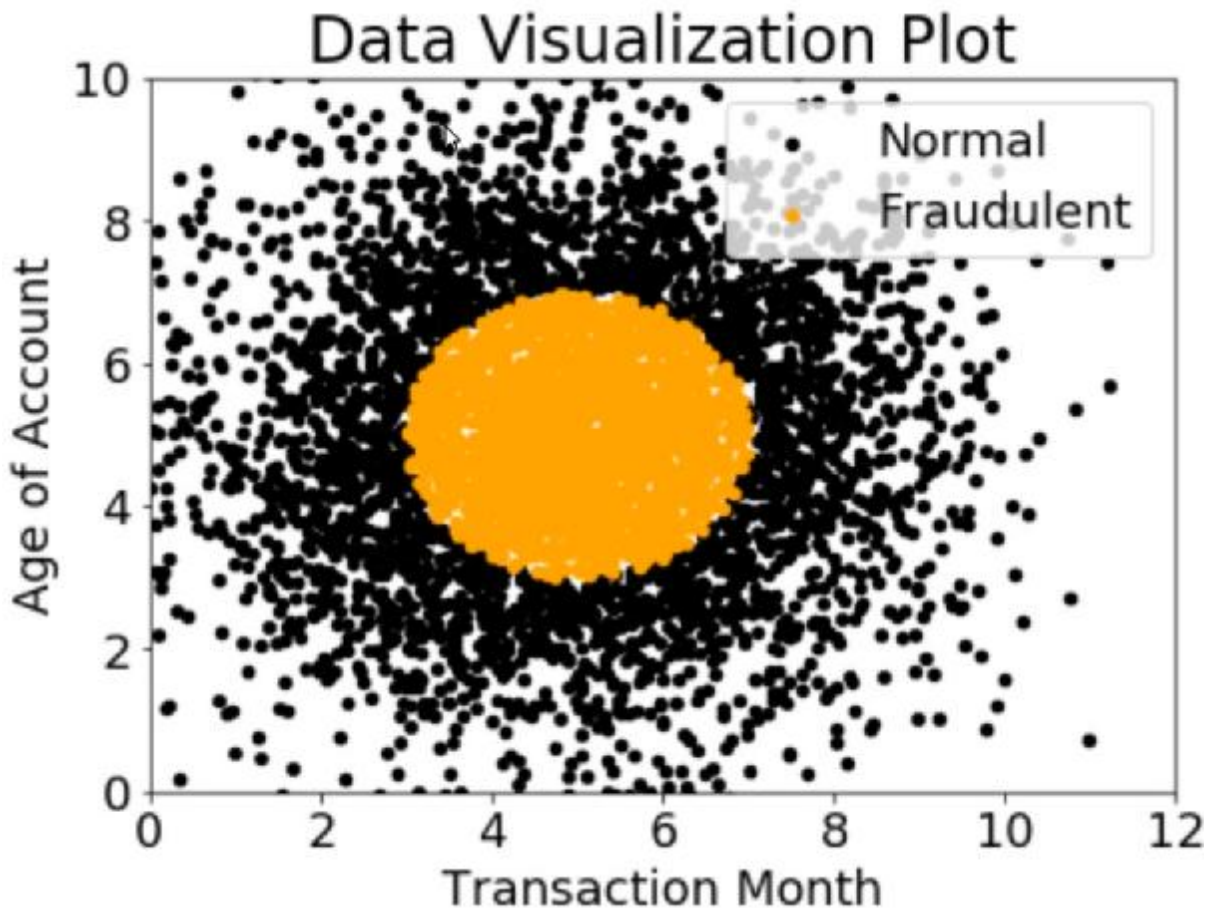
**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 14

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 accuracy?

- A. Long short-term memory (LSTM) model with scaled exponential linear unit (SELU)
- B. Logistic regression
- C. Support vector machine (SVM) with non-linear kernel
- D. Single perceptron with tanh activation function

**Correct Answer: C**

**Section: (none)**

**Explanation**

**Explanation/Reference:**

#### QUESTION 15

A Machine Learning Specialist at a company sensitive to security is preparing a dataset for model training. The dataset is stored in Amazon S3 and contains Personally Identifiable Information (PII).

The dataset:

- Must be accessible from a VPC only.
- Must not traverse the public internet.

How can these requirements be satisfied?

- A. Create a VPC endpoint and apply a bucket access policy that restricts access to the given VPC endpoint and the VPC.
- B. Create a VPC endpoint and apply a bucket access policy that allows access from the given VPC endpoint and an Amazon EC2 instance.
- C. Create a VPC endpoint and use Network Access Control Lists (NACLs) to allow traffic between only the



given VPC endpoint and an Amazon EC2 instance.

- D. Create a VPC endpoint and use security groups to restrict access to the given VPC endpoint and an Amazon EC2 instance

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://docs.aws.amazon.com/AmazonS3/latest/dev/example-bucket-policies-vpc-endpoint.html>

#### QUESTION 16

During mini-batch training of a neural network for a classification problem, a Data Scientist notices that training accuracy oscillates.

What is the MOST likely cause of this issue?

- A. The class distribution in the dataset is imbalanced.
- B. Dataset shuffling is disabled.
- C. The batch size is too big.
- D. The learning rate is very high.

**Correct Answer:** D

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://towardsdatascience.com/deep-learning-personal-notes-part-1-lesson-2-8946fe970b95>

#### QUESTION 17

An employee found a video clip with audio on a company's social media feed. The language used in the video is Spanish. English is the employee's first language, and they do not understand Spanish. The employee wants to do a sentiment analysis.

What combination of services is the MOST efficient to accomplish the task?

- A. Amazon Transcribe, Amazon Translate, and Amazon Comprehend
- B. Amazon Transcribe, Amazon Comprehend, and Amazon SageMaker seq2seq
- C. Amazon Transcribe, Amazon Translate, and Amazon SageMaker Neural Topic Model (NTM)
- D. Amazon Transcribe, Amazon Translate and Amazon SageMaker BlazingText

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 18

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.

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 19

A Machine Learning Specialist is building a logistic regression model that will predict whether or not a person will order a pizza. The Specialist is trying to build the optimal model with an ideal classification threshold.

What model evaluation technique should the Specialist use to understand how different classification thresholds will impact the model's performance?

- A. Receiver operating characteristic (ROC) curve
- B. Misclassification rate
- C. Root Mean Square Error (RMSE)
- D. L1 norm

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

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

#### QUESTION 20

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 vectors that store edit distance with every other word.
- D. Download word embeddings pre-trained on a large corpus.

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://aws.amazon.com/blogs/machine-learning/amazon-sagemaker-object2vec-adds-new-features-that-support-automatic-negative-sampling-and-speed-up-training/>

#### QUESTION 21

A Machine Learning Specialist is configuring Amazon SageMaker so multiple Data Scientists can access notebooks, train models, and deploy endpoints. To ensure the best operational performance, the Specialist needs to be able to track how often the Scientists are deploying models, GPU and CPU utilization on the deployed SageMaker endpoints, and all errors that are generated when an endpoint is invoked.

Which services are integrated with Amazon SageMaker to track this information? (Choose two.)

- A. AWS CloudTrail
- B. AWS Health
- C. AWS Trusted Advisor
- D. Amazon CloudWatch
- E. AWS Config

**Correct Answer:** AD  
**Section:** (none)  
**Explanation**

**Explanation/Reference:**

Reference: <https://aws.amazon.com/sagemaker/faqs/>

#### QUESTION 22

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 to 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.

**Correct Answer:** D  
**Section:** (none)  
**Explanation**

**Explanation/Reference:**

#### QUESTION 23

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)

**Correct Answer:** D  
**Section:** (none)  
**Explanation**

**Explanation/Reference:**

Reference: <https://towardsdatascience.com/building-a-convolutional-neural-network-cnn-in-keras-329fbbadc5f5>

#### QUESTION 24

A Machine Learning Specialist trained a regression model, but the first iteration needs optimizing. The Specialist needs to understand whether the model is more frequently overestimating or underestimating the target.

What option can the Specialist use to determine whether it is overestimating or underestimating the target

value?

- A. Root Mean Square Error (RMSE)
- B. Residual plots
- C. Area under the curve
- D. Confusion matrix

**Correct Answer: C**

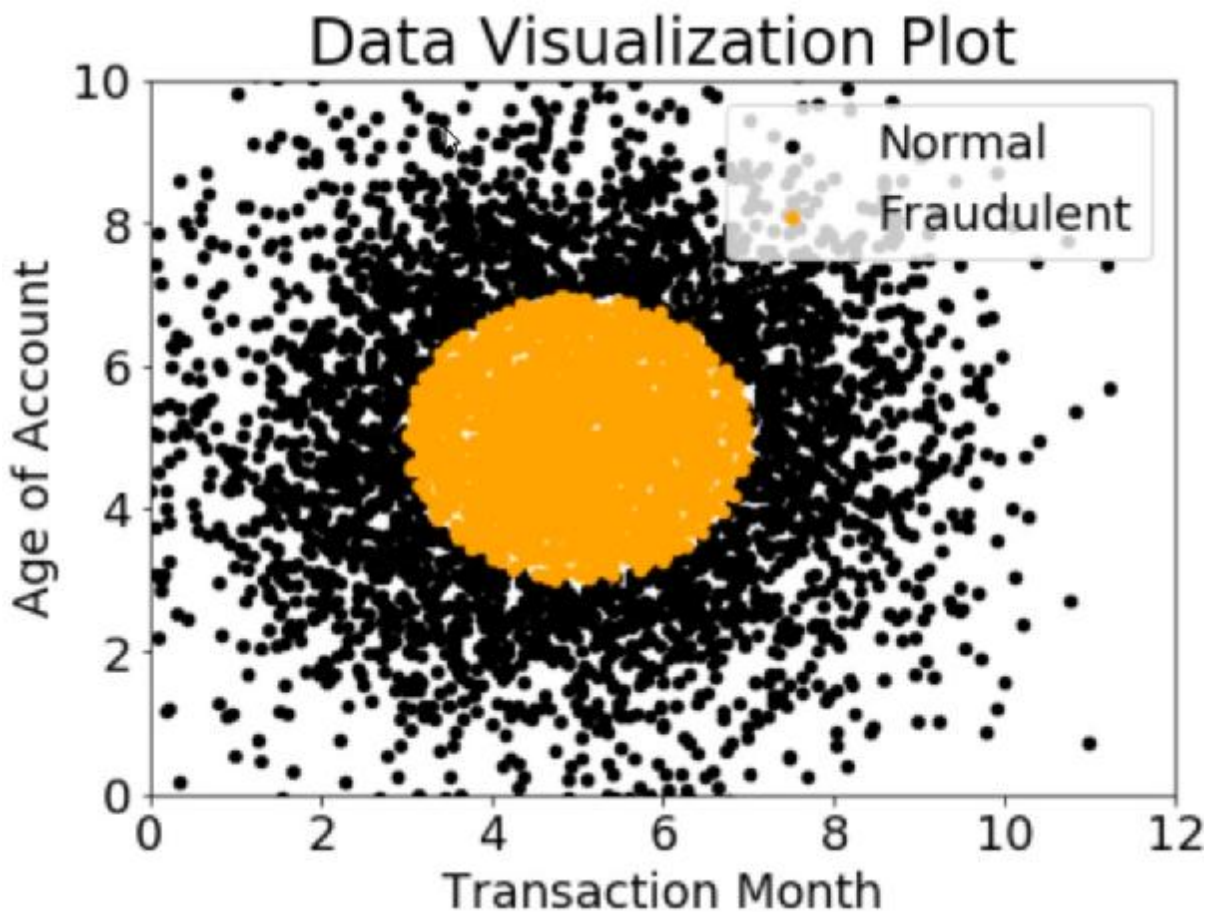
**Section: (none)**

**Explanation**

**Explanation/Reference:**

#### QUESTION 25

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

**Correct Answer: C**

**Section: (none)**

## Explanation

### Explanation/Reference:

#### QUESTION 26

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

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

Which visualization will accomplish this?

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

**Correct Answer:** B

**Section:** (none)

## Explanation

### Explanation/Reference:

#### QUESTION 27

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? (Choose 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 stopword dictionary.
- D. Correct the typography on "quck" to "quick."
- E. One-hot encode all words in the sentence.
- F. Tokenize the sentence into words.

**Correct Answer:** ABD

**Section:** (none)

## Explanation

### Explanation/Reference:

#### QUESTION 28

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

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://docs.aws.amazon.com/polly/latest/dg/ssml.html>

#### QUESTION 29

An insurance company is developing a new device for vehicles that uses a camera to observe drivers' behavior and alert them when they appear distracted. The company created approximately 10,000 training images in a controlled environment that a Machine Learning Specialist will use to train and evaluate machine learning models.

During the model evaluation, the Specialist notices that the training error rate diminishes faster as the number of epochs increases and the model is not accurately inferring on the unseen test images.

Which of the following should be used to resolve this issue? (Choose two.)

- A. Add vanishing gradient to the model.
- B. Perform data augmentation on the training data.
- C. Make the neural network architecture complex.
- D. Use gradient checking in the model.
- E. Add L2 regularization to the model.

**Correct Answer:** BD

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 30

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

**Correct Answer:** AEF

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 31

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

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

### QUESTION 32

Machine Learning Specialist is working with a media company to perform classification on popular articles from the company's website. The company is using random forests to classify how popular an article will be before it is published. A sample of the data being used is below.

Article_Title	Author	Top_Keywords	Day_Of_Week	URL_of_Article	Page_Views
Building a Big Data Platform	Jane Doe	Big Data, Spark, Hadoop	Tuesday	http://examplecorp.com/data_platform.html	1300456
Getting Started with Deep Learning	John Doe	Deep Learning, Machine Learning, Spark	Tuesday	http://examplecorp.com/started_deep_learning.html	1230661
MXNet ML Guide	Jane Doe	Machine Learning, MXNet, Logistic Regression	Thursday	http://examplecorp.com/mxnet_guide.html	937291
Intro to NoSQL Databases	Mary Major	NoSQL, Operations, Database	Monday	http://examplecorp.com/nosql_intro_guide.html	407812

Given the dataset, the Specialist wants to convert the Day\_Of\_Week column to binary values.

What technique should be used to convert this column to binary values?

- A. Binarization
- B. One-hot encoding
- C. Tokenization
- D. Normalization transformation

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

### QUESTION 33

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 999,000 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? (Choose two.)

- A. Add more deep trees to the random forest to enable the model to learn more features.
- B. Include a copy of the samples in the test dataset 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.

**Correct Answer:** BD

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 34

A Data Scientist is developing a machine learning model to predict future patient outcomes based on information collected about each patient and their treatment plans. The model should output a continuous value as its prediction. The data available includes labeled outcomes for a set of 4,000 patients. The study was conducted on a group of individuals over the age of 65 who have a particular disease that is known to worsen with age.

Initial models have performed poorly. While reviewing the underlying data, the Data Scientist notices that, out of 4,000 patient observations, there are 450 where the patient age has been input as 0. The other features for these observations appear normal compared to the rest of the sample population

How should the Data Scientist correct this issue?

- A. Drop all records from the dataset where age has been set to 0.
- B. Replace the age field value for records with a value of 0 with the mean or median value from the dataset
- C. Drop the age feature from the dataset and train the model using the rest of the features.
- D. Use k-means clustering to handle missing features

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 35

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.



**Correct Answer:** A  
**Section:** (none)  
**Explanation**

**Explanation/Reference:**

#### **QUESTION 36**

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.

**Correct Answer:** D  
**Section:** (none)  
**Explanation**

**Explanation/Reference:**

#### **QUESTION 37**

A Machine Learning Specialist working for an online fashion company wants to build a data ingestion solution for the company's Amazon S3-based data lake. The Specialist wants to create a set of ingestion mechanisms that will enable future capabilities comprised of:

- Real-time analytics
- Interactive analytics of historical data
- Clickstream analytics
- Product recommendations

Which services should the Specialist use?

- A. AWS Glue as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for real-time data insights; Amazon Kinesis Data Firehose for delivery to Amazon ES for clickstream analytics; Amazon EMR to generate personalized product recommendations
- B. Amazon Athena as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for near-real-time data insights; Amazon Kinesis Data Firehose for clickstream analytics; AWS Glue to generate personalized product recommendations
- C. AWS Glue as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for historical data insights; Amazon Kinesis Data Firehose for delivery to Amazon ES for clickstream analytics; Amazon EMR to generate personalized product recommendations
- D. Amazon Athena as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for historical data insights; Amazon DynamoDB streams for clickstream analytics; AWS Glue to generate personalized product recommendations

**Correct Answer:** A  
**Section:** (none)  
**Explanation**

**Explanation/Reference:**

**QUESTION 38**

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? (Choose 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.

**Correct Answer:** AD

**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 39**

A Machine Learning Specialist built an image classification deep learning model. However, the Specialist ran into an overfitting problem in which the training and testing accuracies were 99% and 75%, respectively.

How should the Specialist address this issue and what is the reason behind it?

- A. The learning rate should be increased because the optimization process was trapped at a local minimum.
- B. The dropout rate at the flatten layer should be increased because the model is not generalized enough.
- C. The dimensionality of dense layer next to the flatten layer should be increased because the model is not complex enough.
- D. The epoch number should be increased because the optimization process was terminated before it reached the global minimum.

**Correct Answer:** D

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: [https://www.tensorflow.org/tutorials/keras/overfit\\_and\\_underfit](https://www.tensorflow.org/tutorials/keras/overfit_and_underfit)

**QUESTION 40**

A Machine Learning team uses Amazon SageMaker to train an Apache MXNet handwritten digit classifier model using a research dataset. The team wants to receive a notification when the model is overfitting. Auditors want to view the Amazon SageMaker log activity report to ensure there are no unauthorized API calls.

What should the Machine Learning team do to address the requirements with the least amount of code and fewest steps?

- A. Implement an AWS Lambda function to log Amazon SageMaker API calls to Amazon S3. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to

receive a notification when the model is overfitting.

- B. Use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.
- C. Implement an AWS Lambda function to log Amazon SageMaker API calls to AWS CloudTrail. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.
- D. Use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. Set up Amazon SNS to receive a notification when the model is overfitting

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 41

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.

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 42

A Machine Learning Specialist is implementing a full Bayesian network on a dataset that describes public transit in New York City. One of the random variables is discrete, and represents the number of minutes New Yorkers wait for a bus given that the buses cycle every 10 minutes, with a mean of 3 minutes.

Which prior probability distribution should the ML Specialist use for this variable?

- A. Poisson distribution
- B. Uniform distribution
- C. Normal distribution
- D. Binomial distribution

**Correct Answer:** D

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 43

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

How should the Data Science team configure the notebook instance placement to meet these requirements?

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

**Correct Answer:** D

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 44

A Machine Learning Specialist has created a deep learning neural network model that performs well on the training data but performs poorly on the test data.

Which of the following methods should the Specialist consider using to correct this? (Choose three.)

- A. Decrease regularization.
- B. Increase regularization.
- C. Increase dropout.
- D. Decrease dropout.
- E. Increase feature combinations.
- F. Decrease feature combinations.

**Correct Answer:** BDE

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 45

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. 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.
- B. 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.

- 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 inserts it into an Amazon RDS PostgreSQL database. Have the Analysts query and run dashboards from the RDS database.
- D. 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.

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 46

An online reseller has a large, multi-column dataset with one column missing 30% of its data. A Machine Learning Specialist believes that certain columns in the dataset could be used to reconstruct the missing data.

Which reconstruction approach should the Specialist use to preserve the integrity of the dataset?

- A. Listwise deletion
- B. Last observation carried forward
- C. Multiple imputation
- D. Mean substitution

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://worldwidescience.org/topicpages/i/imputing+missing+values.html>

#### QUESTION 47

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.

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf> (46)

#### QUESTION 48

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.

What should the Specialist do to initialize the model to re-train it with the custom data?

- 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.

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 49

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. 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.
- B. 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. On each stream, use Amazon Rekognition Image to detect faces from a collection of known employees and alert when non-employees are detected.
- C. Install AWS DeepLens cameras and use the DeepLens\_Kinesis\_Video module to stream video to Amazon Kinesis Video Streams for each camera. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection on each stream, and alert when non-employees are detected.
- D. Install AWS DeepLens cameras and use the DeepLens\_Kinesis\_Video module to stream video to Amazon Kinesis Video Streams for each camera. 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.

**Correct Answer:** D

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://aws.amazon.com/blogs/machine-learning/video-analytics-in-the-cloud-and-at-the-edge-with-aws-deeplens-and-kinesis-video-streams/>

#### QUESTION 50

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. Find similar profiles on social media.
- D. Use a decision tree classifier engine on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media.

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 51

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

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 52

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

- Support querying old and new data on Amazon S3 through Amazon Athena and Amazon Redshift Spectrum.
- Support event-driven ETL pipelines
- Provide a quick and easy way to understand metadata

Which approach meets these requirements?

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

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 53

A company's Machine Learning Specialist needs to improve the training speed of a time-series forecasting model using TensorFlow. The training is currently implemented on a single-GPU machine and takes

approximately 23 hours to complete. The training needs to be run daily.

The model accuracy is acceptable, but the company anticipates a continuous increase in the size of the training data and a need to update the model on an hourly, rather than a daily, basis. The company also wants to minimize coding effort and infrastructure changes.

What should the Machine Learning Specialist do to the training solution to allow it to scale for future demand?

- A. Do not change the TensorFlow code. Change the machine to one with a more powerful GPU to speed up the training.
- B. Change the TensorFlow code to implement a Horovod distributed framework supported by Amazon SageMaker. Parallelize the training to as many machines as needed to achieve the business goals.
- C. Switch to using a built-in AWS SageMaker DeepAR model. Parallelize the training to as many machines as needed to achieve the business goals.
- D. Move the training to Amazon EMR and distribute the workload to as many machines as needed to achieve the business goals.

**Correct Answer: B**

**Section: (none)**

**Explanation**

**Explanation/Reference:**

#### **QUESTION 54**

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)

**Correct Answer: D**

**Section: (none)**

**Explanation**

**Explanation/Reference:**

#### **QUESTION 55**

A company is running a machine learning prediction service that generates 100 TB of predictions every day. A Machine Learning Specialist must generate a visualization of the daily precision-recall curve from the predictions, and forward a read-only version to the Business team.

Which solution requires the LEAST coding effort?

- A. Run a daily Amazon EMR workflow to generate precision-recall data, and save the results in Amazon S3. Give the Business team read-only access to S3.
- B. Generate daily precision-recall data in Amazon QuickSight, and publish the results in a dashboard shared with the Business team.
- C. Run a daily Amazon EMR workflow to generate precision-recall data, and save the results in Amazon S3. Visualize the arrays in Amazon QuickSight, and publish them in a dashboard shared with the Business team.
- D. Generate daily precision-recall data in Amazon ES, and publish the results in a dashboard shared with the Business team.

**Correct Answer: C**

**Section: (none)**

**Explanation**



**Explanation/Reference:**

**QUESTION 56**

A Machine Learning Specialist is preparing data for training on Amazon SageMaker. The Specialist is using one of the SageMaker built-in algorithms for the training. The dataset is stored in .CSV format and 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.

**Correct Answer: C**

**Section: (none)**

**Explanation**

**Explanation/Reference:**

**QUESTION 57**

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

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

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

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

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

**Correct Answer: B**

**Section: (none)**

**Explanation**

**Explanation/Reference:**

**QUESTION 58**

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

**Correct Answer: C**

**Section: (none)**

## Explanation

### Explanation/Reference:

#### QUESTION 59

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	99,966	34
	1	123

Which combination of steps should the Data Scientist take to reduce the number of false positive predictions by the model? (Choose 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 eval\_metric parameter to optimize based on AUC instead of error.
- E. Decrease the XGBoost max\_depth parameter because the model is currently overfitting the data.

**Correct Answer:** DE

**Section:** (none)

## Explanation

### Explanation/Reference:

#### QUESTION 60

A Machine Learning Specialist is assigned a TensorFlow project using Amazon SageMaker for training, and needs to continue working for an extended period with no Wi-Fi access.

Which approach should the Specialist use to continue working?

- A. Install Python 3 and boto3 on their laptop and continue the code development using that environment.
- B. Download the TensorFlow Docker container used in Amazon SageMaker from GitHub to their local environment, and use the Amazon SageMaker Python SDK to test the code.
- C. Download TensorFlow from tensorflow.org to emulate the TensorFlow kernel in the SageMaker environment.
- D. Download the SageMaker notebook to their local environment, then install Jupyter Notebooks on their laptop and continue the development in a local notebook.

**Correct Answer:** B

**Section:** (none)

## Explanation

### Explanation/Reference:

#### QUESTION 61

A Machine Learning Specialist is working with a large cybersecurity company that manages security events in real time for companies around the world. The cybersecurity company wants to design a solution that will allow it to use machine learning to score malicious events as anomalies on the data as it is being ingested. The company also wants to be able to save the results in its data lake for later processing and analysis.

What is the MOST efficient way to accomplish these tasks?

- A. Ingest the data using Amazon Kinesis Data Firehose, and use Amazon Kinesis Data Analytics Random Cut Forest (RCF) for anomaly detection. Then use Kinesis Data Firehose to stream the results to Amazon S3.
- B. Ingest the data into Apache Spark Streaming using Amazon EMR, and use Spark MLlib with k-means to perform anomaly detection. Then store the results in an Apache Hadoop Distributed File System (HDFS) using Amazon EMR with a replication factor of three as the data lake.
- C. Ingest the data and store it in Amazon S3. Use AWS Batch along with the AWS Deep Learning AMIs to train a k-means model using TensorFlow on the data in Amazon S3.
- D. Ingest the data and store it in Amazon S3. Have an AWS Glue job that is triggered on demand transform the new data. Then use the built-in Random Cut Forest (RCF) model within Amazon SageMaker to detect anomalies in the data.

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 62

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.

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://aws.amazon.com/big-data/real-time-analytics-featured-partners/>

#### QUESTION 63

A retail company intends to use machine learning to categorize new products. A labeled dataset of current products was provided to the Data Science team. The dataset includes 1,200 products. The labeled dataset has 15 features for each product such as title dimensions, weight, and price. Each product is labeled as belonging to one of six categories such as books, games, electronics, and movies.

Which model should be used for categorizing new products using the provided dataset for training?

- A. AnXGBoost model where the objective parameter is set to `multi:softmax`
- B. A deep convolutional neural network (CNN) with a softmax activation function for the last layer
- C. A regression forest where the number of trees is set equal to the number of product categories
- D. A DeepAR forecasting model based on a recurrent neural network (RNN)

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 64**

A Data Scientist is working on an application that performs sentiment analysis. The validation accuracy is poor, and the Data Scientist thinks that the cause may be a rich vocabulary and a low average frequency of words in the dataset.

Which tool should be used to improve the validation accuracy?

- A. Amazon Comprehend syntax analysis and entity detection
- B. Amazon SageMaker BlazingText cbow mode
- C. Natural Language Toolkit (NLTK) stemming and stop word removal
- D. Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizer

**Correct Answer:** D

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://monkeylearn.com/sentiment-analysis/>

**QUESTION 65**

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 bigram (OSB) transformation to apply a fixed-size sliding window to generate new features of a similar magnitude.

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-reference.html>

**QUESTION 66**

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.

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

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.

Reference: <https://www.cloudforecast.io/blog/using-parquet-on-athena-to-save-money-on-aws/>

**QUESTION 67**

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.

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Reference: <https://aws.amazon.com/step-functions/use-cases/>

**QUESTION 68**

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? (Choose 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

**Correct Answer:** CD

**Section:** (none)

**Explanation**

**Explanation/Reference:**

Explanation:

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

**QUESTION 69**

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.

**Correct Answer: B**

**Section: (none)**

**Explanation**

**Explanation/Reference:**

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.

Reference: <https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html>

#### QUESTION 70

A Machine Learning Specialist uploads a dataset to an Amazon S3 bucket protected with server-side encryption using AWS KMS.

How should the ML Specialist define the Amazon SageMaker notebook instance so it can read the same dataset from Amazon S3?

- A. Define security group(s) to allow all HTTP inbound/outbound traffic and assign those security group(s) to the Amazon SageMaker notebook instance.
- B. Configure the Amazon SageMaker notebook instance to have access to the VPC. Grant permission in the KMS key policy to the notebook's KMS role.
- C. Assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. Grant permission in the KMS key policy to that role.
- D. Assign the same KMS key used to encrypt data in Amazon S3 to the Amazon SageMaker notebook instance.

**Correct Answer: D**

**Section: (none)**

**Explanation**

**Explanation/Reference:**

Reference: <https://docs.aws.amazon.com/sagemaker/latest/dg/encryption-at-rest.html>

#### QUESTION 71

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 to 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 real-time 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.

**Correct Answer:** D

**Section:** (none)

**Explanation**

**Explanation/Reference:**

## QUESTION 72

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

**Correct Answer:** BE

**Section:** (none)

**Explanation**

**Explanation/Reference:**

## QUESTION 73

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 analysis. Use Jupyter notebooks from within AWS IoT Analytics to carry out analysis for anomalies.
- B. Use Amazon S3 for ingestion, storage, and further analysis. Use an Amazon EMR cluster to carry out

Apache Spark ML k-means clustering to determine anomalies.

- C. Use Amazon S3 for ingestion, storage, and further analysis. Use the Amazon SageMaker Random Cut Forest (RCF) algorithm to determine anomalies.
- D. Use Amazon Kinesis Data Firehose for ingestion and Amazon Kinesis Data Analytics Random Cut Forest (RCF) to perform anomaly detection. Use Kinesis Data Firehose to store data in Amazon S3 for further analysis.

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 74

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

**Correct Answer:** CE

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 75

A Machine Learning Specialist wants to determine the appropriate `SageMakerVariantInvocationsPerInstance` 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

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 76

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.

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 77

A Machine Learning Specialist needs to move and transform data in preparation for training. Some of the data needs to be processed in near-real time, and other data can be moved hourly. There are existing Amazon EMR MapReduce jobs to clean and feature engineering to perform on the data.

Which of the following services can feed data to the MapReduce jobs? (Choose two.)

- A. AWS DMS
- B. Amazon Kinesis
- C. AWS Data Pipeline
- D. Amazon Athena
- E. Amazon ES

**Correct Answer:** AE

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 78

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 code. Tag the Docker image with the registry hostname and upload it to Amazon ECR.
- B. Serialize the trained model so the format is compressed for deployment. Tag the Docker image with the registry hostname and upload it to Amazon S3.
- C. Serialize the trained model so the format is compressed for deployment. Build the image and upload it to Docker Hub.
- D. Build the Docker image with the inference code. Configure Docker Hub and upload the image to Amazon ECR.

**Correct Answer:** D

**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 79**

A trucking company is collecting live image data from its fleet of trucks across the globe. The data is growing rapidly and approximately 100 GB of new data is generated every day. The company wants to explore machine learning use cases while ensuring the data is only accessible to specific IAM users.

Which storage option provides the most processing flexibility and will allow access control with IAM?

- A. Use a database, such as Amazon DynamoDB, to store the images, and set the IAM policies to restrict access to only the desired IAM users.
- B. Use an Amazon S3-backed data lake to store the raw images, and set up the permissions using bucket policies.
- C. Setup up Amazon EMR with Hadoop Distributed File System (HDFS) to store the files, and restrict access to the EMR instances using IAM policies.
- D. Configure Amazon EFS with IAM policies to make the data available to Amazon EC2 instances owned by the IAM users.

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 80**

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

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 81**

A Data Scientist is training a multilayer perceptron (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

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### **QUESTION 82**

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

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### **QUESTION 83**

A real estate company wants to create a machine learning model for predicting housing prices based on a historical dataset. The dataset contains 32 features.

Which model will meet the business requirement?

- A. Logistic regression
- B. Linear regression
- C. K-means
- D. Principal component analysis (PCA)

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### **QUESTION 84**

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

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

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

Correct Answer: C

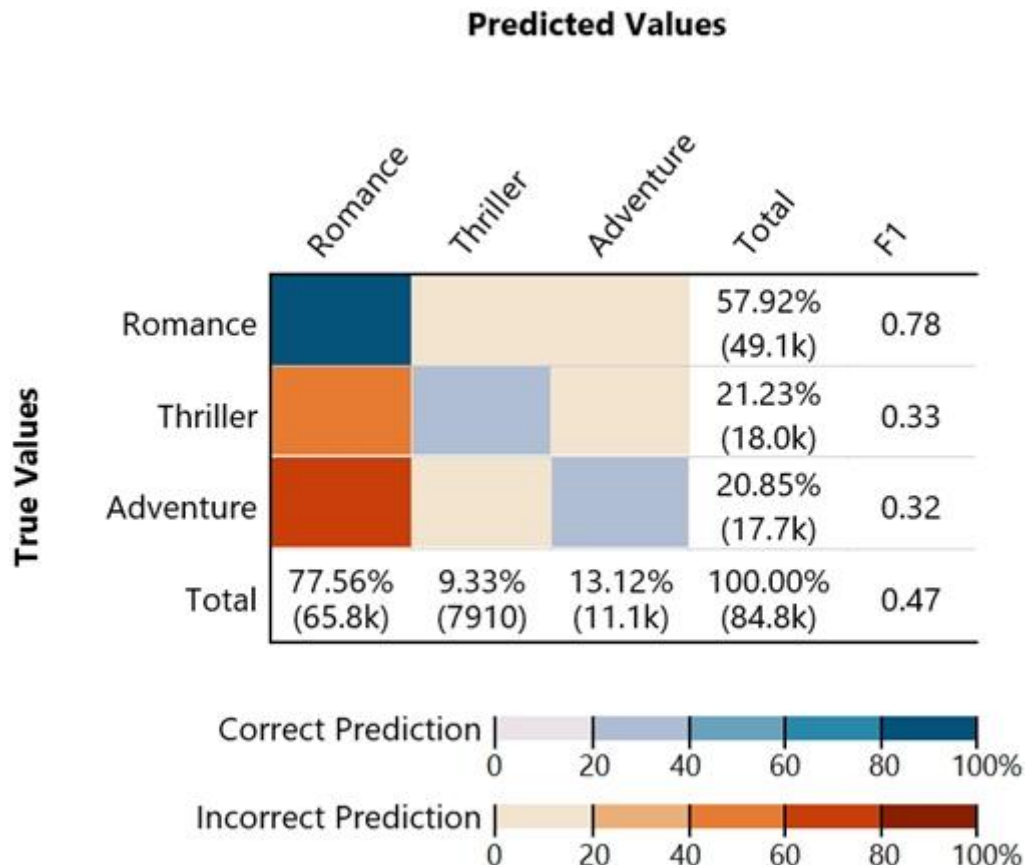
Section: (none)

Explanation

Explanation/Reference:

#### QUESTION 85

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  $77.56\% \times 0.78$  and the predicted class frequency for Adventure is  $20.85\% \times 0.32$

Correct Answer: B

Section: (none)

Explanation

Explanation/Reference:

#### QUESTION 86

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`

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 87

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 shows 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

**Correct Answer:** AB

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 88

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

**Correct Answer:** D

**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 89**

A Machine Learning Specialist is given a structured dataset on the shopping habits of a company's customer base. The dataset contains thousands of columns of data and hundreds of numerical columns for each customer. The Specialist wants to identify whether there are natural groupings for these columns across all customers and visualize the results as quickly as possible.

What approach should the Specialist take to accomplish these tasks?

- A. Embed the numerical features using the t-distributed stochastic neighbor embedding (t-SNE) algorithm and create a scatter plot.
- B. Run k-means using the Euclidean distance measure for different values of k and create an elbow plot.
- C. Embed the numerical features using the t-distributed stochastic neighbor embedding (t-SNE) algorithm and create a line graph.
- D. Run k-means using the Euclidean distance measure for different values of k and create box plots for each numerical column within each cluster.

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 90**

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

Which nodes should the Specialist launch on Spot Instances?

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

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 91**

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.

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 92**

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

**Correct Answer:** D

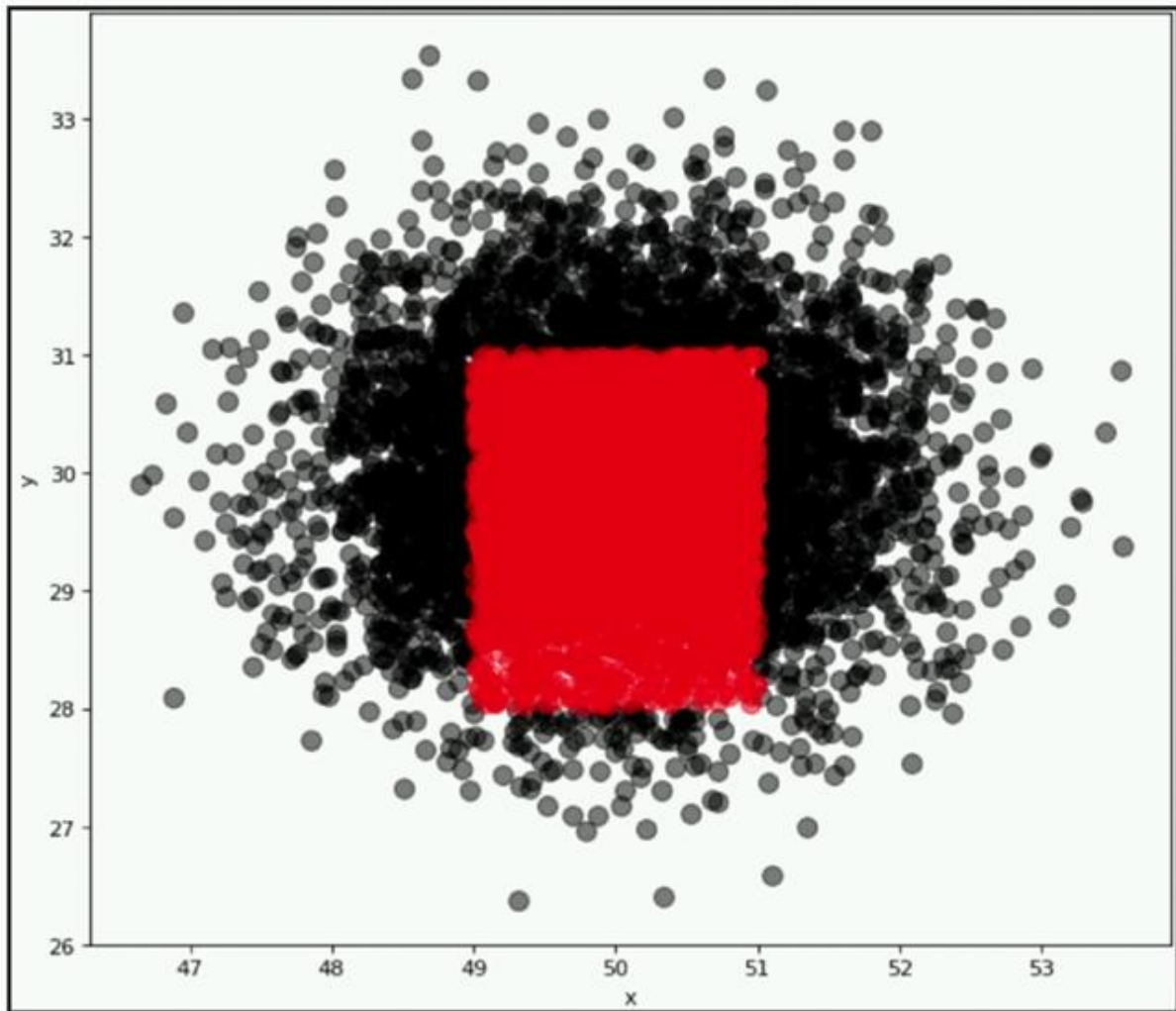
**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 93**

A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a machine learning specialist will build a binary classifier based on two features: age of account, denoted by  $x$ , and transaction month, denoted by  $y$ . The class distributions are illustrated in the provided figure. The positive class is portrayed in red, while the negative class is portrayed in black.



Which model would have the HIGHEST accuracy?

- A. Linear support vector machine (SVM)
- B. Decision tree
- C. Support vector machine (SVM) with a radial basis function kernel
- D. Single perceptron with a Tanh activation function

**Correct Answer: C**

**Section: (none)**

**Explanation**

**Explanation/Reference:**

#### QUESTION 94

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

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

- A. Choose a higher number of layers
- B. Choose a lower number of layers
- C. Choose a smaller learning rate
- D. Enable dropout



- E. Include all the images from the test set in the training set
- F. Enable early stopping

**Correct Answer:** ADE

**Section:** (none)

**Explanation**

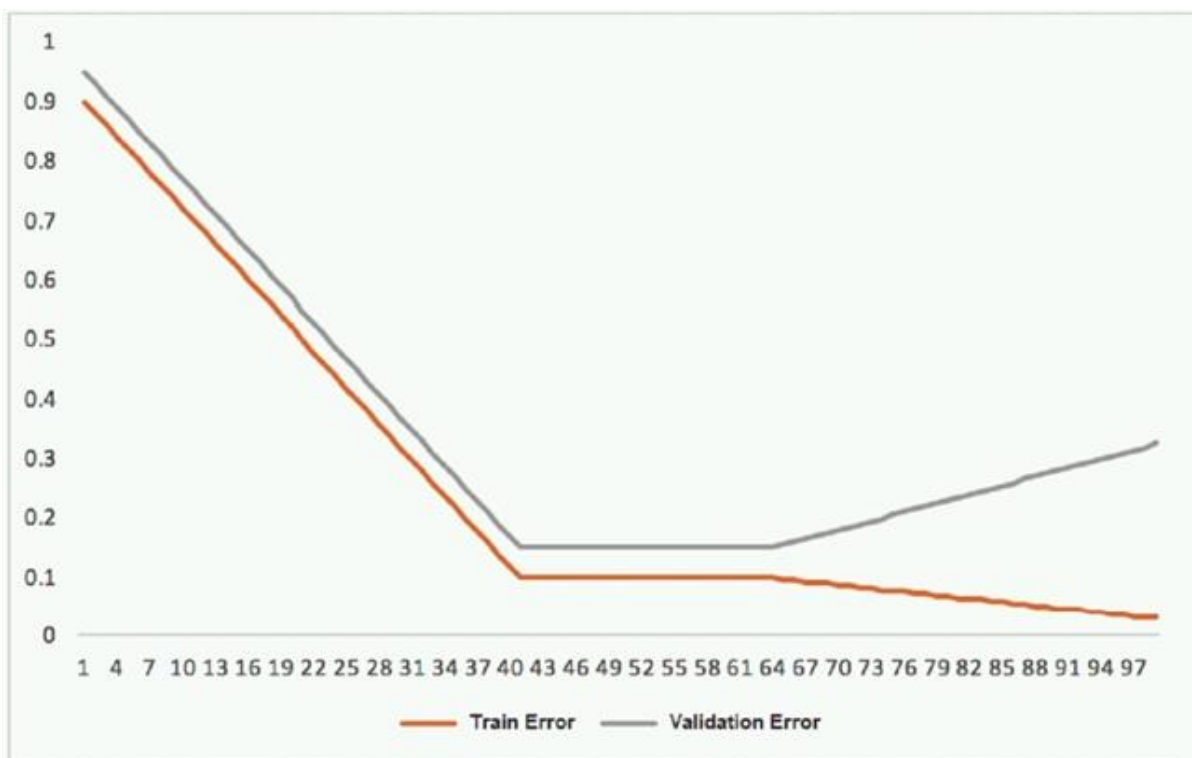
**Explanation/Reference:**

#### QUESTION 95

This graph shows the training and validation loss against the epochs for a neural network.

The network being trained is as follows:

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



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

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

**Correct Answer:** C

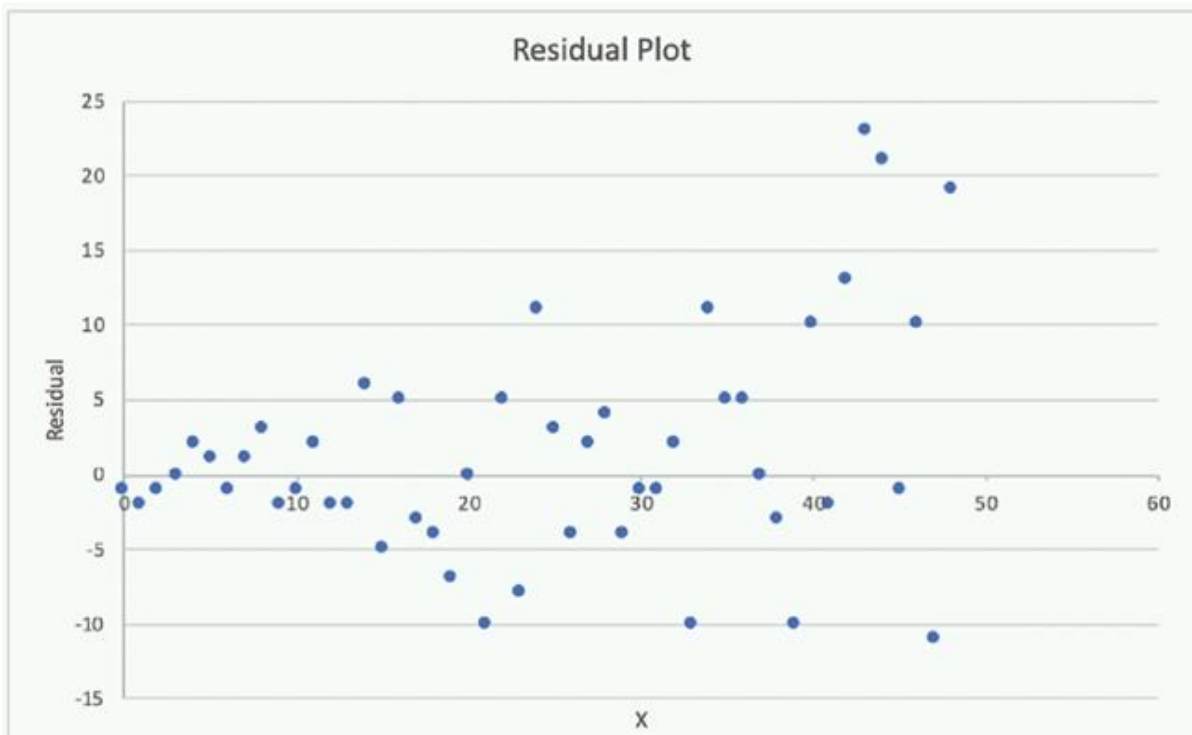
**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 96

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

**Correct Answer:** D

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 97

A large company has developed a BI application that generates reports and dashboards using data collected from various operational metrics. The company wants to provide executives with an enhanced experience so they can use natural language to get data from the reports. The company wants the executives to be able ask questions using written and spoken interfaces.

Which combination of services can be used to build this conversational interface? (Choose three.)

- A. Alexa for Business
- B. Amazon Connect
- C. Amazon Lex
- D. Amazon Polly
- E. Amazon Comprehend
- F. Amazon Transcribe

**Correct Answer:** BEF

**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 98**

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.

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 99**

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

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

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

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 100**

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

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### **QUESTION 101**

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.

**Correct Answer:** C

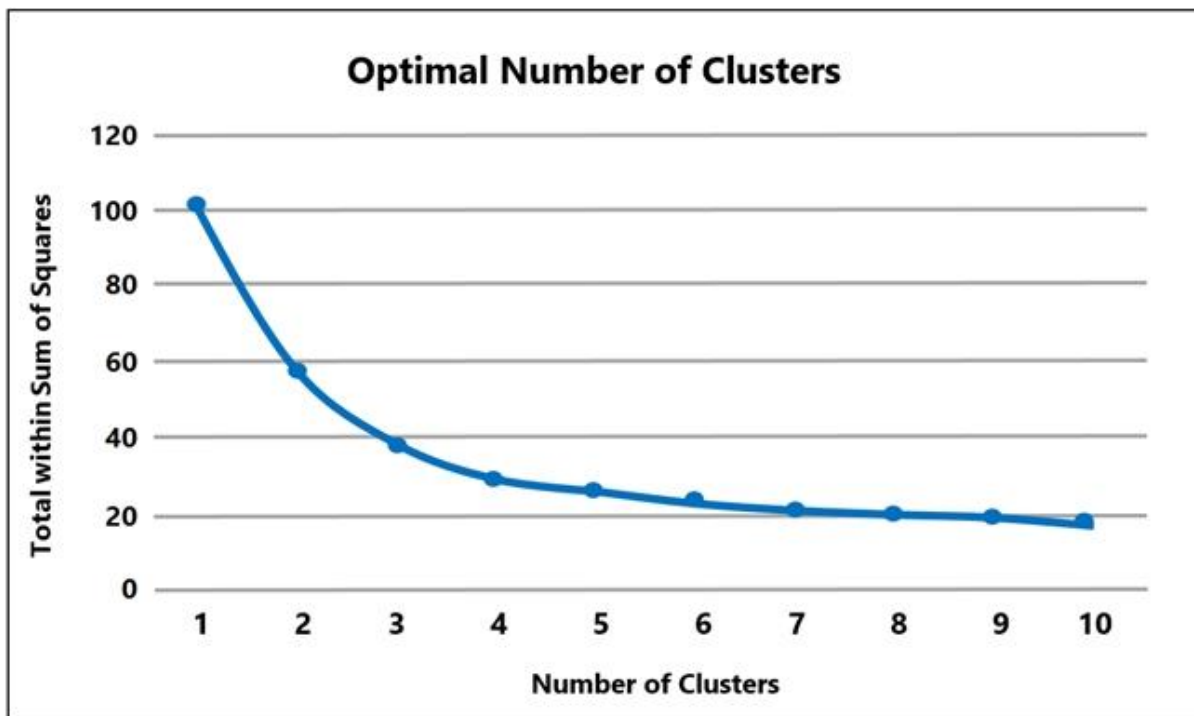
**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### **QUESTION 102**

A Machine Learning Specialist prepared the following graph displaying the results of k-means for  $k = [1..10]$ :



Considering the graph, what is a reasonable selection for the optimal choice of k?

- A. 1
- B. 4
- C. 7
- D. 10

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

### QUESTION 103

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 text. Use the Amazon SageMaker Neural Topic Model (NTM) and Object Detection algorithms to tag data into distinct categories/classes.
- D. 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.

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

**QUESTION 104**

A Machine Learning Specialist is working for an online retailer that wants to run analytics on every customer visit, processed through a machine learning pipeline. The data needs to be ingested by Amazon Kinesis Data Streams at up to 100 transactions per second, and the JSON data blob is 100 KB in size.

What is the MINIMUM number of shards in Kinesis Data Streams the Specialist should use to successfully ingest this data?

- A. 1 shards
- B. 10 shards
- C. 100 shards
- D. 1,000 shards

**Correct Answer: B**

**Section: (none)**

**Explanation**

**Explanation/Reference:**

**QUESTION 105**

A Machine Learning Specialist is deciding between building a naive Bayesian model or a full Bayesian network for a classification problem. The Specialist computes the Pearson correlation coefficients between each feature and finds that their absolute values range between 0.1 to 0.95.

Which model describes the underlying data in this situation?

- A. A naive Bayesian model, since the features are all conditionally independent.
- B. A full Bayesian network, since the features are all conditionally independent.
- C. A naive Bayesian model, since some of the features are statistically dependent.
- D. A full Bayesian network, since some of the features are statistically dependent.

**Correct Answer: C**

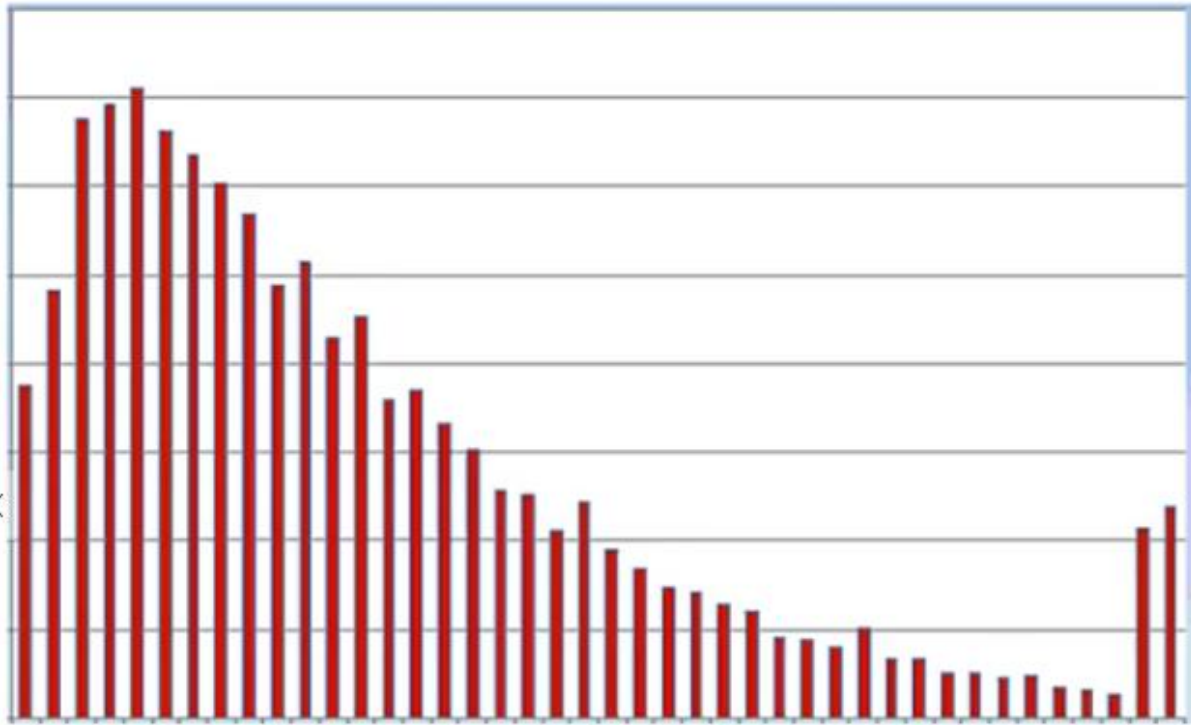
**Section: (none)**

**Explanation**

**Explanation/Reference:**

**QUESTION 106**

A Data Scientist is building a linear regression model and will use resulting p-values to evaluate the statistical significance of each coefficient. Upon inspection of the dataset, the Data Scientist discovers that most of the features are normally distributed. The plot of one feature in the dataset is shown in the graphic.



What transformation should the Data Scientist apply to satisfy the statistical assumptions of the linear regression model?

- A. Exponential transformation
- B. Logarithmic transformation
- C. Polynomial transformation
- D. Sinusoidal transformation

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 107

A Machine Learning Specialist is assigned to a Fraud Detection team and must tune an XGBoost model, which is working appropriately for test data. However, with unknown data, it is not working as expected. The existing parameters are provided as follows.

```
param = {  
    'eta': 0.05, # the training step for each iteration  
    'silent': 1, # logging mode - quiet  
    'n_estimators': 2000,  
    'max_depth': 30,  
    'min_child_weight': 3,  
    'gamma': 0,  
    'subsample': 0.8,  
    'objective': 'multi:softprob', # error evaluation for multiclass training  
    'num_class': 201} # the number of classes that exist in this dataset  
num_round = 60 # the number of training iterations
```

Which parameter tuning guidelines should the Specialist follow to avoid overfitting?

- A. Increase the max\_depth parameter value.

- B. Lower the max\_depth parameter value.
- C. Update the objective to binary:logistic.
- D. Lower the min\_child\_weight parameter value.

**Correct Answer:** B

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 108

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

The solution needs to do the following:

- Calculate an anomaly score for each web traffic entry.
- Adapt unusual event identification to changing web patterns over time.

Which approach should the data scientist implement to meet these requirements?

- A. Use historic web traffic data to train an anomaly detection model using the Amazon SageMaker Random Cut Forest (RCF) built-in model. Use an Amazon Kinesis Data Stream to process the incoming web traffic data. Attach a preprocessing AWS Lambda function to perform data enrichment by calling the RCF model to calculate the anomaly score for each record.
- B. Use historic web traffic data to train an anomaly detection model using the Amazon SageMaker built-in XGBoost model. Use an Amazon Kinesis Data Stream to process the incoming web traffic data. Attach a preprocessing AWS Lambda function to perform data enrichment by calling the XGBoost model to calculate the anomaly score for each record.
- C. Collect the streaming data using Amazon Kinesis Data Firehose. Map the delivery stream as an input source for Amazon Kinesis Data Analytics. Write a SQL query to run in real time against the streaming data with the k-Nearest Neighbors (kNN) SQL extension to calculate anomaly scores for each record using a tumbling window.
- D. Collect the streaming data using Amazon Kinesis Data Firehose. Map the delivery stream as an input source for Amazon Kinesis Data Analytics. Write a SQL query to run in real time against the streaming data with the Amazon Random Cut Forest (RCF) SQL extension to calculate anomaly scores for each record using a sliding window.

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 109

A Data Scientist received a set of insurance records, each consisting of a record ID, the final outcome among 200 categories, and the date of the final outcome. Some partial information on claim contents is also provided, but only for a few of the 200 categories. For each outcome category, there are hundreds of records distributed over the past 3 years. The Data Scientist wants to predict how many claims to expect in each category from month to month, a few months in advance.

What type of machine learning model should be used?

- A. Classification month-to-month using supervised learning of the 200 categories based on claim contents.
- B. Reinforcement learning using claim IDs and timestamps where the agent will identify how many claims in each category to expect from month to month.
- C. Forecasting using claim IDs and timestamps to identify how many claims in each category to expect from month to month.



- D. Classification with supervised learning of the categories for which partial information on claim contents is provided, and forecasting using claim IDs and timestamps for all other categories.

**Correct Answer:** D

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 110

A company that promotes healthy sleep patterns by providing cloud-connected devices currently hosts a sleep tracking application on AWS. The application collects device usage information from device users. The company's Data Science team is building a machine learning model to predict if and when a user will stop utilizing the company's devices. Predictions from this model are used by a downstream application that determines the best approach for contacting users.

The Data Science team is building multiple versions of the machine learning model to evaluate each version against the company's business goals. To measure long-term effectiveness, the team wants to run multiple versions of the model in parallel for long periods of time, with the ability to control the portion of inferences served by the models.

Which solution satisfies these requirements with MINIMAL effort?

- A. Build and host multiple models in Amazon SageMaker. Create multiple Amazon SageMaker endpoints, one for each model. Programmatically control invoking different models for inference at the application layer.
- B. Build and host multiple models in Amazon SageMaker. Create an Amazon SageMaker endpoint configuration with multiple production variants. Programmatically control the portion of the inferences served by the multiple models by updating the endpoint configuration.
- C. Build and host multiple models in Amazon SageMaker Neo to take into account different types of medical devices. Programmatically control which model is invoked for inference based on the medical device type.
- D. Build and host multiple models in Amazon SageMaker. Create a single endpoint that accesses multiple models. Use Amazon SageMaker batch transform to control invoking the different models through the single endpoint.

**Correct Answer:** D

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 111

An agricultural company is interested in using machine learning to detect specific types of weeds in a 100-acre grassland field. Currently, the company uses tractor-mounted cameras to capture multiple images of the field as 10x10 grids. The company also has a large training dataset that consists of annotated images of popular weed classes like broadleaf and non-broadleaf docks.

The company wants to build a weed detection model that will detect specific types of weeds and the location of each type within the field. Once the model is ready, it will be hosted on Amazon SageMaker endpoints. The model will perform real-time inferencing using the images captured by the cameras.

Which approach should a Machine Learning Specialist take to obtain accurate predictions?

- A. Prepare the images in RecordIO format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an image classification algorithm to categorize images into various weed classes.
- B. Prepare the images in Apache Parquet format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an object-detection single-shot multibox detector (SSD) algorithm.

- C. Prepare the images in RecordIO format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an object-detection single-shot multibox detector (SSD) algorithm.
- D. Prepare the images in Apache Parquet format and upload them to Amazon S3. Use Amazon SageMaker to train, test, and validate the model using an image classification algorithm to categorize images into various weed classes.

**Correct Answer:** C

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 112

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

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

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

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

**Correct Answer:** A

**Section:** (none)

**Explanation**

**Explanation/Reference:**

#### QUESTION 113

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

**Correct Answer:** D

**Section:** (none)

**Explanation**

**Explanation/Reference:**