

Exam Questions AWS-Certified-Machine-Learning-Specialty

AWS Certified Machine Learning - Specialty

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



NEW QUESTION 1

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

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

Answer: B

NEW QUESTION 2

A company is building a new version of a recommendation engine. Machine learning (ML) specialists need to keep adding new data from users to improve personalized recommendations. The ML specialists gather data from the users' interactions on the platform and from sources such as external websites and social media.

The pipeline cleans, transforms, enriches, and compresses terabytes of data daily, and this data is stored in Amazon S3. A set of Python scripts was coded to do the job and is stored in a large Amazon EC2 instance. The whole process takes more than 20 hours to finish, with each script taking at least an hour. The company wants to move the scripts out of Amazon EC2 into a more managed solution that will eliminate the need to maintain servers.

Which approach will address all of these requirements with the LEAST development effort?

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

Answer: B

NEW QUESTION 3

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

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

Answer: A

NEW QUESTION 4

An e-commerce company wants to launch a new cloud-based product recommendation feature for its web application. Due to data localization regulations, any sensitive data must not leave its on-premises data center, and the product recommendation model must be trained and tested using nonsensitive data only. Data transfer to the cloud must use IPsec. The web application is hosted on premises with a PostgreSQL database that contains all the data. The company wants the data to be uploaded securely to Amazon S3 each day for model retraining.

How should a machine learning specialist meet these requirements?

- A. Create an AWS Glue job to connect to the PostgreSQL database instance.
- B. Ingest tables without sensitive data through an AWS Site-to-Site VPN connection directly into Amazon S3.
- C. Create an AWS Glue job to connect to the PostgreSQL database instance.
- D. Ingest all data through an AWS Site-to-Site VPN connection into Amazon S3 while removing sensitive data using a PySpark job.
- E. Use AWS Database Migration Service (AWS DMS) with table mapping to select PostgreSQL tables with no sensitive data through an SSL connection.
- F. Replicate data directly into Amazon S3.
- G. Use PostgreSQL logical replication to replicate all data to PostgreSQL in Amazon EC2 through AWS Direct Connect with a VPN connection.
- H. Use AWS Glue to move data from Amazon EC2 to Amazon S3.

Answer: C

NEW QUESTION 5

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

Answer: C

NEW QUESTION 6

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

Which set of steps should the engineer take to build the labeling pipeline with the LEAST effort?

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

Answer: C

Explanation:

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

NEW QUESTION 7

A web-based company wants to improve its conversion rate on its landing page Using a large historical dataset of customer visits, the company has repeatedly trained a multi-class deep learning network algorithm on Amazon SageMaker However there is an overfitting problem training data shows 90% accuracy in predictions, while test data shows 70% accuracy only

The company needs to boost the generalization of its model before deploying it into production to maximize conversions of visits to purchases

Which action is recommended to provide the HIGHEST accuracy model for the company's test and validation data?

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

Answer: C

Explanation:

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

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

NEW QUESTION 8

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

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

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

Which solution provides the MOST accurate result?

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

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

B. Run three different HPO jobs that use different learning rates from the following intervals for MinValue and MaxValue while using the same number of training jobs for each HPO job:[0.01, 0.1][0.001, 0.01][0.0001, 0.001]Select the most accurate hyperparameter configuration from these three HPO jobs.

C. Modify the HPO configuration as follows: C:\Users\Admin\Desktop\Data\Odt data\Untitled.jpg

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

Select the most accurate hyperparameter configuration from this training job.

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

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

Answer: C

NEW QUESTION 9

A Machine Learning Specialist was given a dataset consisting of unlabeled data The Specialist must create a model that can help the team classify the data into different buckets What model should be used to complete this work?

- A. K-means clustering
- B. Random Cut Forest (RCF)
- C. XGBoost
- D. BlazingText

Answer: A

NEW QUESTION 10

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

- A. Download the AWS SDK for the Spark environment
- B. Install the SageMaker Spark library in the Spark environment.
- C. Use the appropriate estimator from the SageMaker Spark Library to train a model.
- D. Compress the training data into a ZIP file and upload it to a pre-defined Amazon S3 bucket.
- E. Use the sageMakerMode
- F. transform method to get inferences from the model hosted in SageMaker
- G. Convert the DataFrame object to a CSV file, and use the CSV file as input for obtaining inferences from SageMaker.

Answer: DEF

NEW QUESTION 10

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

- A. Latent Dirichlet Allocation (LDA)
- B. Recurrent neural network (RNN)
- C. K-means
- D. Convolutional neural network (CNN)

Answer: C

NEW QUESTION 12

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

- A. Add labels over time to indicate which engine faults occur at what time in the future to turn this into a supervised learning problem Use a recurrent neural network (RNN) to train the model to recognize when an engine might need maintenance for a certain fault.
- B. This data requires an unsupervised learning algorithm Use Amazon SageMaker k-means to cluster the data
- C. Add labels over time to indicate which engine faults occur at what time in the future to turn this into a supervised learning problem Use a convolutional neural network (CNN) to train the model to recognize when an engine might need maintenance for a certain fault.
- D. This data is already formulated as a time series Use Amazon SageMaker seq2seq to model the time series.

Answer: B

NEW QUESTION 16

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

Answer: C

NEW QUESTION 18

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

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

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

Answer: CE

NEW QUESTION 19

A company wants to use automatic speech recognition (ASR) to transcribe messages that are less than 60 seconds long from a voicemail-style application. The company requires the correct identification of 200 unique product names, some of which have unique spellings or pronunciations.

The company has 4,000 words of Amazon SageMaker Ground Truth voicemail transcripts it can use to customize the chosen ASR model. The company needs to ensure that everyone can update their customizations multiple times each hour.

Which approach will maximize transcription accuracy during the development phase?

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

Answer: A

NEW QUESTION 21

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

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

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

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

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

Answer: A

NEW QUESTION 22

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? (Select TWO.)

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

Answer: BC

Explanation:

<https://aws.amazon.com/jp/emr/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-car>

NEW QUESTION 25

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

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

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

Answer: B

NEW QUESTION 26

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

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

Answer: A

NEW QUESTION 31

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 analysts and entity detection
- B. Amazon SageMaker BlazingText allow mode
- C. Natural Language Toolkit (NLTK) stemming and stop word removal
- D. Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizers

Answer: A

NEW QUESTION 35

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

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

Which process will improve the testing accuracy the MOST?

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

Answer: B

NEW QUESTION 36

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)

Answer: B

NEW QUESTION 37

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

Answer: B

Explanation:

A/B testing with Amazon SageMaker is required in the Exam. In A/B testing, you test different variants of your models and compare how each variant performs. Amazon SageMaker enables you to test multiple models or model versions behind the `same endpoint` using `production variants`. Each production variant identifies a machine learning (ML) model and the resources deployed for hosting the model. To test multiple models by `distributing traffic` between them, specify the `percentage of the traffic` that gets routed to each model by specifying the `weight` for each `production variant` in the endpoint configuration.
<https://docs.aws.amazon.com/sagemaker/latest/dg/model-ab-testing.html#model-testing-target-variant>

NEW QUESTION 40

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.

Answer: C

NEW QUESTION 41

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.

Answer: C

NEW QUESTION 42

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 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.

Answer: B

NEW QUESTION 44

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

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

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

Answer: A

NEW QUESTION 49

A Machine Learning Specialist needs to create a data repository to hold a large amount of time-based training data for a new model. In the source system, new files are added every hour Throughout a single 24-hour period, the volume of hourly updates will change significantly. The Specialist always wants to train on the last 24 hours of the data

Which type of data repository is the MOST cost-effective solution?

- A. An Amazon EBS-backed Amazon EC2 instance with hourly directories
- B. An Amazon RDS database with hourly table partitions
- C. An Amazon S3 data lake with hourly object prefixes
- D. An Amazon EMR cluster with hourly hive partitions on Amazon EBS volumes

Answer: C

NEW QUESTION 51

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

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

- A. Add an Amazon SageMaker Debugger hook to the script to capture key metric
- B. Run the script as an AWS Glue job.
- C. Add an Amazon SageMaker Experiments tracker to the script to capture key metric
- D. Run the script as an AWS Glue job.
- E. Add an Amazon SageMaker Debugger hook to the script to capture key parameter
- F. Run the script as a SageMaker processing job.
- G. Add an Amazon SageMaker Experiments tracker to the script to capture key parameter
- H. Run the script as a SageMaker processing job.

Answer: B

NEW QUESTION 55

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

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

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

Answer: ACF

NEW QUESTION 57

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

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

Which approach will provide the information required for further analysis?

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

Answer: B

NEW QUESTION 60

A financial company is trying to detect credit card fraud. The company observed that, on average, 2% of credit card transactions were fraudulent. A data scientist trained a classifier on a year's worth of credit card transactions data. The model needs to identify the fraudulent transactions (positives) from the regular ones (negatives). The company's goal is to accurately capture as many positives as possible.

Which metrics should the data scientist use to optimize the model? (Choose two.)

- A. Specificity
- B. False positive rate
- C. Accuracy
- D. Area under the precision-recall curve

E. True positive rate

Answer: DE

NEW QUESTION 62

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

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

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

Answer: AB

NEW QUESTION 64

A Machine Learning Specialist is working for a credit card processing company and receives an unbalanced dataset containing credit card transactions. It contains 99,000 valid transactions and 1,000 fraudulent transactions. The Specialist is asked to score a model that was run against the dataset. The Specialist has been advised that identifying valid transactions is equally as important as identifying fraudulent transactions. What metric is BEST suited to score the model?

- A. Precision
- B. Recall
- C. Area Under the ROC Curve (AUC)
- D. Root Mean Square Error (RMSE)

Answer: A

NEW QUESTION 65

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

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

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

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

- A. Change the XGBoost eval_metric parameter to optimize based on Root Mean Square Error (RMSE).
- B. Increase the XGBoost scale_pos_weight parameter to adjust the balance of positive and negative weights.
- C. Increase the XGBoost max_depth parameter because the model is currently underfitting the data.
- D. Change the XGBoost eval_metric parameter to optimize based on Area Under the ROC Curve (AUC).
- E. Decrease the XGBoost max_depth parameter because the model is currently overfitting the data.

Answer: BD

NEW QUESTION 67

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

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

Answer: A

NEW QUESTION 71

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

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

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

- A. Implement the solution using AWS Deep Learning Containers and run the container as a job using AWS Batch on a GPU-compatible Spot Instance
- B. Implement the solution using a low-cost GPU-compatible Amazon EC2 instance and use the AWS Instance Scheduler to schedule the task

- C. Implement the solution using AWS Deep Learning Containers, run the workload using AWS Fargate running on Spot Instances, and then schedule the task using the built-in task scheduler
- D. Implement the solution using Amazon ECS running on Spot Instances and schedule the task using the ECS service scheduler

Answer: C

NEW QUESTION 76

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

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

Answer: B

NEW QUESTION 80

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

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

Answer: A

NEW QUESTION 83

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

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

Answer: C

Explanation:

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

NEW QUESTION 85

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

How will the data scientist MOST effectively model the problem?

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

Answer: D

NEW QUESTION 88

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? (Select THREE.)

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

Answer: BCD

NEW QUESTION 89

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

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

Answer: AB

NEW QUESTION 92

A company has an ecommerce website with a product recommendation engine built in TensorFlow. The recommendation engine endpoint is hosted by Amazon SageMaker. Three compute-optimized instances support the expected peak load of the website. Response times on the product recommendation page are increasing at the beginning of each month. Some users are encountering errors. The website receives the majority of its traffic between 8 AM and 6 PM on weekdays in a single time zone. Which of the following options are the MOST effective in solving the issue while keeping costs to a minimum? (Choose two.)

- A. Configure the endpoint to use Amazon Elastic Inference (EI) accelerators.
- B. Create a new endpoint configuration with two production variants.
- C. Configure the endpoint to automatically scale with the InvocationsPerInstance metric.
- D. Deploy a second instance pool to support a blue/green deployment of models.
- E. Reconfigure the endpoint to use burstable instances.

Answer: BD

NEW QUESTION 94

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

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

Answer: C

NEW QUESTION 98

A machine learning specialist is running an Amazon SageMaker endpoint using the built-in object detection algorithm on a P3 instance for real-time predictions in a company's production application. When evaluating the model's resource utilization, the specialist notices that the model is using only a fraction of the GPU. Which architecture changes would ensure that provisioned resources are being utilized effectively?

- A. Redeploy the model as a batch transform job on an M5 instance.
- B. Redeploy the model on an M5 instance
- C. Attach Amazon Elastic Inference to the instance.
- D. Redeploy the model on a P3dn instance.
- E. Deploy the model onto an Amazon Elastic Container Service (Amazon ECS) cluster using a P3 instance.

Answer: B

Explanation:

<https://aws.amazon.com/machine-learning/elastic-inference/>

NEW QUESTION 100

A company that runs an online library is implementing a chatbot using Amazon Lex to provide book recommendations based on category. This intent is fulfilled by an AWS Lambda function that queries an Amazon DynamoDB table for a list of book titles, given a particular category. For testing, there are only three categories implemented as the custom slot types: "comedy," "adventure," and "documentary."

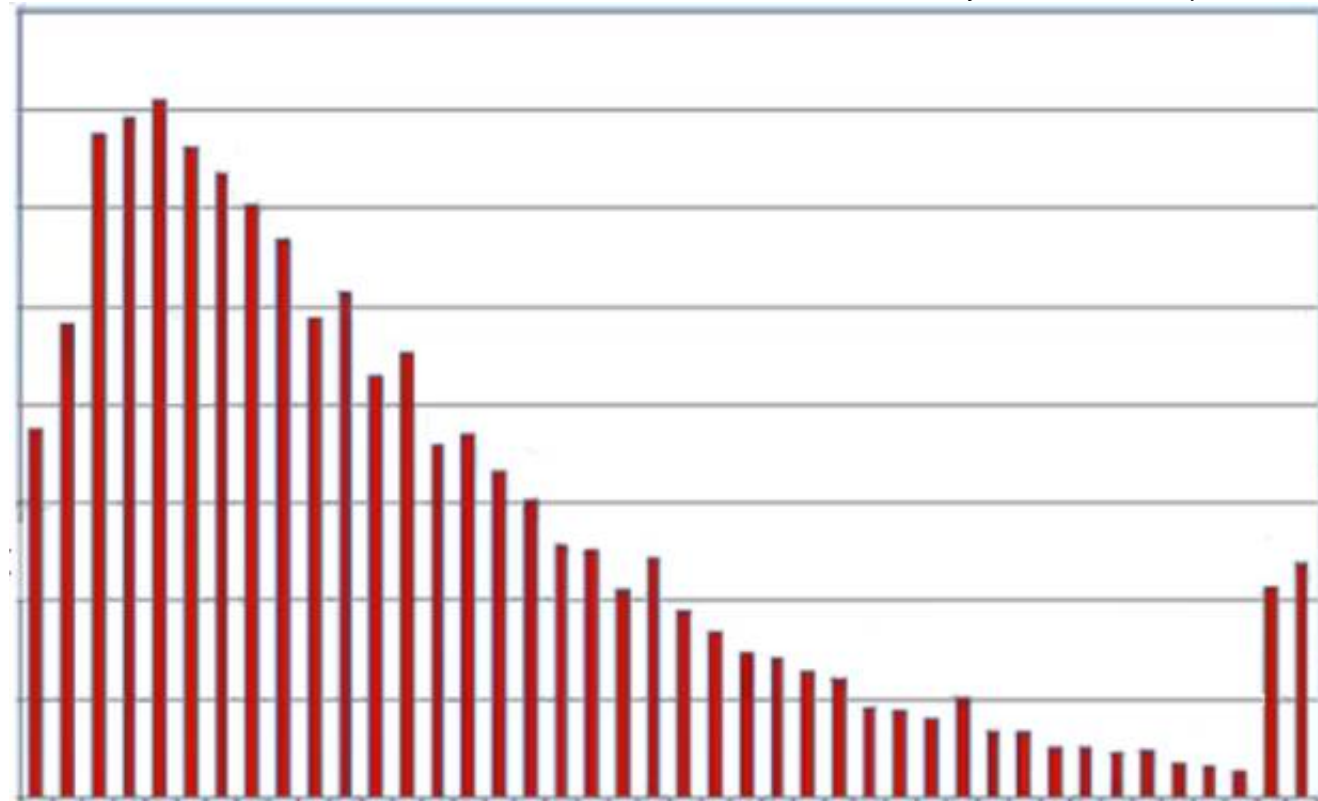
A machine learning (ML) specialist notices that sometimes the request cannot be fulfilled because Amazon Lex cannot understand the category spoken by users with utterances such as "funny," "fun," and "humor." The ML specialist needs to fix the problem without changing the Lambda code or data in DynamoDB. How should the ML specialist fix the problem?

- A. Add the unrecognized words in the enumeration values list as new values in the slot type.
- B. Create a new custom slot type, add the unrecognized words to this slot type as enumeration values, and use this slot type for the slot.
- C. Use the AMAZON.SearchQuery built-in slot types for custom searches in the database.
- D. Add the unrecognized words as synonyms in the custom slot type.

Answer: C

NEW QUESTION 103

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

Answer: A

NEW QUESTION 108

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

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

Answer: D

NEW QUESTION 109

A monitoring service generates 1 TB of scale metrics record data every minute. A Research team performs queries on this data using Amazon Athena. The queries run slowly due to the large volume of data, and the team requires better performance.

How should the records be stored in Amazon S3 to improve query performance?

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

Answer: D

NEW QUESTION 110

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

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

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

Answer: A

NEW QUESTION 112

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.
- B. Confirm that the training code is executing and the model parameters seem reasonable.
- C. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.
- D. Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to the instance.
- E. Train on a small amount of the data to verify the training code and hyperparameter.
- F. Go back to Amazon SageMaker and train using the full dataset.
- G. 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.
- H. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.
- I. Load a smaller subset of the data into the SageMaker notebook and train locally.
- J. Confirm that the training code is executing and the model parameters seem reasonable.
- K. Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to train the full dataset.

Answer: A

NEW QUESTION 116

A machine learning (ML) specialist is administering a production Amazon SageMaker endpoint with model monitoring configured. Amazon SageMaker Model Monitor detects violations on the SageMaker endpoint, so the ML specialist re-trains the model with the latest dataset. This dataset is statistically representative of the current production traffic. The ML specialist notices that even after deploying the new SageMaker model and running the first monitoring job, the SageMaker endpoint still has violations. What should the ML specialist do to resolve the violations?

- A. Manually trigger the monitoring job to re-evaluate the SageMaker endpoint traffic sample.
- B. Run the Model Monitor baseline job again on the new training set.
- C. Configure Model Monitor to use the new baseline.
- D. Delete the endpoint and recreate it with the original configuration.
- E. Retrain the model again by using a combination of the original training set and the new training set.

Answer: B

NEW QUESTION 117

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

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

Answer: C

NEW QUESTION 118

A health care company is planning to use neural networks to classify their X-ray images into normal and abnormal classes. The labeled data is divided into a training set of 1,000 images and a test set of 200 images. The initial training of a neural network model with 50 hidden layers yielded 99% accuracy on the training set, but only 55% accuracy on the test set. What changes should the Specialist consider to solve this issue? (Choose three.)

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

Answer: ADE

NEW QUESTION 122

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

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

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

Answer: BE

NEW QUESTION 127

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 10 × 10 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.

Answer: C

NEW QUESTION 128

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

How can the ML team solve this issue?

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

Answer: A

NEW QUESTION 130

A company offers an online shopping service to its customers. The company wants to enhance the site's security by requesting additional information when customers access the site from locations that are different from their normal location. The company wants to update the process to call a machine learning (ML) model to determine when additional information should be requested.

The company has several terabytes of data from its existing ecommerce web servers containing the source IP addresses for each request made to the web server. For authenticated requests, the records also contain the login name of the requesting user.

Which approach should an ML specialist take to implement the new security feature in the web application?

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

Answer: C

NEW QUESTION 135

A large consumer goods manufacturer has the following products on sale

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

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

Which solution should a Machine Learning Specialist apply?

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

Answer: B

Explanation:

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

NEW QUESTION 140

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.

Answer: B

NEW QUESTION 143

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.
- C. 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.
- D. 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.
- E. 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.

Answer: A

NEW QUESTION 146

A bank's Machine Learning team is developing an approach for credit card fraud detection. The company has a large dataset of historical data labeled as fraudulent. The goal is to build a model to take the information from new transactions and predict whether each transaction is fraudulent or not.

Which built-in Amazon SageMaker machine learning algorithm should be used for modeling this problem?

- A. Seq2seq
- B. XGBoost
- C. K-means
- D. Random Cut Forest (RCF)

Answer: C

NEW QUESTION 150

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

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

Answer: D

NEW QUESTION 154

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

What should the Specialist do to meet this objective?

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

Answer: B

Explanation:

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

NEW QUESTION 159

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

Answer: B

NEW QUESTION 164

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

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

Answer: D

NEW QUESTION 166

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

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

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

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

- A. Change the XGBoost eval_metric parameter to optimize based on rmse instead of error.
- B. Increase the XGBoost scale_pos_weight parameter to adjust the balance of positive and negative weights.
- C. Increase the XGBoost max_depth parameter because the model is currently underfitting the data.
- D. Change the XGBoost 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.

Answer: DE

NEW QUESTION 167

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

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

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

Answer: D

NEW QUESTION 170

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

What should the Specialist do to meet these requirements?

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

Answer: D

NEW QUESTION 172

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

Answer: C

NEW QUESTION 173

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

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

Answer: CD

NEW QUESTION 174

A Machine Learning Specialist is applying a linear least squares regression model to a dataset with 1 000 records and 50 features Prior to training, the ML Specialist notices that two features are perfectly linearly dependent Why could this be an issue for the linear least squares regression model?

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

Answer: C

NEW QUESTION 178

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

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

Answer: C

NEW QUESTION 181

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

Answer: B

NEW QUESTION 186

A data scientist uses an Amazon SageMaker notebook instance to conduct data exploration and analysis. This requires certain Python packages that are not natively available on Amazon SageMaker to be installed on the notebook instance. How can a machine learning specialist ensure that required packages are automatically available on the notebook instance for the data scientist to use?

- A. Install AWS Systems Manager Agent on the underlying Amazon EC2 instance and use Systems Manager Automation to execute the package installation commands.
- B. Create a Jupyter notebook file (.ipynb) with cells containing the package installation commands to execute and place the file under the /etc/init directory of each Amazon SageMaker notebook instance.
- C. Use the conda package manager from within the Jupyter notebook console to apply the necessary conda packages to the default kernel of the notebook.
- D. Create an Amazon SageMaker lifecycle configuration with package installation commands and assign the lifecycle configuration to the notebook instance.

Answer: D

Explanation:

<https://docs.aws.amazon.com/sagemaker/latest/dg/nbi-add-external.html>

NEW QUESTION 191

A data scientist needs to identify fraudulent user accounts for a company's ecommerce platform. The company wants the ability to determine if a newly created

account is associated with a previously known fraudulent user. The data scientist is using AWS Glue to cleanse the company's application logs during ingestion. Which strategy will allow the data scientist to identify fraudulent accounts?

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

Answer: B

NEW QUESTION 196

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

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

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

Answer: B

Explanation:

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

NEW QUESTION 198

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.

Answer: B

NEW QUESTION 203

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

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

Answer: D

NEW QUESTION 207

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.

Answer: B

NEW QUESTION 211

A data scientist is using an Amazon SageMaker notebook instance and needs to securely access data stored in a specific Amazon S3 bucket. How should the data scientist accomplish this?

- A. Add an S3 bucket policy allowing GetObject, PutObject, and ListBucket permissions to the AmazonSageMaker notebook ARN as principal.
- B. Encrypt the objects in the S3 bucket with a custom AWS Key Management Service (AWS KMS) key that only the notebook owner has access to.
- C. Attach the policy to the IAM role associated with the notebook that allows GetObject, PutObject, and ListBucket operations to the specific S3 bucket.
- D. Use a script in a lifecycle configuration to configure the AWS CLI on the instance with an access key ID and secret.

Answer: C

NEW QUESTION 212

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

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

Answer: A

NEW QUESTION 214

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

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

Answer: A

NEW QUESTION 218

.....

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

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

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

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

Money Back Guarantee

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

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