

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 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. 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?

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

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

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

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

Answer: C

NEW QUESTION 5

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

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

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

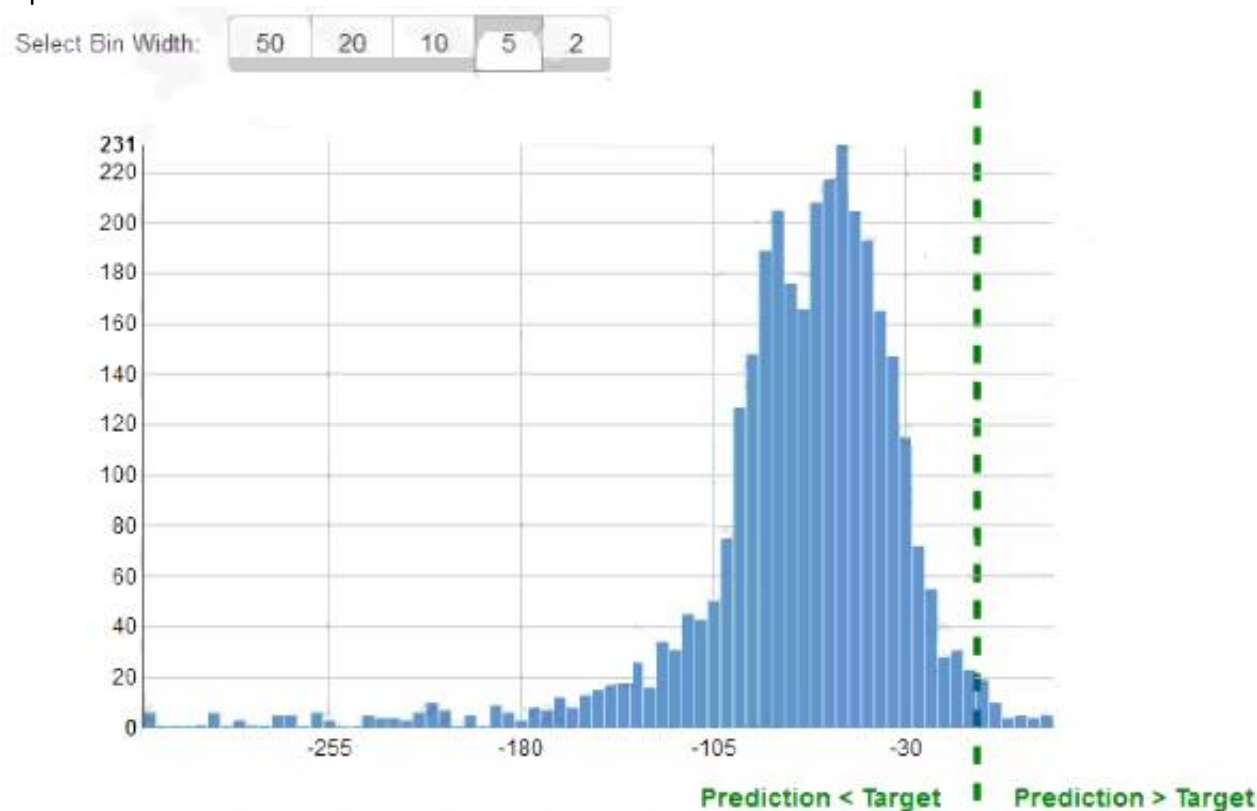
Answer: C

Explanation:

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

NEW QUESTION 6

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



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

Answer: D

NEW QUESTION 7

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

Answer: A

NEW QUESTION 8

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

Here is an example from the dataset:

"The quck BROWN FOX jumps over the lazy dog "

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

- A. Perform part-of-speech tagging and keep the action verb and the nouns only
- B. Normalize all words by making the sentence lowercase
- C. Remove stop words using an English 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.

Answer: BCF

NEW QUESTION 9

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

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

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

- A. The model is 86% accurate and the cost incurred by the company as a result of false negatives is less than the false positives.
- B. The precision of the model is 86%, which is less than the accuracy of the model.
- C. The model is 86% accurate and the cost incurred by the company as a result of false positives is less than the false negatives.
- D. The precision of the model is 86%, which is greater than the accuracy of the model.

Answer: A

NEW QUESTION 10

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

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

- A. Use the SageMaker batch transform feature to transform the training data into a DataFrame
- B. Use AWS Glue to compress the data into the Apache Parquet format
- C. Transform the dataset into the Recordio protobuf format
- D. Use the SageMaker hyperparameter optimization feature to automatically optimize the data

Answer: C

NEW QUESTION 10

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 15

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

The bank's data science team developed a model that classifies account transactions and understands credit eligibility. The data science team used the XGBoost algorithm to train the model. The team used 7 years of bank transaction historical data for training and hyperparameter tuning over the course of several days.

The accuracy of the model is sufficient, but the credit team is struggling to explain accurately why the model denies credit to some customers. The credit team has almost no skill in data science.

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

- A. Use Amazon SageMaker Studio to rebuild the mode
- B. Create a notebook that uses the XGBoost training container to perform model trainin
- C. Deploy the model at an endpoint
- D. Enable Amazon SageMaker Model Monitor to store inference
- E. Use the inferences to create Shapley values that help explain model behavio
- F. Create a chart that shows features and SHapley Additive explanation (SHAP) values to explain to the credit team how the features affect the model outcomes.
- G. Use Amazon SageMaker Studio to rebuild the mode
- H. Create a notebook that uses the XGBoost training container to perform model trainin
- I. Activate Amazon SageMaker Debugger, and configure it to calculate and collect Shapley value
- J. Create a chart that shows features and SHapley Additive explanation (SHAP) values to explain to the credit team how the features affect the model outcomes.
- K. Create an Amazon SageMaker notebook instanc

- L. Use the notebook instance and the XGBoost library to locally retrain the mode
- M. Use the plot_importance() method in the Python XGBoost interface to create a feature importance chart
- N. Use that chart to explain to the credit team how the features affect the model outcomes.
- O. Use Amazon SageMaker Studio to rebuild the mode
- P. Create a notebook that uses the XGBoost training container to perform model trainin
- Q. Deploy the model at an endpoint
- R. Use Amazon SageMakerProcessing to post-analyze the model and create a feature importance explainability chart automatically for the credit team.

Answer: C

NEW QUESTION 18

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

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

Answer: C

NEW QUESTION 21

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

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

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

Answer: B

NEW QUESTION 22

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

Which combination of AWS services will meet these requirements?

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

Answer: A

NEW QUESTION 23

A large company has developed a B1 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 to ask questions using written and spoken interlaces.

Which combination of services can be used to build this conversational interface? (Select THREE)

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

Answer: BEF

NEW QUESTION 28

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

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

- A. Add a VPC endpoint policy to allow access to the IAM users.
- B. Modify the users' IAM policy to allow access to Amazon SageMaker Service API calls only.

- C. Modify the security group on the endpoint network interface to restrict access to the instances.
- D. Modify the ACL on the endpoint network interface to restrict access to the instances.
- E. Add a SageMaker Runtime VPC endpoint interface to the VPC.

Answer: AC

NEW QUESTION 33

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.

Answer: C

NEW QUESTION 36

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 37

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

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

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

- A. Create a service-linked role for Amazon Elastic Container Service (Amazon ECS) with access to the S3 bucket.
- B. Create an ECS cluster that is based on an AWS Deep Learning Containers image.
- C. Write the code to perform the feature engineering.
- D. Train a logistic regression model for predicting the price, pointing to the bucket with the dataset.
- E. Wait for the training job to complete.
- F. Perform the inferences.
- G. Create an Amazon SageMaker notebook with a new IAM role that is associated with the notebook.
- H. Pull the dataset from the S3 bucket.
- I. Explore different combinations of feature engineering transformations, regression algorithms, and hyperparameters.
- J. Compare all the results in the notebook, and deploy the most accurate configuration in an endpoint for predictions.
- K. Create an IAM role with access to Amazon S3, Amazon SageMaker, and AWS Lambda.
- L. Create a training job with the SageMaker built-in XGBoost model pointing to the bucket with the dataset.
- M. Specify the price as the target feature.
- N. Wait for the job to complete.
- O. Load the model artifact to a Lambda function for inference on prices of new houses.
- P. Create an IAM role for Amazon SageMaker with access to the S3 bucket.
- Q. Create a SageMaker AutoML job with SageMaker Autopilot pointing to the bucket with the dataset.
- R. Specify the price as the target attribute.
- S. Wait for the job to complete.
- T. Deploy the best model for predictions.

Answer: A

NEW QUESTION 39

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 43

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

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

Answer: C

NEW QUESTION 47

A large JSON dataset for a project has been uploaded to a private Amazon S3 bucket. The Machine Learning Specialist wants to securely access and explore the data from an Amazon SageMaker notebook instance. A new VPC was created and assigned to the Specialist. How can the privacy and integrity of the data stored in Amazon S3 be maintained while granting access to the Specialist for analysis?

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

Answer: B

NEW QUESTION 50

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 variants.
- 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 models.
- 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 55

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 58

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 60

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 63

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 65

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 allows traffic to Amazon S3.

Answer: ACF

NEW QUESTION 67

Amazon Connect has recently been rolled 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 68

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 70

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

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

- A. Change the notebook instance type to a memory optimized instance with the same vCPU number as the ml.m5.4xlarge instance

- B. Stop the notebook when it is not in use
- C. Run both data preprocessing and feature engineering development on that instance.
- D. Keep the notebook instance type and size the same
- E. Stop the notebook when it is not in use
- F. Run data preprocessing on a P3 instance type with the same memory as the ml.m5.4xlarge instance by using Amazon SageMaker Processing.
- G. Change the notebook instance type to a smaller general purpose instance
- H. Stop the notebook when it is not in use
- I. Run data preprocessing on an ml.r5 instance with the same memory size as the ml.m5.4xlarge instance by using Amazon SageMaker Processing.
- J. Change the notebook instance type to a smaller general purpose instance
- K. Stop the notebook when it is not in use
- L. Run data preprocessing on an R5 instance with the same memory size as the ml.m5.4xlarge instance by using the Reserved Instance option.

Answer: B

NEW QUESTION 73

A Machine Learning Specialist is attempting to build a linear regression model.
Given the displayed residual plot only, what is the MOST likely problem with the model?

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

Answer: D

NEW QUESTION 77

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

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

Answer: D

NEW QUESTION 80

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
- C. Grant permission in the KMS key policy to the notebook's KMS role.
- D. Assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset
- E. Grant permission in the KMS key policy to that role.
- F. Assign the same KMS key used to encrypt data in Amazon S3 to the Amazon SageMaker notebook instance.

Answer: D

NEW QUESTION 84

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

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

Answer: C

Explanation:

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

NEW QUESTION 86

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 92

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

Answer: B

NEW QUESTION 93

A manufacturing company asks its Machine Learning Specialist to develop a model that classifies defective parts into one of eight defect types. The company has provided roughly 100,000 images per defect type for training. During the initial training of the image classification model, the Specialist notices that the validation accuracy is 80%, while the training accuracy is 90%. It is known that human-level performance for this type of image classification is around 90%. What should the Specialist consider to fix this issue?

- A. A longer training time
- B. Making the network larger
- C. Using a different optimizer
- D. Using some form of regularization

Answer: D

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 99

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 100

A credit card company wants to build a credit scoring model to help predict whether a new credit card applicant will default on a credit card payment. The company has collected data from a large number of sources with thousands of raw attributes. Early experiments to train a classification model revealed that many attributes are highly correlated, the large number of features slows down the training speed significantly, and that there are some overfitting issues.

The Data Scientist on this project would like to speed up the model training time without losing a lot of information from the original dataset.

Which feature engineering technique should the Data Scientist use to meet the objectives?

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

Answer: B

NEW QUESTION 105

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 serviceaccounts.
- D. Amazon SageMaker notebook instances are based on AWS ECS instances running within AWS service accounts.

Answer: C

NEW QUESTION 107

A company provisions Amazon SageMaker notebook instances for its data science team and creates Amazon VPC interface endpoints to ensure communication between the VPC and the notebook instances. All connections to the Amazon SageMaker API are contained entirely and securely using the AWS network.

However, the data science team realizes that individuals outside the VPC can still connect to the notebook instances across the internet.

Which set of actions should the data science team take to fix the issue?

- A. Modify the notebook instances' security group to allow traffic only from the CIDR ranges of the VP
- B. Apply this security group to all of the notebook instances' VPC interfaces.
- C. Create an IAM policy that allows the sagemaker:CreatePresignedNotebookInstanceUrl and sagemaker:DescribeNotebookInstance actions from only the VPC endpoint
- D. Apply this policy to all IAM users, groups, and roles used to access the notebook instances.
- E. Add a NAT gateway to the VP
- F. Convert all of the subnets where the Amazon SageMaker notebook instances are hosted to private subnet
- G. Stop and start all of the notebook instances to reassign only private IP addresses.
- H. Change the network ACL of the subnet the notebook is hosted in to restrict access to anyone outside the VPC.

Answer: B

NEW QUESTION 109

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

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

Answer: D

NEW QUESTION 114

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 116

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

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

Answer: A

NEW QUESTION 117

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

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

Answer: B

NEW QUESTION 120

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 125

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

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

Answer: AD

NEW QUESTION 126

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

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

Answer: A

NEW QUESTION 127

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-realtime 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

Answer: A

NEW QUESTION 131

A company that manufactures mobile devices wants to determine and calibrate the appropriate sales price for its devices. The company is collecting the relevant data and is determining data features that it can use to train machine learning (ML) models. There are more than 1,000 features, and the company wants to determine the primary features that contribute to the sales price.

Which techniques should the company use for feature selection? (Choose three.)

- A. Data scaling with standardization and normalization
- B. Correlation plot with heat maps
- C. Data binning
- D. Univariate selection
- E. Feature importance with a tree-based classifier
- F. Data augmentation

Answer: CDF

NEW QUESTION 134

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

Answer: B

NEW QUESTION 137

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 138

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 139

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

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

Answer: CEF

NEW QUESTION 143

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

Answer: A

NEW QUESTION 146

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 147

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 151

A retail company uses a machine learning (ML) model for daily sales forecasting. The company's brand manager reports that the model has provided inaccurate results for the past 3 weeks.

At the end of each day, an AWS Glue job consolidates the input data that is used for the forecasting with the actual daily sales data and the predictions of the model. The AWS Glue job stores the data in Amazon S3. The company's ML team is using an Amazon SageMaker Studio notebook to gain an understanding about the source of the model's inaccuracies.

What should the ML team do on the SageMaker Studio notebook to visualize the model's degradation MOST accurately?

- A. Create a histogram of the daily sales over the last 3 weeks.
- B. In addition, create a histogram of the daily sales from before that period.
- C. Create a histogram of the model errors over the last 3 weeks.
- D. In addition, create a histogram of the model errors from before that period.
- E. Create a line chart with the weekly mean absolute error (MAE) of the model.
- F. Create a scatter plot of daily sales versus model error for the last 3 weeks.
- G. In addition, create a scatter plot of daily sales versus model error from before that period.

Answer: C

NEW QUESTION 156

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

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

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

Answer: C

NEW QUESTION 157

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

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

Answer: B

NEW QUESTION 160

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 165

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

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

Answer: D

NEW QUESTION 170

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 171

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

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

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

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

- A. Change the XGBoost eval_metric parameter to optimize based on rmse instead of error.
- B. Increase the XGBoost scale_pos_weight parameter to adjust the balance of positive and negative weights.
- C. Increase the XGBoost max_depth parameter because the model is currently underfitting the data.
- D. Change the XGBoost 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 175

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 177

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 181

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 183

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? (Select 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.

Answer: BD

NEW QUESTION 187

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 189

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 192

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

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

Answer: B

NEW QUESTION 196

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

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

Answer: A

NEW QUESTION 198

A 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 199

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

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

Answer: B

NEW QUESTION 200

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 201

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

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

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

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

Answer: B

Explanation:

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

NEW QUESTION 206

For the given confusion matrix, what is the recall and precision of the model?

		Actual	
		Yes	No
Predicted	Yes	12	3
	No	1	9

- A. Recall = 0.92 Precision = 0.84
- B. Recall = 0.84 Precision = 0.8
- C. Recall = 0.92 Precision = 0.8
- D. Recall = 0.8 Precision = 0.92

Answer: C

NEW QUESTION 210

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

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

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

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

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

Answer: CD

NEW QUESTION 212

A Machine Learning Specialist is working with multiple data sources containing billions of records that need to be joined. What feature engineering and model

development approach should the Specialist take with a dataset this large?

- A. Use an Amazon SageMaker notebook for both feature engineering and model development
- B. Use an Amazon SageMaker notebook for feature engineering and Amazon ML for model development
- C. Use Amazon EMR for feature engineering and Amazon SageMaker SDK for model development
- D. Use Amazon ML for both feature engineering and model development.

Answer: B

NEW QUESTION 213

.....

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