



Amazon

Exam Questions AWS-Certified-Machine-Learning-Specialty

AWS Certified Machine Learning - Specialty

About ExamBible

[Your Partner of IT Exam](#)

Found in 1998

ExamBible is a company specialized on providing high quality IT exam practice study materials, especially Cisco CCNA, CCDA, CCNP, CCIE, Checkpoint CCSE, CompTIA A+, Network+ certification practice exams and so on. We guarantee that the candidates will not only pass any IT exam at the first attempt but also get profound understanding about the certificates they have got. There are so many alike companies in this industry, however, ExamBible has its unique advantages that other companies could not achieve.

Our Advances

* 99.9% Uptime

All examinations will be up to date.

* 24/7 Quality Support

We will provide service round the clock.

* 100% Pass Rate

Our guarantee that you will pass the exam.

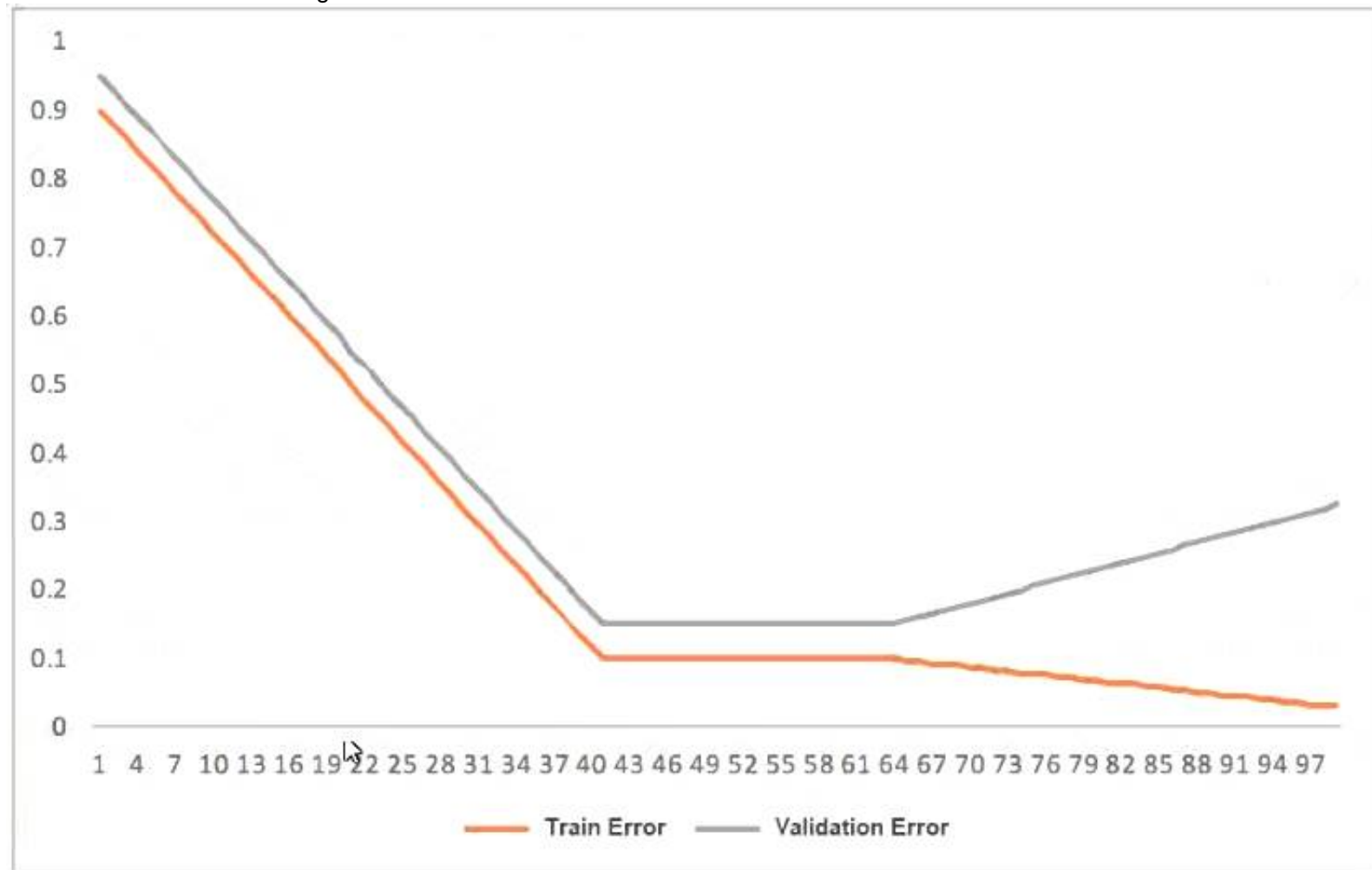
* Unique Gurantee

If you do not pass the exam at the first time, we will not only arrange FULL REFUND for you, but also provide you another exam of your claim, ABSOLUTELY FREE!

NEW QUESTION 1

This graph shows the training and validation loss against the epochs for a neural network. The network being trained is as follows:

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



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

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

Answer: D

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

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 DB 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 DB 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 4

A retail company wants to update its customer support system. The company wants to implement automatic routing of customer claims to different queues to prioritize the claims by category.

Currently, an operator manually performs the category assignment and routing. After the operator classifies and routes the claim, the company stores the claim's record in a central database. The claim's record includes the claim's category.

The company has no data science team or experience in the field of machine learning (ML). The company's small development team needs a solution that requires no ML expertise.

Which solution meets these requirements?

- A. Export the database to a .csv file with two columns: claim_label and claim_text
- B. Use the Amazon SageMaker Object2Vec algorithm and the .csv file to train a model
- C. Use SageMaker to deploy the model to an inference endpoint
- D. Develop a service in the application to use the inference endpoint to process incoming claims, predict the labels, and route the claims to the appropriate queue.
- E. Export the database to a .csv file with one column: claim_text
- F. Use the Amazon SageMaker Latent Dirichlet Allocation (LDA) algorithm and the .csv file to train a model
- G. Use the LDA algorithm to detect labels automatically
- H. Use SageMaker to deploy the model to an inference endpoint
- I. Develop a service in the application to use the inference endpoint to process incoming claims, predict the labels, and route the claims to the appropriate queue.
- J. Use Amazon Textract to process the database and automatically detect two columns: claim_label and claim_text
- K. Use Amazon Comprehend custom classification and the extracted information to train the custom classifier
- L. Develop a service in the application to use the Amazon Comprehend API to process incoming claims, predict the labels, and route the claims to the appropriate queue.
- M. Export the database to a .csv file with two columns: claim_label and claim_text
- N. Use Amazon Comprehend custom classification and the .csv file to train the custom classifier
- O. Develop a service in the application to use the Amazon Comprehend API to process incoming claims, predict the labels, and route the claims to the appropriate queue.

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 data scientist has a dataset of machine part images stored in Amazon Elastic File System (Amazon EFS). The data scientist needs to use Amazon SageMaker to create and train an image classification machine learning model based on this dataset. Because of budget and time constraints, management wants the data scientist to create and train a model with the least number of steps and integration work required.

How should the data scientist meet these requirements?

- A. Mount the EFS file system to a SageMaker notebook and run a script that copies the data to an Amazon FSx for Lustre file system
- B. Run the SageMaker training job with the FSx for Lustre file system as the data source.
- C. Launch a transient Amazon EMR cluster
- D. Configure steps to mount the EFS file system and copy the data to an Amazon S3 bucket by using S3DistC
- E. Run the SageMaker training job with Amazon S3 as the data source.
- F. Mount the EFS file system to an Amazon EC2 instance and use the AWS CLI to copy the data to an Amazon S3 bucket
- G. Run the SageMaker training job with Amazon S3 as the data source.
- H. Run a SageMaker training job with an EFS file system as the data source.

Answer: A

NEW QUESTION 7

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

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

Answer: C

NEW QUESTION 8

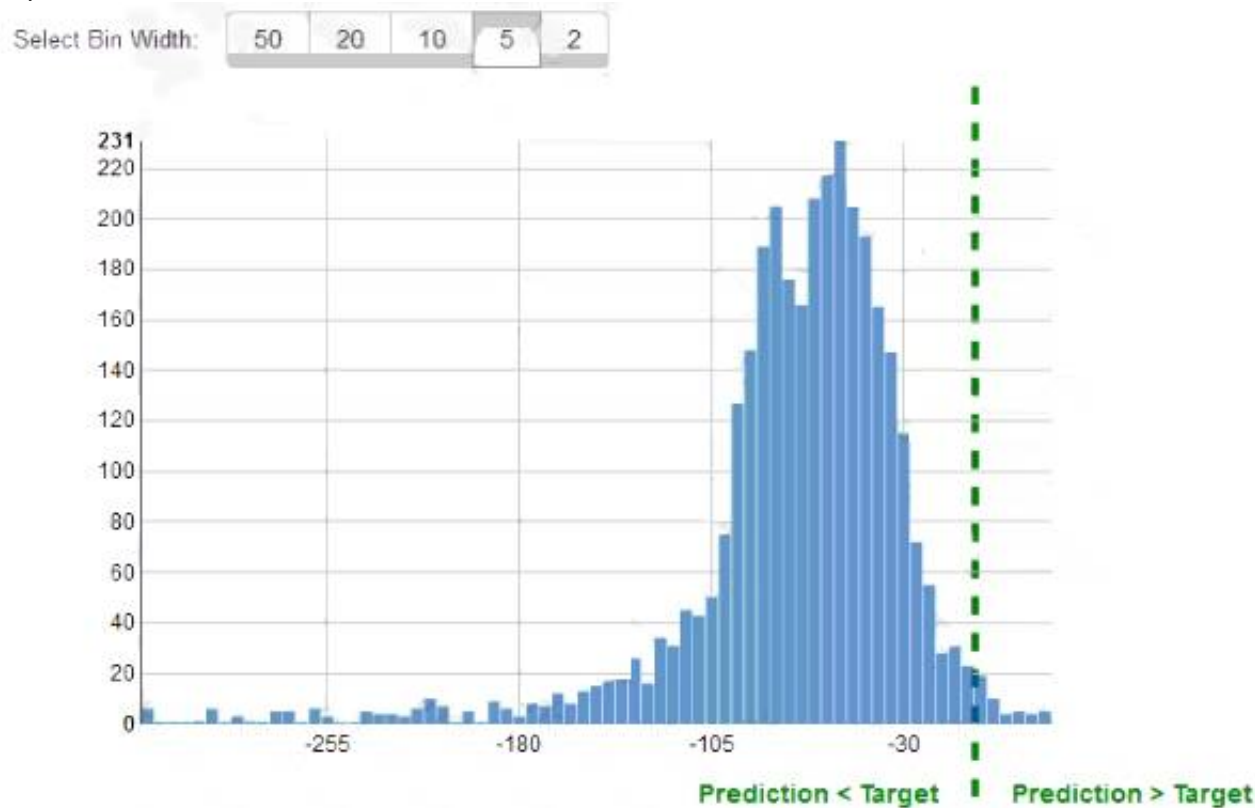
An office security agency conducted a successful pilot using 100 cameras installed at key locations within the main office. Images from the cameras were uploaded to Amazon S3 and tagged using Amazon Rekognition, and the results were stored in Amazon ES. The agency is now looking to expand the pilot into 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 non-employees 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 9

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



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

Answer: D

NEW QUESTION 10

A Machine Learning 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 company will use Amazon SageMaker to train and host a machine learning (ML) model for a marketing campaign. The majority of data is sensitive customer data. The data must be encrypted at rest. The company wants AWS to maintain the root of trust for the master keys and wants encryption key usage to be logged. Which implementation will meet these requirements?

- A. Use encryption keys that are stored in AWS Cloud HSM to encrypt the ML data volumes, and to encrypt the model artifacts and data in Amazon S3.
- B. Use SageMaker built-in transient keys to encrypt the ML data volume
- C. Enable default encryption for new Amazon Elastic Block Store (Amazon EBS) volumes.
- D. Use customer managed keys in AWS Key Management Service (AWS KMS) to encrypt the ML data volumes, and to encrypt the model artifacts and data in Amazon S3.
- E. Use AWS Security Token Service (AWS STS) to create temporary tokens to encrypt the ML storage volumes, and to encrypt the model artifacts and data in Amazon S3.

Answer: C

NEW QUESTION 15

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 20

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 23

A bank wants to launch a low-rate credit promotion. The bank is located in a town that recently experienced economic hardship. Only some of the bank's customers were affected by the crisis, so the bank's credit team must identify which customers to target with the promotion. However, the credit team wants to make sure that loyal customers' full credit history is considered when the decision is made. The bank's data science team developed a model that classifies account transactions and understands credit eligibility. The data science team used the XGBoost algorithm to train the model. The team used 7 years of bank transaction historical data for training and hyperparameter tuning over the course of several days. The accuracy of the model is sufficient, but the credit team is struggling to explain accurately why the model denies credit to some customers. The credit team has almost no skill in data science. What should the data science team do to address this issue in the MOST operationally efficient manner?

- A. Use Amazon SageMaker Studio to rebuild the model
- B. Create a notebook that uses the XGBoost training container to perform model training
- C. Deploy the model at an endpoint
- D. Enable Amazon SageMaker Model Monitor to store inference
- E. Use the inferences to create Shapley values that help explain model behavior
- F. Create a chart that shows features and SHapley Additive explanation (SHAP) values to explain to the credit team how the features affect the model outcomes.
- G. Use Amazon SageMaker Studio to rebuild the model
- H. Create a notebook that uses the XGBoost training container to perform model training
- I. Activate Amazon SageMaker Debugger, and configure it to calculate and collect Shapley values
- J. Create a chart that shows features and SHapley Additive explanation (SHAP) values to explain to the credit team how the features affect the model outcomes.
- K. Create an Amazon SageMaker notebook instance
- L. Use the notebook instance and the XGBoost library to locally retrain the model
- M. Use the `plot_importance()` method in the Python XGBoost interface to create a feature importance chart
- N. Use that chart to explain to the credit team how the features affect the model outcomes.

- O. Use Amazon SageMaker Studio to rebuild the 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 27

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

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

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

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

What can the Specialist from the experiments?

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

Answer: C

NEW QUESTION 31

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

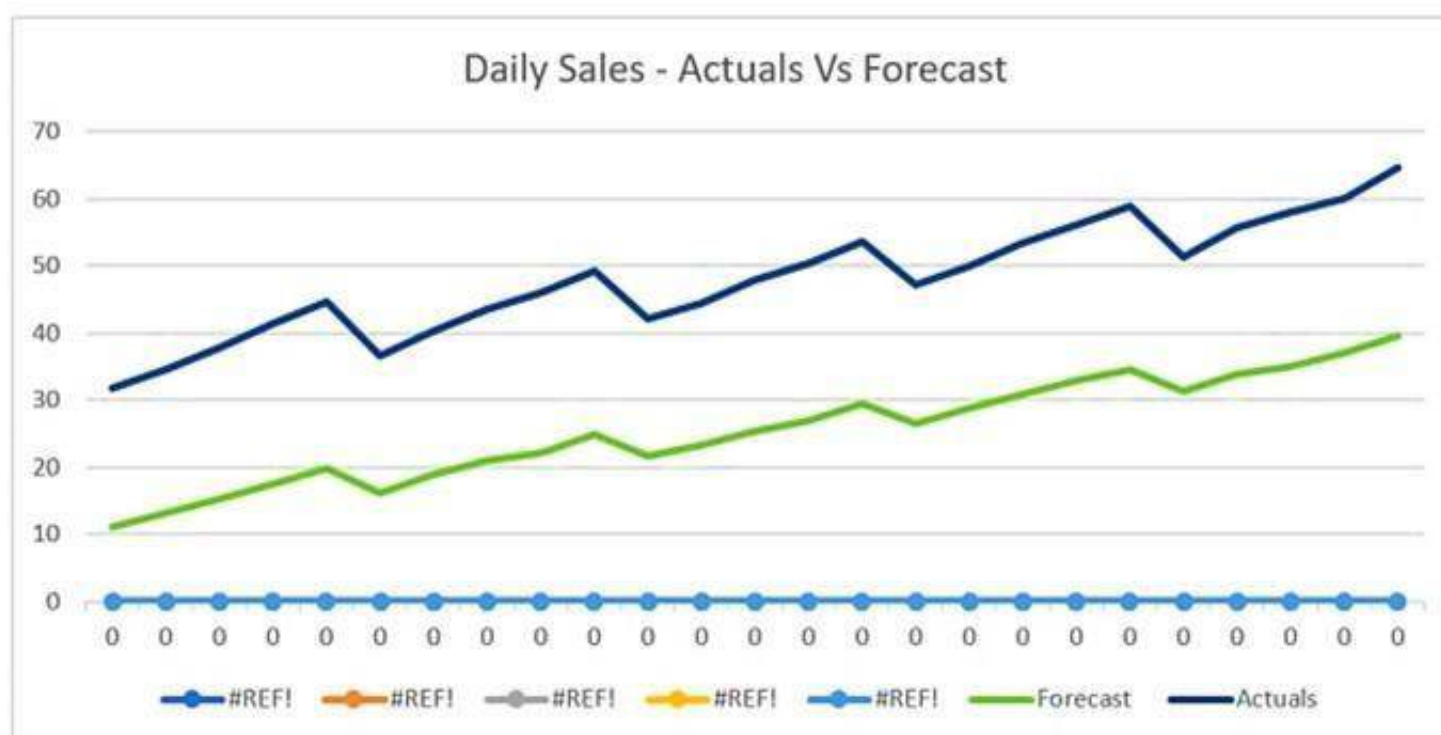
- * Must be accessible from a VPC only.
- * Must not traverse the public internet. How can these requirements be satisfied?

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

Answer: B

NEW QUESTION 32

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



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

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

D. The model does not predict the trend or the seasonality well.

Answer: D

NEW QUESTION 35

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

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

Answer: A

NEW QUESTION 38

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

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

Answer: BCE

NEW QUESTION 41

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 42

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 44

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 46

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 50

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 54

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 55

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: D

NEW QUESTION 59

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 63

A data engineer at a bank is evaluating a new tabular dataset that includes customer data. The data engineer will use the customer data to create a new model to predict customer behavior. After creating a correlation matrix for the variables, the data engineer notices that many of the 100 features are highly correlated with each other.

Which steps should the data engineer take to address this issue? (Choose two.)

- A. Use a linear-based algorithm to train the model.
- B. Apply principal component analysis (PCA).
- C. Remove a portion of highly correlated features from the dataset.
- D. Apply min-max feature scaling to the dataset.
- E. Apply one-hot encoding category-based variables.

Answer: BD

NEW QUESTION 68

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 73

A data scientist is training a text classification model by using the Amazon SageMaker built-in BlazingText algorithm. There are 5 classes in the dataset, with 300 samples for category A, 292 samples for category B, 240 samples for category C, 258 samples for category D, and 310 samples for category E.

The data scientist shuffles the data and splits off 10% for testing. After training the model, the data scientist generates confusion matrices for the training and test sets.

Training data confusion matrix

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

Test data confusion matrix

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

What could the data scientist conclude from these results?

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

Answer: B

NEW QUESTION 76

A data scientist has been running an Amazon SageMaker notebook instance for a few weeks. During this time, a new version of Jupyter Notebook was released along with additional software updates. The security team mandates that all running SageMaker notebook instances use the latest security and software updates provided by SageMaker.

How can the data scientist meet this requirements?

- A. Call the CreateNotebookInstanceLifecycleConfig API operation
- B. Create a new SageMaker notebook instance and mount the Amazon Elastic Block Store (Amazon EBS) volume from the original instance
- C. Stop and then restart the SageMaker notebook instance
- D. Call the UpdateNotebookInstanceLifecycleConfig API operation

Answer: C

NEW QUESTION 77

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 VP
- 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 79

A retail chain has been ingesting purchasing records from its network of 20,000 stores to Amazon S3 using Amazon Kinesis Data Firehose. To support training an improved machine learning model, training records will require new but simple transformations, and some attributes will be combined. The model needs to be retrained daily.

Given the large number of stores and the legacy data ingestion, which change will require the LEAST amount of development effort?

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

Answer: D

NEW QUESTION 84

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 86

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 89

While working on a neural network project, a Machine Learning Specialist discovers that some features in the data have very high magnitude resulting in this data being weighted more in the cost function. What should the Specialist do to ensure better convergence during backpropagation?

- A. Dimensionality reduction
- B. Data normalization
- C. Model regularization
- D. Data augmentation for the minority class

Answer: D

NEW QUESTION 92

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

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

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

Answer: D

NEW QUESTION 94

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 97

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 101

A data scientist has developed a machine learning translation model for English to Japanese by using Amazon SageMaker's built-in seq2seq algorithm with 500,000 aligned sentence pairs. While testing with sample sentences, the data scientist finds that the translation quality is reasonable for an example as short as five words. However, the quality becomes unacceptable if the sentence is 100 words long.

Which action will resolve the problem?

- A. Change preprocessing to use n-grams.
- B. Add more nodes to the recurrent neural network (RNN) than the largest sentence's word count.
- C. Adjust hyperparameters related to the attention mechanism.
- D. Choose a different weight initialization type.

Answer: C

Explanation:

<https://docs.aws.amazon.com/sagemaker/latest/dg/seq-2-seq-howitworks.html>

NEW QUESTION 106

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 110

A data scientist must build a custom recommendation model in Amazon SageMaker for an online retail company. Due to the nature of the company's products, customers buy only 4-5 products every 5-10 years. So, the company relies on a steady stream of new customers. When a new customer signs up, the company collects data on the customer's preferences. Below is a sample of the data available to the data scientist.

timestamp	user_id	product_id	preference_1	...	preference_10
2020-03-04	90	25	0	...	0.374
2020-03-04	90	61	0	...	0.374
2020-02-21	203	56	1	...	0.098

How should the data scientist split the dataset into a training and test set for this use case?

- A. Shuffle all interaction dat
- B. Split off the last 10% of the interaction data for the test set.
- C. Identify the most recent 10% of interactions for each use
- D. Split off these interactions for the test set.
- E. Identify the 10% of users with the least interaction dat
- F. Split off all interaction data from these users for the test set.
- G. Randomly select 10% of the user
- H. Split off all interaction data from these users for the test set.

Answer: B

Explanation:

<https://aws.amazon.com/blogs/machine-learning/building-a-customized-recommender-system-in-amazon-sagem>

NEW QUESTION 115

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

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

Answer: D

NEW QUESTION 118

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 119

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 locall
- B. Confirm that thetraining code is executing and the model parameters seem reasonabl
- 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 theinstanc
- E. Train on a small amount of the data to verify the training code and hyperparameter
- F. Go back toAmazon 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 SageMake
- 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 locall
- J. Confirm that the training code is executing and the model parameters seem reasonabl
- 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 121

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 retrains 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 123

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 124

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 129

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: CDE

NEW QUESTION 132

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

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

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

Answer: BE

NEW QUESTION 136

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 137

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

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

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

Answer: B

NEW QUESTION 138

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 139

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 141

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 145

A Machine Learning Specialist is using Amazon SageMaker to host a model for a highly available customer-facing application.

The Specialist has trained a new version of the model, validated it with historical data, and now wants to deploy it to production To limit any risk of a negative customer experience, the Specialist wants to be able to monitor the model and roll it back, if needed
What is the SIMPLEST approach with the LEAST risk to deploy the model and roll it back, if needed?

- A. Create a SageMaker endpoint and configuration for the new model versio
- B. Redirect production traffic to the new endpoint by updating the client configuratio
- C. Revert traffic to the last version if the model does not perform as expected.
- D. Create a SageMaker endpoint and configuration for the new model versio
- E. Redirect production traffic to the new endpoint by using a load balancer Revert traffic to the last version if the model does not perform as expected.
- F. Update the existing SageMaker endpoint to use a new configuration that is weighted to send 5% of the traffic to the new varian
- G. Revert traffic to the last version by resetting the weights if the model does not perform as expected.
- H. Update the existing SageMaker endpoint to use a new configuration that is weighted to send 100% of the traffic to the new variant Revert traffic to the last version by resetting the weights if the model does not perform as expected.

Answer: A

NEW QUESTION 148

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

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

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

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

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

Answer: B

NEW QUESTION 155

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 week
- 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 week
- 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 week
- G. In addition, create a scatter plot of daily sales versus model error from before that period.

Answer: C

NEW QUESTION 160

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 165

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

Which approach should the Specialist use to continue working?

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

Answer: D

NEW QUESTION 169

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 170

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 173

A company's Machine Learning Specialist needs to improve the training speed of a time-series forecasting model using TensorFlow. The training is currently implemented on a single-GPU machine and takes approximately 23 hours to complete. The training needs to be run daily.

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

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

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

Answer: B

NEW QUESTION 174

A data scientist wants to use Amazon Forecast to build a forecasting model for inventory demand for a retail company. The company has provided a dataset of historic inventory demand for its products as a .csv file stored in an Amazon S3 bucket. The table below shows a sample of the dataset.

timestamp	item_id	demand	category	lead_time
2019-12-14	uni_000736	120	hardware	90
2020-01-31	uni_003429	98	hardware	30
2020-03-04	uni_000211	234	accessories	10

How should the data scientist transform the data?

- A. Use ETL jobs in AWS Glue to separate the dataset into a target time series dataset and an item metadata dataset
- B. Upload both datasets as .csv files to Amazon S3.
- C. Use a Jupyter notebook in Amazon SageMaker to separate the dataset into a related time series dataset and an item metadata dataset
- D. Upload both datasets as tables in Amazon Aurora.
- E. Use AWS Batch jobs to separate the dataset into a target time series dataset, a related time series dataset, and an item metadata dataset
- F. Upload them directly to Forecast from a local machine.
- G. Use a Jupyter notebook in Amazon SageMaker to transform the data into the optimized protobuf recordIO format
- H. Upload the dataset in this format to Amazon S3.

Answer: A

Explanation:

<https://docs.aws.amazon.com/forecast/latest/dg/dataset-import-guidelines-troubleshooting.html>

NEW QUESTION 175

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

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

Which solution will meet these requirements?

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

Answer: A

NEW QUESTION 180

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 185

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 186

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 191

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

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

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

Answer: C

NEW QUESTION 195

A company is building a predictive maintenance model based on machine learning (ML). The data is stored in a fully private Amazon S3 bucket that is encrypted at rest with AWS Key Management Service (AWS KMS) CMKs. An ML specialist must run data preprocessing by using an Amazon SageMaker Processing job that is triggered from code in an Amazon SageMaker notebook. The job should read data from Amazon S3, process it, and upload it back to the same S3 bucket. The preprocessing code is stored in a container image in Amazon Elastic Container Registry (Amazon ECR). The ML specialist needs to grant permissions to ensure a smooth data preprocessing workflow.

Which set of actions should the ML specialist take to meet these requirements?

- A. Create an IAM role that has permissions to create Amazon SageMaker Processing jobs, S3 read and write access to the relevant S3 bucket, and appropriate KMS and ECR permission
- B. Attach the role to the SageMaker notebook instance
- C. Create an Amazon SageMaker Processing job from the notebook.
- D. Create an IAM role that has permissions to create Amazon SageMaker Processing job
- E. Attach the role to the SageMaker notebook instance
- F. Create an Amazon SageMaker Processing job with an IAM role that has read and write permissions to the relevant S3 bucket, and appropriate KMS and ECR permissions.
- G. Create an IAM role that has permissions to create Amazon SageMaker Processing jobs and to access Amazon ECR
- H. Attach the role to the SageMaker notebook instance
- I. Set up both an S3 endpoint and a KMS endpoint in the default VPC
- J. Create Amazon SageMaker Processing jobs from the notebook.
- K. Create an IAM role that has permissions to create Amazon SageMaker Processing job
- L. Attach the role to the SageMaker notebook instance
- M. Set up an S3 endpoint in the default VPC
- N. Create Amazon SageMaker Processing jobs with the access key and secret key of the IAM user with appropriate KMS and ECR permissions.

Answer: D

NEW QUESTION 200

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 204

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 205

A data scientist is using the Amazon SageMaker Neural Topic Model (NTM) algorithm to build a model that recommends tags from blog posts. The raw blog post data is stored in an Amazon S3 bucket in JSON format. During model evaluation, the data scientist discovered that the model recommends certain stopwords such as "a," "an," and "the" as tags to certain blog posts, along with a few rare words that are present only in certain blog entries. After a few iterations of tag review with the content team, the data scientist notices that the rare words are unusual but feasible. The data scientist also must ensure that the tag recommendations of the generated model do not include the stopwords.

What should the data scientist do to meet these requirements?

- A. Use the Amazon Comprehend entity recognition API operation
- B. Remove the detected words from the blog post data
- C. Replace the blog post data source in the S3 bucket.
- D. Run the SageMaker built-in principal component analysis (PCA) algorithm with the blog post data from the S3 bucket as the data source
- E. Replace the blog post data in the S3 bucket with the results of the training job.
- F. Use the SageMaker built-in Object Detection algorithm instead of the NMF algorithm for the training job to process the blog post data.
- G. Remove the stopwords from the blog post data by using the Count Vectorizer function in the scikit-learn library
- H. Replace the blog post data in the S3 bucket with the results of the vectorizer.

Answer: D

NEW QUESTION 208

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 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

.....

Relate Links

100% Pass Your AWS-Certified-Machine-Learning-Specialty Exam with Examible Prep Materials

<https://www.exambible.com/AWS-Certified-Machine-Learning-Specialty-exam/>

Contact us

We are proud of our high-quality customer service, which serves you around the clock 24/7.

Viste - <https://www.exambible.com/>