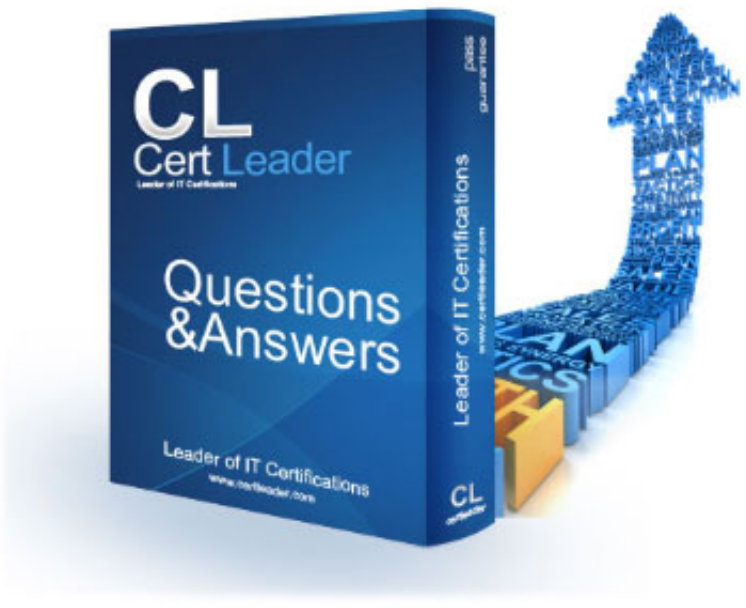
AWS-Certified-Machine-Learning-Specialty Dumps

AWS Certified Machine Learning - Specialty

https://www.certleader.com/AWS-Certified-Machine-Learning-Specialty-dumps.html



NEW QUESTION 1
A Data Scientist needs to create a serverless ingestion and analytics solution for high-velocity, real-time streaming data.
The ingestion process must buffer and convert incoming records from JSON to a query-optimized, columnar format without data loss. The output datastore must be highly available, and Analysts must be able to run SQL queries against the data and connect to existing business intelligence dashboards.
Which solution should the Data Scientist build to satisfy the requirements?

A. Create a schema in the AWS Glue Data Catalog of the incoming data forma
B. Use an Amazon Kinesis Data Firehose delivery stream to stream the data and transform the data to Apache Parquet or ORC format using the AWS Glue Data Catalog before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to Bl tools using the Athena Java Database Connectivity (JDBC) connector.
C. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and writes the data to a processed data location in Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to Bl tools using the Athena Java Database Connectivity (JDBC) connector.
D. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and inserts it into an Amazon RDS PostgreSQL databas
E. Have the Analysts query and run dashboards from the RDS database.
F. Use Amazon Kinesis Data Analytics to ingest the streaming data and perform real-time SQL queries to convert the records to Apache Parquet before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena and connect to Bl tools using the Athena Java Database Connectivity (JDBC) connector.

Answer: A

NEW QUESTION 2
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 3
An agency collects census information within a country to determine healthcare and social program needs by province and city. The census form collects responses for approximately 500 questions from each citizen
Which combination of algorithms would provide the appropriate insights? (Select TWO)

A. The factorization machines (FM) algorithm
B. The Latent Dirichlet Allocation (LDA) algorithm
C. The principal component analysis (PCA) algorithm
D. The k-means algorithm
E. The Random Cut Forest (RCF) algorithm

Answer: CD

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

NEW QUESTION 4
A web-based company wants to improve its conversion rate on its landing page Using a large historical dataset of customer visits, the company has repeatedly trained a multi-class deep learning network algorithm on Amazon SageMaker However there is an overfitting problem training data shows 90% accuracy in predictions, while test data shows 70% accuracy only
The company needs to boost the generalization of its model before deploying it into production to maximize conversions of visits to purchases
Which action is recommended to provide the HIGHEST accuracy model for the company's test and validation data?

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

Answer: C

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

NEW QUESTION 5
A machine learning (ML) specialist is using Amazon SageMaker hyperparameter optimization (HPO) to improve a model’s accuracy. The learning rate parameter is specified in the following HPO configuration:

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

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

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

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

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

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

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

Select the most accurate hyperparameter configuration form this training job.
D. Run three different HPO jobs that use different learning rates form the following intervals for MinValue and MaxValu
E. Divide the number of training jobs for each HPO job by three:[0.01, 0.1][0.001, 0.01][0.0001, 0.001]Select the most accurate hyperparameter configuration form these three HPO jobs.

Answer: C

NEW QUESTION 6
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 7
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	PREDICTED CHURN 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 8
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 9
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 10
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 10
A Data Scientist needs to migrate an existing on-premises ETL process to the cloud The current process runs at regular time intervals and uses PySpark to combine and format multiple large data sources into a single consolidated output for downstream processing
The Data Scientist has been given the following requirements for the cloud solution
* Combine multiple data sources
* Reuse existing PySpark logic
* Run the solution on the existing schedule
* Minimize the number of servers that will need to be managed
Which architecture should the Data Scientist use to build this solution?

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

Answer: A

NEW QUESTION 12
A Machine Learning Specialist is configuring automatic model tuning in Amazon SageMaker
When using the hyperparameter optimization feature, which of the following guidelines should be followed to improve optimization?
Choose the maximum number of hyperparameters supported by

A. Amazon SageMaker to search the largest number of combinations possible
B. Specify a very large hyperparameter range to allow Amazon SageMaker to cover every possible value.
C. Use log-scaled hyperparameters to allow the hyperparameter space to be searched as quickly as possible
D. Execute only one hyperparameter tuning job at a time and improve tuning through successive rounds of experiments

Answer: C

NEW QUESTION 13
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 15
A Machine Learning Specialist is building a convolutional neural network (CNN) that will classify 10 types of animals. The Specialist has built a series of layers in a neural network that will take an input image of an animal, pass it through a series of convolutional and pooling layers, and then finally pass it through a dense and fully connected layer with 10 nodes The Specialist would like to get an output from the neural network that is a probability distribution of how likely it is that the input image belongs to each of the 10 classes
Which function will produce the desired output?

A. Dropout
B. Smooth L1 loss
C. Softmax
D. Rectified linear units (ReLU)

Answer: C

NEW QUESTION 16
A machine learning (ML) specialist must develop a classification model for a financial services company. A domain expert provides the dataset, which is tabular with 10,000 rows and 1,020 features. During exploratory data analysis, the specialist finds no missing values and a small percentage of duplicate rows. There are correlation scores of > 0.9 for 200 feature pairs. The mean value of each feature is similar to its 50th percentile.
Which feature engineering strategy should the ML specialist use with Amazon SageMaker?

A. Apply dimensionality reduction by using the principal component analysis (PCA) algorithm.
B. Drop the features with low correlation scores by using a Jupyter notebook.
C. Apply anomaly detection by using the Random Cut Forest (RCF) algorithm.
D. Concatenate the features with high correlation scores by using a Jupyter notebook.

Answer: C

NEW QUESTION 18
A Machine Learning Specialist is developing a daily ETL workflow containing multiple ETL jobs The workflow consists of the following processes
* Start the workflow as soon as data is uploaded to Amazon S3
* When all the datasets are available in Amazon S3, start an ETL job to join the uploaded datasets with multiple terabyte-sized datasets already stored in Amazon S3
* Store the results of joining datasets in Amazon S3
* If one of the jobs fails, send a notification to the Administrator Which configuration will meet these requirements?

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

Answer: A

NEW QUESTION 19
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 20
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 quenes
D. Use AWS Lambda to transform the data and Amazon Kinesis Data Analytics to run queries

Answer: D

NEW QUESTION 22
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 26
A company uses a long short-term memory (LSTM) model to evaluate the risk factors of a particular energy sector. The model reviews multi-page text documents to analyze each sentence of the text and categorize it as either a potential risk or no risk. The model is not performing well, even though the Data Scientist has experimented with many different network structures and tuned the corresponding hyperparameters.
Which approach will provide the MAXIMUM performance boost?

A. Initialize the words by term frequency-inverse document frequency (TF-IDF) vectors pretrained on a large collection of news articles related to the energy sector.
B. Use gated recurrent units (GRUs) instead of LSTM and run the training process until the validation loss stops decreasing.
C. Reduce the learning rate and run the training process until the training loss stops decreasing.
D. Initialize the words by word2vec embeddings pretrained on a large collection of news articles related to the energy sector.

Answer: C

NEW QUESTION 28
A Data Scientist wants to gain real-time insights into a data stream of GZIP files. Which solution would allow the use of SQL to query the stream with the LEAST latency?

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

Answer: A

NEW QUESTION 32
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 ha
B. Stop the notebook when it is not in us
C. Run both data preprocessing and feature engineering development on that instance.
D. Keep the notebook instance type and size the sam
E. Stop the notebook when it is not in us
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 instanc
H. Stop the notebook when it is not in us
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 instanc
K. Stop the notebook when it is not in us
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 34
A library is developing an automatic book-borrowing system that uses Amazon Rekognition. Images of library members' faces are stored in an Amazon S3 bucket. When members borrow books, the Amazon Rekognition CompareFaces API operation compares real faces against the stored faces in Amazon S3.
The library needs to improve security by making sure that images are encrypted at rest. Also, when the images are used with Amazon Rekognition. they need to be encrypted in transit. The library also must ensure that the images are not used to improve Amazon Rekognition as a service.
How should a machine learning specialist architect the solution to satisfy these requirements?

A. Enable server-side encryption on the S3 bucke
B. Submit an AWS Support ticket to opt out of allowing images to be used for improving the service, and follow the process provided by AWS Support.
C. Switch to using an Amazon Rekognition collection to store the image
D. Use the IndexFaces andSearchFacesByImage API operations instead of the CompareFaces API operation.
E. Switch to using the AWS GovCloud (US) Region for Amazon S3 to store images and for Amazon Rekognition to compare face
F. Set up a VPN connection and only call the Amazon Rekognition API operations through the VPN.
G. Enable client-side encryption on the S3 bucke
H. Set up a VPN connection and only call the Amazon Rekognition API operations through the VPN.

Answer: B

NEW QUESTION 39
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 inappropriat
B. The residuals do not have constant variance.
C. Linear regression is inappropriat
D. The underlying data has outliers.
E. Linear regression is appropriat
F. The residuals have a zero mean.
G. Linear regression is appropriat
H. The residuals have constant variance.

Answer: D

NEW QUESTION 43
An aircraft engine manufacturing company is measuring 200 performance metrics in a time-series. Engineers want to detect critical manufacturing defects in near-real time during testing. All of the data needs to be stored for offline analysis.
What approach would be the MOST effective to perform near-real time defect detection?

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

Answer: B

NEW QUESTION 46
A Data Scientist needs to analyze employment data. The dataset contains approximately 10 million observations on people across 10 different features. During the preliminary analysis, the Data Scientist notices that income and age distributions are not normal. While income levels shows a right skew as expected, with fewer individuals having a higher income, the age distribution also show a right skew, with fewer older individuals participating in the workforce.
Which feature transformations can the Data Scientist apply to fix the incorrectly skewed data? (Choose two.)

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

Answer: AB

NEW QUESTION 49
A company sells thousands of products on a public website and wants to automatically identify products with potential durability problems. The company has 1.000 reviews with date, star rating, review text, review summary, and customer email fields, but many reviews are incomplete and have empty fields. Each review has already been labeled with the correct durability result.
A machine learning specialist must train a model to identify reviews expressing concerns over product durability. The first model needs to be trained and ready to review in 2 days.
What is the MOST direct approach to solve this problem within 2 days?

A. Train a custom classifier by using Amazon Comprehend.
B. Build a recurrent neural network (RNN) in Amazon SageMaker by using Gluon and Apache MXNet.
C. Train a built-in BlazingText model using Word2Vec mode in Amazon SageMaker.
D. Use a built-in seq2seq model in Amazon SageMaker.

Answer: B

NEW QUESTION 51
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 outpu
B. Apply principal component analysis (PCA) with five feature dimensions to predict a segment.
C. The XGBoost model provides a true/false binary outpu
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 algorith
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 algorith
H. Train a k-means model with K = 5 on the same dataset to predict a segment.

Answer: C

NEW QUESTION 53
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 57
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 58
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 61
A financial services company is building a robust serverless data lake on Amazon S3. The data lake should be flexible and meet the following requirements:
* Support querying old and new data on Amazon S3 through Amazon Athena and Amazon Redshift Spectrum.
* Support event-driven ETL pipelines.
* Provide a quick and easy way to understand metadata. Which approach meets trfese requirements?

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

Answer: A

NEW QUESTION 62
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 65
A health care company is planning to use neural networks to classify their X-ray images into normal and abnormal classes. The labeled data is divided into a training set of 1,000 images and a test set of 200 images. The initial training of a neural network model with 50 hidden layers yielded 99% accuracy on the training set, but only 55% accuracy on the test set.
What changes should the Specialist consider to solve this issue? (Choose three.)

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

Answer: ADE

NEW QUESTION 69
A Machine Learning Specialist needs to be able to ingest streaming data and store it in Apache Parquet files for exploration and analysis. Which of the following services would both ingest and store this data in the correct format?

A. AWSDMS
B. Amazon Kinesis Data Streams
C. Amazon Kinesis Data Firehose
D. Amazon Kinesis Data Analytics

Answer: C

NEW QUESTION 72
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 attemp
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 algorith
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 attemp
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 algorith
H. Schedule updates and retraining of the model using new log data nightly.

Answer: C

NEW QUESTION 77
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 1AM 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 78
A bank's Machine Learning team is developing an approach for credit card fraud detection The company has a large dataset of historical data labeled as fraudulent The goal is to build a model to take the information from new transactions and predict whether each transaction is fraudulent or not
Which built-in Amazon SageMaker machine learning algorithm should be used for modeling this problem?

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

Answer: C

NEW QUESTION 80
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 82
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 85
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 cluste
B. Create an Apache Hive external table for the DynamoDB table and S3 dat
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 tabl
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 87
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 datase
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 datase
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 datase
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 forma
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 88
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 bucke
B. Train and evaluate the model by using Amazon SageMake
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 bucke
G. Deploy the model on an Amazon SageMaker hosting services endpoint.
H. Move the training data to an Amazon S3 bucke
I. Train and evaluate the model by using Amazon SageMake
J. Optimize the model by using SageMaker Ne
K. Set up an edge device in the manufacturing facilities with AWS IoT Greengras
L. Deploy the model on the edge device.
M. Train the model on premise

N. Upload the model to an Amazon S3 bucke
O. Set up an edge device in the manufacturing facilities with AWS IoT Greengras
P. Deploy the model on the edge device.

Answer: A

NEW QUESTION 89
The chief editor for a product catalog wants the research and development team to build a machine learning system that can be used to detect whether or not individuals in a collection of images are wearing the company's retail brand. The team has a set of training data.
Which machine learning algorithm should the researchers use that BEST meets their requirements?

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

Answer: D

NEW QUESTION 92
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 (I-SNE) to visualize the large number of input variables in an easier-to-read dimension.
C. A scatter plot showing (he 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 95
A Data Scientist is training a multilayer perception (MLP) on a dataset with multiple classes. The target class of interest is unique compared to the other classes within the dataset, but it does not achieve and acceptable recall metric. The Data Scientist has already tried varying the number and size of the MLP's hidden layers, which has not significantly improved the results. A solution to improve recall must be implemented as quickly as possible.
Which techniques should be used to meet these requirements?

A. Gather more data using Amazon Mechanical Turk and then retrain
B. Train an anomaly detection model instead of an MLP
C. Train an XGBoost model instead of an MLP
D. Add class weights to the MLP's loss function and then retrain

Answer: C

NEW QUESTION 99
A Machine Learning Specialist is applying a linear least squares regression model to a dataset with 1 000 records and 50 features Prior to training, the ML Specialist notices that two features are perfectly linearly dependent
Why could this be an issue for the linear least squares regression model?

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

Answer: C

NEW QUESTION 101
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 instanc
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 instanc
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 EC
H. Attach the role to the SageMaker notebook instanc
I. Set up both an S3 endpoint and a KMS endpoint in the default VP

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 instanc
M. Set up an S3 endpoint in the default VP
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 102
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 103
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 dat
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 sourc
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 NTM 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-learnlibrar
H. Replace the blog post data in the S3 bucket with the results of the vectorizer.

Answer: D

NEW QUESTION 108
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 110
A Machine Learning Specialist is required to build a supervised image-recognition model to identify a cat. The ML Specialist performs some tests and records the following results for a neural network-based image classifier:
Total number of images available = 1,000 Test set images = 100 (constant test set)
The ML Specialist notices that, in over 75% of the misclassified images, the cats were held upside down by their owners.
Which techniques can be used by the ML Specialist to improve this specific test error?

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

Answer: A

NEW QUESTION 112
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 115
.....

Thank You for Trying Our Product

* 100% Pass or Money Back

All our products come with a 90-day Money Back Guarantee.

* One year free update

You can enjoy free update one year. 24x7 online support.

* Trusted by Millions

We currently serve more than 30,000,000 customers.

* Shop Securely

All transactions are protected by VeriSign!

100% Pass Your AWS-Certified-Machine-Learning-Specialty Exam with Our Prep Materials Via below:

https://www.certleader.com/AWS-Certified-Machine-Learning-Specialty-dumps.html