AWS-Certified-Machine-Learning-Specialty Dumps

AWS Certified Machine Learning - Specialty

https://www.certleader.com/AWS-Certified-Machine-Learning-Specialty-dumps.html
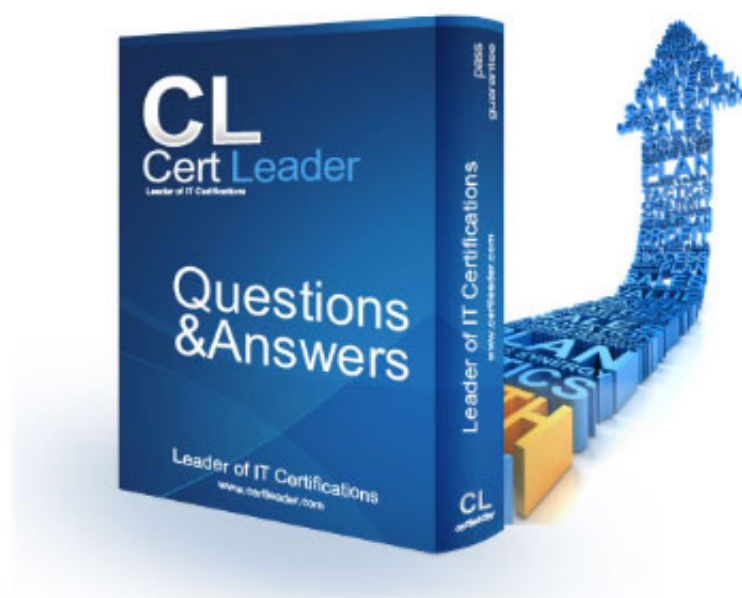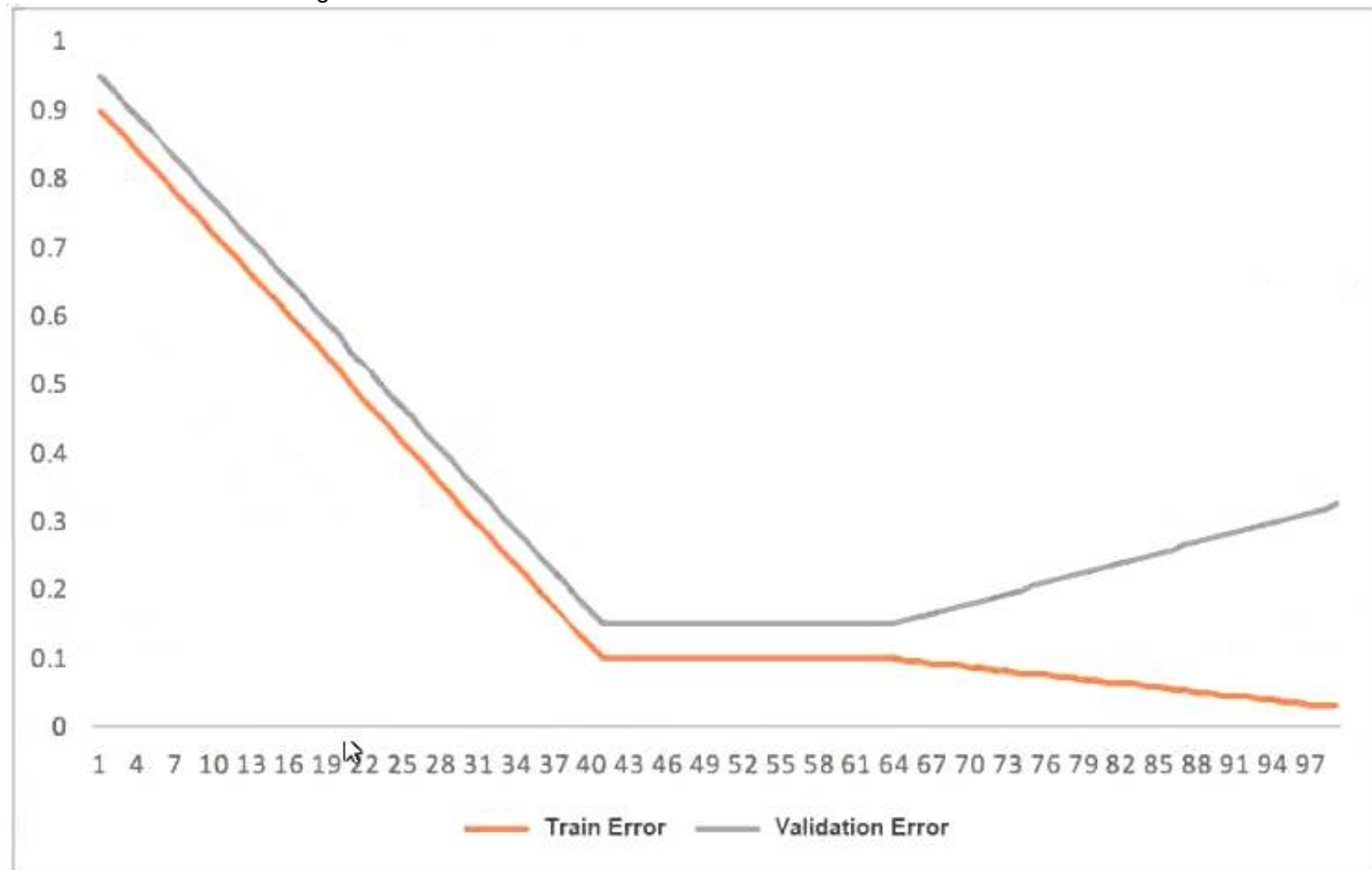
NEW QUESTION 1
This graph shows the training and validation loss against the epochs for a neural network The network being trained is as follows
• Two dense layers one output neuron
• 100 neurons in each layer
• 100 epochs
• Random initialization of weights

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

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

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

How should the Data Scientist correct this issue?

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

Answer: A

NEW QUESTION 4
A company is building a new version of a recommendation engine. Machine learning (ML) specialists need to keep adding new data from users to improve personalized recommendations. The ML specialists gather data from the users' interactions on the platform and from sources such as external websites and social media.

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

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

A. Load the data into an Amazon Redshift cluste
B. Execute the pipeline by using SQ
C. Store the results in Amazon S3.
D. Load the data into Amazon DynamoD
E. Convert the scripts to an AWS Lambda functio
F. Execute the pipeline by triggering Lambda execution
G. Store the results in Amazon S3.
H. Create an AWS Glue jo
I. Convert the scripts to PySpar
J. Execute the pipelin
K. Store the results in Amazon S3.
L. Create a set of individual AWS Lambda functions to execute each of the script
M. Build a step function by using the AWS Step Functions Data Science SD
N. Store the results in Amazon S3.

Answer: B

NEW QUESTION 5
A Machine Learning Specialist works for a credit card processing company and needs to predict which transactions may be fraudulent in near-real time. Specifically, the Specialist must train a model that returns the probability that a given transaction may be fraudulent

How should the Specialist frame this business problem'?

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

Answer: A

NEW QUESTION 6
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 instanc
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 instanc
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 connectio
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 connectio
H. Use AWS Glue to move data from Amazon EC2 to Amazon S3.

Answer: C

NEW QUESTION 7
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 8
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 9
A medical imaging company wants to train a computer vision model to detect areas of concern on patients' CT scans. The company has a large collection of unlabeled CT scans that are linked to each patient and stored in an Amazon S3 bucket. The scans must be accessible to authorized users only. A machine learning engineer needs to build a labeling pipeline.
Which set of steps should the engineer take to build the labeling pipeline with the LEAST effort?

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

Answer: C

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

NEW QUESTION 10
A Machine Learning Specialist is implementing a full Bayesian network on a dataset that describes public transit in New York City. One of the random variables is discrete, and represents the number of minutes New Yorkers wait for a bus given that the buses cycle every 10 minutes, with a mean of 3 minutes.
Which prior probability distribution should the ML Specialist use for this variable?

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

Answer: A

NEW QUESTION 10
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 15

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

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

Answer: B

NEW QUESTION 16
A bank wants to launch a low-rate credit promotion. The bank is located in a town that recently experienced economic hardship. Only some of the bank's customers were affected by the crisis, so the bank's credit team must identify which customers to target with the promotion. However, the credit team wants to make sure that loyal customers' full credit history is considered when the decision is made.
The bank's data science team developed a model that classifies account transactions and understands credit eligibility. The data science team used the XGBoost algorithm to train the model. The team used 7 years of bank transaction historical data for training and hyperparameter tuning over the course of several days.
The accuracy of the model is sufficient, but the credit team is struggling to explain accurately why the model denies credit to some customers. The credit team has almost no skill in data science.
What should the data science team do to address this issue in the MOST operationally efficient manner?

A. Use Amazon SageMaker Studio to rebuild the mode
B. Create a notebook that uses the XGBoost training container to perform model trainin
C. Deploy the model at an endpoin
D. Enable Amazon SageMaker Model Monitor to store inference
E. Use the inferences to create Shapley values that help explain model behavio
F. Create a chart that shows features and SHapley Additive explanation (SHAP) values to explain to the credit team how the features affect the model outcomes.
G. Use Amazon SageMaker Studio to rebuild the mode
H. Create a notebook that uses the XGBoost training container to perform model trainin
I. Activate Amazon SageMaker Debugger, and configure it to calculate and collect Shapley value
J. Create a chart that shows features and SHapley Additive explanation (SHAP) values to explain to the credit team how the features affect the model outcomes.
K. Create an Amazon SageMaker notebook instanc
L. Use the notebook instance and the XGBoost library to locally retrain the mode
M. Use the plot_importance() method in the Python XGBoost interface to create a feature importance char
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 endpoin
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 21
An online reseller has a large, multi-column dataset with one column missing 30% of its data A Machine Learning Specialist believes that certain columns in the dataset could be used to reconstruct the missing data.
Which reconstruction approach should the Specialist use to preserve the integrity of the dataset?

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

Answer: C

NEW QUESTION 24
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 28
A power company wants to forecast future energy consumption for its customers in residential properties and commercial business properties. Historical power consumption data for the last 10 years is available. A team of data scientists who performed the initial data analysis and feature selection will include the historical power consumption data and data such as weather, number of individuals on the property, and public holidays.
The data scientists are using Amazon Forecast to generate the forecasts.
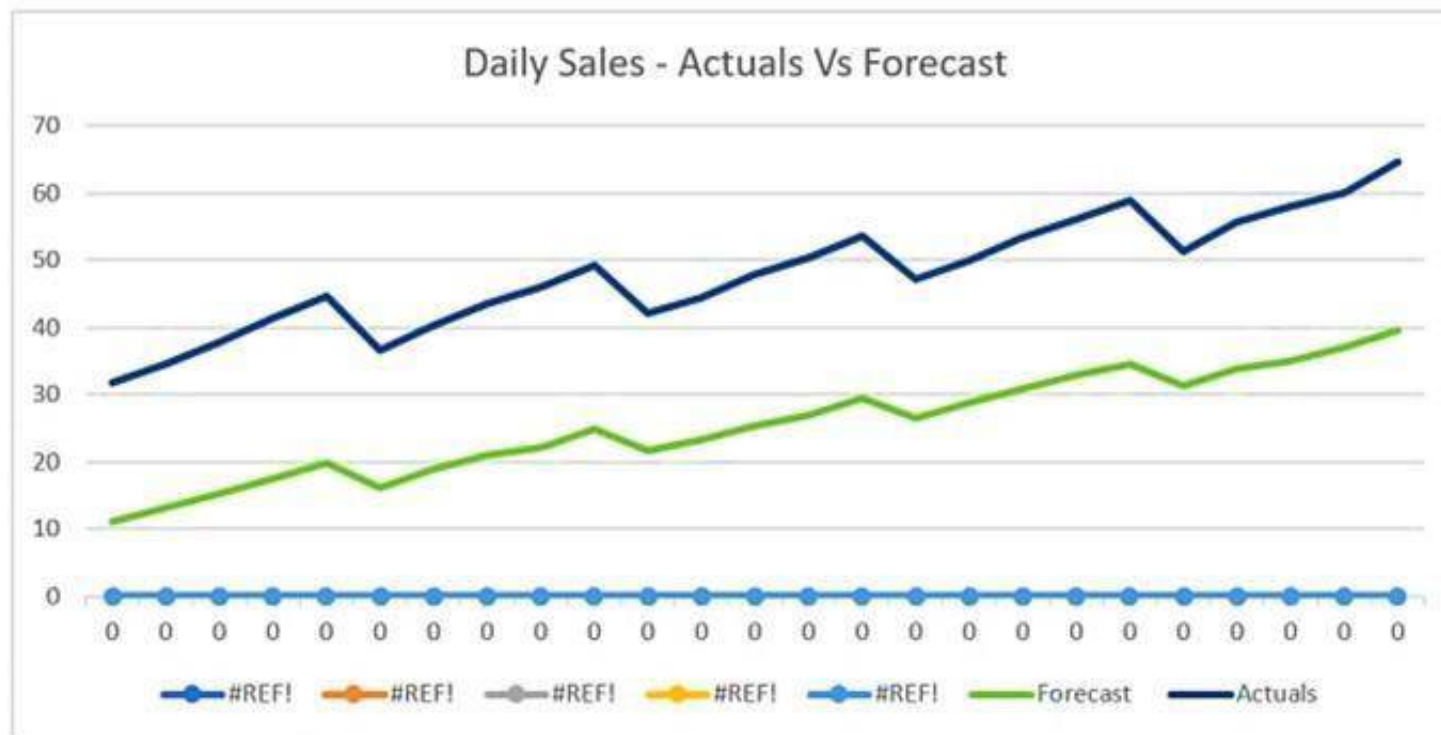Which algorithm in Forecast should the data scientists use to meet these requirements?

A. Autoregressive Integrated Moving Average (AIRMA)
B. Exponential Smoothing (ETS)

C. Convolutional Neural Network - Quantile Regression (CNN-QR)
D. Prophet

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 33
A company needs to quickly make sense of a large amount of data and gain insight from it. The data is in different formats, the schemas change frequently, and new data sources are added regularly. The company wants to use AWS services to explore multiple data sources, suggest schemas, and enrich and transform the data. The solution should require the least possible coding effort for the data flows and the least possible infrastructure management.
Which combination of AWS services will meet these requirements?

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

Answer: A

NEW QUESTION 34
A Machine Learning Specialist needs to move and transform data in preparation for training Some of the data needs to be processed in near-real time and other data can be moved hourly There are existing Amazon EMR MapReduce jobs to clean and feature engineering to perform on the data
Which of the following services can feed data to the MapReduce jobs? (Select TWO)

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

Answer: BC

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

NEW QUESTION 36
A data scientist is developing a pipeline to ingest streaming web traffic data. The data scientist needs to implement a process to identify unusual web traffic patterns as part of the pipeline. The patterns will be used downstream for alerting and incident response. The data scientist has access to unlabeled historic data to use, if needed.
The solution needs to do the following:
⟫ Calculate an anomaly score for each web traffic entry.
⟫ Adapt unusual event identification to changing web patterns over time. Which approach should the data scientist implement to meet these requirements?

A. Use historic web traffic data to train an anomaly detection model using the Amazon SageMaker Random Cut Forest (RCF) built-in mode
B. Use an Amazon Kinesis Data Stream to process the incoming webtrafficdat
C. Attach a preprocessing AWS Lambda function to perform data enrichment by calling the RCF modelto calculate the anomaly score for each record.
D. Use historic web traffic data to train an anomaly detection model using the Amazon SageMaker built-inXGBoost mode
E. Use an Amazon Kinesis Data Stream to process the incoming web traffic dat
F. Attach apreprocessing AWS Lambda function to perform data enrichment by calling the XGBoost model to calculate the anomaly score for each record.
G. Collect the streaming data using Amazon Kinesis Data Firehos
H. Map the delivery stream as an inputsource for Amazon Kinesis Data Analytic
I. Write a SQL query to run in real time against the streaming datawith the k-Nearest Neighbors (kNN) SQL extension to calculate anomaly scores for each record using a tumbling window.
J. Collect the streaming data using Amazon Kinesis Data Firehos
K. Map the delivery stream as an inputsource for Amazon Kinesis Data Analytic
L. Write a SQL query to run in real time against the streaming datawith the Amazon Random Cut Forest (RCF) SQL extension to calculate anomaly scores for each record using a sliding window.

Answer: D

NEW QUESTION 40
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 43
A machine learning (ML) specialist wants to secure calls to the Amazon SageMaker Service API. The specialist has configured Amazon VPC with a VPC interface endpoint for the Amazon SageMaker Service API and is attempting to secure traffic from specific sets of instances and IAM users. The VPC is configured with a single public subnet.
Which combination of steps should the ML specialist take to secure the traffic? (Choose two.)

A. Add a VPC endpoint policy to allow access to the IAM users.
B. Modify the users' IAM policy to allow access to Amazon SageMaker Service API calls only.
C. Modify the security group on the endpoint network interface to restrict access to the instances.
D. Modify the ACL on the endpoint network interface to restrict access to the instances.
E. Add a SageMaker Runtime VPC endpoint interface to the VPC.

Answer: AC

NEW QUESTION 44
A Data Scientist received a set of insurance records, each consisting of a record ID, the final outcome among 200 categories, and the date of the final outcome. Some partial information on claim contents is also provided, but only for a few of the 200 categories. For each outcome category, there are hundreds of records distributed over the past 3 years. The Data Scientist wants to predict how many claims to expect in each category from month to month, a few months in advance.
What type of machine learning model should be used?

A. Classification month-to-month using supervised learning of the 200 categories based on claim contents.
B. Reinforcement learning using claim IDs and timestamps where the agent will identify how many claims ineach category to expect from month to month.
C. Forecasting using claim IDs and timestamps to identify how many claims in each category to expect frommonth to month.
D. Classification with supervised learning of the categories for which partial information on claim contents isprovided, and forecasting using claim IDs and timestamps for all other categories.

Answer: C

NEW QUESTION 46
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 datase
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 require
I. Store the new dataset in Amazon S3.

Answer: BCE

NEW QUESTION 49
A Machine Learning Specialist is configuring Amazon SageMaker so multiple Data Scientists can access notebooks, train models, and deploy endpoints. To

ensure the best operational performance, the Specialist needs to be able to track how often the Scientists are deploying models, GPU and CPU utilization on the deployed SageMaker endpoints, and all errors that are generated when an endpoint is invoked.
Which services are integrated with Amazon SageMaker to track this information? (Select TWO.)

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

Answer: AD

NEW QUESTION 53
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 57
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 enable
E. Generate an S3 pre-signed URL for access to data in the bucket

Answer: B

NEW QUESTION 58
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 60
A Machine Learning Specialist needs to create a data repository to hold a large amount of time-based training data for a new model. In the source system, new files are added every hour Throughout a single 24-hour period, the volume of hourly updates will change significantly. The Specialist always wants to train on the last 24 hours of the data
Which type of data repository is the MOST cost-effective solution?

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

Answer: C

NEW QUESTION 64
A retail company wants to combine its customer orders with the product description data from its product catalog. The structure and format of the records in each dataset is different. A data analyst tried to use a spreadsheet to combine the datasets, but the effort resulted in duplicate records and records that were not properly combined. The company needs a solution that it can use to combine similar records from the two datasets and remove any duplicates.
Which solution will meet these requirements?

A. Use an AWS Lambda function to process the dat
B. Use two arrays to compare equal strings in the fields from the two datasets and remove any duplicates.
C. Create AWS Glue crawlers for reading and populating the AWS Glue Data Catalo
D. Call the AWS Glue SearchTables API operation to perform a fuzzy-matching search on the two datasets, and cleanse the data accordingly.
E. Create AWS Glue crawlers for reading and populating the AWS Glue Data Catalo
F. Use the FindMatches transform to cleanse the data.

G. Create an AWS Lake Formation custom transfor
H. Run a transformation for matching products from the Lake Formation console to cleanse the data automatically.

Answer: D

NEW QUESTION 67
A media company with a very large archive of unlabeled images, text, audio, and video footage wishes to index its assets to allow rapid identification of relevant content by the Research team. The company wants to use machine learning to accelerate the efforts of its in-house researchers who have limited machine learning expertise.
Which is the FASTEST route to index the assets?

A. Use Amazon Rekognition, Amazon Comprehend, and Amazon Transcribe to tag data into distinct categories/classes.
B. Create a set of Amazon Mechanical Turk Human Intelligence Tasks to label all footage.
C. Use Amazon Transcribe to convert speech to tex
D. Use the Amazon SageMaker Neural Topic Model (NTM) and Object Detection algorithms to tag data into distinct categories/classes.
E. Use the AWS Deep Learning AMI and Amazon EC2 GPU instances to create custom models for audio transcription and topic modeling, and use object detection to tag data into distinct categories/classes.

Answer: A

NEW QUESTION 69
An Amazon SageMaker notebook instance is launched into Amazon VPC The SageMaker notebook references data contained in an Amazon S3 bucket in another account The bucket is encrypted using SSE-KMS The instance returns an access denied error when trying to access data in Amazon S3.
Which of the following are required to access the bucket and avoid the access denied error? (Select THREE)

A. An AWS KMS key policy that allows access to the customer master key (CMK)
B. A SageMaker notebook security group that allows access to Amazon S3
C. An 1AM role that allows access to the specific S3 bucket
D. A permissive S3 bucket policy
E. An S3 bucket owner that matches the notebook owner
F. A SegaMaker notebook subnet ACL that allow traffic to Amazon S3.

Answer: ACF

NEW QUESTION 70
Amazon Connect has recently been tolled out across a company as a contact call center The solution has been configured to store voice call recordings on Amazon S3
The content of the voice calls are being analyzed for the incidents being discussed by the call operators Amazon Transcribe is being used to convert the audio to text, and the output is stored on Amazon S3
Which approach will provide the information required for further analysis?

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

Answer: B

NEW QUESTION 73
A financial company is trying to detect credit card fraud. The company observed that, on average, 2% of credit card transactions were fraudulent. A data scientist trained a classifier on a year's worth of credit card transactions data. The model needs to identify the fraudulent transactions (positives) from the regular ones (negatives). The company's goal is to accurately capture as many positives as possible.
Which metrics should the data scientist use to optimize the model? (Choose two.)

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

Answer: DE

NEW QUESTION 74
A Data Scientist is developing a binary classifier to predict whether a patient has a particular disease on a series of test results. The Data Scientist has data on 400 patients randomly selected from the population. The disease is seen in 3% of the population.
Which cross-validation strategy should the Data Scientist adopt?

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

Answer: B

NEW QUESTION 79
Given the following confusion matrix for a movie classification model, what is the true class frequency for Romance and the predicted class frequency for Adventure?

Predicted Values
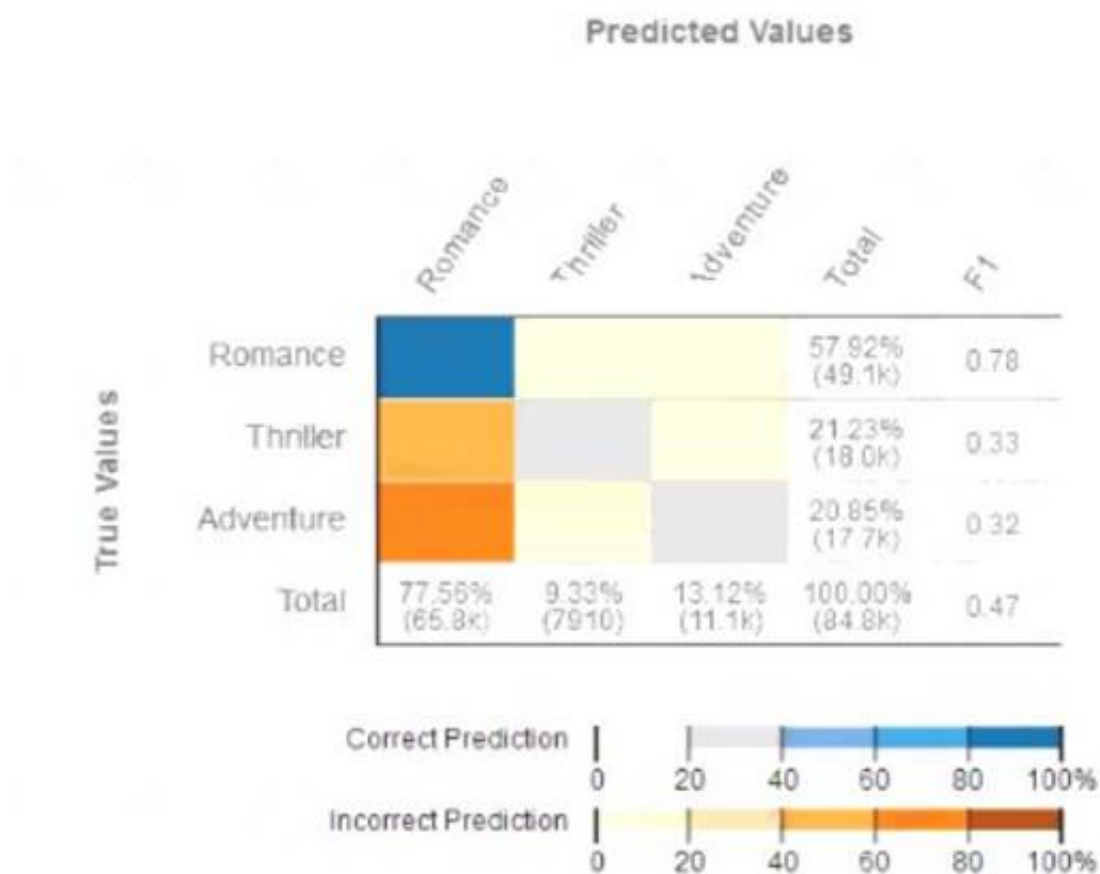
A. The true class frequency for Romance is 77.56% and the predicted class frequency for Adventure is 20 85%
B. The true class frequency for Romance is 57.92% and the predicted class frequency for Adventure is 1312%
C. The true class frequency for Romance is 0 78 and the predicted class frequency for Adventure is (0 47 - 0.32).
D. The true class frequency for Romance is 77.56% * 0.78 and the predicted class frequency for Adventure is 20 85% ' 0.32

Answer: B

Explanation:
https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-model-insights.html

NEW QUESTION 81
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 83
A company is launching a new product and needs to build a mechanism to monitor comments about the company and its new product on social media. The company needs to be able to evaluate the sentiment expressed in social media posts, and visualize trends and configure alarms based on various thresholds. The company needs to implement this solution quickly, and wants to minimize the infrastructure and data science resources needed to evaluate the messages. The company already has a solution in place to collect posts and store them within an Amazon S3 bucket.
What services should the data science team use to deliver this solution?

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

Answer: A

NEW QUESTION 86
A data scientist is working on a public sector project for an urban traffic system. While studying the traffic patterns, it is clear to the data scientist that the traffic behavior at each light is correlated, subject to a small stochastic error term. The data scientist must model the traffic behavior to analyze the traffic patterns and reduce congestion.
How will the data scientist MOST effectively model the problem?

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

Answer: D

NEW QUESTION 87
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 88
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 100000 images per defect type for training During the injial 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 issue1?

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 92
While working on a neural network project, a Machine Learning Specialist discovers thai 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 regulanzation
D. Data augmentation for the minority class

Answer: D

NEW QUESTION 95
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 cod
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 deploymen
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 deploymen
F. Build the image and upload it to Docker Hub.
G. Build the Docker image with the inference cod
H. Configure Docker Hub and upload the image to AmazonECR.

Answer: D

NEW QUESTION 97
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 98
A machine learning specialist is running an Amazon SageMaker endpoint using the built-in object detection algorithm on a P3 instance for real-time predictions in a company's production application. When evaluating the model's resource utilization, the specialist notices that the model is using only a fraction of the GPU.
Which architecture changes would ensure that provisioned resources are being utilized effectively?

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

Answer: B

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

NEW QUESTION 99
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 103
A company is running an Amazon SageMaker training job that will access data stored in its Amazon S3 bucket A compliance policy requires that the data never be transmitted across the internet How should the company set up the job?

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

Answer: D

NEW QUESTION 107
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 111
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 116
IT leadership wants Jo transition a company's existing machine learning data storage environment to AWS as a temporary ad hoc solution The company currently uses a custom software process that heavily leverages SOL as a query language and exclusively stores generated csv documents for machine learning
The ideal state for the company would be a solution that allows it to continue to use the current workforce of SQL experts The solution must also support the storage of csv and JSON files, and be able to query over semi-structured data The following are high priorities for the company:
• Solution simplicity
• Fast development time
• Low cost
• High flexibility
What technologies meet the company's requirements?

A. Amazon S3 and Amazon Athena
B. Amazon Redshift and AWS Glue
C. Amazon DynamoDB and DynamoDB Accelerator (DAX)
D. Amazon RDS and Amazon ES

Answer: B

NEW QUESTION 121
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 123
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 127
A Machine Learning Specialist is working with a large company to leverage machine learning within its products. The company wants to group its customers into categories based on which customers will and will not churn within the next 6 months. The company has labeled the data available to the Specialist.
Which machine learning model type should the Specialist use to accomplish this task?

A. Linear regression
B. Classification
C. Clustering
D. Reinforcement learning

Answer: B

Explanation:
The goal of classification is to determine to which class or category a data point (customer in our case) belongs to. For classification problems, data scientists would use historical data with predefined target variables AKA labels (churner/non-churner) – answers that need to be predicted – to train an algorithm. With classification, businesses can answer the following questions:

> Will this customer churn or not?

> Will a customer renew their subscription?

> Will a user downgrade a pricing plan?

> Are there any signs of unusual customer behavior?

NEW QUESTION 132
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 137
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 141
A company uses camera images of the tops of items displayed on store shelves to determine which items were removed and which ones still remain. After several hours of data labeling, the company has a total of 1,000 hand-labeled images covering 10 distinct items. The training results were poor.
Which machine learning approach fulfills the company's long-term needs?

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

Answer: A

NEW QUESTION 145
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 147
A Machine Learning Specialist trained a regression model, but the first iteration needs optimizing. The Specialist needs to understand whether the model is more frequently overestimating or underestimating the target.
What option can the Specialist use to determine whether it is overestimating or underestimating the target value?

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

Answer: B

NEW QUESTION 148
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 151
A Machine Learning Specialist receives customer data for an online shopping website. The data includes demographics, past visits, and locality information. The Specialist must develop a machine learning approach to identify the customer shopping patterns, preferences and trends to enhance the website for better service and smart recommendations.
Which solution should the Specialist recommend?

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

Answer: C

NEW QUESTION 156
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 157
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 158
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 160
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 161

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 163
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 168
A logistics company needs a forecast model to predict next month's inventory requirements for a single item in 10 warehouses. A machine learning specialist uses Amazon Forecast to develop a forecast model from 3 years of monthly data. There is no missing data. The specialist selects the DeepAR+ algorithm to train a predictor. The predictor means absolute percentage error (MAPE) is much larger than the MAPE produced by the current human forecasters.
Which changes to the CreatePredictor API call could improve the MAPE? (Choose two.)

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

Answer: CD

NEW QUESTION 169
During mini-batch training of a neural network for a classification problem, a Data Scientist notices that training accuracy oscillates What is the MOST likely cause of this issue?

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

Answer: B

NEW QUESTION 170
A data scientist uses an Amazon SageMaker notebook instance to conduct data exploration and analysis. This requires certain Python packages that are not natively available on Amazon SageMaker to be installed on the notebook instance.
How can a machine learning specialist ensure that required packages are automatically available on the notebook instance for the data scientist to use?

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

Answer: D

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

NEW QUESTION 175
A data scientist needs to identify fraudulent user accounts for a company's ecommerce platform. The company wants the ability to determine if a newly created account is associated with a previously known fraudulent user. The data scientist is using AWS Glue to cleanse the company's application logs during ingestion.
Which strategy will allow the data scientist to identify fraudulent accounts?

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

Answer: B

NEW QUESTION 178

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 183

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

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

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

Answer: A

NEW QUESTION 188

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

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

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

A. Increase the max_depth parameter value.
B. Lower the max_depth parameter value.
C. Update the objective to binary:logistic.
D. Lower the min_child_weight parameter value.

Answer: B

NEW QUESTION 191

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

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

Answer: D

NEW QUESTION 193

A manufacturing company has structured and unstructured data stored in an Amazon S3 bucket. A Machine Learning Specialist wants to use SQL to run queries on this data.

Which solution requires the LEAST effort to be able to query this data?

A. Use AWS Data Pipeline to transform the data and Amazon RDS to run queries.
B. Use AWS Glue to catalogue the data and Amazon Athena to run queries.
C. Use AWS Batch to run ETL on the data and Amazon Aurora to run the queries.
D. Use AWS Lambda to transform the data and Amazon Kinesis Data Analytics to run queries.

Answer: B

NEW QUESTION 197
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 202
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 204
A Machine Learning Specialist is developing recommendation engine for a photography blog Given a picture, the recommendation engine should show a picture that captures similar objects The Specialist would like to create a numerical representation feature to perform nearest-neighbor searches
What actions would allow the Specialist to get relevant numerical representations?

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

Answer: A

NEW QUESTION 209
A Machine Learning Specialist is planning to create a long-running Amazon EMR cluster. The EMR cluster will have 1 master node, 10 core nodes, and 20 task nodes. To save on costs, the Specialist will use Spot Instances in the EMR cluster.
Which nodes should the Specialist launch on Spot Instances?

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

Answer: A

NEW QUESTION 210
For the given confusion matrix, what is the recall and precision of the model?

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

Answer: C

NEW QUESTION 215
......

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