AWS-Certified-Machine-Learning-Specialty Dumps

AWS Certified Machine Learning - Specialty

https://www.certleader.com/AWS-Certified-Machine-Learning-Specialty-dumps.html
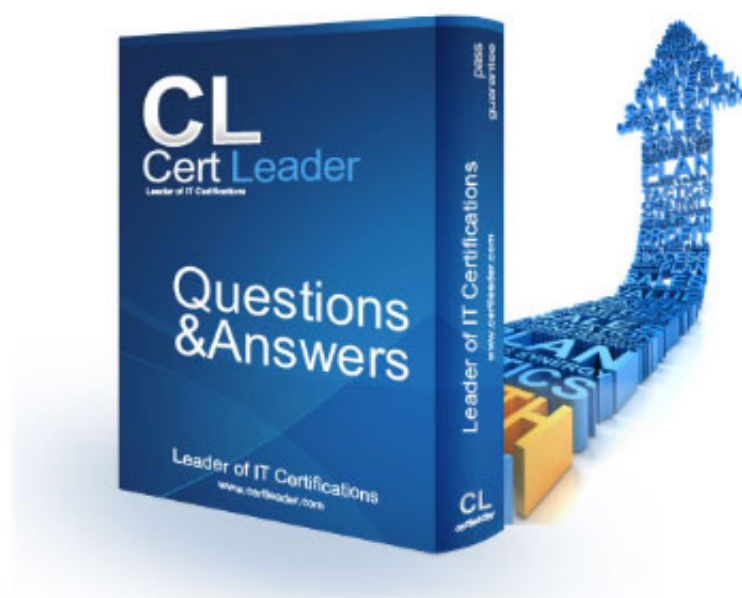
NEW QUESTION 1
A Data Scientist needs to create a serverless ingestion and analytics solution for high-velocity, real-time streaming data.
The ingestion process must buffer and convert incoming records from JSON to a query-optimized, columnar format without data loss. The output datastore must be highly 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 BI 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 BI 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 BI tools using the Athena Java Database Connectivity (JDBC) connector.

Answer: A

NEW QUESTION 2
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
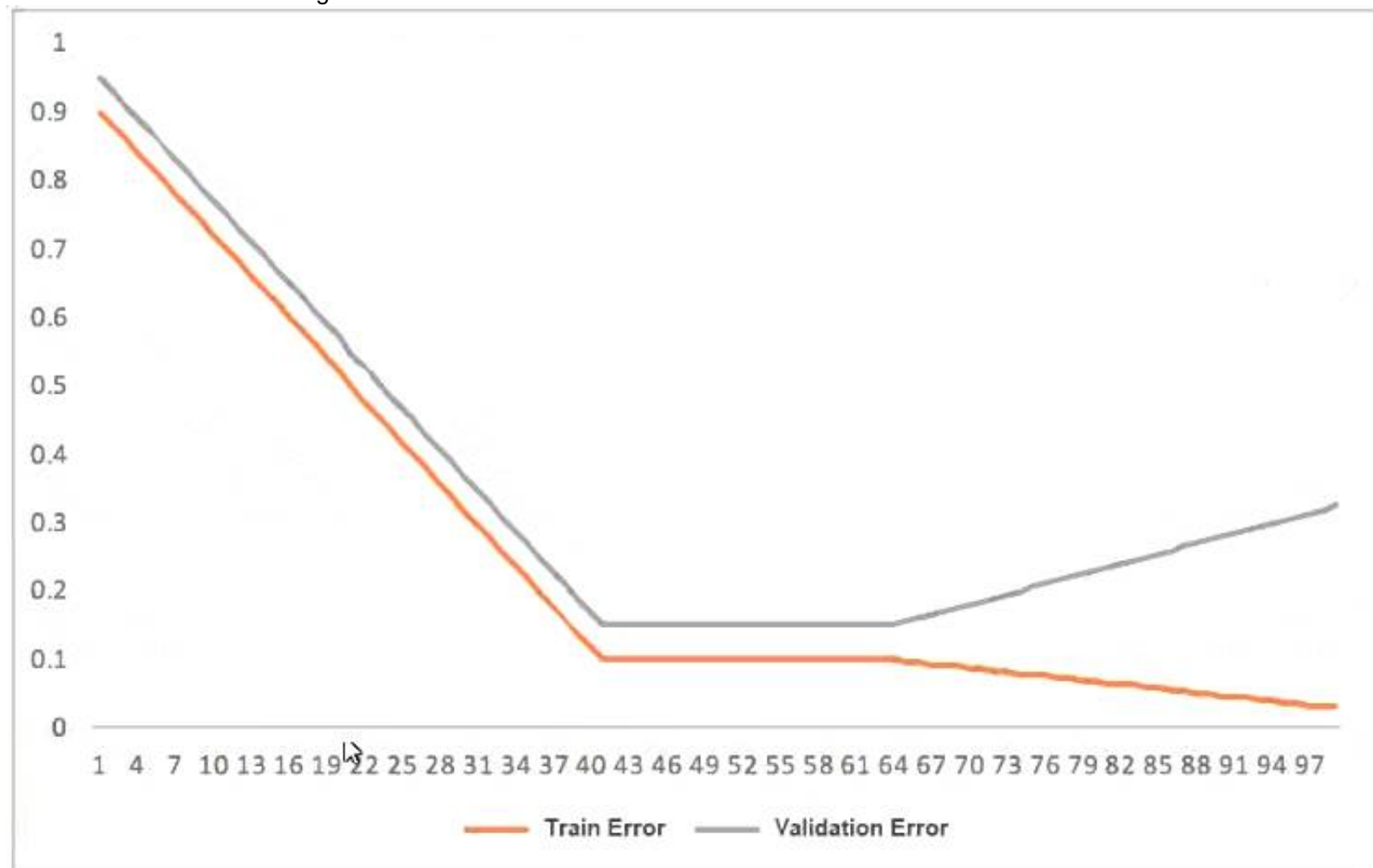
Which technique can be used to improve model performance in terms of accuracy in the validation set?

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 3
A Machine Learning Specialist observes several performance problems with the training portion of a machine learning solution on Amazon SageMaker The solution uses a large training dataset 2 TB in size and is using the SageMaker k-means algorithm The observed issues include the unacceptable length of time it takes before the training job launches and poor I/O throughput while training the model
What should the Specialist do to address the performance issues with the current solution?

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

Answer: B

NEW QUESTION 4

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 5

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

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

Answer: C

NEW QUESTION 6

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 7

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

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

Answer: C

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

NEW QUESTION 8

A Machine Learning Specialist is creating a new natural language processing application that processes a dataset comprised of 1 million sentences The aim is to then run Word2Vec to generate embeddings of the sentences and enable different types of predictions
Here is an example from the dataset
"The quck BROWN FOX jumps over the lazy dog "
Which of the following are the operations the Specialist needs to perform to correctly sanitize and prepare the data in a repeatable manner? (Select THREE)

A. Perform part-of-speech tagging and keep the action verb and the nouns only

B. Normalize all words by making the sentence lowercase
C. Remove stop words using an English stopword dictionary.
D. Correct the typography on "quck" to "quick."
E. One-hot encode all words in the sentence
F. Tokenize the sentence into words.

Answer: BCF

NEW QUESTION 9

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

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

Answer: C

NEW QUESTION 10

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 10

A Machine Learning team uses Amazon SageMaker to train an Apache MXNet handwritten digit classifier model using a research dataset. The team wants to receive a notification when the model is overfitting. Auditors want to view the Amazon SageMaker log activity report to ensure there are no unauthorized API calls. What should the Machine Learning team do to address the requirements with the least amount of code and fewest steps?

A. Implement an AWS Lambda function to long Amazon SageMaker API calls to Amazon S3. Add code to push a custom metric to Amazon CloudWatc
B. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.
C. Use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. Add code to push a custom metric to Amazon CloudWatc
D. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.
E. Implement an AWS Lambda function to log Amazon SageMaker API calls to AWS CloudTrai
F. Add code to push a custom metric to Amazon CloudWatc
G. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.
H. Use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. Set up Amazon SNS to receive a notification when the model is overfitting.

Answer: C

NEW QUESTION 11

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

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

Answer: C

NEW QUESTION 16

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 19
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 22
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 26
A Machine Learning Specialist deployed a model that provides product recommendations on a company's website Initially, the model was performing very well and resulted in customers buying more products on average However within the past few months the Specialist has noticed that the effect of product recommendations has diminished and customers are starting to return to their original habits of spending less The Specialist is unsure of what happened, as the model has not changed from its initial deployment over a year ago
Which method should the Specialist try to improve model performance?

A. The model needs to be completely re-engineered because it is unable to handle product inventory changes
B. The model's hyperparameters should be periodically updated to prevent drift
C. The model should be periodically retrained from scratch using the original data while adding a regularization term to handle product inventory changes
D. The model should be periodically retrained using the original training data plus new data as product inventory changes

Answer: D

NEW QUESTION 31
A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a Machine Learning Specialist would like to build a binary classifier based on two features: age of account and transaction month. The class distribution for these features is illustrated in the figure provided.
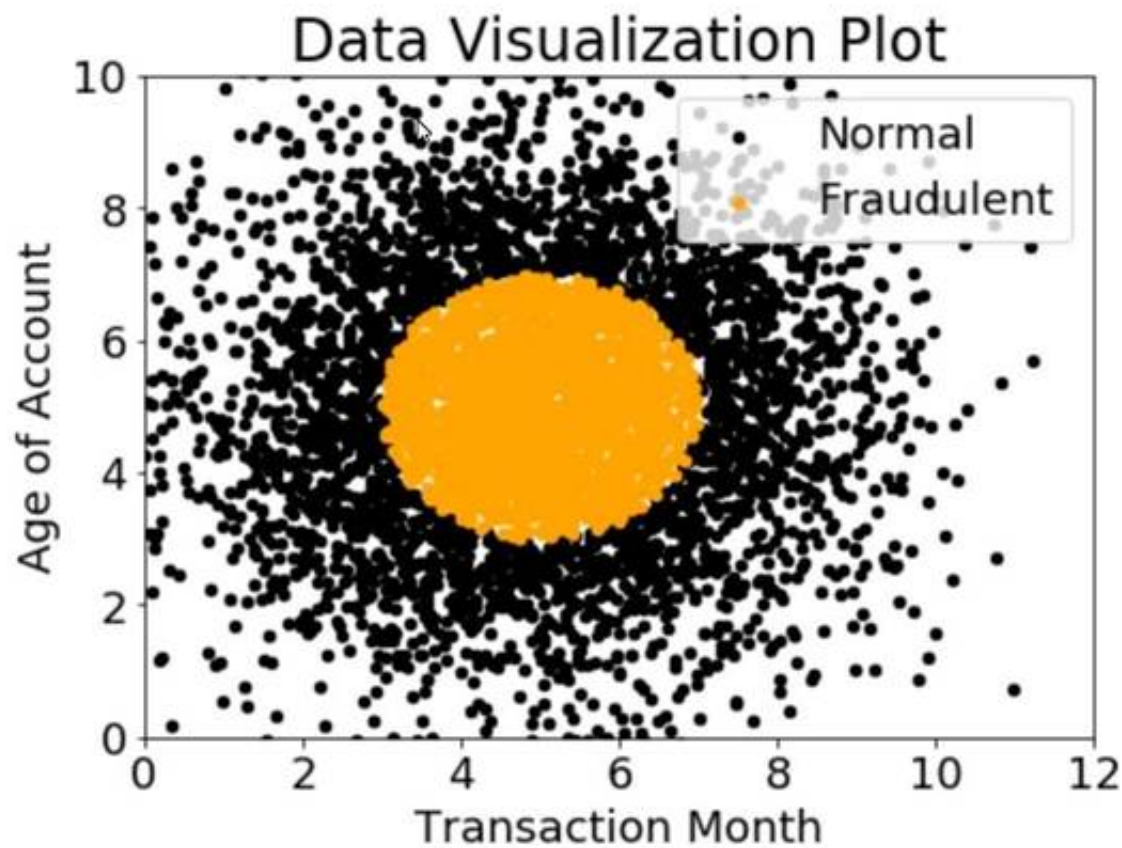
Based on this information which model would have the HIGHEST accuracy?

A. Long short-term memory (LSTM) model with scaled exponential linear unit (SELL))
B. Logistic regression
C. Support vector machine (SVM) with non-linear kernel
D. Single perceptron with tanh activation function

Answer: C

NEW QUESTION 35
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 37
The displayed graph is from a foresting model for testing a time series.
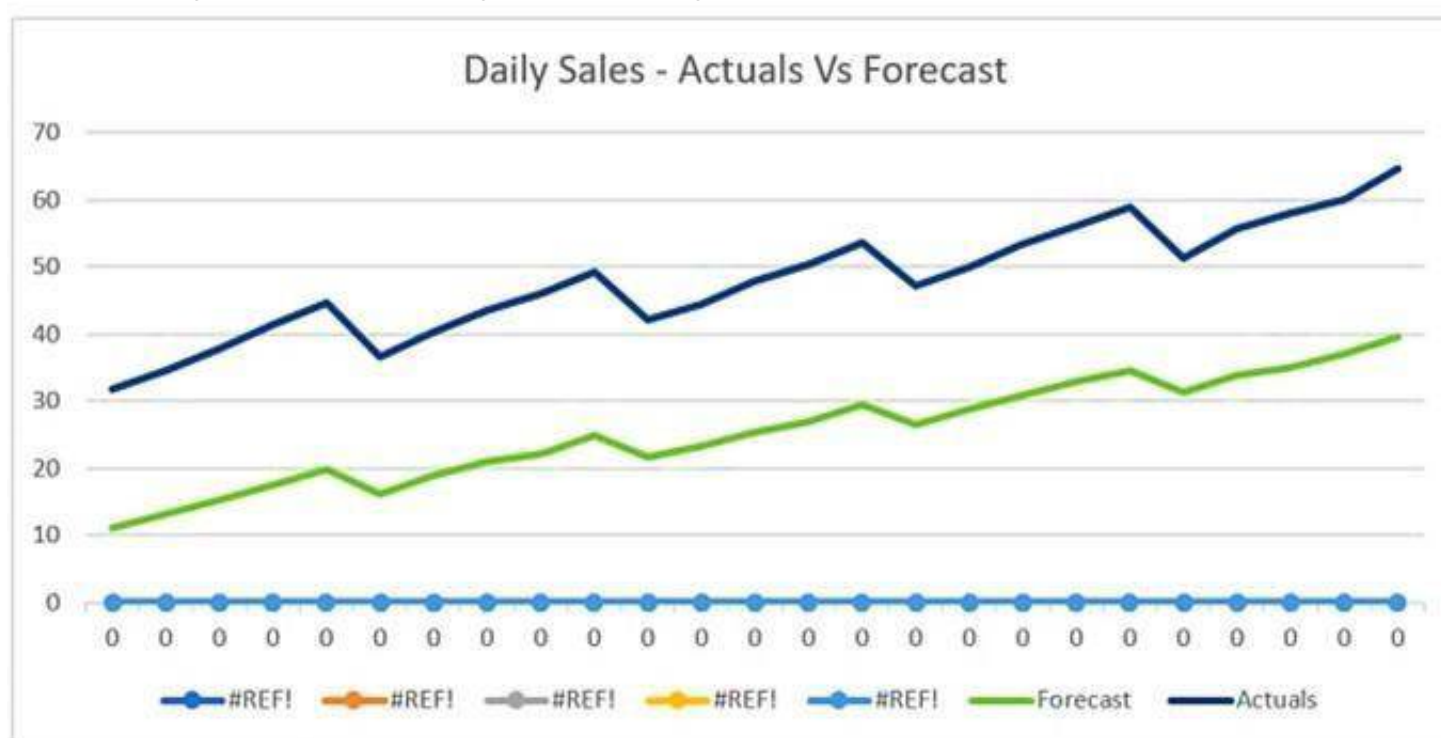
Considering the graph only, which conclusion should a Machine Learning Specialist make about the behavior of the model?

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 40
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 45
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 50
A large company has developed a B1 application that generates reports and dashboards using data collected from various operational metrics The company wants to provide executives with an enhanced experience so they can use natural language to get data from the reports The company wants the executives to be able ask questions using written and spoken interlaces
Which combination of services can be used to build this conversational interface? (Select THREE)

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

Answer: BEF

NEW QUESTION 51
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 53

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 do1?

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

Answer: A

NEW QUESTION 56

A Machine Learning Specialist wants to bring a custom algorithm to Amazon SageMaker. The Specialist implements the algorithm in a Docker container supported by Amazon SageMaker.
How should the Specialist package the Docker container so that Amazon SageMaker can launch the training correctly?

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

Answer: B

NEW QUESTION 57

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 60

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 65

A trucking company is collecting live image data from its fleet of trucks across the globe. The data is growing rapidly and approximately 100 GB of new data is generated every day. The company wants to explore machine learning uses cases while ensuring the data is only accessible to specific IAM users.
Which storage option provides the most processing flexibility and will allow access control with IAM?

A. Use a database, such as Amazon DynamoDB, to store the images, and set the IAM policies to restrict access to only the desired IAM users.
B. Use an Amazon S3-backed data lake to store the raw images, and set up the permissions using bucket policies.
C. Setup up Amazon EMR with Hadoop Distributed File System (HDFS) to store the files, and restrictaccess to the EMR instances using IAM policies.
D. Configure Amazon EFS with IAM policies to make the data available to Amazon EC2 instances owned by the IAM users.

Answer: C

NEW QUESTION 66

A global financial company is using machine learning to automate its loan approval process. The company has a dataset of customer information. The dataset contains some categorical fields, such as customer location by city and housing status. The dataset also includes financial fields in different units, such as account balances in US dollars and monthly interest in US cents.
The company's data scientists are using a gradient boosting regression model to infer the credit score for each customer. The model has a training accuracy of 99% and a testing accuracy of 75%. The data scientists want to improve the model's testing accuracy.
Which process will improve the testing accuracy the MOST?

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

Answer: B

NEW QUESTION 67
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 69
A real estate company wants to create a machine learning model for predicting housing prices based on a historical dataset. The dataset contains 32 features.
Which model will meet the business requirement?

A. Logistic regression
B. Linear regression
C. K-means
D. Principal component analysis (PCA)

Answer: B

NEW QUESTION 72
A company is running a machine learning prediction service that generates 100 TB of predictions every day A Machine Learning Specialist must generate a visualization of the daily precision-recall curve from the predictions, and forward a read-only version to the Business team.
Which solution requires the LEAST coding effort?

A. Run a daily Amazon EMR workflow to generate precision-recall data, and save the results in Amazon S3 Give the Business team read-only access to S3
B. Generate daily precision-recall data in Amazon QuickSight, and publish the results in a dashboard shared with the Business team
C. Run a daily Amazon EMR workflow to generate precision-recall data, and save the results in Amazon S3 Visualize the arrays in Amazon QuickSight, and publish them in a dashboard shared with the Business team
D. Generate daily precision-recall data in Amazon ES, and publish the results in a dashboard shared with the Business team.

Answer: C

NEW QUESTION 74
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 77
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 81
A machine learning (ML) specialist wants to create a data preparation job that uses a PySpark script with complex window aggregation operations to create data for training and testing. The ML specialist needs to evaluate the impact of the number of features and the sample count on model performance.
Which approach should the ML specialist use to determine the ideal data transformations for the model?

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

Answer: B

NEW QUESTION 85
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 86
A Machine Learning Specialist is working for a credit card processing company and receives an unbalanced dataset containing credit card transactions. It contains 99,000 valid transactions and 1,000 fraudulent transactions The Specialist is asked to score a model that was run against the dataset The Specialist has been advised that identifying valid transactions is equally as important as identifying fraudulent transactions
What metric is BEST suited to score the model?

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

Answer: A

NEW QUESTION 89
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 93
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 98
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 100
A data scientist is training a text classification model by using the Amazon SageMaker built-in BlazingText algorithm. There are 5 classes in the dataset, with 300 samples for category A, 292 samples for category B, 240 samples for category C, 258 samples for category D, and 310 samples for category E.
The data scientist shuffles the data and splits off 10% for testing. After training the model, the data scientist generates confusion matrices for the training and test sets.

Training data confusion matrix

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

Test data confusion matrix

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

What could the data scientist conclude form these results?

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

Answer: B

NEW QUESTION 102
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 107
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 109
A Machine Learning Specialist has created a deep learning neural network model that performs well on the training data but performs poorly on the test data.
Which of the following methods should the Specialist consider using to correct this? (Select THREE.)

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

Answer: BCD

NEW QUESTION 113
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 117
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 122
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 124
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 125
A credit card company wants to build a credit scoring model to help predict whether a new credit card applicant will default on a credit card payment. The company has collected data from a large number of sources with thousands of raw attributes. Early experiments to train a classification model revealed that many attributes are highly correlated, the large number of features slows down the training speed significantly, and that there are some overfitting issues.
The Data Scientist on this project would like to speed up the model training time without losing a lot of information from the original dataset.
Which feature engineering technique should the Data Scientist use to meet the objectives?

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

Answer: B

NEW QUESTION 129
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 131
A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a Machine Learning Specialist would like to build a binary classifier based on two features: age of account and transaction month. The class distribution for these features is illustrated in the figure provided.
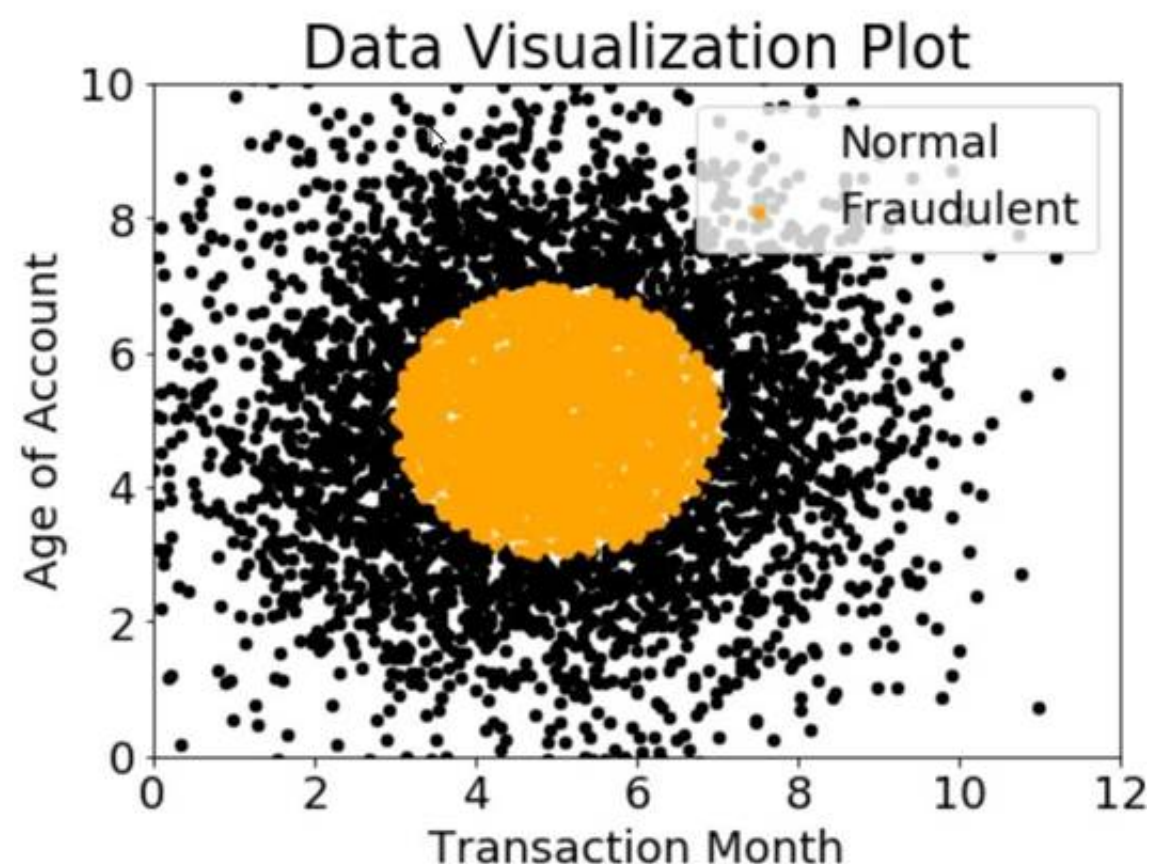
Based on this information, which model would have the HIGHEST recall with respect to the fraudulent class?

A. Decision tree
B. Linear support vector machine (SVM)
C. Naive Bayesian classifier

D. Single Perceptron with sigmoidal activation function

Answer: C

NEW QUESTION 136
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 140
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 145
A Mobile Network Operator is building an analytics platform to analyze and optimize a company's operations using Amazon Athena and Amazon S3
The source systems send data in CSV format in real lime The Data Engineering team wants to transform the data to the Apache Parquet format before storing it on Amazon S3
Which solution takes the LEAST effort to implement?

A. Ingest .CSV data using Apache Kafka Streams on Amazon EC2 instances and use Kafka Connect S3 toserialize data as Parquet
B. Ingest .CSV data from Amazon Kinesis Data Streams and use Amazon Glue to convert data into Parquet.
C. Ingest .CSV data using Apache Spark Structured Streaming in an Amazon EMR cluster and use Apache Spark to convert data into Parquet.
D. Ingest .CSV data from Amazon Kinesis Data Streams and use Amazon Kinesis Data Firehose to convert data into Parquet.

Answer: B

Explanation:
https://medium.com/searce/convert-csv-json-files-to-apache-parquet-using-aws-glue-a760d177b45f https://github.com/ecloudvalley/Building-a-Data-Lake-with-AWS-Glue-and-Amazon-S3

NEW QUESTION 150
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 153
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 155
A company is observing low accuracy while training on the default built-in image classification algorithm in Amazon SageMaker. The Data Science team wants to use an Inception neural network architecture instead of a ResNet architecture.
Which of the following will accomplish this? (Select TWO.)

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

Answer: AD

NEW QUESTION 159
A machine learning (ML) specialist is administering a production Amazon SageMaker endpoint with model monitoring configured. Amazon SageMaker Model Monitor detects violations on the SageMaker endpoint, so the ML specialist retrains the model with the latest dataset. This dataset is statistically representative of the current production traffic. The ML specialist notices that even after deploying the new SageMaker model and running the first monitoring job, the SageMaker endpoint still has violations.
What should the ML specialist do to resolve the violations?

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

Answer: B

NEW QUESTION 164
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 165
A company that manufactures mobile devices wants to determine and calibrate the appropriate sales price for its devices. The company is collecting the relevant data and is determining data features that it can use to train machine learning (ML) models. There are more than 1,000 features, and the company wants to determine the primary features that contribute to the sales price.
Which techniques should the company use for feature selection? (Choose three.)

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

Answer: CDF

NEW QUESTION 167
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 169
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 174
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 175
A telecommunications company is developing a mobile app for its customers. The company is using an Amazon SageMaker hosted endpoint for machine learning model inferences.
Developers want to introduce a new version of the model for a limited number of users who subscribed to a preview feature of the app. After the new version of the model is tested as a preview, developers will evaluate its accuracy. If a new version of the model has better accuracy, developers need to be able to gradually release the new version for all users over a fixed period of time.
How can the company implement the testing model with the LEAST amount of operational overhead?

A. Update the ProductionVariant data type with the new version of the model by using the CreateEndpointConfig operation with the InitialVariantWeight parameter set to 0. Specify the TargetVariant parameter for InvokeEndpoint calls for users who subscribed to the preview featur
B. When the new version of the model is ready for release, gradually increase InitialVariantWeight until all users have the updated version.
C. Configure two SageMaker hosted endpoints that serve the different versions of the mode
D. Create an Application Load Balancer (ALB) to route traffic to both endpoints based on the TargetVariant query string paramete
E. Reconfigure the app to send the TargetVariant query string parameter for users who subscribed to the preview featur
F. When the new version of the model is ready for release, change the ALB's routing algorithm to weighted until all users have the updated version.
G. Update the DesiredWeightsAndCapacity data type with the new version of the model by using the UpdateEndpointWeightsAndCapacities operation with the DesiredWeight parameter set to 0. Specify the TargetVariant parameter for InvokeEndpoint calls for users who subscribed to the preview featur
H. When the new version of the model is ready for release, gradually increase DesiredWeight until all users have the updated version.
I. Configure two SageMaker hosted endpoints that serve the different versions of the mode
J. Create an Amazon Route 53 record that is configured with a simple routing policy and that points to the current version of the mode
K. Configure the mobile app to use the endpoint URL for users who subscribed to the preview feature and to use the Route 53 record for other user
L. When the new version of the model is ready for release, add a new model version endpoint to Route 53, and switch the policy to weighted until all users have the updated version.

Answer: D

NEW QUESTION 179
An employee found a video clip with audio on a company's social media feed. The language used in the video is Spanish. English is the employee's first language, and they do not understand Spanish. The employee wants to do a sentiment analysis.
What combination of services is the MOST efficient to accomplish the task?

A. Amazon Transcribe, Amazon Translate, and Amazon Comprehend
B. Amazon Transcribe, Amazon Comprehend, and Amazon SageMaker seq2seq
C. Amazon Transcribe, Amazon Translate, and Amazon SageMaker Neural Topic Model (NTM)
D. Amazon Transcribe, Amazon Translate, and Amazon SageMaker BlazingText

Answer: A

NEW QUESTION 180
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 185
A company has video feeds and images of a subway train station. The company wants to create a deep learning model that will alert the station manager if any passenger crosses the yellow safety line when there is no train in the station. The alert will be based on the video feeds. The company wants the model to detect the yellow line, the passengers who cross the yellow line, and the trains in the video feeds. This task requires labeling. The video data must remain confidential.
A data scientist creates a bounding box to label the sample data and uses an object detection model. However, the object detection model cannot clearly demarcate the yellow line, the passengers who cross the yellow line, and the trains.
Which labeling approach will help the company improve this model?

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

Answer: B

NEW QUESTION 190
A Machine Learning Specialist is designing a system for improving sales for a company. The objective is to use the large amount of information the company has on users' behavior and product preferences to predict which products users would like based on the users' similarity to other users.
What should the Specialist do to meet this objective?

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

Answer: B

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

NEW QUESTION 191
A company's Machine Learning Specialist needs to improve the training speed of a time-series forecasting model using TensorFlow. The training is currently implemented on a single-GPU machine and takes approximately 23 hours to complete. The training needs to be run daily.
The model accuracy js acceptable, but the company anticipates a continuous increase in the size of the training data and a need to update the model on an hourly, rather than a daily, basis. The company also wants to minimize coding effort and infrastructure changes
What should the Machine Learning Specialist do to the training solution to allow it to scale for future demand?

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

Answer: B

NEW QUESTION 195
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 197
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 198
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 202
An interactive online dictionary wants to add a widget that displays words used in similar contexts. A Machine Learning Specialist is asked to provide word features for the downstream nearest neighbor model powering the widget.
What should the Specialist do to meet these requirements?

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

Answer: D

NEW QUESTION 207
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 211
A Machine Learning Specialist wants to determine the appropriate SageMakerVariant Invocations Per Instance setting for an endpoint automatic scaling configuration. The Specialist has performed a load test on a single instance and determined that peak requests per second (RPS) without service degradation is about 20 RPS As this is the first deployment, the Specialist intends to set the invocation safety factor to 0 5
Based on the stated parameters and given that the invocations per instance setting is measured on a per-minute basis, what should the Specialist set as the sageMakervariantinvocationsPerinstance setting?

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

Answer: C

NEW QUESTION 214
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 219
A machine learning specialist needs to analyze comments on a news website with users across the globe. The specialist must find the most discussed topics in the comments that are in either English or Spanish.
What steps could be used to accomplish this task? (Choose two.)

A. Use an Amazon SageMaker BlazingText algorithm to find the topics independently from language.Proceed with the analysis.
B. Use an Amazon SageMaker seq2seq algorithm to translate from Spanish to English, if necessar
C. Use aSageMaker Latent Dirichlet Allocation (LDA) algorithm to find the topics.
D. Use Amazon Translate to translate from Spanish to English, if necessar
E. Use Amazon Comprehend topic modeling to find the topics.
F. Use Amazon Translate to translate from Spanish to English, if necessar
G. Use Amazon Lex to extract topics form the content.
H. Use Amazon Translate to translate from Spanish to English, if necessar
I. Use Amazon SageMaker Neural Topic Model (NTM) to find the topics.

Answer: B

NEW QUESTION 224
An insurance company is developing a new device for vehicles that uses a camera to observe drivers' behavior and alert them when they appear distracted The company created approximately 10,000 training images in a controlled environment that a Machine Learning Specialist will use to train and evaluate machine learning models
During the model evaluation the Specialist notices that the training error rate diminishes faster as the number of epochs increases and the model is not accurately inferring on the unseen test images
Which of the following should be used to resolve this issue? (Select TWO)

A. Add vanishing gradient to the model
B. Perform data augmentation on the training data
C. Make the neural network architecture complex.
D. Use gradient checking in the model
E. Add L2 regularization to the model

Answer: BD

NEW QUESTION 225
A retail company is selling products through a global online marketplace. The company wants to use machine learning (ML) to analyze customer feedback and identify specific areas for improvement. A developer has built a tool that collects customer reviews from the online marketplace and stores them in an Amazon S3 bucket. This process yields a dataset of 40 reviews. A data scientist building the ML models must identify additional sources of data to increase the size of the dataset.
Which data sources should the data scientist use to augment the dataset of reviews? (Choose three.)

A. Emails exchanged by customers and the company's customer service agents
B. Social media posts containing the name of the company or its products
C. A publicly available collection of news articles
D. A publicly available collection of customer reviews
E. Product sales revenue figures for the company
F. Instruction manuals for the company's products

Answer: BDF

NEW QUESTION 228
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 229
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 231
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 233
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 235
A Machine Learning Specialist must build out a process to query a dataset on Amazon S3 using Amazon Athena The dataset contains more than 800.000 records stored as plaintext CSV files Each record contains 200 columns and is approximately 1 5 MB in size Most queries will span 5 to 10 columns only
How should the Machine Learning Specialist transform the dataset to minimize query runtime?

A. Convert the records to Apache Parquet format
B. Convert the records to JSON format
C. Convert the records to GZIP CSV format
D. Convert the records to XML format

Answer: A

Explanation:
Using compressions will reduce the amount of data scanned by Amazon Athena, and also reduce your S3 bucket storage. It's a Win-Win for your AWS bill.
Supported formats: GZIP, LZO, SNAPPY (Parquet) and ZLIB.

NEW QUESTION 237
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 241
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 243
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 246
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 251
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 255
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 257
......

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