Exam : **MLS-C01**

Title : AWS Certified Machine Learning - Specialty

Vendor : Amazon

Version : V13.25

NO.1 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? (Choose 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
Explanation:
1- Apply words stemming and lemmatization
2- Remove Stop words
3- Tokensize the sentences
https://towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925

NO.2 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? (Choose two.)

A. AWS CloudTrail

B. AWS Health

C. AWS Trusted Advisor

D. Amazon CloudWatch

E. AWS Config

Answer: AD
Explanation:
https://aws.amazon.com/sagemaker/faqs/

NO.3 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:

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

Based on the model evaluation results, why is this a viable model for production?

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

NO.4 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 segments. Find similar profiles on social media.

D. Use a decision tree classifier engine on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media.

Answer: B

Explanation:

https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html

https://docs.aws.amazon.com/sagemaker/latest/dg/algo-kmeans-tech-notes.html

NO.5 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 VPC. Use the SageMaker DeepAR algorithm to randomize the credit card

numbers.

B. 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.

C. Use an Amazon SageMaker launch configuration to encrypt the data once it is copied to the SageMaker instance in a VPC. Use the SageMaker principal component analysis (PCA) algorithm to reduce the length of the credit card numbers.

D. 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

Explanation:

https://docs.aws.amazon.com/sagemaker/latest/dg/pca.html

NO.6 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

NO.7 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 log Amazon SageMaker API calls to Amazon S3. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.

B. Use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.

C. Implement an AWS Lambda function to log Amazon SageMaker API calls to AWS CloudTrail. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.

D. 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: B

Explanation:

Log Amazon SageMaker API Calls with AWS CloudTrail

https://docs.aws.amazon.com/sagemaker/latest/dg/logging-using-cloudtrail.html

NO.8 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: D
Explanation:
https://towardsdatascience.com/deep-learning-personal-notes-part-1-lesson-2-8946fe970b95

NO.9 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.
https://www.cloudforecast.io/blog/using-parquet-on-athena-to-save-money-on-aws/

NO.10 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 these 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
Explanation:

The AWS Glue Data Catalog is your persistent metadata store. It is a managed service that lets you store, annotate, and share metadata in the AWS Cloud in the same way you would in an Apache Hive metastore.
The Data Catalog is a drop-in replacement for the Apache Hive Metastore
https://docs.aws.amazon.com/zh_tw/glue/latest/dg/components-overview.html

NO.11 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. AWS DMS
B. Amazon Kinesis Data Streams
C. Amazon Kinesis Data Firehose
D. Amazon Kinesis Data Analytics
Answer: C

NO.12 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
Explanation:
https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html

NO.13 A Machine Learning Specialist is packaging a custom ResNet model into a Docker container so the company can leverage Amazon SageMaker for training. The Specialist is using Amazon EC2 P3 instances to train the model and needs to properly configure the Docker container to leverage the NVIDIA GPUs.
What does the Specialist need to do?
A. Bundle the NVIDIA drivers with the Docker image.
B. Build the Docker container to be NVIDIA-Docker compatible.
C. Organize the Docker container's file structure to execute on GPU instances.
D. Set the GPU flag in the Amazon SageMaker CreateTrainingJob request body.
Answer: B
Explanation:
https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf
If you plan to use GPU devices, make sure that your containers are nvidia-docker compatible.
Only the CUDA toolkit should be included on containers. Don't bundle NVIDIA drivers with the image.
For more information about nvidia-docker, see NVIDIA/nvidia-docker.

NO.14 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

Explanation:
https://medium.com/data-science-bootcamp/understand-the-softmax-function-in-minutes-f3a59641e86d

NO.15 A Data Scientist is working on optimizing a model during the training process by varying multiple parameters. The Data Scientist observes that, during multiple runs with identical parameters, the loss function converges to different, yet stable, values.
What should the Data Scientist do to improve the training process?

A. Increase the learning rate. Keep the batch size the same.

B. Reduce the batch size. Decrease the learning rate.

C. Keep the batch size the same. Decrease the learning rate.

D. Do not change the learning rate. Increase the batch size.

Answer: B

Explanation:
It is most likely that the loss function is very curvy and has multiple local minima where the training is getting stuck. Decreasing the batch size would help the Data Scientist stochastically get out of the local minima saddles. Decreasing the learning rate would prevent overshooting the global loss function minimum.

NO.16 A large consumer goods manufacturer has the following products on sale:
- 34 different toothpaste variants
- 48 different toothbrush variants
- 43 different mouthwash variants

The entire sales history of all these products is available in Amazon S3. Currently, the company is using custom-built autoregressive integrated moving average (ARIMA) models to forecast demand for these products. The company wants to predict the demand for a new product that will soon be launched.
Which solution should a Machine Learning Specialist apply?

A. Train a custom ARIMA model to forecast demand for the new product.

B. Train an Amazon SageMaker DeepAR algorithm to forecast demand for the new product.

C. Train an Amazon SageMaker k-means clustering algorithm to forecast demand for the new product.

D. Train a custom XGBoost model to forecast demand for the new product.

Answer: B

Explanation:

The Amazon SageMaker DeepAR forecasting algorithm is a supervised learning algorithm for forecasting scalar (one-dimensional) time series using recurrent neural networks (RNN). Classical forecasting methods, such as autoregressive integrated moving average (ARIMA) or exponential smoothing (ETS), fit a single model to each individual time series. They then use that model to extrapolate the time series into the future.

https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html

NO.17 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? (Choose 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: BE

Explanation:

The model must have been overfitted. Regularization helps to solve the overfitting problem in machine learning (as well as data augmentation).

NO.18 A Machine Learning team has several large CSV datasets in Amazon S3. Historically, models built with the Amazon SageMaker Linear Learner algorithm have taken hours to train on similar- sized datasets. The team's leaders need to accelerate the training process.

What can a Machine Learning Specialist do to address this concern?

A. Use Amazon SageMaker Pipe mode.

B. Use Amazon Machine Learning to train the models.

C. Use Amazon Kinesis to stream the data to Amazon SageMaker.

D. Use AWS Glue to transform the CSV dataset to the JSON format.

Answer: A

Explanation:

Amazon SageMaker Pipe mode streams the data directly to the container, which improves the performance of training jobs. In Pipe mode, your training job streams data directly from Amazon S3. Streaming can provide faster start times for training jobs and better throughput. With Pipe mode, you also reduce the size of the Amazon EBS volumes for your training instances. B would not apply in this scenario. C is a streaming ingestion solution, but is not applicable in this scenario. D transforms the data structure

NO.19 A Machine Learning Specialist is building a logistic regression model that will predict whether

or not a person will order a pizza. The Specialist is trying to build the optimal model with an ideal classification threshold.
What model evaluation technique should the Specialist use to understand how different classification thresholds will impact the model's performance?

A. Receiver operating characteristic (ROC) curve

B. Misclassification rate

C. Root Mean Square Error (RMSE)

D. L1 norm

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

NO.20 A Machine Learning Specialist uploads a dataset to an Amazon S3 bucket protected with server- side encryption using AWS KMS.
How should the ML Specialist define the Amazon SageMaker notebook instance so it can read the same dataset from Amazon S3?

A. Define security group(s) to allow all HTTP inbound/outbound traffic and assign those security group(s) to the Amazon SageMaker notebook instance.

B. onfigure the Amazon SageMaker notebook instance to have access to the VPC. Grant permission in the KMS key policy to the notebook's KMS role.

C. Assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. Grant permission in the KMS key policy to that role.

D. Assign the same KMS key used to encrypt data in Amazon S3 to the Amazon SageMaker notebook instance.

Answer: D
Explanation:
https://docs.aws.amazon.com/sagemaker/latest/dg/encryption-at-rest.html

NO.21 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.

https://aws.amazon.com/blogs/big-data/building-a-recommendation-engine-with-spark-ml-on-amazon-emr-using-zeppelin/

NO.22 A retail chain has been ingesting purchasing records from its network of 20,000 stores to Amazon S3 using Amazon Kinesis Data Firehose. To support training an improved machine learning model, training records will require new but simple transformations, and some attributes will be combined. The model needs to be retrained daily.
Given the large number of stores and the legacy data ingestion, which change will require the LEAST amount of development effort?

A. Require that the stores to switch to capturing their data locally on AWS Storage Gateway for loading into Amazon S3, then use AWS Glue to do the transformation.

B. Deploy an Amazon EMR cluster running Apache Spark with the transformation logic, and have the cluster run each day on the accumulating records in Amazon S3, outputting new/transformed records to Amazon S3.

C. Spin up a fleet of Amazon EC2 instances with the transformation logic, have them transform the data records accumulating on Amazon S3, and output the transformed records to Amazon S3.

D. Insert an Amazon Kinesis Data Analytics stream downstream of the Kinesis Data Firehose stream that transforms raw record attributes into simple transformed values using SQL.

Answer: D

NO.23 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 code. Change the machine to one with a more powerful GPU to speed up the training.

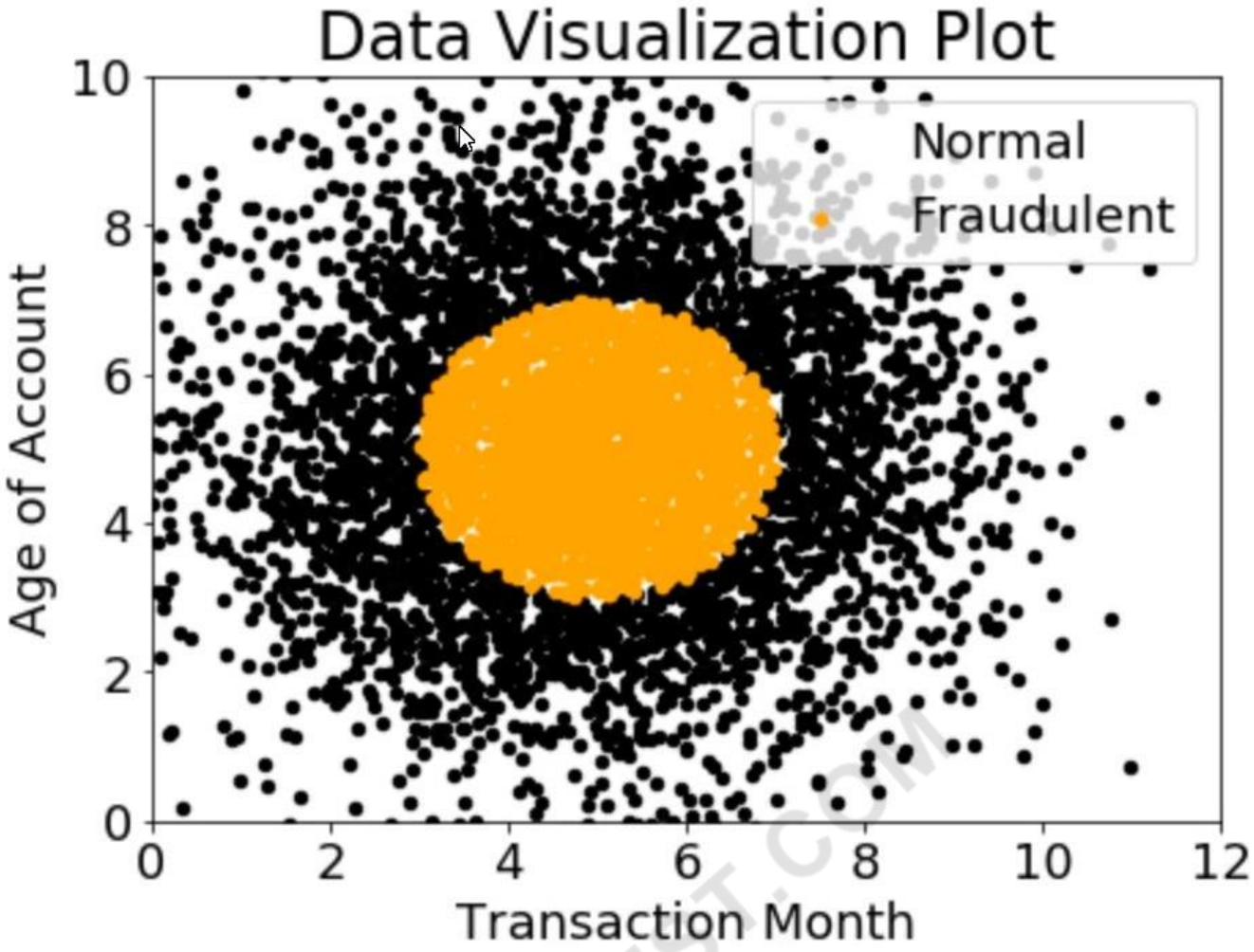
B. Change the TensorFlow code to implement a Horovod distributed framework supported by Amazon SageMaker. Parallelize the training to as many machines as needed to achieve the business goals.

C. Switch to using a built-in AWS SageMaker DeepAR model. Parallelize the training to as many machines as needed to achieve the business goals.

D. Move the training to Amazon EMR and distribute the workload to as many machines as needed to achieve the business goals.

Answer: B

NO.24 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 (SELU)

B. Logistic regression

C. Support vector machine (SVM) with non-linear kernel

D. Single perceptron with tanh activation function

Answer: C

Explanation:

https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html

NO.25 A Machine Learning Specialist is building a model that will perform time series forecasting using Amazon SageMaker. The Specialist has finished training the model and is now planning to perform load testing on the endpoint so they can configure Auto Scaling for the model variant. Which approach will allow the Specialist to review the latency, memory utilization, and CPU utilization during the load test?

A. Review SageMaker logs that have been written to Amazon S3 by leveraging Amazon Athena and Amazon QuickSight to visualize logs as they are being produced.

B. Generate an Amazon CloudWatch dashboard to create a single view for the latency, memory utilization, and CPU utilization metrics that are outputted by Amazon SageMaker.

C. Build custom Amazon CloudWatch Logs and then leverage Amazon ES and Kibana to query and visualize the log data as it is generated by Amazon SageMaker.

D. Send Amazon CloudWatch Logs that were generated by Amazon SageMaker to Amazon ES and use Kibana to query and visualize the log data

Answer: B

Explanation:

https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-cloudwatch.html

NO.26 A Data Scientist is evaluating different binary classification models. A false positive result is 5 times more expensive (from a business perspective) than a false negative result.

The models should be evaluated based on the following criteria:

1) Must have a recall rate of at least 80%

2) Must have a false positive rate of 10% or less

3) Must minimize business costs

After creating each binary classification model, the Data Scientist generates the corresponding confusion matrix.

Which confusion matrix represents the model that satisfies the requirements?

A. TN = 91, FP = 9

FN = 22, TP = 78

B. TN = 99, FP = 1

FN = 21, TP = 79

C. TN = 96, FP = 4

FN = 10, TP = 90

D. TN = 98, FP = 2

FN = 18, TP = 82

Answer: D

Explanation:

The following calculations are required:

TP = True Positive

FP = False Positive

FN = False Negative

TN = True Negative

FN = False Negative

Recall = TP / (TP + FN)

False Positive Rate (FPR) = FP / (FP + TN)

Cost = 5 * FP + FN

	A	B	C	D
Recall	78 / (78 + 22) = 0.78	79 / (79 + 21) = 0.79	90 / (90 + 10) = 0.9	82 / (82 + 18) = 0.82
False Positive Rate	9 / (9 + 91) = 0.09	1 / (1 + 99) = 0.01	4 / (4 + 96) = 0.04	2 / (2 + 98) = 0.02
Costs	5 * 9 + 22 = 67	5 * 1 + 21 = 26	5 * 4 + 10 = 30	5 * 2 + 18 = 28

Options C and D have a recall greater than 80% and an FPR less than 10%, but D is the most cost effective.

NO.27 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.

NO.28 A Machine Learning Specialist is developing a daily ETL workflow containing multiple ETL jobs. The workflow consists of the following processes:
- Start the workflow as soon as data is uploaded to Amazon S3.
- When all the datasets are available in Amazon S3, start an ETL job to join the uploaded datasets with multiple terabyte-sized datasets already stored in Amazon S3.
- Store the results of joining datasets in Amazon S3.
- If one of the jobs fails, send a notification to the Administrator.
Which configuration will meet these requirements?
A. Use AWS Lambda to trigger an AWS Step Functions workflow to wait for dataset uploads to complete in Amazon S3. Use AWS Glue to join the datasets. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.
B. Develop the ETL workflow using AWS Lambda to start an Amazon SageMaker notebook instance. Use a lifecycle configuration script to join the datasets and persist the results in Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.
C. Develop the ETL workflow using AWS Batch to trigger the start of ETL jobs when data is uploaded to Amazon S3. Use AWS Glue to join the datasets in Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.
D. Use AWS Lambda to chain other Lambda functions to read and join the datasets in Amazon S3 as soon as the data is uploaded to Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.
Answer: A
Explanation:
https://aws.amazon.com/step-functions/use-cases/

NO.29 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

Explanation:

https://docs.aws.amazon.com/AmazonS3/latest/dev/example-bucket-policies-vpc-endpoint.html

NO.30 A Data Scientist is working on an application that performs sentiment analysis. The validation accuracy is poor, and the Data Scientist thinks that the cause may be a rich vocabulary and a low average frequency of words in the dataset.

Which tool should be used to improve the validation accuracy?

A. Amazon Comprehend syntax analysis and entity detection

B. Amazon SageMaker BlazingText cbow mode

C. Natural Language Toolkit (NLTK) stemming and stop word removal

D. Scikit-leam term frequency-inverse document frequency (TF-IDF) vectorizer

Answer: D

Explanation:

https://monkeylearn.com/sentiment-analysis/

NO.31 A Machine Learning Specialist is developing a custom video recommendation model for an application. The dataset used to train this model is very large with millions of data points and is hosted in an Amazon S3 bucket. The Specialist wants to avoid loading all of this data onto an Amazon SageMaker notebook instance because it would take hours to move and will exceed the attached 5 GB Amazon EBS volume on the notebook instance.

Which approach allows the Specialist to use all the data to train the model?

A. Load a smaller subset of the data into the SageMaker notebook and train locally. Confirm that the training code is executing and the model parameters seem reasonable. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.

B. Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to the instance. Train on a small amount of the data to verify the training code and hyperparameters. Go back to Amazon SageMaker and train using the full dataset

C. Use AWS Glue to train a model using a small subset of the data to confirm that the data will be compatible with Amazon SageMaker. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.

D. Load a smaller subset of the data into the SageMaker notebook and train locally. Confirm that the training code is executing and the model parameters seem reasonable. Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to train the full dataset.

Answer: A

NO.32 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 999,000 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?
(Choose two.)

A. Add more deep trees to the random forest to enable the model to learn more features.

B. Include a copy of the samples in the test dataset 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: BD

NO.33 An office security agency conducted a successful pilot using 100 cameras installed at key locations within the main office. Images from the cameras were uploaded to Amazon S3 and tagged using Amazon Rekognition, and the results were stored in Amazon ES. The agency is now looking to expand the pilot into a full production system using thousands of video cameras in its office locations globally. The goal is to identify activities performed by non-employees in real time Which solution should the agency consider?

A. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection of known employees, and alert when non-employees are detected.

B. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Image to detect faces from a collection of known employees and alert when non-employees are detected.

C. Install AWS DeepLens cameras and use the DeepLens_Kinesis_Video module to stream video to Amazon Kinesis Video Streams for each camera. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection on each stream, and alert when non-employees are detected.

D. Install AWS DeepLens cameras and use the DeepLens_Kinesis_Video module to stream video to Amazon Kinesis Video Streams for each camera. 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

Explanation:
https://aws.amazon.com/blogs/machine-learning/video-analytics-in-the-cloud-and-at-the-edge- with -aws-deeplens-and-kinesis-video-streams/

NO.34 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

NO.35 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 vectors that store edit distance with every other word.

D. Download word embeddings pre-trained on a large corpus.

Answer: D

Explanation:

As it is a interactive online dictionary, we need pre-trained word embedding thus the answer is D. In addition, there is no mention that the online dictonary is unique and does not have a pre- trained word embedding.

NO.36 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

NO.37 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: C

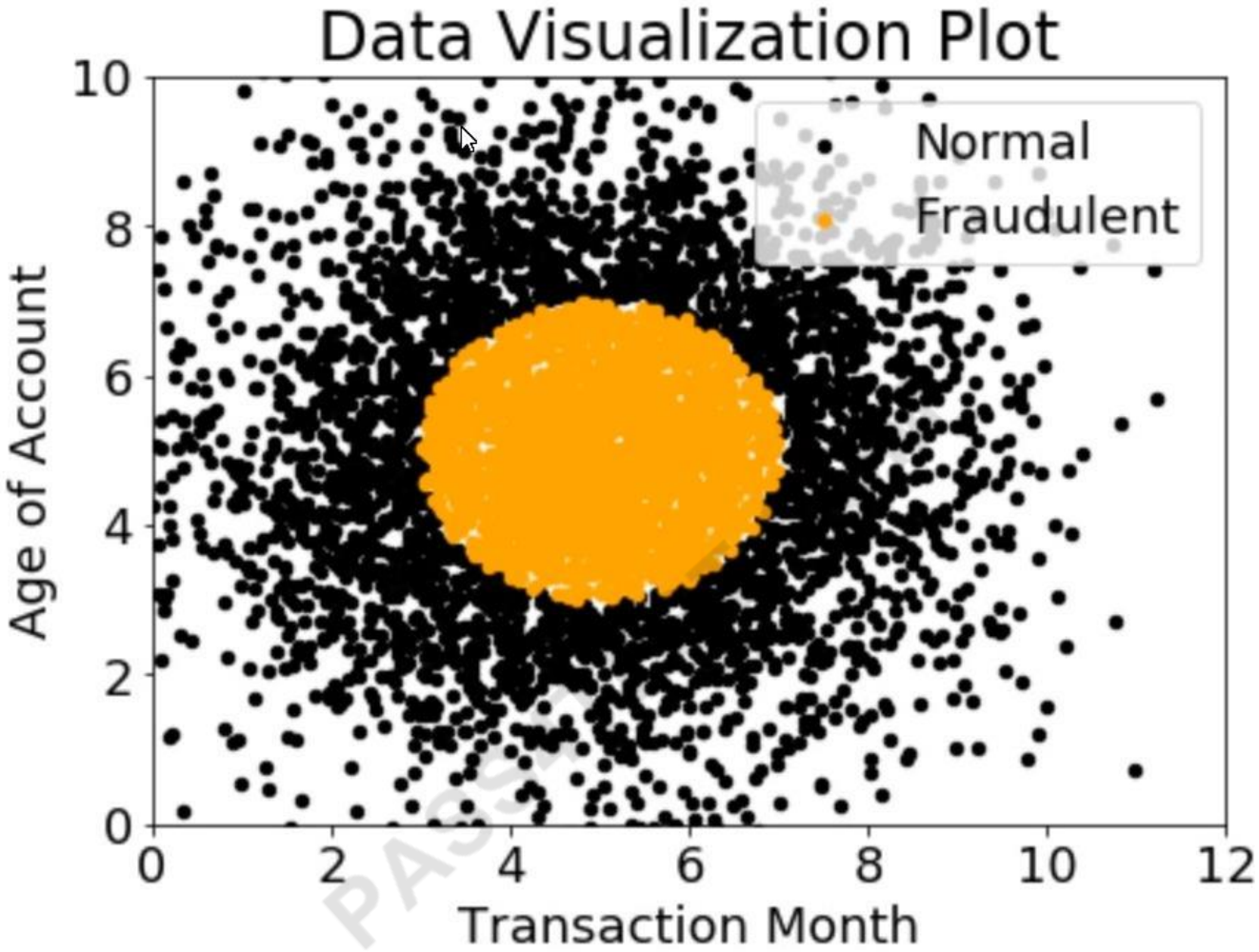
NO.38 Machine Learning Specialist is building a model to predict future employment rates based on a wide range of economic factors. While exploring the data, the Specialist notices that the magnitude of the input features vary greatly. The Specialist does not want variables with a larger magnitude to dominate the model.
What should the Specialist do to prepare the data for model training?
A. Apply quantile binning to group the data into categorical bins to keep any relationships in the data by replacing the magnitude with distribution.
B. Apply the Cartesian product transformation to create new combinations of fields that are independent of the magnitude.
C. Apply normalization to ensure each field will have a mean of 0 and a variance of 1 to remove any significant magnitude.
D. Apply the orthogonal sparse bigram (OSB) transformation to apply a fixed-size sliding window to generate new features of a similar magnitude.
Answer: C
Explanation:
https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-reference.html

NO.39 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 format. 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.
B. 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.
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 inserts it into an Amazon RDS PostgreSQL database. Have the Analysts query and run dashboards from the RDS database.
D. 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

NO.40 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: A

Explanation:

To identity fraudulent or not, you need a model with high accuracy and high recall (low false negative). The data in the graph is non-linear. Generally Decision tree gives better result for non-linear data than Naive Bayes classifier.

https://datascience.stackexchange.com/questions/6787/are-decision-tree-algorithms-linear-or-

nonlinear
https://sebastianraschka.com/Articles/2014_naive_bayes_1.html

NO.41 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 service accounts.
D. Amazon SageMaker notebook instances are based on AWS ECS instances running within AWS service accounts.
Answer: C
Explanation:
https://docs.aws.amazon.com/sagemaker/latest/dg/gs-setup-working-env.html

NO.42 When submitting Amazon SageMaker training jobs using one of the built-in algorithms, which common parameters MUST be specified? (Choose three.)
A. The training channel identifying the location of training data on an Amazon S3 bucket.
B. The validation channel identifying the location of validation data on an Amazon S3 bucket.
C. The IAM role that Amazon SageMaker can assume to perform tasks on behalf of the users.
D. Hyperparameters in a JSON array as documented for the algorithm used.
E. The Amazon EC2 instance class specifying whether training will be run using CPU or GPU.
F. The output path specifying where on an Amazon S3 bucket the trained model will persist.
Answer: AEF

NO.43 Machine Learning Specialist is training a model to identify the make and model of vehicles in images. The Specialist wants to use transfer learning and an existing model trained on images of general objects. The Specialist collated a large custom dataset of pictures containing different vehicle makes and models.
What should the Specialist do to initialize the model to re-train it with the custom data?
A. Initialize the model with random weights in all layers including the last fully connected layer.
B. Initialize the model with pre-trained weights in all layers and replace the last fully connected layer.
C. Initialize the model with random weights in all layers and replace the last fully connected layer.
D. Initialize the model with pre-trained weights in all layers including the last fully connected layer.
Answer: B

NO.44 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? (Choose two.)
A. Customize the built-in image classification algorithm to use Inception and use this for model training.
B. Create a support case with the SageMaker team to change the default image classification algorithm to Inception.
C. Bundle a Docker container with TensorFlow Estimator loaded with an Inception network and use this for model training.
D. Use custom code in Amazon SageMaker with TensorFlow Estimator to load the model with an Inception network, and use this for model training.
E. Download and apt-get install the inception network code into an Amazon EC2 instance and use this instance as a Jupyter notebook in Amazon SageMaker.
Answer: CD

NO.45 A Machine Learning Specialist working for an online fashion company wants to build a data ingestion solution for the company's Amazon S3-based data lake.
The Specialist wants to create a set of ingestion mechanisms that will enable future capabilities comprised of:
- Real-time analytics
- Interactive analytics of historical data
- Clickstream analytics
- Product recommendations
Which services should the Specialist use?
A. AWS Glue as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for real-time data insights; Amazon Kinesis Data Firehose for delivery to Amazon ES for clickstream analytics; Amazon EMR to generate personalized product recommendations
B. Amazon Athena as the data catalog: Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for near-real-time data insights; Amazon Kinesis Data Firehose for clickstream analytics; AWS Glue to generate personalized product recommendations
C. AWS Glue as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for historical data insights; Amazon Kinesis Data Firehose for delivery to Amazon ES for clickstream analytics; Amazon EMR to generate personalized product recommendations
D. Amazon Athena as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for historical data insights; Amazon DynamoDB streams for clickstream analytics; AWS Glue to generate personalized product recommendations
Answer: A

NO.46 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? (Choose three.)
A. Decrease regularization.
B. Increase regularization.
C. Increase dropout.
D. Decrease dropout.

E. Increase feature combinations.

F. Decrease feature combinations.

Answer: BCF

Explanation:

Feature selection: consider using fewer feature combinations, decrease n-grams size, and decrease the number of numeric attribute bins.

Increase the amount of regularization used

https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

NO.47 A Data Science team is designing a dataset repository where it will store a large amount of training data commonly used in its machine learning models. As Data Scientists may create an arbitrary number of new datasets every day, the solution has to scale automatically and be cost-effective. Also, it must be possible to explore the data using SQL.

Which storage scheme is MOST adapted to this scenario?

A. Store datasets as files in Amazon S3.

B. Store datasets as files in an Amazon EBS volume attached to an Amazon EC2 instance.

C. Store datasets as tables in a multi-node Amazon Redshift cluster.

D. Store datasets as global tables in Amazon DynamoDB.

Answer: A

NO.48 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

Explanation:

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

NO.49 A Data Scientist is developing a machine learning model to classify whether a financial transaction is fraudulent. The labeled data available for training consists of 100,000 non- fraudulent observations and 1,000 fraudulent observations.

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

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

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

A. Change the XGBoost eval_metric parameter to optimize based on rmse instead of error.

B. Increase the XGBoost scale_pos_weight parameter to adjust the balance of positive and negative weights.

C. Increase the XGBoost max_depth parameter because the model is currently underfitting the data.

D. Change the XGBoost eval_metric parameter to optimize based on AUC instead of error.

E. Decrease the XGBoost max_depth parameter because the model is currently overfitting the data.

Answer: BD

NO.50 A Machine Learning Specialist is working with a large cybersecurity company that manages security events in real time for companies around the world. The cybersecurity company wants to design a solution that will allow it to use machine learning to score malicious events as anomalies on the data as it is being ingested. The company also wants be able to save the results in its data lake for later processing and analysis.

What is the MOST efficient way to accomplish these tasks?

A. Ingest the data using Amazon Kinesis Data Firehose, and use Amazon Kinesis Data Analytics Random Cut Forest (RCF) for anomaly detection. Then use Kinesis Data Firehose to stream the results to Amazon S3.

B. Ingest the data into Apache Spark Streaming using Amazon EMR, and use Spark MLlib with k-means to perform anomaly detection. Then store the results in an Apache Hadoop Distributed File System (HDFS) using Amazon EMR with a replication factor of three as the data lake.

C. Ingest the data and store it in Amazon S3. Use AWS Batch along with the AWS Deep Learning AMIs to train a k-means model using TensorFlow on the data in Amazon S3.

D. Ingest the data and store it in Amazon S3. Have an AWS Glue job that is triggered on demand transform the new data. Then use the built-in Random Cut Forest (RCF) model within Amazon SageMaker to detect anomalies in the data.

Answer: A

Explanation:

https://aws.amazon.com/tw/blogs/machine-learning/use-the-built-in-amazon-sagemaker-random-cut-forest-algorithm-for-anomaly-detection/

NO.51 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.

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

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?

A. Binarization

B. One-hot encoding

C. Tokenization

D. Normalization transformation

Answer: B

NO.52 A company's Machine Learning Specialist needs to improve the training speed of a time-series forecasting model using TensorFlow. The training is currently implemented on a single-GPU machine and takes approximately 23 hours to complete. The training needs to be run daily.
The model accuracy is acceptable, but the company anticipates a continuous increase in the size of the training data and a need to update the model on an hourly, rather than a daily, basis. The company also wants to minimize coding effort and infrastructure changes.
What should the Machine Learning Specialist do to the training solution to allow it to scale for future demand?

A. Do not change the TensorFlow code. Change the machine to one with a more powerful GPU to speed up the training.

B. Change the TensorFlow code to implement a Horovod distributed framework supported by Amazon SageMaker. Parallelize the training to as many machines as needed to achieve the business goals.

C. Switch to using a built-in AWS SageMaker DeepAR model. Parallelize the training to as many machines as needed to achieve the business goals.

D. Move the training to Amazon EMR and distribute the workload to as many machines as needed to achieve the business goals.

Answer: B

NO.53 A retail company intends to use machine learning to categorize new products. A labeled dataset of current products was provided to the Data Science team. The dataset includes 1,200 products.

The labeled dataset has 15 features for each product such as title dimensions, weight, and price. Each product is labeled as belonging to one of six categories such as books, games, electronics, and movies.
Which model should be used for categorizing new products using the provided dataset for training?
A. AnXGBoost model where the objective parameter is set to multi:softmax
B. A deep convolutional neural network (CNN) with a softmax activation function for the last layer
C. A regression forest where the number of trees is set equal to the number of product categories
D. A DeepAR forecasting model based on a recurrent neural network (RNN)
Answer: A
Explanation:
A XGBoost multi class classification.
https://medium.com/@gabrielziegler3/multiclass-multilabel-classification-with-xgboost-66195e4d9f2d
CNN is used for image classificaiton problems.

NO.54 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
Explanation:
The Poisson distribution is used to model the number of events occurring within a given time interval .

NO.55 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 the customer 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

NO.56 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
Explanation:
The ML Specialist notices that, in over 75% of the misclassified images, the cats were held upside down by their owners.

NO.57 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 to 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 in Amazon S3 that is accessible for 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 PySpark 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 real-time 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: B
Explanation:
Kinesis Data Analytics can not directly stream the input data.

NO.58 A company is using Amazon Polly to translate plaintext documents to speech for automated

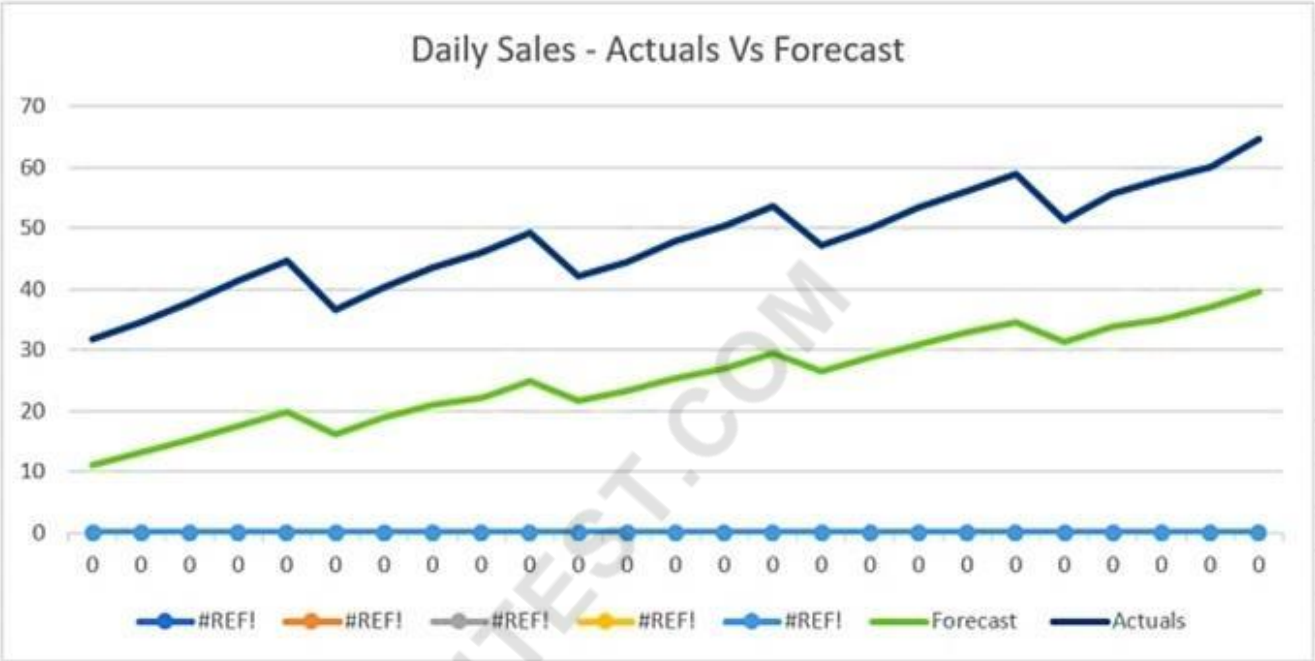
company announcements. However, company acronyms are being mispronounced in the current documents.
How should a Machine Learning Specialist address this issue for future documents?
A. Convert current documents to SSML with pronunciation tags.
B. Create an appropriate pronunciation lexicon.
C. Output speech marks to guide in pronunciation.
D. Use Amazon Lex to preprocess the text files for pronunciation
Answer: A
Explanation:
https://docs.aws.amazon.com/polly/latest/dg/ssml.html

NO.59 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?
https://www.kdnuggets.com/2019/05/churn-prediction-machine-learning.html

NO.60 A company is interested in building a fraud detection model. Currently, the Data Scientist does not have a sufficient amount of information due to the low number of fraud cases.
Which method is MOST likely to detect the GREATEST number of valid fraud cases?
A. Oversampling using bootstrapping
B. Undersampling
C. Oversampling using SMOTE
D. Class weight adjustment
Answer: C
Explanation:
With datasets that are not fully populated, the Synthetic Minority Over-sampling Technique (SMOTE) adds new information by adding synthetic data points to the minority class. This technique would be

the most effective in this scenario. Refer to Section 4.2 at this link for supporting informatio

NO.61 The displayed graph is from a forecasting 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: A

Explanation:

https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/

NO.62 An insurance company needs to automate claim compliance reviews because human reviews are expensive and error-prone. The company has a large set of claims and a compliance label for each.

Each claim consists of a few sentences in English, many of which contain complex related information. Management would like to use Amazon SageMaker built-in algorithms to design a machine learning supervised model that can be trained to read each claim and predict if the claim is compliant or not.

Which approach should be used to extract features from the claims to be used as inputs for the downstream supervised task?

A. Derive a dictionary of tokens from claims in the entire dataset. Apply one-hot encoding to tokens found in each claim of the training set. Send the derived features space as inputs to an Amazon SageMaker builtin supervised learning algorithm.

B. Apply Amazon SageMaker BlazingText in Word2Vec mode to claims in the training set. Send the derived features space as inputs for the downstream supervised task.

C. Apply Amazon SageMaker BlazingText in classification mode to labeled claims in the training set to

derive features for the claims that correspond to the compliant and non-compliant labels, respectively.

D. Apply Amazon SageMaker Object2Vec to claims in the training set. Send the derived features space as inputs for the downstream supervised task.

Answer: D

Explanation:

Amazon SageMaker Object2Vec generalizes the Word2Vec embedding technique for words to more complex objects, such as sentences and paragraphs. Since the supervised learning task is at the level of whole claims, for which there are labels, and no labels are available at the word level, Object2Vec needs be used instead of Word2Vec.

NO.63 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

Explanation:

Another benefit of using AUC is that it is classification-threshold-invariant like log loss.
https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226

NO.64 A company has collected customer comments on its products, rating them as safe or unsafe, using decision trees. The training dataset has the following features: id, date, full review, full review summary, and a binary safe/unsafe tag. During training, any data sample with missing features was dropped. In a few instances, the test set was found to be missing the full review text field.
For this use case, which is the most effective course of action to address test data samples with missing features?

A. Drop the test samples with missing full review text fields, and then run through the test set.

B. Copy the summary text fields and use them to fill in the missing full review text fields, and then run through the test set.

C. Use an algorithm that handles missing data better than decision trees.

D. Generate synthetic data to fill in the fields that are missing data, and then run through the test set.

Answer: B

Explanation:

In this case, a full review summary usually contains the most descriptive phrases of the entire review and is a valid stand-in for the missing full review text field.

NO.65 A city wants to monitor its air quality to address the consequences of air pollution. A Machine Learning Specialist needs to forecast the air quality in parts per million of contaminates for the next 2 days in the city. As this is a prototype, only daily data from the last year is available.
Which model is MOST likely to provide the best results in Amazon SageMaker?

A. Use the Amazon SageMaker k-Nearest-Neighbors (kNN) algorithm on the single time series

consisting of the full year of data with a predictor_type of regressor.

B. Use Amazon SageMaker Random Cut Forest (RCF) on the single time series consisting of the full year of data.

C. Use the Amazon SageMaker Linear Learner algorithm on the single time series consisting of the full year of data with a predictor_type of regressor.

D. Use the Amazon SageMaker Linear Learner algorithm on the single time series consisting of the full year of data with a predictor_type of classifier.

Answer: C

Explanation:

https://aws.amazon.com/blogs/machine-learning/build-a-model-to-predict-the-impact-of-weather-on-urban-air-quality-using-amazon-sagemaker/?ref=Welcome.AI

NO.66 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 time. 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 to serialize 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:

1. Use Glue Crawler to build scheme from the structured csv file.
2. Configure and run a job to transform the data from CSV to Parquet.

https://aws.amazon.com/blogs/big-data/build-a-data-lake-foundation-with-aws-glue-and-amazon-s3/

NO.67 A Data Science team within a large company uses Amazon SageMaker notebooks to access data stored in Amazon S3 buckets. The IT Security team is concerned that internet-enabled notebook instances create a security vulnerability where malicious code running on the instances could compromise data privacy. The company mandates that all instances stay within a secured VPC with no internet access, and data communication traffic must stay within the AWS network.

How should the Data Science team configure the notebook instance placement to meet these requirements?

A. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Place the Amazon SageMaker endpoint and S3 buckets within the same VPC.

B. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Use IAM policies to grant access to Amazon S3 and Amazon SageMaker.

C. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Ensure the VPC has S3 VPC endpoints and Amazon SageMaker VPC endpoints attached to it.

D. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Ensure the VPC has a NAT gateway and an associated security group allowing only outbound connections to Amazon S3 and Amazon SageMaker.

Answer: C

Explanation:

We must use the VPC endpoint (either Gateway Endpoint or Interface Endpoint)to comply with this requirement "Data communication traffic must stay within the AWS network".

https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html

NO.68 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

NO.69 A company is setting up a system to manage all of the datasets it stores in Amazon S3. The company would like to automate running transformation jobs on the data and maintaining a catalog of the metadata concerning the datasets. The solution should require the least amount of setup and maintenance.

Which solution will allow the company to achieve its goals?

A. Create an Amazon EMR cluster with Apache Hive installed. Then, create a Hive metastore and a script to run transformation jobs on a schedule.

B. Create an AWS Glue crawler to populate the AWS Glue Data Catalog. Then, author an AWS Glue ETL job, and set up a schedule for data transformation jobs.

C. Create an Amazon EMR cluster with Apache Spark installed. Then, create an Apache Hive metastore and a script to run transformation jobs on a schedule.

D. Create an AWS Data Pipeline that transforms the data. Then, create an Apache Hive metastore and a script to run transformation jobs on a schedule.

Answer: B

Explanation:

AWS Glue is the correct answer because this option requires the least amount of setup and maintenance since it is serverless, and it does not require management of the infrastructure. A, C, and D are all solutions that can solve the problem, but require more steps for configuration, and require higher operational overhead to run and maintain.

NO.70 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

Explanation:
https://worldwidescience.org/topicpages/i/imputing+missing+values.html

NO.71 A company is setting up an Amazon SageMaker environment. The corporate data security policy does not allow communication over the internet.
How can the company enable the Amazon SageMaker service without enabling direct internet access to Amazon SageMaker notebook instances?

A. Create a NAT gateway within the corporate VPC.

B. Route Amazon SageMaker traffic through an on-premises network.

C. Create Amazon SageMaker VPC interface endpoints within the corporate VPC.

D. Create VPC peering with Amazon VPC hosting Amazon SageMaker.

Answer: A

Explanation:
https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf (46)

NO.72 A term frequency-inverse document frequency (tf-idf) matrix using both unigrams and bigrams is built from a text corpus consisting of the following two sentences:
1. Please call the number below.
2. Please do not call us.
What are the dimensions of the tf-idf matrix?

A. (2, 16)

B. (2, 8)

C. (2, 10)

D. (8, 10)

Answer: A

Explanation:
There are 2 sentences, 8 unique unigrams, and 8 unique bigrams, so the result would be (2,16).
The phrases are "Please call the number below" and "Please do not call us." Each word individually (unigram) is "Please," "call," "the," "number," "below," "do," "not," and "us." The unique bigrams are "Please call," "call the," "the number," "number below," "Please do," "do not," "not call," and "call us."

NO.73 A Machine Learning Specialist is assigned a TensorFlow project using Amazon SageMaker for training, and needs to continue working for an extended period with no Wi-Fi access.
Which approach should the Specialist use to continue working?

A. Install Python 3 and boto3 on their laptop and continue the code development using that

environment.

B. Download the TensorFlow Docker container used in Amazon SageMaker from GitHub to their local environment, and use the Amazon SageMaker Python SDK to test the code.

C. Download TensorFlow from tensorflow.org to emulate the TensorFlow kernel in the SageMaker environment.

D. Download the SageMaker notebook to their local environment, then install Jupyter Notebooks on their laptop and continue the development in a local notebook.

Answer: B

Explanation:

https://aws.amazon.com/blogs/machine-learning/use-the-amazon-sagemaker-local-mode-to-train-on-your-notebook-instance/

NO.74 A Machine Learning Specialist is preparing data for training on Amazon SageMaker. The Specialist is using one of the SageMaker built-in algorithms for the training. The dataset is stored in .CSV format and 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

NO.75 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 t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the large number of input variables in an easier-to-read dimension.

C. A scatter plot showing the performance of the objective metric over each training iteration.

D. A scatter plot showing the correlation between maximum tree depth and the objective metric.

Answer: B

Explanation:

https://medium.com/all-things-ai/in-depth-parameter-tuning-for-random-forest-d67bb7e920d

NO.76 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: B

NO.77 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
Explanation:
https://aws.amazon.com/getting-started/hands-on/analyze-sentiment-comprehend/

NO.78 A Machine Learning Engineer is preparing a data frame for a supervised learning task with the Amazon SageMaker Linear Learner algorithm. The ML Engineer notices the target label classes are highly imbalanced and multiple feature columns contain missing values. The proportion of missing values across the entire data frame is less than 5%.
What should the ML Engineer do to minimize bias due to missing values?
A. Replace each missing value by the mean or median across non-missing values in same row.
B. Delete observations that contain missing values because these represent less than 5% of the data.
C. Replace each missing value by the mean or median across non-missing values in the same column.
D. For each feature, approximate the missing values using supervised learning based on other features.
Answer: D
Explanation:
Use supervised learning to predict missing values based on the values of other features. Different supervised learning approaches might have different performances, but any properly implemented supervised learning approach should provide the same or better approximation than mean or median approximation, as proposed in responses A and C.
Supervised learning applied to the imputation of missing values is an active field of research.

NO.79 A Data Scientist uses logistic regression to build a fraud detection model. While the model accuracy is 99%, 90% of the fraud cases are not detected by the model.
What action will definitively help the model detect more than 10% of fraud cases?
A. Using undersampling to balance the dataset
B. Decreasing the class probability threshold
C. Using regularization to reduce overfitting
D. Using oversampling to balance the dataset
Answer: B

Explanation:
Decreasing the class probability threshold makes the model more sensitive and, therefore, marks more cases as the positive class, which is fraud in this case. This will increase the likelihood of fraud detection. However, it comes at the price of lowering precision.

NO.80 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 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

NO.81 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: B
Explanation:
For k-means you should do additional derivation of feasible number of clusters which is not a trivial task.

NO.82 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

Explanation:

https://aws.amazon.com/big-data/real-time-analytics-featured-partners/

NO.83 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%, 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: B

Explanation:

https://kharshit.github.io/blog/2018/05/04/dropout-prevent-overfitting

NO.84 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: B

NO.85 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