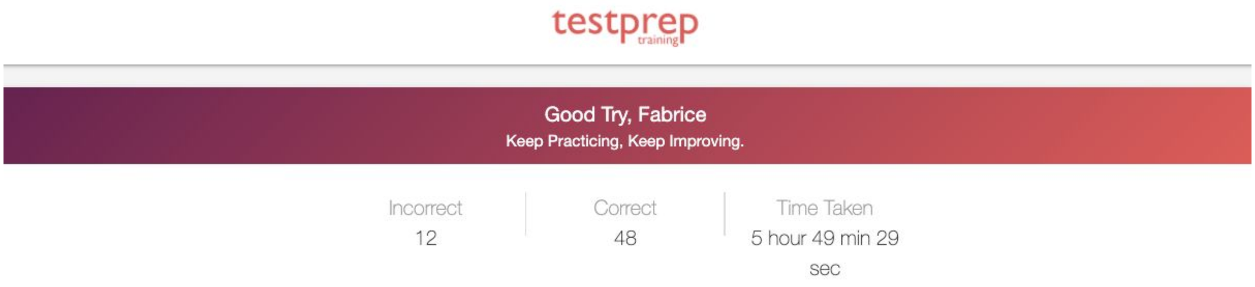
ML Specialty Practice Test - TestPrep #2

11/10/2020



testprep training

Good Try, Fabrice

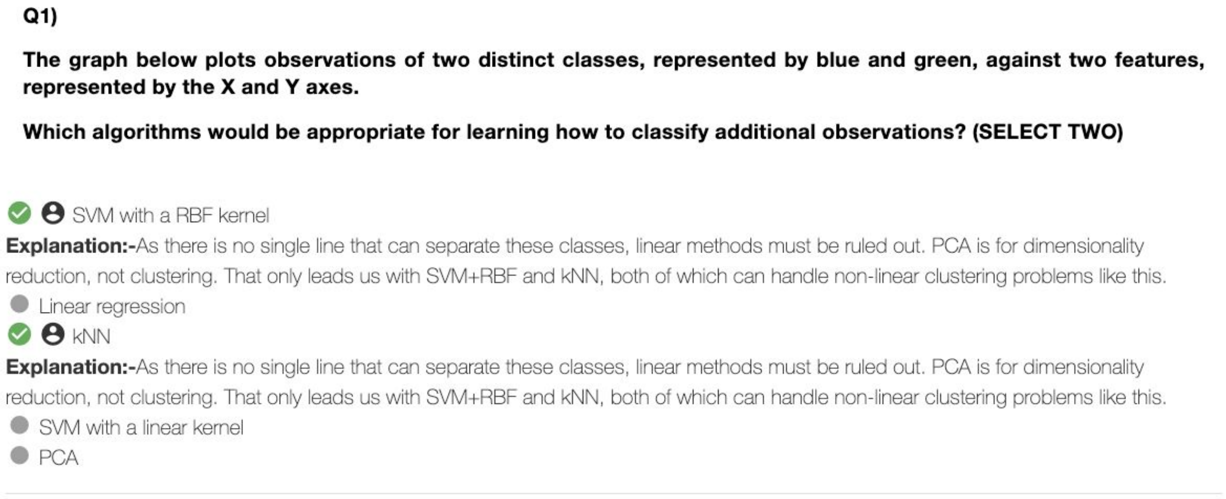
Keep Practicing, Keep Improving.

Incorrect	Correct	Time Taken
12	48	5 hour 49 min 29 sec

Score: 80%

Time: ~1h20

Except for the first 10 questions that I am pretty sure are borrowed from another online exam (cloudguru/udemy), this test had questions that were poorly written, with incorrect syntax. There were also some questions that seem to be more simply around best practice than really applied to a legit ML use case. This test does not reflect the type of AWS question we get.



Q1)

The graph below plots observations of two distinct classes, represented by blue and green, against two features, represented by the X and Y axes.

Which algorithms would be appropriate for learning how to classify additional observations? (SELECT TWO)

✓ 👤 SVM with a RBF kernel

Explanation:-As there is no single line that can separate these classes, linear methods must be ruled out. PCA is for dimensionality reduction, not clustering. That only leads us with SVM+RBF and kNN, both of which can handle non-linear clustering problems like this.

● Linear regression

✓ 👤 kNN

Explanation:-As there is no single line that can separate these classes, linear methods must be ruled out. PCA is for dimensionality reduction, not clustering. That only leads us with SVM+RBF and kNN, both of which can handle non-linear clustering problems like this.

● SVM with a linear kernel

● PCA

Q2)

You wish to use a model built with Tensorflow for training within a SageMaker notebook. To do so, you have created a Dockerfile with which you'll package your model into a SageMaker container, copying your training code with the command COPY train.py /opt/ml/code/train.py.

What further needs to be done to define your train.py as the script entrypoint?

✓ 👤 Include ENV SAGEMAKER_PROGRAM train.py in the Dockerfile

Explanation:-Expect to be tested on the details of using ECR and containers with SageMaker. The details of this question refer to the SageMaker developer guide, under "Use Your Own Algorithms or Models" / "Get Started with Containers".

● Nothing; any script inside /opt/ml/code will be considered the entrypoint automatically.

● Enter train.py as the entrypoint in the SageMaker console

● Nothing; the entrypoint must be named train.py and this is assumed.

Q3)

You want to build a "Universal Translator" that can listen to speech in a variety of languages, translate it to English, and speak the translated text back to you.

What sequence of AWS services would do this?

● Transcribe->Comprehend->Polly

✓ 👤 Transcribe->Translate->Polly

Explanation:-Transcribe could convert the speech to text in a variety of languages. Translate could translate that to English, and the resulting translated text could be spoken back with Polly.

● Translate->Polly

● Translate->Transcribe->Polly

Q4)

You have a large set of encyclopedia articles in text format, but do not have topics already assigned to each article to train with.

Which tool allows you to automatically assign topics to articles with a minimum of human effort?

● Ground Truth

● Random Cut Forest

● Amazon Translate

✓ 👤 LDA

Explanation:-Latent Dirichlet Allocation (LDA) is made for unsupervised topic modeling. SageMaker's Neural Topic Model (NTM) and the Amazon Comprehend service would also be valid choices, if they were offered. Ground Truth could also be used to label the articles, but it involves human effort.

Q5)

Your neural network is underfitting, and in response you've added more layers. Upon adding additional layers, your accuracy no longer converges successfully while training.

What is the most likely cause?

● The learning rate needs to be increased.
● The additional layers are now causing your model to over-fit.
✓ 👤 Use of a sigmoid activation function is leading to the "vanishing gradient" problem; ReLU may work better.
Explanation:-A "vanishing gradient" results from multiplying together many small derivates of the sigmoid activation function in multiple layers. ReLU does not have a small derivative, and avoids this problem.
● Too many training epochs are being used.

Q6)

You are training a linear regression model the predicts income based on age, and a few other features. The training data contains several outliers from billionaires.

How should these outliers be handled in order to maximize accuracy for the non-billionaires?

● Replace the income labels of the outliers with the mean income value.
✓ 👤 Remove the outliers prior to training, identified as being outside some mutliple of a standard deviation from the mean
Explanation:-Outliers can skew linear models. Since we explicitly said we don't care about predicting results for outliers, it's best to just discard them.
● Keep all data intact, in order to produce the most accurate model.
● Replace the income labels of the outliers with the median income value.

Q7) Which are best practices for hyperparameter tuning in SageMaker? (CHOOSE TWO)

● Run training jobs on a single instance
✓ 👤 Run only one training job at a time
Explanation:-For reference, see "Best Practices for HyperParameter Tuning" in the SageMaker developer guide at https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-considerations.html. You should understand this document; some of the recommendations surrounding choosing the best number of training jobs and how many instances to use can seem counter-intuitive.
✓ 👤 Choose the smallest possible ranges for your hyperparameters
Explanation:-For reference, see "Best Practices for HyperParameter Tuning" in the SageMaker developer guide at https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-considerations.html. You should understand this document; some of the recommendations surrounding choosing the best number of training jobs and how many instances to use can seem counter-intuitive.
● Use linear scales for hyperparameters
● Choose a large number of hyperparameters to tune

Q8)

You are building a neural network to estimate medical expenses based on a patient's blood pressure, among other attributes. Blood pressures are reported with two decimal points of precision, but must be encoded into a small number of discrete values for use in the network. Furthermore, the blood pressure values are very unevenly distributed, and we do not want to lose that distribution in the process.

Which technique could address both of these issues?

● Normalization

✓ 👤 Quantile binning

Explanation:-Quantile binning splits data into a fixed number of buckets, with the same number of observations in each bin.

● Boosting

● Interval binning

Q9)

A book publisher is ingesting a data feed into S3 containing features of various books coming from various sources. Many of these sources are sending data that is duplicated by other sources, and the S3 data lake where the data is being stored needs to be de-duplicated prior to further processing.

What mechanism could achieve this goal with minimal development effort and ongoing maintenance?

● Use a Glue Crawler to identify and eliminate duplicate records as its table structure is being inferred

● Import the data into Redshift, using a primary key that prevents duplicate records from being entered.

✓ 👤 Use AWS Glue's FindMatches ML Transform to identify and eliminate duplicate records as they are received.

Explanation:-Glue's FindMatches feature is a new way to perform de-duplication as part of Glue ETL, and is a simple, server-less solution to the problem.

● Periodically load the data into a Spark Dataframe on EMR, and use the dropDuplicates() function to remove the duplicates before passing it on for further processing.

Q10)

You are training SageMaker's supervised BlazingText using file mode.

Which is an example of a properly formatted line within the training file?

● __label__4 linux ready for prime time, intel says.

✓ 👤 __label__4 linux ready for prime time , intel says .

Explanation:-Each line of the input file contains a training sentence per line, along with their labels. Labels must be prefixed with __label__, and the tokens within the sentence - including punctuation - should be space separated.

● __label4 linux ready for prime time, intel says.

● __label__4 Linux ready for prime time, Intel says.

Q11)

A machine learning specialist is working the IT department to increase visibility into the operations of the Machine Learning infrastructure on AWS.

How could the specialist use Cloudwatch to accomplish this goal?

● Cloudwatch can create timers for AWS Lambda functions

✓ 👤 Cloudwatch could be use to enable governance, compliance and operational auditing

Explanation:-Cloudwatch could be used to enable governance, compliance and operational auditing. It can also be used to create visibility into user and resource activity and also security analysis and troubleshooting.

● None of these

● All of these

Q12)

A disaster recovery auditor has asked the machine learning specialist in a company to identify AWS services used in production that have the ability to create snapshots.

What is the best answer?

● RDS, S3, EMR

✓ 👤 RDS, S3, DynamoDB

Explanation:-RDS, S3, and DynamoDB all have the ability to take snapshots.

● RDS, S3, Amazon ElastiCache

● RDS, S3, Athena

Q13)

A machine learning specialist is looking to create a Data Lake for a Fortune 500 company on AWS. An executive has asked for a summary of the benefits.

What benefits should the machine learning specialist highlight?

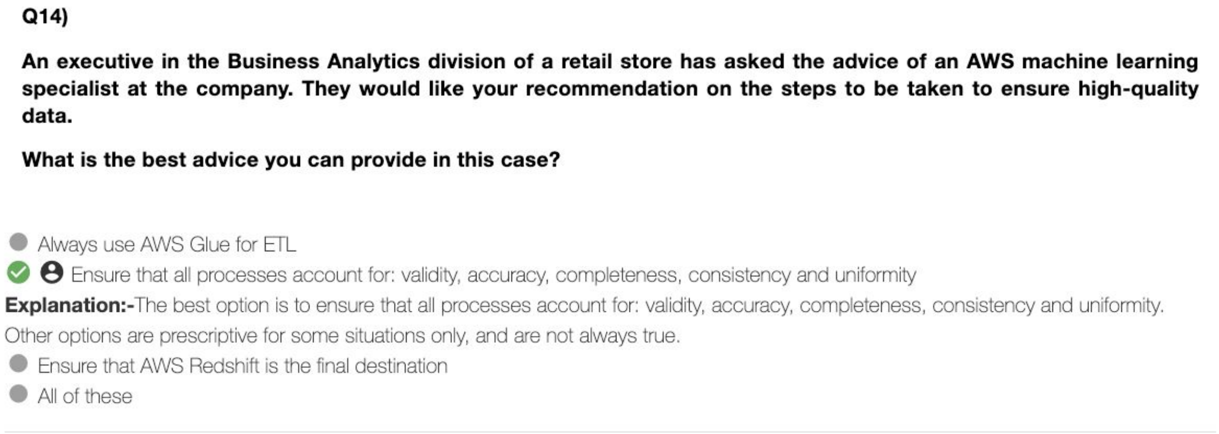
✓ 👤 All of these

Explanation:-A data lake can store structured and unstructured data, can be used for analytics and ML, and also work on data without data movement. Additionally, it is low-cost storage.

● Can store structured and unstructured data

● Used for analytics and ML

● Work on data without data movement



Q14)

An executive in the Business Analytics division of a retail store has asked the advice of an AWS machine learning specialist at the company. They would like your recommendation on the steps to be taken to ensure high-quality data.

What is the best advice you can provide in this case?

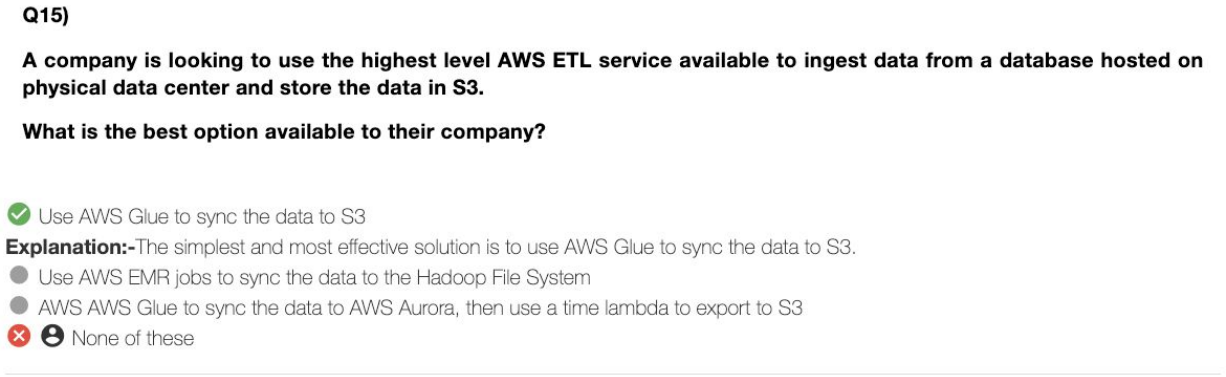
● Always use AWS Glue for ETL

✓ 👤 Ensure that all processes account for: validity, accuracy, completeness, consistency and uniformity

Explanation:-The best option is to ensure that all processes account for: validity, accuracy, completeness, consistency and uniformity. Other options are prescriptive for some situations only, and are not always true.

● Ensure that AWS Redshift is the final destination

● All of these



Q15)

A company is looking to use the highest level AWS ETL service available to ingest data from a database hosted on physical data center and store the data in S3.

What is the best option available to their company?

✓ Use AWS Glue to sync the data to S3

Explanation:-The simplest and most effective solution is to use AWS Glue to sync the data to S3.

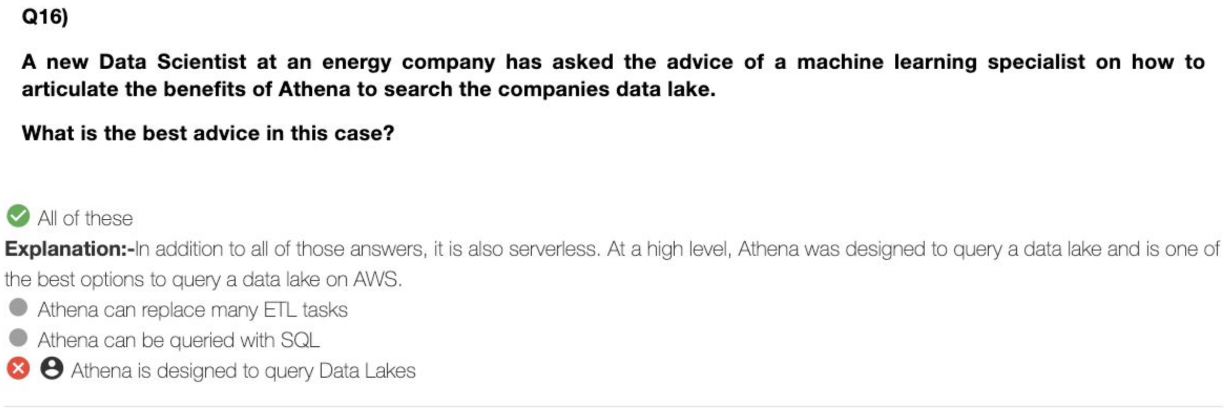
● Use AWS EMR jobs to sync the data to the Hadoop File System

● AWS AWS Glue to sync the data to AWS Aurora, then use a time lambda to export to S3

✗ 👤 None of these

https://aws.amazon.com/blogs/big-data/how-to-access-and-analyze-on-premises-data-stores-using-aws-glue/#:~:text=AWS%20Glue%20can%20also%20connect,Microsoft%20SQL%20Server%2C%20and%20MariaDB.&text=AWS%20Glue%20jobs%20extract%20data,data%20stores%20as%20a%20target.

AWS Glue can also connect to a variety of on-premises JDBC data stores such as PostgreSQL, MySQL, Oracle, Microsoft SQL Server, and MariaDB.



Q16)

A new Data Scientist at an energy company has asked the advice of a machine learning specialist on how to articulate the benefits of Athena to search the companies data lake.

What is the best advice in this case?

✓ All of these

Explanation:-In addition to all of those answers, it is also serverless. At a high level, Athena was designed to query a data lake and is one of the best options to query a data lake on AWS.

● Athena can replace many ETL tasks

● Athena can be queried with SQL

✗ 👤 Athena is designed to query Data Lakes

I would not use Athena for an ETL but you can technically extract/transformload data yes https://aws.amazon.com/blogs/big-data/extract-transform-and-load-data-into-s3-data-lake-using-ctas-and-insert-into-statements-in-amazon-athena/

Amazon Athena is an interactive query service that makes it easy to analyze the data stored in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run.
This blog post discusses how to use Athena for extract, transform and load (ETL) jobs for data processing

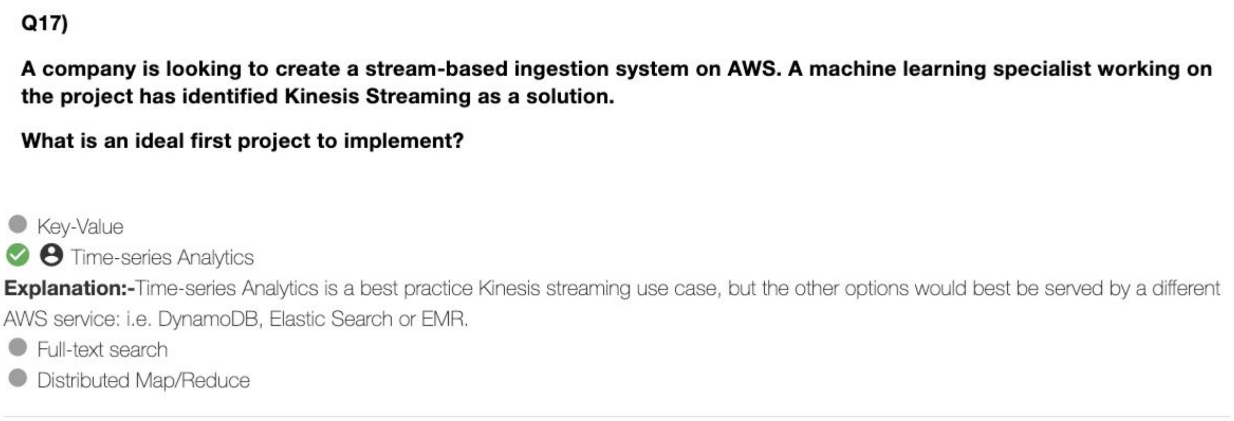
https://aws.amazon.com/blogs/big-data/simplify-etl-data-pipelines-using-amazon-athenas-federated-queries-and-user-defined-functions/
Amazon Athena recently added support for federated queries and user-defined functions (UDFs), both in Preview.

Using Amazon Athena to simplify ETL workflows and enable quicker analytics

Athena, a fully managed serverless interactive service for querying data in Amazon S3 using SQL, has been rapidly adopted by multiple departments across our organization. For our use case, we did not require an always-on EMR cluster waiting for an analytics query. Athena's serverless nature is perfect for our use case. Along the way we discovered that we could use Athena to run extract, transform, and load (ETL) jobs.

However, Athena is a lot more than an interactive service for querying data in Amazon S3. We also found Athena to be a robust, powerful, reliable, scalable, and cost-effective ETL tool. The ability to schedule SQL statements, along with support for Create Table As Select (CTAS) and INSERT INTO statements, helped us accelerate our ETL workloads.

With the addition of query federation and UDFs to Athena, Jornaya has been able to replace many of our unstable data pipelines with Athena to extract and transform data from DynamoDB and write it to Amazon S3.



Q17)

A company is looking to create a stream-based ingestion system on AWS. A machine learning specialist working on the project has identified Kinesis Streaming as a solution.

What is an ideal first project to implement?

● Key-Value

✓ ● Time-series Analytics

Explanation:-Time-series Analytics is a best practice Kinesis streaming use case, but the other options would best be served by a different AWS service: i.e. DynamoDB, Elastic Search or EMR.

● Full-text search

● Distributed Map/Reduce

Q18)

A business analyst has asked a machine learning specialist about the best architecture for an ad-hoc SQL based aggregation pipeline that can both be queried by their team and also used to feed data into an ML pipeline. The data they are querying lives in Amazon S3.

What is the best option in this case?

● Use AWS Glue and AWS DynamoDB

● Use AWS ElasticSearch and AWS Lambda

● Use AWS Sagemaker and AWS Lambda

✓ 👤 Use AWS Athena and AWS Glue

Explanation:-The best solution to support both ad-hoc querying of data via SQL and also to allow that same data to be sent to an ML pipeline would be AWS Athena and AWS Glue. AWS Athena can do ad-hoc queries and AWS Glue can do the ETL.

Q19)

A company uses AWS Batch to run large KMeans clustering jobs. The Machine Learning Specialist has been told to optimize costs.

What next step would be best suited to achieve that goal?

● Setup AWS Batch to create fine-grained access control

● Setup AWS Batch to run single jobs that span multiple EC2 instances

✓ 👤 Setup AWS Batch to dynamically bid on Spot instances

Explanation:-While all of the answers are features of AWS Batch, only "a" optimizes cost savings because AWS Spot instances can save up to 90% from on demand.

● All of these

Q20)

A data scientist has asked for help in identifying services that could be used for preparing data and cleaning data for an upcoming project.

What is the best advice?

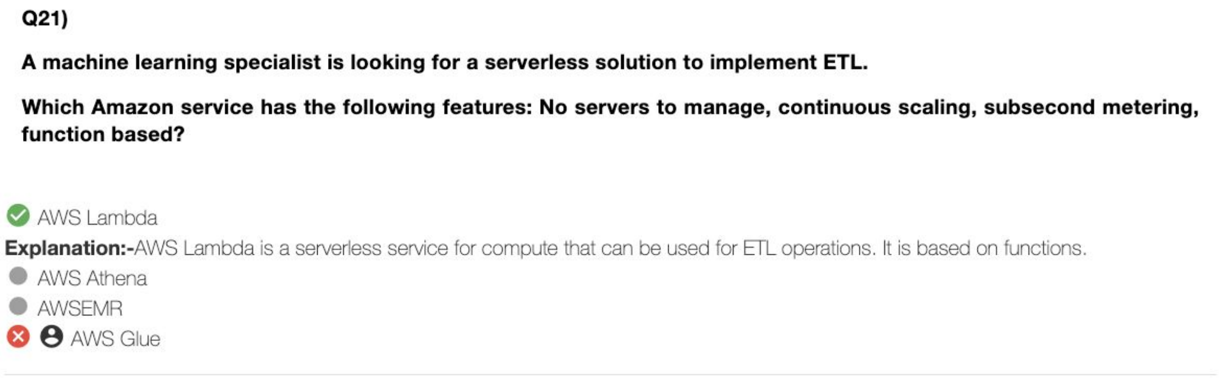
● AWS Lambda is the best solution for most ETL pipelines

● AWS Machine Learning service can be given unstructured data and it will clean it

✓ AWS Glue can be used to clean many types of data sources including S3

Explanation:-AWS Glue can be used to clean many types of data sources including S3. AWS Lambda can perform ETL tasks, but may not be suitable for some tasks that require heavy CPU or advanced automation from many sources to many destinations. AWS Machine Learning can infer schema, but in some cases, the data may need additional preprocessing steps and AWS Machine Learning would be an inappropriate service.

✗ 👤 All of these



Q21)

A machine learning specialist is looking for a serverless solution to implement ETL.

Which Amazon service has the following features: No servers to manage, continuous scaling, subsecond metering, function based?

✓ AWS Lambda

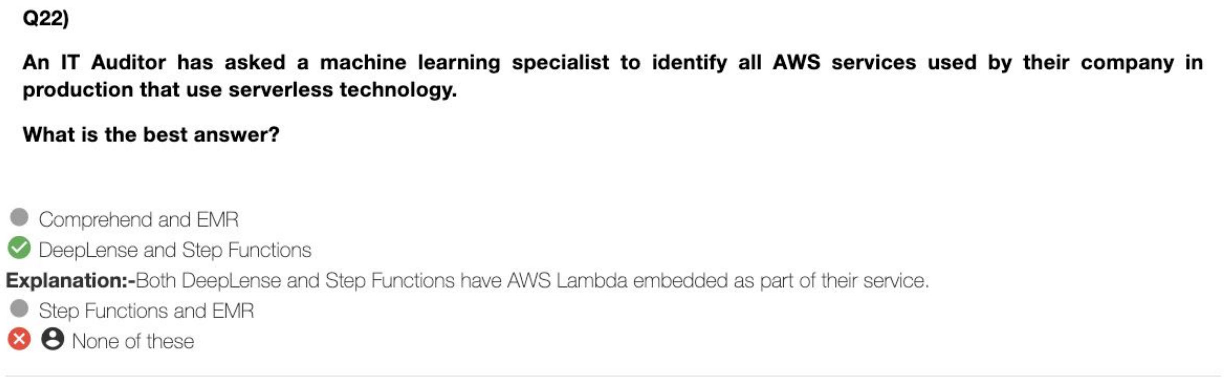
Explanation:-AWS Lambda is a serverless service for compute that can be used for ETL operations. It is based on functions.

● AWS Athena

● AWSEMR

✗ 👤 AWS Glue

I disagree- AWS Glue is ALSO serverless and has code (functiona)



Q22)

An IT Auditor has asked a machine learning specialist to identify all AWS services used by their company in production that use serverless technology.

What is the best answer?

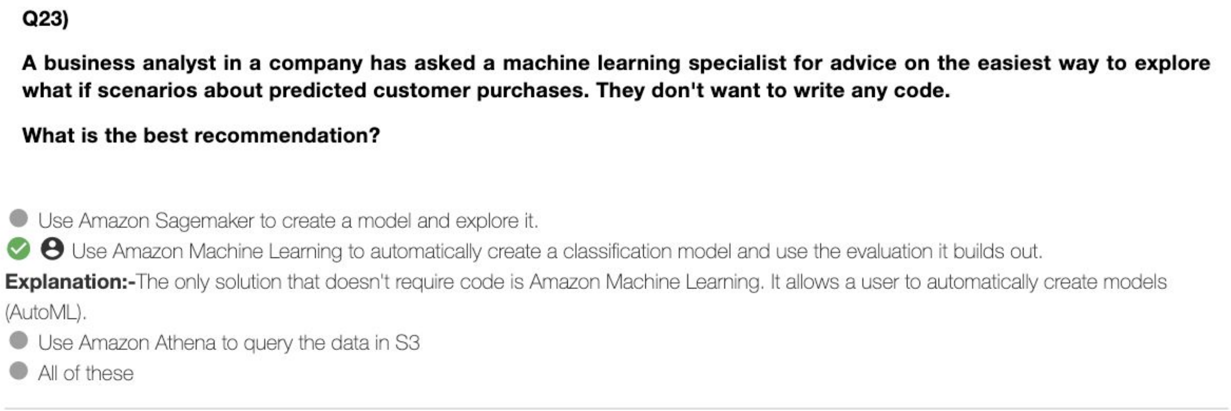
● Comprehend and EMR

✓ DeepLense and Step Functions

Explanation:-Both DeepLense and Step Functions have AWS Lambda embedded as part of their service.

● Step Functions and EMR

✗ 👤 None of these



Q23)

A business analyst in a company has asked a machine learning specialist for advice on the easiest way to explore what if scenarios about predicted customer purchases. They don't want to write any code.

What is the best recommendation?

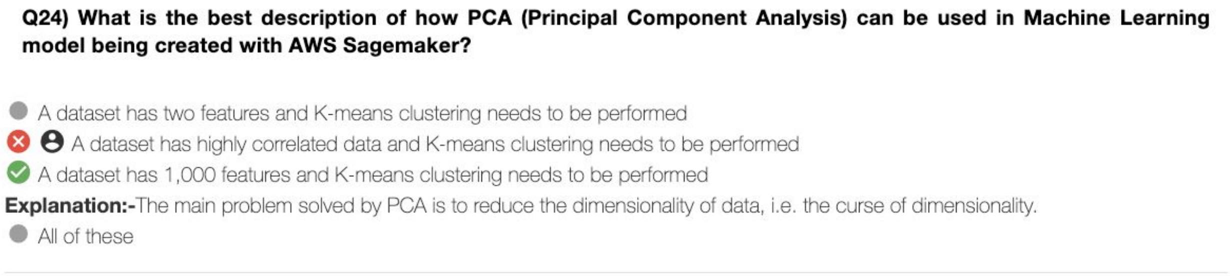
● Use Amazon Sagemaker to create a model and explore it.

✓ 👤 Use Amazon Machine Learning to automatically create a classification model and use the evaluation it builds out.

Explanation:-The only solution that doesn't require code is Amazon Machine Learning. It allows a user to automatically create models (AutoML).

● Use Amazon Athena to query the data in S3

● All of these



Q24) What is the best description of how PCA (Principal Component Analysis) can be used in Machine Learning model being created with AWS Sagemaker?

● A dataset has two features and K-means clustering needs to be performed

✗ 👤 A dataset has highly correlated data and K-means clustering needs to be performed

✓ A dataset has 1,000 features and K-means clustering needs to be performed

Explanation:-The main problem solved by PCA is to reduce the dimensionality of data, i.e. the curse of dimensionality.

● All of these

Q25)

A machine learning specialist has been asked to describe the data that is being selected for a classification model.

What is the correct response in this case?

✓ 👤 It needs to be categorical

Explanation:-Classification models require categorical data and it could be ordinal (ordered) or nominal (unordered).

● It needs to be categorical, but not nominal

● It needs to be categorical and must be ordinal

● None of these

Q26)

A machine learning specialist is writing a manual for other employees to achieve success when creating classification models. Someone has asked for clarification about the purpose of "one hot encoding".

What is the best answer?

● It is a process in which categorical variables are converted into both nominal and ordinal variables

✓ 👤 It is a process in which categorical variables are converted into a form that can be provided to ML models

Explanation:-Categorical variables are converted into a binary form, like, 1 or 0, such that ML models can work with them.

● It is a process in which categorical and numerical values are converted into a form that can be provided to ML models

● None of these

Q27) How can Heatmaps be used as part of a Machine Learning process?

● As part of data cleaning

✓ 👤 As part of feature engineering

Explanation:-Heatmaps are used in the Feature Engineering to identify potential features, i.e. columns that are correlated, and thus able to give a signal to the model,

● As part of modeling

● All of these

Q28) How could a new feature be created by looking at a histogram?

● Select median values

✓ 👤 Look at age distributions in a population and create four categories for a type of sport: Rookie, Prime, Post Prime and Pre-Retirement.

Explanation:-The best answer is that it could be used to create a new feature with the categories described in answer a.

● Dropping outliers

● All of these

Q29) What is the purpose of scaling data before it is clustered?

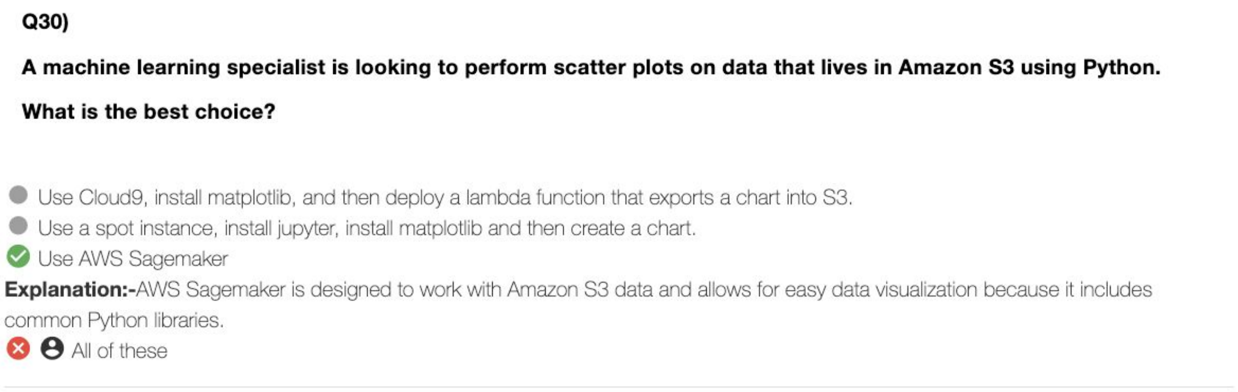
● Creates one-hot encoding

✓ 👤 Enables all variables to be treated the same

Explanation:-Scaling is important in clustering because it makes sure that all variables are within the same range, say 0 to 1. This ensures all attributes/variables have the same importance.

● Adds regularization

● All of these



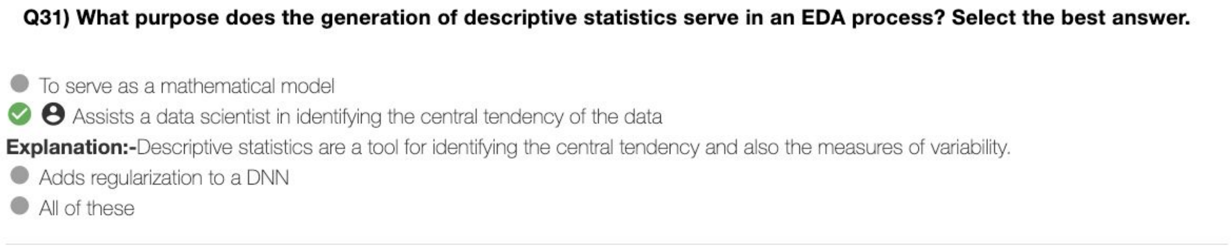
Q30)

A machine learning specialist is looking to perform scatter plots on data that lives in Amazon S3 using Python.

What is the best choice?

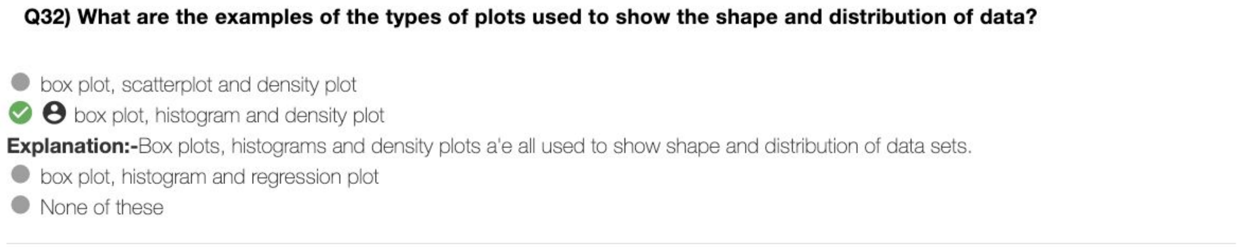
● Use Cloud9, install matplotlib, and then deploy a lambda function that exports a chart into S3.
● Use a spot instance, install jupyter, install matplotlib and then create a chart.
✓ Use AWS Sagemaker
Explanation:-AWS Sagemaker is designed to work with Amazon S3 data and allows for easy data visualization because it includes common Python libraries.
✗ 👤 All of these

Did not read the question properly as “best” choice



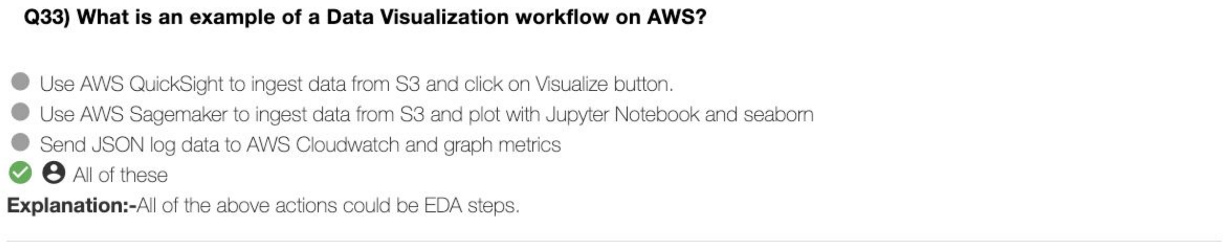
Q31) What purpose does the generation of descriptive statistics serve in an EDA process? Select the best answer.

● To serve as a mathematical model
✓ 👤 Assists a data scientist in identifying the central tendency of the data
Explanation:-Descriptive statistics are a tool for identifying the central tendency and also the measures of variability.
● Adds regularization to a DNN
● All of these



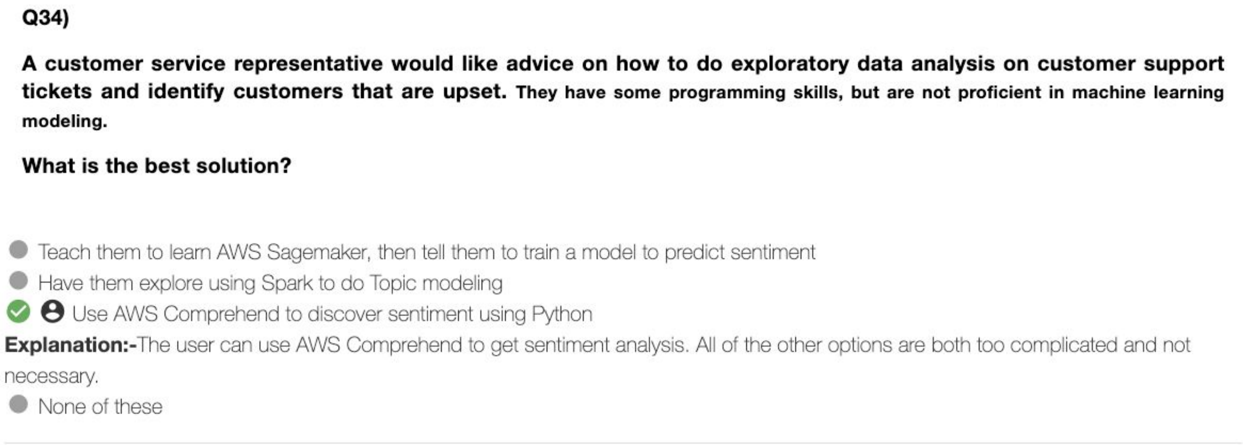
Q32) What are the examples of the types of plots used to show the shape and distribution of data?

● box plot, scatterplot and density plot
✓ 👤 box plot, histogram and density plot
Explanation:-Box plots, histograms and density plots a'e all used to show shape and distribution of data sets.
● box plot, histogram and regression plot
● None of these



Q33) What is an example of a Data Visualization workflow on AWS?

● Use AWS QuickSight to ingest data from S3 and click on Visualize button.
● Use AWS Sagemaker to ingest data from S3 and plot with Jupyter Notebook and seaborn
● Send JSON log data to AWS Cloudwatch and graph metrics
✓ 👤 All of these
Explanation:-All of the above actions could be EDA steps.



Q34)

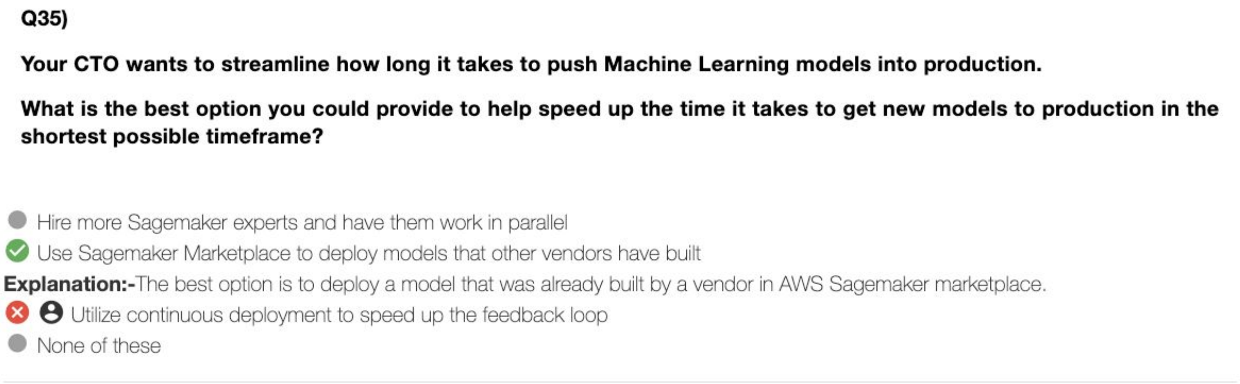
A customer service representative would like advice on how to do exploratory data analysis on customer support tickets and identify customers that are upset. They have some programming skills, but are not proficient in machine learning modeling.

What is the best solution?

- ● Teach them to learn AWS Sagemaker, then tell them to train a model to predict sentiment
- ● Have them explore using Spark to do Topic modeling
- ✓ 👤 Use AWS Comprehend to discover sentiment using Python

Explanation:-The user can use AWS Comprehend to get sentiment analysis. All of the other options are both too complicated and not necessary.

- ● None of these



Q35)

Your CTO wants to streamline how long it takes to push Machine Learning models into production.

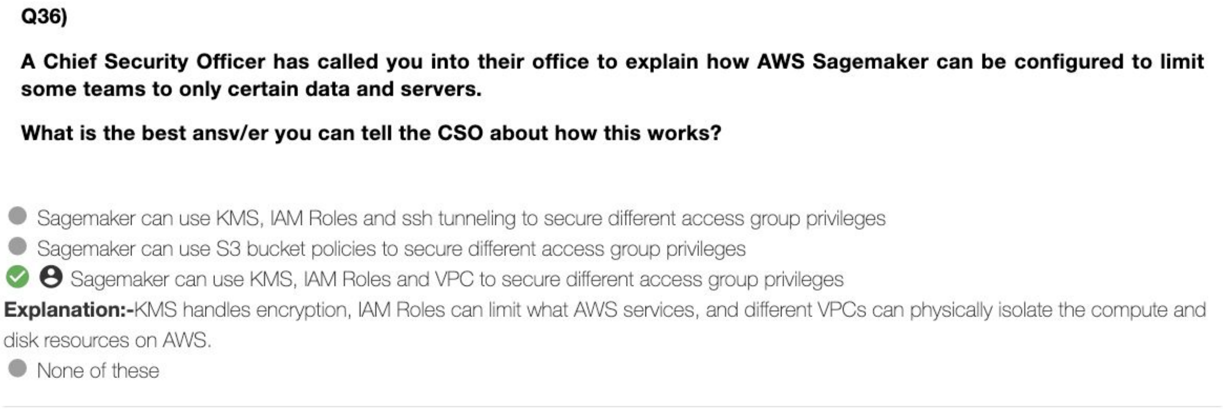
What is the best option you could provide to help speed up the time it takes to get new models to production in the shortest possible timeframe?

- ● Hire more Sagemaker experts and have them work in parallel
- ✓ Use Sagemaker Marketplace to deploy models that other vendors have built

Explanation:-The best option is to deploy a model that was already built by a vendor in AWS Sagemaker marketplace.

- ✗ 👤 Utilize continuous deployment to speed up the feedback loop
- ● None of these

I did not read the question properly with “new models” and “shortest”



Q36)

A Chief Security Officer has called you into their office to explain how AWS Sagemaker can be configured to limit some teams to only certain data and servers.

What is the best ansv/er you can tell the CSO about how this works?

- ● Sagemaker can use KMS, IAM Roles and ssh tunneling to secure different access group privileges
- ● Sagemaker can use S3 bucket policies to secure different access group privileges
- ✓ 👤 Sagemaker can use KMS, IAM Roles and VPC to secure different access group privileges

Explanation:-KMS handles encryption, IAM Roles can limit what AWS services, and different VPCs can physically isolate the compute and disk resources on AWS.

- ● None of these

Q37)

Your head of Technical Operations approaches you in a panic and tells you that the production customer credit prediction model built three months ago has suddenly denying all customers that submit a credit card application.

What could you tell him to potentially solve the problem?

● The only way to fix it is to redo the machine learning pipeline by starting over from scratch (a one month process)
✓ 👤 Retrain the model on the current customer data and redeploy
Explanation:-It is common for a model to be retrained if data has changed. This is the most efficient first approach.
● The system is under attack, and it best to shut down the website
● None of these

Q38)

The CEO of your startup has called you up at 10 PM at night to tell you that investors have been asking to understand how your Machine Learning startup can handle traffic at scale. They will not put an additional 10 million dollars into your company until you explain how it works.

What would be your response?

● The company uses static hosted websites with S3 and route 53. If traffic to prediction models gets too high, you will revert everyone to a maintenance page. This can handle infinite traffic
✓ 👤 The company uses AWS Sagemaker with autoscaling of endpoints. This ensures it can scale up and down with additional prediction traffic.
Explanation:-The only answer that makes sense is b. Sagemaker can automatically scale up to handle predictions. The other answers would result in the company losing investors as they are "janky" best practices employed by naive startups.
● The company has many certified solutions architects and they are on call to spin up more machines to create prediction if they get paged.
● All of these

Q39)

The CFO of an AI company has approached your consulting company with a request to help them lower the cost of their production machine learning systems.

What is the best piece of general advice you can give them?

● Always train models on the most powerful instances available
✓ Inference is 90% of the cost of production Machine Learning
Explanation:-Creating predictions are the most expensive aspects of machine learning and are 90% of the costs. Controlling this cost can be done with elastic interfaces via Sagemaker.
● Always train models on CPUs vs GPUs, as GPUs are more expensive
✗ 👤 None of these

Q40) Why is ECR an important part of Sagemaker marketplace pipelines?

● ECR is the other name for marketplace
● ECR and Sagemaker Marketplace provide the same functionality but one is open source and other is proprietary
✓ 👤 In order for a company to provide machine learning models to AWS Sagemaker marketplace, they must use ECR
Explanation:-To sell Sagemaker models, they must use ECR to register the container
● None of these

Q41)

A new machine learning engineer in your group has asked you to explain why Deep Learning AMIs are valuable.

What is the best answer you can provide?

✓ 👤 All of these

Explanation:-All of the answers are correct. Additionally, they are available as "Pay as you Go" and 'Spot Instances"

● There are three styles: conda, base AMI, and AMIs with source code

● They are preloaded with all of the frameworks for Deep Learning

● They can perform Multi-GPU training

Q42)

The CIO of your company has asked you to create a production readiness checklist for production machine learning deployment on AWS.

What is an example of a statement that would appear on that checklist?

✓ 👤 All of these

Explanation:-All of the answers are checklist items for deploying production ML models on AWS. Another example item would be: Do you have separate environments?

● Do you have alerts setup for prediction threshold failures?

● Are you using a simple enough model?

● Are you using a Data Lake or directly talking to a SQL database?

Q43)

A new startup has hired your consulting company to advise them as machine learning specialists for AWS. They need to quickly implement natural language processing around sentiment analysis in six months, but they do not have any experience with machine learning other than one data scientist.

What would be the most efficient way for them to solve their problem?

● Start with AWS Sagemaker and hire more data scientists

✓ 👤 Start with AWS Comprehend and move to AWS Sagemaker in the future if required

Explanation:-It is important to be able to select between off-the-shelf solutions vs build yourself. Answer b is the most effective solution.

● Advise them this timeframe is not realistic

● None of these

Q44)

A member of the DevOps team has asked for advice on how to integrate continuous delivery for a Machine Learning pipeline running on AWS with minimal effort.

As a machine learning specialist, what is the best advice you can give?

● Use Jenkins along with Opsworks

✓ 👤 Use CodePipeline to create automation for both code, containers.

Explanation:-The best and most relevant option is to use Code Pipeline.

● Use Elastic Beanstalk

● All of these

Q45)

A third-party auditor that inspecting your machine learning pipeline has asked you to explain how your company is safeguarding customer data that is trained using Sagemaker.

What do you tell the auditor that accurately describes the best practice for encryption with Sagemaker?

● Your company uses encryption algorithms designed by your lead engineer, this ensures hackers will never find out how to decrypt the data

● Your company uses Sagemaker a substitution cypher that hides social security numbers you use as a unique id

✓ 👤 Sagemaker uses Amazon KMS to encrypt all data used during the Machine Learning process

Explanation:-The best practice is to use answer A, which ensures that encryption is seamless and secure. The other choices are examples of extreme red flags that naive companies use to secure their data and would be immediately flagged by an auditor.

● All of these

Q46)

As a Deep Learning expert at your company, you have been investigating ways to minimize costs and deployment complexities by using containers.

What are the correct statements you make about available solutions on AWS that would help solve this problem?

● Only ECS, but not EKS could be used in this capacity

● EKS is a better solution because it is open source

✓ 👤 Both ECS and EKS could be used in this capacity

Explanation:-Both ECS (Elastic Container Service) and EKS (Elastic Kubernetes Service) are viable options to use in containerizing AWS ML in production.

● None of these

Q47) Which of the following statements are correct about Reinforcement Learning?

● Inspired by behavioral biology

● Learns to take action to maximize total reward

● Model learns by interacting with its environment

✓ 👤 All of these

Explanation:-All of these statements correctly describe RNN.

Q48)

A machine learning specialist has been asked about a result from a model that an AWS Machine Learning service has created for a binary classification.

What could be a correct description of AUC in the evaluation tab?

✓ 👤 All of these

Explanation:-All of the answers are correct descriptions of AUC. Additionally, it measures the quality of the model's prediction regardless of what the classification threshold has been set to.

● AUC ranges in value from 0 to 1

● An AUC model whose predictions are 100% wrong has an AUC of 0

● AUC is scale-invariant

Q49) What is a common problem when training a neural network using backpropagation?

✓ All of these

Explanation:-All of these are common problems when training a neural network with backpropagation.

● Dead RelU Units

✗ 👤 Vanishing gradients

● Exploding gradients

Q50) Which of the following is a best practice with splitting data?

● Always train on both test and train data

✓ 👤 Never train on test data

Explanation:-It is never acceptable to train a model on the test data. This will most likely create overfitting.

● Only train on test data when better model accuracy is required

● None of these

Q51) Which of the following statements is correct about L1 regularization?

● It not efficient for all models

✓ 👤 It penalizes the absolute values of all the weights

Explanation:-The only correct answer is c, and L1 regularization is efficient for all models.

● L2 and L1 regularization are the same thing

● None of these

Q52) Which of the following statements best describes the purpose of splitting data into 70%/30% sections when modeling?

● The creation of feature engineering data sets

✓ 👤 The creation of test and training data sets

Explanation:-It is a common practice when modeling to split data into test and training sets. A common split is to have 70% of the data be used for training and 30% of the data be used for testing.

● The creation of A/B models

● All of these

Q53)

A data scientists has asked for some guidance on how to use AWS Sagemaker automated hyperparameter tuning.

What is the best general advice?

Statement I. Always select the maximum of 20 metrics for your tuning job, this always Sagemaker to pick the best accuracy.

Statement II. Only pick a few metrics for your tuning job because it is likely to give better results.

Statement III. Running one training job at a time achieves the best results with the least amount of compute time

● Statement I
● Statement II
● Statement III
✓ 👤 Both Statement II and III

Explanation:-Both II and III are the best general advice. Additionally, you will get better results by limiting your search to a smaller hyperparameter value range.

Q54)

An associate in the marketing department has asked for guidance about how to select good features for a machine learning model being prepared.

A machine learning specialist would give which of the following pieces of advice?

● There are no easily abstractable ways to select good features
✓ Avoid rarely used discrete feature values

Explanation:-A good feature should appear more than a handful of times in a data set. This gives a model the ability to see the feature in multiple settings

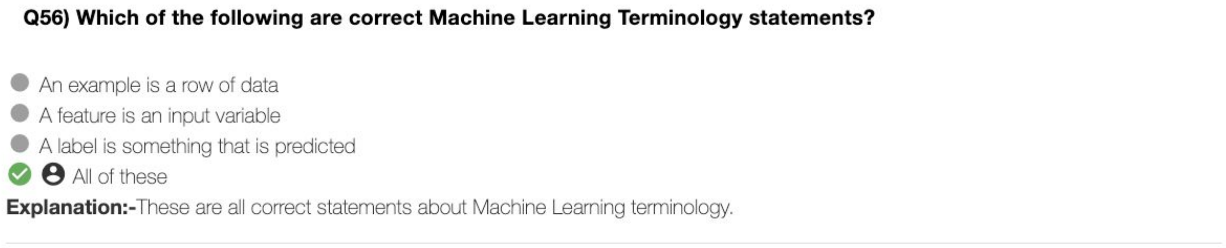
● Always use features with correlations above .8
✗ 👤 None of these

Q55) Which of the following is not a valid usage of machine learning?

● Optimize utility functions
✓ 👤 Sorting data

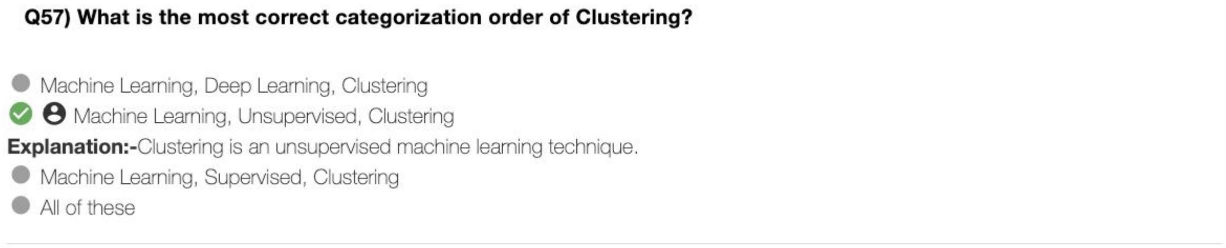
Explanation:-Algorithmic problems, like sorting, where there is one correct way to perform an operation, are not ideal use cases for ML.

● Make predictions
● Classifying data



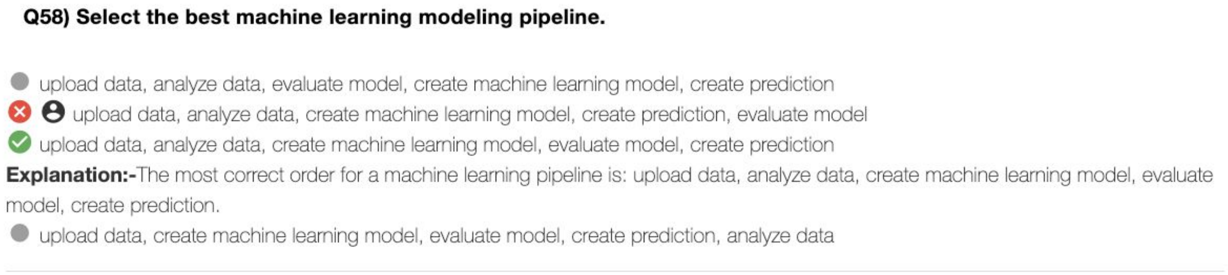
Q56) Which of the following are correct Machine Learning Terminology statements?

● An example is a row of data
● A feature is an input variable
● A label is something that is predicted
✓ 👤 All of these
Explanation:-These are all correct statements about Machine Learning terminology.



Q57) What is the most correct categorization order of Clustering?

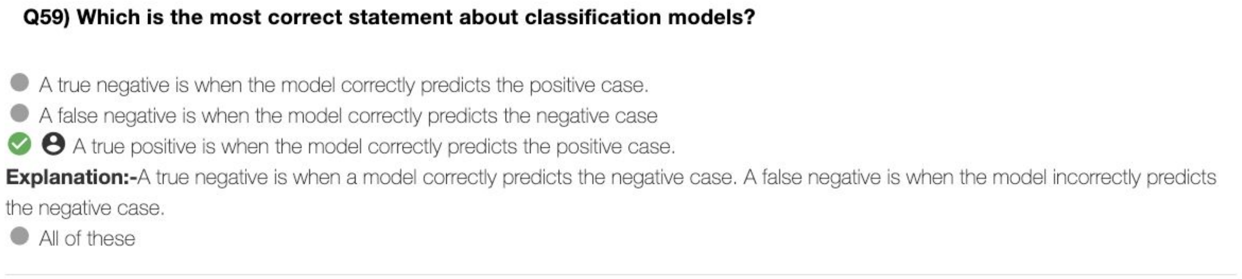
● Machine Learning, Deep Learning, Clustering
✓ 👤 Machine Learning, Unsupervised, Clustering
Explanation:-Clustering is an unsupervised machine learning technique.
● Machine Learning, Supervised, Clustering
● All of these



Q58) Select the best machine learning modeling pipeline.

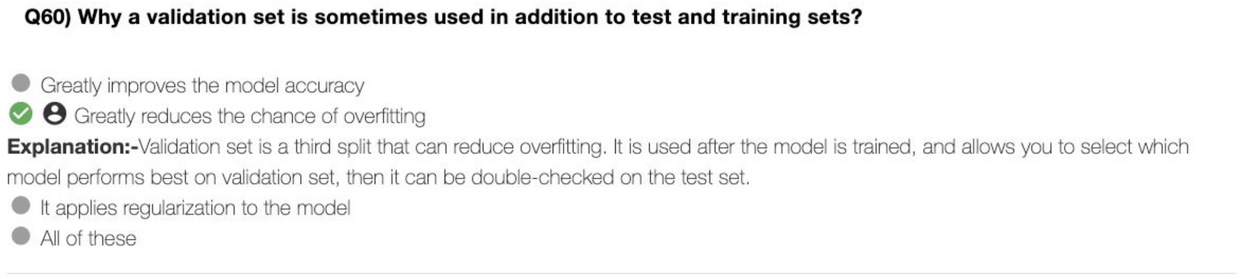
● upload data, analyze data, evaluate model, create machine learning model, create prediction
✗ 👤 upload data, analyze data, create machine learning model, create prediction, evaluate model
✓ upload data, analyze data, create machine learning model, evaluate model, create prediction
Explanation:-The most correct order for a machine learning pipeline is: upload data, analyze data, create machine learning model, evaluate model, create prediction.
● upload data, create machine learning model, evaluate model, create prediction, analyze data

Obviously - I was thinking of creating the model with endpoint configuration so we can then test it. But evaluation happens before with validation set during training.



Q59) Which is the most correct statement about classification models?

● A true negative is when the model correctly predicts the positive case.
● A false negative is when the model correctly predicts the negative case
✓ 👤 A true positive is when the model correctly predicts the positive case.
Explanation:-A true negative is when a model correctly predicts the negative case. A false negative is when the model incorrectly predicts the negative case.
● All of these



Q60) Why a validation set is sometimes used in addition to test and training sets?

● Greatly improves the model accuracy
✓ 👤 Greatly reduces the chance of overfitting
Explanation:-Validation set is a third split that can reduce overfitting. It is used after the model is trained, and allows you to select which model performs best on validation set, then it can be double-checked on the test set.
● It applies regularization to the model
● All of these