Question: 1

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Create data repositories for machine learning**

**Domain:** Data Engineering

**Question text**:

You are a machine learning expert working for a marketing firm. You are supporting a team of data scientists and marketing managers who are running a marketing campaign. Your data scientists and marketing managers need to answer the question “Will this user subscribe to my campaign?” You have been given a dataset in the form of a CSV file which is formatted as such:

UserId, jobId, jobDescription, educationLevel, campaign, duration, willRespondToCampaign

When you build your schema for this dataset, which of the following data descriptors would you use to define the willRespondToCampaign attribute? (Select TWO).

1. CATEGORICAL
2. targetAttributeName
3. TEXT
4. BINARY
5. Numeric
6. rowId

**Answer:** B and D

**Explanation:**

Option A is incorrect because you choose the CATEGORICAL data type for an attribute that holds a limited set of unique strings. For example, a user name, the region, and a product code are categorical values. The willRespondToCampaign attribute takes on either ‘yes’ or ‘no’ values, which are binary in nature.

Option B is correct because for each user observation you are trying to discern “Will this user subscribe to my campaign?” You assign the targetAttributeName field value to the name of the attribute that you are trying to predict. You must assign a targetAttributeName when you create or evaluate your model.

Option C is incorrect because you choose the TEXT data type for an attribute that is a string, or a set of words. Amazon ML converts text attributes into tokens and uses white space as a delimiter. For example, document title becomes document and title, and document-title here becomes document-title and here.

Option D is correct because you choose the BINARY data type for an attribute that only has two possible values, such as yes or no, or true or false. The attribute willRespondToCampaign has only two possible answers: yes or no.

Option E is incorrect because you choose the NUMERIC data type for an attribute that holds a quantity as a number. For example, count, height, and acceleration rate are numeric values.

Option F is incorrect because you choose the rowId field value as an optional flag associated with an attribute in the input data. If you specify an attribute as the rowId, it is included in the prediction output. This attribute allows you to associate each prediction with its observation. The willRespondToCampaign attribute would make a poor identifier for each observation since it only takes two values: yes, or no.

**Reference:**

Please see the AWS Developer Guide titled **Creating a Data Schema for Amazon ML** (<https://docs.aws.amazon.com/machine-learning/latest/dg/creating-a-data-schema-for-amazon-ml.html#assigning-data-types>) for a complete description of the schema attributes.

Question: 2

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Identify and implement a data-ingestion solution**

**Domain:** Data Engineering

**Question text**:

You work for an energy company that buys and sells energy to customers. To get the best prices for their energy customers, your company trades financial energy derivative futures contracts. The trading of these futures contracts requires accurate forecasting of energy prices. You need to build a model that compares spot prices (current commodity price) to future commodity prices (price that a commodity can be bought or sold in the future). Your model needs to assist your company’s futures traders in hedging against future energy price changes based on current price predictions. To source the model with appropriate data you need to gather and process the energy price data automatically.

The data pipeline requires two sources of data:

1. Historic energy spot prices
2. Energy consumption and production rates

Based on the company analysts’ requirements, you have decided you need multiple years of historical data. You also realize you’ll need to update the data feed daily as the market prices change. You can gather the required data through APIs from data provider vendor systems. Your company’s traders require a forecast from your model multiple times per day to help them form their trading strategy. So your pipeline needs to call the data provider APIs multiple times per day. Your data-ingestion pipeline needs to take the data from the API calls, perform preprocessing, and then store the data in an S3 data lake from which your forecasting model will access the data.

Your data-ingestion pipeline has three main steps:

1. Data ingestion
2. Data storage
3. Inference generation

Assuming you have written a lambda function that interacts with the data provider APIs and stores the data in CSV format, which of the following python libraries are the best option to perform the data preprocessing to transform the data by changing raw feature vectors into a format best suited for a SageMaker batch transform job to generate your forecast?

1. matplotlib and plotly
2. boto3 and moto
3. pandas and scikit-learn
4. NLTK and scrapy

**Answer:** C

**Explanation:**

Option A is incorrect because matplotlib and plotly are data visualization python libraries which contain no data transformation functions (see <https://matplotlib.org> and <https://plot.ly/python/>).

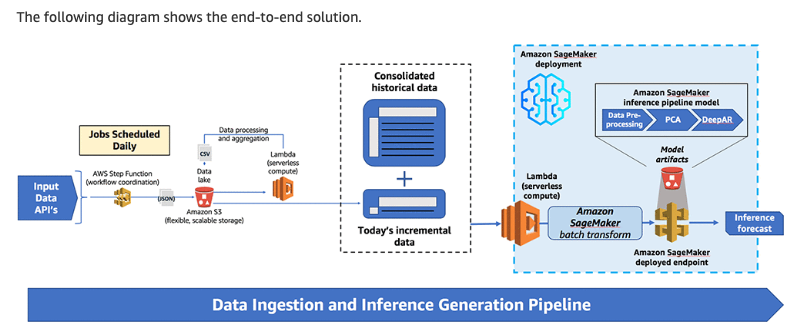
Option B is incorrect because boto3 is a python library that is used to interface with AWS services such as S3, DynamoDB, SQS, etc. Boto3 has no data transformation functions (see <https://aws.amazon.com/sdk-for-python/>). Moto is a python library used to mock interfaces to AWS services such as S3, DynamoDB, SQS, etc. The moto library also contains no data transformation functions (see <https://pypi.org/project/moto/>).

Option C is correct because pandas is the best choice for data wrangling and manipulation of tabular data such as CSV formatted data (see <https://pypi.org/project/pandas/>). Scikit-learn is the best python package to transform raw feature vectors into a format suited to downstream estimators (see <https://scikit-learn.org/stable/modules/preprocessing.html>).

Option D is incorrect because Natural Language Toolkit (NLTK) is best suited to text tagging, classification, and tokenizing, not manipulation of tabular data (see <https://www.nltk.org>). Scrapy is best suited to crawling functionality used to gather structured data from websites, not manipulation of tabular data (see <https://scrapy.org>).

**Diagram:**

Here is a screen shot from the AWS Machine Learning blog depicting the solution:



**Reference:**

Please see the scikit-learn preprocessing data documentation: <https://scikit-learn.org/stable/modules/preprocessing.html>, and a detailed pandas example: <https://towardsdatascience.com/why-and-how-to-use-pandas-with-large-data-9594dda2ea4c>

Question: 3

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Perform hyperparameter optimization**

**Domain:** Modeling

**Question text**:

You work for a retail firm that wishes to conduct a direct mail campaign to attract new customers. Your marketing manager wishes to get answers to questions that can be put into discrete categories, such as “using historical customer email campaign responses, should this customer receive an email from our current campaign?” You decide to use the SageMaker Linear Learner algorithm to build your model. Which hyperparameter setting would you use to get the algorithm to produce discrete results?

1. set the objective hyperparameter to reg:logistic.
2. set the predictor\_type hyperparameter to binary\_classifier.
3. set the predictor\_type hyperparameter to regressor.
4. set the objective hyperparameter to reg:linear.

**Answer:** B

**Explanation:**

Option A is incorrect because the objective hyperparameter is set to reg:logistic when you are using the XGBoost algorithm (See the AWS SageMaker developer documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html>).

Option B is correct because the AWS SageMaker documentation states that for this type of discrete classification problem, when using the Linear Learner algorithm, you set the predictor\_type hyperparameter to binary\_classifier (See the AWS SageMaker documentation: <https://sagemaker.readthedocs.io/en/stable/linear_learner.html>).

Option C is incorrect because the predictor\_type hyperparameter is set to regressor when you are using the Linear Learner algorithm for answers that are quantitative, not discrete (See the AWS SageMaker documentation: <https://sagemaker.readthedocs.io/en/stable/linear_learner.html>).

Option D is incorrect because the objective hyperparameter is set to reg:linear when you are using the XGBoost algorithm for answers that are quantitative in nature (See the AWS SageMaker developer documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html>).

**Reference:**

Please see the AWS SageMaker developer guide titled **Using Amazon SageMaker Built-in Algorithms**: <https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html>) for a complete description of the SageMaker hyperparameter settings.

Question: 4

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Select the appropriate model(s) for a given machine learning problem**

**Domain:** Modeling

**Question text**:

You work for the information security department of a major corporation. You have been asked to build a solution that detects web application log anomalies to protect your organization from fraudulent activity. The system needs to have near-real-time updates to the model where log entry data points dynamically change the underlying model as the log files are updated. Which AWS service component do you use to implement the best algorithm based on these requirements?

1. SageMaker Random Cut Forest
2. Kinesis Data Streams Naive Bayes Classifier
3. Kinesis Data Analytics Random Cut Forest
4. Kinesis Data Analytics Nearest Neighbor

**Answer:** C

**Explanation:**

Option A is incorrect because SageMaker Random Cut Forest is best used for large batch data sets where you don’t need to update the model frequently (See AWS Kinesis Data Analytics documentation: <https://docs.aws.amazon.com/kinesisanalytics/latest/sqlref/sqlrf-random-cut-forest.html>).

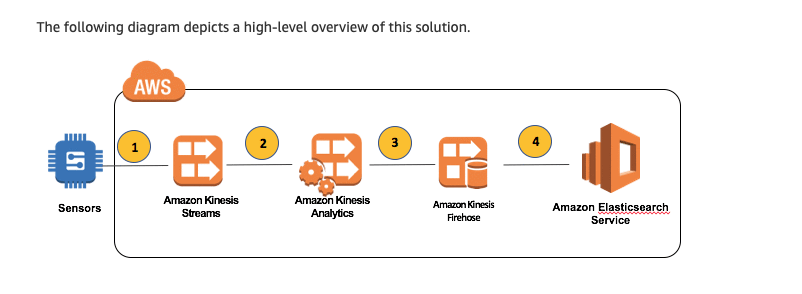
Answer B is incorrect because the Naive Bayes Classifier is used to find independent data points. The Kinesis Data Streams service does not have machine learning algorithm capabilities (See the AWS Kinesis Streams developer documentation: <https://docs.aws.amazon.com/streams/latest/dev/introduction.html>).

Option C is correct. The Kinesis Data Analytics Random Cut Forest algorithm works really well for near-real-time updates to your model (See the AWS Kinesis Data Analytics documentation: <https://docs.aws.amazon.com/kinesisanalytics/latest/sqlref/sqlrf-random-cut-forest.html>).

Option D is incorrect because Kinesis Data Analytics provides a hotspots function that detects higher than normal activity using the distance between a hotspot and its nearest neighbor, but it does not provide ML model update capabilities (See AWS Kinesis Data Analytics documentation: <https://docs.aws.amazon.com/kinesisanalytics/latest/sqlref/sqlrf-hotspots.html>).

**Diagram:**

Here is a screen shot from the AWS Big Data blog:



**Reference:**

For an example, please see the AWS Big Data blog post titled **Perform Near Real-time Analytics on Streaming Data with Amazon Kinesis and Amazon Elasticsearch Service**: <https://aws.amazon.com/blogs/big-data/perform-near-real-time-analytics-on-streaming-data-with-amazon-kinesis-and-amazon-elasticsearch-service/>) for a complete description of the use of Kinesis Data Analytics and the random cut forest algorithm.

Question: 5

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Sanitize and prepare data for modeling**

**Domain:** Exploratory Data Analysis

**Question text**:

You work in the data analytics department of a ride sharing software company. You need to use the K-means machine learning algorithm to separate your company’s optimized ride data into clusters based on ride coordinates. How would you best use AWS Glue to build the data tables needed to classify the ride data?

1. Use Glue crawlers together with a K-means classifier to classify the ride data based on coordinates
2. Use Glue FindMatches to find and remove duplicate records in you data
3. Use Glue to automatically generate code to classify the ride data based on coordinates
4. Use Glue to transform and flatten your data so you can classify the ride data based on coordinates

**Answer:** A

**Explanation:**

Option A is correct. The best way to classify your optimized data is to use a Glue crawler that applies the K-means algorithm. See the AWS Machine Learning documentation (See the AWS SageMaker <https://docs.aws.amazon.com/sagemaker/latest/dg/k-means.html> and AWS Glue crawler <https://docs.aws.amazon.com/glue/latest/dg/add-crawler.html> documentation).

Answer B is incorrect because there is no stated need to remove duplicates from the data.

Option C is incorrect because you don’t need to automatically generate code since Glue will classify your data based on a prioritized list of classifieds without custom code (See the AWS Glue developers guide: (<https://docs.aws.amazon.com/glue/latest/dg/add-classifier.html>).

Option D is incorrect because there is no stated requirement to flatten the ride data.

**Reference:**

For an example, please see the AWS Machine Learning blog post titled **Serverless unsupervised machine learning with AWS Glue and Amazon Athena**: <https://aws.amazon.com/blogs/machine-learning/serverless-unsupervised-machine-learning-with-aws-glue-and-amazon-athena/>.

Question: 6

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Build machine learning solutions for performance, availability, scalability, resiliency, and fault tolerance**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

You work in the security department of your company’s IT division. Your company has decided to try to use facial recognition to improve security on their campus. You have been asked to design a system that augments your company’s building access security by scanning the faces of people entering their buildings and recognizing the person as either an employee/contractor/consultant, who is in the company’s database, or visitor, who is not in their database.

Across their many campus locations worldwide your company has over 750,000 employees and over 250,000 contractors and consultants. These workers are all registered in their HR database. Each of these workers has an image of their face stored in the HR database. You have decided to use Amazon Rekognition for your facial recognition solution. On occasion, the Rekognition model fails to recognize visitors to the buildings. What could be the source of the problem?

1. Face landmarks filters set to a max sharpness
2. Bounding box and confidence score for face comparison threshold tolerances set to max values
3. Confidence threshold tolerance set to the default
4. Face collection contents

**Answer:** D

**Explanation:**

Option A is incorrect. From the [Amazon Rekognition FAQs](https://aws.amazon.com/rekognition/faqs/): “Face landmarks are a set of salient points, usually located on the corners, tips or mid points of key facial components such as the eyes, nose, and mouth. Amazon Rekognition [DetectFaces API](http://docs.aws.amazon.com/rekognition/latest/dg/API_DetectFaces.html) returns a set of face landmarks that can be used to crop faces, morph one face into another, overlay custom masks to create custom filters, and more.” Face landmarks don’t have a sharpness parameter.

Option B is incorrect. The bounding box and confidence score are used to determine confidence in the Rekognition comparison result. A maximum confidence score tolerance would not cause failures to recognize faces, a low confidence score tolerance would.

Option C is incorrect. Similar to option C, the default threshold would not be a common source a recognition failure. A confidence threshold tolerance that is set too low would cause a failure in recognition.

Option D is correct. A suboptimal face collection can be the source of recognition failure. Our face collection has only one image per person. The recommendation, from the [Amazon Rekognition FAQs](https://aws.amazon.com/rekognition/faqs/) is “Besides video resolution, the quality and representative faces part of the face collections to search has major impact. Using multiple face instances per person with variations like beard, glasses, poses (profile and frontal) will significantly improve the performance.”

**Reference:**

Please see the Amazon Recognition developer guide titled [Detecting and Analyzing Faces](https://docs.aws.amazon.com/rekognition/latest/dg/faces.html)

Question: 7

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Recommend and implement the appropriate machine learning services and features for a given problem**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

Your marketing department wishes to understand how their products are being represented in the various social media services in which they have active content streams. They would like insights into the reception of a current product line so they can plan for the roll out of a new product in the line in the new future. You have been tasked with creating a service that organizes the social media content by sentiment across all languages so that your marketing department can determine how best to introduce the new product.

How would you quickly and most efficiently design and build a service for your marketing team that gives insight into the social media sentiment?

1. Use the scikit-learn python library to build a sentiment analysis service to provide insight data to the marketing team’s internal application platform. Build a dashboard into the application platform using React or Angular.
2. Use the DetectSentiment Amazon Comprehend API as a service to provide insight data to the marketing team’s internal application platform. Build a dashboard into the application platform using React or Angular.
3. Use the Amazon Lex API as a service to implement the to provide insight data to the marketing team’s internal application platform. Build a dashboard into the application platform using React or Angular.
4. Use Amazon Translate, Amazon Comprehend, Amazon Kinesis, Amazon Athena, and Amazon QuickSight to build a natural-language-processing (NLP)-powered social media dashboard

**Answer:** D

**Explanation:**

Option A is incorrect since this option is not the quickest to implement nor is it the most efficient, since developers will have to code a react UI and build an end-point to connect the sentiment service to the React or Angular UI.

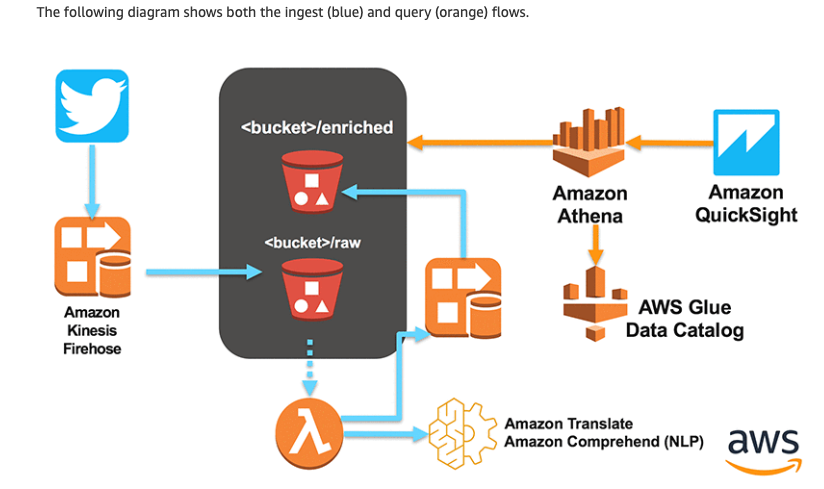
Option B is incorrect since this option is also not the quickest to implement nor is it the most efficient, since developers will have to code a react UI and build an end-point to connect the sentiment service to the React or Angular UI.

Option C is incorrect since Amazon Lex is used primarily for building conversational interfaces into an application using voice or text. This would not give you the most efficient solution to the problem.

Option D is correct since it is the most efficient and quickest way to implement the solution. Amazon Kinesis Data Firehose is used to capture and prepare the social media content. A lambda function can be used to analyze the social media content using Amazon Translate and Amazon Comprehend. Amazon Athena to query the data produced by the lambda function. Use Amazon QuickSight to produce the dashboard. (See the AWS Machine Learning blog post titled: **Build a social media dashboard using machine learning and BI services:**[**https://aws.amazon.com/blogs/machine-learning/build-a-social-media-dashboard-using-machine-learning-and-bi-services/**](https://aws.amazon.com/blogs/machine-learning/build-a-social-media-dashboard-using-machine-learning-and-bi-services/)**)**

**Diagram:**

Here is a screen shot from the AWS Machine Learning blog that shows the desired solution:



**Reference:**

Please see the Amazon Comprehend documentation: <https://aws.amazon.com/comprehend/>.

Question: 8

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Perform feature engineering**

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a financial services firm that wishes to further enhance their fraud detection capabilities. The firm has implemented fine grained transaction logging for all transactions their customers make using their credit cards. The fraud prevention department would like to use this data to produce dashboards to give them insight into their customer’s transaction activity and to provide real-time fraud prediction.

You plan to build a fraud detection model using the transaction observation data with Amazon SageMaker. Each transaction observation has a date-time stamp. In its raw form, the date-time stamp is not very useful in your prediction model since it is unique. Can you make use of the date-time stamp in your fraud prediction model, and if so how?

1. No, you cannot use the date-time stamp since this data point will never occur again. Unique features like this will not help identify patterns in your data.
2. Yes, you can use the date-time stamp data point. You can just use feature selection to deselect the date-time stamp data point, thus dropping it from the learning process.
3. Yes, you can use the date-time stamp data point. You can transform the date-time stamp into features for the hour of the day, the day of the week, and the month.
4. No you cannot use the date-time feature since there is no way to transform it into a unique data point.

**Answer:** C

**Explanation:**

Option A is incorrect since you can use the date-time stamp if you use feature engineering to transform the data point into useful form.

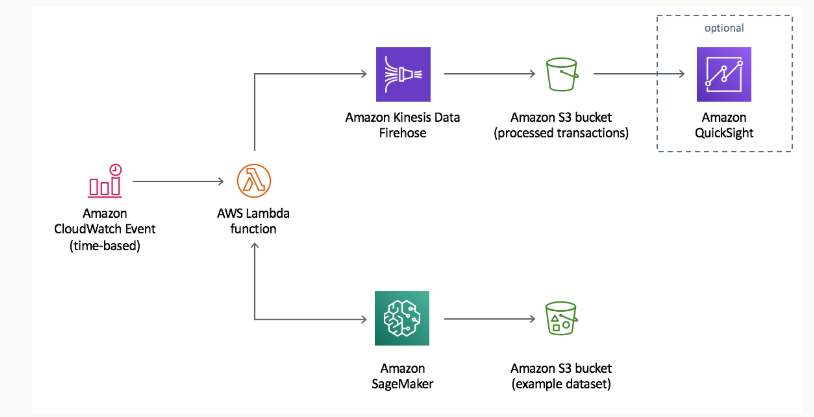
Option B is incorrect since this option is really just another way of ignoring, thus not using, the date-time stamp data point.

Option C is correct. You can transform the data point using feature engineering and thus gain value from it for the learning process of your model. (See the AWS Machine Learning blog post: **Simplify machine learning with XGBoost and Amazon SageMaker:** [**https://aws.amazon.com/blogs/machine-learning/simplify-machine-learning-with-xgboost-and-amazon-sagemaker/**](https://aws.amazon.com/blogs/machine-learning/simplify-machine-learning-with-xgboost-and-amazon-sagemaker/)**)**

Option D is incorrect since we can transform the data point into unique features that represent the hour of the day, the day of the week, and the month, these variables could be useful to learn if the fraudulent activity tends to happen at a particular hour, day of the week, or month.

**Diagram:**

Here is a screen shot from the AWS Machine Learning documentation depicting a typical fraud detection machine learning solution:



**Reference:**

Please see the Amazon Machine Learning developer documentation: <https://docs.aws.amazon.com/machine-learning/latest/dg/feature-processing.html>.

Question: 9

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Train machine learning models**

**Domain:** Modeling

**Question text**:

You work for a real estate company where you are building a machine learning model to predict the prices of houses. You are using a regression decision tree. As you train your model you see that it is overfitted to your training data and that it doesn’t generalize well to unseen data. How can you improve your situation and get better training results in the most efficient way?

1. Use a random forest by building multiple randomized decision trees and averaging their outputs to get the predictions of the housing prices.
2. Gather additional training data that gives a more diverse representation of the housing price data.
3. Use the “dropout” technique to penalize large weights and prevent overfitting.
4. Use feature selection to eliminate irrelevant features and iteratively train your model until you eliminate the overfitting.

**Answer:** A

**Explanation:**

Option A is correct because the random forest algorithm is well known to increase the prediction accuracy and prevent overfitting that occurs with a single decision tree. (See these articles comparing the decision tree and random forest algorithms: <https://medium.com/datadriveninvestor/decision-tree-and-random-forest-e174686dd9eb> and <https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991>)

Option B is incorrect since gathering additional data will not necessarily improve the overfitting problem, especially if the additional data has the same noise level of the original data.

Option C is incorrect since while the “dropout” technique improves models that are overfitted, it is a technique used with neural networks, not decision trees.

Option D is incorrect since it requires significantly more effort than using the random forest algorithm approach.

**Reference:**

Please see this overview of the random forest machine learning algorithm: <https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb>

Question: 10

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Analyze and visualize data for machine learning**

**Domain:** Exploratory Data Analysis

**Question text**:

You work for the data analytics department of your company where you have been asked to build a visualization of the company’s corporate performance for the company’s annual report. The visualization needs to demonstrate company performance by showing how likely it is for a customer to recommend your company’s products, and how much profit a customer brings to the business net acquisition and retention costs. Which types of charts would you use to create this visualization? (Select TWO)

1. Use a distribution scatter chart to show the customer recommendation rate
2. Use a Conversion Rate KPI chart to show conversion rate of customers
3. Use a Relative Market Share KPI chart to show competitive market share
4. Use a Net Promoter Score KPI chart to graph customer recommendations
5. Use a Customer Profitability Score KPI chart to show customer profitability

**Answers:** D and E

**Explanation:**

Option A is incorrect because a distribution scatter chart would show the size of recommendations but not the likelihood of a recommendation.

Option B is incorrect since a Conversion Rate KPI shows how many leads were converted to customers, not likelihood of recommendation or customer profitability.

Option C is incorrect since the Relative Market Share KPI shows how much market share your company owns versus your company’s competitors.

Option D is correct since the Net Promoter Score KPI shows how likely it is a current customer would recommend you company’s products.

Option E is correct since the Customer Profitability Score KPI shows how much profit a customer contributes to your company’s profits after the expenses of acquiring the customer and the expenses associated with retaining the customer.

**Reference:**

Please see this AWS overview of analyzing and visualizing your data for business analytics: <https://aws.amazon.com/data-visualization/>

Question: 11

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Frame business problems as machine learning problems**

**Domain:** Modeling

**Question text**:

You work for a manufacturing company that produces retail apparel, such as shoes, dresses, blouses, etc. Your head of manufacturing has asked you to use your data science skills to determine which product, among a list of potential next products, your company should invest its resources to produce. You decide you need to predict the sales levels of each of the potential next products and select the one with the highest predicted purchase rate. Which type of machine learning approach should you use?

1. You are trying to solve for the greatest number of sales across the potential next products. Therefore, you are solving a multiclass classification problem and you should use multinomial logistic regression.
2. You are trying to solve for the greatest number of sales across the potential next products. Therefore, you are solving a classification problem and you should use the random cut forest model.
3. You are trying to solve for the greatest number of sales across the potential next products. Therefore, you are solving a regression problem and you should use a linear regression model.
4. You are trying to solve for the greatest number of sales across the potential next products. Therefore, you are solving a binary classification problem and you should use a logistic regression model.

**Answer:** C

**Explanation:**

Option A is incorrect. This is not a multiple classification problem where you are trying to solve for more than two outcomes, so a multinomial logistic regression would be the wrong choice for your machine learning model.

Option B is incorrect. You are trying to solve for a numeric result: the number of purchases customers will make for each next potential product. From the [Amazon SageMaker developer guide titled How RCF Works](https://docs.aws.amazon.com/sagemaker/latest/dg/rcf_how-it-works.html): “Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a dataset.”

Option C is correct. You are trying to solve for a numeric result: the number of purchases customers will make for each next potential product. This numeric result case calls for the use of a regression model such as the linear regression model.

Option D is incorrect. This is not a classification problem where you’re solving for a binary (yes or no) result, so a logistic regression model would be the wrong choice for your machine learning.

**Reference:**

Please see this AWS overview of machine learning concepts: <https://docs.aws.amazon.com/machine-learning/latest/dg/machine-learning-concepts.html>, and the Amazon Machine Learning developer guide titled: [Types of ML Models](https://docs.aws.amazon.com/machine-learning/latest/dg/types-of-ml-models.html)

Question: 12

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Apply basic AWS security practices to machine learning solutions**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

You work as a manager of data scientists at a large financial services firm where your team is responsible for building machine learning solutions such as price prediction of equities, futures, and options. You need petabytes of data from dozens of sources internal and external to your organization. All of the external data sources are contractually constrained as to where the data is used and who has access to the data. Your machine learning models require storage of these data in a data lake to allow for quick retrieval of data to fuel your ML models. You have chosen to use S3 to house your data lake. How will you most efficiently protect this data lake, your machine learning data source, against internal threats to data confidentiality and security?

1. Create IAM resource-based policies for each data lake S3 bucket resource. Use bucket policies and Access Control Lists (ACLs) to control the resources at the bucket level and at the object level.
2. Create IAM user policies so that permissions to access your S3 data lake assets are linked to user roles and permissions. Place your data scientists into IAM groups and assign the user policies to those groups. These policies and permissions will define access to the data processing and analytics services which your data scientists will use.
3. Create an access key ID and a secret access key for each internal user of your S3 data lake. Your internal users will then only be able to gain access to your data lake using these keys.
4. Use the AWS CloudHSM cloud-based hardware security module (HSM) to secure your S3 data lake. Internal users of your data lake will use the encryption keys generated by the CloudHSM module to gain access to the data needed for their machine learning models.

**Answer:** B

**Explanation:**

Option A is incorrect because this is a very inefficient approach to the problem of securing a data lake. Most large data lakes contain large numbers of buckets and objects. Using resource-based policies would mean creating a very large set of policies to secure the data lake.

Option B is correct. Per the AWS white paper [Building Big Data Storage Solutions (Data Lakes) for Maximum Flexibility - Securing, Protecting, and Managing Data](https://docs.aws.amazon.com/whitepapers/latest/building-data-lakes/securing-protecting-managing-data.html), “for most data lake environments, we recommend using user policies, so that permissions to access data assets can also be tied to user roles and permissions for the data processing and analytics services and tools that your data lake users will use.”

Option C is incorrect because access keys are used primarily for applications running outside the AWS environment. Resources running inside AWS, as is the case in this scenario, the best practice is to use IAM roles and policies. (See AWS Security blog entry **Guidelines for protecting your AWS account while using programmatic access:** [**https://aws.amazon.com/blogs/security/guidelines-for-protecting-your-aws-account-while-using-programmatic-access/**](https://aws.amazon.com/blogs/security/guidelines-for-protecting-your-aws-account-while-using-programmatic-access/)**)**

Option D is incorrect. The CloudHSM module is used to generate encrypted access keys. However, since we’re dealing with users internal to AWS, IAM role-based security is the best practice for this scenario.

**Reference:**

Please see the AWS white paper [Building Big Data Storage Solutions (Data Lakes) for Maximum Flexibility - Securing, Protecting, and Managing Data](https://docs.aws.amazon.com/whitepapers/latest/building-data-lakes/securing-protecting-managing-data.html)

Question: 13

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Sanitize and prepare data for modeling**

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a financial services company where you are building a model to analyze equity futures prices to predict price movement for your firm’s hedging strategy. You receive several data feeds, some of which contain missing values for some data points. The missing data points in your data feeds are of the categorical type, such as the expiration month or the exchange on which the futures contract is traded. Which strategy should you employ to deal with the missing data point values while attempting to maximize the accuracy of your model without introducing bias into the model?

1. Remove the observations that have the missing data.
2. Impute the missing values using the Mean/Median strategy.
3. Impute the missing values using the Most Frequent strategy.
4. Impute the missing values using a Deep Learning strategy

**Answer:** D

**Explanation:**

Option A is incorrect because this approach will lead to the loss of data points with potentially useful information.

Option B is incorrect because, by definition, it can only be used with numeric data. It is not advisable to use the Mean/Median approach with categorical data points.

Option C is incorrect because while working with categorical data, this method can introduce bias into your data.

Option D is correct. A deep learning approach, using a library such as the datawig python library, uses deep neural networks to impute missing data values and is the most accurate strategy, in the list of given options, at imputing categorical values. (See the datawig documentation: <https://github.com/awslabs/datawig>)

**Reference:**

Please see the article [6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples)](https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779)

Question: 14

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Evaluate machine learning models**

**Domain:** Modeling

**Question text**:

You are a data scientist working for a cancer screening center. The center has gathered data on many patients that have been screened over the years. The data is obviously skewed toward true negative results, as most screened patients don’t have cancer. You are evaluating several machine learning models to decide which model best predicts true positives when using your cancer screening data. You have split your data into a 70/30 ratio of training set to test set. You now need to decide which metric to use to evaluate your models.

Which metric will most accurately determine the model best suited to solve your classification problem?

1. ROC Curve
2. Precision
3. Recall
4. PR Curve

**Answer:** D

**Explanation:**

Option A is incorrect because it is best used when both outcomes have equal importance. Due to the importance of true negative in this equation, it will not differentiate models well for the cancer screening problem, since this data set is skewed to true negatives. The true negative cases are heavily weighted in the equation, thus amplifying the impact of the imbalance.

Option B is incorrect because it only takes into account the percentage of positive cases out of the total predicted positive.

Option C is incorrect because it only takes into account the percentage of positive cases out of the total actual positive.

Option D is correct because the PR Curve is best used to evaluate models on data sets where most of the cases are negative, as in the cancer screening data set. The true negative cases are not weighted heavily in the equation, thus reducing the impact of the imbalance.

**Reference:**

Please see the article [Various ways to evaluate a machine learning model’s performance](https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15)

Question: 15

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Deploy and operationalize machine learning solutions**

**Domain:** ML Implementation and Operations

**Question text**:

You work for a web retailer where you need to analyze data produced for your company by an outside market data provider. You need to produce recommendations based on patterns in user preferences by demographic found in the supplied data. You have stored the data in one of your company’s S3 buckets. You have created a Glue crawler that you have configured to crawl the data on S3 and you have written a custom classifier. Unfortunately, the crawler failed to create a schema. Why might the Glue crawler have failed in this way?

1. You did not add an exclude pattern when you configured the data store
2. The IAM role you assigned to the crawler has the AWSGlueServiceRole managed policy attached plus an inline policy that allows read access to your S3 bucket.
3. All the classifiers returned a certainty of 0.0
4. You chose to create a single schema for each S3 path

**Answer:** C

**Explanation:**

Option A is incorrect. This configuration option is used to exclude objects from the crawler. From the help text on the Add a Data Store screen in the Add Crawler console flow: “The exclude pattern is relative to the include path. Objects that match the exclude pattern are not crawled. For example, with include path **s3://mybucket/** and exclude pattern, **mydir/\*\*** , then all objects in the include path below the **mydir** directory are skipped. In this example, any object whose path matches **s3://mybucket/mydir/\*\*** is not crawled. For more information about patterns see [Cataloging Tables with a Crawler](http://docs.aws.amazon.com/glue/latest/dg/add-crawler.html)”

Option B is incorrect. The IAM role assigned to your crawler needs exactly this managed policy and S3 bucket access. From the Choose an IAM Role screen on the Add Crawler console flow: “Create an IAM role named ‘AWSGlueServiceRole-rolename’ and attach the AWS managed policy, AWSGlueServiceRole, plus an inline policy that allows read access to: s3://yourbucketname”

Option C is correct. The data from the market data provider did not match with certainty any of the built-in classifiers that are part of Glue or your custom classifier. Therefore, Glue returned the default classification string of UNKNOWN. (See the Amazon Glue doc [Adding Classifiers to a Crawler](https://docs.aws.amazon.com/glue/latest/dg/add-classifier.html))

Option D is incorrect. This setting allows you to group compatible schemas. Choosing this option would not prevent the crawler from producing the schema. From the Configure the Crawler’s Output screen in the Add Crawler console flow: “This crawler configuration groups compatible schemas into a single table definition across all S3 objects under the provided include path. Other criteria will still be considered to determine proper grouping.”

**Reference:**

Please see the AWS developer guides [AWS Glue: How It Works](https://docs.aws.amazon.com/glue/latest/dg/how-it-works.html) and [AWS Glue Concepts](https://docs.aws.amazon.com/glue/latest/dg/components-key-concepts.html)

Question: 16

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Perform Feature Engineering**

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a mining company where you are responsible for the data science behind identifying the origin of mineral samples. Your data origins are Canada, Mexico, and the US. Your training data set is imbalanced as such:

Canada | Mexico | US |

1,210 | 120 | 68 |

You run a Random Forest classifier on the training data and get the following results for your test data set (your test data set is balanced):

Confusion matrix:

Predicted \_

Observed | Canada | Mexico | US | Accuracy |

Canada | 45 | 3 | 0 | 94% |

Mexico | 5 | 38 | 5 | 79% |

US | 19 | 8 | 21 | 44% |

In order to address the imbalance in your training data you will need to use a preprocessing step before you create your SageMaker training job. Which technique should you use to address the imbalance?

1. Run your training data through a preprocessing script that uses the SMOTE (Synthetic Minority Over-sampling Technique) approach
2. Run your training data through a Spark pipeline in AWS Glue to one-hot encode the features
3. Run your training data through a preprocessing script that uses the feature-split technique
4. Run your training data through a preprocessing script that uses the min-max normalization technique

**Answer:** A

**Explanation:**

Option A is correct. The SMOTE sampling technique uses the k-nearest neighbors algorithm to create synthetic observations to balance a training data set. (See the article [SMOTE Explained for Noobs](http://rikunert.com/SMOTE_explained))

Option B is incorrect because the Spark pipeline creates one-hot encoded columns in your data. One-hot encoding is a process for converting categorical data points into numeric form. This won’t do anything to address the imbalance in your training data. (See this [explanation of one-hot encoding](https://www.kaggle.com/dansbecker/using-categorical-data-with-one-hot-encoding))

Option C is incorrect because it splits a feature (data point) in your observations into multiple features per observation. This also will have no impact on your imbalanced training data. (See the article [Fundamental Techniques of Feature Engineering for Machine Learning](https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114))

Option D is incorrect because the min-max normalization technique is used to normalize data points into a range of 0 to 1, for example. (See the wikipedia article [Feature Scaling](https://en.wikipedia.org/wiki/Feature_scaling))

**Reference:**

Please see the article [How to Handle Imbalanced Classification Problems in machine learning](https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/)

Question: 17

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Recommend and implement the appropriate machine learning services and features for a given problem**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

You work for a scientific research company where you need to gather data on tree specimens. You have scientist peers who go out in the field across the globe and photograph tree species. The images that they gather need to be classified and labeled so you can use them in your training datasets in your machine learning models. What is the best way to label your image data most accurately and in the most cost efficient manner?

1. Hire human image labelers to process all of your images and label them.
2. Use Amazon Rekognition to analyze all of your images. For the ones that the Rekognition cannot label, have human labelers that you hire attempt to label them.
3. Use an open source labeling tool such as BBox-Label-Tool to process all of your images. For the ones that the tool cannot label, have human labelers that you hire attempt to label them.
4. Use AWS SageMaker Ground Truth to automatically label your images and use the AWS Ground Truth human labelers to label the images that the automatic labeling cannot label.

**Answer:** D

**Explanation:**

Option A is correct. Human labelers may be able to correctly label all of your images, but they will be slow and expensive.

Option B is incorrect. While the Amazon Rekognition service analyzes image data, it does not have the human labeler to active learning model loop that trains an automatic labeling model that Amazon SageMaker Ground Truth has. Therefore, a labeling process based on Rekognition will be more costly and less accurate than a process based on Amazon SageMaker Ground Truth. (See the [Amazon Rekognition overview](https://aws.amazon.com/rekognition/) and the [Amazon SageMaker Ground Truth overview](https://aws.amazon.com/sagemaker/groundtruth/))

Option C is incorrect. While an open source image labeling solution may label some images automatically and a human labeling team that you hire can label the ones the open source software cannot label, this process lacks the human labeler to active learning model loop that trains an automatic labeling model that Amazon SageMaker Ground Truth has. Therefore, a labeling process based on an open source image labeling solution will be less accurate than a process based on Amazon SageMaker Ground Truth.

Option D is correct. As documented in the Amazon SageMaker Ground Truth overview: “Amazon SageMaker Ground Truth uses a process that starts with an active learning model that is trained from human labeled data. Any image that it understands is automatically labeled. Ambiguous data is sent to human labelers for annotation. Then the human labeled images is sent back to the active learning model to retrain the model to incrementally improve its accuracy. (See the [Amazon SageMaker Ground Truth service overview](https://aws.amazon.com/sagemaker/groundtruth/))

**Reference:**

See the [Amazon SageMaker Ground Truth service overview](https://aws.amazon.com/sagemaker/groundtruth/)) and the [Amazon Rekognition overview](https://aws.amazon.com/rekognition/)

Question: 18

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Identify and implement a data-ingestion solution**

**Domain:** Data Engineering

**Question text**:

You need to use machine learning to produce real-time analysis of streaming data from IoT devices out in the field. These devices monitor oil well rigs for malfunction. Due to the safety and security nature of these IoT events, the events must be analyzed by your safety engineers in real-time. You also have an audit requirement to retain your IoT device events for 7 days since you can not fail to process any of the events. Which approach would give you the best solution for processing your streaming data?

1. Use Amazon Kinesis Data Streams and its Kinesis Producer Library to pass your events from your consumers to your Kinesis stream.
2. Use Amazon Kinesis Data Streams and its Kinesis API PutRecords call to pass your events from your consumers to your Kinesis stream.
3. Use Amazon Kinesis Data Streams and its Kinesis Client Library to pass your events from your consumers to your Kinesis stream.
4. Use Amazon Kinesis Data Firehose pass your events directly to your S3 bucket where you store your machine learning data.

**Answer:** B

**Explanation:**

Option A is incorrect. The Amazon Kinesis Data Streams Producer Library is not meant to be used for real-time processing of event data since, according to the AWS developer documentation “it can incur an additional processing delay of up to RecordMaxBufferedTime within the library”. Therefore, it is not the best solution for a real-time analytics solution. (See the AWS developer documentation titled [Developing Producers Using the Amazon Kinesis Producer Library](https://docs.aws.amazon.com/streams/latest/dev/developing-producers-with-kpl.html))

Option B is correct. The Amazon Kinesis Data Streams API PutRecords call is the best choice for processing in real-time since it sends its data synchronously and does not have the processing delay of the Producer Library. Therefore, it is better suited to real-time applications. (See the AWS developer documentation titled [Developing Producers Using the Amazon Kinesis Data Streams API with the AWS SDK for Java](https://docs.aws.amazon.com/streams/latest/dev/developing-producers-with-sdk.html))

Option C is incorrect. The Amazon Kinesis Data Streams Client Library interacts with the Kinesis Producer Library to process its event data. Therefore, you’ll have the same processing delay problem with this option. (See the AWS developer documentation titled [Developing Consumers Using the Kinesis Client Library 1.x](https://docs.aws.amazon.com/streams/latest/dev/developing-consumers-with-kcl.html#kinesis-record-processor-kcl-role))

Option D is incorrect. The Amazon Kinesis Data Firehose service directly streams your event data to your S3 bucket for use in your real-time analytics model. However, Amazon Kinesis Data Firehose retries to send your data for a maximum of 24 hours, but you have a 7 day retention requirement. (See the [Amazon Kinesis Data Firehose FAQs](https://aws.amazon.com/kinesis/data-firehose/faqs/))

**Reference:**

## Please see the [Amazon Kinesis Data Streams documentation](https://aws.amazon.com/kinesis/data-streams/).

Question: 19

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Identify and implement a data-transformation solution**

**Domain:** Data Engineering

**Question text**:

You work as a machine learning specialist for the department of defense in the NSA (National Security Agency). The NSA is responsible for security in the ports of entry around the United States. You need to process real-time video streams from airports around the country to identify questionable activity within the airport facilities and send the streaming data to SageMaker to be used as training data for your model. Your model needs to trigger an alert system when a security event is detected. What AWS services would you use to create this system in the most accurate and cost effective manner?

1. Use AWS Rekognition to process your video streams and send the processed data to your SageMaker model. When the model detects a security event a lambda function is triggered to publish an SNS message to the alert system.
2. Use AWS Elastic Transcoder to process the video streams and send the processed data to your SageMaker model. When the model detects a security event a lambda function is triggered to publish an SNS message to the alert system.
3. Use Amazon Kinesis Video Streams to stream the video to a set of processing workers running in ECS Fargate. The workers send the video data to your SageMaker machine learning model which identifies alert situations. These alerts are processed by Kinesis Data Streams which uses a lambda function to trigger the alert system.
4. Use Amazon Kinesis Data Streams to process your video data using lambda functions which push out an SNS notification to the alert system when a security event is detected.

**Answer:** B

**Explanation:**

Option A is incorrect. The AWS Rekognition service is not meant to process streams. It works with Kinesis Video Streams to provide this capability. Also it needs another component to send its output to your SageMaker model. This part of the solution is missing.

Option B is incorrect. The Amazon Elastic Transcoder service is used to convert video files from one format to another. It would not be useful to stream video to a processing service. (See the AWS documentation titled [Amazon Elastic Transcoder](https://aws.amazon.com/elastictranscoder/))

Option C is correct. The Amazon Kinesis Video Streams service will stream your videos to a processing service which feeds your machine learning model running in SageMaker. Kinesis Streams using lambda to trigger event consumption. (See the AWS machine learning blog titled [Analyze live video at scale in real time using Amazon Kinesis Video Streams and Amazon SageMaker](https://aws.amazon.com/blogs/machine-learning/analyze-live-video-at-scale-in-real-time-using-amazon-kinesis-video-streams-and-amazon-sagemaker/))

Option D is incorrect. This option lacks the machine learning component of the solution.

**Reference:**

## Please see the [Amazon Kinesis Video Streams documentation](https://aws.amazon.com/kinesis/video-streams/).

See a depiction of the proposed solution (in the AWS machine Learning blog titled: [Analyze live video at scale in real time using Amazon Kinesis Video Streams and Amazon SageMaker](https://aws.amazon.com/blogs/machine-learning/analyze-live-video-at-scale-in-real-time-using-amazon-kinesis-video-streams-and-amazon-sagemaker/))

Question: 20

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Sanitize and prepare data for modeling**

**Domain:** Exploratory Data Analysis

**Question text**:

You work as a machine learning specialist at a ride sharing software company. You need to analyze the streaming ride data of your firm’s drivers. First you need to clean, organize, and transform the drive data and load it into your firm’s data lake so you can then use the data in your machine learning models in SageMaker. Which AWS services would give you the simplest solution?

1. Use Amazon Kinesis Data Streams to capture the streaming ride data. Use Amazon Kinesis Data Analytics to clean, organize, and transform the drive data and then output the data to your S3 data lake.
2. Use Amazon Kinesis Data Streams to capture the streaming ride data. Have Amazon Kinesis Data Streams trigger a lambda function to clean, organize, and transform the drive data and then output the data to your S3 data lake.
3. Use Use Amazon Kinesis Data Streams to capture the streaming ride data. Have Kinesis Data Streams stream the data to a set of processing workers running in ECS Fargate. The workers send the data to your S3 data lake.
4. Use Amazon Kinesis Data Firehose to stream the data directly to your S3 data lake.

**Answer:** A

**Explanation:**

Option A is correct. Amazon Kinesis Data Analytics is a very efficient service for taking streams from Amazon Kinesis Data Streams and transforming them with sql or Apache Flink. (See the [Amazon Kinesis Data Analytics overview](https://aws.amazon.com/kinesis/data-analytics/))

Option B is incorrect. Amazon Kinesis Data Analytics does not integrate directly with lambda so you would have to integrate the two services with custom code. This would not be the simplest solution of the options given.

Option C is incorrect. Using ECS Fargate as an intermediary between Amazon Kinesis Data Streams and your data lake would require you to write the transformation logic in your ECS workers. This would not be the simplest solution of the options given.

Option D is incorrect. This option lacks the transformation aspect of the solution.

**Reference:**

## Please see the [Amazon Kinesis Data Analytics documentation](https://aws.amazon.com/kinesis/data-analytics/).

Question: 21

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Identify and implement a data-transformation solution**

**Domain:** Data Engineering

**Question text**:

You work as a machine learning specialist at a marketing company. Your team has gathered market data about your users into an S3 bucket. You have been tasked to write an AWS Glue job to convert the files from json to a format that will be used to store Hive data. Which data format is the most efficient to convert the data for use with Hive?

1. ion
2. grokLog
3. xml
4. orc

**Answer:** D

**Explanation:**

Option A is incorrect. Currently, AWS Glue does not support ion for output. (See the AWS developer guide documentation titled [Format Options for ETL Inputs and Outputs in AWS Glue](https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-etl-format.html))

Option B is incorrect. Currently, AWS Glue does not support grokLog for output. (See the AWS developer guide documentation titled [Format Options for ETL Inputs and Outputs in AWS Glue](https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-etl-format.html))

Option C is incorrect. Currently, AWS Glue does not support xml for output. (See the AWS developer guide documentation titled [Format Options for ETL Inputs and Outputs in AWS Glue](https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-etl-format.html))

Option D is correct. From the Apache Hive Language Manual: “The *Optimized Row Columnar* ([ORC](https://orc.apache.org/)) file format provides a highly efficient way to store Hive data. It was designed to overcome limitations of the other Hive file formats. Using ORC files improves performance when Hive is reading, writing, and processing data.” Also, AWS Glue supports orc for output. (See the [Apache Hive Language Manual](https://cwiki.apache.org/confluence/display/Hive/LanguageManual+ORC) and he AWS developer guide documentation titled [Format Options for ETL Inputs and Outputs in AWS Glue](https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-etl-format.html))

**Reference:**

## Please see the AWS developer guide documentation titled [General Information about Programming AWS Glue ETL Scripts](https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-general.html).

Question: 22

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Create data repositories for machine learning**

**Domain:** Data Engineering

**Question text**:

You work for a software company that has developed a popular mobile gaming app that has a large, active user base. You want to run a predictive model on real-time data generated by the users of the app to see how to structure an upcoming marketing campaign. The data you need for the model is the age of the user, their location, and their level of activity in the game as measured by playing time. You need to filter the data for users who are not yet signed up for you company’s premium service. You’ll also need to deliver your data in json format and convert the playing time into a string format and finally put the data onto an S3 bucket.

Which of the following is the simplest, most cost effective, performant, and scalable way to architect this data pipeline?

1. Create a Kinesis Data Streams application running on an EC2 instance that gathers the mobile user data from its log files; use Kinesis Analytics to transform the log data into the subset you need; connect the Kinesis Data Stream to a Kinesis Firehose which puts the data onto your S3 bucket
2. Create a Kinesis Data Streams application running on EC2 instances in an Auto Scaling Group that gathers the mobile user data from its log files; use Kinesis Analytics to transform the log data into the subset you need; connect the Kinesis Data Stream to a Kinesis Firehose which uses a lambda function to convert the playing time; Kinesis Firehose then puts the data onto your S3 bucket
3. Create a Kinesis Firehose which gathers the data and puts it onto your S3 bucket
4. Create a Kinesis Data Streams application running on EC2 instances in an Auto Scaling Group that gathers the mobile user data from its log files and puts the data onto your S3 bucket

**Answer:** B

**Explanation:**

Option A is incorrect. This option has a bottleneck at the single EC2 instance used to gather the log data from the application log files. This solution would not be the most scalable.

Option B is correct. This option scales well at the Kinesis Data Streams application level because of the Auto Scaling Group. It also uses Kinesis Data Analytics to transform the data into the subset you need and uses the Kinesis Firehose lambda option to convert the playing time to the proper format.

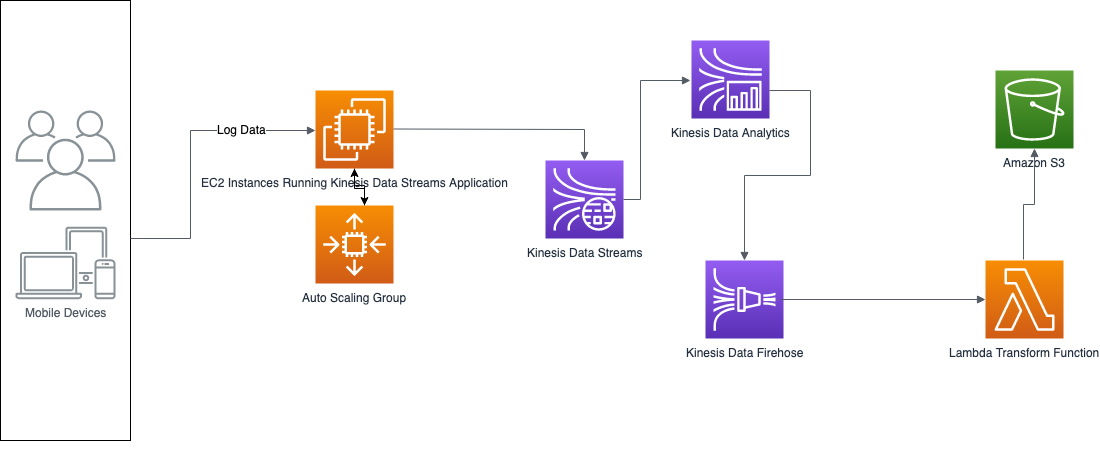
Option C is incorrect. This option does not transform the log data gathered by the Kinesis Firehose before writing the data to the S3 bucket.

Option D is incorrect. This option does not transform the log data gathered by the Kinesis Data Streams application before writing the data to the S3 bucket.

**Reference:**

## Please see the AWS developer guide documentation titled [What is Kinesis Data Streams](https://docs.aws.amazon.com/streams/latest/dev/introduction.html), the [AWS Auto Scaling documentation](https://aws.amazon.com/autoscaling/), the [Amazon Kinesis Data Firehose documentation](https://aws.amazon.com/kinesis/data-firehose/), and the [Amazon Kinesis Data Analytics documentation](https://aws.amazon.com/kinesis/data-analytics/).

Here is a diagram of the solution:



Question: 23

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Deploy and operationalize machine learning solutions**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

You are deploying your data streaming pipeline for your machine learning environment. Your cloud formation stack has a Kinesis Data Firehose using the Data Transformation feature where you have configured Firehose to write to your S3 data lake. When you stream data through your Kinesis Firehose you notice that no data is arriving on your S3 bucket. What might be the problem that is causing the failure?

1. Your lambda memory setting is set to the maximum value allowed
2. Your S3 bucket is in the same region as your Kinesis Data Firehose
3. Your Kinesis Data Firehose buffer setting is set to the default value
4. Your lambda timeout value is set to the default value

**Answer:** D

**Explanation:**

Option A is incorrect. The maximum memory setting for lambda is 3 MB. Using the maximum memory would not cause Firehose to fail to write to S3. It will increase the cost of your solution however, since per the AWS documentation “Lambda allocates CPU power linearly in proportion to the amount of memory configured.”

Option B is incorrect. Your S3 bucket used by Kinesis Data Firehose to output your data must be in the same region as your Firehose. Since they are in the same region, this would not cause a failure to write to the S3 bucket.

Option C is incorrect. The Kinesis Data Firehose documentation states that “Kinesis Data Firehose buffers incoming data before delivering it to Amazon S3. You can choose a buffer size (1–128 MBs) or buffer interval (60–900 seconds). The condition that is satisfied first triggers data delivery to Amazon S3.” Using the default setting would not prevent Firehose from writing to S3.

Option D is correct. The lambda timeout value default is 3 seconds. For many Kinesis Data Firehose implementations, 3 seconds is not enough time to execute the transformation function.

**Reference:**

## Please see the Amazon Kinesis Data Firehose developer guide documentation titled [Configure Settings](https://docs.aws.amazon.com/firehose/latest/dev/create-configure.html), the Amazon Kinesis Data Firehose developer guide documentation titled [Amazon Kinesis Data Firehose Data Transformation](https://docs.aws.amazon.com/firehose/latest/dev/data-transformation.html), and the AWS Lambda developer guide documentation titled [AWS Lambda Function Configuration](https://docs.aws.amazon.com/lambda/latest/dg/resource-model.html).

Question: 24

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Apply basic AWS security practices to machine learning solutions**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

You work as a machine learning specialist at credit card transaction processing company. You have built a data streaming pipeline using Kinesis Data Firehose and S3. Due to the personal identifiable information contained in your data stream your data must be encrypted in flight and at rest. How should you configure your solution to achieve encryption at rest?

1. Encrypt the data at the data consumer application level
2. Encrypt the data by configuring Firehose to use S3-managed encryption keys (SSE-S3)
3. Encrypt the data by configuring Firehose to use S3 server-side encryption with AWS Key Management Service (SSE-KMS)
4. Encrypt the data by configuring Firehose to use S3 server-side encryption with 256-bit AES-GCM with HKDF

**Answer:** C

**Explanation:**

Option A is incorrect. Encrypting the data at the Kinesis consumer application level does not allow for encryption at the S3 bucket. Once the data has reached the consumer application it has already been stored in S3 without being encrypted.

Option B is incorrect. Kinesis Data Firehose does not use SSE-S3, it uses SSE-KMS. (See the Amazon Kinesis Data Firehose developer documentation titled [Configure Settings](https://docs.aws.amazon.com/firehose/latest/dev/create-configure.html))

Option C is correct. The Kinesis Data Firehose documentation states that “Kinesis Data Firehose supports Amazon S3 server-side encryption with AWS Key Management Service (AWS KMS) for encrypting delivered data in Amazon S3. You can choose to not encrypt the data or to encrypt with a key from the list of AWS KMS keys that you own. For more information, see [Protecting Data Using Server-Side Encryption with AWS KMS–Managed Keys (SSE-KMS)](https://docs.aws.amazon.com/AmazonS3/latest/dev/UsingKMSEncryption.html).”

Option D is incorrect. Kinesis Data Firehose does not use 256-bit AES-GCM with HKDF, it uses SSE-KMS. (See the Amazon Kinesis Data Firehose developer documentation titled [Configure Settings](https://docs.aws.amazon.com/firehose/latest/dev/create-configure.html))

**Reference:**

## Please see the Amazon Kinesis Data Firehose developer guide documentation titled [Creating an Amazon Kinesis Data Firehose Delivery Stream](https://docs.aws.amazon.com/firehose/latest/dev/basic-create.html), and the Amazon Kinesis Data Streams developer guide documentation titled [What is Server-Side Encryption for Kinesis Data Streams](https://docs.aws.amazon.com/streams/latest/dev/what-is-sse.html).

Question: 25

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Deploy and operationalize machine learning solutions**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

You are working as a machine learning specialist at a medical research facility. You have setup a data pipeline delivery stream using Amazon Kinesis Data Firehose as your data streaming service and Amazon Redshift as your data warehouse. Your researchers have setup the S3 bucket, in their own account, that you have used for your Kinesis Data Firehose. Your researchers need to access the data using BI tools such as Amazon QuickSight to build dashboards and use metrics in their research. However, when you implement your solution you notice that your streaming data does not load into your Redshift data warehouse. What could be a reason why this is happening? Choose 2 answers.

1. You have not created an IAM role for your Kinesis Firehose to access the S3 bucket
2. You defined a cluster security group and associated it with your Redshift cluster
3. The access policy associated with your Kinesis Firehose does not have lambda:InvokeFunction specified in the Allow Action section of the Lambda actions
4. The access policy associated with your Kinesis Firehose does not have kms:GenerateDataKey specified in the Allow Action section of the KMS actions
5. The access policy associated with your Kinesis Firehose does not have S3:PutObjectAcl specified in the Allow Action section of the S3 actions

**Answers:** A and E

**Explanation:**

Option A is correct. As documented in the [Amazon Kinesis Data Firehose developer guide](https://docs.aws.amazon.com/firehose/latest/dev/controlling-access.html#using-iam-rs) “Kinesis Data Firehose uses the specified Amazon Redshift user name and password to access your cluster, and uses an IAM role to access the specified bucket, key, CloudWatch log group, and streams. You are required to have an IAM role when creating a delivery stream.”

Option B is incorrect. The cluster security group is used to grant users inbound access to the Redshift cluster. Defining a cluster security group would not prevent Kinesis Firehose from accessing your Redshift cluster. (See the Amazon Redshift database developer guide titled [Amazon Redshift Security Overview](https://docs.aws.amazon.com/redshift/latest/dg/c_security-overview.html))

Option C is incorrect. Since you are not using the Lambda function feature of Kinesis Data Firehose, this Lambda action is not needed in the access policy.

Option D is incorrect. Since you are not using the data encryption feature of Kinesis Data Firehose, this KMS action is not needed in the access policy.

Option E is correct. Since you are not the owner of the S3 bucket used by Kinesis Data Firehose, you need to specify the S3:PutObjectAcl in the S3 actions of the access policy. (See the Amazon Kinesis Data Firehose developers guide titled [Grant Kinesis Data Firehose Access to Amazon Redshift Destination](https://docs.aws.amazon.com/firehose/latest/dev/controlling-access.html#using-iam-rs))

**Reference:**

## Please see the Amazon Kinesis Data Firehose developers guide titled [Grant Kinesis Data Firehose Access to Amazon Redshift Destination](https://docs.aws.amazon.com/firehose/latest/dev/controlling-access.html#using-iam-rs), and the [Amazon Kinesis Data Firehose overview page](https://aws.amazon.com/kinesis/data-firehose/), and the Amazon Redshift database developer guide titled [Amazon Redshift Security Overview](https://docs.aws.amazon.com/redshift/latest/dg/c_security-overview.html).

Question: 26

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Create data repositories for machine learning**

**Domain:** Data Engineering

**Question text**:

You work as a machine learning specialist at a hedge fund. You are working on a time-series price prediction model for the firm and you have setup a data delivery stream using Amazon Kinesis Data Streams. You are creating the data producer application code to take trade data from your trade system and send the trade records to your Kinesis Data Stream. Your python code is structured as follows:

import boto3

import requests

import json

client = boto3.client(‘kinesis’, region\_name=’us-east1’)

while True:

r = requests.get(‘<https://trading-applicatio-url>’)

data = json.dumps(r.json())

client.put\_record(  
 *parameters needed for put\_record api call*

)

...

Which of the following options are valid put\_record request parameters? Select 3.

1. Data
2. ImplicitHashKey
3. ExplicitHashKey
4. PartitionKeys
5. SequenceNumberForOrdering
6. ShardId

**Answers:** A, C, and E

**Explanation:**

Options A, C, and E are correct. As documented in the [Amazon Kinesis Data Streams API reference guide titled PutRecord](https://docs.aws.amazon.com/kinesis/latest/APIReference/API_PutRecord.html) “The request accepts the following data in JSON format: Data, ExplicitHashKey, PartitionKey, SequenceNumberForOrdering, and StreamName”

Option B is incorrect. There is no ImplicitHashKey request parameter. (See the [Amazon Kinesis Data Streams API reference guide titled PutRecord](https://docs.aws.amazon.com/kinesis/latest/APIReference/API_PutRecord.html))

Option D is incorrect. There is no PartitionKeys request parameter. However, there is a PartitionKey request parameter. (See the [Amazon Kinesis Data Streams API reference guide titled PutRecord](https://docs.aws.amazon.com/kinesis/latest/APIReference/API_PutRecord.html))

Option F is incorrect. There is no ShardId request parameter. However, there is a ShardId response element. (See the [Amazon Kinesis Data Streams API reference guide titled PutRecord](https://docs.aws.amazon.com/kinesis/latest/APIReference/API_PutRecord.html))

**Reference:**

## Please see the Amazon Kinesis Data Streams developers guide titled [Kinesis Data Stream Producers](https://docs.aws.amazon.com/streams/latest/dev/amazon-kinesis-producers.html), and the [Amazon Kinesis Data Streams API reference guide titled PutRecord](https://docs.aws.amazon.com/kinesis/latest/APIReference/API_PutRecord.html).

Question: 27

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Sanitize and prepare data for modeling**

**Domain:** Exploratory Data Analysis

**Question text**:

You work as a machine learning specialist at a firm that runs a web application that allows users to research and compare real estate properties across the globe. You are working on a property foreclosure model to predict potential price drops. You have decided to use the SageMaker Linear Learner algorithm. Here is a small sample of the data you’ll have to work with:

| Type | Bedrooms | Area | Solar\_Rating | Price | Foreclosed |

| condo | 2 | 2549 | H | 125400 | N |

| house | 4 | 4124 | M | 250250 | Y |

| house | 3 | 3250 | | 200000 | N |

| condo | 1 | 900 | N | 90250 | N |

| condo | 2 | ? | L | 125400 | Y |

In order to feed this data into your model you will first need to clean and format your data.

Which of the following SageMaker built in scikit-learn library transformers would you use to clean and format your data? Select 4.

1. StandardScaler to encode the Solar\_Rating feature
2. OneHotEncoder to encode the Area feature
3. SimpleImputer to complete the missing values in the Solar\_Rating and Area features
4. OneHotEncoder to encode the Type feature
5. OrdinalEncoder to complete the missing values in the Solar\_Rating and Area features
6. OrdinalEncoder to encode the Solar\_Rating feature
7. LabelBinarizer to encode the Foreclosed feature
8. MinMaxScaler to encode the Foreclosed feature

**Answers:** C, D, F, and G

**Explanation:**

Options A, is incorrect. From the [scikit-learn API Reference](https://scikit-learn.org/stable/modules/classes.html): the StandardScaler transformer is used to Standardize features by removing the mean and scaling to unit variance. The OrdinalEncoder transformer would be the better choice for this feature since H > M > L > N, therefore this feature has ordinal values.

Option B is incorrect. The OneHotEncoder transforms nominal categorical features and creates new binary columns for each observation. The Area feature holds numerical or quantitative data, which does not need to be transformed.

Option C is correct. The Solar\_Rating and Area features have missing data in some observations. From the [scikit-learn API Reference](https://scikit-learn.org/stable/modules/classes.html): the SimpleImputer transformer is used to complete missing values.

Option D is correct. The Type feature is a good candidate for the OneHotEncoder transformer since the Type feature holds a limited number of categorical types. The OneHotEncoder transforms nominal categorical features and creates new binary columns for each observation.

Option E is incorrect. From the [scikit-learn API Reference](https://scikit-learn.org/stable/modules/classes.html): the OrdinalEncoder transformer encodes categorical features as an integer array. This encoder does not complete missing values.

Option F is correct. From the [scikit-learn API Reference](https://scikit-learn.org/stable/modules/classes.html): the OrdinalEncoder transformer encodes categorical features as an integer array which maintains the ordinal nature of the data. Since H > M > L > N, this feature has ordinal values.

Option G is correct. The Foreclosed feature holds one of two choices, either a ‘Y’ or a ‘N’. Therefore, this feature is a good candidate for the LabelBinarizer. From the [scikit-learn API Reference](https://scikit-learn.org/stable/modules/classes.html): the LabelBinarizer transformer binarizes label in a one-versus-all fashion.

Option H is incorrect. From the [scikit-learn API Reference](https://scikit-learn.org/stable/modules/classes.html): the MinMaxScaler transformer transforms features by scaling each feature to a given range. The Foreclosed feature has binary data: either ‘Y’ or ‘N’ so it is better suited to the LabelBinarizer transformer.

**Reference:**

## Please see the Amazon SageMaker developer guide titled [Use Scikit-learn with Amazon SageMaker](https://docs.aws.amazon.com/sagemaker/latest/dg/sklearn.html), and the [scikit-learn API Reference](https://scikit-learn.org/stable/modules/classes.html).

Question: 28

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Perform Feature Engineering**

**Domain:** Exploring Data Analysis

**Question text**:

You work as a machine learning specialist for a polling company. For the upcoming election you need to classify the over 500,000 registered voters in you voter database by age for a campaign your team is about to launch. Your data is structured as such:

| voter\_id | voter\_age | voter\_occupation | voter\_income | …

| 1 | 21 | student | 0 | …

| 2 | 35 | nurse | 25000 | …

| 3 | 49 | manager | 150000 | …

| 4 | 63 | truck driver | 45000 | …

| 5 | 55 | teacher | 65000 | …

…

Because you have continuous data for your voter age feature, classifying your observations by age would result in too many classifications, i.e. one for every possible voter age from 21 though probably over 90. You need to have uniform classifications that are limited in number in order to make the best use of your data in you machine learning model.

What numerical feature engineering technique will give you the best distribution of classifications?

1. Cartesian Product Transformation
2. N-Gram Transformation
3. Orthogonal Sparse Bigram (OSB) Transformation
4. Normalization Transformation
5. Quantile Binning Transformation

**Answer:** E

**Explanation:**

Options A is incorrect. From the Amazon Machine Learning developer guide titled [Data Transformations Reference](https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-reference.html#cartesian-product-transformation) “The Cartesian product transformation takes categorical variables or text as input, and produces new features that capture the interaction between these input variables.” Because this transformation is for transforming text it would not give you uniform age classifications that are limited in number.

Option B is incorrect. From the Amazon Machine Learning developer guide titled [Data Transformations Reference](https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-reference.html#cartesian-product-transformation) “The n-gram transformation takes a text variable as input and produces strings corresponding to sliding a window of (user-configurable) n words, generating outputs in the process.” Because this transformation is also for transforming text it would not give you uniform age classifications that are limited in number.

Option C is incorrect. From the Amazon Machine Learning developer guide titled [Data Transformations Reference](https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-reference.html#cartesian-product-transformation) “The OSB transformation is intended to aid in text string analysis and is an alternative to the bi-gram transformation (n-gram with window size 2). OSBs are generated by sliding the window of size n over the text, and outputting every pair of words that includes the first word in the window.” Because this transformation is also for transforming text it would not give you uniform age classifications that are limited in number.

Option D is incorrect. From the Amazon Machine Learning developer guide titled [Data Transformations Reference](https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-reference.html#cartesian-product-transformation) “The normalization transformer normalizes numeric variables to have a mean of zero and variance of one. Normalization of numeric variables can help the learning process if there are very large range differences between numeric variables because variables with the highest magnitude could dominate the ML model, no matter if the feature is informative with respect to the target or not.” Because this transformation is for normalizing continuous data it would not give you uniform age classifications that are limited in number.

Option E is correct. From the Amazon Machine Learning developer guide titled [Data Transformations Reference](https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-reference.html#cartesian-product-transformation) “The quantile binning processor takes two inputs, a numerical variable and a parameter called *bin number*, and outputs a categorical variable. The purpose is to discover non-linearity in the variable's distribution by grouping observed values together.” Because Quantile binning is used to create uniform bins of classifications it would be the right choice to give you uniform age classifications that are limited in number. For example, you could create classification bins such as: Under 30, 30 to 50, Over 50. Or even better: Millennial, Generation X, Baby Boomer, etc.

**Reference:**

Please see the Amazon Machine Learning developer guide titled [Data Transformations for Machine Learning](https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-for-machine-learning.html), and the article [Feature Engineering in Machine Learning (Part 1) Handling Numeric Data with Binning](https://medium.com/hacktive-devs/feature-engineering-in-machine-learning-part-1-a3904769cd93)

## 

Question: 29

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Analyze and visualize data for machine learning**

**Domain:** Exploring Data Analysis

**Question text**:

You work as a machine learning specialist for a consulting firm where you are analyzing data about the consultants who work there in preparation for using the data in you machine learning models. The features you have in your data are things like employee id, specialty, practice, job description, billing hours, and principle. The principle attribute is represented as ‘yes’ or ‘no’, whether the consultant has made principle level or not. For your initial analysis you need to identify the distribution of consultants and their billing hours for the given period. What visualization best describes this relationship?

1. Scatter plot
2. Histogram
3. Line chart
4. Box plot
5. Bubble chart

**Answer:** B

**Explanation:**

Options A is incorrect. You are looking for a distribution on a single dimension: the consultants billing hours. From the Amazon QuickSite User Guide titled [Working with Visual Types in Amazon QuickSight](https://docs.aws.amazon.com/quicksight/latest/user/working-with-visual-types.html) “A scatter chart shows a multiple distribution, i.e. two or three measures for a dimension.”

Option B is correct. You are looking for a distribution of a single dimension: the consultants billing hours. From the [wikipedia article titled Histogram](https://en.wikipedia.org/wiki/Histogram) “A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable.” The continuous variable in this question: the billing hours, binned into ranges (x axis), at a frequency: the number of consultants at a billing hour range (y axis).

Option C is incorrect. From the Amazon QuickSite User Guide titled [Working with Visual Types in Amazon QuickSight](https://docs.aws.amazon.com/quicksight/latest/user/working-with-visual-types.html) “Use line charts to compare changes in measured values over a period of time.” You are looking for a distribution not a comparison of changes over a period of time.

Option D is incorrect. From the Statistics How To article titled [Types of Graphs Used in Math and Statistics](https://www.statisticshowto.datasciencecentral.com/types-graphs/) “A boxplot, also called a box and whisker plot, is a way to show the spread and centers of a data set. Measures of spread include the interquartile range and the mean of the data set. Measures of center include the mean or average and median (the middle of a data set).” A Box Plot shows the distribution of multiple dimensions of the data. Once again, you are looking for a distribution of a single dimension, not a distribution on multiple dimensions.

Option E is incorrect. From the [wikipedia article titled Bubble Chart](https://en.wikipedia.org/wiki/Bubble_chart) “A bubble chart is a type of chart that displays three dimensions of data. Each entity with its triplet (*v*1, *v*2, *v*3) of associated data is plotted as a disk that expresses two of the *vi* values through the disk's *xy* location and the third through its size.” Once again, you are looking for a distribution of a single dimension, not a distribution on three dimensions.

**Reference:**

Please see the Amazon QuickSight user guide titled [Working with Amazon QuickSight Visuals](https://docs.aws.amazon.com/quicksight/latest/user/working-with-visuals.html), and the Statistics How To article titled [Types of Graphs Used in Math and Statistics](https://www.statisticshowto.datasciencecentral.com/types-graphs/)

## 

Question: 30

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Train machine learning models**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a robotics manufacturer where you are attempting to use unsupervised learning to train your robots to perform their prescribed tasks. You have engineered your data and produced a CSV file and placed it on S3.

Which of the following input data channel specifications are correct for your data?

1. Metadata Content-Type is identified as text/csv
2. Metadata Content-Type is identified as application/x-recordio-protobuf;boundary=1
3. Metadata Content-Type is identified as application/x-recordio-protobuf;label\_size=1
4. Metadata Content-Type is identified as text/csv;label\_size=0

**Answer:** D

**Explanation:**

Option A is incorrect. The Content-Type of text/csv without specifying a label\_size is used when you have target data, usually in column one, since the default value for label\_size is 1 meaning you have one target column. (See the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html))

Option B is incorrect. The boundary content type is not relevant to CSV files, it is used for multipart form data.

Option C is incorrect. For unsupervised learning the label\_size should equal 0, indicating the absence of a target. (See the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html))

Option D is correct. For unsupervised learning the label\_size equals 0, indicating the absence of a target. (See the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html))

**Reference:**

Please see the Amazon SageMaker developer guide, specifically [Common Data Formats for Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-algo-common-data-formats.html) and [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html)

## 

Question: 31

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Train machine learning models**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a manufacturing plant where you are attempting to use supervised learning to train assembly line image recognition to categorize malformed parts. You have engineered your data and produced a CSV file and placed it on S3.

Which of the following input data channel specifications are correct for your data? (Select TWO)

1. Metadata Content-Type is identified as text/csv
2. Metadata Content-Type is identified as text/csv;label\_size=0
3. Target value should be in the first column with no header
4. Target value should be in the last column with no header
5. Target value should be in the last column with a header
6. Target value should be in the first column with a header

**Answers:** A and C

**Explanation:**

Option A is correct. The Content-Type of text/csv without specifying a label\_size is used when you have target data, usually in column one, since the default value for label\_size is 1 meaning you have one target column. (See the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html))

Option B is incorrect. The Content-Type of text/csv specifying a label\_size of 0 is used when you do not have target data. You usually choose this setting when using unsupervised learning. (See the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html))

Option C is correct. From the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html) “Amazon SageMaker requires that a CSV file doesn't have a header record and that the target variable is in the first column”

Option D is incorrect. From the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html) “Amazon SageMaker requires that a CSV file doesn't have a header record and that the target variable is in the first column”

Option E is incorrect. From the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html) “Amazon SageMaker requires that a CSV file doesn't have a header record and that the target variable is in the first column”

Option F is incorrect. From the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html) “Amazon SageMaker requires that a CSV file doesn't have a header record and that the target variable is in the first column”

**Reference:**

Please see the Amazon SageMaker developer guide, specifically [Common Data Formats for Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-algo-common-data-formats.html) and [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html)

## 

Question: 32

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Hyperparameter tuning**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a marketing firm. Your firm wishes to determine which customers in a dataset of their registered users will respond to a new proposed marketing campaign. You plan to use the XGBoost algorithm on the binary classification problem. In order to find the optimal model you plan to run many hyperparameter tuning jobs to reach the best hyperparameter values. Which of the following hyperparameters must you use in your tuning jobs if your objective is set to multi:softprob? (Select TWO)

1. alpha
2. base\_score
3. eta
4. num\_round
5. gamma
6. num\_class

**Answers:** D and F

**Explanation:**

Option A is incorrect. The alpha hyperparameter is used to adjust the L1 regulation term on weights. This term is optional. (See the Amazon SageMaker developer guide titled [XGBoost Hyperparameters](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html))

Option B is incorrect. The base\_score hyperparameter is used to set the initial prediction score of all instances. This term is optional. (See the Amazon SageMaker developer guide titled [XGBoost Hyperparameters](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html))

Option C is incorrect. The eta hyperparameter is used to prevent overfitting. This term is optional. (See the Amazon SageMaker developer guide titled [XGBoost Hyperparameters](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html))

Option D is correct. The num\_round hyperparameter is used to set the number of rounds to run in your hyperparameter tuning jobs. This term is required. (See the Amazon SageMaker developer guide titled [XGBoost Hyperparameters](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html))

Option E is incorrect. The gamma hyperparameter is used to set the minimum loss reduction required to make a further partition on a leaf node of the tree. This term is optional. (See the Amazon SageMaker developer guide titled [XGBoost Hyperparameters](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html))

Option F is correct. This hyperparameter is used to set the number of classes. This term is required if the objective is set to multi:softmax or multi:softprob. (See the Amazon SageMaker developer guide titled [XGBoost Hyperparameters](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html))

**Reference:**

Please see the Amazon SageMaker developer guide titled [Automatic Model Tuning](https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning.html), and the Amazon SageMaker developer guide titled [How Hyperparameter Tuning Works](https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-how-it-works.html)

## 

Question: 33

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Hyperparameter tuning**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a healthcare insurance company. Your company wishes to determine which registered plan participants will choose a new health care option your company plans to release. The roll-out plan for the new option is compressed, so you need to produce results quickly. You plan to use a binary classification algorithm on this problem. In order to find the optimal model quickly you plan to run the maximum number of concurrent hyperparameter training jobs to reach the best hyperparameter values. Which of the following types of hyperparameters tuning techniques will best suit your needs?

1. Bayesian Search
2. Hidden Markov Models
3. Conditional Random Fields
4. Random Search

**Answer:** D

**Explanation:**

Option A is incorrect. Bayesian Search uses regression to iteratively choose sets hyperparameters to test. Due to this iterative approach, this method cannot run the maximum number of concurrent training jobs without impacting the performance of the search. Therefore, this method will take longer than the Random Search method.

Option B is incorrect. The Hidden Markov Model is a class of probabilistic graphical model. It is not used by SageMaker for hyperparameter tuning.

Option C is incorrect. Conditional Random Fields is a type of discriminative classifier. It is not used by SageMaker for hyperparameter tuning.

Option D is correct. The Random Search technique allows for you to run the maximum number of concurrent training jobs without impacting the performance of the search. Therefore, getting you to your optimized hyperparameters quickly.

**Reference:**

Please see the Amazon SageMaker developer guide titled [How Hyperparameter Tuning Works](https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-how-it-works.html)

## 

Question: 34

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Train machine learning models**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a financial services company. You are building a machine learning model to perform futures price prediction. You have trained your model and you now want to evaluate it to make sure it is not overtrained and can generalize.

Which of the following techniques is the appropriate method to cross validate your machine learning model?

1. Leave One Out Cross Validation (LOOCV)
2. K-Fold Cross Validation
3. Stratified Cross Validation
4. Time Series Cross Validation

**Answer:** D

**Explanation:**

Option A is incorrect. Since we are trying to validate a time series set of data, we need to use a method that uses a rolling origin with day n as training data and day n+1 as test data. The LOOCV approach doesn’t give us this option. (See the article [K-Fold and Other Cross-Validation Techniques](https://medium.com/datadriveninvestor/k-fold-and-other-cross-validation-techniques-6c03a2563f1e))

Option B is incorrect. The K-Fold cross validation technique randomizes the test dataset. We cannot randomize our test dataset since we are trying to validate a time series set of data. Randomized time series data loses its time related value.

Option C is incorrect. We are trying to cross validate time series data. We cannot randomize the test data because it will lose its time related value.

Option D is correct. The Time Series Cross Validation technique is the correct choice for cross validating a time series dataset. Time series cross validation uses forward chaining where the origin of the forecast moves forward in time. Day n is training data and day n+1 is test data.

**Reference:**

Please see the Amazon Machine Learning developer guide titled [Cross Validation](https://docs.aws.amazon.com/machine-learning/latest/dg/cross-validation.html), and he article [K-Fold and Other Cross-Validation Techniques](https://medium.com/datadriveninvestor/k-fold-and-other-cross-validation-techniques-6c03a2563f1e)

## 

Question: 35

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Select the appropriate model for a given machine learning problem**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a bank. Your bank management team is concerned about a recent increase in fraudulent transactions. You need to build a machine learning model to recognize fraudulent transactions in real time. The following is a sample of the dataset you are using to train your model:

| Transaction ID | Account ID | type | amount | source | … | fraud |

| 12576477 | 37564378 | debit | 350.00 | ATM | … | not\_fraud |

| 39844569 | 74897544 | credit | 756.23 | ATM | … | not\_fraud |

| 54986984 | 55656753 | credit | 243.90 | ATM | … | undetermined |

| 34567863 | 27564378 | debit | 1250.00 | ATM | … | fraud |

You are using the fraudulent feature as your label. You have decided to use the linear learner built-in algorithm for your model. Which predictor type should you use for your linear learner?

1. binary\_classifier
2. regressor
3. cross\_entropy\_loss
4. multiclass\_classifier

**Answer:** D

**Explanation:**

Option A is incorrect. The binary\_classifier predictor type is used when your target feature is binary, either yes or no, 1 or 0, etc. Your target feature has three potential values, therefore it is classified across multiple classes.

Option B is incorrect. The regressor predictor type is used for regression models where your target feature is a continuous numeric value. Your target feature has three potential values, therefore it is classified across multiple classes.

Option C is incorrect. The cross\_entropy\_loss is an option for the binary\_classifier\_model\_selection\_criteria parameter of the linear learner SageMaker algorithm. This parameter is used when you are using a binary classifier as your predictor type.

Option D is correct. The multiclass\_classifier predictor type is used when your target feature has more than two potential values. Your target feature has three potential values, therefore it is classified across multiple classes.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Linear Learner Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html), the Amazon SageMaker read the docs api doc titled [LinearLearner](https://sagemaker.readthedocs.io/en/stable/linear_learner.html), and the Amazon Machine Learning developer guide titled [Multiclass Classification](https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-classification.html)

## 

Question: 36

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Select the appropriate model for a given machine learning problem**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a robotics product manufacturer. Your company is trying to use machine learning to help its automatic vacuuming robot determine the most efficient path across the floor of a room. You need to build a machine learning model to accomplish this problem.

Which modeling approach best fits your problem?

1. Multi-Class Classification
2. Simulation-based Reinforcement Learning
3. Binary Classification
4. Heuristic Approach

**Answer:** B

**Explanation:**

Option A is incorrect. Multi-Class Classification is used when your model needs to has many class outcomes from which to choose, as in a car model classification image recognition problem. In this strategy determination problem we need to learn a strategy that optimizes an objective. A Multi-Class Classification approach wouldn’t give you this result.

Option B is correct. Simulation-Based Reinforcement Learning is used in problems where your model needs to learn through trial and error. This is how a robot would best learn the optimal path through a given environment.

Option C is incorrect. Binary Classification is the approach you us when you are trying to predict a binary outcome. This strategy determination problem would not fit a binary classification model.

Option D is incorrect. The Heuristic Approach is used when a machine learning approach is not necessary. An example is the rate of acceleration of a particle through space. There are well known formulas for speed, inertia, and friction that can solve a problem such as this.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Linear Learner Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html), the Amazon SageMaker developer guide titled [Reinforcement Learningwith Amazon SageMaker RL](https://docs.aws.amazon.com/sagemaker/latest/dg/reinforcement-learning.html), the Amazon Machine Learning developer guide titled [Multiclass Classification](https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-classification.html), and the article titled [What is the difference between a machine learning algorithm and a heuristic, and when to use each?](https://www.quora.com/What-is-the-difference-between-a-machine-learning-algorithm-and-a-heuristic-and-when-to-use-each)

Question: 37

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Select the appropriate model for a given machine learning problem**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a state highway administration department. Your department is trying to use machine learning to help determine the make and model of cars as they pass a camera on the state highways. You need to build a machine learning model to accomplish this problem.

Which modeling approach best fits your problem?

1. Multi-Class Classification
2. Simulation-based Reinforcement Learning
3. Binary Classification
4. Heuristic Approach

**Answer:** A

**Explanation:**

Option A is correct. Multi-Class Classification is used when your model needs to choose from a finite set of outcomes, such as this car make and model classification image recognition problem.

Option B is incorrect. Simulation-Based Reinforcement Learning is used in problems where your model needs to learn through trial and error. An image recognition problem with a finite set of outcomes is better suited to a multi-class classification model.

Option C is incorrect. Binary Classification is the approach you us when you are trying to predict a binary outcome. This strategy determination problem would not fit a binary classification model since you have a finite set from which to choose that is greater than 2.

Option D is incorrect. The Heuristic Approach is used when a machine learning approach is not necessary. An example is the rate of acceleration of a particle through space. There are well known formulas for speed, inertia, and friction that can solve a problem such as this.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Linear Learner Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html), the Amazon SageMaker developer guide titled [Reinforcement Learningwith Amazon SageMaker RL](https://docs.aws.amazon.com/sagemaker/latest/dg/reinforcement-learning.html), the Amazon Machine Learning developer guide titled [Multiclass Classification](https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-classification.html), and the article titled [What is the difference between a machine learning algorithm and a heuristic, and when to use each?](https://www.quora.com/What-is-the-difference-between-a-machine-learning-algorithm-and-a-heuristic-and-when-to-use-each)

## 

Question: 38

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Select the appropriate model for a given machine learning problem**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a retail pet products chain. Your company is trying to use machine learning to help determine the breed of dogs in the photos your customers tag on Instagram and Twitter. You need to build a machine learning model to accomplish this problem.

Which SageMaker model would you use to best fit your machine learning problem?

1. K-Means
2. Linear Learner
3. Sequence-to-Sequence
4. Neural Topic Model

**Answer:** B

**Explanation:**

Option A is incorrect. K-Means is used to find discrete groupings in data. It is mostly used on numeric data that is continuous. Image data is not numeric and is not continuous, so K-Means would not be a good model for your dog image classification problem. (See the Amazon SageMaker developer guide titled [K-Means Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/k-means.html))

Option B is correct. The Linear Learner model is used to solve classification problems such as image classification. (See the Amazon SageMaker developer guide titled [Linear Learner Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html))

Option C is incorrect. The Sequence-to-Sequence model is used to take a sequence of tokens and produces another sequence of tokens. It is used for problems like language translation, text summarization, and speech-to-text. (See the Amazon SageMaker developer guide titled [Sequence-to-Sequence Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/seq-2-seq.html))

Option D is incorrect. The Neural Topic Model algorithm is used to organize documents into topics. This type of model is not suited to image classification. (See the Amazon SageMaker developer guide titled [Neural Topic Model (NTM) Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/ntm.html))

**Reference:**

Please see the Amazon SageMaker developer guide titled [Use Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html)

## 

Question: 39

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Recommend and Implement the appropriate machine learning services and features for a given problem**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

You are building a machine learning model for your user behavior prediction problem using your company’s user interaction data stored in DynamoDB. You want to get your data into CSV format and load it into an S3 bucket so you can use it for your machine learning algorithm to give personalized recommendations to your users. Your data set needs to be updated automatically in order to produce real-time recommendations. Your business analysts also want to have the ability to run ad hoc queries on your data.

Which of the following architectures will be the most efficient way to achieve this?

1. Use AWS Data Pipeline to coordinate the following set of tasks: export DynamoDB data to S3 as JSON; Convert JSON to CSV; SageMaker model uses the data to produce real-time predictions; analysts use Amazon Athena to perform ad hoc queries against the CSV data in S3
2. Create a custom classifier in an AWS Glue ETL job that extracts the DynamoDB data to CSV format on your S3 bucket; run your SageMaker model on the new data to produce real-time recommendations; analysts use Amazon Athena to perform ad hoc queries against the CSV data in S3
3. Use AWS DMS to connect to your DynamoDB database and export the data to S3 in CSV format; run your SageMaker model on the new data to produce real-time recommendations; analysts use Amazon Athena to perform ad hoc queries against the CSV data in S3
4. Use Kinesis Data Streams to receive the data from DynamoDB; use an ETL job running on an EC2 instance to consume the data and produce the CSV representation; run your SageMaker model on the new data to produce real-time recommendations; analysts use Amazon Athena to perform ad hoc queries against the CSV data in S3

**Answer:** A

**Explanation:**

Option A is correct. AWS Data Pipeline is used here to schedule frequent runs of the described workflow: DynamoDB export, transformation, and running the model to give real-time predictions.

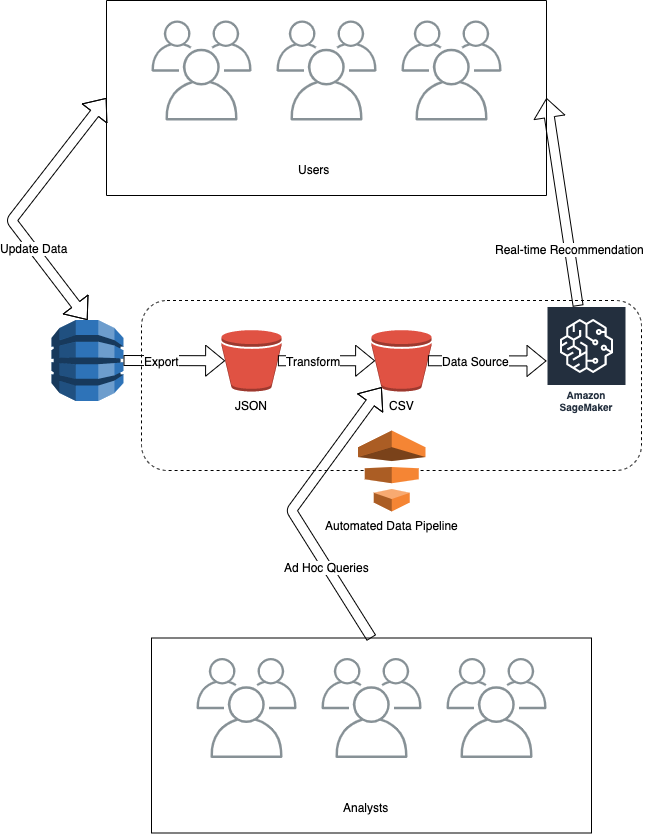
Option B is incorrect. This approach lacks the pipeline coordination described in Option A.

Option C is incorrect. AWS DMS does not support DynamoDB as a data source. Also, this approach lacks the pipeline coordination described in Option A.

Option D is incorrect. You would have to write more code to make this option work when compared to option A. You would need to write an extraction job to make the DynamoDB data into a Kinesis producer. You would also have to write the consumer ETL job. Also, this approach lacks the pipeline coordination described in Option A.

**Reference:**

Please see the AWS Data Pipeline developer guide titled [What is AWS Data Pipeline](https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/what-is-datapipeline.html), and AWS Database Migration Service user guide titled [How AWS Data Migration Service Works](https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Introduction.html) specifically the section on sources, the Amazon Kinesis Data Streams developer guide titled [Amazon Kinesis Data Streams Terminology and Concepts](https://docs.aws.amazon.com/streams/latest/dev/key-concepts.html)

Here is a diagram of the best option:  


## 

Question: 40

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Recommend and Implement the appropriate machine learning services and features for a given problem**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

You are building a machine learning model to use your web server logs to predict which users are most likely to buy a given product. Using your company’s unstructured web server log data stored in S3, you want to get your data into CSV format and load it into another S3 bucket so you can use it for your machine learning algorithm.

Which of the following architectures will be the most efficient way to achieve this?

1. Load the log data into a Redshift cluster; use the UNLOAD Redshift command with a select statement to unload the data in CSV format to S3; SageMaker model uses the data to produce product purchase predictions
2. Use a built-in classifier in an AWS Glue crawler that crawls the web server logs and outputs the log data to CSV format on your ML S3 bucket; SageMaker model uses the data to produce product purchase predictions.
3. Use AWS Schema Conversion tool to convert your web log data to CSV format and output it to your ML S3 bucket; run your SageMaker model on the new data to produce product purchase predictions.
4. Use AWS Snowball Edge and its lambda function capability to convert and then move the web log to S3 in CSV format; run your SageMaker model on the new data toto produce product purchase predictions.

**Answer:** B

**Explanation:**

Option A is incorrect. Using Redshift as an intermediary step in this architecture is an expensive, in terms of implementation effort, extraneous design decision that makes this option less efficient than Option B.

Option B is correct. AWS Glue has built-in classifiers designed specifically for web server log crawling. The crawler will generate CSV formatted data and output it to your ML S3 bucket. This option is the simplest to implement, and therefore the most efficient.

Option C is incorrect. The AWS Schema Conversion tool is used to convert a database from one database engine to another database engine, such as from PostgreSQL to MySQL. The AWS Schema Conversion tool will not work with unstructured web log data.

Option D is incorrect. AWS Snowball Edge is used to move data into and out of AWS. It would not be the most efficient way to transform your web log data to CSV and store it in your ML S3 bucket.

**Reference:**

Please see the Amazon Redshift Database developer guide titled [Unloading Data](https://docs.aws.amazon.com/redshift/latest/dg/c_unloading_data.html), and Amazon Machine Learning developer guide titled [Creating an Amazon ML Datasource from Data in Amazon Redshift](https://docs.aws.amazon.com/machine-learning/latest/dg/using-amazon-redshift-with-amazon-ml.html), the AWS Schema Conversion Tool user guide titled [What is the AWS Schema Conversion Tool?](https://docs.aws.amazon.com/SchemaConversionTool/latest/userguide/CHAP_Welcome.html), and the [Cloud Data Migration Guide](https://aws.amazon.com/cloud-data-migration/), specifically the section on AWS Snowball Edge, and the AWS Glue developer guide titled [Adding Classifiers to a Crawler](https://docs.aws.amazon.com/glue/latest/dg/add-classifier.html)

Question: 41

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Perform feature engineering**

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a language translation software company. Your company needs to move from traditional translation software to a machine learning model based approach that produces the translations accurately. One of your first tasks is to take text given in the form of a document and use a histogram to measure the occurrence of individual words in the document for use in document classification.

Which of the following text feature engineering techniques is the best solution for this task?

1. Orthogonal Sparse Bigram (OSB)
2. Term Frequency-Inverse Document Frequency (tf-idf)
3. Bag-of-Words
4. N-Gram

**Answer:** C

**Explanation:**

Option A is incorrect. The Orthogonal Sparse Bigram natural language processing algorithm creates groups of words and outputs the pairs of words that includes the first word. You are trying to measure the occurrence of individual words.

Option B is incorrect. Term Frequency-Inverse Document Frequency determines how important a word is in a document by giving weights to words that are common and less common in the document. You are not trying to determine the importance of the words in your document, just the count of the individual words.

Option C is correct. The Bag-of-Words natural language processing algorithm creates tokens of the input document text and outputs a statistical depiction of the text. The statistical depiction, such as a histogram, shows the count of each word in the document.

Option D is incorrect. The N-Gram natural language processing algorithm is used to find multi-word phrases in the text of a document. You are not trying to find multi-word phrases, you are just trying to find the count of the individual words.

**Reference:**

Please see the article titled [Introduction to Natural Language Processing for Text](https://towardsdatascience.com/introduction-to-natural-language-processing-for-text-df845750fb63)

## 

Question: 42

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Perform feature engineering**

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a marketing firm that wants to analyze Twitter user stream data to find popular subjects among users who buy products produced by the firm’s clients. You need to analyze the streamed text to find important or relevant repeated common words and phrases and correlate this data to client products. You’ll then include these topics in your client product marketing material.

Which of the following text feature engineering techniques is the best solution for this task?

1. Orthogonal Sparse Bigram (OSB)
2. Term Frequency-Inverse Document Frequency (tf-idf)
3. Bag-of-Words
4. N-Gram

**Answer:** C

**Explanation:**

Option A is incorrect. The Orthogonal Sparse Bigram natural language processing algorithm creates groups of words and outputs the pairs of words that includes the first word. You are trying to determine how important a word is in a document by finding relevant repeated common words.

Option B is correct. Term Frequency-Inverse Document Frequency determines how important a word is in a document by giving weights to words that are common and less common in the document. You can use this information to select the most important repeated phrases in the user’s tweets in your client marketing material.

Option C is incorrect. The Bag-of-Words natural language processing algorithm creates tokens of the input document text and outputs a statistical depiction of the text. The statistical depiction, such as a histogram, shows the count of each word in the document. You are looking for relevant common repeated phrases, not individual words.

Option D is incorrect. The N-Gram natural language processing algorithm is used to find multi-word phrases in the text of a document. However, it does not weight common words or phrases. You need the weighting aspect of the tf-idf algorithm to find the relevant, important repeated phrases used in the tweets.

**Reference:**

Please see the article titled [Introduction to Natural Language Processing for Text](https://towardsdatascience.com/introduction-to-natural-language-processing-for-text-df845750fb63)

## 

Question: 43

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Select the appropriate model(s) for a given machine learning problem**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for major phone and internet provider. Your customer support department needs to upgrade their phone response systems to reduce the number of human service representatives needed to handle their dramatically increasing call volume. Your senior management team has leveraged off-shore call center services to reduce costs, but they now want to take advantage of voice recognition to automate many of the most frequent support call types, such as “I forgot my password”, or “my internet is down.”

Your management team has assigned you to the team that will implement the machine learning model behind the voice recognition system. Which SageMaker built-in algorithm is the best choice for this problem?

1. Sequence-to-Sequence
2. K-Means
3. Semantic Segmentation
4. Neural Topic Model (NTM)

**Answer:** A

**Explanation:**

Option A is correct. The Sequence-to-Sequence algorithm takes audio as input data and generates a sequence of tokens, such as the words in the audio. This can then be used to provide automated responses to users’ requests.

Option B is incorrect. The K-Means algorithm is used to find groups within data where the members of the group are similar to each other but different from members of other groups. This algorithm will not help you encode speech audio streams.

Option C is incorrect. The semantic segmentation algorithm is used to develop computer vision applications. You are trying to solve a speech recognition problem, so this algorithm would not work for this problem.

Option D is incorrect. The Neural Topic Model algorithm is used to group documents into topics using the statistical distribution of words within the documents. You are trying to solve a speech recognition problem, so this algorithm would not work for this problem.

**Reference:**

Please see the SageMaker developer guide titled [Using Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html)

Question: 44

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Deploy and operationalize machine learning solutions**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

You work as a machine learning specialist for an eyewear manufacturing plant where you have used XGBoost to train a model that uses assembly line image data to categorize contact lenses as malformed or correctly formed. You have engineered your data and used CSV as your Training ContentType. You are now ready to deploy your model using the Amazon SageMaker hosting service.

Assuming you used the default configuration settings, which of the following are true statements about your hosted model? (Select THREE)

1. The training instance class is GPU
2. The algorithm is not parallelizable for distributed training
3. The training data target value should be in the first column of the CSV with no header
4. The training data target value should be in the last column of the CSV with no header
5. The inference data target value should be in the first column of the CSV with no header
6. The inference CSV data has no label column
7. The training instance class is CPU

**Answers:** C, F, G

**Explanation:**

Option A is incorrect. The SageMaker XGBoost currently only supports a CPU instance type for training. (See the Amazon SageMaker developer guide titled [XGBoost Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html) particularly the EC2 Instance Recommendation for the XGBoost Algorithm section)

Option B is incorrect. The XGBoost algorithm is parallelizable and therefore can be deployed on multiple instances for distributed training. (See the Amazon SageMaker developer guide titled [Common Parameters for Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-algo-docker-registry-paths.html))

Option C is correct. From the Amazon SageMaker developer guide titled [XGBoost Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html#xgboost-modes) “For CSV training, the algorithm assumes that the target variable is in the first column and that the CSV does not have a header record”

Option D is incorrect. From the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html) “Amazon SageMaker requires that a CSV file doesn't have a header record and that the target variable is in the first column”

Option E is incorrect. From the Amazon SageMaker developer guide titled [XGBoost Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html#xgboost-modes) “For CSV inference, the algorithm assumes that CSV input does not have the label column”

Option F is correct. From the Amazon SageMaker developer guide titled [XGBoost Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html#xgboost-modes) “For CSV inference, the algorithm assumes that CSV input does not have the label column”

Option G is correct. The SageMaker XGBoost currently only supports a CPU instance type for training. (See the Amazon SageMaker developer guide titled [XGBoost Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html) particularly the EC2 Instance Recommendation for the XGBoost Algorithm section)

**Reference:**

Please see the Amazon SageMaker developer guide titled [Deploy a Model on Amazon SageMaker Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-hosting.html) for an overview of the deployment of a SageMaker model.

## 

Question: 45

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Perform hyperparameter optimization**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for the highway toll collection division of the regional state area. The toll collection division uses cameras to identify car license plates as the cars pass through the various toll gates on the state highways. You are on the team that is using SageMaker Image Classification machine learning to read and classify license plates by state and then identify the actual license plate number.

Very rarely, cars pass through the toll gates with plates from foreign countries, for example Great Britain, or Mexico. The outliers must not adversely affect your model’s predictions.

Which hyperparameter should you set, and to what value, to ensure your model is not adversely impacted by these outliers?

1. feature\_dim set to 5
2. feature\_dim set to 1
3. sample\_size set to 10
4. sample\_size set to 100
5. learning\_rate set to 0.1
6. learning\_rate set to 0.75

**Answer:** E

**Explanation:**

Option A is incorrect. The feature\_dim hyperparameter is a setting on the K-Means and K-Nearest Neighbors algorithms, not the Image Classification algorithm.

Option B is incorrect. The feature\_dim hyperparameter is a setting on the K-Means and K-Nearest Neighbors algorithms, not the Image Classification algorithm.

Option C is incorrect. The sample\_size hyperparameter is a setting on the K-Nearest Neighbors algorithm, not the Image Classification algorithm.

Option D is incorrect. The sample\_size hyperparameter is a setting on the K-Nearest Neighbors algorithm, not the Image Classification algorithm.

Option E is correct. The learning\_rate hyperparameter governs how quickly the model adapts to new or changing data. Valid values range from 0.0 to 1.0. Setting this hyperparameter to a low value, such as 0.1, will make the model learn more slowly and be less sensitive to outliers. This is what you want, you want your model to not be adversely impacted by outlier data.

Option F is incorrect. The learning\_rate hyperparameter governs how quickly the model adapts to new or changing data. Valid values range from 0.0 to 1.0. Setting this hyperparameter to a high value, such as 0.75, will make the model learn more quickly but be sensitive to outliers. This is not what you want, you want your model to not be adversely impacted by outlier data.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Image Classification Hyperparameters](https://docs.aws.amazon.com/sagemaker/latest/dg/IC-Hyperparameter.html), and the Amazon SageMaker developer guide titled [Use Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html)

Question: 46

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Deploy and operationalize machine learning solutions**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

You work as a machine learning specialist for a major oil refinery company. Your company needs to do complex analysis on its crude and oil chemical compound structures. You have selected an algorithm for your machine learning model that is not one of the SageMaker built-in algorithms. You have created your model using CreateModel and you have created your HTTPS endpoint. Your docker container running your model is now ready to receive inference requests for real-time inferences. When SageMaker returns the inference result from a client's request which of the following are true? (Select THREE)

1. To receive inference requests your inference container must have a web server running on port 8080
2. Your inference container must accept GET requests to the /invocations endpoint
3. Your inference container must accept PUT requests to the /inferences endpoint
4. Amazon SageMaker strips all POST headers except those supported by InvokeEndpoint. Amazon SageMaker might add additional headers. Your inference container must be able to safely ignore these additional headers
5. Your inference container must accept POST requests to the /inferences endpoint
6. Your inference container must accept POST requests to the /invocations endpoint

**Answers:** A, D, F

**Explanation:**

Option A is correct. From the Amazon SageMaker developer guide titled [Use You Own Inference Code with Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-inference-code.html) “To receive inference requests, the container must have a web server listening on port 8080”

Option B is incorrect. The inference container must accept POST requests to the /invocations endpoint. (See the Amazon SageMaker developer guide titled [Use You Own Inference Code with Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-inference-code.html))

Option C is incorrect. The inference container must accept POST requests to the /invocations endpoint. (See the Amazon SageMaker developer guide titled [Use You Own Inference Code with Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-inference-code.html))Maker developer guide titled [Deploy a Model on Amazon SageMaker Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-hosting.html))

Option D is correct. From the Amazon SageMaker developer guide titled [Use You Own Inference Code with Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-inference-code.html) “Amazon SageMaker strips all POST headers except those supported by InvokeEndpoint. Amazon SageMaker might add additional headers. Inference containers must be able to safely ignore these additional headers.”

Option E is incorrect. The inference container must accept POST requests to the /invocations endpoint. (See the Amazon SageMaker developer guide titled [Use You Own Inference Code with Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-inference-code.html))

Option F is correct. The inference container must accept POST requests to the /invocations endpoint. (See the Amazon SageMaker developer guide titled [Use You Own Inference Code with Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-inference-code.html))

**Reference:**

Please see the Amazon SageMaker developer guide titled [Deploy a Model](https://docs.aws.amazon.com/sagemaker/latest/dg/deploy-model.html)

## 

Question: 47

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Frame business problems as machine learning problems**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a personal care product manufacturer. You are creating a binary classification model that you want to use to predict whether a customer is likely to positively respond to toothbrush and toothpaste sample mailed to their house. Since your company incurs expenses for the products and the shipping when sending samples, you only want to send your samples to customers who you believe have a high probability of buying your products. When analyzing if a customer will follow up with a purchase, which outcome will you want to minimize in your confusion matrix to save costs?

1. True Negative
2. False Negative
3. False Affirmative
4. True Positive
5. False Positive

**Answer:** E

**Explanation:**

Option A is incorrect. True Negatives are definitely not an outcome you want to minimize because you definitely don’t want to send samples to customers who will not respond.

Option B is incorrect. You don’t need to limit False Negatives as much as false positives, since False Negatives only omit customers with a higher probability of following up. Not sending a sample to these customers won’t save costs.

Option C is incorrect. The terms used in a confusion matrix are: True Positive, False Negative, True Negative, and False Positive.

Option D is incorrect. True Positives are the ones to which you want to send your samples.

Option E is correct. You use a confusion matrix, or table, to describe the performance of a classification model on a set of test data when you know the true values. It’s called a confusion matrix because it shows when one class is mislabeled (or confused) as another. For example, when the observation is negative but the model prediction is positive (a False Positive). To reduce the number of mailings to customers who probably won’t follow up with a purchase, you want to limit False Positives.

**Reference:**

Please see the Wikipedia article titled [Confusion Matrix](https://en.wikipedia.org/wiki/Confusion_matrix)

Question: 48

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Train machine learning models**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a clothing manufacturer. You have built a linear regression model using SageMaker’s built-in linear learner algorithm to predict sales for a given year. Your training dataset observations are based on a number of features such as marketing dollars spent, number of active stores, traffic per store, online traffic to the company website, overall market indicators, etc. You have decided to use the k-fold method of cross-validation to assess how the results of your model will generalize beyond your training data.

Which of these will indicate that you don’t have biased training data?

1. The variance of the estimate increases as you increase k
2. You shouldn’t have to worry about bias because your error function removes bias in the data
3. Every k-fold cross-validation round increases the training error rate
4. Every k-fold cross-validation round has a very similar error rate to the rate of all the other rounds
5. You would not normally use k-fold with linear regression models

**Answer:** D

**Explanation:**

Option A is incorrect. When using k-fold for cross-validation the variance of the estimate is reduced as you increase k. So a 10-fold cross-validation should have lower variance than a 5-fold cross-validation.

Option B is incorrect. The k-fold error function just gives you the error rate of the cross-validation round, it doesn’t resolve bias.

Option C is incorrect. The goal of k-fold cross validation is to produce relatively equal error rates for each round (indicating proper randomization of the data) not to reduce the error rate for each round.

Option D is correct. If you have relatively equal error rates for all k-fold rounds it is an indication that you have properly randomized your test data, therefore reducing the chance of bias.

Option E is incorrect. The k-fold cross-validation technique is commonly used with linear regression analysis.

**Reference:**

Please see the Amazon Machine Learning developer guide titled [Evaluating ML Models](https://docs.aws.amazon.com/machine-learning/latest/dg/evaluating_models.html), and the Amazon Machine Learning developer guide titled [Cross Validation](https://docs.aws.amazon.com/machine-learning/latest/dg/cross-validation.html)

Question: 49

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Select the appropriate model for a given machine learning problem**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for the National Oceanic and Atmospheric Administration (NOAA Research). NOAA has developed a great white shark detection program to help warn shore populations when the sharks are in the area of a populated beach. You have the assignment to use your machine learning expertise to decide where to place 10 high tech shark detection sensors on the oceanic floor as part of a pilot to determine if the NOAA should invest broadly in these sensors, which are very expensive. You have great white sightings data from around the globe gathered over the past several years to use as your model training and test data. The model dataset contains several useful features such as the longitude and latitude of each sighting.

You have decided to use an unsupervised learning algorithm that attempts to find discrete groupings within the data. Specifically, you want to find similarities in the longitude and latitude and find groupings of these. You need to produce 10 longitude and latitude pairs to determine where to place the sensors.

Which algorithm can you use in SageMaker that best suits this task?

1. Linear Learner
2. Neural Topic Model
3. K-Means
4. Random Cut Forest
5. Semantic Segmentation
6. XGBoost

**Answer:** C

**Explanation:**

Option A is incorrect. From the Amazon SageMaker developer guide titled [Linear Learner Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html) “*Linear models* are supervised learning algorithms used for solving either classification or regression problems.” But you are trying to solve a data clustering problem so you can find the ten best clustered sightings in order to determine where to place your shark detection sensors.

Option B is incorrect. From the Amazon SageMaker developer guide titled [Neural Topic Model (NTM) Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/ntm.html) “Amazon SageMaker NTM is an unsupervised learning algorithm that is used to organize a corpus of documents into *topics* that contain word groupings based on their statistical distribution.” So this algorithm is used for natural language processing, not data clustering.

Option C is correct. The k-means algorithm is a clustering algorithm. From the Amazon SageMaker developer guide titled [K-Means Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/k-means.html) “K-means is an unsupervised learning algorithm. It attempts to find discrete groupings within data, where members of a group are as similar as possible to one another and as different as possible from members of other groups.” By setting the k hyperparameter to 10, this algorithm will allow you to find the 10 best groupings of shark sightings around the world.

Option D is incorrect. From the Amazon SageMaker developer guide titled [Random Cut Forest (RCF) Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/randomcutforest.html) “Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a data set.” But you are trying to solve a data clustering problem so you can find the ten best clustered sightings in order to determine where to place your shark detection sensors.

Option E is incorrect. From the Amazon SageMaker developer guide titled [Semantic Segmentation Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/semantic-segmentation.html) “The Amazon SageMaker semantic segmentation algorithm provides a fine-grained, pixel-level approach to developing computer vision applications.” So the Semantic Segmentation algorithm is used for computer vision applications, but you are trying to solve a data clustering problem.

Option F is incorrect. The XGBoost algorithm is a gradient boosting algorithm. From the Amazon SageMaker developer guide titled XGBoost Algorithm “gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable by combining an ensemble of estimates from a set of simpler, weaker models.” You are not trying to predict a target value, you are trying to find discrete groupings in your dataset.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Use Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html)

Question: 50

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Select the appropriate model for a given machine learning problem**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a sports analytics company. Your company has been contracted by the Major League Baseball Association to perform real-time analytics on baseball statistics as baseball plays unfold live on national television. Your first assignment is to predict the outcome of situational set plays (such as stolen bases or pitch results) as they are about to unfold. Therefore, your model must deliver its predictions in close to real-time.

You have decided to use a SageMaker built-in algorithm. You have looked at classical forecasting methods like autoregressive integrated moving average (ARIMA) and exponential smoothing (ETS) which use one model for each time series in your data. However, you have many time series over which to train.

Based on your performance requirements and your training requirements, which SageMaker built-in algorithm should you use?

1. Linear Learner
2. Neural Topic Model
3. K-Means
4. Random Cut Forest
5. DeepAR Forecasting
6. XGBoost

**Answer:** E

**Explanation:**

Option A is incorrect. From the Amazon SageMaker developer guide titled [Linear Learner Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html) “*Linear models* are supervised learning algorithms used for solving either classification or regression problems.” But you are trying to solve a one-dimensional time series problem so you can extrapolate baseball play time series into the future.

Option B is incorrect. From the Amazon SageMaker developer guide titled [Neural Topic Model (NTM) Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/ntm.html) “Amazon SageMaker NTM is an unsupervised learning algorithm that is used to organize a corpus of documents into *topics* that contain word groupings based on their statistical distribution.” So this algorithm is used for natural language processing, not time series problems.

Option C is incorrect. The k-means algorithm is a clustering algorithm. From the Amazon SageMaker developer guide titled [K-Means Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/k-means.html) “K-means is an unsupervised learning algorithm. It attempts to find discrete groupings within data, where members of a group are as similar as possible to one another and as different as possible from members of other groups.” You are trying to solve a one-dimensional time series problem so you can extrapolate play time series into the future, not a data clustering problem.

Option D is incorrect. From the Amazon SageMaker developer guide titled [Random Cut Forest (RCF) Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/randomcutforest.html) “Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a data set.” But you are trying to solve a one-dimensional time series problem so you can extrapolate baseball play time series into the future.

Option E is correct. From the Amazon SageMaker developer guide titled [DeepAR Forecasting Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html) “... you have many similar time series across a set of cross-sectional units. For example, you might have time series groupings for demand for different products, server loads, and requests for webpages. For this type of application, you can benefit from training a single model jointly over all of the time series. DeepAR takes this approach. When your dataset contains hundreds of related time series, DeepAR outperforms the standard ARIMA and ETS methods. You can also use the trained model to generate forecasts for new time series that are similar to the ones it has been trained on.” Also from the same developer guide “The training input for the DeepAR algorithm is one or, preferably, more target time series that have been generated by the same process or similar processes. Based on this input dataset, the algorithm trains a model that learns an approximation of this process/processes and uses it to predict how the target time series evolves.” So the DeepAR algorithm is used for one-dimensional time series problems for complex analysis like baseball play prediction.

Option F is incorrect. The XGBoost algorithm is a gradient boosting algorithm. From the Amazon SageMaker developer guide titled XGBoost Algorithm “gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable by combining an ensemble of estimates from a set of simpler, weaker models.” You are not trying to predict a target value, you are trying to solve a one-dimensional time series problem.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Use Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html), the AWS Machine Learning Blog titled [Now Available in Amazon SageMaker: DeepAR algorithm for more accurate time series forecasting](https://aws.amazon.com/blogs/machine-learning/now-available-in-amazon-sagemaker-deepar-algorithm-for-more-accurate-time-series-forecasting/), and the AWS StatCast AI page titled [See how AI on AWS gives baseball fans new insights into the game](https://aws.amazon.com/statcastai/)

Question: 51

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Build machine learning solutions for performance, availability, scalability, resiliency, and fault tolerance**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

You work as a machine learning specialist for a flight data company. Your company has a contract with the US National Defence to produce real-time prediction capabilities for fighter jet flight assist software. Due to the nature of the use case, the implementation of the algorithm you choose for your machine learning model must be able to perform predictions in as close to real-time as possible.

You are in the development stages and have chosen to use the DeepAR SageMaker built-in deep learning model. You are setting up your jupyter notebook instance in SageMaker. Which of the following jupyter notebook settings will allow you to test and evaluate production performance when you are building your models?

1. Notebook instance type
2. Lifecycle configuration
3. Volume size
4. Elastic inference
5. Primary container

**Answer:** E

**Explanation:**

Option A is incorrect. This is the type of EC2 instance on which your notebook will run. This won’t help you understand production performance.

Option B is incorrect. The lifecycle configuration allows you to customize your notebook environment with default scripts and plugins. Default jupyter notebook scripts and plugins won’t give you any insight into production performance.

Option C is incorrect. The volume size is just the size of the jupyter instance in GBs. This won’t give you any insight into production performance.

Option D is correct. From the Amazon SageMaker developer guide titled [Amazon SageMaker Elastic Inference (EI)](https://docs.aws.amazon.com/sagemaker/latest/dg/ei.html#ei-how-it-works) “By using Amazon Elastic Inference (EI), you can speed up the throughput and decrease the latency of getting real-time inferences from your deep learning models … You can also add an EI accelerator to an Amazon SageMaker [notebook instance](https://docs.aws.amazon.com/sagemaker/latest/dg/nbi.html) so that you can test and evaluate inference performance when you are building your models” Therefore, while you are in the development stage using jupyter notebooks, Elastic Inference allows you to gain insight into the production performance of your model once it is deployed.

Option E is incorrect. From the Amazon SageMaker developer guide titled [CreateModel](https://docs.aws.amazon.com/sagemaker/latest/dg/API_CreateModel.html) “... you name the model and describe a primary container. For the primary container, you specify the docker image containing inference code, artifacts (from prior training), and custom environment map that the inference code uses when you deploy the model for predictions.

Use this API to create a model if you want to use Amazon SageMaker hosting services or run a batch transform job.” So the primary container is a parameter used in the CreateModel request when you are creating a model in SageMaker. It is not used when setting up your jupyter notebook.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Amazon SageMaker Elastic Inference (EI)](https://docs.aws.amazon.com/sagemaker/latest/dg/ei.html#ei-how-it-works), the AWS FAQ titled [Amazon Elastic Inference FAQs](https://aws.amazon.com/machine-learning/elastic-inference/faqs/), and the AWS Machine Learning blog titled [Optimizing costs in Amazon Elastic Inference with TensorFlow](https://aws.amazon.com/blogs/machine-learning/optimizing-costs-in-amazon-elastic-inference-with-amazon-tensorflow/)

Question: 52

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Build machine learning solutions for performance, availability, scalability, resiliency, and fault tolerance**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

You work as a machine learning specialist for a polling research company. You have national polling data for the last 10 presidential elections that you have engineered, randomized, partitioned into various training and test datasets, and stored on S3. You have selected a SageMaker built-in algorithm to use for your model. Your training datasets are very large. As you repeatedly run your training job with different large datasets you find your training is taking a very long time.

How can you improve the performance of your training runs? (Select TWO)

1. Use the protobuf recordIO format
2. Convert your data to XML and use file mode to load your data to the EBS training instance volumes
3. Use pipe mode to stream the training data directly to your EBS training instance volumes
4. Convert your data to CSV and use file mode to load your data to the EBS training instance volumes
5. Change your Elastic Inference accelerator type to a larger instance type

**Answers:** A, C

**Explanation:**

Option A is correct. The protobuf recordIO format, used for training data, is the optimal way to load data into your model for training. (See the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html))

Option B is incorrect. XML is not a supported data format for training in SageMaker. (See the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html))

Option C is correct. When you use the protobuf recordIO format you can also take advantage of pipe mode when training your model. Pipe mode, used together with the protobuf recordIO format, gives you the best data load performance by streaming your data directly from S3 to your EBS volumes used byu your training instance. (See the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html))

Option D is incorrect. When you use the CSV format and file mode all of your data is loaded from S3 to the EBS volumes used by your training instance. This is much less efficient from a performance perspective than streaming the training data directly from S3 to your EBS volumes used by your training instance. (See the Amazon SageMaker developer guide titled [Common Data Formats for Training](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html))

Option E is incorrect. Elastic Inference is used to speed up the throughput of retrieving real-time inferences from models deployed as SageMaker hosted models. Elastic Inference accelerators accelerate your inference calls, they aren’t used while training. (See the Amazon SageMaker developer guide titled [Amazon SageMaker Elastic Inference (EI)](https://docs.aws.amazon.com/sagemaker/latest/dg/ei.html))

**Reference:**

Please see the Amazon SageMaker developer guide titled [Common Data Formats for Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-algo-common-data-formats.html) and the AWS FAQ titled [Amazon Elastic Inference FAQs](https://aws.amazon.com/machine-learning/elastic-inference/faqs/)

Question: 53

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Identify and implement a data ingestion solution**

**Domain:** Data Engineering

**Question text**:

You work for a financial services company where you have a large Hadoop cluster hosting a data lake in your on premises data center. Your department has loaded your data lake with financial services operational data from your corporate actions, order management, cash management, reconciliations, and trade management systems. Your investment management operations team now wants to use data from the data lake to build financial prediction models. You want to use data from the Hadoop cluster in your machine learning training jobs. Your Hadoop cluster has Hive, Spark, Sqoop, and Flume installed.

How can you most effectively load data from your Hadoop cluster into you SageMaker model for training?

1. Use the distcp utility to copy your dataset from your hadoop platform to the S3 bucket where your SageMaker training job can use it
2. Use the HadoopActivity command with AWS Data Pipeline to move your dataset from your hadoop platform to the S3 bucket where your SageMaker training job can use it
3. Use the SageMaker Spark library using the data frames in your Spark clusters to train your model
4. Use the Sqoop export command to export your dataset from your Hadoop cluster to the S3 bucket where your SageMaker training job can use it

**Answer:** C

**Explanation:**

Option A is incorrect. The Hadoop distcp utility is used for inter/intra cluster data movement. It is not an efficient method to get data into your SageMaker training instance. (See the [Apache Hadoop distcp guide](https://hadoop.apache.org/docs/current/hadoop-distcp/DistCp.html))

Option B is incorrect. The HadoopActivity command is used to run a job on a cluster. You would have to write the job to extract and load the data onto S3. This would not be the most efficient method of the options listed. (See AWS Data Pipeline developer guide titled [HadoopActivity](https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-object-hadoopactivity.html))

Option C is correct. The SageMaker Spark library that makes it so you can easily train models using data frames in your Spark clusters. This is the most efficient method of the options listed. (See the Amazon SageMaker developer guide titled [Use Apache Spark with Amazon SageMaker](https://docs.aws.amazon.com/sagemaker/latest/dg/apache-spark.html))

Option D is incorrect. The Sqoop export command is used for exporting files from HDFS to an RDBMS. This would not help you load your data into your SageMaker training instance. (See the [Sqoop User Guide](http://sqoop.apache.org/docs/1.4.2/SqoopUserGuide.html#_purpose_3))

**Reference:**

Please see the Amazon SageMaker developer guide titled [Use Machine Learning Frameworks with Amazon SageMaker](https://docs.aws.amazon.com/sagemaker/latest/dg/frameworks.html)

Question: 54

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Sanitize and prepare data for modeling**

**Domain:** Exploratory Data Analysis

**Question text**:

You are working for a consulting firm in their machine learning practice. Your current client is a sports equipment manufacturer. You are building a linear regression model to predict ski and snowboard sales based on the daily snowfall in various regions around the country.

After you have cleaned your CSV data, which of the following tasks would you perform next?

1. Use the scikit-learn cross\_validate method to evaluate the estimation precision of your model
2. Load your data into a pandas DataFrame and remove header rows and any superfluous features
3. Use one-hot encoding to convert categorical values, such as ‘region of the country’ to numerical values
4. Randomize your data using a shuffling technique

**Answer:** D

**Explanation:**

Option A is incorrect. The scikit-learn cross\_validate method is used to evaluate your model’s precision while tuning the model’s hyperparameters. (See Scikit-Learn user guide titled [cross\_validate](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html#sklearn.model_selection.cross_validate))

Option B is incorrect. Using a Pandas DataFrame to remove superfluous rows and features is part of cleaning you data, which you have already done.

Option C is incorrect. One-hot encoding is another way to clean your data in preparation for training. You have already completed the cleaning of your data.

Option D is correct. For a linear regression model, once you have cleaned your data you need to randomize the data to prevent overfitting and to reduce variance. (See Amazon Machine Learning developer guide titled [The Amazon Machine Learning Process](https://docs.aws.amazon.com/machine-learning/latest/dg/the-machine-learning-process.html))

**Reference:**

Please see the Amazon Machine Learning developer guide titled [Machine Learning Concepts](https://docs.aws.amazon.com/machine-learning/latest/dg/machine-learning-concepts.html), and the Amazon Machine Learning developer guide titled [The Amazon Machine Learning Process](https://docs.aws.amazon.com/machine-learning/latest/dg/the-machine-learning-process.html)

Question: 55

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Sanitize and prepare data for modeling**

**Domain:** Exploratory Data Analysis

**Question text**:

You work as a machine learning specialist at a retail shoe manufacturer. Your marketing department wants to do a promotion for a new running shoe they are about to release into their product pipeline. They need a model to predict sales of the new shoe using the purchase history of their registered customers based on past releases of new shoes.

You have decided to use a linear regression algorithm for your model. Your data has thousands of observations and 35 numeric features. While doing analysis to better understand your data you find 25 observations that have what looks like outlier data points. After speaking to your marketing department you learn that these values are valid. You also find several hundred observations that have some blank feature values.

How should you correct the outlier and blank feature problems?

1. Remove the observations with the outlier data points and replace the blank values with the null value
2. Remove the outlier and blank value observations
3. Remove the observations with the outlier data points and replace the blank values with the mean value
4. Remove the observations with the outlier data points and replace the blank values with the value 0

**Answer:** C

**Explanation:**

Option A is incorrect. Null values in an observation should be replaced since linear regression calculations will have a problem with null values. Therefore, you would not replace empty fields with null.

Option B is incorrect. Removing the observations with blank values will reduce the accuracy of your model’s predictions since you have removed many features from the training dataset.

Option C is correct. You should remove the outlier observations. You should also replace the blank values with a meaningful value. The mean value is the best option of those listed.

Option D is incorrect. You should remove the outlier observations. You should also replace the blank values with a meaningful value. The 0 value is not the best option of those listed because the mean is invariably a better approximation than 0 for a continuous numeric value.

**Reference:**

Please see the Amazon Machine Learning developer guide titled [Feature Processing](https://docs.aws.amazon.com/machine-learning/latest/dg/feature-processing.html)

Question: 56

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Create data repositories for machine learning**

**Domain:** Data Engineering

**Question text**:

You work as a machine learning specialist at a hedge fund firm. Your firm is working on a new quant algorithm to predict when to enter and exit holdings in their portfolio. You are building a machine learning model to predict these entry and exit points in time. You have cleaned your data and you are now ready to split the data into training and test datasets.

Which splitting technique is best suited to your model’s requirements?

1. Use k-fold cross validation to split the data
2. Sequentially splitting the data
3. Randomly splitting the data
4. Categorically splitting the data by holding

**Answer:** B

**Explanation:**

Option A is incorrect. Using k-fold cross validation will randomly split your data, but you need to consider the time-series nature of your data when splitting. So randomizing the data would eliminate the time element of your observations, making the datasets unusable for predicting price changes over time.

Option B is correct. By sequentially splitting the data you preserve the time element of your observations.

Option C is incorrect. Randomly splitting the data would eliminate the time element of your observations, making the datasets unusable for predicting price changes over time.

Option D is incorrect. If you split the data by a category such as the holding attribute you would create imbalanced training and test dataset since some holdings would only be in the training dataset and others would only be in the test dataset.

**Reference:**

Please see the Amazon Machine Learning developer guide titled [Splitting Your Data](https://docs.aws.amazon.com/machine-learning/latest/dg/splitting-types.html)

Question: 57

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Create data repositories for machine learning**

**Domain:** Data Engineering

**Question text**:

You work as a machine learning specialist for a software company that is developing a movie rating social media site where users can rate movies. You want to use your companies data to predict the ratings distribution of a movie based on the genre of the movie. Your training data contains a genre feature with a set of categories such as documentary, romance, etc. You have sorted your data by the genre feature and then used the Amazon ML sequential split option to split your data into training and test datasets.

When using your test dataset to verify your genre-prediction model you discover that the accuracy rate is very low. What could be the underlying problem?

1. You should have sorted by a different feature before you used the sequential split option
2. You should have split your data categorically by genre
3. You should have split your data sequentially by year
4. You should not have used the sequential split option

**Answer:** D

**Explanation:**

Option A is incorrect. Sorting the data by a different feature wouldn’t solve the problem. You used the sequential option when splitting the data so you have not properly randomized your data.

Option B is incorrect. By categorically splitting the data by definition you will have some genre movies only in the training dataset and others only in the test dataset. This reduces the genre feature to a meaningless datapoint.

Option C is incorrect. Sequentially splitting the data by year wouldn’t solve the problem. You used the sequential option when splitting the data so you have not properly randomized your data.

Option D is correct. You should not have used the sequential option when splitting your data. For this type of problem, in order to get proper generalization from your data, you need to randomize it.

**Reference:**

Please see the Amazon Machine Learning developer guide titled [Splitting Your Data](https://docs.aws.amazon.com/machine-learning/latest/dg/splitting-types.html)

Question: 58

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Analyze and visualize data for machine learning**

**Domain:** Exploratory Data Analysis

**Question text**:

You work as a machine learning specialist for a real estate company. You are using the kaggle housing prices data as your experimentation data to optimize your model before you use your model on the real estate data for your area of the country. You have a hypothesis that you can predict the price of a real estate property based on the foundation type. You have your data from kaggle but you want to make sure your model is not overly influenced by outliers.

What is the quickest way to identify outliers in your data?

1. Arrange your data points from lowest to highest; calculate the median of the data set; use a qualitative assessment to determine whether to remove outliers
2. Calculate the Z-Score for your data points
3. Visualize your data using scatter plots and/or box plots
4. Visualize your data using network and correlation matrices

**Answer:** C

**Explanation:**

Option A is incorrect. You can find your outliers using a quantitative assessment, but it will involve more effort and therefore more time than visualizing your data.

Option B is incorrect. The z-score of a data point shows how many standard deviations the data point is from the mean. This would help you find your outliers but it will involve more effort and therefore more time than visualizing your data.

Option C is correct. With large datasets, such as the real estate data you are using in this problem, the quickest way to find outliers is to visualize your data. The best plots for this task are the scatter plot and the box plot. (See the article titled [How to Make your Machine Learning Models Robust to Outliers](https://heartbeat.fritz.ai/how-to-make-your-machine-learning-models-robust-to-outliers-44d404067d07))

Option D is incorrect. Visualization is the quickest and easiest way to find outliers, but the network and/or correlation matrix charting choices will not show outliers. They are used to represent relations between data points as nodes. These relationships would not give you any information about the extremity of a data point.

**Reference:**

Please see the article titled [How to Make your Machine Learning Models Robust to Outliers](https://heartbeat.fritz.ai/how-to-make-your-machine-learning-models-robust-to-outliers-44d404067d07), and the article titled [A Brief Overview of Outlier Detection Techniques](https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561)

Question: 59

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Select the appropriate model(s) for a given machine learning problem**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a company that runs car rating website. Your company wants to build a price prediction model that is more accurate than their current model, which is a linear regression model using the age of the car as the single independent variable in the regression to predict the price. You have decided to add the horse power, fuel type, city mpg (miles per gallon), drive wheels, and number of doors as independent variables in your model. You believe that adding these additional independent variables will give you a more accurate prediction of price.

Which type of algorithm will you now use for your prediction?

1. Logistic Regression
2. Decision Tree
3. Naive Bayes
4. Multivariate Regression

**Answer:** D

**Explanation:**

Option A is incorrect. Logistic regression is used for problems where you are trying to classify and estimate a discrete value (on or off, 1 or 0) based on a set of independent variables. In your problem you are trying to estimate a continuous numerical value: price, not a binary classification.

Option B is incorrect. A decision tree is a classification algorithm, so it is not a good fit for your continuous numerical value prediction problem.

Option C is incorrect. Naive Bayes is another classification algorithm, so it is not a good fit for your continuous numerical value prediction problem.

Option D is correct. You are trying to predict the price of a car (dependent variable) based on a number of independent variables (horse power, fuel type, city mpg, drive wheels, and number of doors, etc.) The Multivariate Regression algorithm is the best choice for this type of problem. (See the article [Data Science Simplified Part 5: Multivariate Regression Models](https://towardsdatascience.com/data-science-simplified-part-5-multivariate-regression-models-7684b0489015))

**Reference:**

Please see the Amazon Machine Learning developer guide titled [Regression Model Insights](https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html), and the article titled [Commonly Used Machine Learning Algorithms (with Python and R codes)](https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/)

Question: 60

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Analyze and visualize data for machine learning**

**Domain:** Exploratory Data Analysis

**Question text**:

You work as a machine learning specialist for a company that produces a polling data and uses it for predictive modeling. Your company wants to build an election prediction model that uses multiple independent variables such as age of voter, religion, sex, registered affiliation, etc. to predict the candiate for which each observed voter will vote in the upcoming election.

Which type of algorithm is NOT a good choice to use for your prediction? (Select THREE)

1. Ordinary Least Squares Regression (OLSR)
2. Local Outlier Factor (LOF)
3. Naive Bayes
4. Least-Angle Regression (LARS)
5. K-Means

**Answers:** B, C, E

**Explanation:**

Option A is incorrect. Ordinary Least Squares Regression (OLSR) is a regression technique that predicts a dependent variable using one or more independent variables. This is exactly what you are trying to solve.

Option B is correct. The Local Outlier Factor (LOF) algorithm is used to discover outlier data points. So this would NOT be a good choice for your algorithm where you are trying to solve for a dependent variable based on multiple independent variables.

Option C is correct. The Naive Bayes algorithm is used as a classifier. So this would NOT be a good choice for your algorithm where you are trying to solve for a dependent variable based on multiple independent variables.

Option D is incorrect. Least-Angle Regression (LARS) is also a regression technique that predicts a dependent variable using one or more independent variables. This is exactly what you are trying to solve.

Option E is correct. The K-Means algorithm is used as a clustering algorithm, so it would NOT be a good choice for your algorithm where you are trying to solve for a dependent variable based on multiple independent variables.

**Reference:**

Please see the Amazon Machine Learning developer guide titled [Regression Model Insights](https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html), and the article titled [A Tour of the Most Popular Machine Learning Algorithms](https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/)

Question: 61

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Identify and Implement a data-transformation solution**

**Domain:** Data Engineering

**Question text**:

You are a machine learning specialist for a research firm. Your team is using Amazon SageMaker and it’s built-in scikit-learn library for feature transformation in your machine learning process. When using the SimpleImputer transformer to replace missing values in your observations, which strategy is the default strategy that your SageMaker scikit-learn code will use if you don’t explicitly pass a strategy parameter?

1. constant
2. most\_frequent
3. median
4. mean
5. mode

**Answer:** D

**Explanation:**

Option A is incorrect. The default strategy is mean. The constant strategy replaces the missing values with a constant you supply.

Option B is incorrect. The default strategy is mean. The most\_frequent strategy replaces the missing values with the most frequent value along each column.

Option C is incorrect. The default strategy is mean. The median strategy replaces the missing values with the median along each column.

Option D is correct. The default strategy is mean. The mean strategy replaces the missing values with the mean along each column.

Option E is incorrect. There is no mode strategy in the SimpleImputer scikit-learn transformer.

**Reference:**

Please see the Amazon Machine Learning blog titled [Preprocess input data before making predictions using Amazon SageMaker inference pipelines and Scikit-learn](https://aws.amazon.com/blogs/machine-learning/preprocess-input-data-before-making-predictions-using-amazon-sagemaker-inference-pipelines-and-scikit-learn/)

Question: 62

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Identify and Implement a data-transformation**

**Domain:** Data Engineering

**Question text**:

You are a machine learning specialist for a gaming software startup. Your company is investigating ways to use machine learning to enhance their game software platform. The team has selected the Amazon SageMaker platform for their machine learning efforts. You are participating in the feature transformation process in preparation to creating your machine learning models. Instead of transforming your data before you use it in your SageMaker models, you and your team have decided to use the built-in transformations of SageMaker. Specifically, you and your team have decided to use the built-in OneHotEncoder transformer to transform your categorical data.

You have decided to drop one of the categories per feature because you suspect you may have perfectly collinear features. Which of the following is NOT a drop methodology used in the OnHotEncoder transformer?

1. None
2. Last
3. Array
4. First

**Answer:** B

**Explanation:**

Option A is incorrect. The OneHotEncoder transformer has the following methodologies you can use to drop one of the categories per feature: None, first, array. None is the default methodology.

Option B is correct. The OneHotEncoder transformer has the following methodologies you can use to drop one of the categories per feature: None, first, array. None is the default methodology. The OneHotEncoder transformer drop parameter does not offer a last methodology.

Option C is incorrect. The OneHotEncoder transformer has the following methodologies you can use to drop one of the categories per feature: None, first, array. None is the default methodology.

Option D is incorrect. The OneHotEncoder transformer has the following methodologies you can use to drop one of the categories per feature: None, first, array. None is the default methodology.

**Reference:**

Please see the Amazon Machine Learning blog titled [Preprocess input data before making predictions using Amazon SageMaker inference pipelines and Scikit-learn](https://aws.amazon.com/blogs/machine-learning/preprocess-input-data-before-making-predictions-using-amazon-sagemaker-inference-pipelines-and-scikit-learn/), and the Scikit-learn api documentation [OneHotEncoder](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html)

Question: 63

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Evaluate machine learning models**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a consulting firm that has the NFL as a client. You are working on the passer completion probability model using statistics from in-play metrics. You are running your linear learner model in Amazon SageMaker using a CSV file representation of your passer completion probability statistics. You are now running your inference.

Some of the features and their data types are listed below:

| Feature Name | Data Type |

| Passer age | Numeric |

| Length of pass | Numeric |

| Complete (yes/no) | Categorical |

| Feature Name | Data Type |

| Distance between receiver and nearest defender| Numeric |

| Play called (post, crossing, screen, etc.) | Categorical |

You are using the Complete feature as your prediction response feature. You are now making predictions on new data. When you interrogate the response of your model, which of the following do you expect to find?

1. score: the prediction produced by the model
2. score: the prediction produced by the model AND predicted\_class which is an integer from 0 to num\_classes-1
3. score: single floating point number measuring the strength of the prediction AND predicted\_label which is 0 or 1
4. score: the prediction produced by the model OR predicted\_label which is 0 or 1

**Answer:** C

**Explanation:**

Option A is incorrect. For a binary classification (complete yes or no) the model produces a score denoting the strength of the prediction AND a predicted\_label denoting complete or not complete

Option B is incorrect. This option describes the response for a multiclass classification, but you are working with a binary classification.

Option C is correct. For a binary classification (complete yes or no) the model produces a score denoting the strength of the prediction AND a predicted\_label denoting complete or not complete.

Option D is incorrect. For a binary classification (complete yes or no) the model produces a score denoting the strength of the prediction AND a predicted\_label denoting complete or not complete.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Linear Learner Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html)

Question: 64

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Perform hyperparameter optimization**

**Domain:** Modeling

**Question text**:

You work in the machine learning department of a major retail company. Your team is working on a model to predict the region that will have the highest sales for a given quarter. You have selected your observations from past sales cycles for all regions and split your data into training and evaluation datasets. You are now training your linear learner model in Amazon SageMaker and you are trying to select the model hyperparameters that give your team the best predictions.

You have set the predictor\_type hyperparameter to binary\_classifier. Which loss function hyperparameter setting is NOT one of your options?

1. auto
2. logistic
3. hinge\_loss
4. softmax\_loss

**Answer:** D

**Explanation:**

Option A is incorrect. The three hyperparameters values that you can set for the loss function are auto, logistic, and hinge\_loss. The default for auto is logistic.

Option B is incorrect. The three hyperparameters values that you can set for the loss function are auto, logistic, and hinge\_loss. The default for auto is logistic.

Option C is incorrect. The three hyperparameters values that you can set for the loss function are auto, logistic, and hinge\_loss. The default for auto is logistic.

Option D is correct. The three hyperparameters values that you can set for the loss function are auto, logistic, and hinge\_loss. The default for auto is logistic. The softmax\_loss setting is an option if your predictor\_type is set to multiclass\_classifier.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Linear Learner Hyperparameters](https://docs.aws.amazon.com/sagemaker/latest/dg/ll_hyperparameters.html)

Question: 65

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Perform hyperparameter optimization**

**Domain:** Modeling

**Question text**:

You work in the machine learning department of a major retail company. Your team is working on a model to classify customers by purchase history. Your marketing department wants to use the results of your model predictions to determine which customers should receive a new campaign offer. You have selected your observations and cleaned your data. You have also split your data into training and evaluation datasets. You are now training your k-means model in Amazon SageMaker and you are trying to select the model hyperparameters that give your marketing team the best predictions.

You have set the feature\_dim hyperparameter to equal the number of features in your input data. You have set the k hyperparameter to 10, the number of clusters you estimate is appropriate for your model. You have set the epochs hyperparameter to 1 so that the model performs one pass over your data.

You need to report a score for your model. Which k-means hyperparameter allows you to select the metric types to report this scoring, and what are the available metric options?

1. extra\_center\_factor with msd, ssd, or [msd, ssd] as the available metric type values
2. score\_metrics with mse, ssd, or [mse, ssd] as the available metric type values
3. eval\_method with mse, ssd, or [mse, ssd] as the available metric type values
4. eval\_metrics with msd, ssd, or [msd, ssd] as the available metric type values

**Answer:** D

**Explanation:**

Option A is incorrect. The hyperparameter you chose to report a score for your model is the eval\_metrics hyperparameter. The eval\_metrics hyperparameter has the allowed values of msd for Mean Square Error, ssd for Sum of Square Distance, and the option of both msd and ssd. The extra\_center\_factor is used to control the number of clusters.

Option B is incorrect. The hyperparameter you chose to report a score for your model is the eval\_metrics hyperparameter. The eval\_metrics hyperparameter has the allowed values of msd for Mean Square Error, ssd for Sum of Square Distance, and the option of both msd and ssd. The Amazon SageMaker k-means algorithm does not have a score\_metrics hyperparameter.

Option C is incorrect. The hyperparameter you chose to report a score for your model is the eval\_metrics hyperparameter. The eval\_metrics hyperparameter has the allowed values of msd for Mean Square Error, ssd for Sum of Square Distance, and the option of both msd and ssd. The Amazon SageMaker k-means algorithm does not have a eval\_method hyperparameter.

Option D is correct. The hyperparameter you chose to report a score for your model is the eval\_metrics hyperparameter. The eval\_metrics hyperparameter has the allowed values of msd for Mean Square Error, ssd for Sum of Square Distance, and the option of both msd and ssd.

**Reference:**

Please see the Amazon SageMaker developer guide titled [K-Means Hyperparameters](https://docs.aws.amazon.com/sagemaker/latest/dg/k-means-api-config.html)

Question: 66

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Select the appropriate model(s) for a given machine learning problem**

**Domain:** Modeling

**Question text**:

You are a machine learning specialist at a large online retailer. Your team is working on a recommender model for your online purchase workflow. The recommender will suggest similar items to the items the user has viewed or placed in their shopping cart. To find items that are similar to the item your customer is viewing, you want to compare other users who like each item. If these similar users like the same two items, then the probability the items are similar is higher.

Which Amazon SageMaker built-in algorithm is best suited to your use case?

1. Semantic Segmentation
2. K-Nearest Neighbor
3. Linear Learner
4. Random Cut Forest

**Answer:** B

**Explanation:**

Option A is incorrect. The semantic segmentation algorithm is used to develop computer vision applications. You are trying to find items that are similar to each other.

Option B is correct. The k-nearest neighbor algorithm is used to find items that are similar to each other. This is what you need to find similar items to recommend to a user in the online purchase workflow.

Option C is incorrect. The linear learner algorithm is used to show how a change in an independent variable affects a dependent variable. You are trying to find items that are similar to each other.

Option D is incorrect. The random cut forest algorithm is predominantly used to classify observations, such as whether a transaction is fraudulent or not. You are trying to find items that are similar to each other.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Use Amazon Sagemaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html)

Question: 67

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Identify and implement a data-transformation solution**

**Domain:** Data Engineering

**Question text**:

You have just landed a position as a machine learning specialist at a large financial services firm. Your new team is working on a fraud detection model using the SageMaker built-in linear learner algorithm. You are gathering the data required for your machine learning model. The dataset you intend to produce will contain well over 5,000 objects that need to be labeled. Your team wants to control the costs of cleaning your data. Therefore, the team has decided to use SageMaker Ground Truth active learning to automate the labeling of your data.

The Ground Truth automated labeling job initially follows this set of steps:

* Selects a random sample of data
* sends the sample data to human workers
* uses the human-labeled data as validation data
* runs a SageMaker batch transform using the validation set which generates a quality metric used to estimate the potential quality of auto-labeling the rest of the unlabeled data
* runs a SageMaker batch transform on the unlabeled data
* data where the expected quality of automatically labeling the data is above the requested level of accuracy is labeled

After performing the above steps, what does Ground Truth do next to complete the labeling of ALL of your data?

1. Selects a new sample of unlabeled data and sends it to human workers; it uses the existing labeled data to verify the new human-labeled data; repeats this later set of steps until all the data in the dataset is labeled
2. Selects a new sample of unlabeled data and sends it to human workers; it uses the existing labeled data and the new human-labeled data to train a new model; repeats this later set of steps until all the data in the dataset is labeled
3. Selects a new sample of the most hard to identify unlabeled data and sends it to human workers; it uses the existing labeled data to verify the new human-labeled data; repeats this later set of steps until all the data in the dataset is labeled
4. Selects a new sample of the most hard to identify unlabeled data and sends it to human workers; it uses the existing labeled data and the new human-labeled data to train a new model; repeats this later set of steps until all the data in the dataset is labeled

**Answer:** D

**Explanation:**

Option A is incorrect. This option doesn’t articulate that the selection of a new sample looks for the most hard to identify unlabeled data. It also doesn’t state that the new human-labeled data is used with the existing labeled data to train a new model.

Option B is incorrect. This option doesn’t articulate that the selection of a new sample looks for the most hard to identify unlabeled data.

Option C is incorrect. This option doesn’t state that the new human-labeled data is used with the existing labeled data to train a new model.

Option D is correct. This is the set of steps Ground Truth uses to iterate over the unlabeled data using human labelers and model training to complete the labeling of your large dataset.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Amazon SageMaker Ground Truth](https://docs.aws.amazon.com/sagemaker/latest/dg/sms.html), and the Amazon SageMaker developer guide titled [Using Automated Data Labeling](https://docs.aws.amazon.com/sagemaker/latest/dg/sms-automated-labeling.html)

Question: 68

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Perform feature engineering**

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a major banking firm as a machine learning specialist. As part of the bank’s fraud detection team, you are building a machine learning model to detect fraudulent transactions. Using your training dataset you have produced a Receiver Operating Characteristic (ROC) curve and it shows 99.99% accuracy. Your transaction dataset is very large, but 99.99% of the observations in your dataset represent non-fraudulent transactions. Therefore, the fraudulent observations are a minority class. Your dataset is very imbalanced.

Given you have the approval from your management team to produce the most accurate model possible, even if it means spending more time perfecting the model, what is the most effective technique to address the imbalance in your dataset?

1. Synthetic Minority Oversampling Technique (SMOTE) oversampling
2. Random oversampling
3. Generative Adversarial Networks (GANs) oversampling
4. Edited Nearest Neighbor undersampling

**Answer:** C

**Explanation:**

Option A is incorrect. The SMOTE technique creates new observations of the underrepresented class, in this case the fraudulent observations. These synthetic observations are almost identical to the original fraudulent observations. This technique is expeditious, but the types of synthetic observations it produces are not as useful as the unique observations created by other oversampling techniques.

Option B is incorrect. Random oversampling uses copies of some of the minority class observations (randomly selected) to augment the minority class observation set. These observations are exact replicas of existing minority class observations, making them less effective than observations created by other techniques that produce unique synthetic observations.

Option C is correct. The Generative Adversarial Networks (GANs) technique generates unique observations that more closely resemble the real minority observations without being so similar that they are almost identical. This results in more unique observations of your minority class that improve your model’s accuracy by helping to correct the imbalance in your data.

Option D is incorrect. Using an undersampling technique would remove potentially useful majority class observations. Additionally, you would have to remove a very large number of your majority class observations to correct your imbalance that you would render your entire training dataset useless.

**Reference:**

Please see the wikipedia article titled [Oversampling and undersampling in data analysis](https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis), and the article titled [Imbalanced data and credit card fraud](https://medium.com/@hazy_ai/imbalanced-data-and-credit-card-fraud-ad1c1ed011ea)

Question: 69

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Evaluate machine learning models**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a car manufacturer that has developed driverless technology for their new line of cars. These cars require real-time machine learning models to perform all of the tasks of driving. You have trained multiple models, using different algorithms and/or different hyperparameters, as candidates to assist in lane line crossover detection using live data from sensors on the undercarriage of the car. You want to select one of these models as the model to go to production in the line of cars.

Using the various options available from SageMaker, which are the most effective method steps you should use to select the correct model? (Select TWO)

1. Use online testing with historical data
2. Deploy your trained models to beta endpoints, then using a jupyter notebook in your SageMaker instance, send inference requests to each model in turn using the AWS SDK for python or the SageMaker high-level python library and finally evaluate each model.
3. Use online testing with live data
4. Deploy your models to a SageMaker training instance, then train each model on a portion of the live data and finally evaluate each model
5. Deploy your models to a SageMaker endpoint, then send a portion of the live data to each model and finally evaluate each model

**Answers:** C, E

**Explanation:**

Option A is incorrect. For online testing you use live data. For offline testing you use historical data.

Option B is incorrect. When performing offline testing of your models, you deploy your trained models to alpha endpoints, not beta endpoints.

Option C is correct. For online testing you use live data. Testing with live data will allow you to perform the steps listed in option E.

Option D is incorrect. To use online testing, you deploy your models to a SageMaker endpoint, not a SageMaker training instance.

Option E is correct. To perform online testing of your models you deploy the models to a SageMaker endpoint and then send a portion of the data to each model (or production variant) allowing you to evaluate the models.

**Reference:**

Please see the SageMaker developer guide titled [Validate a Machine Learning Model](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-model-validation.html)

Question: 70

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Select the appropriate model(s) for a given machine learning problem**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a large auto parts manufacturing company. You have been tasked with building a machine learning model to analyze images of car parts on your company’s production lines to automatically classify the parts. The classified parts will then be placed in their appropriate warehouse containers by classification.

Some examples of the classifications are: electronics, trim, gasket, hose, etc. Since your company has many manufacturing plants across the globe, your classification model needs to be able to classify millions of high resolution images.

Which algorithm best fits your problem?

1. Object Detection
2. Convolutional Neural Network
3. Latent Dirichlet Allocation (LDA)
4. Factorization Machine

**Answer:** B

**Explanation:**

Option A is incorrect. The Object Detection algorithm is used to identify all instances of an object within an image. While this may be used in a naive approach to the image classification problem, it is not meant for image classification in the way and scale needed for your problem.

Option B is correct. The SageMaker built-in Image Classification algorithm uses a Convolutional Neural Network to classify images that supports multi-label classification. It scales to millions of images at high resolution. It solves this problem through convolution and multiple layers in the neural network. (See the article [AWS SageMaker and CNN for Dog Breed Classification](https://hiregion.org/2018/09/02/aws-sagemaker-and-cnn-for-dog-breed-categorization/))

Option C is incorrect. The Latent Dirichlet Allocation algorithm is used for topic discovery within documents.

Option D is incorrect. The Factorization Machine algorithm can be used to classify observations, but it is used primarily to detect interactions between features. Examples include reaction to ads on a web page, or item recommendation.

**Reference:**

Please see the SageMaker developer guide titled [Using Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html)

Question: 71

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Select the appropriate model(s) for a given machine learning problem**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for medical research facility. Your research team is working on a brain tumor detection scanner to be used in hospitals across the country. The team has decided to use machine learning to detect tumors in the scans and to catalog the findings in a database that can be shared across medical facilities.

You have millions of brain scan data to use in your model. Also, you will have an incoming stream of new scans every day, so your volume is very high. Your research team requires that the model perform at scale and with very high accuracy due to the nature of the consequences of false negative predictions.

Which algorithm best fits your problem?

1. Object Detection
2. K-Means
3. Convolutional Neural Network
4. Random Cut Forest

**Answer:** C

**Explanation:**

Option A is incorrect. The Object Detection algorithm is used to identify all instances of an object within an image. You are trying to classify a high resolution image as either containing a tumor or not. You are not trying to identify, and surrounding with a bounding box, all elements in an image.

Option B is incorrect. The K-Means algorithm is used to find groups within data where the members of the group are similar. This would not work for our image classification problem.

Option C is correct. The SageMaker built-in Image Classification algorithm uses a Convolutional Neural Network to classify images. It breaks up each image into a series of tiles and then predicts what each tile contains. This is the optimal way to find a tumor within a larger brain scan image. (See the article [Image Classification using Deep Neural Networks - A beginner friendly approach using TensorFlow](https://medium.com/@tifa2up/image-classification-using-deep-neural-networks-a-beginner-friendly-approach-using-tensorflow-94b0a090ccd4))

Option D is incorrect. The Random Cut Forest algorithm is used to find abnormal data points with your dataset. It would not be the best choice for your image classification problem with large numbers of high resolution images in which you are trying to detect an anomaly.

**Reference:**

Please see the SageMaker developer guide titled [Using Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html), and the article titled [How might companies use random forest models for predictions?](https://www.techopedia.com/how-might-companies-use-random-forest-models-for-predictions/7/32995)

Question: 72

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Recommend and implement the appropriate machine learning services and features for a given problem**

**Domain:** ML Implementation and Operations

**Question text**:

You work as a machine learning specialist for an online retail company that sells health products. Your company allows users to enter reviews of the products they buy from the website. You want to make sure the reviews do not contain any offensive or unsafe content, such as obscenities or threatening language.

Which Amazon SageMaker algorithm or service will allow you to scan your user’s review text in the simplest way?

1. BlazingText
2. Neural Topic Model (NTM)
3. Semantic Segmentation
4. Comprehend

**Answer:** D

**Explanation:**

Option A is incorrect. The BlazingText algorithm is used for natural language processing tasks like sentiment analysis, and named entity recognition. You should use all of these features when scanning your user’s review text, however the BlazingText algorithm requires more developer effort and time than using the Comprehend service.

Option B is incorrect. The Neural Topic Model algorithm is used to group documents into topics using the statistical distribution of words within the documents. This algorithm would not be the most efficient choice for detecting offensive or unsafe language.

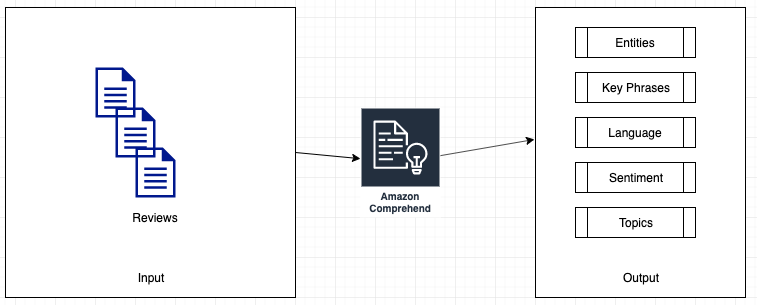
Option C is incorrect. The Semantic Segmentation algorithm is used for computer vision application, so it is not an algorithm you would use for text analysis.

Option D is correct. The Comprehend service scans your unstructured review text and analyzes it using SageMaker Natural Language Processing (NLP) algorithms to find key phrases, entities, and sentiments. This is the most expeditious and efficient option.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Using Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html), and the Amazon Machine Learning blog titled [Analyze content with Amazon Comprehend and Amazon SageMaker notebooks](https://aws.amazon.com/blogs/machine-learning/analyze-content-with-amazon-comprehend-and-amazon-sagemaker-notebooks/)

Here is a diagram of the solution:



Question: 73

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Recommend and implement the appropriate machine learning services and features for a given problem**

**Domain:** ML Implementation and Operations

**Question text**:

You work as a machine learning specialist for a news organization that has a very active online community who contributes comments on your organization’s news articles very frequently. Your news editors wish to use the comments from their users to gain insight into what interests them the most. Instead of just relying on the raw count of comments per article, the editors would like to use machine learning to find the underlying intent of the comments. This will allow them to understand their readers better so that they can provide more tailored articles for the most popular subjects.

You have decided to use Amazon Comprehend as your machine learning platform for this task. Which of the listed Comprehend APIs would give you the information your editors have requested? (Select THREE)

1. CreateDocumentClassifier
2. DetectSentiment
3. DetectSyntax
4. DetectEntities
5. DetectKeyPhrases
6. DetectDominantLanguage

**Answers:** B, D, E

**Explanation:**

Option A is incorrect. The CreateDocumentClassifier Comprehend API creates a document classifier that you use to categorize documents. Your editors want you to find the underlying intent of the comments.

Option B is correct. The DetectSentiment Comprehend API gives you the underlying sentiment (positive, neutral, mixed, or negative) of a string, such as a comment.

Option C is incorrect. The DetectSyntax Comprehend API gives you the part of speech of each word in a string. This would not help you understand the underlying intent of a comment.

Option D is correct. The DetectEntities Comprehend API finds named entities in text. This would help you find entities such as a news organization, politicians, celebrities, companies, etc. This information will help you identify the subject matter of the comments.

Option E is correct. The DetectKeyPhrases Comprehend API finds key noun phrases in text. This will also help you identify the subject matter of a comment.

Option F is incorrect. The DetectDominantLanguage Comprehend API finds the language (English, French, Spanish, etc.) used most frequently in the comments. This would not offer you much insight into the intent of a comment.

**Reference:**

Please see the Amazon Comprehend developer guide titled [Amazon Comprehend](https://docs.aws.amazon.com/comprehend/latest/dg/API_Operations_Amazon_Comprehend.html)

Question: 74

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Select the appropriate model(s) for a given machine learning problem**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a marketing consulting firm. Your firm has an online retailer as a client that wants to apply different marketing strategies per segment of their customer base. They have decided that the best way to segment their customers is by their purchase history. You have all of the online retailer purchase history from the last 5 years that you can use for your machine learning model.

Which type of machine learning algorithm would give you segmentation based on purchase history in the most expeditious manner?

1. K-Nearest Neighbors (KNN)
2. K-Means
3. Semantic Segmentation
4. Neural Topic Model (NTM)

**Answer:** B

**Explanation:**

Option A is incorrect. The k-nearest neighbor algorithm is used to find items that are similar to each other. This may find purchases that are similar to each other, but not customers that have similar purchase history. You would have to do additional modeling to use this algorithm.

Option B is correct. The K-Means algorithm is used to find groups within data where the members of the group are similar to each other but different from members of other groups. This is exactly what you are trying to solve: find groups of customers with similar purchase history.

Option C is incorrect. The semantic segmentation algorithm is used to develop computer vision applications. You are trying to solve a clustering problem, so this algorithm would not work for this problem.

Option D is incorrect. The Neural Topic Model algorithm is used to group documents into topics using the statistical distribution of words within the documents. You are trying to solve a clustering problem, so this algorithm would not work for this problem.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Using Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html), and the article titled [The 5 Clustering Algorithms Data Scientists Need to Know](https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68)

Question: 75

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Perform feature engineering**

**Domain:** Exploratory Data Analysis

**Question text**:

You work for the security department of your firm. As part of securing your firm’s email activity from phishing attacks you need to build a machine learning model that analyzes incoming email text to find word phrases like “you’re a winner” or “click here now” to find potential phishing emails.

Which of the following text feature engineering techniques is the best solution for this task?

1. Orthogonal Sparse Bigram (OSB)
2. Term Frequency-Inverse Document Frequency (tf-idf)
3. Bag-of-Words
4. N-Gram

**Answer:** D

**Explanation:**

Option A is incorrect. The Orthogonal Sparse Bigram natural language processing algorithm creates groups of words and outputs the pairs of words that includes the first word. You are trying to classify an email as a phishing attack by having your model learn based on the presence of multi-word phrases in the email text, not pairs of words from the email text stream using the first word as the key.

Option B is incorrect. Term Frequency-Inverse Document Frequency determines how important a word is in a document by giving weights to words that are common and less common in the document. You are trying to classify an email as a phishing attack by having your model learn based on the presence of multi-word phrases in the email text, you are not trying to determine the importance of a word of phrase in the email text.

Option C is incorrect. The Bag-of-Words natural language processing algorithm creates tokens of the input document text and outputs a statistical depiction of the text. The statistical depiction, such as a histogram, shows the count of each word in the document. You are trying to classify an email as a phishing attack by having your model learn based on the presence of multi-word phrases in the email text, not individual words.

Option D is correct. The N-Gram natural language processing algorithm is used to find multi-word phrases in text, in this case an email. This suits your phishing detection task since you are trying to classify an email as a phishing attack by having your model learn based on the presence of multi-word phrases.

**Reference:**

Please see the article titled [Introduction to Natural Language Processing for Text](https://towardsdatascience.com/introduction-to-natural-language-processing-for-text-df845750fb63), and the article titled [Document Classification Part 2: Text Processing (N-Gram Model & TF-IDF Model)](https://medium.com/machine-learning-intuition/document-classification-part-2-text-processing-eaa26d16c719)

## 

Question: 76

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Perform feature engineering**

**Domain:** Identify and implement a data-transformation solution

**Question text**:

You work for a car manufacturer as a machine learning specialist. Your marketing team wants to use a marketing strategy to market to different consumer segments based on how the features of each of their cars resonate with their customer base.

The dataset with which you have to work contains many features about each car, such as color, size, number of doors, number of speakers, type of roof, type of auto-assist, etc. Through your exploratory modeling you have found many of these features are redundant, meaning they don’t offer anything further to your algorithm’s performance.

Your dataset contains a large number of observations and a large number of features. How would you solve this redundant feature problem in the most efficient and expeditious manner?

1. Keep all the features and use the XGBoost algorithm to account for redundant features
2. Use Sparse Feature Graph to remove the redundant features
3. Use Principal Component Analysis to reduce the number of features
4. Keep all the features and use the Random Cut Forest algorithm to account for redundant features

**Answer:** C

**Explanation:**

Option A is incorrect. The XGBoost algorithm is used to predict a target variable in a very fast and efficient manner. However, the XGBoost will not automatically adjust for redundant features. The redundant features will act as a performance drag since you have a large number of features and a large number of observations.

Option B is incorrect. Removing the redundant features outright creates the risk of information loss. A better solution is to find composites of features that are uncorrelated, which is the technique used by Principal Component Analysis.

Option C is correct. Principal Component Analysis is a machine learning algorithm that reduces dimensionality within your data without sacrificing information. It does this by finding composites of features that are uncorrelated

Option D is incorrect. The Random Cut Forest algorithm is used to find atypical data points in a dataset, therefore it will not help find redundant features. The redundant features will act as a performance drag since you have a large number of features and a large number of observations.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Using Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html), the Amazon SageMaker developer guide titled [Principal Component Analysis (PCA) Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/pca.html), and the article titled [Automatically Redundant Features Removal for Unsupervised Feature Selection via Sparse Feature Graph](https://arxiv.org/abs/1705.04804)

## 

Question: 77

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ ML Implementation and Operations**

**Domain:** Deploy and operationalize machine learning solutions

**Question text**:

You work for an auto parts manufacturer as a machine learning specialist. You need to build a machine learning model that categorizes proprietary auto parts as they traverse your plant’s production lines. You do not have any existing trained models from which to start your work. You plan to use an image classification algorithm such as ResNet to classify the auto parts with one or more labels. The classified image data will then be used by your accounting department to dynamically keep the company’s parts database updated with the newly produced units.

Since you are building a model to classify images of proprietary auto parts, which technique can you use within SageMaker to expedite the deployment and operation of your model?

1. Online learning
2. Incremental learning
3. Transfer learning
4. Out-of-core learning

**Answer:** C

**Explanation:**

Option A is incorrect. Online learning refers to the process of training your model incrementally by giving it data observations as individual observations or in mini-batches. This will train your model, but it won’t expedite the process.

Option B is incorrect. Incremental learning would help expedite the training process if you are starting with an existing model and extending it with new data, specifically your proprietary auto parts images. However, you don’t have any existing trained models from which to start your work.

Option C is correct. When you use transfer learning you start with an existing trained model, usually ‘off the shelf’ from a source such as [ONNX Model Zoo](https://github.com/onnx/models#image_classification). You take the existing trained model and apply it to your different but closely aligned observations. This saves you time in deploying and operationalizing your machine learning solution since you are starting from a pretrained model.

Option D is incorrect. Out-of-code learning is used to train huge datasets that you can’t load into your server’s memory. This algorithm loads some of the data, trains on that subset, loads another subset of observations, trains on that subset, and repeats this process until it has completed the training of all the observations. This process will not help you deploy and operationalize your model more expeditiously.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Using Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html), the Amazon SageMaker machine learning blog titled [Now easily perform incremental learning on Amazon SageMaker](https://aws.amazon.com/blogs/machine-learning/now-easily-perform-incremental-learning-on-amazon-sagemaker/), and the article titled [Transfer learning with MXNet Gluon](https://hackernoon.com/transfer-learning-with-mxnet-gluon-8203005afafe)

Question: 78

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Perform hyperparameter optimization**

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist on a team tasked with designing an image recognition system that can adapt to new observations very quickly. You team is designing automated driving software for cars in a ride-share fleet. Your company wants to implement a service where when users hail a ride through your app on their mobile device, a nearby self-driving car arrives at the user’s location. It has the desired route preloaded and is ready to take the user to their destination. Your team has decided to use the SageMaker Image Classification algorithm in your image recognition model.

The machine learning models powering this self-driving car fleet need to react very quickly to new observations, such as previously not encountered obstacles like different types and sized animals, etc. Which hyperparameter would you set, and to what value, to obtain the desired outcome?

1. early\_stopping set to True
2. early\_stopping set to False
3. learning\_rate set to 0.1
4. learning\_rate set to 0.8
5. use\_pretrained\_model set to 0
6. use\_pretrained\_model set to 1

**Answer:** D

**Explanation:**

Option A is incorrect. The early\_stopping hyperparameter is used to decide whether to use early stopping during training. This hyperparameter allows you to terminate a training job early if it is observed that further training will not be necessary. Tuning this hyperparameter would not help your model react very quickly to new observations.

Option B is incorrect. The early\_stopping hyperparameter is used to decide whether to use early stopping during training. This hyperparameter allows you to terminate a training job early if it is observed that further training will not be necessary. Tuning this hyperparameter would not help your model react very quickly to new observations.

Option C is incorrect. The learning\_rate hyperparameter governs how quickly the model adapts to new or changing data. Valid values range from 0.0 to 1.0. Setting this hyperparameter to a low value, such as 0.1, will make the model learn more slowly. This is not what you want, you want your model to learn very rapidly.

Option D is correct. The learning\_rate hyperparameter governs how quickly the model adapts to new or changing data. Valid values range from 0.0 to 1.0. Setting this hyperparameter to a high value, such as 0.8, will make the model learn quickly. This is what you want, you want your model to learn very rapidly.

Option E is incorrect. The use\_pretrained\_model hyperparameter defines whether you want a pre-trained model to be loaded before training. This will not help you adapt quickly to new or changing observations.

Option F is incorrect. The use\_pretrained\_model hyperparameter defines whether you want a pre-trained model to be loaded before training. This will not help you adapt quickly to new or changing observations.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Image Classification Hyperparameters](https://docs.aws.amazon.com/sagemaker/latest/dg/IC-Hyperparameter.html), and the Amazon Machine Learning blog titled [Amazon SageMaker Automatic Model Tuning now supports early stopping of training jobs](https://aws.amazon.com/blogs/machine-learning/amazon-sagemaker-automatic-model-tuning-now-supports-early-stopping-of-training-jobs/)

Question: 79

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Deploy and operationalize machine learning solutions**

**Domain:** Machine Learning Implementation and Operations

**Question text**:

You work as a machine learning specialist for a gaming software company. You have trained and tested a machine learning model to predict gaming users likelihood of buying in-app purchases based on their player characteristics, such as playing time, levels achieved, etc. You are now ready to deploy your trained model onto the Amazon SageMaker Hosting service.  
  
What are the three steps for deploying a model using Amazon SageMaker Hosting Services? (Select THREE)

1. Create a model in Amazon SageMaker including the S3 path where the model artifacts are stored and the Docker registry path for the inference image
2. Create a model in Amazon SageMaker including the S3 path where the model artifacts are stored and the Kubernetes registry path for the inference image
3. Create an endpoint configuration for a REST endpoint
4. Create an endpoint configuration for an HTTPS endpoint
5. Create an HTTPS endpoint
6. Create a REST endpoint

**Answers:** A, D, E

**Explanation:**

Option A is correct. From the Amazon SageMaker developer guide titled [Deploy a Model on Amazon SageMaker Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-hosting.html) “By creating a model, you tell Amazon SageMaker where it can find the model components. This includes the S3 path where the model artifacts are stored and the Docker registry path for the image that contains the inference code.”

Option B is incorrect. The Amazon SageMaker Hosting Service expects to find the inference code in a Docker container, not in Kubernetes. (See the Amazon SageMaker developer guide titled [Deploy a Model on Amazon SageMaker Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-hosting.html))

Option C is incorrect. The Amazon SageMaker Hosting Service uses an HTTPS endpoint (not a REST endpoint) to provide inferences from the model. (See the Amazon SageMaker developer guide titled [Deploy a Model on Amazon SageMaker Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-hosting.html))

Option D is correct. The Amazon SageMaker Hosting Service uses an HTTPS endpoint to provide inferences from the model. This endpoint is configured to provide models to launch and instances on which to run them. (See the Amazon SageMaker developer guide titled [Deploy a Model on Amazon SageMaker Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-hosting.html))

Option E is correct. The Amazon SageMaker Hosting Service uses an HTTPS endpoint to provide inferences from the model. Client applications send requests to the SageMaker runtime HTTPS endpoint to get inferences, in your case to get inferences on the probability that a gamer will buy in-app purchases.

Option F is incorrect. The Amazon SageMaker Hosting Service uses an HTTPS endpoint (not a REST endpoint) to provide inferences from the model.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Deploy a Model on Amazon SageMaker Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-hosting.html) for an overview of the deployment of a SageMaker model.

## 

Question: 80

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Create data repositories for machine learning**

**Domain:** Data Engineering

**Question text**:

You are building a data repository for your company’s social media website that allows users to upload photos and videos to their personal stream. These photos and videos need to be labeled and classified so your company can use them to build direct marketing capabilities into your application based on machine learning. The direct marketing capability will be used to send targeted advertisements to users who have uploaded videos or photos of content that relates to a given products.

You are using Amazon SageMaker Ground Truth to label you user’s photos and videos. Sometimes your Ground Truth human workers mislabel images and/or videos. Which SageMaker Ground Truth feature helps you continue to get high quality labeling in an automated way even when your workers occasionally mislabel?

1. Chaining labeling jobs
2. Label verification and adjustment
3. Batches for labeling tasks
4. Annotation consolidation

**Answer:** D

**Explanation:**

Option A is incorrect. Ground Truth chaining labeling jobs allows you to reuse datasets from previous labeling jobs. This feature would not help you address mislabeled images or videos.

Option B is incorrect. The Ground Truth label verification and adjustment feature allows you to have workers verify and correct labels that were mislabeled. This would help you correct mislabeled items, but it is not an automated process, it is manual.

Option C is incorrect. The Ground Truth batches for labeling tasks feature is used to send objects to your workers in batches. This would not help you correct mislabeled objects.

Option D is correct. The Ground Truth annotation consolidation feature allows you to combine the annotations of multiple workers to produce an automated probabilistic estimate of what the correct label should be.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Data Labeling](https://docs.aws.amazon.com/sagemaker/latest/dg/sms-data-labeling.html), and the Amazon Machine Learning blog titled [Use the wisdom of crowds with Amazon SageMaker Ground Truth to annotate data more accurately](https://aws.amazon.com/blogs/machine-learning/use-the-wisdom-of-crowds-with-amazon-sagemaker-ground-truth-to-annotate-data-more-accurately/)

Question: 81

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Create data repositories for machine learning**

**Domain:** Data Engineering

**Question text**:

You work as a machine learning specialist for a media sharing service. The media sharing service will be used by healthcare professionals to share images of x-rays, MRIs, and other medical imagery. The accuracy of labeling these images is of primary importance, since the labeling will be used in autodiagnostic software. As your team builds the data repository to be used by your machine learning algorithms, you need to use human manual labelers. You have decided to use Amazon Ground Truth for this purpose. Since accuracy is of prime importance, you have decided to use the annotation consolidation feature of Ground Truth to ensure proper labeling of the medical images.

Which of the Ground Truth annotation consolidation functions should you use for ensuring the accuracy of your labeling tasks? (Select TWO)

1. Bounding box
2. Semantic segmentation
3. Named entity
4. Output manifest
5. Mechanical turk

**Answers:** A, B

**Explanation:**

Option A is correct. The bounding box finds the most similar bounding boxes from workers and averages them, thus using the power of multiple workers to annotate your images more accurately.

Option B is correct. The semantic segmentation feature fuses the pixel annotations of multiple workers and applying a smoothing function to the image, thus using the power of multiple workers to annotate your images more accurately.

Option C is incorrect. The named entity feature is used with text annotation work, not image annotation.

Option D is incorrect. The Ground Truth output manifest allows the output of a labeling job to be used as the input to a machine learning model. This feature will not help ensure accuracy of worker annotations.

Option E is incorrect. The Ground Truth Mechanical Turk feature gives you access to a large pool of labeling workers. While increasing the number of workers at your disposal, this feature will not help ensure accuracy of worker annotations.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Annotation Consolidation](https://docs.aws.amazon.com/sagemaker/latest/dg/sms-annotation-consolidation.html), and the Amazon Machine Learning blog titled [Use the wisdom of crowds with Amazon SageMaker Ground Truth to annotate data more accurately](https://aws.amazon.com/blogs/machine-learning/use-the-wisdom-of-crowds-with-amazon-sagemaker-ground-truth-to-annotate-data-more-accurately/), and GitHub repository titled [Amazon Sagemaker Examples Introduction to Ground Truth Labeling Jobs](https://github.com/awslabs/amazon-sagemaker-examples/tree/master/ground_truth_labeling_jobs)

Question: 82

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Recommend and implement the appropriate machine learning services and features for a given problem**

**Domain:** ML Implementation and Operations

**Question text**:

You work as a machine learning specialist for a large software company that has several huge data centers around the world. Your company has realized they could do a better job managing their data center power usage effectiveness (PUE) by implementing a machine learning system to automate the management of the many controls used to control their data center power usage. The machine learning model needs to take as inputs data from building management systems such as chillers, pumps, cooling units, the actual load from systems usage, etc. You have trained your model based on historical data of these inputs and the desired outcomes in these historical observations. Now you want to run your model to process real-time inferences while also continuing to learn from the new inferences.

Which combination of SageMaker algorithms and learning techniques should you use for your model to predict settings that optimize PUE on an ongoing basis?

1. Supervised learning using a Convolutional Neural Network algorithm
2. Unsupervised learning using a Multilayer Perceptron algorithm
3. Reinforcement learning using a Convolutional Neural Network algorithm
4. Unsupervised learning using a Sequence-to-Sequence Neural Network algorithm
5. Supervised learning using a Feedforward Neural Network algorithm

**Answer:** C

**Explanation:**

Option A is incorrect. In order to benefit from the trained model and then perform inferences while continuing to learn from the inferences, you cannot use supervised learning, you need to use reinforcement learning.

Option B is incorrect. The Multilayer Perceptron algorithm is used for speech recognition and translation.

Option C is correct. Reinforcement learning is used to continually update your model as new inference observations are encountered. Also, the Convolutional Neural Network algorithm is typically used in scenarios like this. (See the article titled [A Practical Guide to Artificial Intelligence for the Data Center](https://blog.equinix.com/blog/2019/01/10/a-practical-guide-to-artificial-intelligence-for-the-data-center-2/))

Option D is incorrect. The Sequence-to-Sequence Neural Network algorithm is used for machine translation and question answering systems.

Option E is incorrect. The Feedforward Neural Network algorithm is a simple neural network not capable of handling a complex problem like data center power usage effectiveness management. (See the article titled [A Practical Guide to Artificial Intelligence for the Data Center](https://blog.equinix.com/blog/2019/01/10/a-practical-guide-to-artificial-intelligence-for-the-data-center-2/))

**Reference:**

Please see the article titled [A Practical Guide to Artificial Intelligence for the Data Center](https://blog.equinix.com/blog/2019/01/10/a-practical-guide-to-artificial-intelligence-for-the-data-center-2/), the article titled [Demystifying reinforcement learning and convolutional neural network](https://www.ericsson.com/en/blog/2018/9/demystifying-reinforcement-learning-and-convolutional-neural-network), the wikipedia article titled [Reinforcement learning](https://en.wikipedia.org/wiki/Reinforcement_learning), the wikipedia article titled [Convolutional neural network](https://en.wikipedia.org/wiki/Convolutional_neural_network), and the article titled [A Comprehensive Guide to Types of Neural Networks](https://www.digitalvidya.com/blog/types-of-neural-networks/)

Question: 83

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​ Recommend and implement the appropriate machine learning services and features for a given problem**

**Domain:** ML Implementation and Operations

**Question text**:

You work as a machine learning specialist for a home maintenance automation company that produces robots to vacuum the floor, mow the lawn, and other automated worker tools. You have built and trained your model (starting from a pre-trained model from [ImageNet](http://www.image-net.org/)) using the SageMaker built-in Object Detection algorithm. The Object Detection algorithm is used by the robots to detect objects that are obstacles or boundaries in their work area. You now need to have the robots run in real home settings using your model. You also want your robots to be able to communicate with each other if there is more than one robot in the operating area.

Which set of Amazon services will give you the most cost effective solution?

1. Amazon Elastic Inference and AWS IoT Greengrass
2. AWS RoboMaker and Amazon Sumerian
3. Amazon Rekognition and AWS IoT Greengrass
4. Amazon Rekognition and Amazon Sumerian

**Answer:** A

**Explanation:**

Option A is correct. Amazon Elastic Inference allows you to reduce the cost of your inference learning by up to 75% while giving you the inference processing (CPU, GPU, etc.) you need to process your obstacle and boundary observations. AWS IoT Greengrass gives you the capability to run inference on your robot devices and communicate with other IoT devices.

Option B is incorrect. Amazon Sumerian is used for augmented reality, which is not needed to solve your machine learning scenarios.

Option C is incorrect. Amazon Rekognition is used for image and video analysis. It would identify objects in your domain, but it wouldn’t contribute to lowering the cost of your inference implementation.

Option D is incorrect. Amazon Rekognition is used for image and video analysis. It would identify objects in your domain, but it wouldn’t contribute to lowering the cost of your inference implementation. Also, Amazon Sumerian is used for augmented reality, which is not needed to solve your machine learning scenarios.

**Reference:**

Please see the [Amazon SageMaker Overview](https://aws.amazon.com/sagemaker/), particularly the Deploy and manage models in production section, the [Amazon Elastic Inference Overview](https://aws.amazon.com/machine-learning/elastic-inference/), the AWS News blog titled [Amazon Elastic Inference – GPU-Powered Deep Learning Inference Acceleration](https://aws.amazon.com/blogs/aws/amazon-elastic-inference-gpu-powered-deep-learning-inference-acceleration/), the Amazon SageMaker developer guide titled [Object Detection Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/object-detection.html), the [AWS IoT Greengrass Overview](https://aws.amazon.com/greengrass/), the [Amazon Sumerian Overview](https://aws.amazon.com/sumerian/), and the [Amazon Rekognition Overview](https://aws.amazon.com/rekognition/)

Question: 84

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Analyze and visualize data for machine learning

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a city government in their shared bike program as a machine learning specialist. You need to visualize the bike share location predictions you are producing on an hourly basis using your model inference you created using the SageMaker built-in K-Means algorithm. Your inference endpoint takes IoT data from your shared bikes as they are used throughout the city. You also want to enrich your shared bike data with external data sources such as current weather and road conditions.

Which set of Amazon services would you use to create your visualization with the least amount of effort?

1. IoT Core -> IoT Analytics ->SageMaker -> QuickSight
2. IoT Core -> Kinesis Firehose -> SageMaker -> QuickSight
3. IoT Core -> Lambda -> SageMaker -> QuickSight
4. IoT Core -> IoT Greengrass -> QuickSight

**Answer:** A

**Explanation:**

Option A is correct. IoT Core collects data from each shared bike, IoT Analytics retrieves messages from the shared bikes as they stream data data, IoT Analytics also enriches the streaming data with your external data sources and sends the streaming data to your K-Means machine learning inference endpoint, QuickSight is then used to create your visualization. This approach requires the least amount of effort mainly because of the data enrichment feature of IoT Analytics.

Option B is incorrect. With this option you would have to create a lambda function to gather the data enrichment information (weather, road conditions) and enrich the data streams in your own code.

Option C is incorrect. Also, with this option you would have to add code to your lambda function to gather the data enrichment information (weather, road conditions) and enrich the data streams in your own code.

Option D is incorrect. IoT Greengrass is a service that you use to run local machine learning inference capabilities on connected devices. This approach would not easily integrate with your QuickSight visualization.

**Reference:**

Please see the [AWS IoT Analytics overview](https://aws.amazon.com/iot-analytics/), the Amazon SageMaker developer guide titled [K-Means Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/k-means.html), the AWS Big Data blog titled [Build a Visualization and Monitoring Dashboard for IoT Data with Amazon Kinesis Analytics and Amazon QuickSight](https://aws.amazon.com/blogs/big-data/build-a-visualization-and-monitoring-dashboard-for-iot-data-with-amazon-kinesis-analytics-and-amazon-quicksight/), the AWS IoT Analytics User Guide titled [What IS AWS IoT Analytics?](https://docs.aws.amazon.com/iotanalytics/latest/userguide/welcome.html), and the [AWS IoT Greengrass FAQs](https://aws.amazon.com/greengrass/faqs/)

Question: 85

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Identify and implement a data-transformation solution

**Domain:** Data Engineering

**Question text**:

You work for a logistics company that specializes in the storage, movement, and control of massive amounts of packages. You are on the machine learning team assigned the task of building a machine learning model to assist in the control of your company’s package logistics. Specifically, your model needs to predict the routes your package movers should take for optimal delivery and resource usage. The model requires various transformations to be performed on the data. You also want to get inferences on entire datasets once you have your model in production. Additionally, you won’t need a persistent endpoint for applications to call to get inferences.

Which type of production deployment would you use to get predictions from your model in the most expeditious manner?

1. SageMaker Hosting Services
2. SageMaker Batch Transform
3. SageMaker Containers
4. SageMaker Elastic Inference

**Answer:** B

**Explanation:**

Option A is incorrect. SageMaker Hosting Services is used for applications to send requests to an HTTPS endpoint to get inferences. This type of deployment is used when you need a persistent endpoint for applications to call to get inferences.

Option B is correct. SageMaker Batch Transform is used to get inferences for an entire dataset and you don’t need a persistent endpoint for applications to call to get inferences.

Option C is incorrect. SageMaker Containers is a service you can use to create your own Docker containers to deploy your models. This would not be the most expeditious option.

Option D is incorrect. SageMaker Elastic Interface is used to accelerate deep learning inference workloads. This service alone would not give you the batch transform capabilities you need.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Deploy a Model on Amazon SageMaker Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-hosting.html), the Amazon SageMaker developer guide titled [Get Inferences for an Entire Dataset with Batch Transform](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-batch.html), the Amazon Elastic Inference developer guide titled [What Is Amazon Elastic Inference?](https://docs.aws.amazon.com/elastic-inference/latest/developerguide/what-is-ei.html), and the Amazon SageMaker developer guide titled [Amazon SageMaker Containers: a Library to Create Docker Containers](https://docs.aws.amazon.com/sagemaker/latest/dg/amazon-sagemaker-containers.html)

Question: 86

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Perform feature engineering

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a flight diagnostics company that builds instrumentation for airline manufacturers. Your company’s instrumentation hardware and software is used to detect flight pattern information such as flight path deviation, as well as airline component malfunction. Your team of machine learning specialists has created a model using the Random Cut Forest algorithm to be used to identify anomalies in the data. The streaming data that your instrumentation processes needs to be cleaned and transformed via feature engineering before passing it to your inference endpoint. You have created the pre-processing and post-processing steps (for cleaning and feature engineering) in your training process.

How can you implement the cleaning and feature engineering steps in your inference processing in the most efficient manner?

1. Execute the pre-processing in a client application before sending the data to your inference endpoint
2. Bundle and export the training pre-processing steps and deploy them to your inference container
3. Bundle and export the training pre-processing steps and deploy them as part of your Inference Pipeline
4. Bundle and export the training pre-processing steps and deploy them to IoT Core on the data emitting devices.

**Answer:** C

**Explanation:**

Option A is incorrect. Although you could execute your pre-processing steps in a client application before sending the data on to your inference end-point, this would require additional work on your part to build that client application and then incorporate your feature engineering scripts from your training process into it.

Option B is incorrect. You could also include your pre-processing steps in your inference container, however this requires more work on your part than using the SageMaker Inference Pipelines feature.

Option C is correct. SageMaker Inference Pipelines allows you to bundle and export your pre and post-processing steps from your training process and deploy them as part of your Inference Pipeline. Inference Pipelines are fully managed by AWS.

Option D is incorrect. Amazon IoT Core is used to facilitate device intercommunication. It is not a service you would use for pre-processing data streams for machine learning inference endpoints.

**Reference:**

Please see the Amazon announcement titled [Announcing Enhancements for Data Processing and Feature Engineering, and Improved Framework Support with Amazon SageMaker](https://aws.amazon.com/about-aws/whats-new/2018/11/amazon-sagemaker-announces-several-enhancements-for-developer-pr/), the Amazon SageMaker developer guide titled [Deploy an Inference Pipeline](https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipelines.html), the AWS Machine Learning blog titled [Use the built-in Amazon SageMaker Random Cut Forest algorithm for anomaly detection](https://aws.amazon.com/blogs/machine-learning/use-the-built-in-amazon-sagemaker-random-cut-forest-algorithm-for-anomaly-detection/), and the [AWS IoT Core Overview page](https://aws.amazon.com/iot-core/)

Question: 87

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Frame business problems as machine learning

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for a farming corporation that wants to use in-ground soil sensors together with enrichment from geolocation, rainfall, and other weather information for the growing area to help identify crop growth stages. They want to use the crop growth information to increase yield and produce more product year over year. They also hope to increase the crop quality through this effort.

The machine learning models that you build for this solution will analyze various growing conditions, such as temperature and humidity so the farming corporation can schedule watering appropriately for the area.

What collection of AWS services would you use to implement a solution that first trains your model, then gathers the information from the in-ground sensors, then enriches the sensor data, and finally deploys the model to run inference on connected devices in the field?

1. SageMaker, IoT Core, IoT Analytics, IoT Greengrass
2. SageMaker, IoT Core, Kinesis Data Analytics, IoT Greengrass
3. SageMaker, IoT Code, Kinesis Data Streams, IoT Greengrass
4. SageMaker, IoT Core, IoT Analytics, Inference Pipeline

**Answer:** A

**Explanation:**

Option A is correct. SageMaker is used to create your model and train it initially. IoT Core sends the sensor data to IoT Analytics for enrichment and analysis. The pre-trained model is deployed into the field using IoT Greengrass so you can perform ML inference using the enriched data on the farm local devices in the field.

Option B is incorrect. You could use Kinesis Data Analytics to analyze your IoT device data streams, but IoT Analytics is built specifically for analyzing the highly unstructured IoT data, so it is a better choice.

Option C is incorrect. You could use Kinesis Data Streams to stream your IoT device data, but you would have to write lambda functions to perform the enrichment step. IoT Analytics is built specifically for analyzing and enriching the highly unstructured IoT data, so it is a better choice.

Option D is incorrect. Inference Pipeline is used to define and deploy pretrained SageMaker algorithms. Inference Pipeline does not have the IoT inference integration that IoY Greengrass has, so IoT Greengrass is a better choice for this problem.

**Reference:**

Please see the [AWS IoT Greengrass ML Inference overview](https://aws.amazon.com/greengrass/ml/), the [AWS IoT Analytics overview](https://aws.amazon.com/iot-analytics/), the [Amazon Kinesis Data Analytics overview](https://aws.amazon.com/kinesis/data-analytics/), and Amazon SageMaker developer guide titled [Deploy an Inference Pipeline](https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipelines.html)

Question: 88

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Perform hyperparameter optimization

**Domain:** Modeling

**Question text**:

You work for a transportation company as a machine learning specialist. You are currently working on a project to optimize container truck routes with the objective of minimizing empty container travel. For example, as a truck delivers its payload to a destination you want to have the container loaded for another route, you don’t want the truck to move to another destination with an empty container. You have selected the SageMaker XGBoost algorithm for your model. You now need to tune your hyperparameters to get the optimum performance out of your model. You have chosen the Area Under the Curve (AUC) metric as your objective metric for your hyperparameter tuning job.

Which algorithm should you use as the SageMaker hyperparameter tuning algorithm to get your results in the minimal number of training jobs?

1. Random search
2. Bayesian Search
3. Linear Search
4. Depth First Search

**Answer:** B

**Explanation:**

Option A is incorrect. SageMaker uses two types of models to search for the optimum hyperparameters for your model: Random Search and Bayesian Search. For most models, Bayesian Search requires less training jobs to reach your optimal hyperparameter settings. (See the Amazon Machine Learning blog titled [Amazon SageMaker automatic model tuning now supports random search and hyperparameter scaling](https://aws.amazon.com/blogs/machine-learning/amazon-sagemaker-automatic-model-tuning-now-supports-random-search-and-hyperparameter-scaling/))

Option B is correct. SageMaker uses two types of models to search for the optimum hyperparameters for your model: Random Search and Bayesian Search. For most models, Bayesian Search requires less training jobs to reach your optimal hyperparameter settings. (See the Amazon Machine Learning blog titled [Amazon SageMaker automatic model tuning now supports random search and hyperparameter scaling](https://aws.amazon.com/blogs/machine-learning/amazon-sagemaker-automatic-model-tuning-now-supports-random-search-and-hyperparameter-scaling/))

Option C is incorrect. SageMaker hyperparameter tuning does not use Linear Search as a hyperparameter tuning model.

Option D is incorrect. SageMaker hyperparameter tuning does not use Depth First Search as a hyperparameter tuning model.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Configure and Launch a Hyperparameter Tuning Job](https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-ex-tuning-job.html), the Amazon SageMaker developer guide titled [Automatic Model Tuning](https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning.html), and the Amazon SageMaker developer guide titled [How Hyperparameter Tuning Works](https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-how-it-works.html)

Question: 89

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Perform hyperparameter optimization

**Domain:** Modeling

**Question text**:

You work for a software company that produces an online sports betting app. You are on the machine learning team responsible for building a model that predicts the likelihood of registered users to wager on a given event based on several features of sports events offered in the app. You and your team have selected the Linear Learner algorithm and have trained your model. You now wish to find the best set of hyperparameters for your model. You have chosen to use SageMaker’s automatic model tuning and you have set your objective to validation:precision in your hyperparameter tuning job.

How do pass your tuning job settings into your hyperparameter tuning job? (Select THREE)

1. Define a JSON object and pass it as the value of the HyperParameterConfig to the [HyperParameterTuningJob](https://docs.aws.amazon.com/sagemaker/latest/dg/API_CreateHyperParameterTuningJob.html)
2. Define a JSON object and pass it as the value of the HyperParameterTuningJobConfig to the [CreateHyperParameterTuningJob](https://docs.aws.amazon.com/sagemaker/latest/dg/API_CreateHyperParameterTuningJob.html)
3. In the JSON object specify the ranges of the hyperparameters you want to tune
4. In the JSON object specify the limits of the hyperparameters you want to tune
5. In the JSON object specify the objective metric for the hyperparameter tuning job
6. In the JSON object specify the MaxSequentialTrainingJobs parameter in the ResourceLimits section

**Answers:** B, C, E

**Explanation:**

Option A is incorrect. The correct name of the value you use to pass your JSON object is HyperParameterTuningJobConfig and the name of the job is [CreateHyperParameterTuningJob](https://docs.aws.amazon.com/sagemaker/latest/dg/API_CreateHyperParameterTuningJob.html).

Option B is correct. To specify the hyperparameter settings for your hyperparameter tuning job you pass a JSON object as the HyperParameterTuningJobConfig parameter to the job named [CreateHyperParameterTuningJob](https://docs.aws.amazon.com/sagemaker/latest/dg/API_CreateHyperParameterTuningJob.html)

Option C is correct. You specify the ranges of the hyperparameters you want to tune in the ParameterRanges section of the HyperParameterTuningJobConfig.

Option D is incorrect. You specify the ranges of the hyperparameters you want to tune in the ParameterRanges section of the HyperParameterTuningJobConfig, not the limits of the hyperparameters.

Option E is correct. In the HyperParameterTuningJobObjective section of the HyperParameterTuningJobConfig you set MetricName to the objective metric for the hyperparameter tuning job.

Option F is incorrect. There is no MaxSequentialTrainingJobs parameter in the ResourceLimits section of the HyperParameterTuningJobConfig.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Automatic Model Tuning](https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning.html), and the Amazon SageMaker developer guide titled [Configure and Launch a Hyperparameter Tuning Job](https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-ex-tuning-job.html)

Question: 90

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Analyze and visualize data for machine learning

**Domain:** Exploratory Data Analysis

**Question text**:

You are working on a Linear Learner algorithm based model used to predict the quarterly sales for each region of your company’s global sales force. The model needs to use data from your sales team’s past sales performance, such as quantity of products sold, revenue generated, expenses incurred, sales force size, etc.

You and your team are in the process of training the model based on the SageMaker built-in Linear Learner algorithm. You want to track and monitor metrics, such as test objective loss and test precision as the model trains. Which AWS service(s) would you use to track and monitor these metrics? (Select THREE)

1. Specify the metrics you want to track using the AWS Management Dashboard for SageMaker
2. Specify the metrics you want to track using the AWS Management Console for SageMaker
3. Specify the metrics you want to track using the SageMaker Javascript SDK APIs
4. Specify the metrics you want to track using the SageMaker Python SDK APIs
5. Use the CloudWatch console for visualizing time-series curves of your metrics
6. Use the SageMaker Javascript SDK APIs to visualize your metrics programmatically

**Answers:** B, D, E,

**Explanation:**

Option A is incorrect. You can specify the metrics you want to track using the AWS Management Console for SageMaker, not the AWS Management Dashboard for SageMaker.

Option B is correct. To specify the metrics you want to track you use the AWS Management Console for SageMaker or the SageMaker Python SDK APIs.

Option C is incorrect. To specify the metrics you want to track you use the AWS Management Console for SageMaker or the SageMaker Python SDK APIs, not the SageMaker Javascript SDK APIs.

Option D is correct. To specify the metrics you want to track you use the AWS Management Console for SageMaker or the SageMaker Python SDK APIs.

Option E is correct. Once the model training starts, SageMaker streams the metrics you specified to CloudWatch where you can visualize time-series curves of your metrics.

Option F is incorrect. You can visualize your metrics either via the CloudWatch console, or the SageMaker Python SDK APIs, not the SageMaker Javascript SDK APIs.

**Reference:**

Please see the AWS Machine Learning Blog titled [Easily monitor and visualize metrics while training models on Amazon SageMaker](https://aws.amazon.com/blogs/machine-learning/easily-monitor-and-visualize-metrics-while-training-models-on-amazon-sagemaker/)

Question: 91

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Analyze and visualize data for machine learning

**Domain:** Exploratory Data Analysis

**Question text**:

You work for an oil refinery company where you are on one of their machine learning teams. Your team is responsible for building models that help the company decide where to place their exploratory drilling teams across the globe. Your team lead has decided to build your model based on the K-Means built-in SageMaker algorithm. The team lead has tasked you with providing metric visualization charts for the training runs of your team’s model.

How would you go about visualizing the training metrics? (Select TWO)

1. In your SageMaker jupyter notebook, using the SageMaker python module called pandas.analytics, import TrainingAnalytics .
2. In your SageMaker jupyter notebook, using the SageMaker python module called sagemaker.analytics, import TrainingAnalytics.
3. In your SageMaker jupyter notebook, using the SageMaker python module called sagemaker.analytics, import TrainingJobAnalytics.
4. In your SageMaker jupyter notebook, using the SageMaker python module called pandas.analytics, import TrainingJobAnalytics.
5. Set one of the metric names to test:cross\_entropy’
6. Set one of the metric names to test:msd’

**Answers:** C, F

**Explanation:**

Option A is incorrect. You use the SageMaker python module called sagemaker.analytics (not pandas.analytics) from which you import TrainingJobAnalytics (not TrainingAnalytics) to gain access to the python methods that allow you to visualize you metrics in charts.

Option B is incorrect. You use the SageMaker python module called sagemaker.analytics from which you import TrainingJobAnalytics (not TrainingAnalytics) to gain access to the python methods that allow you to visualize you metrics in charts.

Option C is correct. You use the SageMaker python module called sagemaker.analytics from which you import TrainingJobAnalytics to gain access to the python methods that allow you to visualize you metrics in charts.

Option D is incorrect. You use the SageMaker python module called sagemaker.analytics (not pandas.analytics) from which you import TrainingJobAnalytics to gain access to the python methods that allow you to visualize you metrics in charts.

Option E is incorrect. To set the metric name that you wish to visualize you need to give a valid metric for the algorithm you are training. The test:cross\_entropy metric is not valid for a K-Means training run.

Option F is correct. To set the metric name that you wish to visualize you need to give a valid metric for the algorithm you are training. The test:msd metric is one of the two valid for a K-Means training run. The other valid metric for K-Means is test:ssd.

**Reference:**

Please see the AWS Machine Learning Blog titled [Easily monitor and visualize metrics while training models on Amazon SageMaker](https://aws.amazon.com/blogs/machine-learning/easily-monitor-and-visualize-metrics-while-training-models-on-amazon-sagemaker/), and the Amazon SageMaker developer guide titled [Tune a K-Means model](https://docs.aws.amazon.com/sagemaker/latest/dg/k-means-tuning.html)

Question: 92

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Analyze and visualize data for machine learning

**Domain:** Exploratory Data Analysis

**Question text**:

You work for an online retailer as a machine learning specialist. Your team is building a deep learning model based on the Keras Sequential model to categorize the clothing your company’s users post on their instagram feeds when they use one of the hashtags that refers to your company. You are the machine learning specialist assigned to building the training run visualization code to allow the team to monitor training metrics of the model as it trains.

How would you go about visualizing the training metrics? (Select TWO)

1. When creating your model training job in the SageMaker console, specify a regex pattern for the metrics that you want your model training script to write to your logs
2. When creating your model training job in the SageMaker console, specify the metrics that you want your model training script to write to your logs
3. Use the CloudWatch metrics dashboard to visualize the metrics that SageMaker automatically parsed from your logs and published for graphing and visualization.
4. Use the SageMaker metrics dashboard to visualize the metrics that SageMaker automatically parsed from your logs and published for graphing and visualization.
5. Write a python script in your SageMaker jupyter notebook to visualize the metrics that SageMaker automatically parsed from your logs and published for graphing and visualization.

**Answers:** A, C

**Explanation:**

Option A is correct. While creating your model training job in the SageMaker console, you specify a regex pattern that is used for the metrics that your model training script writes to your logs.

Option B is incorrect. While creating your model training job in the SageMaker console, you specify a regex pattern that is used for the metrics that your model training script writes to your logs. You can’t specify the metrics directly, you must use a regex pattern.

Option C is correct. SageMaker parses from your logs the metrics which you wish to track and publishes them to CloudWatch. The CloudWatch metrics dashboard allows you to visualize your SageMaker training job metrics as graphs for visualization.

Option D is incorrect. The CloudWatch metrics dashboard allows you to visualize your SageMaker training job metrics as graphs for visualization, not the SageMaker metrics dashboard.

Option E is incorrect. You would not need to write a python script to visualize your metrics data since the CloudWatch metrics dashboard gives you this functionality.

**Reference:**

Please see the AWS Machine Learning Blog titled [Easily monitor and visualize metrics while training models on Amazon SageMaker](https://aws.amazon.com/blogs/machine-learning/easily-monitor-and-visualize-metrics-while-training-models-on-amazon-sagemaker/), the Amazon SageMaker developer guide titled [Use TensorFlow with Amazon SageMaker](https://docs.aws.amazon.com/sagemaker/latest/dg/tf.html), and the Tensorflow.org page titled [Basic classification: Classify images of clothing](https://www.tensorflow.org/tutorials/keras/classification)

Question: 93

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Identify and implement a data-ingestion solution

**Domain:** Data Engineering

**Question text**:

You work for a manufacturer of wifi connected radios. You company wants to use data captured when these radios are in use by their customers (such as how the hardware is performing, the applications that are running on the radio, and the content that’s being streamed) to better serve their customers. You and your team of machine learning specialists have been asked to use the data captured when users play their radios to build a model that detects anomalies with the hardware performance.

What AWS service and function within that service will allow you to identify anomalies in the data stream?

1. Kinesis Data Analytics and its Hotspots function
2. Kinesis Data Analytics and its Random Cut Forest function
3. Kinesis Data Firehose and its Hotspots function
4. Kinesis Data Streams and its Random Cut Forest function
5. Kinesis Data Streams and its Hotspots function
6. Kinesis Data Firehose and its Random Cut Forest function

**Answer:** B

**Explanation:**

Option A is incorrect. The Kinesis Data Analytics Hotspot function is used to get information about dense regions in your data, not to identify outlier data, or anomalies, in your streaming data.

Option B is correct. The Kinesis Data Analytics Random\_Cut\_Forest function is used to identify outlier data, or anomalies, in your streaming data.

Option C is incorrect. Kinesis Data Firehose does not have functions like Hotspots or Random\_Cut\_Forest.

Option D is incorrect. Kinesis Data Streams does not have functions like Hotspots or Random\_Cut\_Forest.

Option E is incorrect. Kinesis Data Streams does not have functions like Hotspots or Random\_Cut\_Forest.

Option F is incorrect. Kinesis Data Firehose does not have functions like Hotspots or Random\_Cut\_Forest.

**Reference:**

Please see the Amazon Kinesis Data Analytics for SQL Applications Developer Guide titled [Examples: Machine Learning](https://docs.aws.amazon.com/kinesisanalytics/latest/dev/examples-machine.html), the Amazon Kinesis Data Analytics for SQL Applications Developer Guide titled [Example: Detecting Data Anomalies on a Stream (RANDOM\_CUT\_FOREST Function)](https://docs.aws.amazon.com/kinesisanalytics/latest/dev/app-anomaly-detection.html), and the Amazon Kinesis Data Analytics for SQL Applications Developer Guide titled [Example: Detecting Hotspots on a Stream (HOTSPOTS Function)](https://docs.aws.amazon.com/kinesisanalytics/latest/dev/app-hotspots-detection.html)

Question: 94

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Perform feature engineering

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a city electric scooter rental company. Your company supplies a fleet of electric scooters to different cities around the country. These scooters need to be managed as far as their location, their rental miles, their need for maintenance, etc. The company accumulates hundreds of data points on each scooter every day. You are on the machine learning team of your company where you have been assigned the job of building a machine learning model to track each scooter and decide when they are ready for maintenance. One would assume the decision for maintenance would be based predominantly on miles accumulated. Since you have so many features captured for a given scooter, you have decided you need to find the most predictive features in your model in order to avoid low model performance due to collinearity.

You have built your model in SageMaker using the built-in XGBoost algorithm. Using the XGBoost python API package, which type of booster and which API call would you use if you wanted to select the most predictive features based on the total gain across all splits in which the feature is used?

1. booster = gblinear using the get\_fscore with importance\_type parameter set to total\_gain
2. booster = gblinear using the get\_score with importance\_type parameter set to gain
3. booster = gbtree using the get\_score with importance\_type parameter set to total\_gain
4. booster = gbtree using the get\_fscore with importance\_type parameter set to gain
5. booster = dart using the get\_fscore with importance\_type parameter set to gain
6. booster = dart using the get\_score with importance\_type parameter set to total\_gain

**Answer:** C

**Explanation:**

Option A is incorrect. To get the features based on the total gain across all splits in which the feature is used you need to use the gbtree booster and call get\_score passing the parameter importance\_type set to total\_gain. Feature importance is defined only for base learner, or tree boosters. Feature importance is not defined for linear learners. The importance\_type parameter is defined for the get\_score API call, not the get\_fscore API call.

Option B is incorrect. To get the features based on the total gain across all splits in which the feature is used you need to use the gbtree booster and call get\_score passing the parameter importance\_type set to total\_gain. Feature importance is defined only for base learner, or tree boosters. Feature importance is not defined for linear learners. The importance\_type parameter needs to be set to total\_gain to get the total gain across all splits in which the feature is used. The importance\_type parameter of gain gives you the average gain across all splits in which the feature is used.

Option C is correct. To get the features based on the total gain across all splits in which the feature is used you need to use the gbtree booster and call get\_score passing the parameter importance\_type set to total\_gain.

Option D is incorrect. To get the features based on the total gain across all splits in which the feature is used you need to use the gbtree booster and call get\_score passing the parameter importance\_type set to total\_gain. The importance\_type parameter needs to be set to total\_gain to get the total gain across all splits in which the feature is used. The importance\_type parameter of gain gives you the average gain across all splits in which the feature is used.

Option E is incorrect. To get the features based on the total gain across all splits in which the feature is used you need to use the gbtree booster and call get\_score passing the parameter importance\_type set to total\_gain. Feature importance is defined only for base learner, or tree boosters. Feature importance is not defined for dart boosters. The importance\_type parameter needs to be set to total\_gain to get the total gain across all splits in which the feature is used. The importance\_type parameter of gain gives you the average gain across all splits in which the feature is used.

Option F is incorrect. To get the features based on the total gain across all splits in which the feature is used you need to use the gbtree booster and call get\_score passing the parameter importance\_type set to total\_gain. Feature importance is defined only for base learner, or tree boosters. Feature importance is not defined for dart boosters.

**Reference:**

Please see the Amazon SageMaker developer guide titled [XGBoost Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html), the Amazon SageMaker developer guide titled [XGBoost Hyperparameters](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html), and the [XGBoost Python API Reference](https://xgboost.readthedocs.io/en/latest/python/python_api.html)

Question: 95

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Identify and implement a data-transformation solution

**Domain:** Data Engineering

**Question text**:

You work as a machine learning specialist for a retail chain that has recently purchased another retail chain and is in the process of merging the two chain’s systems. Both retail chains have customer databases. Some of the firm’s customers overlap, meaning that the same customer registered with both chains in the past. When merging the customer data stores of the two, now merged retail chains, you need to link duplicate customer data so that you can have one accurate customer data source.

You have been assigned the task of creating the new customer data source for the now merged retail chain. Instead of trying to find the duplicate customer data manually through traditional programming techniques, you have decided to use machine learning techniques to solve the problem.

You have determined that the AWS Glue Machine Learning FindMatches Transform is the best solution to this problem. Knowing that incorrectly linking what appear to be duplicate customers must be avoided at all costs, how should you configure the AWS Glue FindMatches ML Transform parameters to achieve the most efficient and accurate duplicate customer detection process?

1. Set the FindMatches precision-recall parameter to ‘precision’ and the accuracy-cost parameter to ‘accuracy’
2. Set the FindMatches precision-recall parameter to ‘precision’ and the accuracy-cost parameter to ‘lower cost’
3. Set the FindMatches precision-recall parameter to ‘recall’ and the accuracy-cost parameter to ‘accuracy’
4. Set the FindMatches precision-recall parameter to ‘recall’ and the accuracy-cost parameter to ‘lower cost’

**Answer:** A

**Explanation:**

Option A is correct. Setting the FindMatches precision-recall parameter to ‘precision’ minimizes false positives (when you don’t have a match of a duplicate customer but mark it as a match mistakenly). This is what you want. Setting the FindMatches accuracy-cost parameter to ‘accuracy’ maximizes the transform accuracy of finding matching records as duplicate. This is also what you want.

Option B is incorrect. Setting the FindMatches precision-recall parameter to ‘precision’ minimizes false positives (when you don’t have a match of a duplicate customer but mark it as a match mistakenly). This is what you want. But, setting the accuracy-cost parameter to ‘lower cost’ favors cost or the speed of running the transform at the expense of the transform’s accuracy. This may make your transform more performant, but your primary concern is avoiding linking customers incorrectly so you should set the accuracy-cost parameter to ‘accuracy’.

Option C is incorrect. Setting the FindMatches precision-recall parameter to ‘recall’ minimizes false negatives (when you have a match of a duplicate customer but fail to detect it). This may cause customer frustration, but your primary concern is avoiding linking customers incorrectly.

Option D is incorrect. Setting the FindMatches precision-recall parameter to ‘recall’ minimizes false negatives (when you have a match of a duplicate customer but fail to detect it). This may cause customer frustration, but your primary concern is avoiding linking customers incorrectly.

**Reference:**

Please see the AWS Glue developer guide titled [Machine Learning Transforms in AWS Glue](https://docs.aws.amazon.com/glue/latest/dg/machine-learning.html), and the AWS Glue developer guide titled [Tuning Machine Learning Transforms in AWS Glue](https://docs.aws.amazon.com/glue/latest/dg/add-job-machine-learning-transform-tuning.html)

Question: 96

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Identify and implement a data-transformation solution

**Domain:** Data Engineering

**Question text**:

You work for a car rental firm in their car tracking department. Your team is responsible for building the machine learning solutions to track the company's fleet of cars. Each car is equipped with a GPS vehicle tracking device that emits IoT data. You are building a data transformation solution to take the GPS IoT data and transform it before storing it in S3 for use in your machine learning models.

You have decided to use Kinesis Data Firehose data transformation to pre-process your IoT data before storing it in S3. You have written your lambda function that pre-processes the data and you are now testing your data transformation process flow. When running your tests you see that Kinesis Data Firehose is rejecting every record as a data transformation failure. What could be the reason for the failure?

1. In your lambda function you have set the result to OK or Dropped for each record processed.
2. The transformed records from your lambda function contain the recordId and result parameters.
3. When creating your lambda function you used a lambda blueprint for data transformation from the AWS Serverless Application Repository.
4. When creating your lambda function you used a lambda blueprint for data transformation from the AWS Lambda console.

**Answer:** B

**Explanation:**

Option A is incorrect. The status of your transformed record produced by your lambda function can be Ok (the record was transformed successfully), Dropped (the record was dropped intentionally by your transformation logic), or ProcessingFailed (the record could not be transformed). A status of Ok or Dropped indicates to Kinesis Data Firehose that the record was successfully processed. A status of ProcessingFailed indicates a failed transformation. Your lambda function has set each record’s status to either Ok or Dropped, so this option is incorrect.

Option B is correct. Transformed records received by Kinesis Data Firehose from lambda must contain the recordId, result, and data parameters. Your transformed records only contain the recordId and result parameters.

Option C is incorrect. You can use lambda blueprints from either the AWS Serverless Application Repository or the AWS Lambda console to create you transformation lambda function.

Option D is incorrect. You can use lambda blueprints from either the AWS Serverless Application Repository or the AWS Lambda console to create you transformation lambda function.

**Reference:**

Please see the Amazon Kinesis Data Firehose developer guide titled [Amazon Kinesis Data Firehose Data Transformation](https://docs.aws.amazon.com/firehose/latest/dev/data-transformation.html)

Question: 97

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Apply basic AWS security practices to machine learning solutions

**Domain:** ML Implementation and Operations

**Question text**:

You work for a healthcare data provider company that gathers real-time streaming data from healthcare plan participants who have agreed to allow their insurance company use their health data gathered by their wearable technology, such as internet connected watches and step counters. The plan participants receive discounts on their healthcare plan fees when participating in the data streaming effort. You are on the machine learning team that will use this data to better predict healthcare issues based on the gathered wearable data. Due to the secure nature of this personal information, you need to build encryption into your data pipeline for this effort.

How would you construct your data pipeline in the most secure way to ensure your data is encrypted as it moves from the IoT wearable devices to your machine learning data source?

1. Use IoT Analytics to gather the streaming data from the IoT devices, encrypt the data, and send it to your machine learning data source.
2. Use Kinesis Data Streams to gather the streaming data from the IoT devices. Have Kinesis Data Streams be the source of a Kinesis Data Firehose delivery stream which encrypts your data using an AWS Key Management Service (AWS KMS) key before storing the data at rest and then delivers the data to your S3 bucket used for your machine learning models.
3. Use Kinesis Data Streams to gather the streaming data from the IoT devices and encrypt your data using an AWS Key Management Service (AWS KMS) key before storing the data at rest. Then have Kinesis Data Streams be the source of a Kinesis Data Firehose delivery stream which delivers the data to your S3 bucket used for your machine learning models.
4. Use Kinesis Data Analytics to gather the streaming data from the IoT devices, encrypt the data, and send it to your machine learning data source.

**Answer:** C

**Explanation:**

Option A is incorrect. IoT Analytics is used to filter, transform, and enrich IoT data before storing the data in a time-series data store for analysis. IoT Analytics doesn’t encrypt your data.

Option B is incorrect. Using Kinesis Data Streams to gather your IoT data and be the source for a Kinesis Data Firehose delivery stream is the correct choice. However, you would leverage Kinesis Data Streams to encrypt your data using an AWS Key Management Service (AWS KMS) key before storing the data at rest, not Kinesis Data Firehose. When you use a Kinesis data stream as the source of a Kinesis Data Firehose delivery stream, Kinesis Data Firehose does not store the data at rest. The data is stored at rest in the Kinesis Data Stream.

Option C is correct. You use Kinesis Data Streams to gather your IoT data and be the source for a Kinesis Data Firehose delivery stream. You also leverage Kinesis Data Streams to encrypt your data using an AWS Key Management Service (AWS KMS) key before storing the data at rest. Then Kinesis Data Streams is used as the source of your to Kinesis Data Firehose delivery stream, which delivers the data to your S3 bucket used for your machine learning models.

Option D is incorrect. You would have to use Kinesis Data Streams together with Kinesis Data Analytics to get the encryption needed for your solution.

**Reference:**

Please see the Amazon Kinesis Data Firehose developer guide titled [Data Protection in Amazon Kinesis Data Firehose](https://docs.aws.amazon.com/firehose/latest/dev/encryption.html), the [Amazon Kinesis Data Analytics overview page](https://aws.amazon.com/kinesis/data-analytics/), the [AWS IoT Analytics overview page](https://aws.amazon.com/iot-analytics/), the AWS IoT Analytics user guide titled [What Is AWS IoT Analytics](https://docs.aws.amazon.com/iotanalytics/latest/userguide/welcome.html), and the Amazon Kinesis Data Analytics for SQL Applications developers guide titled [Data Protection in Amazon Kinesis Data Analytics for SQL Applications](https://docs.aws.amazon.com/kinesisanalytics/latest/dev/data-protection.html)

Question: 98

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Evaluate machine learning models

**Domain:** Modeling

**Question text**:

You work for a fantasy sports wagering software company as a machine learning specialist. You are the leader of a team of machine learning specialists who have been given the assignment of building a model to predict the over/under line for every professional football game each week of the NFL season. Due to the complex nature of the problem and its many feature combinations, you have your team experimenting with different datasets, algorithms, and hyperparameters to find the best combination for your machine learning problem. You don’t want to limit the number of experiments your team can perform. Since you have a relatively large team of talented machine learning specialists, they will generate several hundred to over a thousand experiments over the course of your modeling effort.

Which Amazon machine learning service(s)/feature(s) should you use to help manage your team’s experiments at scale?

1. Use Amazon SageMaker Inference Pipeline
2. Use Amazon SageMaker model tracking capability
3. Use Amazon SageMaker model experiments capability
4. Use Amazon SageMaker model containers capability

**Answer:** B

**Explanation:**

Option A is incorrect. The Amazon Inference Pipeline is used to deploy pretrined SageMaker algorithms packaged in Docker containers. You would not use Amazon Inference Pipeline to manage experiments at scale.

Option B is correct. You can use the Amazon SageMaker model tracking capability to search key model attributes such as hyperparameter values, the algorithm used, and tags associated with your team’s models. This SageMaker capability allows you to manage your team’s experiments at the scale of up to thousands of model experiments.

Option C is incorrect. There is no Amazon SageMaker feature called ‘model experiments capability’

Option D is incorrect. There is no Amazon SageMaker feature called ‘model containers capability’

**Reference:**

Please see the AWS announcement titled [New Model Tracking Capabilities for Amazon SageMaker Are Now Generally Available](https://aws.amazon.com/about-aws/whats-new/2019/08/new-model-tracking-capabilities-for-amazon-sagemaker-now-generally-available/), the Amazon SageMaker developer guide titled [Manage Machine Learning Experiments](https://docs.aws.amazon.com/sagemaker/latest/dg/search.html), the AWS Machine Learning blog titled [Using model attributes to track your training runs on Amazon SageMaker](https://aws.amazon.com/blogs/machine-learning/amazon-sagemaker-now-comes-with-new-capabilities-for-accelerating-machine-learning-experimentation/), the Amazon SageMaker developer guide titled [Monitor and Analyze Training Jobs Using Metrics](https://docs.aws.amazon.com/sagemaker/latest/dg/training-metrics.html), and the Amazon SageMaker developer guide titled [Deploy an Inference Pipeline](https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipelines.html)

Question: 99

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Evaluate machine learning models

**Domain:** Modeling

**Question text**:

You work on an application development team for a new start-up social media site. Your team is made up of data scientists and machine learning specialists, of which you are the lead machine learning specialist. Your team has built a model in SageMaker using the built-in linear learner algorithm. The team has performed several training runs in an effort to find the best datasets and hyperparameters. You have decided to use the SageMaker model tracking capability to manage the many training runs your team has produced.

You have asked your team to show you the results of their efforts to help you lead them in making the decision on which hyperparameters and test datasets to use. They have used the AWS SDK API for SageMaker to produce the data for your decision. The following is a section of code from their use of the SageMaker model tracking capability. What does the code do?

search\_params = {

“MaxResults”: 10,

“Resource”: “TrainingJob”,

“SearchExpression”: {

“Filters”: [{

“Name”: “Tags.Model”,

“Operator”: “Equals”,

“Value”: ”Model\_Social\_Media\_Classifier”,

}]},

“SortBy”: “”Metrics.train:precision”,

“SortOrder”: “Descending”

}

smclient = boto3.client(service\_name=’sagemaker’)

results = smclient.search(\*\*search\_params)

1. It uses the SageMaker API to run at most 10 training jobs for a model called Model\_Social\_Media\_Classifier and sorts the results by the model precision in descending order
2. It uses the SageMaker API to find the the 10 best hyperparameters (based on the precision metric) of a model that has been tagged as Model: Model\_Social\_Media\_Classifier
3. It uses the SageMaker API to find the the 10 best training runs (based on their precision metric) of a model that has been tagged as Model: Model\_Social\_Media\_Classifier
4. It uses the SageMaker API to run a training job called Model\_Social\_Media\_Classifier and sorts the results by the precision metric in descending order for the 10 best results.

**Answer:** C

**Explanation:**

Option A is incorrect. The code uses the SageMaker python client API to search your team’s SageMaker resources (such as training run results) for a specific model’s training run results. It does not run any training jobs.

Option B is incorrect. The code uses the SageMaker python client API to search your team’s SageMaker resources (such as training run results) for a specific model’s training run results. It does not search for the best hyperparameters.

Option C is correct. The code uses the SageMaker python client API to search your team’s SageMaker resources (such as training run results) for a specific model’s training run results. It then sorts the results by the precision metric in descending order. This will allow you to see which training model run performed the best from a precision perspective. The results give that model’s algorithm, data sources, hyperparameter values, and metrics results.

Option D is incorrect. The code uses the SageMaker python client API to search your team’s SageMaker resources (such as training run results) for a specific model’s training run results. It does not run any training jobs.

**Reference:**

Please see the AWS announcement titled [New Model Tracking Capabilities for Amazon SageMaker Are Now Generally Available](https://aws.amazon.com/about-aws/whats-new/2019/08/new-model-tracking-capabilities-for-amazon-sagemaker-now-generally-available/), the AWS Machine Learning blog titled [Using model attributes to track your training runs on Amazon SageMaker](https://aws.amazon.com/blogs/machine-learning/amazon-sagemaker-now-comes-with-new-capabilities-for-accelerating-machine-learning-experimentation/), the Amazon SageMaker developer guide titled [Search](https://docs.aws.amazon.com/sagemaker/latest/dg/API_Search.html), the Amazon SageMaker developer guide titled [Manage Machine Learning Experiments](https://docs.aws.amazon.com/sagemaker/latest/dg/search.html), the [AWS SageMaker Client boto3 docs](https://boto3.amazonaws.com/v1/documentation/api/latest/reference/services/sagemaker.html#SageMaker.Client.search), and the Amazon SageMaker developer guide titled [Tune a Linear Learner Model](https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner-tuning.html)

Question: 100

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Apply basic AWS security practices to machine learning solutions

**Domain:** ML Implementation and Operations

**Question text**:

You work on an application development team for a financial services firm. You and your team are working on a mission critical project with a very aggressive timeline for implementation. For this project you are building a machine learning model to predict customer retention where you are using customer PII (Personal Identifiable Information) data. This data is very sensitive and is also controlled by SEC (Securities Exchange Commision) compliance regulations. Therefore, your data ingestion process and data storage must be highly secure. For this reason, you have a mandate to use encryption for all data storage.

How do you use SageMaker features to make sure all of your model artifacts are highly secure with the least amount of effort on your team’s part?

1. Use SSL to encrypt your data on your S3 bucket (where you store your model artifacts and data) and your SageMaker jupyter notebooks. Then run your SageMaker training jobs, hyperparameter tuning jobs, batch transform jobs, and your inference endpoint using the default SageMaker IAM roles and policies.
2. Use SageMaker Neo, which encrypts your data at rest in your S3 bucket where you store your model artifacts and data. Then pass an AWS Key Management Service key to your SageMaker jupyter notebooks, training jobs, hyperparameter tuning jobs, batch transform jobs, and your inference endpoint to encrypt the S3 bucket.
3. Use encrypted S3 buckets for your model artifacts and data. Then pass an AWS Key Management Service key to your SageMaker jupyter notebooks, training jobs, hyperparameter tuning jobs, batch transform jobs, and your inference endpoint to encrypt the attached machine learning storage volume.
4. Use your customer owned AWS Key Management Service key to store your data on the ML EBS volume or in your S3 buckets, which you encrypt using your customer owned Key Management Service key. Pass your customer owned Key Management Service key to your SageMaker jupyter notebooks, training jobs, hyperparameter tuning jobs, batch transform jobs, and your inference endpoint to encrypt the attached machine learning storage volume.

**Answer:** D

**Explanation:**

Option A is incorrect. To ensure your data is secure you use an AWS Key Management Service key to store your data and to pass to your SageMaker resources. You don’t use SSL for this purpose.

Option B is incorrect. SageMaker Neo is a SageMaker service that allows you to train your model once and run it anywhere in the cloud and at the edge. SageMaker Neo does not provide encryption services.

Option C is incorrect. You should use AWS Key Management Service keys for your data and SageMaker resource encryption, but since your project requires encryption for regulatory compliance reasons, you need to use a customer owned KMS key.

Option D is correct. Since your project requires encryption for regulatory compliance reasons, you need to use a customer owned KMS key. You should use your customer owned AWS KMS key to store your data on the ML EBS volume or in your S3 buckets, which you encrypt using your customer managed KMS keys. You also should pass your customer owned KMS key to your SageMaker jupyter notebooks, training jobs, hyperparameter tuning jobs, batch transform jobs, and your inference endpoint to encrypt the attached machine learning storage volume.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Protecting Data at Rest Using Encryption](https://docs.aws.amazon.com/sagemaker/latest/dg/encryption-at-rest.html), and the [Amazon SageMaker Neo overview page](https://aws.amazon.com/sagemaker/neo/)

Question: 101

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Identify and implement a data-ingestion solution

**Domain:** Data Engineering

**Question text**:

You work for a telecommunications service and internet provider company that has been in business for decades. Over the decades the company has built various types of application systems and database technologies on the evolving platforms of the time. Therefore, you have massive amounts of customer and company operational data on legacy mainframe systems and their associated data stores, such as aging relational databases.

Your team is attempting to build a machine learning model to use streaming data from the company’s in-home routers, functioning as IoT (Internet of Things) devices, and use that data to help the company sell additional services to its customer base. The IoT data is unstructured, so you need to transform it to CSV format before you ingest it into your S3 buckets that you use to house your datasets for your SageMaker model. You also need to enrich the IoT data with real-time data from your legacy mainframe systems as the data streams into your AWS cloud environment.

Which set of Amazon services would you use to setup this data transformation and ingestion pipeline?

1. Use Kinesis Data Firehose to receive the streaming data from the IoT devices. Use the Kinesis Data Firehose lambda integration capability to enrich the IoT data with your legacy mainframe systems data and transform it to CSV before writing it to the S3 bucket used by your SageMaker model.
2. Have your legacy mainframe systems write to S3 and use AWS Storage Gateway to enrich the IoT data with your legacy system data and transform it to CSV before writing it to the S3 bucket used by your SageMaker model.
3. Have your legacy mainframe systems write to AWS Storage Gateway using the File Gateway configuration via an NFS (Network File System) connection. Use Kinesis Data Firehose to receive the streaming data from the IoT devices. Use the Kinesis Data Firehose lambda integration capability to enrich the IoT data with your legacy mainframe systems data and convert it to CSV before writing it to the S3 bucket used by your SageMaker model.
4. Use AWS Snowball to migrate your legacy mainframe data to your AWS account. Use Kinesis Data Firehose to receive the streaming data from the IoT devices. Use the Kinesis Data Firehose lambda integration capability to enrich the IoT data with your legacy mainframe systems data and convert it to CSV before writing it to the S3 bucket used by your SageMaker model.

**Answer:** C

**Explanation:**

Option A is incorrect. You can’t enrich your IoT data with your mainframe data without first getting your mainframe data into your AWS cloud environment.

Option B is incorrect. You can’t write directly from your mainframe systems to S3. You could use AWS Storage Gateway to get your mainframe data into your AWS cloud environment, but AWS Storage Gateway doesn’t have the capability to enrich your IoT data.

Option C is correct. You can use AWS Storage Gateway using the File Gateway configuration via an NFS (Network File System) connection to move your data from your legacy mainframe systems into your AWS cloud environment. You can then use Kinesis Data Firehose lambda integration to to enrich the IoT data with your legacy mainframe systems data and convert it to CSV. Finally, you can have your lambda function write your transformed data to your S3 bucket used by your SageMaker model.

Option D is correct. AWS Snowball moves data from your on premises environment to your AWS cloud environment in a one time batch. This wouldn’t work since you need real-time integration of your legacy data with your IoT data.

**Reference:**

Please see the AWS whitepaper titled [Building Big Data Storage Solutions (Data Lakes) for Maximum Flexibility](https://docs.aws.amazon.com/whitepapers/latest/building-data-lakes/data-ingestion-methods.html)

Question: 102

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Train machine learning models

**Domain:** Modeling

**Question text**:

You work for a ride share software company as a machine learning specialist. You are working on a model to predict driver capacity based on a number of factors, such as location, time of day, weather, population density, age of car, etc. You have several million observations stretching back over 5 years across several geographic locations worldwide. You have performed feature engineering on your data and you have transformed it into 5 CSV files (one for each year) which you have uploaded to your S3 bucket training prefix.

Due to the large number of observations your management team anticipates that training this model could get costly, so they have asked you to keep the costs of your project as low as possible.

You have written the following python code using the SageMaker Python SDK in your SageMaker jupyter notebook:

s3\_train =

sagemaker.s3\_input(s3\_data='s3://{}/{}'.format(bucket, path\_train),

content\_type='csv',distribution='ShardedByS3Key')

my\_container = get\_image\_uri(boto3.Session().region\_name, 'xgboost')

my\_session = sagemaker.Session()

role = get\_execution\_role()

xgb = sagemaker.estimator.Estimator(my\_container,

role,

train\_instance\_count=5,

train\_instance\_type='ml.m4.xlarge',

output\_path=output\_path,

sagemaker\_session=my\_session)

xgb.set\_hyperparameters( max\_depth=10,

eta=0.2,

gamma=4,

min\_child\_weight=40,

subsample=0.8,

silent=0,

objective='reg:linear',

early\_stopping\_rounds=10,

num\_round=200 )

xgb.fit({'train': s3\_train,

'validation': s3\_input\_validation})

Using this code, how does SageMaker replicate your dataset to your Machine Learning instances for training?

1. SageMaker replicates the entire dataset on each of the 10 ML instances that are launched for training
2. SageMaker replicates the entire dataset on each of the 5 ML instances that are launched for training
3. SageMaker replicates a subset of your dataset on each of the 10 ML instances that are launched for training
4. SageMaker replicates a subset of your dataset on each of the 5 ML instances that are launched for training

**Answer:** D

**Explanation:**

Option A is incorrect. In the SageMaker API, when you set the distribution type parameter to ShardedByS3Key, SageMaker replicates a subset of your dataset on each of the ML instances you’ve defined.

Option B is incorrect. In the SageMaker API, when you set the distribution type parameter to ShardedByS3Key, SageMaker replicates a subset of your dataset on each of the ML instances you’ve defined. You define the quantity of the ML instances (in this case 5) in the train\_instance\_count parameter of the Estimator API call.

Option C is incorrect. It is correct that in the SageMaker API, when you set the distribution type parameter to ShardedByS3Key, SageMaker replicates a subset of your dataset on each of the ML instances you’ve defined. You define the quantity of the ML instances (in this case 5) in the train\_instance\_count parameter of the Estimator API call.

Option D is correct. In the SageMaker API, when you set the distribution type parameter to ShardedByS3Key, SageMaker replicates a subset of your dataset on each of the ML instances you’ve defined. You define the quantity of the ML instances (in this case 5) in the train\_instance\_count parameter of the Estimator API call. Distributing your dataset across several instances, making your training much faster and therefore less expensive.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Train a Model with Amazon SageMaker](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html), the Amazon SageMaker developer guide titled [S3DataSource](https://docs.aws.amazon.com/sagemaker/latest/dg/API_S3DataSource.html), and the AWS Machine Learning blog titled [Amazon SageMaker Automatic Model Tuning becomes more efficient with warm start of hyperparameter tuning jobs](https://aws.amazon.com/blogs/machine-learning/amazon-sagemaker-automatic-model-tuning-becomes-more-efficient-with-warm-start-of-hyperparameter-tuning-jobs/) )specifically the ‘create a training estimator’ section of the blog)

Question: 103

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Train machine learning models

**Domain:** Modeling

**Question text**:

You work for a large healthcare diagnostics company. You are on the machine learning team responsible for predicting various anomalies in blood samples. You have data samples from all of the corporation’s many testing facilities across the country. You have performed feature engineering and data cleaning on your dataset. You have also written the python code to split your dataset into training and test datasets. You are now ready to train your model for the first time.

You have written the following python code in your SageMaker jupyter notebook:

import sagemaker

from sagemaker.amazon.amazon\_estimator import get\_image\_uri

from sagemaker import get\_execution\_role

container = get\_image\_uri(boto3.Session().region\_name, 'xgboost')

role = get\_execution\_role()

s3\_train = 's3://{}/{}/{}'.format(bucket, prefix, 'train')

s3\_validation = 's3://{}/{}/{}'.format(bucket, prefix, 'validation')

s3\_output = 's3://{}/{}/{}'.format(bucket, prefix, xgb\_output)

xgb\_model = sagemaker.estimator.Estimator(container,

role,

train\_instance\_count=1,

train\_instance\_type='ml.m4.xlarge',

train\_volume\_size = 5,

output\_path=s3\_output,

sagemaker\_session=sagemaker.Session())

xgb\_model.set\_hyperparameters(max\_depth = 2,

eta = 2,

gamma = 2,

min\_child\_weight = 2,

silent = 0,

objective = "multi:softmax",

num\_class = 10,

num\_round = 10)

train\_channel = sagemaker.session.s3\_input(s3\_train, content\_type='text/csv')

valid\_channel = sagemaker.session.s3\_input(s3\_validation, content\_type='text/csv')

data\_channels = {'train': train\_channel, 'validation': valid\_channel}

xgb\_model.fit(inputs=data\_channels, logs=True)

When you attempt to run this code in your SageMaker jupyter notebook it fails. You check the CloudWatch logs and find this error message:  
  
AlgorithmError: u'2' is not valid under any of the given

schemas\n\nFailed validating u'oneOf' in

schema[u'properties'][u'feature\_dim']:\n {u'oneOf':

[{u'pattern': u'^([0]\.[0-9])$', u'type': u'string'},\n

{u'minimum': 0, u'type': u'integer'}]}\

What is the cause of your error?

1. You have used an invalid hyperparameter
2. You have used an invalid hyperparameter value
3. You have used an invalid train content\_type
4. You have used an invalid objective

**Answer:** B

**Explanation:**

Option A is incorrect. If you had specified an invalid hyperparameter you would get an error such as:  
 ERROR 139623806805824 train.py:48]

Additional properties are not allowed (u'min\_child\_weigh' was

unexpected)

Option B is correct. You specified the value of 2 for the eta hyperparameter, but the valid range for this hyperparameter for the XGBoost algorithm is float range: [0,1]

Option C is incorrect. The valid content types for the XGBoost algorithm are text/libsvm (default) or text/csv. You have used text/csv, so your content type is valid.

Option D is incorrect. The objective multi:softmax is a valid setting for the XGBoost algorithm.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Logs for Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/common-info-all-sagemaker-models-logs.html), the Amazon SageMaker developer guide titled [XGBoost Hyperparameters](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html), and the [XGBoost Parameters GitHub page](https://github.com/dmlc/xgboost/blob/master/doc/parameter.rst) (especially the Learning Task Parameters section)

Question: 104

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Sanitize and prepare data for modeling

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a retail clothing manufacturer that has a very active online web store. You have been assigned the task of building a model to contact customers for a direct marketing campaign based on their predicted receptiveness to the campaign. Some of your customers have been contacted in the past for other marketing campaigns. You don’t want to contact these customers who have been contacted in the past for this latest campaign.

Before training this model, you need to clean your data and prepare it for the XGBoost algorithm you are going to use. You have written your cleaning/preparation code in your SageMaker notebook. Based on the following code, what happens on lines 19, 21, 22? (Select THREE)

1 import sagemaker

2 import boto3

3 from sagemaker.predictor import csv\_serializer

4 import numpy as np

5 import pandas as pd

6 from time import gmtime, strftime

7 import os

8 region = boto3.Session().region\_name

9 smclient = boto3.Session().client('sagemaker')

10 from sagemaker import get\_execution\_role

11 role = get\_execution\_role()

12 bucket = 'sagemakerS3Bucket'

13 prefix = 'sagemaker/xgboost'

14 !wget -N https://.../bank.zip

15 !unzip -o bank.zip

16 data = pd.read\_csv('./bank/bank-full.csv', sep=';')

17 pd.set\_option('display.max\_columns', 500)

18 pd.set\_option('display.max\_rows', 5)

19 data['no\_previous\_campaign'] = np.where(data['contacted'] == 999, 1, 0)

20 data['not\_employed'] = np.where(np.in1d(data['job'], ['student', 'retired', 'unempl']), 1, 0)

21 model\_data = pd.get\_dummies(data)

22 model\_data = model\_data.drop(['duration', 'employee.rate', 'construction.price.idex',

'construction.confidence.idx','lifetime.rate', 'region'], axis=1)

23 train\_data, validation\_data, test\_data = np.split(model\_data.sample(frac=1,

random\_state=1729), [int(0.7 \* len(model\_data)), int(0.9\*len(model\_data))])

24 pd.concat([train\_data['y\_yes'], train\_data.drop(['y\_no', 'y\_yes'], axis=1)],

axis=1).to\_csv('train.csv', index=False, header=False)

25 pd.concat([validation\_data['y\_yes'], validation\_data.drop(['y\_no', 'y\_yes'], axis=1)],

axis=1).to\_csv('validation.csv', index=False, header=False)

26 pd.concat([test\_data['y\_yes'], test\_data.drop(['y\_no', 'y\_yes'], axis=1)],

axis=1).to\_csv('test.csv', index=False, header=False)

27 boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix,

'train/train.csv')).upload\_file('train.csv')

28 boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix,

'validation/validation.csv')).upload\_file('validation.csv')

1. Splits bank dataset into train, validation, and test datasets
2. Sets the attribute no\_previous\_campaign to 999, 0, or 1 depending if the customer in the observation has been contacted via a previous campaign
3. Sets the attribute no\_previous\_campaign to 1 if the customer in the observation has not been contacted via a previous campaign or 0 if they have been contacted via a previous campaign
4. Converts categorical data to a set of indicator variables
5. Converts empty attributes to dummy variables
6. Removes features deemed inconsequential
7. Removes observations deemed inconsequential

**Answers:** C, D, F

**Explanation:**

Option A is incorrect. This option describes what happens on line 23, not what happens on lines 20, 21, or 22.

Option B is incorrect. Line 19 does not set the attribute no\_previous\_campaign to 999. It sets the attribute no\_previous\_campaign to 1 or 0 depending on whether the customer in the observation has been contacted via a previous campaign, as indicated by the value 999.

Option C is correct. Line 19 sets the attribute no\_previous\_campaign to 1 or 0 depending if the customer in the observation has been contacted via a previous campaign, as indicated by the value 999.

Option D is correct. Line 21 uses the pandas library get\_dummies method to convert the categorical attributes in the dataframe to dummy (or indicator) variables.

Option E is incorrect. Line 21 does not convert empty attributes to dummy variables, it uses the pandas library get\_dummies method to convert the categorical attributes in the dataframe to dummy (or indicator) variables.

Option F is correct. Line 22 removes (or drops) several features presumably because you have deemed the features inconsequential to the training of your model.

Option G is incorrect. Line 22, in this usage, calls the pandas drop method to remove features, not observations.

**Reference:**

Please see the [SciPy numpy.where](https://docs.scipy.org/doc/numpy/reference/generated/numpy.where.html) documentation (for line 19), the [pandas get\_dummies](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html) documentation (for line 21), and the [pandas DataFrame.drop](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.drop.html) documentation (for line 22)

Question: 105

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Perform feature engineering

**Domain:** Exploratory Data Analysis

**Question text**:

You work for the credit card division of a large financial services firm. You are a machine learning specialist working on a credit card transaction classification model. Your model will be used to classify your firm’s customer transactions for use in direct marketing campaigns by your firm’s marketing department. You have built your model based on the SageMaker pre-built Linear Learner algorithm. You have also deployed your model to an inference endpoint using an inference pipeline. You are performing your feature engineering via the SageMaker built-in feature transformers so you don’t need to write your own feature engineering logic.

You have defined the containers for your pipeline using the CreateModel SageMaker API and you have created an inference endpoint using the SageMaker CreateEndpointConfig and CreateEndpoint APIs. You have decided to change your pipeline to use a different SageMaker feature transformer strategy (change the strategy from the default None to SingleRecord).

How do you make this change to your inference pipeline?

1. Your pipeline model is mutable, meaning you can change it while it is running.
2. Your pipeline is immutable, but you can update your inference pipeline by deleting the old one and redeploying the new one using the SageMaker CreateEndpointConfig and CreateEndpoint APIs.
3. Your pipeline is immutable, but you can change your inference pipeline by deploying a new one using the ReplaceEndpoint API.
4. Your pipeline is immutable, but you can change your inference pipeline by deploying a new one using the UpdateEndpoint API.

**Answer:** D

**Explanation:**

Option A is incorrect. SageMaker inference pipelines are immutable, so you cannot change them while they are running.

Option B is incorrect. It is true that your inference pipeline is immutable, but you change it via the UpdateEndpoint API. You do not have to delete your pipeline and recreate it.

Option C is incorrect. Your inference pipeline is immutable, but you change it via the UpdateEndpoint API not a ReplaceEndpoint API. The is no ReplaceEndpoint API.

Option D is correct. Your inference pipeline is immutable. You change it by deploying a new one via the UpdateEndpoint API. SageMaker deploys the new inference pipeline, then switches incoming requests to the new one. SageMaker then deletes the resources associated with the old pipeline.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Deploy an Inference Pipeline](https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipelines.html), the Amazon SageMaker developer guide titled [CreateModel](https://docs.aws.amazon.com/sagemaker/latest/dg/API_CreateModel.html), the Amazon SageMaker developer guide titled [UpdateModel](https://docs.aws.amazon.com/sagemaker/latest/dg/API_UpdateEndpoint.html), the Amazon SageMaker developer guide titled [Use Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html), and the SageMaker docs page titled [Transformer](https://sagemaker.readthedocs.io/en/stable/transformer.html)

Question: 106

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Identify and implement a data-transformation solution

**Domain:** Data Engineering

**Question text**:

You work for an online retailer as a machine learning specialist. Your team has been tasked with creating a machine learning model to identify similar products for a product comparison chart on many of the product pages on your website. Your website designers want to show a grid of a product compared to similar products. The grid will show price, review summary (stars), and key features of each product. You are at the stage in your development where you are gathering, cleaning, and transforming your data and training your model.

Using machine learning techniques, how can you determine similar product data for use in this grid in the most efficient manner?

1. Use the Linear Learner built-in SageMaker algorithm and set its predictor\_type hyperparameter to binary\_classifier
2. Use the XGBoost built-in SageMaker algorithm and set its objective hyperparameter to reg:logistic
3. Use the Linear Learner built-in SageMaker algorithm and set its predictor\_type hyperparameter to regressor
4. Use the AWS Glue FindMatches ML Transform and set its precision-recall parameter to precision
5. Use the XGBoost built-in SageMaker algorithm and set its objective hyperparameter to reg:linear
6. Use the AWS Glue FindMatches ML Transform and set its precision-recall parameter to recall

**Answer:** D

**Explanation:**

Option A is incorrect. Using a Linear Learner algorithm based model with the binary\_classifier predictor\_type may help you find similar products, but it is not the most efficient technique listed in the options.

Option B is incorrect. Using a XGBoost algorithm based model with the reg:logistic objective may help you find similar products, but it is not the most efficient technique listed in the options.

Option C is incorrect. Using the Linear Learner algorithm with the regressor predictor\_type would not be a good choice for a discrete categorization problem such as matching similar products.

Option D is correct. The AWS Glue FindMatches ML Transform uses machine learning capabilities to find matching records in your database, even when the records don’t have exactly matching fields. This type of matching is perfect for finding similar products in a products table. Setting the FindMatches ML Transform precision\_recall parameter to precision is the correct parameter setting. You use this setting when you want to minimize false positives. Meaning, you don’t want to show two items as similar when they are not similar.

Option E is incorrect. Using the XGBoost algorithm with the reg:linear objective would not be a good choice for a discrete categorization problem such as matching similar products.

Option F is incorrect. The AWS Glue FindMatches ML Transform uses machine learning capabilities to find matching records in your database, even when the records don’t have exactly matching fields. Setting the FindMatches ML Transform precision\_recall parameter to recall is incorrect since this setting is used when you want to minimize false negatives. Meaning, the ML transform failed to find a match when a match actually existed. This is not an optimal result, but it is a better outcome than incorrectly identifying two items as similar when they really aren’t (false positive).

**Reference:**

Please see the AWS Glue developer guide titled [Machine Learning Transforms in AWS Glue](https://docs.aws.amazon.com/glue/latest/dg/machine-learning.html), the AWS Glue developer guide titled [Tuning Machine Learning Transforms in AWS Glue](https://docs.aws.amazon.com/glue/latest/dg/add-job-machine-learning-transform-tuning.html), and the Amazon SageMaker developer guide titled [Use Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html)

Question: 107

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Create data repositories for machine learning

**Domain:** Data Engineering

**Question text**:

You work for an online fashion retailer as a machine learning specialist. You are on a team of machine learning specialists and data scientists who have been given the responsibility of centralizing your company’s product, customer, supplier, and materials data in one source. This new data source will be used for analytics and for making informed business decisions using KPIs (Key Performance Indicators). Your company has many different data sources where their product, customer, supplier, and materials data is stored. These data repositories are also housed on several different database technologies.

When you load the various data sources into your new centralized data source, you need to clean and classify the data as well. What is the most expeditious and efficient way to create this new centralized data source?

1. Use Amazon EMR and its built-in machine learning tool Apache Spark MLlib to extract the data from your disparate data sources, transform (clean and classify) the data, and load it into an S3 data lake.
2. Use AWS Glue crawlers to crawl your disparate data sources and create a metastore for your S3 data lake. Use AWS Glue to then extract, transform (clean and classify), and load the source data into your S3 data lake.
3. Use Amazon Kinesis Data Firehose to send the data from your disparate data sources to you S3 data lake. Use lambda integration with Kinesis Data Firehose to transform (clean and classify) your data as it loads into your S3 data lake.
4. Use AWS Lake Formation to collect and catalog the data from your disparate data sources, transform (clean and classify) your data, and load the data into your S3 data lake.

**Answer:** D

**Explanation:**

Option A is incorrect. Using Amazon EMR and its built-in machine learning tools will work to extract, transform, and load your disparate data sources into your S3 data lake, but it is not the quickest or simplest option given.

Option B is incorrect. Using AWS Glue and its crawlers will work to extract, transform, and load your disparate data sources into your S3 data lake, but it is not the quickest or simplest option given.

Option C is incorrect. Using Amazon Kinesis Data Firehoe and its lambda integration will work to extract, transform, and load your disparate data sources into your S3 data lake, but it is not the quickest or simplest option given.

Option D is correct. AWS Lake Formation builds on the capabilities of AWS Glue to simplify the creation of an S3 data lake. Once you define your disparate data sources to AWS Lake Formation, it crawls your data sources and moves the data into your S3 data lake. It uses machine learning algorithms to clean and classify your data. This is the simplest and most efficient option listed.

**Reference:**

Please see the [AWS Lake Formation overview page](https://aws.amazon.com/lake-formation/), the [Amazon EMR overview page](https://aws.amazon.com/emr/), the AWS Big Data blog titled [Build a Data Lake Foundation with AWS Glue and Amazon S3](https://aws.amazon.com/blogs/big-data/build-a-data-lake-foundation-with-aws-glue-and-amazon-s3/), and the [Amazon Kinesis overview page](https://aws.amazon.com/kinesis/)

Question: 108

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Identify and implement a data-transformation solution

**Domain:** Data Engineering

**Question text**:

You work for a major banking and financial services firm as a machine learning specialist. Your firm has decided to improve their fraud detection for specialized cases where fraudulent actors attempt to open accounts through your firm’s banking and trading services. These services have websites where potential customers can open accounts by completing online forms. These services make use of your firm’s highly secure customer and account data stores.

You have been assigned the task of determining when a known fraudulent actor attempts to open a new account. You have decided to build a machine learning solution to solve this problem. Since your firm has a very large customer base, several million customer accounts, you need to consider the performance as well as the precision of your fraud detection process.

You have decided to use the AWS Glue FindMatches ML Transform to process your online form data to find matching known fraudulent accounts in your firm’s data stores. Knowing that detecting a fraudulent actor is of primary importance, how should you configure the AWS Glue FindMatches ML Transform parameters to achieve the most performant and accurate fraud detection process?

1. Set the FindMatches precision-recall parameter to ‘precision’ and the accuracy-cost parameter to ‘accuracy’
2. Set the FindMatches precision-recall parameter to ‘precision’ and the accuracy-cost parameter to ‘lower cost’
3. Set the FindMatches precision-recall parameter to ‘recall’ and the accuracy-cost parameter to ‘accuracy’
4. Set the FindMatches precision-recall parameter to ‘recall’ and the accuracy-cost parameter to ‘lower cost’

**Answer:** C

**Explanation:**

Option A is incorrect. Setting the FindMatches precision-recall parameter to ‘precision’ minimizes false positives (when you don’t have a match of a fraudulent account but mark it as a match mistakenly). But you are more concerned about minimizing false negatives (when you have a match of a fraudulent account but fail to detect it).

Option B is incorrect. Setting the FindMatches precision-recall parameter to ‘precision’ minimizes false positives (when you don’t have a match of a fraudulent account but mark it as a match mistakenly). But you are more concerned about minimizing false negatives (when you have a match of a fraudulent account but fail to detect it).

Option C is correct. Setting the FindMatches precision-recall parameter to ‘recall’ minimizes false negatives (when you have a match of a fraudulent account but fail to detect it). This is what you want. Also, setting the FindMatches accuracy-cost parameter to ‘accuracy’ maximizes the transform accuracy of finding matching records as fraudulent.

Option D is incorrect. Setting the FindMatches precision-recall parameter to ‘recall’ minimizes false negatives (when you have a match of a fraudulent account but fail to detect it). This is what you want. But, setting the accuracy-cost parameter to ‘lower cost’ favors cost or the speed of running the transform at the expense of the transform’s accuracy. This may make your transform more performant, but your primary concern is detecting a fraudulent actor so you should set the accuracy-cost parameter to ‘accuracy’.

**Reference:**

Please see the AWS Glue developer guide titled [Machine Learning Transforms in AWS Glue](https://docs.aws.amazon.com/glue/latest/dg/machine-learning.html), and the AWS Glue developer guide titled [Tuning Machine Learning Transforms in AWS Glue](https://docs.aws.amazon.com/glue/latest/dg/add-job-machine-learning-transform-tuning.html)

Question: 109

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Deploy and operationalize machine learning solutions

**Domain:** ML Implementation and Operations

**Question text**:

You work for a government census bureau in their machine learning group. Your team is working on a model that will be used to predict population movement based on many attributes of the population and the geographic regions in which they live and move to and from. Some of the dataset features are id, age, height, weight, family size, country of origin, etc. You have built your model using the SageMaker built-in linear learner algorithm. You have trained your model and deployed it using SageMaker Hosting Services. You are now ready to send inference requests to your inference endpoint. You have chosen to use CSV file data stored on one of your S3 buckets as your inference request data. Since you are processing large census data files you don’t need sub-second latency.

Here is an example of the CSV file data:

| id | age | height (in.) | weight (lb) | family size | country of origin | ...

| 6185 | 23 | 75 | 145 | 3 | USA | ...

| 5437 | 54 | 80 | 187 | 7 | Canada | ...

...

You know that the id attribute in your dataset is not relevant to your model’s prediction results and you didn’t use it when training your model. What is the simplest way you exclude this attribute when you send prediction requests to your inference endpoint, but have the id attribute associated with the prediction results that your model outputs so you can easily analyze the prediction results?

1. Use SageMaker Batch Transform to run the predictions from your CSV file on your S3 bucket and have it exclude the id from the prediction request. Also have Batch Transform join the id attribute to the prediction results.
2. Use Kinesis Data Analytics to stream your prediction requests from your CSV file on your S3 bucket to your inference endpoint. Transform the prediction requests by removing the id attribute. Use Kinesis Data Analytics to join the id attribute to the prediction results.
3. Use Kinesis Data Analytics to stream your prediction requests from your CSV file on your S3 bucket to your inference endpoint. Transform the prediction requests by removing the id attribute. Use Kinesis Data Streams to join the id attribute to the prediction results.
4. Use Kinesis Data Firehose to run the predictions from your CSV file on your S3 bucket and have it exclude the id from the prediction request. Use Kinesis Data Streams to join the id attribute to the prediction results.

**Answer:** A

**Explanation:**

Option A is correct. The simplest way to first exclude the id attribute from the inference prediction requests and then join the id attribute to the prediction results is to use Amazon SageMaker Batch Transform.

Option B is incorrect. While you could use Kinesis DataAnalytics to exclude the id attribute from your prediction request and then to join the attribute with the prediction results, this would not be as simple a solution as just using Batch Transform pre and post processing.

Option C is incorrect. While you could use Kinesis Data Analytics to exclude the id attribute from your prediction requests and then use Kinesis Data Streams to join the id attribute to the prediction results possibly using a lambda function you would have to write, this approach would not be as simple as just using Batch Transform pre and post processing.

Option D is incorrect. You could use Kinesis Data Firehose Data Transformation to exclude your id attribute from your prediction requests and then use Kinesis Data Streams to join the id attribute to the prediction results possibly using a lambda function you would have to write, this approach would not be as simple as just using Batch Transform pre and post processing.

**Reference:**

Please see the AWS announcement titled [SageMaker Batch Transform now enables associating prediction results with input attributes](https://aws.amazon.com/about-aws/whats-new/2019/07/sagemaker-batch-transform-enable-associating-prediction-results-with-input-attributes/), the Amazon SageMaker developer guide titled [Associate Prediction Results with Input Records](https://docs.aws.amazon.com/sagemaker/latest/dg/batch-transform-data-processing.html), the Amazon SageMaker developer guide titled [Deploy a Model on Amazon SageMaker Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-hosting.html), the AWS Lambda developer guide titled [Using AWS Lambda with Amazon Kinesis](https://docs.aws.amazon.com/lambda/latest/dg/with-kinesis.html), and the Amazon Kinesis Data Firehose developer guide titled [Amazon Kinesis Data Firehose Data Transformation](https://docs.aws.amazon.com/firehose/latest/dev/data-transformation.html)

Question: 110

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Train machine learning models

**Domain:** Modeling

**Question text**:

You work as a machine learning specialist for an auto manufacturer who produces several car models in several product lines. Example models include an LX model, an EX model, a Sport model, etc. These models have many similarities, but of course they also have defining differences. Each model has its own parts list entries in your company’s parts database. When ordering commodity parts for these car models from auto parts manufacturers you want to produce the most efficient orders for each parts manufacturer by combining orders for similar parts lists. This will save your company money. You have decided to use the AWS Glue FindMatches Machine Learning Transform to find your matching parts lists.

You have created your data source file as a CSV, and you have also created your labeling file used to train your FindMatches transform. When you run your AWS Glue transform job it fails. Which of the following could be the root of the problem?

1. The labeling file is in the CSV format
2. The labeling file has labeling\_set\_id and label as its first two columns with the remaining columns matching the schema of the parts list data to be processed
3. Records in the labeling file that don’t have any matches have unique labels
4. The labeling file is not encoded in UTF-8 without BOM (byte order mark)

**Answer:** D

**Explanation:**

Option A is incorrect. When using the AWS Glue FindMatches ML Transform, the labeling file must be in CSV format.

Option B is incorrect. When using the AWS Glue FindMatches ML Transform, the first two columns of the labeling file are required to be labeling\_set\_id and label. Also, the remaining columns must match the schema of the data to be processed.

Option C is correct. When using the AWS Glue FindMatches ML Transform, if a record doesn’t have a match, it is assigned a unique label.

Option D is correct. When using the AWS Glue FindMatches ML Transform, the labeling file must be encoded as UTF-8 without BOM.

**Reference:**

Please see the AWS Glue developer guide titled [Machine Learning Transforms in AWS Glue](https://docs.aws.amazon.com/glue/latest/dg/machine-learning.html)

Question: 111

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Sanitize and prepare data for modeling

**Domain:** Exploratory Data Analysis

**Question text**:

You work as a machine learning specialist for an electric bicycle company. The electric bicycles your company produces have IoT sensors on them that transmit usage and maintenance information to your company data lake. You are using Kinesis Data Streams to gather the bicycle IoT data and store it into an S3 data store that you can use for your machine learning models. You are on the team that has the assignment of using the IoT data to predict when your customer’s electric bicycles need maintenance.

The IoT data that the electric bicycles produce is unstructured, and sometimes, depending on the manufacturer of the IoT part, the data has a different schema structure. You need to clean and classify the IoT data before using it in your machine learning model. How can you build an ETL script to perform the necessary cleaning and classification knowing you have message data with differing schema structures?

1. Use AWS Glue to build a series of transforms that use Apache Spark SparkSQL DataRecord to pass the data from transform to transform. Each transform performing a different cleaning and/or transforming task.
2. Use AWS Glue to build a series of transforms that use Apache Spark SparkSQL DataFrames to pass the data from transform to transform. Each transform performing a different cleaning and/or transforming task.
3. Use AWS Glue to build a series of transforms that uses DynamicFrames to pass the data from transform to transform. Each transform performing a different cleaning and/or transforming task.
4. Use AWS Glue to build a series of transforms that uses DynamicRecord to pass the data from transform to transform. Each transform performing a different cleaning and/or transforming task.

**Answer:** C

**Explanation:**

Option A is incorrect. There is no DataRecord construct in Apache Spark SparkSQL.

Option B is incorrect. The Apache Spark SparkSQL DataFrame does not efficiently handle data with unknown schema structure. This option would produce suboptimal results.

Option C is correct. The AWS Glue DynamicFrame allows for each record to be self-describing so it can handle unknown or changing schemas.

Option D is incorrect. DynamicRecord represents a logical record within a DynamicFrame. It is a row in a DynamicFrame. So you wouldn’t pass individual DynamicRecords from transform to transform, you pass a DynamicFrame.

**Reference:**

Please see the AWS Glue developer guide titled [Machine Learning Transforms in AWS Glue](https://docs.aws.amazon.com/glue/latest/dg/machine-learning.html), and the AWS Glue developer guide titled [DynamicFrame Class](https://docs.aws.amazon.com/glue/latest/dg/aws-glue-api-crawler-pyspark-extensions-dynamic-frame.html)

Question: 112

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Identify and implement a data-ingestion solution

**Domain:** Data Engineering

**Question text**:

You work as a machine learning specialist for a book publishing company. Your company has several publishing data stores housed in relational databases across its infrastructure. Your company recently purchased another publishing company and are in the process of merging the two company’s systems infrastructure. A part of this merger activity is joining the two publisher book databases. Your team has been given the assignment to build a data lake sourced from the two company’s relational data stores.

How would you construct an ETL pipeline to achieve this goal? (Select FOUR)

1. Use AWS DataSync to ingest the relational data from your book data stores and store it in S3
2. Use an AWS Glue crawler to build your AWS Glue catalog
3. Have a lambda function triggered by an S3 trigger to start your AWS Glue crawler
4. Use an AWS Glue trigger to start your AWS Glue ETL job that processes/transforms your data and places it into your S3 data lake
5. Use a lambda function triggered by a CloudWatch event trigger to start your AWS Glue ETL job that processes/transforms your data and places it into your S3 data lake
6. Use AWS Database Migration Service to ingest the relational data from your book data stores and store it in S3

**Answers:** B, C, E, F

**Explanation:**

Option A is incorrect. AWS DataSync is used to ingest data from a Network File System (NFS), not relational databases.

Option B is correct. Once your data has been ingested from your databases, you need to catalog the data using an AWS Glue crawler.

Option C is correct. The AWS Glue crawler can be started by a lambda function that is triggered by an S3 object create event.

Option D is incorrect. It is not possible to start an AWS Glue ETL job from an AWS Glue trigger.

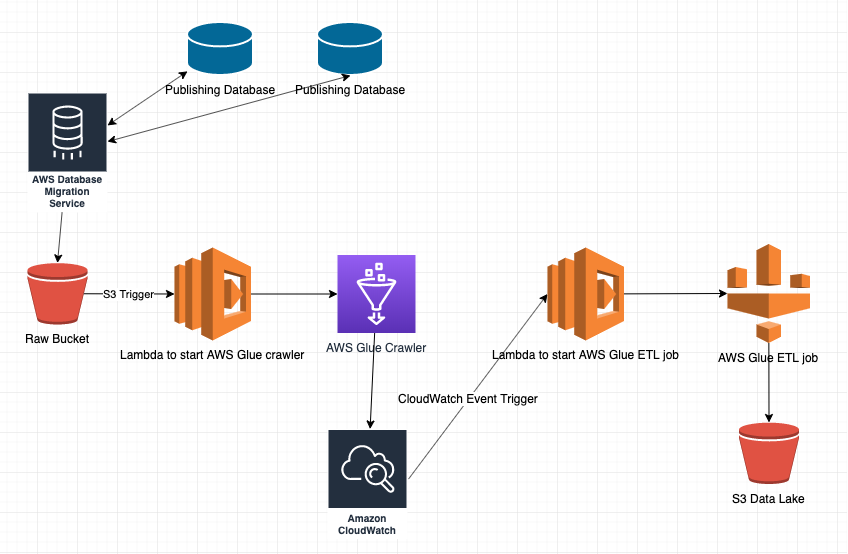
Option E is correct. You can have your AWS Glue ETL job started by a lambda function that is triggered by a CloudWatch event trigger.

Option F is correct. You can use the AWS Database Migration Service to ingest your data from your relational databases and then store the data in an S3 bucket.

**Reference:**

Please see the AWS Big Data blog titled [Build and automate a serverless data lake using an AWS Glue trigger for the Data Catalog and ETL jobs](https://aws.amazon.com/blogs/big-data/build-and-automate-a-serverless-data-lake-using-an-aws-glue-trigger-for-the-data-catalog-and-etl-jobs/), and the AWS article titled [How can I automatically start an AWS Glue job when a crawler run completes?](https://aws.amazon.com/premiumsupport/knowledge-center/start-glue-job-run-end/)

Here is a diagram representing the proposed solution:



Question: 113

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Deploy and operationalize machine learning solutions

**Domain:** ML Implementation and Operations

**Question text**:

You work for a sports wagering company as a machine learning specialist. Your team is responsible for building the machine learning models that produce the sports wager line for the NFL (National Football League) games each week. You are working on the line versus the spread model. For this model you have chosen the XGBoost algorithm. You have trained your model and deployed it to Amazon SageMaker Hosting Services where you are now ready to send inference requests to your model.

You are sending requests to your inference endpoint, but you are seeing that your inferences are failing. Which of these would NOT be the source of the problem? (Select TWO)

1. You have serialized your inference request in the text/csv format
2. You have serialized your inference request in the application/x-recordio-protobuf format
3. You have serialized your inference request in the text/libsvm format
4. You have serialized your inference request in the application/json format

**Answers:** A, C

**Explanation:**

Option A is correct. Inference endpoints built using the XGBoost algorithm only support the text/csv and text/libsvm request formats.

Option B is incorrect. Inference endpoints built using the XGBoost algorithm only support the text/csv and text.libsvm request formats. Your inference request will fail if you serialize your inference request using the application/x-recordio-protobuf format.

Option C is correct. Inference endpoints built using the XGBoost algorithm only support the text/csv and text/libsvm request formats.

Option D is incorrect. Inference endpoints built using the XGBoost algorithm only support the text/csv and text.libsvm request formats. Your inference request will fail if you serialize your inference request using the application/json format.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Deploy a Model on Amazon SageMaker Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-hosting.html), the Amazon SageMaker developer guide titled [CreateEndpoint](https://docs.aws.amazon.com/sagemaker/latest/dg/API_CreateEndpoint.html), and the Amazon SageMaker developer guide titled [Common Data Formats for Inference](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-inference.html)

Question: 114

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Train machine learning models

**Domain:** Modeling

**Question text**:

You work for a computer peripheral manufacturer that builds printers, external hard drives, etc. You are on the machine learning team where you are currently building a machine learning model to be used to find anomalies in the functional behavior of your company’s line of printers. The printers generate IoT device messages that are streamed to your madel S3 bucket using Amazon Kinesis Data Streams. You have performed your data cleansing and data engineering of you IoT printer data. You are now ready to start training your model. You have chosen the Random Cut Forest SageMaker built-in algorithm for your model. You hope to find anomalies in your customer’s printer activity by looking for outlier observations using your Random Cut Forest based model. Finding these anomalies will help your company provide better customer service.

You have started your first training job, but you see that your training job is failing. What may be the cause of this failure?

1. You have selected compute resources of the GPU compute instance class
2. You have selected compute resources of the CPU compute instance class
3. You have built your training data files using the CSV file type
4. You have built your training data files using the recordio-protobuf file type

**Answer:** A

**Explanation:**

Option A is correct. SageMaker only supports the CPU instance class for the Random Cut Forest algorithm.

Option B is incorrect. SageMaker only supports the CPU instance class for the Random Cut Forest algorithm. So selecting the instance class of CPU would not cause your training job to fail.

Option C is incorrect. SageMaker supports both the CSV and recordio-protobuf file types for training data files. So using the CSV file type for your training data would not cause your training job to fail.

Option D is incorrect. SageMaker supports both the CSV and recordio-protobuf file types for training data files. So using the recordio-protobuf file type for your training data would not cause your training job to fail.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Train a Model with Amazon SageMaker](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html), the Amazon SageMaker developer guide titled [Common Parameters for Built-In Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-algo-docker-registry-paths.html)

Question: 115

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Evaluate machine learning models

**Domain:** Modeling

**Question text**:

You work for a power tool manufacturer as a machine learning specialist. You work in the battery powered power tool division where your team of machine learning specialists and data scientists have been tasked with building a model that predicts the lifespan of particular models of power tools. You have selected the Linear Learner algorithm on which to build your model. You have cleaned and engineered your features for your training and test data. Your feature engineering transformations convert all feature attributes to integers or real numbers. You have also trained your model and have deployed it to Amazon SageMaker Hosting Services.

Your training dataset has this structure:

| model | power | battery Ah | use pattern | region | country |

For your client application inference requests, how would you structure the body argument for your invoke\_endpoint call?

1. A string with this value: “547,3.5,1.5,23.4,2,43,1”
2. A string with this value: “547,3.5,1.5,23.4,2,43”
3. A string with this value: “Quite strike,battery,1.5,frequent,North America,US”
4. An array set to these values: [547,3.5,1.5,23.4,2,43]
5. A list set to these values: [547,3.5,1.5,23.4,2,43]

**Answer:** B

**Explanation:**

Option A is incorrect. The Linear Learner algorithm expects either CSV or recordio-protobuf as the inference request content type. For text/csv the value of the body argument for the invoke\_endpoint API call should be a string with with comma separated values for each feature. This option has a comma separated string, but it has 7 values, when you only have 6 features in your data used to train your model.

Option B is correct. The Linear Learner algorithm expects either CSV or recordio-protobuf as the inference request content type. For text/csv the value of the body argument for the invoke\_endpoint API call should be a string with with comma separated values for each feature. This option has a comma separated string and it has 6 values. You also have 6 features in your data used to train your model, so this inference request is structured correctly.

Option C is incorrect. The Linear Learner algorithm expects either CSV or recordio-protobuf as the inference request content type. Also, any transforms performed on the data for training also must be performed on inference request data before attempting to obtain an inference. The body argument in this option has not been transformed in the way your training data was transformed.

Option D is incorrect. The Linear Learner algorithm expects either CSV or recordio-protobuf as the inference request content type. Also, the body argument to the invoke\_endpoint should be a string, not an array.

Option E is incorrect. The Linear Learner algorithm expects either CSV or recordio-protobuf as the inference request content type. Also, the body argument to the invoke\_endpoint should be a string, not a list.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Train a Model with Amazon SageMaker](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html), the Amazon SageMaker developer guide titled [Common Parameters for Built-In Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-algo-docker-registry-paths.html), and the Amazon SageMaker developer guide titled [Common Data Formats for Inference](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-inference.html)

Question: 116

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Evaluate machine learning models

**Domain:** Modeling

**Question text**:

You work for an Internet of Things (IoT) component manufacturer which builds servos, engines, sensors, etc. The IoT devices transmit usage and environment information back to AWS IoT Core via the MQTT protocol. You want to use a machine learning model to show how/where the use of your products is clustered in various regions around the world. This information will help your data scientists build KPI dashboards for use in improving your component engineering quality and performance. You have created, trained, and deployed to Amazon SageMaker Hosting Services your model based on the XGBoost algorithm. Your model is set up to receive inference requests from a lambda function that is triggered by the receipt of an IoT Core MQTT message via your Kinesis Data Streams instance.

What transform steps need to be done for each inference request. Also which steps are handled by your code versus by the inference algorithm? (Select TWO)

1. Inference request serialization (handled by the algorithm)
2. Inference request serialization (handled by your lambda code)
3. Inference request deserialization (handled by your lambda code)
4. Inference request deserialization (handled by the algorithm)
5. Inference request post serialization (handled by the algorithm)

**Answers:** B, D

**Explanation:**

Option A is incorrect. The inference request serialization must be completed by your lambda code. The algorithm needs to receive the inference request in serialized form.

Option B is correct. The inference request serialization must be completed by your lambda code.

Option C is incorrect. The inference request is deserialized by the algorithm in the response to the inference request. Your lambda code is responsible for serializing the inference request.

Option D is correct. The inference request is deserialized by the algorithm in the response to the inference request.

Option E is incorrect. There is no inference request post serialization step in the SageMaker inference request/response process.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Common Data Formats for Inference](https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-inference.html), the [AWS IoT Core overview page](https://aws.amazon.com/iot-core/), the AWS IoT developer guide titled [Creating an AWS Lambda Rule](https://docs.aws.amazon.com/iot/latest/developerguide/iot-lambda-rule.html)

Question: 117

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Recommend and implement the appropriate machine learning services and features for a given problem

**Domain:** ML Implementation and Operations

**Question text**:

You work for a farming equipment component manufacturer which builds farm product containers like corn silos, mike containers, etc. These containers have IoT sensors built into them that transmit information such as fill rate, capacity usage, etc. The IoT devices transmit their data back to your cloud environment via the MQTT protocol. You want to use a machine learning model to predict container usage buy region and by product stored. This information will help your management team use real-time dashboards to better understand their product marketing campaigns by region. You have created, trained, and deployed to Amazon SageMaker Hosting Services your model based on the Linear Learner algorithm.

What AWS services would you use to create your pipeline to feed your inference requests to you model? (Select THREE)

1. IoT Greengrass to receive the IoT device MQTT messages
2. IoT Core to receive the IoT device MQTT messages
3. Elastic Beanstalk to stream the IoT messages
4. Kinesis Data Streams to stream the IoT messages
5. A Lambda function to transform the IoT message data to the inference request serialization format
6. API Gateway to transform the IoT message data to the inference request serialization format
7. Route 53 to stream the IoT messages

**Answers:** B, D, E

**Explanation:**

Option A is incorrect. AWS IoT Greengrass is used to extend AWS to edge devices, such as your sensors in your farming containers. Greengrass is used to perform prediction directly on the devices themselves. IoT Core is a better option for receiving your MQTT IoT messages for processing via your machine learning inference running in Amazon SageMaker Hosting Services.

Option B is correct. IoT Core is designed to allow IoT devices interact with other AWS services, such as Kinesis Data Streams.

Option C is incorrect. Elastic Beanstalk is used to host web applications or worker nodes for web applications. You wouldn’t use Elastic Beanstalk to stream IoT messages.

Option D is correct. Kinesis Data Streams can receive IoT messages from IoT Core and then stream to your SageMaker inference endpoint via a Lambda function.

Option E is correct. You can write a lambda function that is triggered by Kinesis Data Streams that transforms you IoT messages into the inference serialization format required by your inference endpoint.

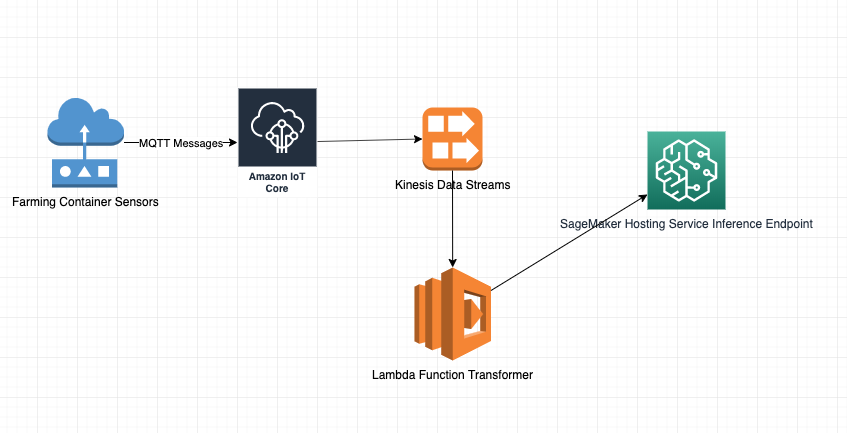
Option F is incorrect. API Gateway is used to create an API request endpoint. You wouldn’t ise API Gateway to transform IoT message data. You would have to have a lambda function behind your API Gateway to accomplish this.

Option G is incorrect. Route 53 is Amazon’s DNS server implementation. You can’t use Route 53 to transform messages.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Train a Model with Amazon SageMaker](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html), the [AWS IoT Core overview](https://aws.amazon.com/iot-core/), the AWS IoT developer guide titled [Creating an AWS Lambda Rule](https://docs.aws.amazon.com/iot/latest/developerguide/iot-lambda-rule.html), and the [AWS IoT Greengrass overview](https://aws.amazon.com/greengrass/)

Here is a diagram of the proposed solution:



Question: 118

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Analyze and visualize data for machine learning

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a management consulting firm as a machine learning specialist. You are on a team of data scientists and other machine learning specialists. Your team has been assigned the task of building a machine learning model to predict Return On Investment (ROI) for new potential engagements that your management consults may wish to take onto their book of business.

You have a dataset of past engagements that has many features that can help you define your problem as a machine learning problem. Before you decide on which machine learning algorithms to evaluate you wish to visualize the historical data to get an idea of the relationships between three of the key features of your dataset: ROI, investment time, and investment size.

Which type of visualization would best give you an idea of the relationship between these three features?

1. Pie chart
2. Tree map
3. Column histogram
4. Bar chart
5. Bubble chart
6. Line chart

**Answer:** E

**Explanation:**

Option A is incorrect. A pie chart is best used to show the portion of the total for each slice of the pie. This type of chart doesn’t work well with three dimensions, such as ROI, investment time, and investment size.

Option B is incorrect. A tree map chart also shows the portion of the total. This type of chart is good for data with a long tail. But it also would not work well on three dimensions.

Option C is incorrect. Column histograms are distribution charts. They show how data is distributed of intervals. But you a looking for a visualization to show the relationship between three variables.

Option D is incorrect. A bar chart is a comparison chart. These types of charts are good for showing how feature values change over time or to show a static snapshot of how different variables compare with each other. But you are looking for the relationship between three variables, not change over time or a static snapshot comparison.

Option E is correct. A bubble chart is a relationship chart. For a relationship between two variables you could use a scatter chart. For the relationship between 3 variables, a bubble chart shows the relationships as such: x-axis for investment time, y-axis for ROI, and the bubble size for investment size.

Option F is incorrect. A line chart is used to show a comparison of variables changing over time. You are looking for a relationship between three variables, not how they change over time.

**Reference:**

Please see the [AWS Data Visualization page](https://aws.amazon.com/data-visualization/), and the [Amazon QuickSite overview page](https://aws.amazon.com/quicksight/)

Question: 119

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Select the appropriate model(s) for a given machine learning problem

**Domain:** Modeling

**Question text**:

You work for a major retail chain in their web development area. You are on the machine learning team responsible for building a recommendation engine for the company’s retail website where they sell many different items across many different categories. The recommendation engine will use customer data such as purchase history, credit rating, geographic location, household income, response to past marketing mailings, etc. Your marketing team has decided to send a marketing mailing to customers who have responded to past mailings. They have two different content templates to use depending on the classification category of each customer. Your model needs to recommend which mailing template to use for each customer in the target customer dataset.

Which SageMaker built-in algorithm is best suited to this problem, and what value should you use for the predictor\_type hyperparameter for the desired outcome? (Select TWO)

1. Linear Learner
2. classifier
3. regressor
4. Factorization Machine
5. multiclass\_classifier
6. K-Means
7. binary\_classifier
8. Neural Topic Model

**Answers:** D, G

**Explanation:**

Option A is incorrect. The Linear Learner is best suited for discrete classification problems. But you have already classified your customers, you are now trying to provide a discrete recommendation. The Factorization Machine algorithm is better suited for this type of problem.

Option B is incorrect. The classifier predictor\_type hyperparameter value is not a valid choice for the Factorization Machine algorithm. The classifier predictor\_type hyper parameter value is a valid choice for the K-Nearest-Neighbor algorithm.

Option C is incorrect. The regressor predictor\_type hyperparameter value setting is used for regression type problems and therefore is not the correct choice for this type of problem. The regressor predictor\_type hyperparameter setting is used when you are solving for a quantitative value. You are trying to solve for a discrete value.

Option D is correct. The Factorization Machine algorithm is a good choice for problems where you are trying to solve for a discrete recommendation.

Option E is incorrect. The multiclass\_classifier predictor\_type hyperparameter value is not a valid choice for the Factorization Machine algorithm. The multiclass\_classifier predictor\_type hyperparameter value is a valid choice for the Linear Learner algorithm.

Option F is incorrect. The K-Means algorithm is best used for grouping observations. You are trying to solve a discrete recommendation problem, you are not trying to group customers.

Option G is correct. The binary\_classifier predictor\_type hyperparameter value is the correct choice for this discrete recommendation problem where you are attempting to choose one of two possible outcomes (one of the two content templates).

Option H is incorrect. The Neural Topic Model algorithm is best suited to organizing documents into topics using groupings of words based on their statistical distribution within the documents. This algorithm is not a good choice for a discrete recommendation problem.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Use Amazon SageMaker Built-in Algorithms](https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html), and the AWS Machine Learning blog titled [Build a movie recommender with factorization machines on Amazon SageMaker](https://aws.amazon.com/blogs/machine-learning/build-a-movie-recommender-with-factorization-machines-on-amazon-sagemaker/)

Question: 120

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Analyze and visualize data for machine learning

**Domain:** Exploratory Data Analysis

**Question text**:

You work for the city planning department of a major metropolitan city in the United States. You are on the city’s machine learning team where you are responsible for creating a model that assists in the resource planning for police officers in the city. Each day the city has to assign police officers to each precinct according to varying parameters. You have data from the past several years for your city and other US cities of a similar makeup. You are in the process of deciding which algorithm to use for your police officer allocation model. Your goal is to predict the police officer allocation size for a given shift based on your dataset features.

Your city dataset has the following features:

1. Infrastructure average age
2. Square feet
3. Citizens
4. Precincts
5. Residences
6. Population density
7. Police officers

Before you select an algorithm you need to perform feature selection and dimensionality reduction of your features. You only want to select features that are relevant to your training dataset, i.e. dimensionality reduction. This process will help you prevent overfitting, and increase computation efficiency through simplification of the feature set.

You have chosen to use visualization techniques to decide which of your 7 features are the most important, or most relevant. In other words, which of your 7 features are needed to properly train your model.

Which visualization techniques are the best to use for this purpose? (Choose TWO)

1. Cat plot
2. Swarm plot
3. Pairs plot
4. Covariance matrix
5. Entropy matrix

**Answers:** C, D

**Explanation:**

Option A is incorrect. A catplot is used to show the relationship between a numerical value and one or more categorical variables using a visualization such as violinplot, boxenplot, etc. But you are trying to show relationships between pairs of data, such as police officers to population density, or police officers to precincts.

Option B is incorrect. A swarm plot is used to show categorical scatter plot data that shows the distribution of values for each feature. But you are trying to show relationships between pairs of data, such as police officers to population density, or police officers to precincts.

Option C is correct. A pairs plot is used to show the relationship between pairs of features as well as the distribution of one of the variables in relation to the other. This is what you need to analyze. You want to see which features correlate well with your police officers feature.

Option D is correct. A covariance matrix shows the degree of correlation between two features. This visualization gives you a numerical representation of the correlation, where the pairs plot gives you a visual representation as points plotted in two dimensional space.

Option E is incorrect. Entropy represents the measure of randomness in your features. This measure would not help you find the correlation between your target feature, police officers, and the potential training features.

**Reference:**

Please see the article titled [Feature Selection and Dimensionality Reduction Using Covariance Matrix Plot](https://medium.com/towards-artificial-intelligence/feature-selection-and-dimensionality-reduction-using-covariance-matrix-plot-b4c7498abd07), the article titled [Visualizing Data with Pairs Plots in Python](https://towardsdatascience.com/visualizing-data-with-pair-plots-in-python-f228cf529166), and the article titled [What is Entropy and why Information gain matter in Decision Trees?](https://medium.com/coinmonks/what-is-entropy-and-why-information-gain-is-matter-4e85d46d2f01)

Question: 121

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Frame business problems as machine learning problems

**Domain:** Modeling

**Question text**:

You work for a firm that produces cameras that can be used for research studies of animals in the wild. When placed in the wild, these cameras are used to identify individual animals and groups of animals as they pass in front of the camera. Researchers use your company’s cameras to catalog animal traffic and specific animal counts in geographic areas where these animals are suspected to live. An example is the identification and counting of wolves in Canada and the far reaches of North America.

Using your company’s cameras, you and your team of machine learning specialists have been contracted by the Wolf Conservation Center of North America to build a machine learning model to identify and count a specific species of wolf in remote areas of the Arctic Circle.

What type of machine learning problem are you trying to solve?

1. Linear regression
2. Binary classification
3. Multidimensional regression
4. Multiclass classification

**Answer:** D

**Explanation:**

Option A is incorrect. A linear regression is used to model the relationship between a dependent variable and one or more independent variables. For example: what will the sales in the north american region be when the GDP (Gross Domestic Product) is trending up and interest rates are trending down. You are trying to solve a classification problem with images as your inference data.

Option B is incorrect. A binary classification is used to classify an observation into one of two categories. For example: based on the image data, is the animal in the image a wolf or not a wolf. You are trying to solve a multiclass classification problem, what type of animal is in the image? You are looking for a specific species of wolf.

Option C is incorrect. A multidimensional regression is used to find more than one real number values. For example: what is the height and width of the animal in the image? You are trying to solve a multiclass classification problem: what type of animal is in the image? You are looking for a specific species of wolf.

Option D is correct. A multiclass classification solves a classification problem where you have more than one class for your answer. For example: of all the animals identified in a given region, what type of animal is in the image? This is the type of problem you are trying to solve. Of all the types of wolves identified to live in the Arctic Circle, what specific species of wolf is in the image?

**Reference:**

Please see the Amazon Machine Learning developer guide titled [Formulating the Problem](https://docs.aws.amazon.com/machine-learning/latest/dg/formulating-the-problem.html), and the article titled [Frame a problem as a machine learning problem or otherwise](https://www.datasciencecentral.com/profiles/blogs/frame-a-problem-as-a-machine-learning-problem-or-otherwise)

Question: 122

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Select the appropriate model(s) for a given machine learning problem

**Domain:** Modeling

**Question text**:

You work for a manufacturing firm that is attempting to build a rechargeable battery that has capacity multiple times greater than the current rechargeable batteries on the market. As a machine learning specialist on the team responsible for building a machine learning model that can predict the chemical component interaction that maximizes battery capacity, you have decided that none of the built-in algorithms available in SageMaker fit your problem as well as you would like. So you and your team have decided to create your own SageMaker algorithm resource. You’ll use this custom algorithm to train and run inferences on your model.

Which of the following steps do you NOT need to complete to create your custom algorithm for use in SageMaker?

1. Create Docker containers for your training and inference code
2. Specify the hyperparameters that your algorithm supports
3. Specify the metrics that your algorithm sends to CloudWatch when training
4. The instance types your algorithm supports for training and inference
5. Whether your algorithm supports distributed inference across multiple instances

**Answer:** E

**Explanation:**

Option A is incorrect. SageMaker uses Docker containers for your custom algorithm training and hosting your algorithm.

Option B is incorrect. When you create a custom algorithm resource in SageMaker you need to specify the hyperparameters your algorithm will support.

Option C is incorrect. When you create a custom algorithm resource in SageMaker you need to specify the metrics that your algorithm will send to CloudWatch when running your training jobs.

Option D is incorrect. When you create a custom algorithm resource in SageMaker you need to specify the EC2 instance types your algorithm supports for training and inference.

Option E is correct. When you create a custom algorithm resource in SageMaker you need to specify whether it supports distributed training across multiple instances, not distributed inference.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Use Your Own Algorithms or Models with Amazon SageMaker](https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms.html), and the Amazon SageMaker developer guide titled [Create an Algorithm Resource](https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-mkt-create-algo.html)

Question: 123

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Perform Feature Engineering

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a company that manufactures cell phone peripherals such as bluetooth headphones and bluetooth selfie sticks. Your company has designed their products so that they act as IoT devices that send usage and diagnostic MQTT messages to your IoT Core service running in AWS. Your machine learning team wants to use this IoT message data to run inferences through your machine learning inference end-point. However, the IoT data is unstructured so you need to preprocess the data by performing feature engineering on the observations before they are fed into your inference endpoint.

You have decided to use a SageMaker Inference Pipeline to construct this machine learning solution. As you are defining the containers for your pipeline, one for feature engineering preprocessing, and one for inference predictions, which SageMaker CLI command and which parameter on that command do you need to run using the SageMaker CLI in order to build your inference pipeline?

1. CreateModel command with the EndpointArn request parameter
2. UpdateEndpoint command with the Containers parameter
3. CreateModel command with the PrimaryContainer request parameter
4. CreateModel command with the Containers request parameter
5. UpdateEndpoint command with the ModelArn parameter

**Answer:** D

**Explanation:**

Option A is incorrect. The SageMaker CLI CreateModel command is the correct command but EndpointArn is a response element of the UpdateEndpoint command.

Option B is incorrect. The SageMaker CLI UpdateEndpoint command is used to switch from an existing endpoint to a new endpoint. You would not use this command to create a new inference pipeline container.

Option C is incorrect. The SageMaker CLI CreateModel command is the correct command but you use thePrimaryContainer request parameter when you want to create a single container, not when you want to create an inference pipeline.

Option D is correct. The SageMaker CLI CreateModel command is the correct command and the Containers parameter is the correct parameter. The Containers request parameter is used to set the containers that make up your pipeline.

Option E is incorrect. The SageMaker CLI UpdateEndpoint command is used to switch from an existing endpoint to a new endpoint. You would not use this command to create a new inference pipeline container.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Deploy an Inference Pipeline](https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipelines.html), the Amazon SageMaker developer guide titled [CreateModel](https://docs.aws.amazon.com/sagemaker/latest/dg/API_CreateModel.html), the Amazon SageMaker developer guide titled [UpdateEndpoint](https://docs.aws.amazon.com/sagemaker/latest/dg/API_UpdateEndpoint.html), the AWS CLI Command Reference titled [create-model](https://docs.aws.amazon.com/cli/latest/reference/sagemaker/create-model.html), and the [AWS IoT Core overview page](https://aws.amazon.com/iot-core/)

Question: 124

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Perform Feature Engineering

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a large manufacturer of consumer electronic devices. Your company wishes to build a machine learning model to predict which product has the most dedicated following among its consumer base. This product will receive funding for future investment in new models and/or enhancements to existing models. You and your machine learning team have a vast amount of observations of the use of the current product base. You know you and your team need to perform feature engineering on the large dataset before you use it to train your XGBoost algorithm based model for predictions.

What SageMaker feature can you use to perform the required feature engineering of your dataset in the most efficient way?

1. Automatic Model Tuning
2. Built-In Transforms
3. Batch Transform
4. Hosting Services

**Answer:** C

**Explanation:**

Option A is incorrect. The SageMaker Automatic Model Tuning feature is used to automatically adjusting thousands of different combinations of hyperparameters to give you the most accurate predictions for your model. But you are trying to perform feature engineering transformation prior to training, so this option is not correct.

Option B is incorrect. The Built-In Transforms feature is part of the AWS Glue service, not SageMaker.

Option C is correct. The SageMaker Batch Transform feature can be used to preprocess your data before using the data in your training runs.

Option D is incorrect. The SageMaker Hosting Services feature is used to allow your model to provide inferences once you’ve trained your model. But you are trying to perform feature engineering transformation prior to training, so this option is not correct.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Run Batch Transforms with Inference Pipelines](https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipeline-batch.html), the Amazon SageMaker developer guide titled [Get Inferences for an Entire Dataset with Batch Transform](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-batch.html), the [Amazon SageMaker Features overview page](https://aws.amazon.com/sagemaker/features/), the Amazon SageMaker developer guide titled [Deploy a Model on Amazon SageMaker Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-hosting.html), and the AWS Glue developer guide titled [Built-In Transforms](https://docs.aws.amazon.com/glue/latest/dg/built-in-transforms.html)

Question: 125

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Evaluate Machine Learning Models

**Domain:** Modeling

**Question text**:

You work for a real estate ecommerce company. Your machine learning team is building a house price prediction model to be used on your company’s site. This model will be used as a guide to users as an unbiased objective estimate of a given house’s value. Your company has gathered an enormous dataset of house observations from across the United States. The observations in the dataset are categorized by region of the country. The housing data prices are mainly clustered by region across the dataset. However, each region has several outlier priced houses.

Since you have defined the housing price prediction work as a regression problem, you have selected the XGBoost SageMaker built-in algorithm on which to base your model. You are now ready to do your hyperparameter tuning so you need a good regression evaluation metric. Which of the following evaluation metrics best fit your problem?

1. MSE (Mean Squared Error)
2. AUC (Area Under the Curve)
3. ROC curve (Receiver Operating Characteristic) curve
4. MAE (Mean Absolute Error)

**Answer:** D

**Explanation:**

Option A is incorrect. The MSE metric is useful for measuring regression problems, however it does not handle outliers as well as the MAE metric. Your dataset has several outliers per region.

Option B is incorrect. The AUC metric is best used for classification type machine learning algorithms. You are using a regression algorithm.

Option C is incorrect. The AUC metric is best used for classification type machine learning algorithms. You are using a regression algorithm.

Option D is correct. The MAE is the correct regression metric to use when your dataset can be significantly influenced by outliers. Your dataset contains several outliers per region.

**Reference:**

Please see the article titled [20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics](https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce), the Amazon SageMaker developer guide titled [XGBoost Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html), and the Amazon SageMaker developer guide titled [Tune an XGBoost Model](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-tuning.html)

Question: 126

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Create data repositories for machine learning

**Domain:** Data Engineering

**Question text**:

You work for a startup ecommerce site that sells various consumer products. Your company has just launched their ecommerce website. The site provides the capability for your users to rate their purchases and the products they have purchased from your ecommerce site. You would like to use the review data to build a recommender machine learning model.

Since your ecommerce site is very new, you don’t yet have a very large review dataset to use for your recommendation model. You have decided to use the Amazon Customer Reviews dataset available from the AWS website as a first data source for your machine learning model. Since your website sells similar products to the products sold on Amazon, you will use the Amazon Customer Reviews dataset as the basis for your initial training runs of your model. Once you have enough data from your own ecommerce site you’ll use that data.

Your goal is to perform sentiment analysis on the review dataset to create your own dataset that will be the source used for your recommender machine learning model. Which set of AWS services would you use to build your data pipeline to produce your sentiment dataset for use by your SageMaker model?

1. S3 -> AWS Glue ETL -> Comprehend -> S3 -> SageMaker
2. S3 -> AWS Glue ETL -> Comprehend -> S3 -> Athena -> QuickSite -> SageMaker
3. S3 -> Kinesis Data Firehose -> Comprehend -> S3 -> SageMaker
4. S3 -> Kinesis Data Firehose -> Lambda -> S3 -> SageMaker

**Answer:** A

**Explanation:**

Option A is correct. The Amazon Customer Reviews dataset is stored on S3. You can use an AWS Glue ETL job to read the reviews from the Amazon dataset. The ETL job calls Comprehend for each review to get the sentiment for that review. The ETL job stores the sentiment enriched review data onto another S3 bucket in your account. Your SageMaker model uses the S3 bucket in your account as its dataset source for training your recommender model.

Option B is incorrect. This option has unnecessary steps. Specifically, you don’t need Athena and QuickSite to produce your sentiment enriched dataset for your machine learning model.

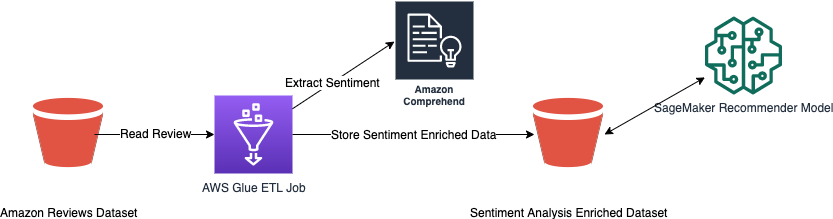
Option C is incorrect. The option uses Kinesis Data Firehose unnecessarily. The Amazon Customer Reviews dataset is stored on S3, there is no need to stream the data when you can simply read it using an ETL job. Also, if you used Kinesis Data Firehose to stream the data you would have to write a lambda function to call Comprehend for each streamed review data row.

Option D is incorrect. The option uses Kinesis Data Firehose unnecessarily. The Amazon Customer Reviews dataset is stored on S3, there is no need to stream the data when you can simply read it using an ETL job. That being said, this option does correctly combine Kinesis Data Firehose and lambda. However it lacks the Comprehend service. You would have to write your own sentiment analysis in your lambda function.

**Reference:**

Please see the data repository titled [Registry of Open Data on AWS](https://registry.opendata.aws/?source=post_page-----bb6d0dc3378b----------------------), the AWS Machine Learning blog titled [How to scale sentiment analysis using Amazon Comprehend, AWS Glue and Amazon Athena](https://aws.amazon.com/blogs/machine-learning/how-to-scale-sentiment-analysis-using-amazon-comprehend-aws-glue-and-amazon-athena/), and the data set titled [Amazon Customer Reviews Dataset](https://s3.amazonaws.com/amazon-reviews-pds/readme.html)

Here is a diagram of the proposed solution:



Question: 127

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Sanitize and prepare data for modeling

**Domain:** Exploratory Data Analysis

**Question text**:

You work for a retail athletic footwear company. Your company has just completed production of a new running shoe that contains IoT sensors in the shoe. These sensors are used to enhance the runner’s running experience by giving detailed data about foot plant, distance, acceleration, gait, and other data points for use in personal running performance analysis.

You are on the machine learning team assigned the task of building a machine learning model to use the shoe IoT sensor data to make predictions of shoe life expectancy based on user wear and tear of the shoes. Instead of just using raw running miles as the predictor of shoe life, your model will use all of the IoT sensor data to produce a much more accurate prediction of remaining life of the shoes.

You are in the process of building your dataset for training your model and running inferences from your model. You need to clean the IoT sensor data before you use it for training or use it to provide inferences from your inference endpoint. You have decided to use Spark ML jobs within AWS Glue to build your feature transformation code. Which machine learning packages are the best choices for building your IoT sensor data transformer tasks in the simplest way possible? (Select THREE)

1. MLeap
2. MLib
3. SparkML Serving Container
4. SparkML Batch Transform
5. MLTransform
6. SparkML MapReduce

**Answers:** A, B, C

**Explanation:**

Option A is correct. AWS Glue serializes Spark ML jobs into MLeap containers. You add these MLeap containers to your inference pipeline.

Option B is correct. Apache Spark MLib is a machine learning library that lets you build machine learning pipeline components where you can transform your data using the full suite of standard transformers such as tokenizers, OneHotEncoders, normalizers, etc.

Option C is correct. The SparkML Serving Container allows you to deploy an Apache Spark ML pipeline in SageMaker.

Option D is incorrect. Batch Transformer is a feature of SageMaker that allows you to get inferences for an entire dataset. Batch Transform is not an Apache SparkML feature.

Option E is incorrect. There is no Apache SparkML feature called MLTransform.

Option F is incorrect. There is no Apache SparkML feature called MapReduce.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Feature Processing with Spark ML and Scikit-learn](https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipeline-mleap-scikit-learn-containers.html), the [MLeap documentation](http://mleap-docs.combust.ml/), the [SageMaker SparkML Serving Container GitHub repo](https://github.com/aws/sagemaker-sparkml-serving-container), the [Apache Spark MLib overview page](http://spark.apache.org/mllib/), the Apache Spark MLib docs page titled [Extracting, transforming, and selecting features](http://spark.apache.org/docs/latest/ml-features.html), the Amazon SageMaker developer guide titled [Deploy a Model on Amazon SageMaker Hosting Services](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-hosting.html), and the Amazon SageMaker developer guide titled [Get Inferences for an Entire Dataset with Batch Transform](https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-batch.html)

Question: 128

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Frame business problems as machine learning problems

**Domain:** Modeling

**Question text**:

You work for a robotics company that is building a new product that allows commuters to ride electric skateboards to work. These skateboards are equipped with IoT sensors for safety measures. The sensors detect obstacles in the path of the skateboard and alert the rider with haptics and sound. The onboard software also uses the IoT sensor data to adjust the skateboard’s performance based on its surroundings. This allows the rider who follows similar paths to work on their daily commute to have their skateboard become more adept at handling the surroundings commonly encountered on this path.

Which type of machine learning model would you use to build the onboard software for these commuter skateboards?

1. Unsupervised Learning model
2. Supervised Learning model
3. Reinforcement Learning model
4. Semi-Supervised Learning model

**Answer:** C

**Explanation:**

Option A is incorrect. Unsupervised learning is used to find patterns in your training dataset when you don’t have preexisting labels. It is self-organizing. This type of model is not the best choice for learning an environment through exploration, which is what you are trying to do using your skateboard IoT sensor data. The better choice is Reinforcement Learning.

Option B is incorrect. Supervised learning is used when you have a training dataset that is labeled. In your skateboard learning example, you don’t have any labels for your IoT sensor observations. Therefore, you could not use supervised learning for this type of problem.

Option C is correct. Reinforcement learning is used when you want to find the best way to achieve a goal or improve performance of a task. Your IoT sensor driven model is trying to improve the performance of the task of alerting for safety hazards as you ride the board through your daily commute environment.

Option D is incorrect. Semi-Supervised learning is used when you have a dataset with both labeled and unlabeled data from which to train your model. In the IoT sensor driven environment exploration use case, you will not have labeled data since you will discover new observations as your IoT sensor equipped skateboard moves through its environment. Therefore, this type of machine learning model is not the best choice for this type of problem.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Reinforcement Learning with Amazon SageMaker RL](https://docs.aws.amazon.com/sagemaker/latest/dg/reinforcement-learning.html), and the NVIDIA blog titled [SuperVize Me: What’s the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning?](https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/)

Question: 129

**Main​ ​Topic​ ​:​** Machine Learning

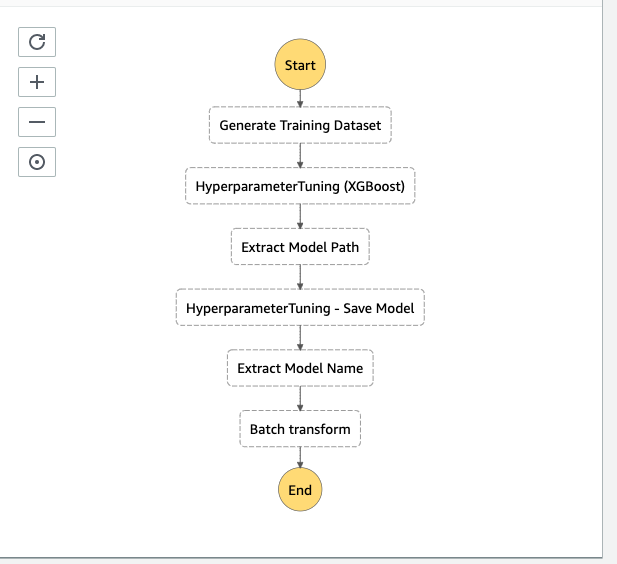
**Sub​ ​Topic​ ​:​** Build machine learning solutions for performance, availability, scalability, resiliency, and fault tolerance

**Domain:** ML Implementation and Operation

**Question text**:

You work for a mining company in their machine learning department. You and your team are working on a model to predict the minimum depth at which to drill in order to find various mineral deposits. You are building a model based on the XGBoost algorithm and your team is at the stage where you are running various models based on different hyperparameters in order to find the best hyperparameter settings. Because of the complexity of the problem you may have to run hundreds, or even thousands of hyperparameter tuning jobs to get the best result.

Your machine learning pipeline also includes a batch transform step to be executed after every hyperparameter tuning job. Your team lead has suggested that you use the Amazon Step Functions SageMaker integration capability to automate the execution of your many hyperparameter tuning jobs. You have setup your Step Functions environment and you have configured it as such:



You have written the following JSON-based Amazon States Language (ASL) for your State Machine (partial listing):

{

"StartAt": "Generate Training Dataset",

"States": {

"Generate Training Dataset": {

"Resource": "<GENERATE\_LAMBDA\_FUNCTION\_ARN>",

"Type": "Task",

"Next": "HyperparameterTuning (XGBoost)"

},

"HyperparameterTuning (XGBoost)": {

"Resource": "arn:<PARTITION>:states:::sagemaker:createHyperParameterTuningJob.sync",

"Parameters": {

"HyperParameterTuningJobName.$": "<JOB\_NAME\_FROM\_LAMBDA>",

"HyperParameterTuningJobConfig": {

"Strategy": "Bayesian",

"HyperParameterTuningJobObjective": {

"Type": "Minimize",

"MetricName": "validation:rmse"

},

"ResourceLimits": {

"MaxNumberOfTrainingJobs": 2,

"MaxParallelTrainingJobs": 2

},

"ParameterRanges": {

"ContinuousParameterRanges": [{

"Name": "alpha",

"MinValue": "0",

"MaxValue": "1000",

"ScalingType": "Auto"

},

{

"Name": "gamma",

"MinValue": "0",

"MaxValue": "5",

"ScalingType": "Auto"

}

],

...

Based on your Step Functions code, what is the type of metric you are using for your regression evaluation? Additionally, in the HyperparameterTuning (XGBoost) step, what happens when the alpha parameter increases through its range of 0 to 1,000? (Select TWO)

1. Relative Mean Square Error
2. Gamma
3. Alpha
4. Root Mean Square Error
5. Mean Square Error
6. As alpha increases, the model becomes more conservative
7. As alpha increases, the model becomes less conservative
8. As alpha increases the model gains precision but sacrifices accuracy

**Answers:** D, F

**Explanation:**

Option A is incorrect. The rmse metric acronym stands for Root Mean Square Error, not Relative Mean Square Error.

Option B is incorrect. The Gamma parameter defines the minimum loss reduction used to partition leaf nodes of the tree within the algorithm. This parameter is not used as a regression evaluation objective.

Option C is incorrect. The Alpha parameter defines regularization terms on weights within the algorithm. This parameter is not used as a regression evaluation objective.

Option D is correct. Your code specifies the rmse metric as the objective on which to evaluate the tuning model run. The rmse acronym stands for Root Mean Square Error.

Option E is incorrect. Your code specifies the rmse metric as the objective on which to evaluate the tuning model run. The rmse acronym stands for Root Mean Square Error, not Mean Square Error.

Option F is correct. As the value of the alpha parameter increases, makes the model more conservative.

Option G is incorrect. As the value of the alpha parameter increases, makes the model more conservative, not less conservative.

Option H is incorrect. As the value of the alpha parameter increases, makes the model more conservative, it does not make the model gain precision while sacrificing accuracy.

**Reference:**

Please see the Amazon announcement titled [Amazon SageMaker Announces New Machine Learning capabilities for Orchestration, Experimentation and Collaboration](https://aws.amazon.com/about-aws/whats-new/2018/11/amazon-sagemaker-announces-new-machine-learning-capabilities-for-orchestration-experimentation-and-collaboration/), the AWS Step Functions developer guide titled [Manage Amazon SageMaker with Step Functions](https://docs.aws.amazon.com/step-functions/latest/dg/connect-sagemaker.html), the Amazon SageMaker developer guide titled [Tune an XGBoost Model](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-tuning.html), and the XGBoost docs page titled [XGBoost Parameters](https://xgboost.readthedocs.io/en/latest/parameter.html)

Question: 130

**Main​ ​Topic​ ​:​** Machine Learning

**Sub​ ​Topic​ ​:​** Evaluate machine learning models

**Domain:** Modeling

**Question text**:

You work for a startup shirt manufacturer that has come up with a new manufacturing process for shirts that is very stylish and has become very popular since your company ran an online kickstarter fundraiser and shipped its first line of shirts. You now want to use machine learning to classify your shirt styles as either conservative or not based on customer feedback on your website. This classification information will help your designers target new designs based on the customer perception of your current offerings.

You have gathered your data from your website comments and ratings. You have also performed feature engineering of your data. You are now ready to run several model tuning jobs, as many as needed even if you have to run hundreds of them, to find the best version of your XGBoost model. You plan to do this by running many hyperparameter tuning jobs that test the range of hyperparameters you have available to you. Since you have decided on using a binary classifier algorithm and based on the business problem you are trying to solve, you have decided you need to measure the success of a hyperparameter tuning job based on precision and recall.

Which XGBoost metric is the best objective on which to evaluate your model?

1. accuracy
2. error
3. F1
4. MAE (Mean Absolute Error)
5. MAP (Mean Average Precision)
6. merror

**Answer:** C

**Explanation:**

Option A is incorrect. The accuracy metric only measures (right cases)/(all cases), which doesn’t give you precision or recall, which are the two metrics you wish to use to evaluate your model.

Option B is incorrect. The error metric only measures (wrong cases)/(all cases), which doesn’t give you precision or recall, which are the two metrics you wish to use to evaluate your model.

Option C is correct. The f1 metric combines precision and recall into one metric. It represents the harmonic mean of precision and recall. Its formula: 2\*precision\*recall/(precision+recall).

Option D is incorrect. The mae metric finds the absolute value of the error between the predicted and target values. This doesn’t give you precision or recall, which are the two metrics you wish to use to evaluate your model.

Option E is incorrect. The map metric finds the mean average precision. This doesn’t give you recall, which is one of the two metrics you wish to use to evaluate your model.

Option F is incorrect. The merror metric is a multiclass classification error rate, which is represented as (wrong cases)/(all cases). This doesn’t give you recall, which is one of the two metrics you wish to use to evaluate your model. Also, this metric is used for multiclass classification problems, you are trying to solve a binary classification problem.

**Reference:**

Please see the Amazon SageMaker developer guide titled [Tune an XGBoost Model](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-tuning.html), the XGBoost docs page titled [XGBoost Parameters](https://xgboost.readthedocs.io/en/latest/parameter.html), and the article titled [20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics](https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce)