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?

A) Run your training data through a preprocessing script that uses the SMOTE (Synthetic Minority Over-sampling Technique) approach
B) Run your training data through a Spark pipeline in AWS Glue to one-hot encode the features
C) Run your training data through a preprocessing script that uses the feature-split technique
D) 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)

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)

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)

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)

Reference:
Please see the article How to Handle Imbalanced Classification Problems in machine learning

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?

A) Hire human image labelers to process all of your images and label them.
B) 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.
C) 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.
D) 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

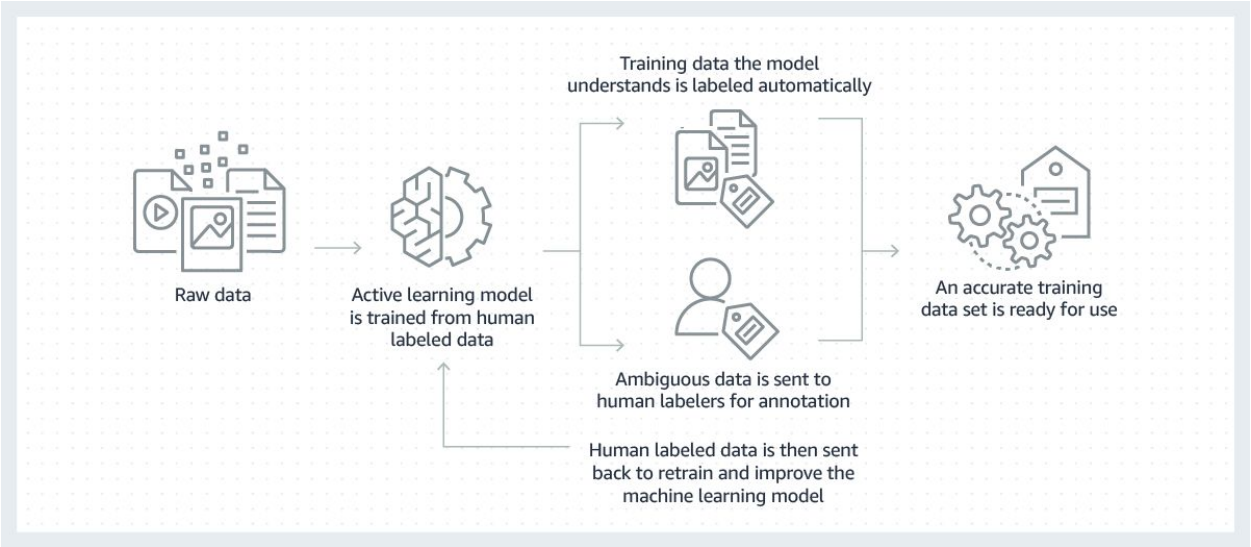
Explanation:
Option A is correct. Human labelers may be able to correctly label all of your images, but they will be slow and expensive.

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 and the Amazon SageMaker Ground Truth overview)

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)

Reference:
Here is a diagram of how the Amazon SageMaker Ground Truth service works:



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?

A) Use Amazon Kinesis Data Streams and its Kinesis Producer Library to pass your events from your consumers to your Kinesis stream.
B) Use Amazon Kinesis Data Streams and its Kinesis API PutRecords call to pass your events from your consumers to your Kinesis stream.
C) Use Amazon Kinesis Data Streams and its Kinesis Client Library to pass your events from your consumers to your Kinesis stream.
D) 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)

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)

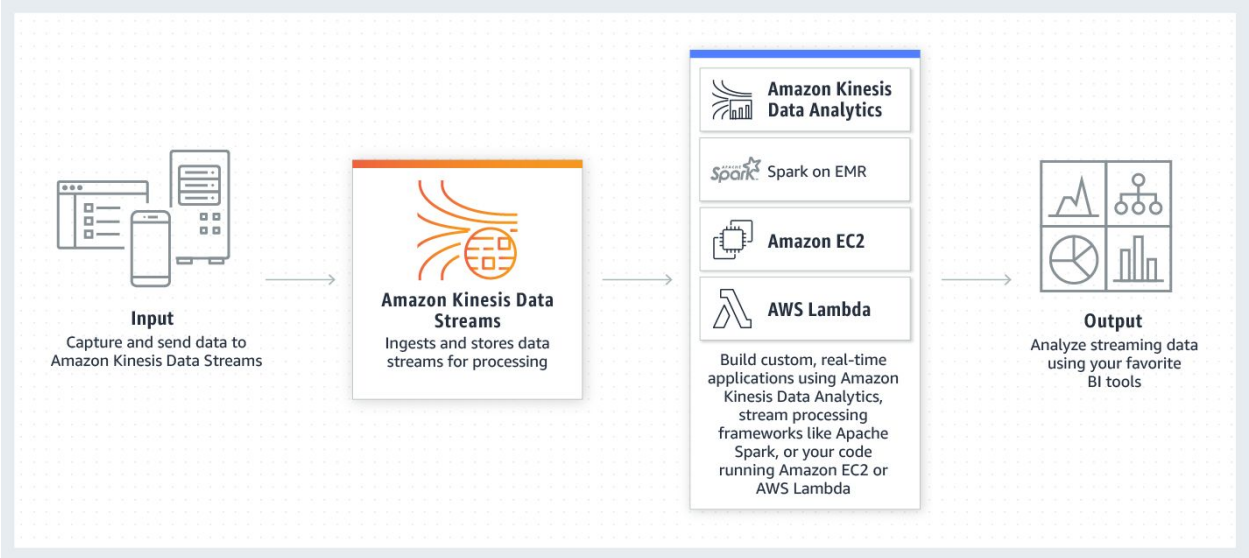
Option C is incorrect. The Amazon Kinesis Data Streams Client Library interacts with the Kinesis Producer Library to process its event data. Therefore, you’ll have the same processing delay problem with this option. (See the AWS developer documentation titled Developing Consumers Using the Kinesis Client Library 1.x)

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)

Reference:

Please see the Amazon Kinesis Data Streams documentation.

Here’s a depiction from the documentation of how it works:



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. 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?

A) 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.
B) 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.
C) 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.
D) 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

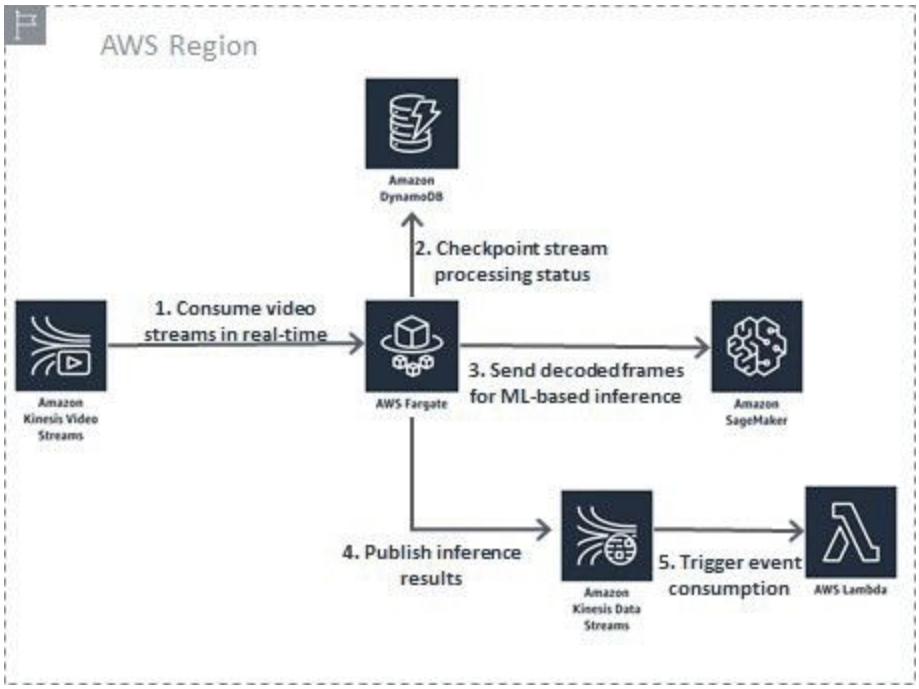
Explanation:
Option A is incorrect. The AWS Rekognition service is not meant to process streams. It works with Kinesis Video Streams to provide this capability. Also it needs another component to send its output to your SageMaker model. This part of the solution is missing.

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)

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)

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

Reference:
Please see the Amazon Kinesis Video Streams documentation.
Here's a depiction of the proposed solution (from the AWS machine Learning blog titled: 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?

A) 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.
B) 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.
C) 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.
D) Use Amazon Kinesis Data Firehose to stream the data directly to your S3 data lake.

Answer: A

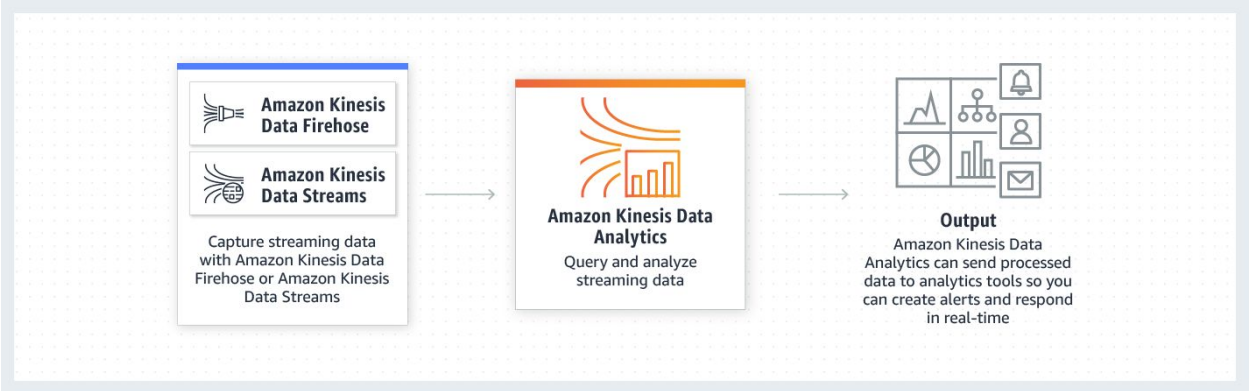
Explanation:
Option A is correct. Amazon Kinesis Data Analytics is a very efficient service for taking streams from Amazon Kinesis Data Streams and transforming them with sql or Apache Flink. (See the Amazon Kinesis Data Analytics overview)

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.
Here is a high level depiction of the best option (from the AWS documentation):



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?

A) ion
B) grokLog
C) xml
D) 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)

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)

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)

Option D is correct. From the Apache Hive Language Manual: "The *Optimized Row Columnar* (ORC) 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 and he AWS developer guide documentation titled Format Options for ETL Inputs and Outputs in AWS Glue)

Reference:
Please see the AWS developer guide documentation titled General Information about Programming AWS Glue ETL Scripts.