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:
See the Amazon SageMaker Ground Truth service overview) and the Amazon Rekognition overview

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.

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.
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)

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.

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.

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?

A) 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
B) 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
C) Create a Kinesis Firehose which gathers the data and puts it onto your S3 bucket

D) 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.

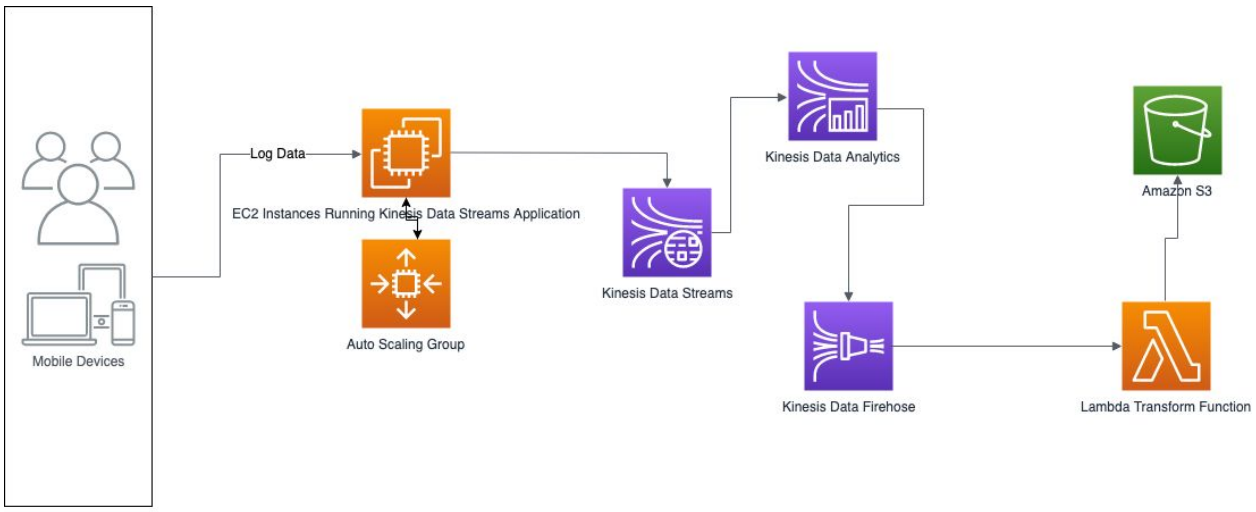
Option B is correct. This option scales well at the Kinesis Data Streams application level because of the Auto Scaling Group. It also uses Kinesis Data Analytics to transform the data into the subset you need and uses the Kinesis Firehose lambda option to convert the playing time to the proper format.

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, the AWS Auto Scaling documentation, the Amazon Kinesis Data Firehose documentation, and the Amazon Kinesis Data Analytics documentation.

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?

A) Your lambda memory setting is set to the maximum value allowed
B) Your S3 bucket is in the same region as your Kinesis Data Firehose
C) Your Kinesis Data Firehose buffer setting is set to the default value
D) 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, the Amazon Kinesis Data Firehose developer guide documentation titled Amazon Kinesis Data Firehose Data Transformation, and the AWS Lambda developer guide documentation titled AWS Lambda Function Configuration.

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?

A) Encrypt the data at the data consumer application level
B) Encrypt the data by configuring Firehose to use S3-managed encryption keys (SSE-S3)
C) Encrypt the data by configuring Firehose to use S3 server-side encryption with AWS Key Management Service (SSE-KMS)
D) 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)

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).”

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)

Reference:
Please see the Amazon Kinesis Data Firehose developer guide documentation titled Creating an Amazon Kinesis Data Firehose Delivery Stream, and the Amazon Kinesis Data Streams developer guide documentation titled What is Server-Side Encryption for Kinesis Data Streams.

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.

A) You have not created an IAM role for your Kinesis Firehose to access the S3 bucket
B) You defined a cluster security group and associated it with your Redshift cluster
C) The access policy associated with your Kinesis Firehose does not have lambda:InvokeFunction specified in the Allow Action section of the Lambda actions
D) The access policy associated with your Kinesis Firehose does not have kms:GenerateDataKey specified in the Allow Action section of the KMS actions
E) 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
“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)

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)

Reference:
Please see the Amazon Kinesis Data Firehose developers guide titled Grant Kinesis Data Firehose Access to Amazon Redshift Destination, and the Amazon Kinesis Data Firehose overview page, and the Amazon Redshift database developer guide titled Amazon Redshift Security Overview.

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.

A) Data
B) ImplicitHashKey
C) ExplicitHashKey
D) PartitionKeys
E) SequenceNumberForOrdering
F) 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 "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)

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)

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)

Reference:
Please see the Amazon Kinesis Data Streams developers guide titled Kinesis Data Stream Producers, and the Amazon Kinesis Data Streams API reference guide titled PutRecord.

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.

A) StandardScaler to encode the Solar_Rating feature
B) OneHotEncoder to encode the Area feature
C) SimpleImputer to complete the missing values in the Solar_Rating and Area features
D) OneHotEncoder to encode the Type feature
E) OrdinalEncoder to complete the missing values in the Solar_Rating and Area features
F) OrdinalEncoder to encode the Solar_Rating feature
G) LabelBinarizer to encode the Foreclosed feature
H) MinMaxScaler to encode the Foreclosed feature

Answers: C, D, F, and G

Explanation:
Options A, is incorrect. From the scikit-learn API Reference: 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: 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: 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: 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: the LabelBinarizer transformer binarizes label in a one-versus-all fashion.

Option H is incorrect. From the scikit-learn API Reference: 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, and the scikit-learn API Reference.

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?

A) Cartesian Product Transformation
B) N-Gram Transformation
C) Orthogonal Sparse Bigram (OSB) Transformation
D) Normalization Transformation
E) Quantile Binning Transformation

Answer: E

Explanation:

Options A is incorrect. From the Amazon Machine Learning developer guide titled Data Transformations Reference "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 "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 "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 "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 "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, and the article Feature Engineering in Machine Learning (Part 1) Handling Numeric Data with Binning

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?

A) Scatter plot
B) Histogram
C) Line chart
D) Box plot
E) 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 "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 "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 "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 "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 "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 v_i 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, and the Statistics How To article titled Types of Graphs Used in Math and Statistics

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?

A) Metadata Content-Type is identified as text/csv
B) Metadata Content-Type is identified as application/x-recordio-protobuf;boundary=1
C) Metadata Content-Type is identified as application/x-recordio-protobuf;label_size=1
D) 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)

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)

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)

Reference:
Please see the Amazon SageMaker developer guide, specifically Common Data Formats for Built-in Algorithms and Common Data Formats for Training

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)

A) Metadata Content-Type is identified as text/csv
B) Metadata Content-Type is identified as text/csv;label_size=0
C) Target value should be in the first column with no header
D) Target value should be in the last column with no header
E) Target value should be in the last column with a header
F) 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)

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)

Option C is correct. From the Amazon SageMaker developer guide titled Common Data Formats for Training "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 "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 "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 "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 and Common Data Formats for Training

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)

A) alpha
B) base_score
C) eta
D) num_round
E) gamma
F) 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)

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)

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)

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)

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)

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)

Reference:
Please see the Amazon SageMaker developer guide titled Automatic Model Tuning, and the Amazon SageMaker developer guide titled How Hyperparameter Tuning Works

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?

A) Bayesian Search
B) Hidden Markov Models
C) Conditional Random Fields
D) 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

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?

A) Leave One Out Cross Validation (LOOCV)
B) K-Fold Cross Validation
C) Stratified Cross Validation
D) 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)

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, and he article K-Fold and Other Cross-Validation Techniques

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.

Which modeling approach best fits your problem?

A) Multi-Class Classification
B) Simulation-based Reinforcement Learning
C) Binary Classification
D) Heuristic Approach

Answer: C

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 transaction fraud detection problem we only have two outcomes: fraud or not fraud. This is a binary classification problem.

Option B is incorrect. Simulation-Based Reinforcement Learning is used in problems where your model needs to learn through trial and error. This is not a good choice for a binary classification problem such as predicting fraud or not fraud.

Option C is correct. Binary Classification is the right approach since you are trying to predict a binary outcome: a transaction is fraudulent or it is not fraudulent.

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, the Amazon SageMaker developer guide titled Reinforcement Learningwith Amazon SageMaker RL, the Amazon Machine Learning developer guide titled Multiclass Classification, and the article titled What is the difference between a machine learning algorithm and a heuristic, and when to use each?

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?

A) Multi-Class Classification
B) Simulation-based Reinforcement Learning
C) Binary Classification
D) 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, the Amazon SageMaker developer guide titled Reinforcement Learningwith Amazon SageMaker RL, the Amazon Machine Learning developer guide titled Multiclass Classification, and the article titled 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?

A) Multi-Class Classification
B) Simulation-based Reinforcement Learning
C) Binary Classification
D) 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, the Amazon SageMaker developer guide titled Reinforcement Learningwith Amazon SageMaker RL, the Amazon Machine Learning developer guide titled Multiclass Classification, and the article titled 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?

A) K-Means
B) Linear Learner
C) Sequence-to-Sequence
D) 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)

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)

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)

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)

Reference:
Please see the Amazon SageMaker developer guide titled Use Amazon SageMaker Built-in Algorithms

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?

A) 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
B) 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
C) 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
D) 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.

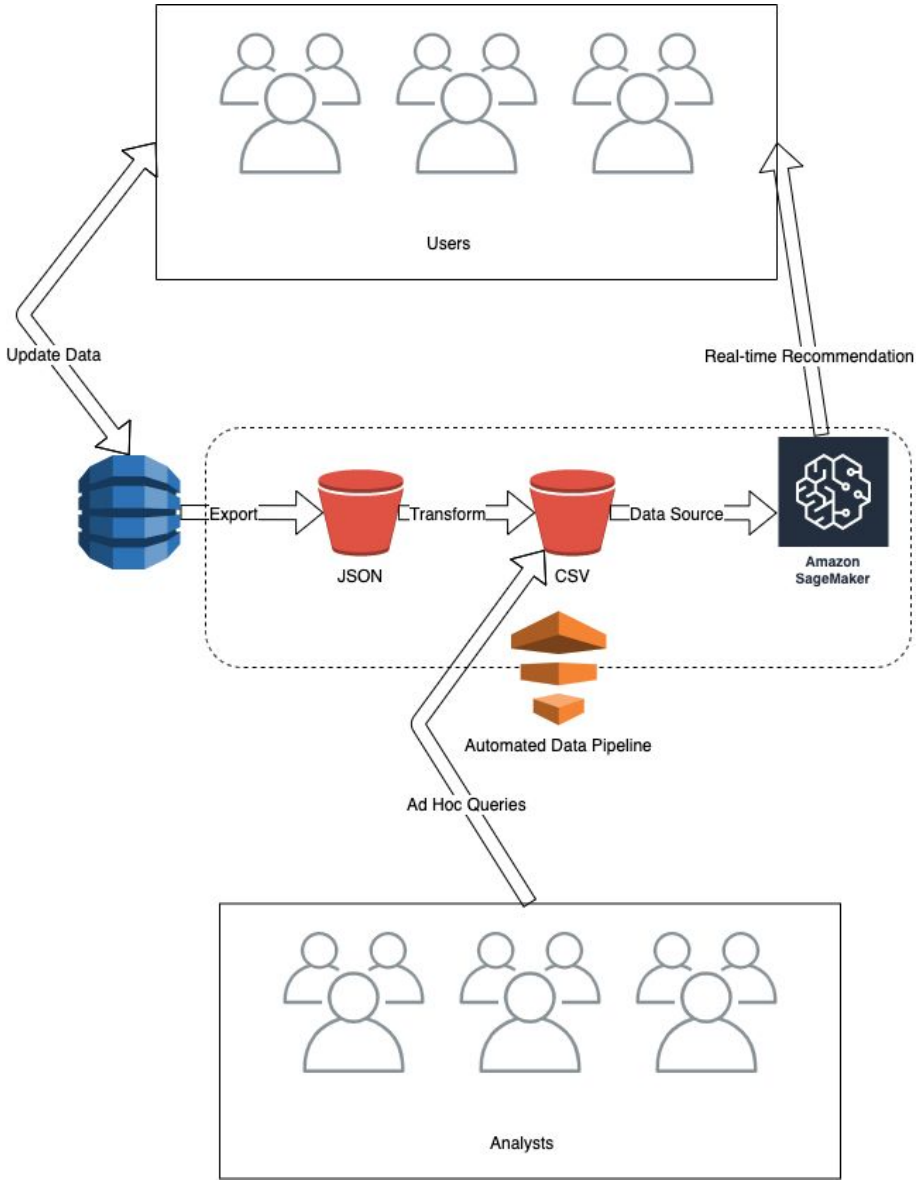
Option C is incorrect. AWS DMS does not support DynamoDB as a data source. Also, this approach lacks the pipeline coordination described in Option A.

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, and AWS Database Migration Service user guide titled How AWS Data Migration Service Works specifically the section on sources, the Amazon Kinesis Data Streams developer guide titled Amazon Kinesis Data Streams Terminology and Concepts

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?

A) 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
B) 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.
C) 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.
D) 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, and Amazon Machine Learning developer guide titled Creating an Amazon ML Datasource from Data in Amazon Redshift, the AWS Schema Conversion Tool user guide titled What is the AWS Schema Conversion Tool?, and the Cloud Data Migration Guide, specifically the section on AWS Snowball Edge, and the AWS Glue developer guide titled Adding Classifiers to a Crawler

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?

A) Orthogonal Sparse Bigram (OSB)
B) Term Frequency-Inverse Document Frequency (tf-idf)
C) Bag-of-Words
D) 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

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?

A) Orthogonal Sparse Bigram (OSB)
B) Term Frequency-Inverse Document Frequency (tf-idf)
C) Bag-of-Words
D) 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

Question: 43
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?

A) Orthogonal Sparse Bigram (OSB)
B) Term Frequency-Inverse Document Frequency (tf-idf)
C) Bag-of-Words
D) 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, and the article titled Document Classification Part 2: Text Processing (N-Gram Model & TF-IDF Model)