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