AWS Exam Readiness - Study Questions - Results

aws training and certification

You reached the end of these comprehensive study questions. Your overall score is a 70%.

The breakdown of how you did in each domain is below:
Domain 1: 83%
Domain 2: 70%
Domain 3: 61%
Domain 4: 78%

Below you can see how you scored on each question. **For detailed explanations of each answer as well as average class scores to compare yourself to, click the link below.** The page accessed from the link below will also have explanations for each of the quiz questions you encountered earlier in the course.

https://amazonmr.au1.qualtrics.com/reports/RC/public/YW1hem9ubXItNWQ0YWQzYTViZjczMDIwMDBmNTc4ZWY3LVVSX2RvRFVNUlZZS1NtQjUyWg==

30/43	69.8%

Compared to Average, my results are above average
https://amazonmr.au1.qualtrics.com/reports/RC/public/YW1hem9ubXItNWQ0YWQzYTViZjczMDIwMDBmNTc4ZWY3LVVSX2RvRFVNUlZZS1NtQjUyWg==

Class Scores on Comprehensive Study/Practice Questions

Below are the running averages for all students of the ML Exam Readiness course that have taken the comprehensive study/practice questions found at the end of the course. Averages are broken out by overall score and domain scores. You are free to compare your performance on these study questions to the below averages.

-	Value
Average Overall Score (%)	64.4
Average Domain 1 Score (%)	64.7
Average Domain 2 Score (%)	66.5
Average Domain 3 Score (%)	65.5
Average Domain 4 Score (%)	59.5

My results

A financial planning company is using the Amazon SageMaker endpoint with an Auto Scaling policy to serve its forecasting model to the company's customers to help them plan for retirement. The team wants to update the endpoint with its latest forecasting model, which has been trained using Amazon SageMaker training jobs. The team wants to do this without any downtime and with minimal change to the code.

0/1

What steps should the team take to update this endpoint?

Create a new endpoint using a new configuration with the latest model. Then, register the endpoint as a scalable target.

✕ **Use a new endpoint configuration with the latest model Amazon S3 path in the UpdateEndpoint API.**

De-register the endpoint as a scalable target. Update the endpoint using a new endpoint configuration with the latest model Amazon S3 path. Finally, register the endpoint as a scalable target again.

Update the endpoint using a new configuration with the latest model Amazon S3 path. Then, register the endpoint as a scalable target.

1. Use a new endpoint configuration with the latest model Amazon S3 path in the *UpdateEndpoint* API. **(Using a new endpoint configuration will not have Auto Scaling enabled.)**
2. De-register the endpoint as a scalable target. Update the endpoint using a new endpoint configuration with the latest model Amazon S3 path. Finally, register the endpoint as a scalable target again.*
3. Update the endpoint using a new configuration with the latest model Amazon S3 path. Then, register the endpoint as a scalable target. **(Before you can update the endpoint, you need to deregister the endpoint as a scalable target.)**
4. Create a new endpoint using a new configuration with the latest model. Then, register the endpoint as a scalable target. **(Creating a new endpoint will mean that they have to update the code every time the model changes, which isn't scalable.)**

A navigation and transportation company is using satellite images to model weather around the world in order to create optimal routes for its ships and planes. The company is using Amazon SageMaker training jobs to build and train its models.

However, during training, it takes too long to download the company's 100 GB data from Amazon S3 to the training instance before the training starts.

What should the company do to speed up its training jobs while keeping the costs low?

Create an Amazon EBS volume with the data on it and attach it to the training job

Increase the instance size for training

✓ **Change the input mode to Pipe**

Increase the batch size in the model

An ML Engineer at a real estate startup wants to use a new quantitative feature for an existing ML model that predicts housing prices. Before adding the feature to the cleaned dataset, the Engineer wants to visualize the feature in order to check for outliers and overall distribution and skewness of the feature.

What visualization technique should the ML Engineer use? (Select TWO.)

Scatterplot

Heatmap

✓ **Histogram**

✓ **Box Plot**

T-SNE

A Data Scientist wants to include "month" as a categorical column in a training dataset for an ML model that is being built. However, the ML algorithm gives an error when the column is added to the training data.

What should the Data Scientist do to add this column?

✗ **Map the "month" column data to the numbers 1 to 12 and use this new numerical mapped column.**

Scale the months using StandardScaler.

Use pandas fillna() to convert the column to numerical data.

Convert the "month" column to 12 different columns, one for each month, by using one-hot encoding.

1. Convert the "month" column to 12 different columns, one for each month, by using one-hot encoding.*
2. Map the "month" column data to the numbers 1 to 12 and use this new numerical mapped column. **(This is not a good option, because the numerical mapping of the months would imply magnitudes of difference between the different months (for instance, April could be looked at as twice as much as February).)**
3.

1. Scale the months using *StandardScaler*. (***StandardScaler*** **is used to scale numerical data. It will will not work with categorical data, which is what this question is about.)**
2. Use pandas *fillna()* to convert the column to numerical data. **(This approach, which deals with missing data, is not relevant to this question.)**

A financial organization uses multiple ML models to detect irregular patterns in its data to combat fraudulent activity such as money laundering. They use a TensorFlow-based Docker container on GPU-enabled Amazon EC2 instances to concurrently train the multiple models for this workload.

1/1

However, they want to automate the batch data preprocessing and ML training aspects of this pipeline, scheduling them to take place automatically every 24 hours.

What AWS service can they use to do this?

Kinesis Data Analytics

Amazon EMR

AWS Glue

✓ **AWS Batch**

A security and networking company wants to use ML to flag certain IP addresses that have been known to send spam and phishing information. The company wants to build an ML model based on previous user feedback indicating whether specific IP addresses have been connected to a website designed for spam and phishing.

0/1

What is the simplest solution that the company can implement?

Regression

✗ **Classification**

A rule-based solution should be used instead of ML

Natural language processing (NLP)

1. Regression **(Regression needs a historical dataset with a numerical output which fails the requirement of the use case in this question.)**
2. Classification **(Classification can work in this situation, but it's not the simplest solution.)**
3. Natural language processing (NLP) **(ML for natural language processing and text analytics is used to understand the meaning of text documents. It can be one part of the solution in this context, but it is not the simplest solution.)**
4. A rule-based solution should be used instead of ML*

A company is using its genomic data to classify how different human DNA affects cell growth, so that they can predict a person's chances of getting cancer. Before creating and preparing the training and validation datasets for the model, the company wants to reduce the high dimensionality of the data.

What technique should the company use to achieve this goal? (Select TWO.)

Use L2 regularization to reduce the features used in the data. Visualize the data using matplotlib.

✓ **Use T-SNE to reduce the dimensionality of the data. Visualize the data using matplotlib.**

Calculate the eigenvectors. Use a scatter matrix to choose the best features.

Use seaborn distribution plot (distplot) to visualize the correlated data. Remove the unrelated features.

✓ **Use Principle Component Analysis (PCA) to reduce the dimensionality of the data. Visualize the data using matplotlib.**

An analytics company wants to use a fully managed service that automatically scales to handle the transfer of its Apache web logs, syslogs, text and videos on their webserver to Amazon S3 with minimum transformation.

What service can be used for this process?

✓ **Kinesis Firehose**

Kinesis Video Streams

Kinesis Data Analytics

Kinesis Data Streams

An ad tech company is using an XGBoost model to classify its clickstream data. The company's Data Scientist is asked to explain how the model works to a group of non-technical colleagues.

1/1

What is a simple explanation the Data Scientist can provide?

XGBoost is a robust, flexible, scalable algorithm that uses logistic regression to classify data into bucket

XGBoost is an efficient and scalable neural network architecture

✓ **XGBoost is an Extreme Gradient Boosting algorithm that is optimized for boosted decision trees**

XGBoost is a state-of-the-art algorithm that uses logistic regression to split each feature of the data based on certain conditions

---

A real estate startup wants to use ML to predict the value of homes in various cities. To do so, the startup's data science team is joining real estate price data with other variables such as weather, demographic, and standard of living data.

0/1

However, the team is having problems with slow model convergence. Additionally, the model includes large weights for some features, which is causing degradation in model performance.

What kind of data preprocessing technique should the team use to more effectively prepare this data?

One hot encoder

Max absolute scaler

✗ **Normalizer**

Standard scaler

---

1. Standard scaler* **(Standard scaler is the best option, because it performs scaling and shifting/centering.)**
2. Normalizer **(This would perform row normalization. This situation requires *column* normalization.)**
3. Max absolute scaler **(This would scale each column by its max value, but would not shift/center the data.)**
4. One hot encoder **(There is no symbolic/string data being mentioned here to perform 1-hot encoding on.)**

A Data Scientist is using stochastic gradient descent (SGD) as the gradient optimizer to train a machine learning model. However, the model training error is taking longer to converge to the optimal solution than desired.

1/3

What optimizer can the Data Scientist use to improve training performance? (Select THREE.)

✓ **Adam**

RMSProp

✗ **Mini-batch gradient descent**

✗ **Gradient Descent**

Xavier

Adagrad

1. Adam* (**Adam stands for adaptive momentum which can help the model converge faster and get out of being stuck in local minima.**)
2. Adagrad* (**Adagrad is an algorithm for gradient-based optimization that adapts the learning rate to the parameters by performing smaller updates and, in turn, helps with convergence.**)
3. Gradient Descent (**Gradient descent will take longer to converge than SGD does since it needs the whole dataset for every step calculation.**)
4. RMSProp* (**RMSProp uses a moving average of squared gradients to normalize the gradient itself, which helps with faster convergence.**)
5. Mini-batch gradient descent (**Mini batch gradient descent will suffer from some of the same problems as SGD.**)
6. Xavier (**Xavier is an initialization technique and not an optimization technique.**)

A Machine Learning Engineer is creating a regression model for forecasting company revenue based on an internal dataset made up of past sales and other related data.   **1/1**

What metric should the Engineer use to evaluate the ML model?

✓ **Root Mean squared error (RMSE)**

Precision

Sigmoid

Cross-entropy log loss

A log analytics company wants to provide a history of Amazon SageMaker API calls made on its client's account for security analysis and operational troubleshooting purposes.   **1/1**

What must be done in the client's account to ensure that the company can analyze the API calls?

Use the Amazon SageMaker SDK to call the 'sagemaker_history()' function.

Enable CloudWatch logs.

Use IAM roles. "logs:*" are added to those IAM roles.

✓ **Enable AWS CloudTrail.**

A multi-national banking organization provides loan services to customers worldwide. Many of its customers still submit loan applications in paper form in one of the bank's branch locations. The bank wants to speed up the loan approval process for this set of customers by using machine learning. More specifically, it wants to create a process in which customers submit the application to the clerk, who scans and uploads it to the system. The system then reads and provides an approval or denial of the application in a matter of minutes.

1/1

What can the bank use to read and extract the necessary data from the loan applications without needing to manage the process?

A custom CNN model

✓ **Amazon Textract**

An LSTM model

Amazon Personalize

An ML scientist has built a decision tree model using scikit-learn with 1,000 trees. The training accuracy for the model was 99.2% and the test accuracy was 70.3%.

1/1

Should the Scientist use this model in production?

Yes, because it is not generalizing well on the test set

Yes, because it is generalizing well on the training set

✓ **No, because it is not generalizing well on the test set**

No, because it is generalizing well on the training set

A data and analytics company is expanding its platform on AWS. The company wants to build a serverless product that preprocesses large structured data, while minimizing the cost for data storage and compute. The company also wants to integrate the new product with an existing ML product that uses Amazon EMR with Spark.

1/1

What solution should the company use to build this new product?

Use AWS Lambda for data preprocessing. Save the data in Amazon S3 in Parquet format.

Use AWS Lambda for data preprocessing. Save the data in Amazon S3 in CSV format.

Use AWS Glue for data preprocessing. Save the data in Amazon S3 in CSV format.

✓ **Use AWS Glue for data preprocessing. Save the data in Amazon S3 in Parquet format.**

A video streaming company wants to create a searchable video library that provides a personalized searching experience and automated content moderation for its users, so that when the users search for a keyword, they get all the videos that map to that keyword. The company wants to do this with minimal cost and limited need for management.

1/1

What approach should the company take to building this solution?

✓ **Use Amazon Rekognition Video to extract metadata from the videos**

Use AWS Batch to transform a batch of video files into metadata

Use Amazon SageMaker to create an ML model that extracts metadata from the videos

Use Amazon Kinesis Video Streams to stream the videos to Amazon EMR in order to create an ML model

A Data Scientist for a credit card company is creating a solution to predict credit card fraud at the time of transaction. To that end, the Data Scientist is looking to create an ML model to predict fraud and will do so by training that model on an existing dataset of credit card transactions. That dataset contains 1,000 examples of transactions in total, only 50 of which are labeled as fraud.

How should the Data Scientist deal with this class imbalance?

Undersample the non-fraudulent records to improve the class imbalance

Use K-fold cross validation when training the mode

✓ **Use the Synthetic Minority Oversampling Technique (SMOTE) to oversample the fraud records**

Drop all the fraud examples, and use a One-Class SVM to classify

---

An online news organization wants to expand its reach globally by translating some of its most commonly read articles into different languages using ML. The organization's data science team is gathering all the news articles that they have published in both English and at least one other language. They want to use this data to create one machine learning model for each non-English language that the organization is targeting. The models should only require minimum management.

What approach should the team use to building these models?

Use Amazon SageMaker Object2Vec to create a vector. Use Amazon EC2 instances with the Deep Learning Amazon Machine Image (AMI) to create a language encoder-decoder model

✓ **Use Amazon SageMaker Object2Vec to create a vector. Use the Amazon SageMaker built-in Sequence to Sequence model (Seq2Seq)**

Use Amazon SageMaker Object2Vec to create a vector. Use the SockEye model in Amazon SageMaker using Building Your Own Containers (BYOC)

Use Amazon SageMaker Object2Vec to create a vector. Then use a Long Short-term Memory (LSTM) model using Building Your Own Containers (BYOC)

A Data Scientist wants to tune the hyperparameters of a machine learning model to improve the model's F1 score.

What technique can be used to achieve this desired outcome on Amazon SageMaker? (Select TWO.)

✗ **Grid Search**

Bayesian optimization

✓ **Random Search**

Depth first search

Breadth First Search

A logistics company is tracking the arrival and departure of their ships at various ports around the world using image recognition. They have collected 2.3 TBs of high quality images for training their convolutional neural network (CNN) model. This training data will also be used concurrently by other data scientists in the company for additional data science projects. They want the solution to provide high throughput and consistent low latency to process the ML training datasets.

What service should the company use to store this training data?

Amazon S3

Amazon EBS

✓ **Amazon FSx for Lustre**

Amazon Kinesis

What factors lead to the wide adoption of neural networks in the last decade? (Select THREE.)

Efficient algorithms

✓ **An orders of magnitude increase in data collected**

✓ **Cheaper GPUs**

✗ **Wide adoption of cloud-based services**

Cheaper CPUs

1. Efficient algorithms*
2. Cheaper GPUs*
3. An orders of magnitude increase in data collected* **(Over the last two decades, the amount of available data of all sorts and the power of our data storing and processing machines (GPUs) have exponentially increased. Combined with better computing capabilities, the massive increases in the amount of available data to train models have allowed the creation of larger, deeper neural networks, which just perform better than smaller ones.)**
4. Cheaper CPUs **(*GPUs* are needed to train neural networks efficiently, so cheaper CPUs don't have much to do with the wide adoption of neural networks in the last decade.)**
5. Wide adoption of cloud-based services **(While cloud-based services made it easy for everyone to do machine learning, they are based on the actual factors like efficient algorithms and cheaper GPUs)**

A Data Scientist at an ad-tech startup wants to update an ML model that uses an Amazon SageMaker endpoint using the canary deployment methodology, in which the production variant 1 is the production model and the production variant 2 is the updated model.

How can the Data Scientist *efficiently* configure this endpoint configuration to deploy the two different versions of the model while monitoring the Amazon CloudWatch invocations?

Create two Amazon SageMaker endpoints and change the endpoint URL after testing the new endpoint.

Create an endpoint configuration with production variants for the two models with equal weights.

Create an endpoint configuration with production variants for the two models with a weight ratio of 10:90.

✓ **Create an endpoint configuration with production variants for the two models with a weight ratio of 0:1. Update the weights periodically.**

A Data Scientist wants to use the Amazon SageMaker hyperparameter tuning job to automatically tune a random forest model.

What API does the Amazon SageMaker SDK use to create and interact with the Amazon SageMaker hyperparameter tuning jobs?

✗ **HyperparameterTunerJob()**

Hyperparameter()

HyperparameterTuner()

HyperparameterTuningJobs()

A Data Scientist at a credit card company trained a classification model to predict fraud at the time of a transaction. The Data Scientist used a confusion matrix to evaluate the performance of the model.

	True Positive	True Negative
Predicted Positive	100	90
Predicted Negative	25	250

Using the confusion matrix below, determine the percent of positive records that were classified correctly. Choose the answer that also labels this evaluation metric correctly.

80%; Precision

✓ **80%; Recall**

52.6%; Recall

52.6%; Precision

A Machine Learning Engineer wants to use Amazon SageMaker and the built-in XGBoost algorithm for model training. The training data is currently stored in CSV format, with the first 10 columns representing features and the 11th column representing the target label.

What should the ML Engineer do to prepare the data for use in an Amazon SageMaker training job?

The data should be split into training, validation, and test sets. The datasets should then be uploaded to Amazon S3.

✓ **The target label should be changed to the first column. The data should be split into training, validation, and test sets. Finally, the datasets should be uploaded to Amazon S3.**

The target label should be changed to the first column. The dataset should then be uploaded to Amazon S3. Finally, Amazon SageMaker can be used to split the data into training, validation, and test sets.

The dataset should be uploaded directly to Amazon S3. Amazon SageMaker can then be used to split the data into training, validation, and test sets.

A Data Scientist at a retail company is using Amazon SageMaker to classify social media posts that mention the company into one of two categories: Posts that require a response from the company, and posts that do not. The Data Scientist is using a training dataset of 10,000 posts, which contains the timestamp, author, and full text of each post.

However, the Data Scientist is missing the target labels that are required for training.

Which approach can the Data Scientist take to create valid target label data? (Select TWO)

Use the a priori probability distribution of the two classes. Then, use Monte-Carlo simulation to generate the labels

✓ **Use Amazon Mechanical Turk to publish Human Intelligence Tasks that ask Turk workers to label the posts**

✓ **Ask the social media handling team to review each post using Amazon SageMaker GroundTruth and provide the label**

Use the sentiment analysis natural language processing library to determine whether a post requires a response

Use K-Means to cluster posts into various groups, and pick the most frequent word in each group as its label

An advertising and analytics company uses machine learning to predict user response to online advertisements using a custom XGBoost model. The company wants to improve its ML pipeline by porting its training and inference code, written in R, to Amazon SageMaker, and do so with minimal changes to the existing code.

How should the company set up this new pipeline?

Use Amazon in-built algorithms to run their training and inference jobs.

Create a new Amazon SageMaker notebook instance. Copy the existing code into an Amazon SageMaker notebook. Then, run the pipeline from this notebook.

✓ **Use the Build Your Own Container (BYOC) Amazon SageMaker option. Create a new Docker container with the existing code. Register the container in Amazon Elastic Container Registry (ECR). Finally, run the training and inference jobs using this container.**

Use the Amazon pre-built R container option and port the existing code over to the container. Register the container in Amazon Elastic Container Registry (Amazon ECR). Finally, run the training and inference jobs using this container.

A Data Scientist wants to create a linear regression model to train on a housing dataset to predict home prices. As part of that process, the Data Scientist created a correlation matrix between the dataset's features and the target variable. The correlations between the target and two of the features, feature 3 and feature 7, are 0.64 and -0.85, respectively.

Which feature has a stronger correlation with the target variable?

Feature 7 and feature 3 both have weak correlations to the target

✗ **Feature 7**

Feature 3

There is not sufficient enough data to determine which variable has a stronger correlation to the target

---

1. Feature 3*
2. Feature 7 (**Feature 7 has a negative correlation with the target variable, and even though the magnitude is higher than the correlation with Feature 3, the question asks for a *stronger* correlation, which is associated with a positive correlation.**)
3. There is not sufficient enough data to determine which variable has a stronger correlation to the target (**There is sufficient data. Even though the magnitude of the correlation with Feature 7 is higher, "stronger" is associated with a positive correlation.**)
4. Feature 7 and feature 3 both have weak correlations to the target (**This is not true, as 0.64 is not considered very weak.**)

A ride-share company wants to create intelligent conversational chatbots that will serve as first responders to customers who call to report an issue with their ride. The company wants these chatbot-customer calls to mimic natural conversations that provide personalized experiences for the customers.

1/1

What combination of AWS services can the company use to create this workflow without a lot of ongoing management?

✓ **Amazon Lex to parse the utterances and intent of customer comments, Amazon Polly to reply to the customers**

Amazon Polly to parse the utterances and intent of customer comments, Amazon Lex to reply to the customers

Amazon Transcribe to parse the utterances and intent of customer comments, Amazon Polly to reply to the customers

Amazon Transcribe to parse the utterances and intent of customer comments, Amazon Lex to reply to the customers

---

A video streaming company is looking to create a personalized experience for its customers on its platform. The company wants to provide recommended videos to stream based on what other similar users watched previously. To this end, it is collecting its platform's clickstream data using an ETL pipeline and storing the logs and syslogs in Amazon S3.

0/1

What kind of algorithm should the company use to create the simplest solution in this situation?

Regression

Recommender system

✗ **Classification**

Reinforcement learning

1. Regression **(Regression will not provide personalized recommendations to customer, because regression can either predict a number or classify based on historical data.)**
2. Classification **(Classification cannot deliver a personalized recommendation for every user.)**
3. Recommender system* **(A *recommendation system* is a subclass of an information filtering system that seeks to predict the "rating" or "preference" a user would give to an item. It is ideal for the situation in this question.)**
4. Reinforcement learning **(Reinforcement learning is a relatively new field and while there are solutions that use it, it is not the simplest solution, since we already have historical data.)**

A video streaming company wants to analyze its VPC flow logs to build a real-time anomaly detection pipeline. The pipeline must be minimally managed and enable the business to build a near real-time dashboard.

0/1

What combination of AWS service and algorithm can the company use for this pipeline?

Amazon SageMaker with RandomCutForest

Kinesis Data Analytics with RandomCutForest

Apache Spark on Amazon EMR with MLLib

✗ **Amazon QuickSight with ML Insights**

1. Amazon SageMaker with RandomCutForest **(Amazon SageMaker cannot be used with streamed data.)**
2. Kinesis Data Analytics with RandomCutForest*
3. Amazon QuickSight with ML Insights **(Amazon QuickSight can only be used with structured datasets that need to be stored in Amazon S3 or a database.)**
4. Apache Spark on Amazon EMR with MLLib **(Amazon EMR can be used for this use case, but it doesn't satisfy the minimally managed requirement.)**

A team of Data Scientists wants to use Amazon SageMaker training jobs to run two different versions of the same model in parallel to compare the long-term effectiveness of the different versions in reaching the related business outcome. **1/1**

How should the team deploy these two model versions with minimum management?

  Create an endpoint configuration with production variants for the two models with a weight ratio of 90:10.

✓ **Create an endpoint configuration with production variants for the two models with equal weights.**

  Create a Lambda function that downloads the models from Amazon S3 and calculates and returns the predictions of the two models.

  Create a Lambda function that preprocesses the incoming data, calls the two Amazon SageMaker endpoints for the two models, and finally returns the prediction.

A retail company currently relies heavily on the data it collects to derive important business insights. Its current pipeline consists of a nightly batch extract, transform, and load (ETL) of its data extracted from various sources like company web servers and third party sources, on an Amazon EMR instance. The processed data is then stored in Amazon S3. **1/1**

The company, however, is finding it increasingly difficult to manage this pipeline, because they don't have enough Data Engineers on staff.

What AWS service could the company use to help solve this issue?

  Amazon Athena

  AWS Lambda

✓ **AWS Glue**

  Amazon EC2

A healthcare organization has an application that takes in sensitive user data. This data is encrypted at rest and stored in an Amazon S3 bucket using customer-managed encryption with AWS Key Management Service (AWS KMS). A Data Scientist in the organization wants to use this encrypted data as features in a Amazon SageMaker training job. However, the following error continues to occur: "Data download failed." **0/1**

What should the Data Scientist do to fix this issue?

  Make sure the AWS Identity and Access Management (IAM) role used for Amazon S3 access has permissions to encrypt and decrypt the data with the AWS KMS key.

  Specify the "VolumeKmsKeyId" in the Amazon SageMaker training job.

  Add "EnableKMS" to the Amazon SageMaker training job. Then, specify the Amazon S3 bucket that includes the data.

✗ **Add "S3:*" to the IAM role that is attached to the Amazon SageMaker training job.**

What should the Data Scientist do to fix this issue?

1. Make sure the AWS Identity and Access Management (IAM) role used for Amazon S3 access has permissions to encrypt and decrypt the data with the AWS KMS key. *
2. Add "S3:*" to the IAM role that is attached to the Amazon SageMaker training job. **(Adding "S3:*" wildcard is not a good security practice. It will not read encrypted AWS KMS data into Amazon SageMaker.)**
3. Specify the "VolumeKmsKeyId" in the Amazon SageMaker training job. **(VolumeKmsKeyId helps in encrypting data on the training job instance storage, not on Amazon S3.)**
4. Add "EnableKMS" to the Amazon SageMaker training job. Then, specify the Amazon S3 bucket that includes the data. **(There is no option to enable AWS KMS that will read at-rest data in an encrypted Amazon S3 bucket.)**