training and certification

AWS Certified Machine Learning – Specialty (MLS-C01) Exam Guide

Introduction

The AWS Certified Machine Learning – Specialty (MLS-C01) exam is intended for individuals who perform an artificial intelligence/machine learning (AI/ML) development or data science role. The exam validates a candidate's ability to design, build, deploy, optimize, train, tune, and maintain ML solutions for given business problems by using the AWS Cloud.

The exam also validates a candidate's ability to complete the following tasks:

- Select and justify the appropriate ML approach for a given business problem
- Identify appropriate AWS services to implement ML solutions
- Design and implement scalable, cost-optimized, reliable, and secure ML solutions

Target candidate description

The target candidate is expected to have 2 or more years of hands-on experience developing, architecting, and running ML or deep learning workloads in the AWS Cloud.

Recommended AWS knowledge

The target candidate should have the following knowledge:

- The ability to express the intuition behind basic ML algorithms
- Experience performing basic hyperparameter optimization
- Experience with ML and deep learning frameworks
- The ability to follow model-training best practices
- The ability to follow deployment best practices
- The ability to follow operational best practices

What is considered out of scope for the target candidate?

The following is a non-exhaustive list of related job tasks that the target candidate is not expected to be able to perform. These items are considered out of scope for the exam:

- Extensive or complex algorithm development
- Extensive hyperparameter optimization
- Complex mathematical proofs and computations
- Advanced networking and network design
- Advanced database, security, and DevOps concepts
- DevOps-related tasks for Amazon EMR

For a detailed list of specific tools and technologies that might be covered on the exam, as well as lists of in-scope and out-of-scope AWS services, refer to the Appendix.

Exam content

Response types

There are two types of questions on the exam:

- **Multiple choice:** Has one correct response and three incorrect responses (distractors)
- **Multiple response:** Has two or more correct responses out of five or more response options

Select one or more responses that best complete the statement or answer the question. Distractors, or incorrect answers, are response options that a candidate with incomplete knowledge or skill might choose. Distractors are generally plausible responses that match the content area.

Unanswered questions are scored as incorrect; there is no penalty for guessing. The exam includes 50 questions that will affect your score.

Unscored content

The exam includes 15 unscored questions that do not affect your score. AWS collects information about candidate performance on these unscored questions to evaluate these questions for future use as scored questions. These unscored questions are not identified on the exam.

Exam results

The AWS Certified Machine Learning – Specialty (MLS-C01) exam is a pass or fail exam. The exam is scored against a minimum standard established by AWS professionals who follow certification industry best practices and guidelines.

Your results for the exam are reported as a scaled score of 100–1,000. The minimum passing score is 750. Your score shows how you performed on the exam as a whole and whether or not you passed. Scaled scoring models help equate scores across multiple exam forms that might have slightly different difficulty levels.

Your score report could contain a table of classifications of your performance at each section level. This information is intended to provide general feedback about your exam performance. The exam uses a compensatory scoring model, which means that you do not need to achieve a passing score in each section. You need to pass only the overall exam.

Each section of the exam has a specific weighting, so some sections have more questions than other sections have. The table contains general information that highlights your strengths and weaknesses. Use caution when interpreting section-level feedback.

Content outline

This exam guide includes weightings, test domains, and objectives for the exam. It is not a comprehensive listing of the content on the exam. However, additional context for each of the objectives is available to help guide your preparation for the exam. The following table lists the main content domains and their weightings. The table precedes the complete exam content outline, which includes the additional context. The percentage in each domain represents only scored content.

Domain	% of Exam
Domain 1: Data Engineering	20%
Domain 2: Exploratory Data Analysis	24%
Domain 3: Modeling	36%
Domain 4: Machine Learning Implementation and Operations	20%
TOTAL	**100%**

Domain 1: Data Engineering

1.1 Create data repositories for machine learning.

- Identify data sources (e.g., content and location, primary sources such as user data)
- Determine storage mediums (e.g., DB, Data Lake, S3, EFS, EBS)

1.2 Identify and implement a data ingestion solution.

- Data job styles/types (batch load, streaming)
- Data ingestion pipelines (Batch-based ML workloads and streaming-based ML workloads)
    - Kinesis
    - Kinesis Analytics
    - Kinesis Firehose
    - EMR
    - Glue
- Job scheduling

1.3 Identify and implement a data transformation solution.

- Transforming data transit (ETL: Glue, EMR, AWS Batch)
- Handle ML-specific data using map reduce (Hadoop, Spark, Hive)

Domain 2: Exploratory Data Analysis

2.1 Sanitize and prepare data for modeling.

- Identify and handle missing data, corrupt data, stop words, etc.
- Formatting, normalizing, augmenting, and scaling data
- Labeled data (recognizing when you have enough labeled data and identifying mitigation strategies [Data labeling tools (Mechanical Turk, manual labor)])

2.2 Perform feature engineering.

- Identify and extract features from data sets, including from data sources such as text, speech, image, public datasets, etc.
- Analyze/evaluate feature engineering concepts (binning, tokenization, outliers, synthetic features, 1 hot encoding, reducing dimensionality of data)

2.3 Analyze and visualize data for machine learning.

- Graphing (scatter plot, time series, histogram, box plot)
- Interpreting descriptive statistics (correlation, summary statistics, p value)
- Clustering (hierarchical, diagnosing, elbow plot, cluster size)

Domain 3: Modeling

3.1 Frame business problems as machine learning problems.

- Determine when to use/when not to use ML
- Know the difference between supervised and unsupervised learning
- Selecting from among classification, regression, forecasting, clustering, recommendation, etc.

3.2 Select the appropriate model(s) for a given machine learning problem.

- Xgboost, logistic regression, K-means, linear regression, decision trees, random forests, RNN, CNN, Ensemble, Transfer learning
- Express intuition behind models

3.3 Train machine learning models.

- Train validation test split, cross-validation
- Optimizer, gradient descent, loss functions, local minima, convergence, batches, probability, etc.
- Compute choice (GPU vs. CPU, distributed vs. non-distributed, platform [Spark vs. non-Spark])
- Model updates and retraining
  - Batch vs. real-time/online

3.4 Perform hyperparameter optimization.

- Regularization
  - Drop out
  - L1/L2
- Cross validation
- Model initialization
- Neural network architecture (layers/nodes), learning rate, activation functions
- Tree-based models (# of trees, # of levels)
- Linear models (learning rate)

3.5 Evaluate machine learning models.

- Avoid overfitting/underfitting (detect and handle bias and variance)
- Metrics (AUC-ROC, accuracy, precision, recall, RMSE, F1 score)
- Confusion matrix
- Offline and online model evaluation, A/B testing
- Compare models using metrics (time to train a model, quality of model, engineering costs)
- Cross validation

Domain 4: Machine Learning Implementation and Operations

4.1 Build machine learning solutions for performance, availability, scalability, resiliency, and fault tolerance.

- AWS environment logging and monitoring
  - CloudTrail and CloudWatch
  - Build error monitoring
- Multiple regions, Multiple AZs
- AMI/golden image
- Docker containers
- Auto Scaling groups
- Rightsizing
  - Instances
  - Provisioned IOPS
  - Volumes
- Load balancing
- AWS best practices

4.2 Recommend and implement the appropriate machine learning services and features for a given problem.

- ML on AWS (application services)
  - Poly
  - Lex
  - Transcribe
- AWS service limits
- Build your own model vs. SageMaker built-in algorithms
- Infrastructure: (spot, instance types), cost considerations
  - Using spot instances to train deep learning models using AWS Batch

4.3 Apply basic AWS security practices to machine learning solutions.

- IAM
- S3 bucket policies
- Security groups
- VPC
- Encryption/anonymization

4.4 Deploy and operationalize machine learning solutions.

- Exposing endpoints and interacting with them
- ML model versioning
- A/B testing
- Retrain pipelines
- ML debugging/troubleshooting
  - Detect and mitigate drop in performance
  - Monitor performance of the model

Appendix

Which key tools, technologies, and concepts might be covered on the exam?

The following is a non-exhaustive list of the tools and technologies that could appear on the exam. This list is subject to change and is provided to help you understand the general scope of services, features, or technologies on the exam. The general tools and technologies in this list appear in no particular order. AWS services are grouped according to their primary functions. While some of these technologies will likely be covered more than others on the exam, the order and placement of them in this list is no indication of relative weight or importance:

- Ingestion/Collection
- Processing/ETL
- Data analysis/visualization
- Model training
- Model deployment/inference
- Operational
- AWS ML application services
- Language relevant to ML (for example, Python, Java, Scala, R, SQL)
- Notebooks and integrated development environments (IDEs)

AWS services and features

Analytics:

- Amazon Athena
- Amazon EMR
- Amazon Kinesis Data Analytics
- Amazon Kinesis Data Firehose
- Amazon Kinesis Data Streams
- Amazon QuickSight

Compute:

- AWS Batch
- Amazon EC2

Containers:

- Amazon Elastic Container Registry (Amazon ECR)
- Amazon Elastic Container Service (Amazon ECS)
- Amazon Elastic Kubernetes Service (Amazon EKS)

Database:

- AWS Glue
- Amazon Redshift

Internet of Things (IoT):

- AWS IoT Greengrass

Machine Learning:

- Amazon Comprehend
- AWS Deep Learning AMIs (DLAMI)
- AWS DeepLens
- Amazon Forecast
- Amazon Fraud Detector
- Amazon Lex
- Amazon Polly
- Amazon Rekognition
- Amazon SageMaker
- Amazon Textract
- Amazon Transcribe
- Amazon Translate

Management and Governance:

- AWS CloudTrail
- Amazon CloudWatch

Networking and Content Delivery:

- Amazon VPC

Security, Identity, and Compliance:

- AWS Identity and Access Management (IAM)

Serverless:

- AWS Fargate
- AWS Lambda

Storage:

- Amazon Elastic File System (Amazon EFS)
- Amazon FSx
- Amazon S3

Out-of-scope AWS services and features

The following is a non-exhaustive list of AWS services and features that are not covered on the exam. These services and features do not represent every AWS offering that is excluded from the exam content. Services or features that are entirely unrelated to the target job roles for the exam are excluded from this list because they are assumed to be irrelevant.

Out-of-scope AWS services and features include the following:

- AWS Data Pipeline
- AWS DeepRacer
- Amazon Machine Learning (Amazon ML)

aws training and certification

Machine Learning – Specialty (MLS-C01)
Sample Exam Questions

1) A machine learning team has several large CSV datasets in Amazon S3. Historically, models built with the Amazon SageMaker Linear Learner algorithm have taken hours to train on similar-sized datasets. The team's leaders need to accelerate the training process.

What can a machine learning specialist do to address this concern?

A) Use Amazon SageMaker Pipe mode.
B) Use Amazon Machine Learning to train the models.
C) Use Amazon Kinesis to stream the data to Amazon SageMaker.
D) Use AWS Glue to transform the CSV dataset to the JSON format.

2) A term frequency–inverse document frequency (tf–idf) matrix using both unigrams and bigrams is built from a text corpus consisting of the following two sentences:

1. Please call the number below.
2. Please do not call us.

What are the dimensions of the tf–idf matrix?

A) (2, 16)
B) (2, 8)
C) (2, 10)
D) (8, 10)

3) A company is setting up a system to manage all of the datasets it stores in Amazon S3. The company would like to automate running transformation jobs on the data and maintaining a catalog of the metadata concerning the datasets. The solution should require the least amount of setup and maintenance.

Which solution will allow the company to achieve its goals?

A) Create an Amazon EMR cluster with Apache Hive installed. Then, create a Hive metastore and a script to run transformation jobs on a schedule.
B) Create an AWS Glue crawler to populate the AWS Glue Data Catalog. Then, author an AWS Glue ETL job, and set up a schedule for data transformation jobs.
C) Create an Amazon EMR cluster with Apache Spark installed. Then, create an Apache Hive metastore and a script to run transformation jobs on a schedule.
D) Create an AWS Data Pipeline that transforms the data. Then, create an Apache Hive metastore and a script to run transformation jobs on a schedule.

4) A data scientist is working on optimizing a model during the training process by varying multiple parameters. The data scientist observes that, during multiple runs with identical parameters, the loss function converges to different, yet stable, values.

What should the data scientist do to improve the training process?

A) Increase the learning rate. Keep the batch size the same.
B) Reduce the batch size. Decrease the learning rate.
C) Keep the batch size the same. Decrease the learning rate.
D) Do not change the learning rate. Increase the batch size.

5) A data scientist is evaluating different binary classification models. A false positive result is 5 times more expensive (from a business perspective) than a false negative result.

The models should be evaluated based on the following criteria:

1) **Must have a recall rate of at least 80%**
2) **Must have a false positive rate of 10% or less**
3) **Must minimize business costs**

After creating each binary classification model, the data scientist generates the corresponding confusion matrix.

Which confusion matrix represents the model that satisfies the requirements?

A) TN = 91, FP = 9
FN = 22, TP = 78
B) TN = 99, FP = 1
FN = 21, TP = 79
C) TN = 96, FP = 4
FN = 10, TP = 90
D) TN = 98, FP = 2
FN = 18, TP = 82

6) A data scientist uses logistic regression to build a fraud detection model. While the model accuracy is 99%, 90% of the fraud cases are not detected by the model.

What action will definitively help the model detect more than 10% of fraud cases?

A) Using undersampling to balance the dataset
B) Decreasing the class probability threshold
C) Using regularization to reduce overfitting
D) Using oversampling to balance the dataset

7) A company is interested in building a fraud detection model. Currently, the data scientist does not have a sufficient amount of information due to the low number of fraud cases.

Which method is MOST likely to detect the GREATEST number of valid fraud cases?

A) Oversampling using bootstrapping
B) Undersampling
C) Oversampling using SMOTE
D) Class weight adjustment

8) A machine learning engineer is preparing a data frame for a supervised learning task with the Amazon SageMaker Linear Learner algorithm. The ML engineer notices the target label classes are highly imbalanced and multiple feature columns contain missing values. The proportion of missing values across the entire data frame is less than 5%.

What should the ML engineer do to minimize bias due to missing values?

A) Replace each missing value by the mean or median across non-missing values in same row.
B) Delete observations that contain missing values because these represent less than 5% of the data.
C) Replace each missing value by the mean or median across non-missing values in the same column.
D) For each feature, approximate the missing values using supervised learning based on other features.

9) A company has collected customer comments on its products, rating them as safe or unsafe, using decision trees. The training dataset has the following features: id, date, full review, full review summary, and a binary safe/unsafe tag. During training, any data sample with missing features was dropped. In a few instances, the test set was found to be missing the full review text field.

For this use case, which is the most effective course of action to address test data samples with missing features?

A) Drop the test samples with missing full review text fields, and then run through the test set.
B) Copy the summary text fields and use them to fill in the missing full review text fields, and then run through the test set.
C) Use an algorithm that handles missing data better than decision trees.
D) Generate synthetic data to fill in the fields that are missing data, and then run through the test set.

10) An insurance company needs to automate claim compliance reviews because human reviews are expensive and error-prone. The company has a large set of claims and a compliance label for each. Each claim consists of a few sentences in English, many of which contain complex related information. Management would like to use Amazon SageMaker built-in algorithms to design a machine learning supervised model that can be trained to read each claim and predict if the claim is compliant or not.

Which approach should be used to extract features from the claims to be used as inputs for the downstream supervised task?

A) Derive a dictionary of tokens from claims in the entire dataset. Apply one-hot encoding to tokens found in each claim of the training set. Send the derived features space as inputs to an Amazon SageMaker built-in supervised learning algorithm.
B) Apply Amazon SageMaker BlazingText in Word2Vec mode to claims in the training set. Send the derived features space as inputs for the downstream supervised task.
C) Apply Amazon SageMaker BlazingText in classification mode to labeled claims in the training set to derive features for the claims that correspond to the compliant and non-compliant labels, respectively.
D) Apply Amazon SageMaker Object2Vec to claims in the training set. Send the derived features space as inputs for the downstream supervised task.

Answers

1) A – Amazon SageMaker Pipe mode streams the data directly to the container, which improves the performance of training jobs. (Refer to this link for supporting information.) In Pipe mode, your training job streams data directly from Amazon S3. Streaming can provide faster start times for training jobs and better throughput. With Pipe mode, you also reduce the size of the Amazon EBS volumes for your training instances. B would not apply in this scenario. C is a streaming ingestion solution, but is not applicable in this scenario. D transforms the data structure.

2) A – There are 2 sentences, 8 unique unigrams, and 8 unique bigrams, so the result would be (2,16). The phrases are "Please call the number below" and "Please do not call us." Each word individually (unigram) is "Please," "call," "the," "number," "below," "do," "not," and "us." The unique bigrams are "Please call," "call the," "the number," "number below," "Please do," "do not," "not call," and "call us." The tf–idf vectorizer is described at this link.

3) B – AWS Glue is the correct answer because this option requires the least amount of setup and maintenance since it is serverless, and it does not require management of the infrastructure. Refer to this link for supporting information. A, C, and D are all solutions that can solve the problem, but require more steps for configuration, and require higher operational overhead to run and maintain.

4) B – It is most likely that the loss function is very curvy and has multiple local minima where the training is getting stuck. Decreasing the batch size would help the data scientist stochastically get out of the local minima saddles. Decreasing the learning rate would prevent overshooting the global loss function minimum. Refer to the paper at this link for an explanation.

5) D – The following calculations are required:

TP = True Positive
FP = False Positive
FN = False Negative
TN = True Negative
FN = False Negative

Recall = TP / (TP + FN)

False Positive Rate (FPR) = FP / (FP + TN)

Cost = 5 * FP + FN

	A	B	C	D
Recall	78 / (78 + 22) = 0.78	79 / (79 + 21) = 0.79	90 / (90 + 10) = 0.9	82 / (82 + 18) = 0.82
False Positive Rate	9 / (9 + 91) = 0.09	1 / (1 + 99) = 0.01	4 / (4 + 96) = 0.04	2 / (2 + 98) = 0.02
Costs	5 * 9 + 22 = 67	5 * 1 + 21 = 26	5 * 4 + 10 = 30	5 * 2 + 18 = 28

Options C and D have a recall greater than 80% and an FPR less than 10%, but D is the most cost effective. For supporting information, refer to this link.

6) B – Decreasing the class probability threshold makes the model more sensitive and, therefore, marks more cases as the positive class, which is fraud in this case. This will increase the likelihood of fraud detection. However, it comes at the price of lowering precision. This is covered in the Discussion section of the paper at this link.

7) C – With datasets that are not fully populated, the Synthetic Minority Over-sampling Technique (SMOTE) adds new information by adding synthetic data points to the minority class. This technique would be the most effective in this scenario. Refer to Section 4.2 at this link for supporting information.

8) D – Use supervised learning to predict missing values based on the values of other features. Different supervised learning approaches might have different performances, but any properly implemented supervised learning approach should provide the same or better approximation than mean or median approximation, as proposed in responses A and C. Supervised learning applied to the imputation of missing values is an active field of research. Refer to this link for an example.

9) B – In this case, a full review summary usually contains the most descriptive phrases of the entire review and is a valid stand-in for the missing full review text field. For supporting information, refer to page 1627 at this link, and this link and this link.

10) D – Amazon SageMaker Object2Vec generalizes the Word2Vec embedding technique for words to more complex objects, such as sentences and paragraphs. Since the supervised learning task is at the level of whole claims, for which there are labels, and no labels are available at the word level, Object2Vec needs be used instead of Word2Vec. For supporting information, refer to this link and this link.