AWS Practice Exam

MLS-P01: AWS Certified Machine Learning - Specialty Practice

Candidate Name: Fabrice Guillaume

Date: November 06, 2020

Exam Title: AWS Certified Machine Learning - Specialty Practice

Language: English

Thank you for taking the AWS Certified Machine Learning - Specialty Practice exam. Please examine the following information to determine which topics may require additional preparation.

Overall Score: 75%

Topic Level Scoring:

1.0. Data Engineering	100%
2.0. Exploratory Data Analysis	80%
3.0. Modeling	50%
4.0. Machine Learning Implementation and Operations	100%

Exam Guide

Click here for more information about the AWS Certified Machine Learning - Specialty exam. There will be a link to download the AWS Certified Machine Learning - Specialty exam guide from this page. These links provide information which will assist you in preparing for the certification exam.

If you have any questions about the AWS Certification program, please contact us for assistance.

A Machine Learning Specialist is developing a model that classifies defective parts from a manufacturing process into one of eight defect types. The training data consists of 100,000 images per defect type. During the initial training of the image classification model, the Specialist notices that the validation accuracy is 89.5%, while the training accuracy is 90%. It is known that human-level performance for this type of image classification is around 97%.

What should the Specialist consider to improve the performance of the model? (Select TWO.)

- [x] A. A longer training time
- [x] B. Data augmentation
- [ ] C. Getting more training data
- [ ] D. A different optimizer
- [ ] E. L2 regularization

=> The model is underfitting

B. Data augmentation and provide more details about images may help the neural

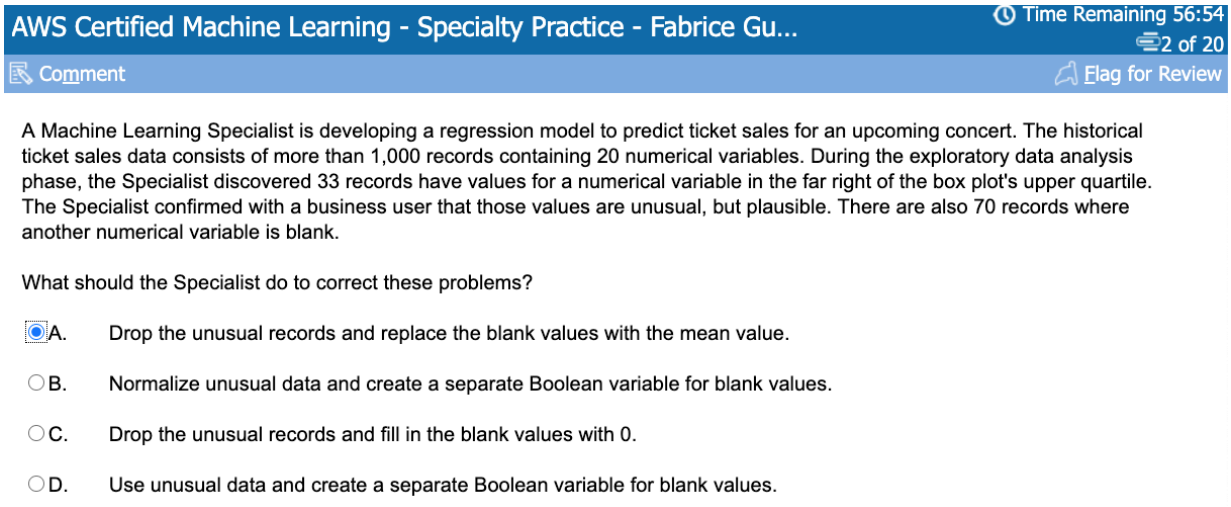
network earn more
A. A longer training time with more epochs too

MAYBE:
D. A different optimizer

NOT
E: Regularization is for overfitting
C. Need more training data - not sure, we already have 100,000 which is pretty large dataset. And this is to address overfitting pls

Techniques to reduce underfitting :

1. Increase model complexity.
2. Increase number of features, performing feature engineering. ==> B
3. Remove noise from the data.
4. Increase the number of epochs or increase the duration of training to get better results. ==> A



AWS Certified Machine Learning - Specialty Practice - Fabrice Gu...

Time Remaining 54:39

3 of 20

Comment

Flag for Review

A company is building a website that offers a variety of comedy content for adults and children. The company intends to automate the process of ingesting the content and tagging it as safe for viewing by children. Showing inappropriate content to children would cause severe harm to the company's reputation.

Which is the MOST relevant metric when evaluating the machine learning model for this task?

- A. Recall
- B. Accuracy
- C. AUC/ROC
- D. Precision

Need to minimize false positives (where we say it's safe but it is not)
=> Precision to minimize during evaluation

AWS Certified Machine Learning - Specialty Practice - Fabrice Gu...

Time Remaining 52:01

4 of 20

Comment

Flag for Review

A digital newspaper owns a large collection of articles and human-written associated article summaries. The summaries are used as headers for each article posted online, and the newspaper editors would like to find a way to produce the summaries automatically. A Machine Learning Specialist needs to apply machine learning to automate the summary-generation process.

Which solution addresses this need?

- A. Apply a neural topic model on the raw article sequence formatted as a bag of words.
- B. Project articles to a low-dimensional, compact representation by average word2vec embeddings of article words.
- C. Downsample article words to a desired summary size with the word dropout probability inverse to the word term frequency-inverse document frequency (TF-IDF).
- D. Apply seq2seq recurrent neural networks (RNNs) under the encoder-decoder framework.

D. Seq2seq is good at summarization

AWS Certified Machine Learning - Specialty Practice - Fabrice Gu...

Time Remaining 50:46

5 of 20

Comment

Flag for Review

A Machine Learning Specialist has a large number of everyday voice recordings in English stored in Amazon S3 that need to be analyzed for their conversation topics.

How should the Specialist accomplish this with the LEAST amount of effort?

- A. Run an Amazon Transcribe job, then apply a custom natural language processing (NLP) algorithm with Amazon SageMaker to the Transcribe output.
- B. Apply the Amazon SageMaker BlazingText algorithm, then run an Amazon Transcribe job.
- C. Apply a custom natural language processing (NLP) algorithm with Amazon SageMaker, then run an Amazon Transcribe job.
- D. Run an Amazon Transcribe job, then execute an Amazon Comprehend job on the Transcribe output.

AWS Certified Machine Learning - Specialty Practice - Fabrice Gu...

Time Remaining 48:17

6 of 20

Comment

Flag for Review

A Machine Learning team is trying to train a model on image classification using TensorFlow on Amazon SageMaker. The business objectives are to have a small model size and a greater than 80% accuracy. The data quality is good and sufficiently large for the use case; however, the team has been experiencing low accuracy on the test data during training. The confidence score is below 40% and, most of the time, the wrong labels are being predicted.

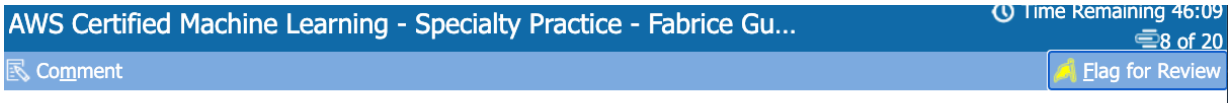
What should be done to increase the accuracy of the model?

- A. Use automatic model tuning in Amazon SageMaker. Specify the business objectives in the tuning API to optimize for these requirements. Take the best performing parameters suggested by the service and use them for training the final model.
- B. Use automatic model tuning in Amazon SageMaker. Take the best performing parameters suggested by the service and manually fine-tune the parameters to meet the business objectives.
- C. Use automatic model tuning in Amazon SageMaker. Take the best performing parameters and use those to run many training jobs in parallel with different numbers of machines using AWS Batch.
- D. Use a greater capacity of compute resources to spin up many training jobs with randomly initialized hyperparameters. Use the AWS Deep Learning AMI for experimentation. Once the best parameters are identified, use those in Amazon SageMaker.

A Machine Learning Specialist needs to monitor Amazon SageMaker in a production environment by analyzing performance metrics, setting alarms, and automatically reacting to changes in production traffic.

Which service should the Specialist use to meet these needs?

- A. AWS CloudTrail
- B. Amazon CloudWatch
- C. AWS Systems Manager
- D. AWS Config

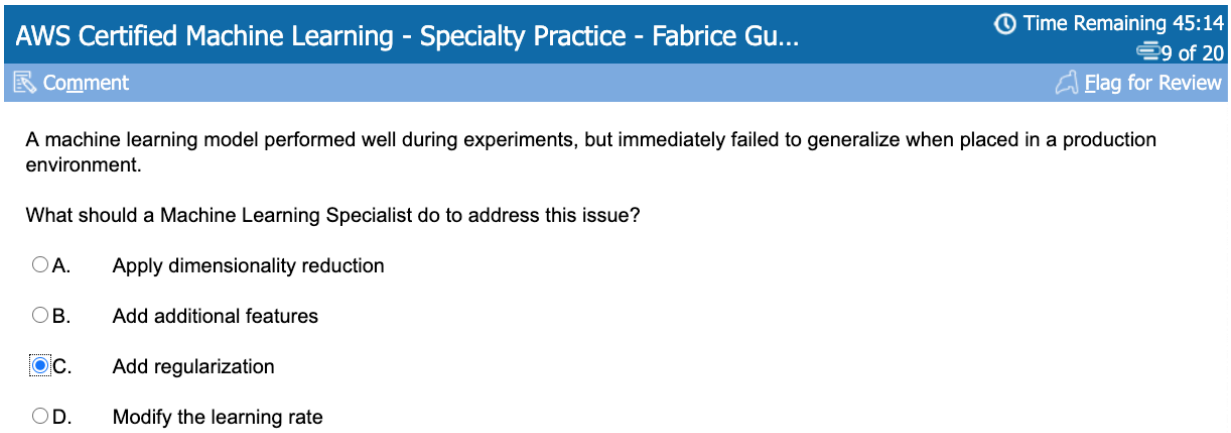


A Machine Learning Specialist is setting up a machine learning environment that will be accessed by multiple Data Scientists. The Specialist will deploy one Amazon SageMaker notebook instance for each Data Scientist and needs to ensure that each Data Scientist has access to their personal instance only.

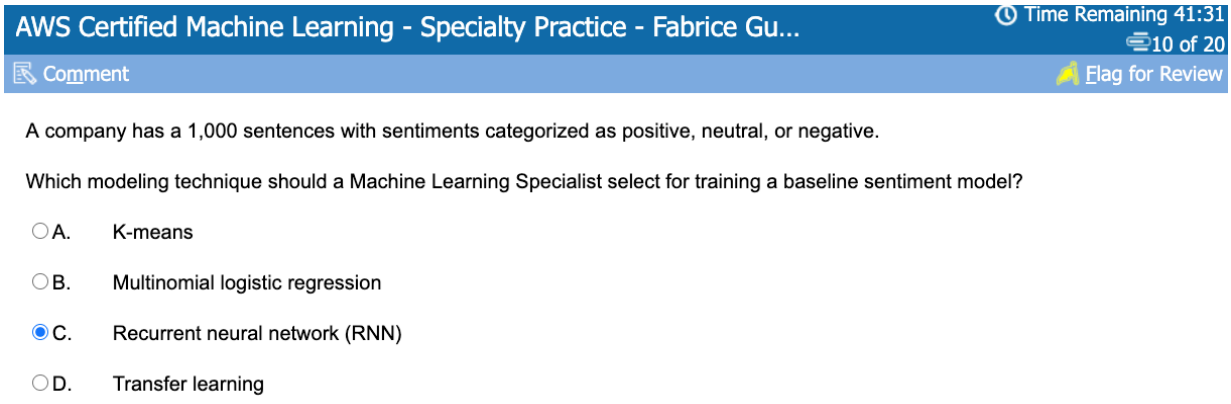
How should the Specialist manage access to SageMaker notebook instances?

- A. Attach an IAM policy to the Data Scientists' IAM users that grants access to their personal notebooks instance only.
- B. Use port forwarding to prevent all internet traffic from being forwarded to the notebook instances.
- C. Use Amazon CloudWatch to trigger an AWS Lambda function that restricts unauthorized access.
- D. Attach an Amazon S3 bucket policy to restrict access to the buckets that contain other users' notebook instances.

B. Could make it impossible to access at atll
C. Is far fetched
D. That just restricts access to DATA, but not Notebooks



The model overfits the training data => regularization



This is a multi-class classifier
==> B is the correct answer
Although C could work, but amount of data is too small for deep neural net

(1) K-means attempts to find discrete groupings within data, where members of a group are as similar as possible to one another and as different as possible from members of other groups. You define the attributes that you want the algorithm to use to determine similarity. This is used for clustering rather than classification.
(2) Multinomial Logistic Regression is a classification method that generalizes logistic regression to *multiclass* problems. It is a model that is used to predict the probabilities of the different possible outcomes of a categorical data.
(3) RNN is used in deep learning and can be used for sentiment models, but again I think the amount of data is just too small for a deep learning problem.
(4) Transfer Learning is the idea of using a model developed for a task and then reusing it as the starting point for a model on a second task. This really does not make sense for creating a *baseline* model.

AWS Certified Machine Learning - Specialty Practice - Fabrice Gu...

Time Remaining 39:39

11 of 20

Comment

Flag for Review

A Data Scientist needs to create a model for fraud detection. The dataset is composed of 2 years' worth of logged transactions, each with a small set of features. All of the transactions in the dataset were manually labeled. Since fraud does not occur frequently, the dataset is highly imbalanced. Less than 2% of the dataset was labeled as fraudulent.

Which solution provides the optimal predictive power for classifying fraudulent activity?

- A. Oversample the dataset using a clustering technique, use accuracy as the objective metric, and apply Random Cut Forest (RCF).
- B. Undersample the majority class in the dataset using a clustering technique, use precision as the objective metric, and apply Random Cut Forest (RCF).
- C. Resample the dataset (oversampling/undersampling), use the F1 score as the objective metric, and apply XGBoost.
- D. Resample the dataset (oversampling/undersampling), use accuracy as the objective metric, and apply XGBoost.

We will look at false positives (and false negatives) with F1 score, and not so much accuracy which is useless in imbalanced dataset

AWS Certified Machine Learning - Specialty Practice - Fabrice Gu...

Time Remaining 37:48

12 of 20

Comment

Flag for Review

A Machine Learning Specialist is training a model using a supervised learning algorithm. The Specialist split the dataset to use 80% of the data for training and reserved 20% of the data for testing. While evaluating the model, the Specialist discovers that the model is 97% accurate for the training dataset and 75% accurate for the test dataset.

What is the reason for the discrepancy and what action should the Specialist take?

- A. The high accuracy for the larger amount of training data means that the model is finished. Deploy the model to production.
- B. The model is currently overfitting the training data and not performing as well as it should on data it has not seen before. Change the hyperparameters to simplify and generalize the model, then retrain.
- C. Additional data in the test set is needed to balance the scoring of the model. Redistribute the data more evenly across training and test sets.
- D. The model is currently underfitting and does not have enough complexity to capture the full scope of the dataset. Change the hyperparameters to make the model more specific and complex, then retrain.

AWS Certified Machine Learning - Specialty Practice - Fabrice Gu... Time Remaining 35:03
13 of 20
Comment Flag for Review

A Data Scientist is working on a predictive maintenance model and received a company's dataset with 500,000 measurements of machine behavior during both normal operations and failures. In the dataset, 98% of the samples were collected during normal operations and 2% were collected during failures.

Which of the following actions should address the imbalance while minimizing information loss? (Select TWO.)

- [x] A. Request more data from the company, focusing on failure samples.
- [x] B. Use an approach to create synthetic samples, such as oversampling.
- [ ] C. Remove normal operations samples until the sample amount matches the number of failure samples.
- [ ] D. Run a Latent Dirichlet Allocation (LDA) algorithm on the dataset.
- [ ] E. Remove all failure samples and perform classification training using the normal operations samples only.

I could also go with C ut we need to minimize information loss and that would mean we lose a lot of good positive samples

AWS Certified Machine Learning - Specialty Practice - Fabrice Gu... Time Remaining 33:52
14 of 20
Comment Flag for Review

A city government wants to track cars in its parking lots for automatic payment. The city is looking to ingest videos of cars parking in near-real time, use machine learning to identify license plates, and store the results in a database.

Which solution meets these requirements with the LEAST amount of development effort?

- ○ A. Use Amazon Kinesis Data Streams to ingest the videos in near-real time, use the Kinesis Data Streams consumer integration with Amazon Rekognition Video to identify the license plate information, and then store the results in Amazon DynamoDB.
- ◉ B. Use Amazon Kinesis Video Streams to ingest the videos in near-real time, use the Kinesis Video Streams integration with Amazon Rekognition Video to identify the license plate information, and then store the results in Amazon DynamoDB.
- ○ C. Use Amazon Kinesis Data Streams to ingest the videos in near-real time, call Amazon Rekognition to identify the license plate information, and then store the results in Amazon DynamoDB.
- ○ D. Use Amazon S3 to ingest the videos in near-real time, trigger AWS Lambda with S3 event notifications to make a call to Amazon Rekognition Video to identify the license plate information, and then store the results in Amazon DynamoDB.

AWS Certified Machine Learning - Specialty Practice - Fabrice Gu...

Time Remaining 32:46

15 of 20

Comment | Flag for Review

A Machine Learning Specialist is optimizing a solution to define whether or not online payment transactions are fraudulent. The historical data of manually classified transactions includes the customer name, customer type, transaction amount, customer tenure, and transaction type. The transaction type is either "normal" or "abnormal."

What data preprocessing action should the Specialist take?

- A. Drop both the customer type and the transaction type before beginning to train the model.
- B. Drop the customer name and perform label encoding on the transaction type before beginning to train the model.
- C. Drop the transaction type and perform label encoding on the customer type before beginning to train the model.
- D. Train the model normally as the data is ready to be used — without any preprocessing tasks — during the model training phase.

AWS Certified Machine Learning - Specialty Practice - Fabrice Gu...

Time Remaining 28:52

16 of 20

Comment | Flag for Review

A Machine Learning Specialist is approached by a customer with an MP3 audio file of a press conference recorded in Spanish. The company wants to provide a report of the press conference to English-speaking senior managers that details the questions asked, the timestamp of each question, and the sentiment of each person who asked a question.

What is the MOST efficient order of AWS machine learning technologies that should be used to accomplish this?

- A. Amazon Translate → Amazon Comprehend → Amazon Transcribe
- B. Amazon Transcribe → Amazon Comprehend → Amazon Translate
- C. Amazon Translate → Amazon Comprehend → Amazon Transcribe → Amazon Translate
- D. Amazon Transcribe → Amazon Translate → Amazon Comprehend → Amazon Translate

Audio Spanish > **Transcribe** > Text Spanish + time > **Translate** > Text English > **Comprehend** > sentiment > English
Translate works against TEXT files only, correct?

Audio Spanish > **Transcribe** > Text Spanish + time > **Comprehend** > sentiment > **Translate** > English
Does Comprehend works against Spanish?

AWS Certified Machine Learning - Specialty Practice - Fabrice Gu...

Time Remaining 28:10

17 of 20

Comment

Flag for Review

A Machine Learning Specialist has 1 TB of files and the associated metadata stored in a data lake within an S3 bucket. The Specialist wants to search the metadata to better evaluate the dataset. The Specialist expects to search through the metadata multiple times.

Which solution meets the requirements with the LEAST amount of effort?

- A. Enable S3 analytics, and then review and search through the file metadata.
- B. Use Amazon Athena to review and query the file metadata.
- C. Deploy an Amazon EMR cluster to process and search through the metadata on the data lake.
- D. Use AWS Lambda to send the metadata to Amazon Kinesis Data Streams, and use Amazon Kinesis Data Analytics to run searches on the metadata.

AWS Certified Machine Learning - Specialty Practice - Fabrice Gu...

Time Remaining 27:09

18 of 20

Comment

Flag for Review

A Machine Learning Specialist is building a new model and wants to test multiple production variants using live data in a beta environment, where customers will interact with the model. Based on these interactions, the Specialist will A/B test the models, then select and deploy the best model.

What is the SIMPLEST method for testing the multiple model variants?

- A. Deploy multiple versions of the model on different Amazon EC2 instances using the AWS Deep Learning AMI, evaluate the model performance, and terminate the instances that are not hosting the best-performing model.
- B. Use Amazon SageMaker to deploy different versions of the model behind a single endpoint and route a percentage of traffic to each, select the best performing model, and reroute 100% of traffic to that model.
- C. Use Amazon SageMaker to deploy an endpoint for each model and then use an Application Load Balancer to route a percentage of traffic to each model, and gradually route 100% of traffic to the best model.
- D. Use Amazon SageMaker to deploy an endpoint for each model and then use a Network Load Balancer to route a percentage of traffic to each model, and route 100% of traffic to the best model.

AWS Certified Machine Learning - Specialty Practice - Fabrice Gu... Time Remaining 21:21

19 of 20

Comment Flag for Review

A manufacturer has deployed an array of 50,000 sensors throughout its plant to predict failures in components. Data Scientists have built a long short-term memory (LSTM) model in Gluon and are training it using the Amazon SageMaker Python API. The Data Scientists are training the model using a time series with 10 million examples. Training is currently taking 100 hours and Data Scientists are attempting to speed it up by using multiple GPUs. However, when they modified the code to use 8 GPUs, it is running slightly slower than on 1 GPU. The current hyperparameter settings are:

Hyperparameter	Value
Batch size	128
Clip gradient	10
Autoregressive window	160
Learning rate	0.01
Epochs	80

Which of the following changes together are recommended to speed up training on 8 GPUs while maintaining test accuracy? (Select TWO.)

- [ ] A. Increase the batch size by a factor of 8.
- [x] B. Increase the clip gradient by 8.
- [x] C. Increase the autoregressive window by a factor of 8.
- [ ] D. Increase the learning rate by a factor of 8.
- [ ] E. Decrease the number of epochs by 20.

A. Batch size can be increase by 8, to feed more training data to the 8 GPUs and this will require less batches to complete one epoch. However it can also cause the model to be stuck in a local minima and hence impact the accuracy
D. Learning rate increase may cause jumping over and oscillating around minima and hence impact accuracy
E. Decreasing epochs may results into less accuracy (model learns epic after epocc)

B and C. Not sure what clip gradient and autoregressive window do????

By elimination, going with B and C

INCORRECT
=> correct answers are:
- increasing batch size by a factor of 8
- increasing learning rate by 8
Increase the mini-batch size by multiplying by the number of GPUs to keep the mini-batch size per GPU constant. For example, if a mini-batch size of 128 keeps a single GPU fully utilized, you should increase to a mini-batch size of 512 when using four GPUs.
To increase the rate of convergence with larger mini-batch size, you must increase the learning rate of the SGD optimizer.

A Machine Learning Specialist is setting up a machine learning pipeline. The goal is to enable the ETL part of the pipeline to trigger machine learning training jobs in Amazon SageMaker. In particular, the Specialist intends to use an Amazon EMR cluster to handle the ETL tasks, and would like it to interface with Amazon SageMaker without writing code specifically to connect Amazon SageMaker and the EMR cluster.

Which framework enables this goal?

- A. Apache Hive
- B. Apache Flink
- C. Apache Spark
- D. Apache Pig