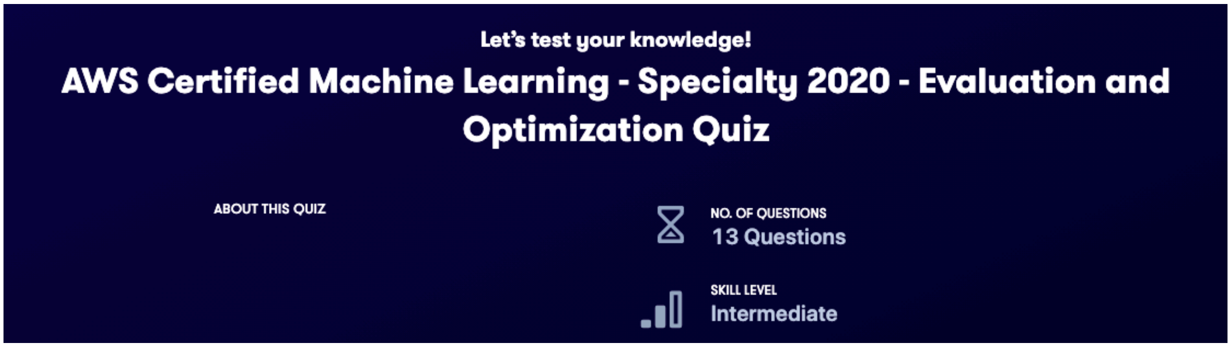
Cloud Guru - 7 - Evaluation and Optimization Quiz

https://acloud.guru/course/aws-certified-machine-learning-specialty/learn/c865bf80-7377-1ff0-4077-ff6cb94bf923/chapter-8/watch?backUrl=~2Fcourses&backUrl=%2Fcourses



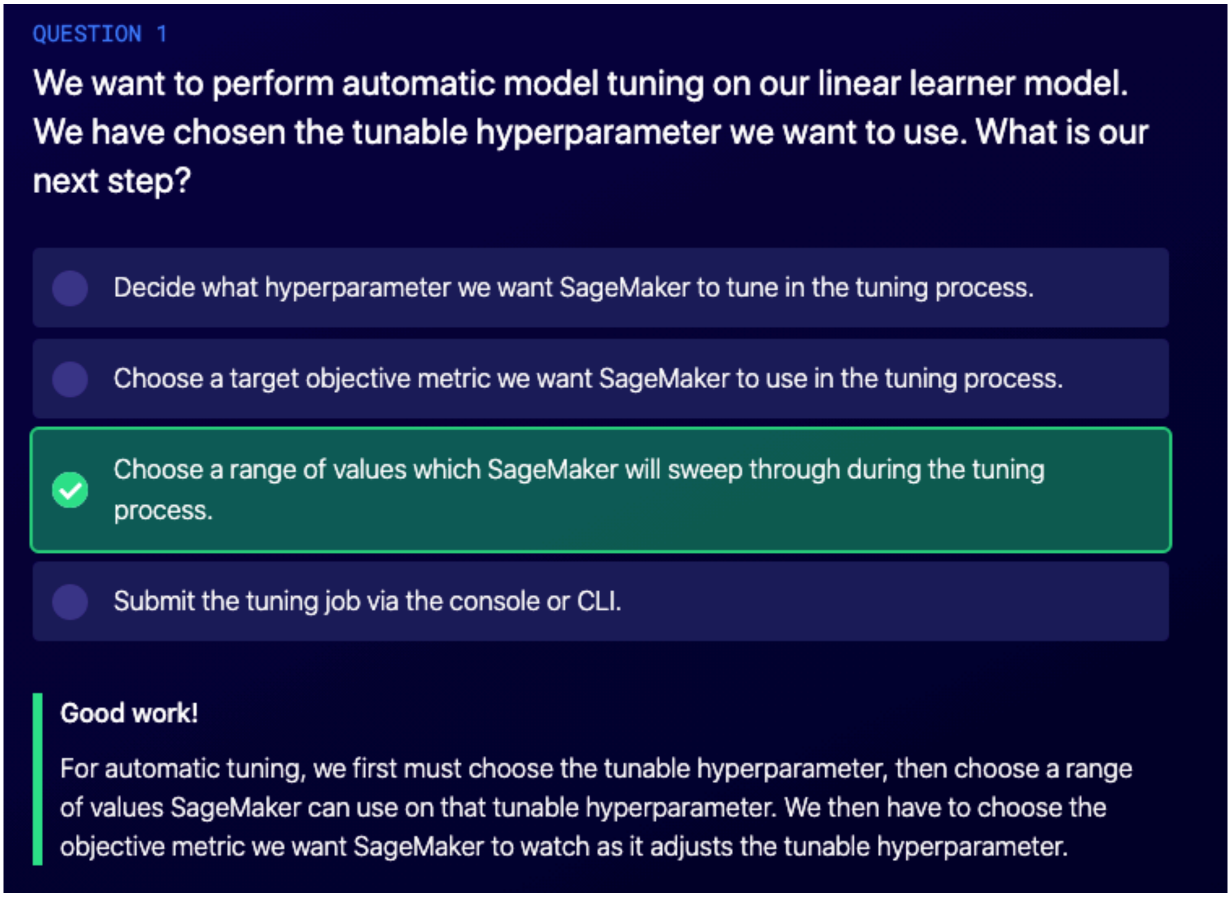
Let's test your knowledge!

AWS Certified Machine Learning - Specialty 2020 - Evaluation and Optimization Quiz

ABOUT THIS QUIZ

NO. OF QUESTIONS
13 Questions

SKILL LEVEL
Intermediate



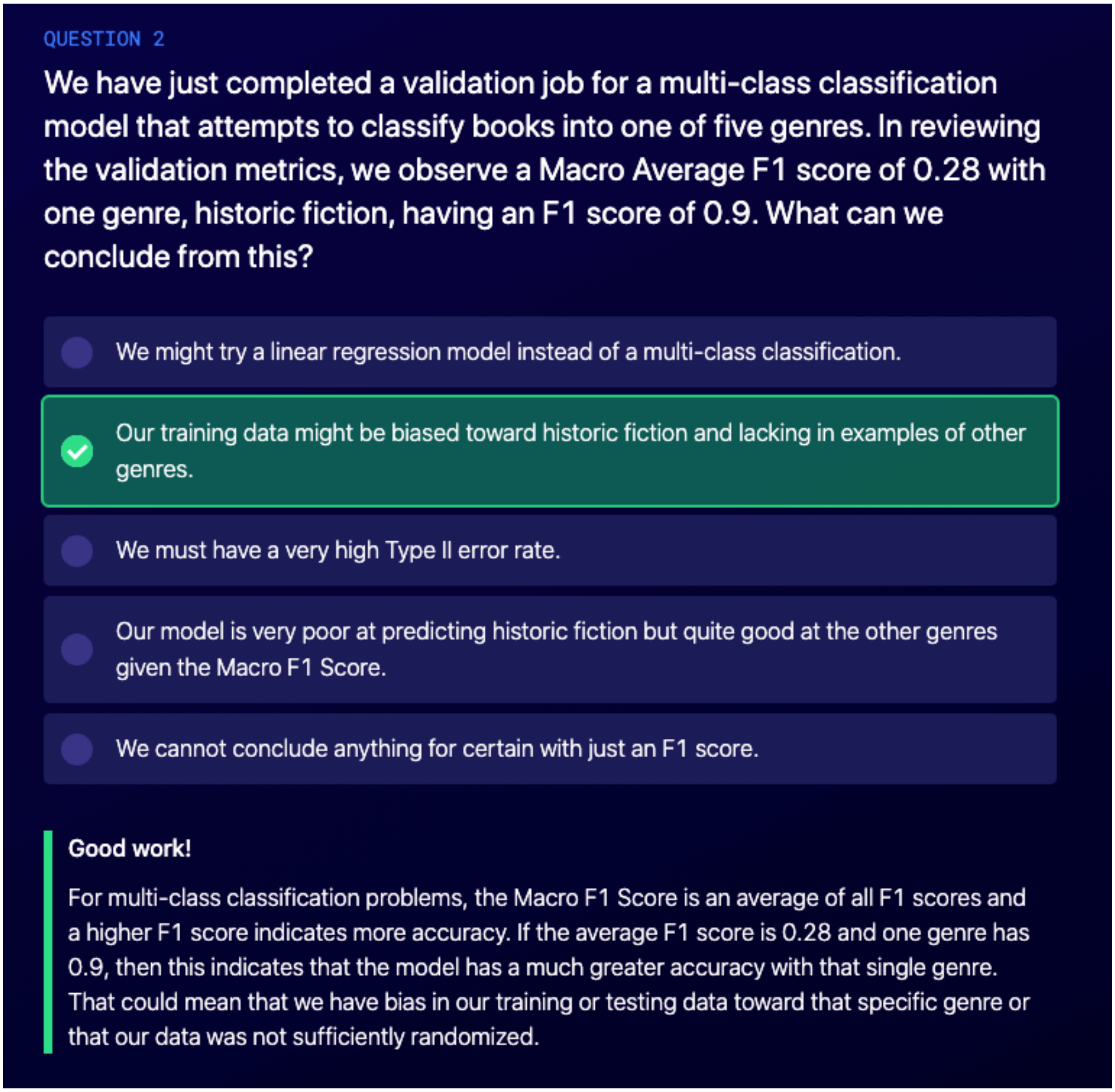
QUESTION 1

We want to perform automatic model tuning on our linear learner model. We have chosen the tunable hyperparameter we want to use. What is our next step?

- ○ Decide what hyperparameter we want SageMaker to tune in the tuning process.
- ○ Choose a target objective metric we want SageMaker to use in the tuning process.
- ✓ Choose a range of values which SageMaker will sweep through during the tuning process.
- ○ Submit the tuning job via the console or CLI.

Good work!

For automatic tuning, we first must choose the tunable hyperparameter, then choose a range of values SageMaker can use on that tunable hyperparameter. We then have to choose the objective metric we want SageMaker to watch as it adjusts the tunable hyperparameter.



QUESTION 2

We have just completed a validation job for a multi-class classification model that attempts to classify books into one of five genres. In reviewing the validation metrics, we observe a Macro Average F1 score of 0.28 with one genre, historic fiction, having an F1 score of 0.9. What can we conclude from this?

○ We might try a linear regression model instead of a multi-class classification.

✓ Our training data might be biased toward historic fiction and lacking in examples of other genres.

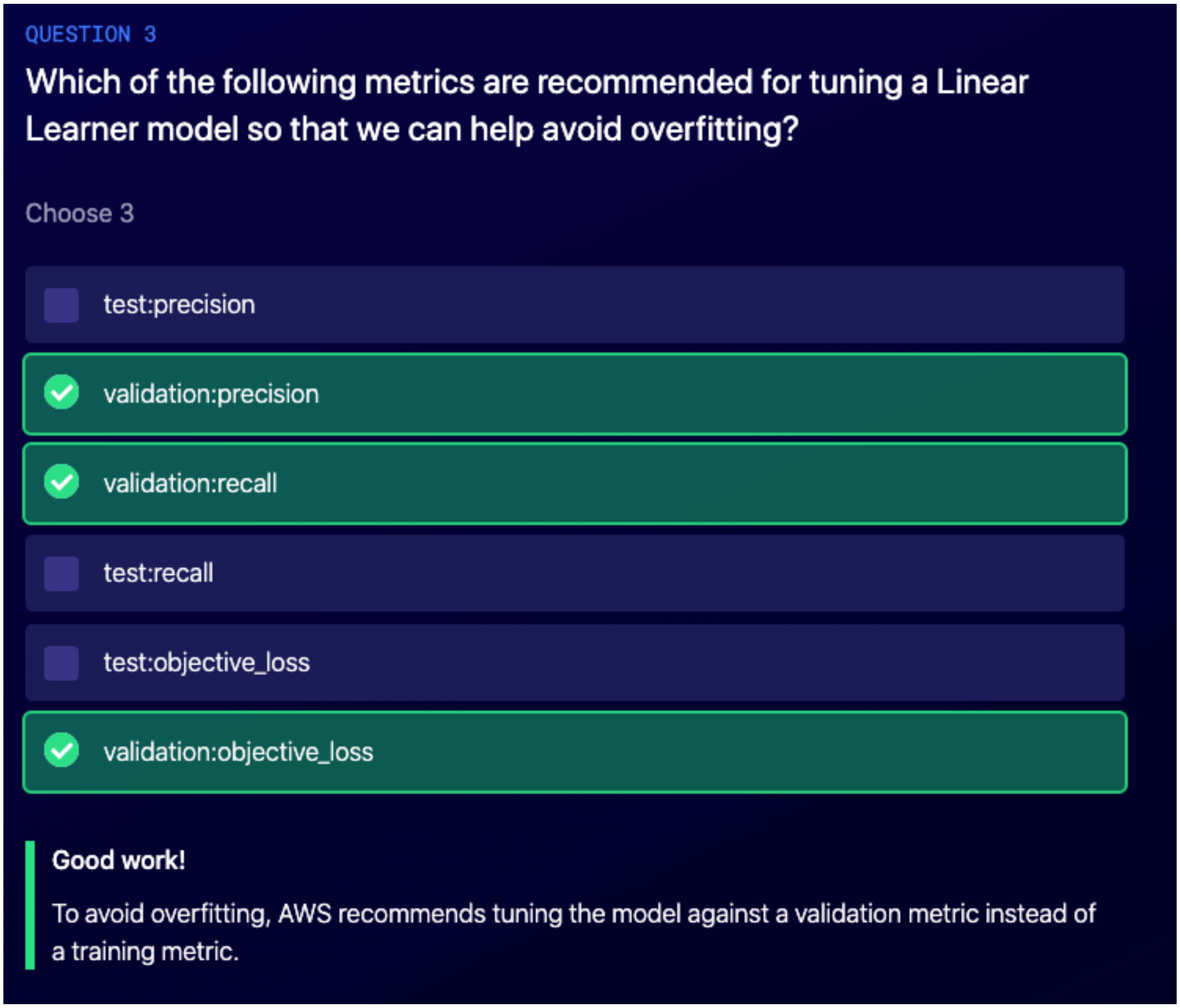
○ We must have a very high Type II error rate.

○ Our model is very poor at predicting historic fiction but quite good at the other genres given the Macro F1 Score.

○ We cannot conclude anything for certain with just an F1 score.

Good work!

For multi-class classification problems, the Macro F1 Score is an average of all F1 scores and a higher F1 score indicates more accuracy. If the average F1 score is 0.28 and one genre has 0.9, then this indicates that the model has a much greater accuracy with that single genre. That could mean that we have bias in our training or testing data toward that specific genre or that our data was not sufficiently randomized.



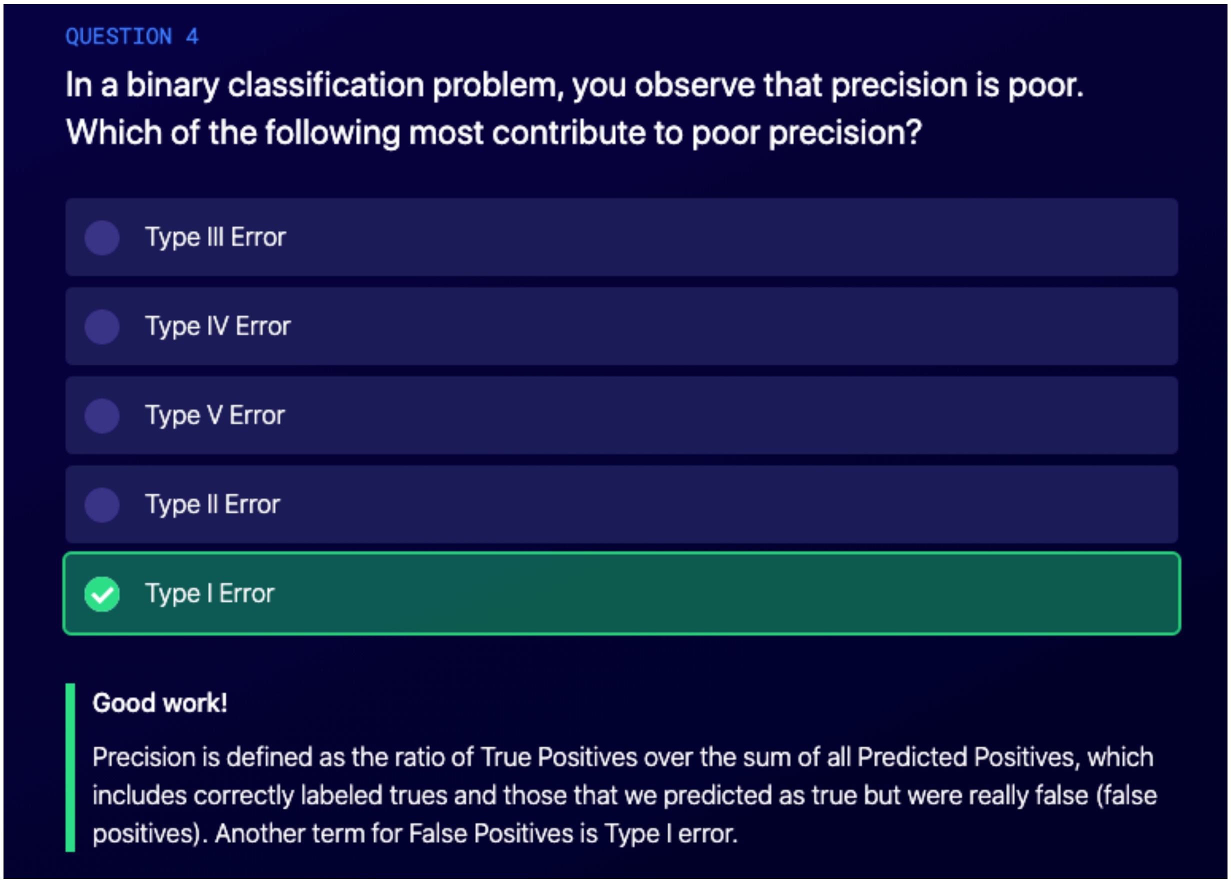
QUESTION 3

Which of the following metrics are recommended for tuning a Linear Learner model so that we can help avoid overfitting?

Choose 3

- □ test:precision
- ☑ validation:precision
- ☑ validation:recall
- □ test:recall
- □ test:objective_loss
- ☑ validation:objective_loss

> **Good work!**
>
> To avoid overfitting, AWS recommends tuning the model against a validation metric instead of a training metric.



QUESTION 4

In a binary classification problem, you observe that precision is poor. Which of the following most contribute to poor precision?

○ Type III Error

○ Type IV Error

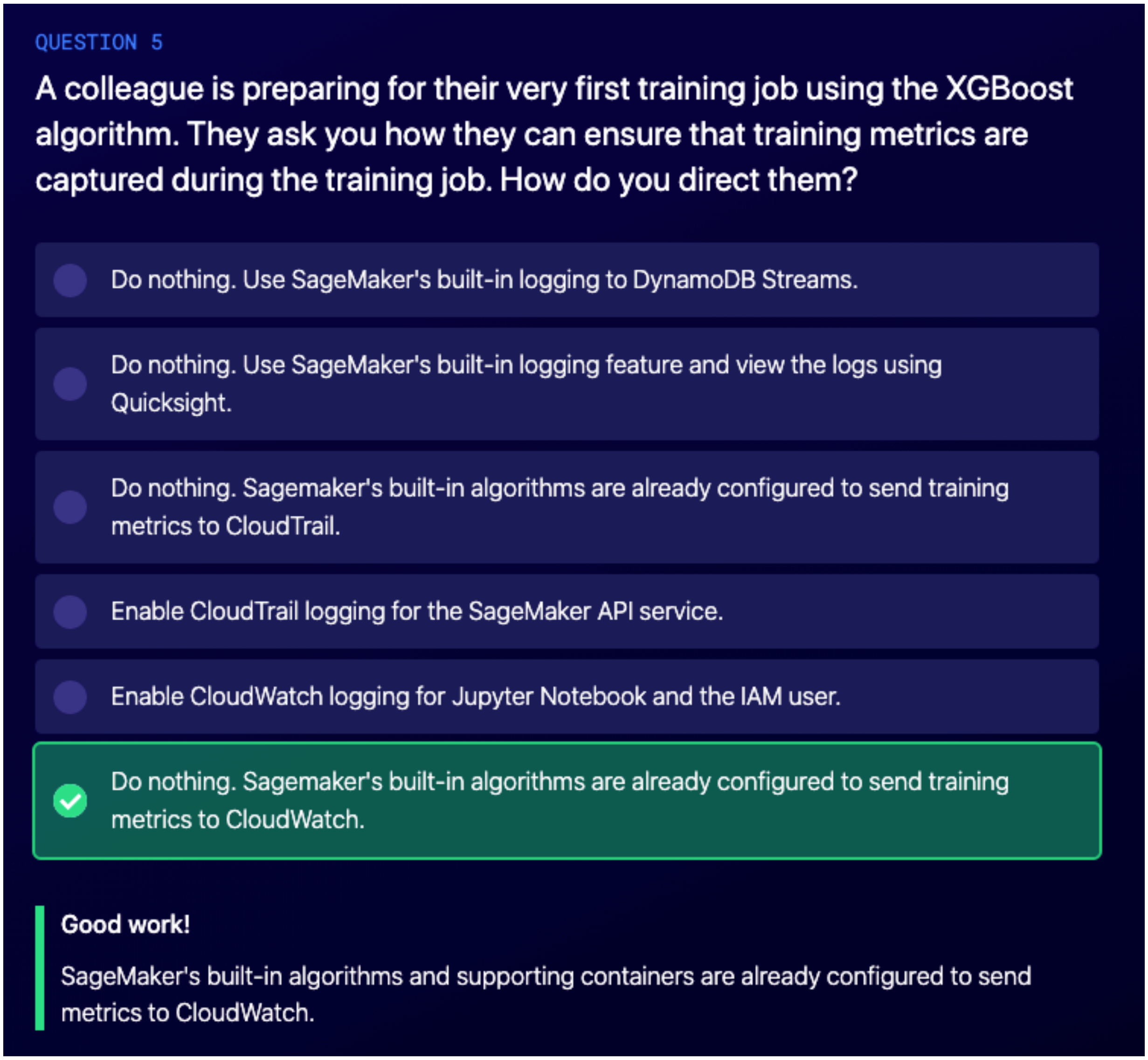
○ Type V Error

○ Type II Error

✓ Type I Error

Good work!

Precision is defined as the ratio of True Positives over the sum of all Predicted Positives, which includes correctly labeled trues and those that we predicted as true but were really false (false positives). Another term for False Positives is Type I error.



QUESTION 5

A colleague is preparing for their very first training job using the XGBoost algorithm. They ask you how they can ensure that training metrics are captured during the training job. How do you direct them?

○ Do nothing. Use SageMaker's built-in logging to DynamoDB Streams.

○ Do nothing. Use SageMaker's built-in logging feature and view the logs using Quicksight.

○ Do nothing. Sagemaker's built-in algorithms are already configured to send training metrics to CloudTrail.

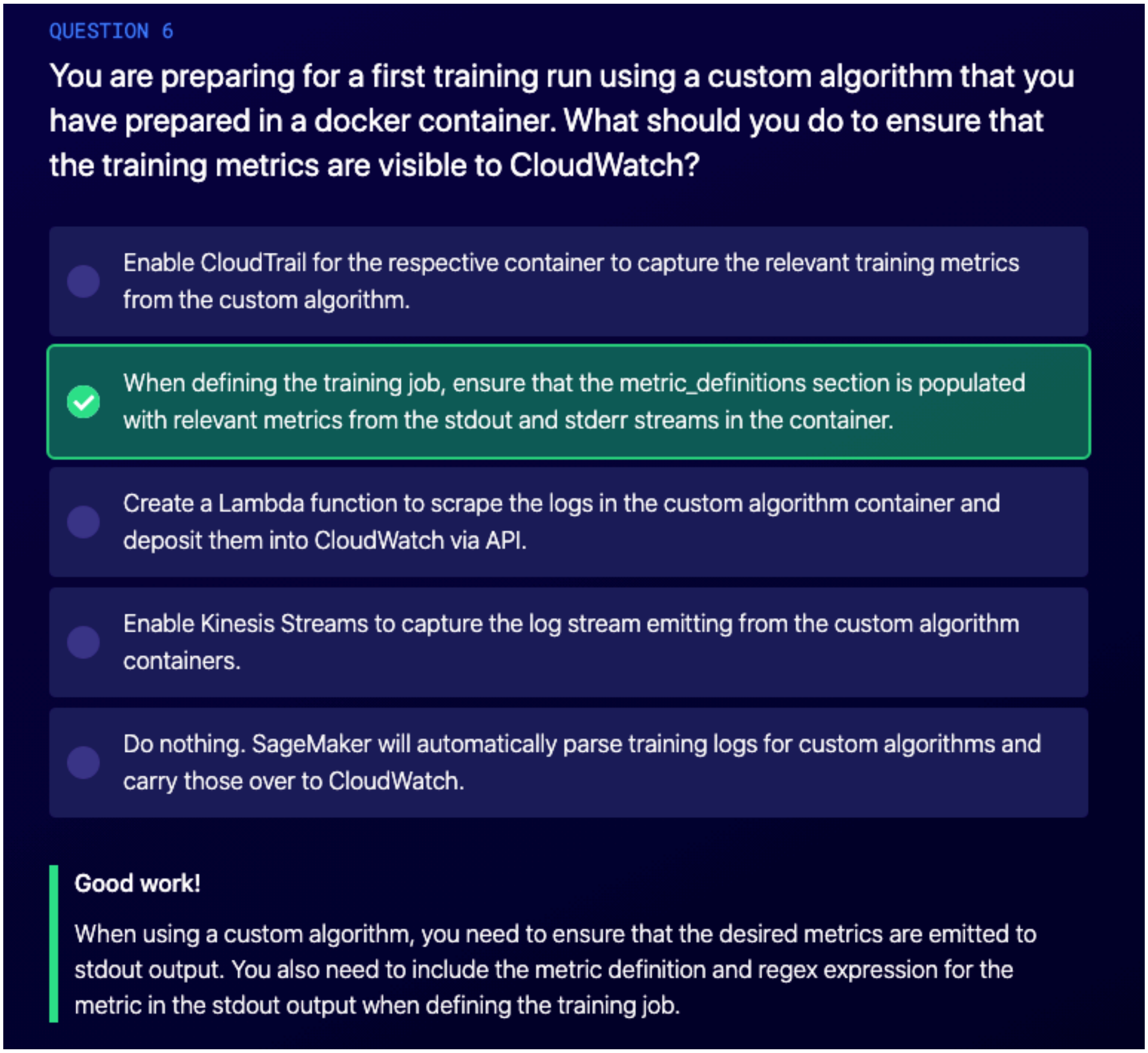
○ Enable CloudTrail logging for the SageMaker API service.

○ Enable CloudWatch logging for Jupyter Notebook and the IAM user.

✓ Do nothing. Sagemaker's built-in algorithms are already configured to send training metrics to CloudWatch.

Good work!

SageMaker's built-in algorithms and supporting containers are already configured to send metrics to CloudWatch.



QUESTION 6

You are preparing for a first training run using a custom algorithm that you have prepared in a docker container. What should you do to ensure that the training metrics are visible to CloudWatch?

○ Enable CloudTrail for the respective container to capture the relevant training metrics from the custom algorithm.

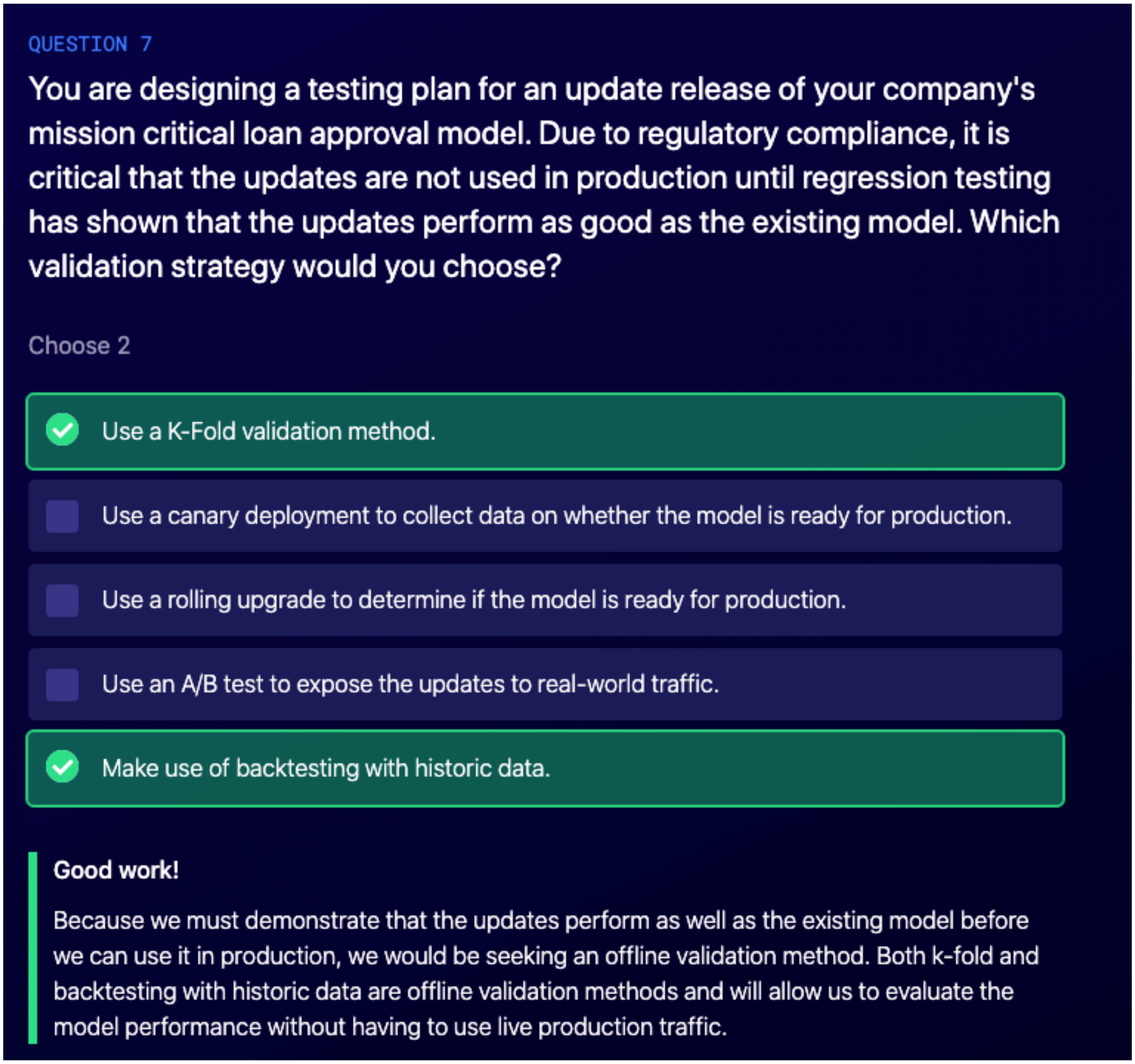
✓ When defining the training job, ensure that the metric_definitions section is populated with relevant metrics from the stdout and stderr streams in the container.

○ Create a Lambda function to scrape the logs in the custom algorithm container and deposit them into CloudWatch via API.

○ Enable Kinesis Streams to capture the log stream emitting from the custom algorithm containers.

○ Do nothing. SageMaker will automatically parse training logs for custom algorithms and carry those over to CloudWatch.

> **Good work!**
>
> When using a custom algorithm, you need to ensure that the desired metrics are emitted to stdout output. You also need to include the metric definition and regex expression for the metric in the stdout output when defining the training job.



QUESTION 7

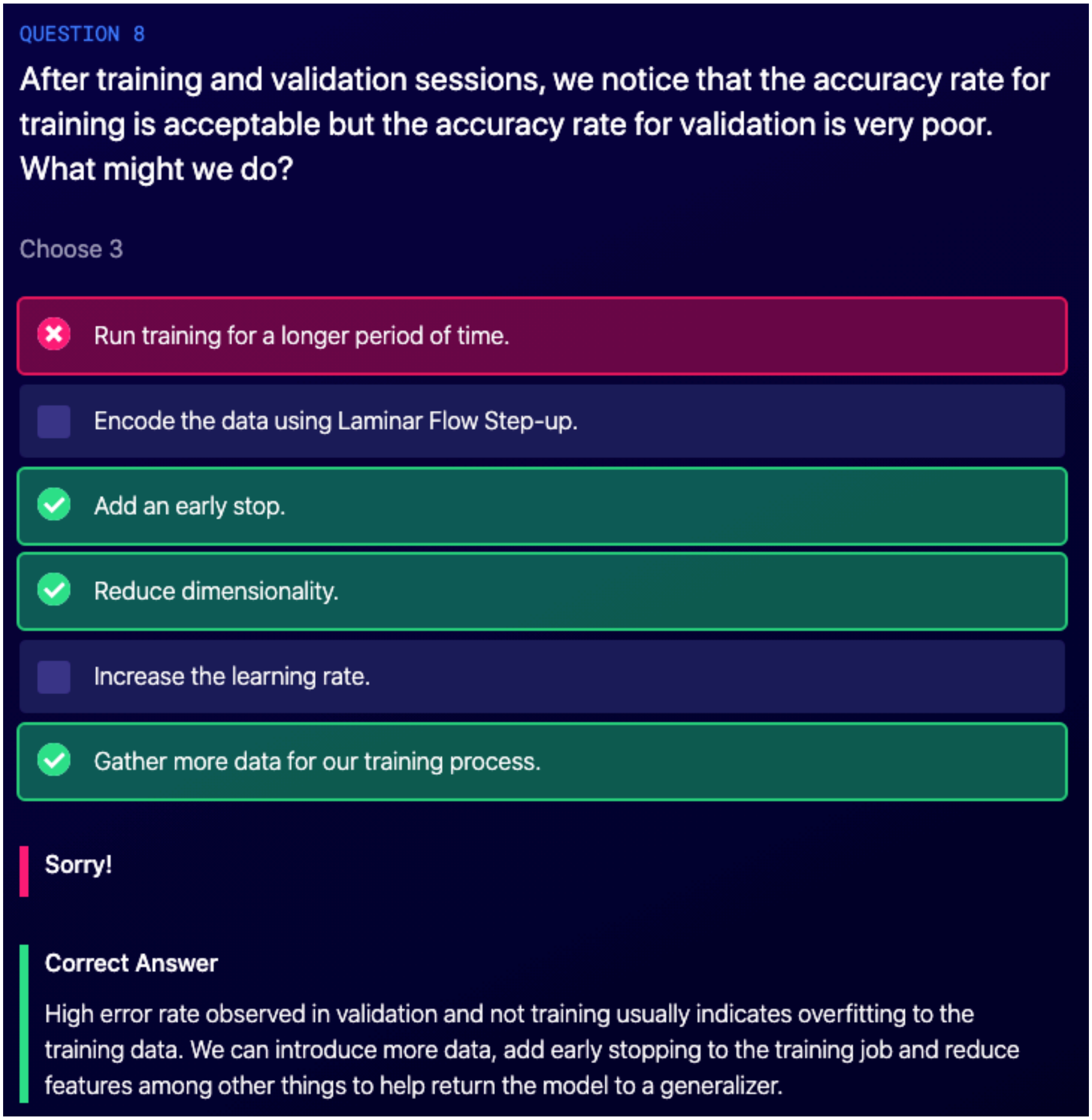
You are designing a testing plan for an update release of your company's mission critical loan approval model. Due to regulatory compliance, it is critical that the updates are not used in production until regression testing has shown that the updates perform as good as the existing model. Which validation strategy would you choose?

Choose 2

- ✓ Use a K-Fold validation method.
- □ Use a canary deployment to collect data on whether the model is ready for production.
- □ Use a rolling upgrade to determine if the model is ready for production.
- □ Use an A/B test to expose the updates to real-world traffic.
- ✓ Make use of backtesting with historic data.

Good work!

Because we must demonstrate that the updates perform as well as the existing model before we can use it in production, we would be seeking an offline validation method. Both k-fold and backtesting with historic data are offline validation methods and will allow us to evaluate the model performance without having to use live production traffic.



QUESTION 8

After training and validation sessions, we notice that the accuracy rate for training is acceptable but the accuracy rate for validation is very poor. What might we do?

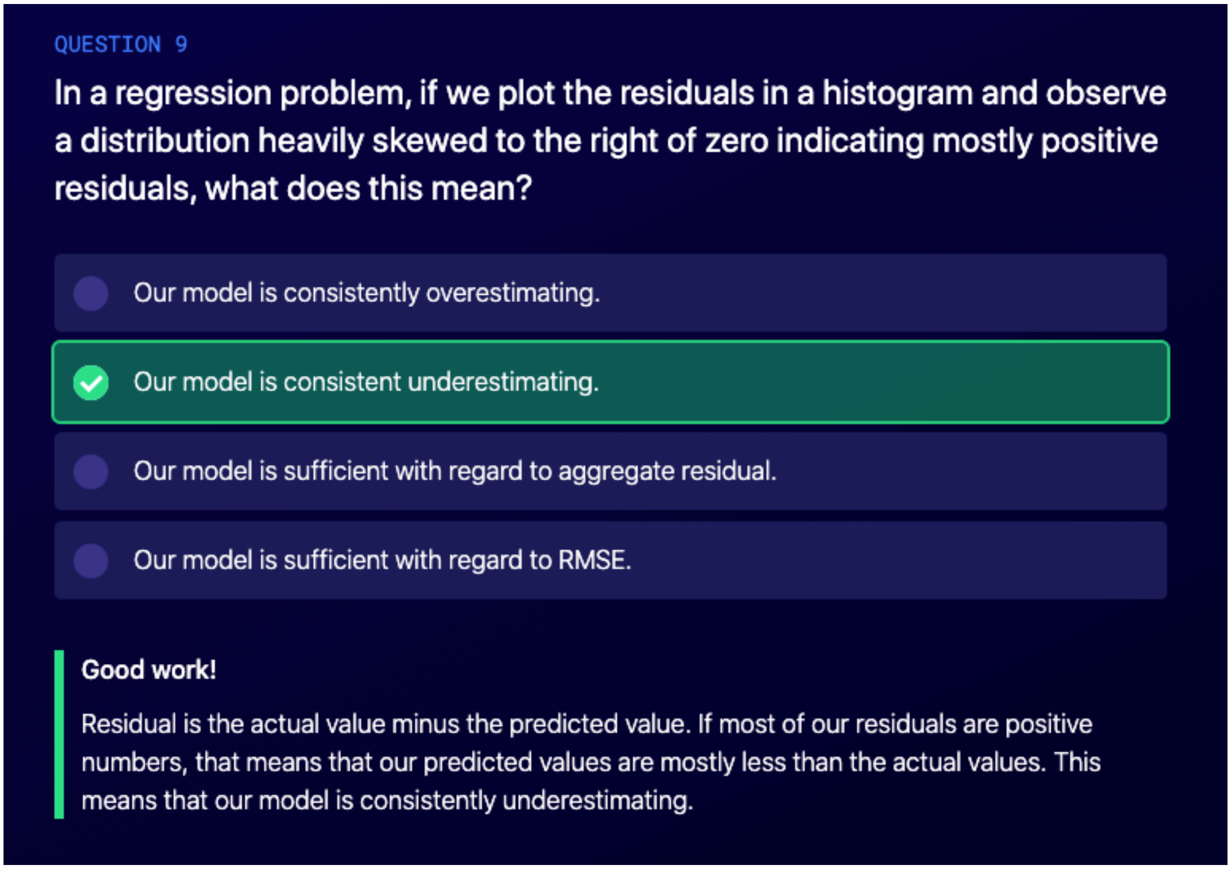
Choose 3

- ☒ Run training for a longer period of time.
- □ Encode the data using Laminar Flow Step-up.
- ✓ Add an early stop.
- ✓ Reduce dimensionality.
- □ Increase the learning rate.
- ✓ Gather more data for our training process.

Sorry!

Correct Answer

High error rate observed in validation and not training usually indicates overfitting to the training data. We can introduce more data, add early stopping to the training job and reduce features among other things to help return the model to a generalizer.

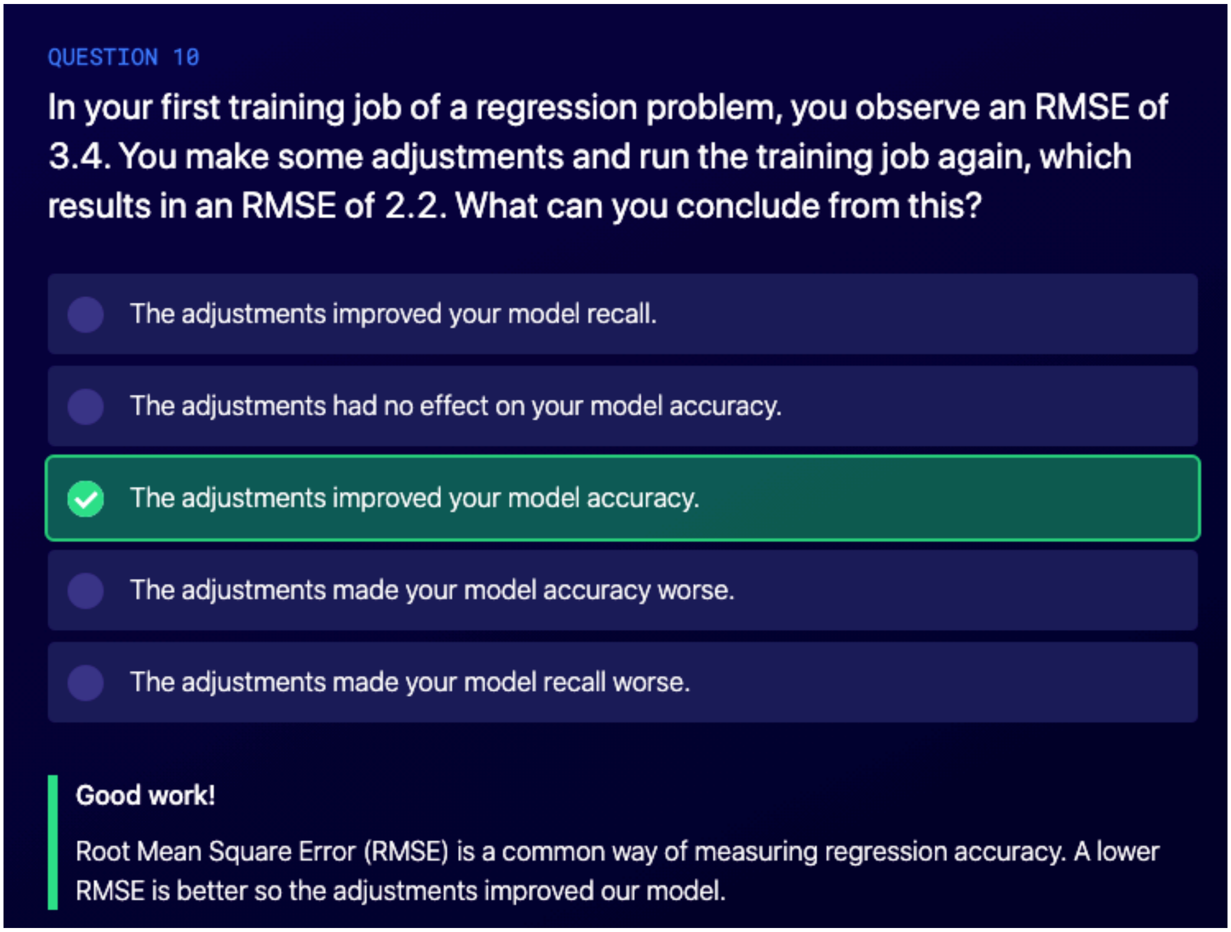


QUESTION 9

In a regression problem, if we plot the residuals in a histogram and observe a distribution heavily skewed to the right of zero indicating mostly positive residuals, what does this mean?

- ○ Our model is consistently overestimating.
- ✓ Our model is consistent underestimating.
- ○ Our model is sufficient with regard to aggregate residual.
- ○ Our model is sufficient with regard to RMSE.

> **Good work!**
>
> Residual is the actual value minus the predicted value. If most of our residuals are positive numbers, that means that our predicted values are mostly less than the actual values. This means that our model is consistently underestimating.



QUESTION 10

In your first training job of a regression problem, you observe an RMSE of 3.4. You make some adjustments and run the training job again, which results in an RMSE of 2.2. What can you conclude from this?

○ The adjustments improved your model recall.

○ The adjustments had no effect on your model accuracy.

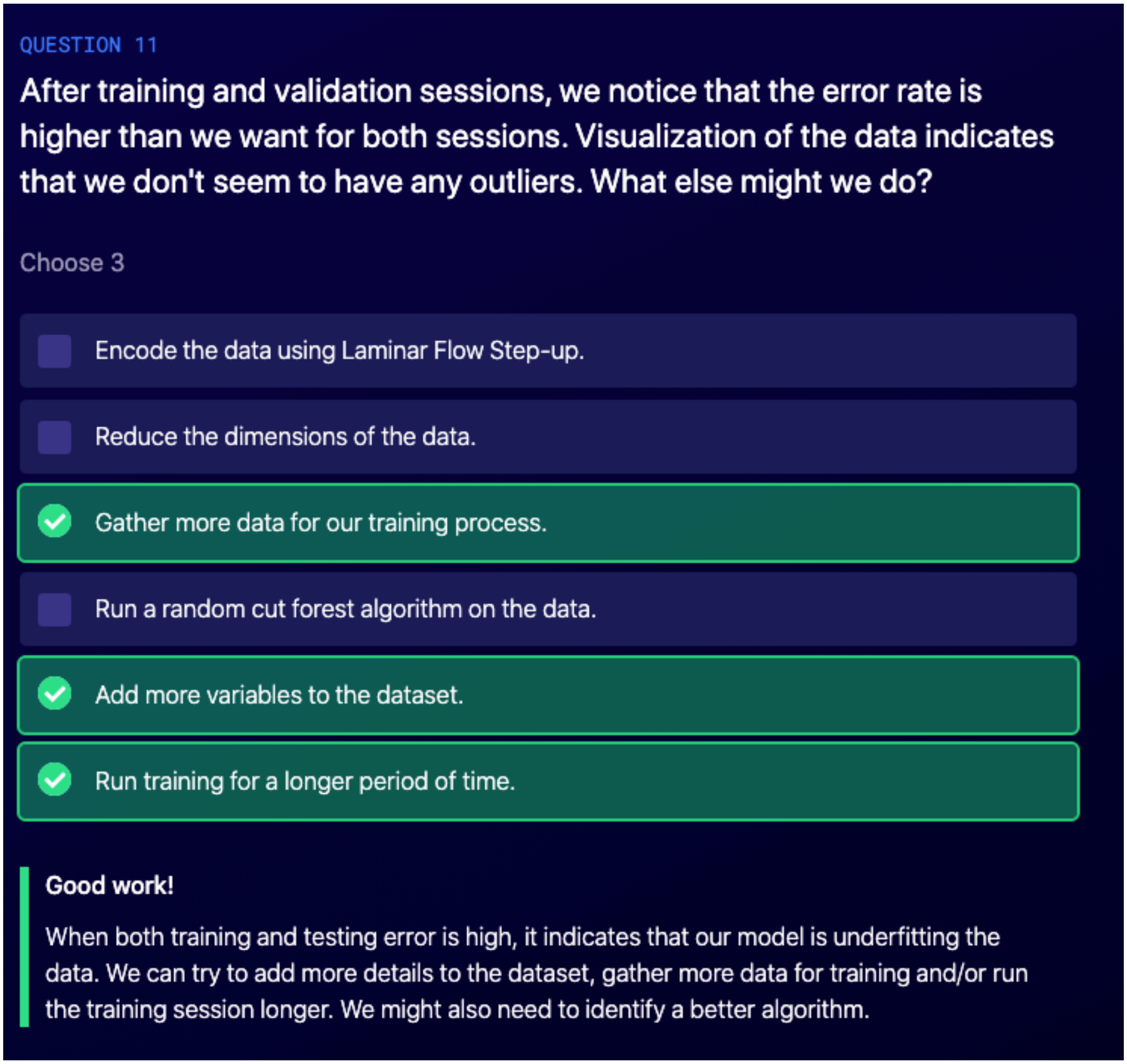
✓ The adjustments improved your model accuracy.

○ The adjustments made your model accuracy worse.

○ The adjustments made your model recall worse.

Good work!

Root Mean Square Error (RMSE) is a common way of measuring regression accuracy. A lower RMSE is better so the adjustments improved our model.



QUESTION 11

After training and validation sessions, we notice that the error rate is higher than we want for both sessions. Visualization of the data indicates that we don't seem to have any outliers. What else might we do?

Choose 3

□ Encode the data using Laminar Flow Step-up.

□ Reduce the dimensions of the data.

☑ Gather more data for our training process.

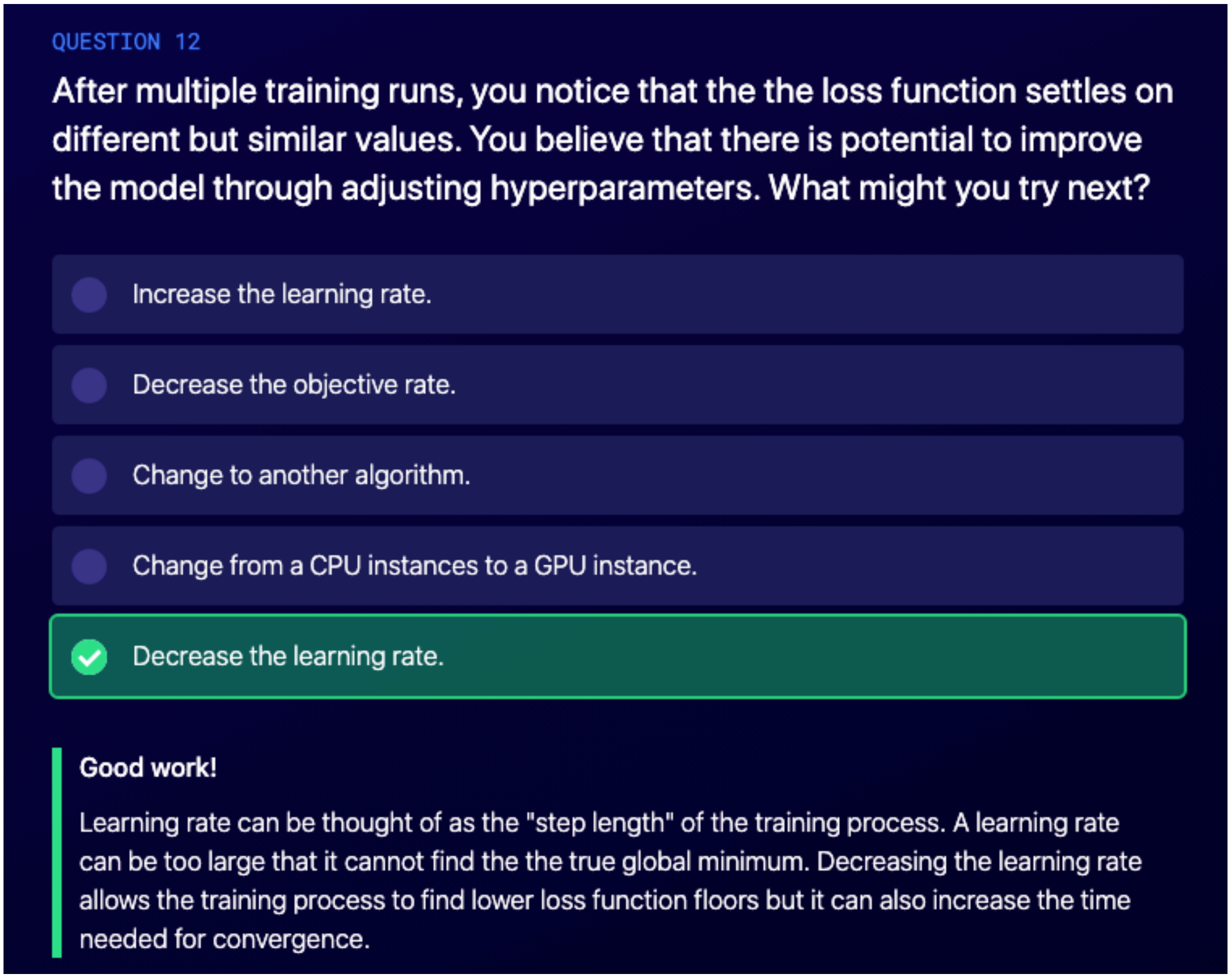
□ Run a random cut forest algorithm on the data.

☑ Add more variables to the dataset.

☑ Run training for a longer period of time.

Good work!

When both training and testing error is high, it indicates that our model is underfitting the data. We can try to add more details to the dataset, gather more data for training and/or run the training session longer. We might also need to identify a better algorithm.



QUESTION 12

After multiple training runs, you notice that the the loss function settles on different but similar values. You believe that there is potential to improve the model through adjusting hyperparameters. What might you try next?

○ Increase the learning rate.

○ Decrease the objective rate.

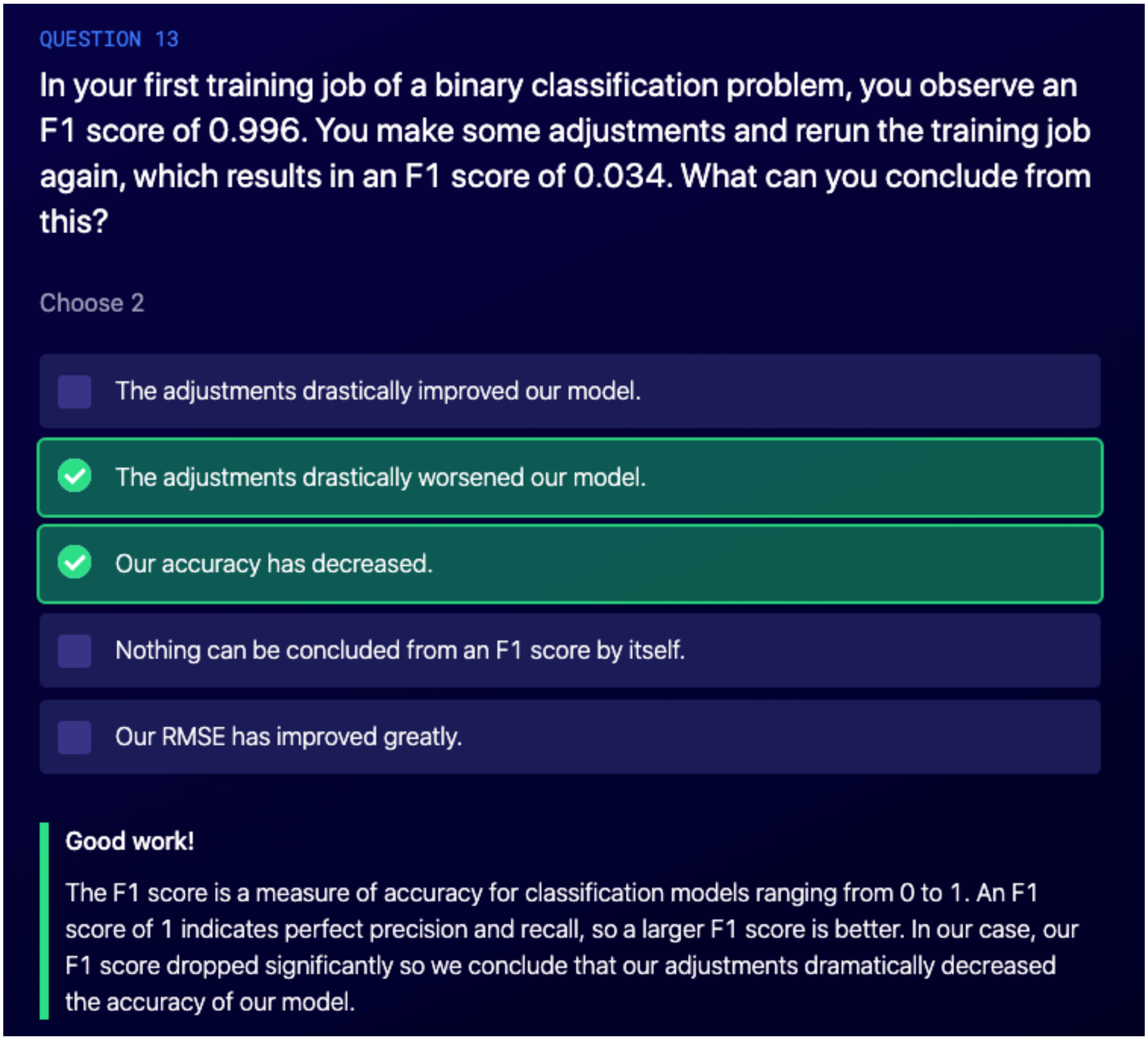
○ Change to another algorithm.

○ Change from a CPU instances to a GPU instance.

✓ Decrease the learning rate.

Good work!

Learning rate can be thought of as the "step length" of the training process. A learning rate can be too large that it cannot find the the true global minimum. Decreasing the learning rate allows the training process to find lower loss function floors but it can also increase the time needed for convergence.



QUESTION 13

In your first training job of a binary classification problem, you observe an F1 score of 0.996. You make some adjustments and rerun the training job again, which results in an F1 score of 0.034. What can you conclude from this?

Choose 2

- □ The adjustments drastically improved our model.
- ✓ The adjustments drastically worsened our model.
- ✓ Our accuracy has decreased.
- □ Nothing can be concluded from an F1 score by itself.
- □ Our RMSE has improved greatly.

Good work!

The F1 score is a measure of accuracy for classification models ranging from 0 to 1. An F1 score of 1 indicates perfect precision and recall, so a larger F1 score is better. In our case, our F1 score dropped significantly so we conclude that our adjustments dramatically decreased the accuracy of our model.

92%

Congratulations!

You passed AWS Certified Machine Learning - Specialty 2020 - Evaluation and Optimization Quiz!