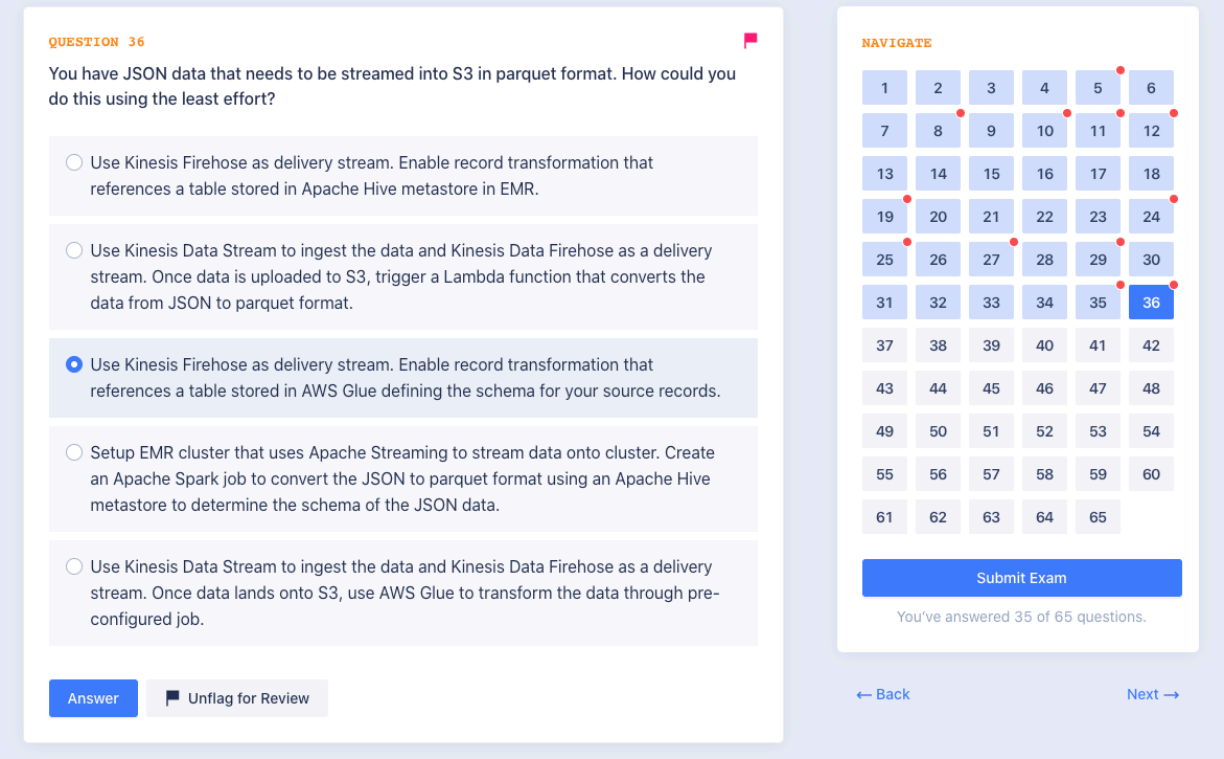
CloudGuru - Practice Exam - part 2

Questions 36-65



QUESTION 36

You have JSON data that needs to be streamed into S3 in parquet format. How could you do this using the least effort?

- Use Kinesis Firehose as delivery stream. Enable record transformation that references a table stored in Apache Hive metastore in EMR.
- Use Kinesis Data Stream to ingest the data and Kinesis Data Firehose as a delivery stream. Once data is uploaded to S3, trigger a Lambda function that converts the data from JSON to parquet format.
- **Use Kinesis Firehose as delivery stream. Enable record transformation that references a table stored in AWS Glue defining the schema for your source records.**
- Setup EMR cluster that uses Apache Streaming to stream data onto cluster. Create an Apache Spark job to convert the JSON to parquet format using an Apache Hive metastore to determine the schema of the JSON data.
- Use Kinesis Data Stream to ingest the data and Kinesis Data Firehose as a delivery stream. Once data lands onto S3, use AWS Glue to transform the data through pre-configured job.

Answer | Unflag for Review

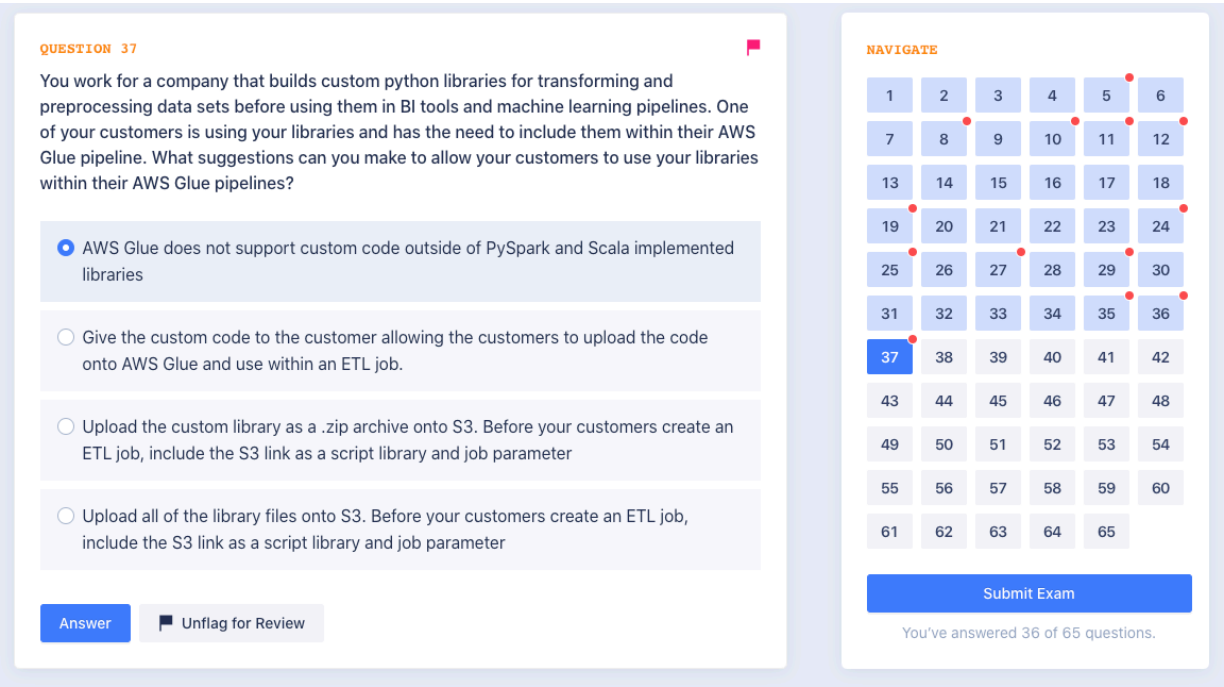
NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 35 of 65 questions.

← Back Next →



QUESTION 37

You work for a company that builds custom python libraries for transforming and preprocessing data sets before using them in BI tools and machine learning pipelines. One of your customers is using your libraries and has the need to include them within their AWS Glue pipeline. What suggestions can you make to allow your customers to use your libraries within their AWS Glue pipelines?

- **AWS Glue does not support custom code outside of PySpark and Scala implemented libraries**
- Give the custom code to the customer allowing the customers to upload the code onto AWS Glue and use within an ETL job.
- Upload the custom library as a .zip archive onto S3. Before your customers create an ETL job, include the S3 link as a script library and job parameter
- Upload all of the library files onto S3. Before your customers create an ETL job, include the S3 link as a script library and job parameter

Answer | Unflag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 36 of 65 questions.

QUESTION 38

What is a good target metric to use generally when comparing different regression models?

- Area Under the Curve (AUC)
- Root Mean Squared Error (RMSE)
- F1 score
- Recall

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 37 of 65 questions.

QUESTION 39

You are a machine learning specialist that needs to setup an ETL pipeline for your organization using Amazon Elastic Map Reduce (EMR). You must connect the EMR cluster to Amazon SageMaker without writing any specific code. Which framework allows you to achieve this?

- Apache Spark
- Apache Flink
- Apache Mahout
- Apache Pig
- Apache Hive

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 38 of 65 questions.

QUESTION 40

You are apply normalization techniques to a column in your dataset. The column has the following values {1, 5, 7}. When we apply normalization what will the respective output results be?

- {-1.00, -0.66, 0.00}
- {0.00, 0.66, 1.00}
- {-1.00, 0.66, 1.00}
- {0.00, -0.66, 1.00}
- {-1.33, 0.26, 1.06}

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 39 of 65 questions.

QUESTION 41

You have setup a group of SageMaker Notebook instances for your company's data scientists. You wanted to uphold your company's philosophy on least privilege and disabled Internet access for the notebooks. However, the data scientists report that they are unable to import certain key libraries from the Internet into their notebooks. What is the most efficient path?

- Create a series of EC2 instances outside of the VPC and install Jupyter Notebook on those instances. Have the scientists use those instances instead of SageMaker.
- Advise the data scientists that it is not possible to import libraries from the internet given the company's least privilege philosophy.
- Suggest that the scientists choose different libraries that are open source and do not pose a threat to company policy.
- Create a NAT gateway within the Notebook VPC and associated default route to the NAT gateway.
- Create a VPC Gateway Endpoint that bridges between the VPC and the desired Internet location of the required libraries.

Answer | Unflag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 40 of 65 questions.

← Back | Next →

QUESTION 42

What is the best way to split time series data when using Amazon Machine Learning?

- ○ Allow Amazon Machine Learning to trigger a Lambda function to split on upload
- ◉ Allow Amazon Machine Learning to split sequentially
- ○ Allow Amazon Machine Learning to randomize and split
- ○ Split before upload

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 41 of 65 questions.

QUESTION 43

You have developed a very complex deep learning model but your accuracy levels are still not at your desired target levels, even after hyperparameter optimization. What is the most likely cause?

- ○ Hyperparameter optimization is not performing as documented.
- ○ You did not employ a warm start method for your first optimization job.
- ◉ Hyperparameter tuning is not flawless and can still fail to converse on the best answer.
- ○ A random approach for hyperparameter optimization versus a Bayesian approach.

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 42 of 65 questions.

QUESTION 44

You have been tasked with collecting 100-byte events from hundreds or thousands of low power devices and writing records into a Kinesis stream. You have Amazon Elastic Compute Cloud (EC2) instances serving as a proxy for these events. You must add logic for batching or multithreading, in addition to retry logic and record de aggregation at the consumer side. Which service can you use to handle all of this for you?

- Using the Kinesis Producer Library (KPL)
- Using a combination of SQS and Lambda for retry logic and batching respectively.
- Using the APIs for Kinesis Streams
- Using the Kinesis Client Library (KCL)
- Using EMR cluster as intermediate logic mechanism
- Using the Amazon Kinesis Agent

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 43 of 65 questions.

QUESTION 45

You are a machine learning specialist building a model to determine the location (latitude and longitude) from different images taken and posted on a social media site. You've been provided with millions of images to use for training stored in Amazon S3. You've written a Java script to read the images from Amazon S3, extract pixels, latitude and longitude data into CSV format to train the model with. Which service is the best candidate to distribute the workload and create the training dataset?

- Amazon EMR
- Amazon Athena
- SageMaker GroundTruth
- AWS Glue

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 44 of 65 questions.

QUESTION 46

You are working for a company with strict compliance and data security requirements that requires that data is encrypted at all times, including at rest and in transit within the AWS cloud. You have been tasked with setting up a streaming data pipeline to move their data into the AWS cloud. What combination of tools allows this to be implemented with minimum amount of custom code?

- Encrypt data with Amazon Kinesis Consumer Library (KCL), decrypt data with the Amazon Kinesis Producer Library (KPL), and use AWS KMS to manage keys
- (●) Encrypt data with the Amazon Kinesis Producer Library (KPL), decrypt data with Amazon Kinesis Consumer Library (KCL), and use AWS KMS to manage keys
- Kinesis only supports encryption of data once it is loaded into AWS cloud. Use Kinesis Data Analytics and a KMS key to transform and decrypt data before using Kinesis Firehose to output results into Amazon S3
- Use SHA-256 custom managed keys to encrypt data before using the PutRecord or PutRecords API call. Use Kinesis Data Analytics to transform and decrypt data before using Kinesis Firehose to output results into Amazon S3

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 45 of 65 questions.

← Back Next →

QUESTION 47

You have setup autoscaling for your deployed model using SageMaker Hosting Services. You notice that in times of heavy load spikes, it takes a long time for the hosted model to scale out in response to the load. How might you increase the reaction time of auto-scaling?

- Disable CloudWatch advanced tracking metrics.
- (●) Change the scale metric from InvocationsPerInstance to MemoryUtilization.
- Create a new target metric based on time since last scale event.
- Reduce the cooldown period for automatic scaling.
- Change the timeout in the auto-scaling Lambda function.

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 46 of 65 questions.

QUESTION 48

Which of the following is NOT a valid use-case for incremental training?

- Train several variants of a model, either with different hyperparameter settings or using different datasets.
- Rebuilt model artifacts which you have accidentally deleted.
- **Train a new model using an expanded dataset that contains an underlying pattern that was not accounted for in the previous training and which resulted in poor model performance.**
- Resume a training job that was stopped.
- Use the model artifacts or a portion of the model artifacts from a popular publicly available model in a training job. You don't need to train a new model from scratch.
- Train several variants of a model, either with different hyperparameter settings or using different datasets.

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 47 of 65 questions.

← Back Next →

QUESTION 49

You are helping a customer with some significant modernization efforts. They want to implement a way to forecast future production demand based on historical data. However, they do not presently have budget for a full-time data scientist or machine learning expert. What might you recommend in this situation?

- Recommend they investigate Amazon Personalize
- Recommend they send a supply chain planner to get a degree in Data Science.
- Recommend they deploy a model based on DeepAR using competitor data that was offered by an ex-employee.
- Recommend they continue to pay you indefinitely while you develop and adjust a linear regression forecasting model.
- **Recommend they investigate Amazon Forecast**

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 48 of 65 questions.

QUESTION 50

A ski equipment company is trying to predict the expected sales from a line of ski goggles. They have never sold this type of product before, but they do have some historic sales data for other products which they believe have similar market adoption curves. What would be your first algorithm of choice among the built-in SageMaker algorithms for this use-case?

- Factorization Machines
- DeepAR
- Linear Learner
- K-Nearest Neighbor
- XGBoost

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 49 of 65 questions.

QUESTION 51

You are consulting for a large intelligence organization that has very strict rules around how data must be handled. One such rule is that data cannot be allowed to transit the public internet. What might you suggest as they are setting up SageMaker Notebook instances?

- VPC Log Monitoring
- AWS CloudTrail
- AWS Macie
- API Gateway
- VPC Interface Endpoints
- Route 53 Weighted Routing

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 50 of 65 questions.

QUESTION 52

You have prepared a machine learning model using Amazon SageMaker and have the model artifact stored onto S3. The client you are working for wants to have to ability to run new datasets into your implemented model and make inferences on the entire dataset. The client is not worried about sub second latency, as these jobs will run late at night. What is the best option to use to achieve this?

- Setup Lambda function and SQS to batch the inference calls for the entire dataset. Use SageMaker hosted endpoint to make batch inference calls on queued data.
- Setup API gateway endpoint that triggers a Lambda function. Setup the Lambda function to call the SageMaker hosted endpoint to make batch inference calls on entire dataset
- Setup SageMaker hosted endpoint to make batch inference calls on entire dataset
- Create a batch transform job using the trained model and the entire dataset
- Create a Amazon Machine Learning Batch prediction using the Amazon ML model and a set of input observations

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 51 of 65 questions.

← Back Next →

QUESTION 53

Your organization has the need to set up a petabyte scaled BI and dashboard analysis tool that will query millions of rows of data spread across thousands of files stored in S3. Your organization wants to save as much money as possible. Which solution will allows developers to run dozens if not hundreds or thousands of queries per day, and possibly scanning many TBs of data each, while still being cost effective?

- Data Pipeline and RDS
- AWS Glue Data Catalog and Amazon Athena
- EC2 Spot instances and Presto
- S3 Analytics

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 52 of 65 questions.

QUESTION 54

A machine learning specialist is running a training job on a single EC2 instance using their own Tensorflow code on a Deep Learning AMI. The specialist wants to run distributed training and inference using SageMaker. What should the machine learning specialist do?

- (●) Use Tensorflow in SageMaker and edit your code to run using the SageMaker Python SDK
- ( ) It is not possible to run custom Tensorflow code in SageMaker
- ( ) Ensure both the SageMaker Notebook instance and EC2 instance have the same role assigned to them. Use Notebook peering to gain access to run scripts from SageMaker on the EC2 instance
- ( ) Use Tensorflow in SageMaker and modify the AWS Deep Learning Docker containers
- ( ) Use Tensorflow in SageMaker and run your code as a script

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 53 of 65 questions.

QUESTION 55

When evaluating a model after the training and testing process, you notice that the error rate during training is high but the error rate during testing is low. Which of the following could be the reason for obtaining these error rates?

(Choose 2)

- [x] You should train for a longer period of time.
- [x] You have a data issue with both your training and testing datasets.
- [ ] Your model is overfitting the training data.
- [ ] You have a programmatic issue with your algorithm.
- [ ] Your model is underfitting the testing data.
- [ ] You need to re-evaluate the section of your algorithm.

Answer | Unflag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 54 of 65 questions.

QUESTION 56

You need to extract keywords from a collection of news stories. Which of the following algorithms could you use for this?

(Choose 2)

- [x] LDA
- [x] NTM
- [ ] RCF
- [ ] PCA
- [ ] WWF

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 55 of 65 questions.

QUESTION 57

You are a machine learning specialist evaluating a current model that has been deployed into production. It has been deployed for a few weeks now and the results are not accurate and sometimes the inference data is missing values. What are some techniques you can review to help solve this problem?

(Choose 3)

- [x] Ensure the extraction methods used to generate the training datasets are the same as for the production inference data.
- [ ] Ensure the inference data has placeholder values for any of the missing values
- [ ] Ensure the training data has a 50/50 distribution of the target attribute.
- [x] Ensure the training datasets are large, representative samples of the populations that the model needs to make predictions.
- [ ] Ensure the inference data is in the exact same for as the training and testing data
- [x] Ensure the target variable used as the predictor during training represents the actual outcome that the machine learning model is trying to predict.

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 56 of 65 questions.

← Back Next →

QUESTION 58

After training and validation sessions, you notice that the error rate is higher than expected for both sessions. What could you do to reduce the error rates for your model?
(Choose 3)

- [x] Run training for a longer period of time.
- [ ] Run a random cut forest algorithm on the data.
- [x] Gather more data for your training process.
- [ ] Encode the data using Laminar Flow Step-up
- [ ] Reduce the dimensions of the data.
- [x] Add more variables to the dataset

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 57 of 65 questions.

QUESTION 59

You are building out a machine learning model using multiple algorithms. You are at the point where you feel like one of the models is ready for production but you want to test difference variants of the model and compare the inference results in a testing environment before launching into production. What is the simplest way for you to test different model variants before launching into production.

- ( ) Use Amazon SageMaker to deploy the different versions of the model to a multiple endpoints. Use a Application Load Balancer to route a percentage of traffic to each model. Evaluate the results and use Route53 to route 100% of traffic to higher evaluated model.
- (•) Use Amazon SageMaker to deploy the different versions of the model to a single endpoint and route a percentage of traffic to each model. Evaluate the results and reroute 100% of traffic to higher evaluated model.
- ( ) Use Amazon SageMaker to deploy the different versions of the model to a multiple endpoints. Use a Network Load Balancer to route a percentage of traffic to each model. Evaluate the results and use Route53 to route 100% of traffic to higher evaluated model.
- ( ) Use multiple EC2 instances to deploy the model on Deep Learning AMIs. Evaluate the results and reroute 100% of traffic to higher evaluated model.

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 58 of 65 questions.

← Back Next →

QUESTION 60

You are preparing for the deployment of an ML model based on DeepAR deployed using SageMaker Hosting Services. This model is used for real-time inferences in the financial sector. Management is concerned about a damaged reputation if the service suffers an outage. Which of the following should you do to increase fault-tolerance?

- ○ Create a duplicate endpoint in another region using Amazon Forecast.
- ○ Recommend that they deploy using EKS in addition to the SageMaker Hosting deployment.
- ○ Keep a copy of all the DeepAR code in a Glacier Vault for safe keeping.
- ◉ Ensure that InitialInstanceCount is at least 2 or more in the endpoint production variant.
- ○ Include Elastic Inference in the endpoint configuration.

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 59 of 65 questions.

QUESTION 61

After a training and testing session, you notice that your training accuracy is 98% while you accuracy during testing was only 67%. What might you do to improve the model?
(Choose 2)

- ☑ Reduce the number of features being analyzed in the model.
- ☐ Rerun the training process with a larger learning rate.
- ☐ Change the approach from Linear Regression to Logistic Regression.
- ☑ Ensure the data was properly randomized before the split.
- ☐ Reduce the amount of data in the training and testing dataset.

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 60 of 65 questions.

QUESTION 62

A machine learning specialist is working with a dataset to create a model using a supervised learning algorithm. The specialist initially splits the data into two different sets, one for training and reserves the other for testing. The ratio of the dataset is broken into 80% training and 20% testing. After training the model the evaluation is yielding odd results. The model is 95% accurate for the training data and 64% accurate for the testing data. What is the reason for these odd results and what action needs to be taken to resolve issue?

- (●) The testing dataset has highly imbalanced labels. Reshuffle the data more evenly across both training and testing datasets.
- ( ) The model is currently underfitting on the dataset and does not know how to generalize for new data seen. This can be fixed by changing the hyperparameters to make the model (plus/minus the number of epochs), then retrain.
- ( ) Use accuracy as the objective metric for the training set and use a different objective metric to measure accuracy on the training set (F1, Precision, Recall).
- ( ) The model is currently underfitting on the dataset and does not know how to generalize for new data seen. This can be fixed by changing the hyperparameters to make the model (plus/minus the learning rate), then retrain.

Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 61 of 65 questions.

← Back Next →

QUESTION 63

You are trying to classify a number of items based on different features into one of 6 groups (books, electronics, movies, etc.) based on features. Which algorithm would be best suited for this type of problem?

- ( ) Use Linear Learner with predictor set to regressor
- ( ) Use a stochastic approach when choosing target parameters and recommended ranges
- ( ) Use regression forest with the number of trees set to the number of categories
- (●) Use K-Means algorithm with k set to the number of classes
- ( ) Use XGBoost with objective set to multi:softmax

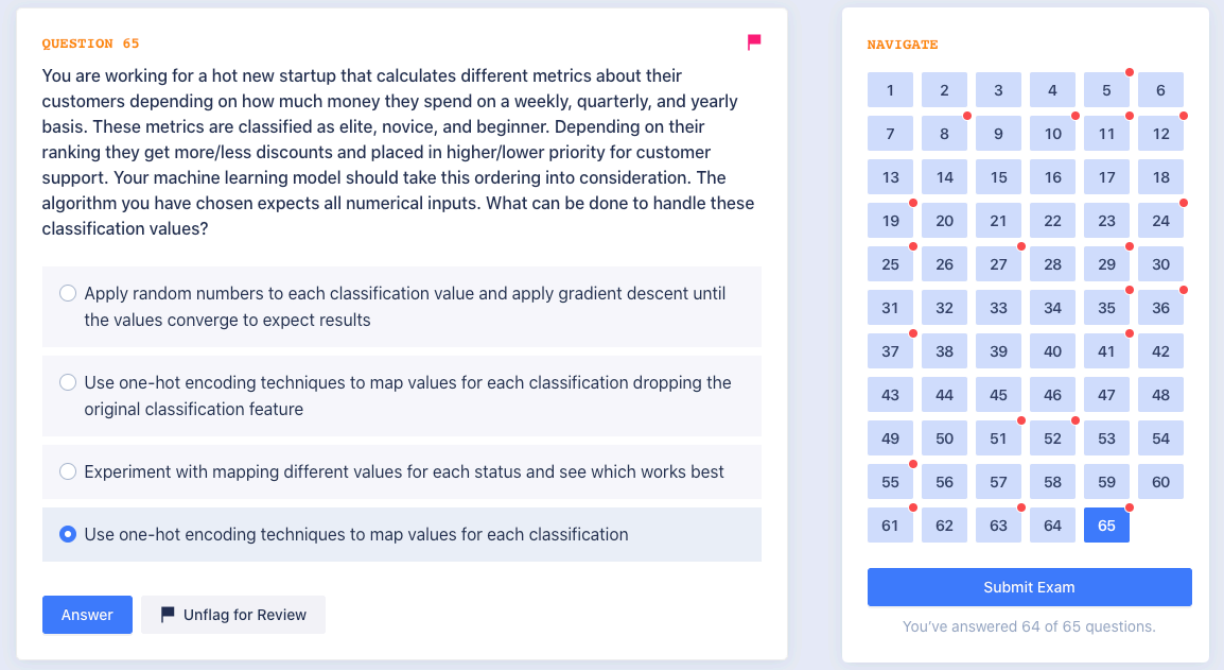
Answer | Flag for Review

NAVIGATE

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	

Submit Exam

You've answered 62 of 65 questions.



Time left aft 1st pass: 1h26 - so I have more than 1 hour to review my questions
20 questions flagged by me where I was not sure - I could go back to them but for this test, I just submitted to see my score

01:26:43

RESULTS: