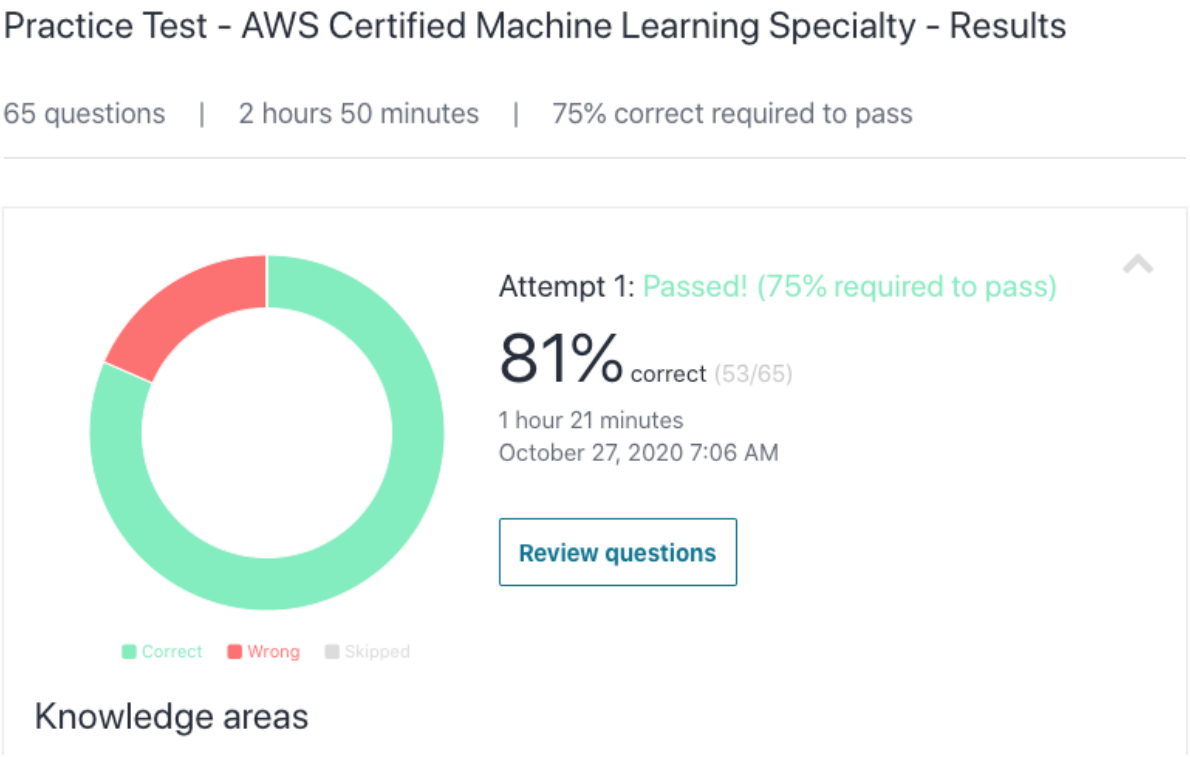
Udemy - Practice Test 2

Test from “AWS SageMaker and Certified ML Specialty Exam” Course
https://www.udemy.com/course/aws-machine-learning-a-complete-guide-with-python/

Passed on 10/27
https://www.udemy.com/course/aws-machine-learning-a-complete-guide-with-python/learn/quiz/4774522/test#overview



Knowledge areas

AWS (12 questions)

Machine Learning Concepts (33 questions)

AWS SageMaker, AI and Frameworks (20 questions)

Correct Wrong Skipped

Question 1:

You want to test new values for hyperparameters for an algorithm. At what point in the model lifecycle can you change hyperparameters?

- (●) Training
- ( ) Validation
- ( ) Hosting
- ( ) Testing

Question 2:

A model has the following errors: Training Error is 2%, Test Error is 5%. The benchmark is human-level performance, and the human error is 1%.

The model is:

- ○ Normal
- ○ Underfitting
- ○ Performing close to human-level performance
- ◉ Overfitting

Question 3:

You need to configure the SageMaker Endpoint to Scale on demand. Based on load testing, you have determined that one instance can handle 150 requests per second. Assume a safety factor of 0.5.

What value do you need to set for SageMakerVariantInvocationsPerInstance to trigger auto-scaling action?

Note: SageMakerVariantInvocationsPerInstance is a per minute metric.

- ◉ 4,500
- ○ 2,250
- ○ 18,000
- ○ 9,000

Question 4:

When training a deep learning network, what is the impact of using smaller mini-batch sizes?

- ○ Smaller mini-batch will force the algorithm to converge and get stuck in local minima
- ○ It will make smoother and more gradual adjustments to the weight
- ● It can help optimization algorithm jump local minima and explore other areas for global minima
- ○ Optimization algorithm uses all samples for every weight adjustment

A utility company wants to forecast water consumption per household. The historical data set contains the following attributes:

*** Year - Numeric**

*** Month - Numeric**

*** Floor Size SqFt – numeric**

*** Lot Size SqFt - numeric**

*** Number of Bathrooms – numeric**

*** Lawn – categorical with values YES or NO**

*** Consumption – numeric (target)**

To train using XGBoost, what data transformation step do you need to perform?

- ○ Transform non-numeric categories to equivalent numeric categories
- ○ One-Hot encode categorical features
- ○ Scale all numeric features to similar range and scale
- ● Normalize all numeric features

Question 5:

A utility company wants to forecast water consumption per household. The historical data set contains the following attributes:

*** Year - Numeric**

*** Month - Numeric**

*** Floor Size SqFt – numeric**

*** Lot Size SqFt - numeric**

*** Number of Bathrooms – numeric**

*** Lawn – categorical with values YES or NO**

*** Consumption – numeric (target)**

To train using XGBoost, what data transformation step do you need to perform?

- (●) Transform non-numeric categories to equivalent numeric categories
- ( ) One-Hot encode categorical features
- ( ) Scale all numeric features to similar range and scale
- ( ) Normalize all numeric features

Question 6:

You have launched an EC2 instance using Deep Learning AMI. Under AWS Shared Responsibility Model, who is responsible for applying critical security patches on EC2 instances?

- ( ) AWS
- ( ) EC2 Support
- (●) Customer
- ( ) AMI Provider

Question 7:

You are using SageMaker Automatic Hyperparameter tuning to search for optimal parameters for a learning algorithm.

What are the best practices when running a hyperparameter tuning job? (Choose three)

- [ ] Configure the tuning job to explore all hyperparameters supported by the algorithm
- [ ] Use Linear Scaling for hyperparameter that spans several orders of magnitude
- [x] Use Logarithmic Scaling for hyperparameter that spans several orders of magnitude
- [x] Configure the tuning job to search a smaller number of hyperparameters
- [x] Use fewer concurrent tuning jobs

Question 8:

When you increase the mini-batch size, for every iteration of the training set, the weights of features are adjusted

- ( ) More often
- ( ) Weight adjustment depends on the number of examples
- (•) Less often
- ( ) Weight adjustment is not dependent on mini-batch size

Question 9:

You are using AWS provided services for maintaining metadata about your data files stored in S3. The incoming files to S3 have additional attributes that are collected, and they are not showing up in the metadata. What is the recommended approach to address this issue?

- (●) Ensure Glue Crawlers are configured as a scheduled job to scan the files and update metadata
- ( ) Ensure Athena queries are scheduled to run periodically to update metadata
- ( ) Configure the Lambda function to monitor S3 and to capture the metadata changes
- ( ) Create a new table in the Glue Catalog to capture the changes

Question 10:

The training error is low, but the test error is high. Among the choices presented, which one of these options can correct the issue? (Choose Three)

- [x] Train with more data
- [ ] Increase the number of epochs
- [x] Increase Regularization
- [x] New neural network architecture
- [ ] Decrease regularization

Question 11:

A grocery store has a robust online presence. The store wants to improve product recommendations using machine learning and suggest products that are purchased together.

Which of these algorithms can be used for this requirement?

- ○ BlazingText
- ○ DeepAR
- ● Factorization Machines
- ○ Comprehend

Question 12:

An online marketplace wants to help customers make an informed choice when purchasing products. They would like to present the most positive and most critical customer reviews side-by-side in the product summary page.

Which capability can you use for this purpose?

- ○ Rekognition
- ○ Textract
- ● Sentiment Analysis with Comprehend
- ○ Custom Classification with Comprehend

I would use Comprehend to extract positive reviews

Question 13:

When training a deep learning model, if you increase the batch size, you should also

- (●) Increase the learning rate
- ( ) Learning rate and batch size are independent of each other
- ( ) Keep the learning rate same as batch size
- ( ) Decrease the learning rate

Question 14:

Which of these services require you to select an AWS region when using it (choose three)?

- [x] IAM
- [ ] S3
- [x] CloudWatch
- [x] SageMaker

Incorrect

- [x] IAM (Incorrect)
- [ ] S3 (Correct)
- [x] CloudWatch (Correct)
- [x] SageMaker (Correct)

Explanation

IAM is a global resource, and any policy or user or group or role that you create are available across all regions. With SageMaker, you need to pick a region to launch notebook instances, or for training and hosting models. S3 requires you to specify a region to create a bucket. CloudWatch is a repository of all metrics for monitoring

Question 15:

Which one of the services may be impacted when a single availability zone goes down in an AWS region?

- ( ) S3
- ( ) Artificial Intelligence Services like Rekognition
- ( ) SageMaker Endpoint with multiple instances
- (•) SageMaker Endpoint with a single instance

Question 16:

Which activation function would you use in the output layer for a Multi-class Classification neural network that predicts a single label from a set of possible labels?

- ○ None
- ◉ Softmax
- ○ Sigmoid
- ○ ReLU

Question 17:

You have a dense dataset with 1000s of features. You are using a custom training algorithm that has difficulty handling large datasets; you would like to reduce this dataset to a few important features.

The transformed dataset needs to retain as much information as possible from the original dataset.

What approach can you use for this problem?

- ◉ Reduce Dimension using Principal Component Analysis
- ○ Use algorithms like Factorization Machines that are optimized for very large datasets
- ○ Store data in Parquet format
- ○ Compress using GZIP algorithm

Question 18:

Under the AWS Shared Responsibility Model, the customer is responsible for which of these tasks?

- ○ Patching Host Operating System
- ○ Physical security of hardware
- ● Configuring Access to S3 bucket based on job role
- ○ Virtualization infrastructure

Question 19:

What privileges does a newly created Identity and Access Management (IAM) user have? This User does not have any policy attached and does not belong to any IAM Groups.

- ○ Read-only access in the region where IAM user was created
- ○ Read-only access to all resources in your account
- ○ Read-Write access to all resources in your account
- ● User cannot access AWS resources until explicit allow access is granted

Question 20:

The human level error rate is 2%, and the model training error rate is 8%. What steps can you take to optimize the model? (Choose Three)

- [ ] Increase regularization
- [x] Build a more complex model
- [x] Train longer
- [x] New neural network architecture

Question 21:

You are using CSV formatted files to train on SageMaker's built-in XGBoost algorithm.

SageMaker expects your training and validation to follow this convention:

- ( ) CSV must not have a column header record. Target variable must be the first column
- (•) CSV must have column headers with the target variable in the first column
- ( ) CSV must not have a column header record. Target variable must be the last column
- ( ) CSV must have column headers and target variable must be the last column

Incorrect

- CSV must not have a column header record. Target variable must be the first column (Correct)
- CSV must have column headers with the target variable in the first column (Incorrect)
- CSV must not have a column header record. Target variable must be the last column
- CSV must have column headers and target variable must be the last column

Explanation

With CSV format, SageMaker XGBoost expects the target variable in the first column and without a column header

Question 22:

You are working on developing a solution to identify specific breeds of cats and dogs from an image. The dataset you have is small. You noticed that an existing image classification neural network that was trained on a large dataset has an excellent ability to classify images. You would like to reuse the network to make it work for the new problem. What steps can you take to accomplish this?

- [x] Use Transfer learning by removing the output layer of the image classification model, reinitialize the weights of last hidden layer and retrain the model
- [ ] Retrain the image classification model with new data
- [ ] Use Transfer learning by removing the output layer of the image classification model, reinitialize the weights of all layers and retrain the model
- [ ] Use Transfer learning and remove the first hidden layer of image classification model and retrain the model

Question 23:

You are building a neural network for image analysis – What type of network would you use?

- [ ] Try different neural network architectures
- [ ] General Purpose Neural Network
- [ ] Recurrent Neural Network
- [x] Convolutional Neural Network

Question 24:

You are developing a deep learning network for converting speech to text. The dataset has recordings of 1,000 individuals, with everyone providing five different audio files along with the transcribed text. (for a total 5,000 audio samples). The trained model must generalize well for new individuals. How would you use this data for developing a model?

- ○ Ensure some individuals are only in the test set – use the remaining data for training and validation
- ○ For each individual, keep four audio files in the training set and one in the test set
- ○ Randomly split data between training and test set
- ◉ For each individual, keep three audio files in the training set, one in validation set and one in the test set

Incorrect

- ○ Ensure some individuals are only in the test set – use the remaining data for training and validation **(Correct)**
- ○ For each individual, keep four audio files in the training set and one in the test set
- ○ Randomly split data between training and test set
- ◉ For each individual, keep three audio files in the training set, one in validation set and one in the test set **(Incorrect)**

Explanation

The objective is to ensure the model generalizes well for unheard voices. So, the test set should not contain any individuals from the training or validation set. If we have the same individuals in the training and test set – the model may memorize voice for that individual and may artificially show improved performance. Reference: NIPS 2016 tutorial: Nuts and bolts of building AI applications by Dr. Andrew Ng.

Question 25:

A data scientist is working on a problem to classify incoming data into one of five categories: Good, DefectA, DefectB, DefectC, and DefectD. The dataset consists of primarily numeric features, and some of the samples have missing values for features. This missing values in features can help predict the defect class.

How do you train the model to learn from missing values?

- ○ Do nothing – algorithms can handle missing values if you provide examples in the training set
- ○ Replace missing values with the average value for that feature
- ○ Add substitute variables for each feature – when the feature has a missing value for a sample, set the substitute variable to 1 for that feature, and when the feature has a valid value, set the variable to 0
- ◉ Replace missing values with 0

Incorrect

- Do nothing – algorithms can handle missing values if you provide examples in the training set
- Replace missing values with the average value for that feature
- Add substitute variables for each feature – when the feature has a missing value for a sample, set the substitute variable to 1 for that feature, and when the feature has a valid value, set the variable to 0 **(Correct)**
- Replace missing values with 0 **(Incorrect)**

Explanation

Substitute variables are Boolean features that capture if a feature contains a missing value for the sample. This allows the algorithm to learn from missing values

https://docs.aws.amazon.com/machine-learning/latest/dg/data-insights.html#missing-values

https://docs.aws.amazon.com/machine-learning/latest/dg/data-insights.html#missing-values

Missing Values

The missing values report lists the attributes in the input data for which values are missing. Only attributes with numeric data types can have missing values. Because missing values can affect the quality of training an ML model, we recommend that missing values be provided, if possible.

During ML model training, if the target attribute is missing, Amazon ML rejects the corresponding record. If the target attribute is present in the record, but a value for another numeric attribute is missing, then Amazon ML overlooks the missing value. In this case, Amazon ML creates a substitute attribute and sets it to 1 to indicate that this attribute is missing. This allows Amazon ML to learn patterns from the occurrence of missing values.

Question 26:

A company has received an email from a customer with product feedback. Feedback is in an unknown language, and the company's product team has requested a German version of the email.

What steps are needed to accomplish this?

- ○ Transcribe to German with Source language set to auto-detect
- ○ Translate to English with source language set to auto-detect and then translate the output to German
- ○ Transcribe to English, Translate to German
- ● Translate to German with source language set to auto-detect

Question 27:

An organization is using TensorFlow Machine Learning Framework for building models and would like to migrate the machine learning infrastructure to AWS.

Which one of these options takes the least effort to train, host, and manage TensorFlow models in AWS?

- ○ Built custom docker image that conforms to SageMaker specification to develop and host models using SageMaker infrastructure
- ○ Launch EC2 instance with Deep Learning AMIs
- ● Use pre-built TensorFlow docker images provided by SageMaker to train and host models on SageMaker infrastructure
- ○ Launch EC2 instance, download and install required Machine Learning Frameworks

Question 28:

A machine learning specialist needs to get inference for the entire dataset that is stored in S3. The Machine Learning Model was trained on SageMaker.

Which of these options provides a managed infrastructure that is cost-effective for large scale inference?

- ○ Autoscaling
- ○ S3 Analytics
- ◉ SageMaker Batch Transform
- ○ SageMaker Endpoint

Question 29:

For a regression problem, which of these algorithms cap the output to a range of values seen in the training set? (Choose two)

- ☑ decision tree
- ☐ linear regression
- ☐ neural network
- ☑ xgboost

Question 30:

An Auto Show organizer wants to detect celebrities who are among the audience. The event center has several cameras that are recording the event live. What combination of service and order of processing can help achieve this task?

- ○ Kinesis Data Streams, Amazon Rekognition, Kinesis Video Stream
- ◉ Kinesis Video Streams, Amazon Rekognition, Amazon Data Stream
- ○ Kinesis Firehose, Lambda, and Amazon Rekognition
- ○ Kinesis Firehose, Kinesis Analytics, and Amazon Rekognition

Question 31:

A customer has 1000s of documents, and they would like to create a summary of each document.

Which of these services is best suited for this requirement?

- ○ Rekognition Text Extraction
- ○ Textract
- ◉ Comprehend
- ○ Transcribe

Question 32:

You have a collection of documents that has text about a variety of different topics: animals, plants, transportation, travel, food, and so forth. You want to train an algorithm to categorize the documents into one of the above categories.

Which of these algorithms can you use for this requirement?

- ○ Comprehend
- ◉ Seq2Seq
- ○ Neural Topic Modeling (NTM)
- ○ LDA

This is a classification problem
I would convert the text into embeddings and label each with a category and use seq2seq to find the vectors that are similar to pick the right category

Question 32: Incorrect

You have a collection of documents that has text about a variety of different topics: animals, plants, transportation, travel, food, and so forth. You want to train an algorithm to categorize the documents into one of the above categories.

Which of these algorithms can you use for this requirement?

- ○ Comprehend (Correct)
- ◉ Seq2Seq (Incorrect)
- ○ Neural Topic Modeling (NTM)
- ○ LDA

Explanation

LDA and NTM are used for topic modeling; however, they are unsupervised and generally used in exploratory setting for understanding data.

You have the flexibility to specify the number of topics – however, the algorithms automatically assign topics – it may not match with what we consider as topics: travel, food, transportation, and so forth. It will automatically generate appropriate topics.

For example, LDA/NTM may come with a topic that groups travel and food together.

For this problem, Comprehend service can be used to train a classifier that can map text content to a topic. Seq2Seq is used for translation, summarization and so forth

⭐ Question 33:

A team of machine learning experts is building a speech recognition system that can work in a noisy factory environment. The dataset consists of 10,000 hours of clean speech data and another dataset with 100 hours of noisy speech data recorded inside the factory.

How do you define training, validation, and test set? (Select Two)

- [ ] Split the 10,000 hours of clean speech data into training and validation set. Divide 100 hours of noisy speech data, add some to the validation set and keep the rest in the test set
- [ ] Split the 10,000 hours of clean speech data into training and validation sets. Optimize the model to improve validation performance. Use 100 hours of noisy data for final testing
- [x] Use 10,000 hours of clean speech data for training the model. Divide 100 hours of noisy data into validation and test sets. Optimize the model to improve validation performance and perform the final test using the test set
- [ ] Use 100 hours of noisy data for training and split the general speech data for validation and testing

I forgot to select one more answer!

A team of machine learning experts is building a speech recognition system that can work in a noisy factory environment. The dataset consists of 10,000 hours of clean speech data and another dataset with 100 hours of noisy speech data recorded inside the factory.

How do you define training, validation, and test set? (Select Two)

- [ ] Split the 10,000 hours of clean speech data into training and validation set. Divide 100 hours of noisy speech data, add some to the validation set and keep the rest in the test set **(Correct)**

- [ ] Split the 10,000 hours of clean speech data into training and validation sets. Optimize the model to improve validation performance. Use 100 hours of noisy data for final testing

- [x] Use 10,000 hours of clean speech data for training the model. Divide 100 hours of noisy data into validation and test sets. Optimize the model to improve validation performance and perform the final test using the test set **(Correct)**

Explanation

The objective of this model is to recognize speech in a noisy environment. Since there is very little noisy data available when compared to clean data, one approach that can be used is to train the model on clean data, split the noisy data into validation and test set. Use the noisy validation data to tune the model performance and perform the final check with test data.

Another option is to split the clean speech data into training and validation sets. Add some of the noisy data to the validation dataset and keep the remaining noisy data for the test set.

If you keep split the clean data into training and validation sets and tune model based on validation performance, this model only performs well with clean data and would perform poorly with noisy test data. That is because the distribution of clean and noisy data is different.

Just training on 100 hours of noisy data may not be enough for this use case.
Reference: NIPS 2016 tutorial: Nuts and bolts of building AI applications by Dr. Andrew Ng

Question 34:

A binary classifier metrics for validation data has the following values:

TP: 8, FN: 2, TN: 3, FP: 5

What is the Precision for this model?

- ○ 0.5
- ○ 0.3
- ○ 0.8
- ◉ 0.6

Question 35:

A dataset contains a large number of features. You would like the algorithm to aggressively prune features that are not relevant. What hyperparameter can you use for this?

- ○ Either L1 or L2 Regularization
- ○ Learning Rate
- ◉ L1 Regularization
- ○ L2 Regularization

Question 36:

A machine learning specialist needs to come up with an approach to automatically summarize the content of large text documents. Which algorithm can be used for this use case?

- ○ Random Cut Forest
- ○ K-Means
- ○ Seq2Seq
- ◉ LDA

Incorrect

- ○ Random Cut Forest
- ○ K-Means
- ○ Seq2Seq (Correct)
- ◉ LDA (Incorrect)

Explanation

Seq2Seq algorithm is used for text summarization – It accepts a series of tokens as input and outputs another sequence of tokens. LDA is an unsupervised algorithm for topic modeling – it can generate probabilities of a document belonging to a number of specified topics. K-Means is a clustering algorithm that is used for identifying grouping within data. Random Cut Forest is used for detecting anomalous data points

⭐ Question 37:

A dataset consists of following features along with the type of values it can contain

*** DayOfWeek – Sunday, Monday, Tuesday and so forth**

*** Holiday – True or False**

*** Temperature – in Fahrenheit**

*** Humidity – 0 to 100**

*** Precipitation – 0 to 100**

*** Windspeed – 0 to 150**

*** Pollen – 0 to 1**

*** AirQuality – Good, Bad**

AirQuality is the label

The Machine Learning Analyst is planning to compare a variety of algorithms and would like to reuse the same transformed dataset for training and testing.

What data transformation is recommended? (Select Three)

- [ ] Use numeric data without any transformation, and one hot encode categorical features
- [ ] Transform using Principal Component Analysis
- [x] One-Hot encode Day of Week
- [x] Label encode AirQuality and Holiday features
- [x] Scale Temperature, Humidity, Precipitation, Windspeed, Pollen features

Question 38:

A customer is using Polly to generate audio for text. However, Polly is not pronouncing some of the words correctly. What option would help you control the speech output?

- ○ Use batch streaming for highest quality outputs
- ◉ Use Speech Synthesis Markup Language
- ○ Use correct Region and Language
- ○ Use real-time streaming for highest quality output

Question 39:

You are using SageMaker's Automatic Hyperparameter tuning to find an optimal set of parameters for a deep learning network. You are using the Bayesian search with a maximum number of training jobs set to 100. What is the recommended amount of concurrent tuning jobs that you can run for the best results?

- ○ 32
- ○ 4
- ◉ 1
- ○ 100

Question 40:

A binary classifier metrics for validation data has the following values:

TP: 8, FN: 2, TN: 3, FP: 5

What is the Recall for this model?

- ○ 0.3
- ○ 0.5
- ○ 0.6
- ◉ 0.8

Question 41:

A startup is analyzing social media trends with data stored in S3. For analysis, it is common to access a subset of attributes across a large number of records. Which of these formats can lower the cost of storage while improving query performance?

- ○ CSV
- ○ Avro
- ◉ Parquet
- ○ JSON

Question 42:

For a binary classification problem, the cost of misclassifying a positive sample is three times more than the cost of misclassifying a negative example.

Which model has the lowest cost with at least 60% recall?

Model 1 – TP: 10, FN: 5, TN: 25, FP: 10

Model 2 – TP: 5, FN: 10, TN: 20, FP: 15

Model 3 – TP: 1, FN: 14, TN: 30, FP: 5

Model 4 – TP: 9, FN: 6, TN: 20, FP: 15

- (●) Model 1
- ( ) Model 4
- ( ) Model 2
- ( ) Model 3

We need to minimize recall yet keep it above 60%
Model1 => Recall = TP / (TP+FN) = 0.666
Other models are below 60%

Question 43:

Training data has values for all features. With Test data, some of the features have missing values. If you build a neural network with training data and use test data to verify performance, how would the neural network behave?

- (●) The response depends on activation function
- ( ) The network would automatically learn insights about missing values
- ( ) The network would not learn from missing values
- ( ) Behavior depends on the number of layers

- The response depends on activation function (Incorrect)
- The network would automatically learn insights about missing values
- The network would not learn from missing values (Correct)
- Behavior depends on the number of layers

Explanation

The system would not learn insights from missing values. You would need to create new examples in training data with missing values so that the model can learn to ignore missing values

Question 44:

You need to read the CSV files in S3, transform the content to Parquet format, and store the processed data back in S3. Which of these options is recommended for this solution?

- Use Kinesis Datastreams for collecting the data from S3 and use built-in transformation to store the results in Parquet format
- Use Kinesis Firehose for reading the data from S3 and use built-in transformation to store the results in Parquet format
- Configure S3 to invoke Lambda function when a new file is added, perform the transformation in Lambda, and store the results back in S3
- Use Glue ETL to run Spark ETL scripts and configure it as a scheduled job

Firehose has built-in transform to parquet. But it works against streams of data.

Glue works against bath and also has built-in

Question 45:

A manufacturing company has a collection of images that contains examples of normal and defective products. These images need to be manually labeled by human experts for model training, and they need a solution to manage the workflow to distribute images among human experts for manual labeling.

What capability can you for this?

- ○ ImageClassification
- ○ SageMaker Neo
- ● SageMaker GroundTruth
- ○ Rekognition

Question 46:

A labeled dataset contains a lot of duplicate examples. How should you handle duplicate data?

- ○ Ensure all duplicates are in train data
- ○ Ensure all duplicates are in test data
- ○ Ensure data is shuffled before creating train and test set
- ● Ensure there are no duplicates

Question 47:

You are working on a model to differentiate positive and negative classes – the dataset that was provided to you is highly unbalanced. 99% of the data is normal, with only 1% positive. What steps can you go through to handle this unbalanced dataset? (select two)

- [ ] Oversample by duplicating positive data
- [x] Collect more positive samples
- [ ] Use Accuracy as a measure for the unbalanced dataset
- [ ] Use ROC AUC as a metric for the unbalanced dataset
- [x] Oversample positive data using techniques like SMOTE

Question 48:

A machine learning specialist is using a SageMaker algorithm to train a model. The dataset is large, and the training job is distributed across multiple training instances. What mechanism does SageMaker provide to minimize temporary storage required in the training instance volumes?

- (•) Pipe Mode
- ( ) SageMaker does not copy data to local instance volumes – all data resides in S3
- ( ) Explore compressed storage
- ( ) File Mode

Question 49:

You are using a lambda function to invoke SageMaker Endpoints. This function can accept a batch of records as input and returns the list of predicted values. You are testing a new model that requires compute-intensive pre-processing of incoming data. You want to use a higher-performing instance for your lambda function. What option does AWS provide to improve performance?

- ○ Use a compute-optimized instance
- ○ Increase allocated memory
- ○ Increase timeout
- ◉ Increase allocated vCPU

Incorrect

- ○ Use a compute-optimized instance
- ○ Increase allocated memory (Correct)
- ○ Increase timeout
- ◉ Increase allocated vCPU (Incorrect)

Explanation

With Lambda, you must choose the amount of memory needed to execute your function. Based on the memory configuration, proportional CPU capacity is allocated. You can also increase the timeout for up to a maximum of 15 minutes.

Question 50:

You are using unigram text transformation to convert words to the frequency of occurrence. There are two sentences in the text.

"this is working - not disappointed"

"this is not working - disappointed"

How many features would the transformed dataset have?

- ○ 10
- ○ 5
- ○ 8
- ◉ 6

I am counting the "-" sign
[This, is, not, working, -, disappointed]
Incorrect

- ○ 10
- ○ 5 (Correct)
- ○ 8
- ◉ 6 (Incorrect)

With unigram transformation, each unique word is a feature. There are five unique words: disappointed, is, not, this, working. With bigram transformation, you need to include consecutive two-word combinations like "this is", "is working" and so forth.

https://en.wikipedia.org/wiki/N-gram

The items can be phonemes, syllables, letters, words or base pairs according to the application.

Question 51:

A data scientist has a large dataset that needs to be trained on the AWS SageMaker service. The training algorithm is optimized for GPU processing and can benefit from substantial speed-up when trained on instances with GPUs. Which instance family can you use for a training job for the best performance?

- ○ Compute Optimized family
- ◉ Accelerated Computing family
- ○ General Purpose family
- ○ Memory-Optimized family

Question 52:

Your company has a portfolio of machine learning models that are used by web applications and mobile apps. What is the best mechanism to integrate machine learning models with your application? The solution also needs to scale on demand.

- ○ Use Lambda function to invoke machine learning models and invoke the Lambda function from the client application
- ◉ API Gateway, Lambda, SageMaker Endpoint with Auto Scaling
- ○ Invoke Machine Learning model endpoint from your Client application
- ○ Host your models on EC2 web server instances, and load balance using Elastic Load Balancing. Setup autoscaling to scale web servers

Question 53:

You are exploring different parameters for tuning the model. What dataset should you use to guide with this tuning exercise?

- ○ Train
- ○ Use a random sample from train, validation and test sets
- ○ Test
- ● Validation

This is about parameters - we use validation metrics to tune it during Training time

Question 54:

A highly unbalanced dataset has 95% normal data and 5% positive data. What is a good performance metric to use for assessing the quality of the model?

- ● F1 Score
- ○ Recall
- ○ Accuracy
- ○ Precision

This is for a classification exercise.
However I don't know how "normal" compares to "positive" data
I can assume the normal data is "negative"
And we only have 5% positive.
F1-score can be a good metric when we are not sure if we need to minimize FN or FP
We don't know from the business if we need to optimize TP or FN...
Keep F1-score

Question 55:

Your company uses S3 for storing data collected from a variety of sources. The users are asking for a feature similar to a trash can or recycle bin. Deleted files should be available for restore for up to 30 days. How would you implement this? (Choose Two)

- [x] Enable Lifecycle Policies on the bucket
- [x] Enable Versioning on the bucket
- [ ] Enable Cross-Region Replication and restore objects from the replicated site
- [ ] Move the deleted object to a temporary bucket and use it for restoring

Question 56:

Your legal department has asked your team to ensure that historical manufacturing data are not deleted or tampered for a 5-year period. Your team is currently using Glacier for long term storage. What option would you pick to enforce this policy?

- ( ) Use Vault Lock to implement write once, read many type policies
- ( ) Enforce controls like these at the application level
- ( ) Replicate Data to another read-only bucket
- (•) Implement IAM Access Policy to remove delete access or modify access

Incorrect

- [ ] Use Vault Lock to implement write once, read many type policies **(Correct)**
- [ ] Enforce controls like these at the application level
- [ ] Replicate Data to another read-only bucket
- [x] Implement IAM Access Policy to remove delete access or modify access **(Incorrect)**

Explanation

Vault Lock allows you to set immutable policies to enforce compliance controls. With the IAM Access policy, you can define who has access to storage and type of access. However, the IAM policy on its own is not sufficient for compliance-related controls as someone could change the policy to grant write permissions

https://docs.aws.amazon.com/amazonglacier/latest/dev/vault-lock.html

Vault Locking Overview

S3 Glacier Vault Lock allows you to easily deploy and enforce compliance controls for individual S3 Glacier vaults with a vault lock policy. You can specify controls such as "write once read many" (WORM) in a vault lock policy and lock the policy from future edits. Once locked, the policy can no longer be changed.

S3 Glacier enforces the controls set in the vault lock policy to help achieve your compliance objectives, for example, for data retention. You can deploy a variety of compliance controls in a vault lock policy using the AWS Identity and Access Management (IAM) policy language. For more information about vault lock policies, see Amazon S3 Glacier Access Control with Vault Lock Policies.

Question 57:

An organization is consolidating data in S3, and data scientists need access to this data for initial exploration. They are well versed in SQL and would prefer to access the data in S3 using SQL. Which of these options provides the lowest cost without requiring to provision any servers?

- ○ EMR Hive
- ○ EMR Spark
- ◉ Athena
- ○ Redshift Spectrum

Question 58:

A company has several audio files that must be converted to other languages.

What is the best way to complete this task?

- ○ Transcribe, Polly, Translate
- ○ Translate
- ○ Translate, Polly
- ◉ Transcribe, Translate, Polly

Question 59:

An organization has human experts who perform manual classification of products by visual inspection. A Machine Learning specialist is building a classification system to match human-level performance. When reviewing the error rate of humans, the specialist observes the following:

Newly trained employees had a misclassification error rate of 5%, Experienced employee had an error rate of 2.5%, and when a team of experienced employees worked together, they had a misclassification rate of 1%.

What should be considered as human-level performance?

- ○ 2.5%
- ○ 5%
- ● 1%
- ○ Average of the error rates

Question 60:

You have a requirement to convert temperature from Celsius to Fahrenheit. You have a dataset of a few hundred rows that contain examples of Celsius and equivalent Fahrenheit. These are results observed using different approaches.

Which option would you pick?

- ○ When using XGBoost Regression algorithm, it easily handles this dataset with very low RMSE error on the validation dataset
- ○ When using Linear Regression algorithm, it easily handles this dataset with very low RMSE error on the validation dataset
- ○ Use either Linear Regression or XGBoost
- ● Instead of using Machine Learning, implement the logic in code as the conversion logic is simple

Question 61:

You are training a model to predict the probability of leaving the mobile operator. You would like to assess the quality of the metrics at various cut-off thresholds. Which metric gives you insight into the model performance over a range of tradeoffs between true positive rate and false-positive rate?

- ○ Squared Error
- ○ Accuracy
- ◉ ROC AUC Metric
- ○ F1 Score

Question 62:

A binary classifier metrics for validation data has the following values:

TP: 8, FN: 2, TN: 4, FP: 5

How many positive and negative samples are there in the validation dataset?

- ◉ Positive: 10, Negative: 9
- ○ Positive: 13, Negative: 6
- ○ Positive: 12, Negative: 7
- ○ Positive: 6, Negative: 13

TP+FN = Actual Positive = 8 + 2 = 10
FP+TN = Actual Negative = 5 + 4 = 9

Question 63:

A team of students is building an application that can blur or remove unwanted objects in an image. Users can pick the objects on which action needs to be performed.

Which one of the AWS machine learning capabilities can you use for this?

- ○ Rekognition
- ○ ImageClassification
- ● Semantic Segmentation
- ○ ObjectDetection

Semantic segmentation - pixel level object detection => great for shapes of object

Question 64:

A Machine Learning Expert is working on a time series forecasting problem to predict future demand for products. The dataset consists of two years' worth of historical data. What is the recommended way to split the training and test set?

- ○ Shuffle data and perform a random split to keep 80% for training and 20% for testing
- ○ Split data in such a way that first 80% of the days in a month are part of the training set and the remaining 20% of each month is set aside in the test set
- ● Order data by time and set aside first 80% for training and the remaining 20% for testing
- ○ Split data into 80% for training and 20% for testing

Question 65:

A data scientist is exploring the use of the XGBoost algorithm for a regression problem.

The dataset consists of numeric features.

Some of the features are highly correlated, and almost all the features are on different orders of magnitude.

What data-transformation is required to train on XGBoost?

- (●) Scaling
- ( ) Data transformation is not needed for this dataset
- ( ) Remove one feature from every highly correlated feature pairs
- ( ) Normalization

Incorrect

- (●) Scaling (Incorrect)
- ( ) Data transformation is not needed for this dataset (Correct)
- ( ) Remove one feature from every highly correlated feature pairs
- ( ) Normalization

Explanation

Decision Tree-based algorithms like XGBoost automatically handles correlated features, numeric features on a different scale, and numeric-categorical variables. Other algorithms like a neural network and the linear model would require features on a similar scale and range, and you need to keep only one feature in every highly correlated feature pairs and one-hot encode categorical features.