Exam Readiness: AWS Certified Machine Learning - Specialty
Study/Quiz Question Solutions & Class Averages

Class Scores on Comprehensive Study/Practice Questions

Below are the running averages for all students of the ML Exam Readiness course that have taken the comprehensive study/practice questions found at the end of the course. Averages are broken out by overall score and domain scores. You are free to compare your performance on these study questions to the below averages.

-	Value
Average Overall Score (%)	64.9
Average Domain 1 Score (%)	65.7
Average Domain 2 Score (%)	67.1
Average Domain 3 Score (%)	66.1
Average Domain 4 Score (%)	59.3

Double click to edit

Double click to edit

Double click to edit

Q3. An analytics company wants to use a fully managed service that automatically scales to handle the transfer of its Apache web logs, syslogs, text and videos on their webserver to Amazon S3 with minimum transformation.

What service can be used for this process?

1. Kinesis Data Streams **(Kinesis Data Streams is not a fully managed service and therefore does not meet one of the requirements of this question.)**
2. Kinesis Firehose*
3. Kinesis Data Analytics **(Kinesis Data Analytics is not an appropriate service here, because it helps with streaming analytics.)**
4. Amazon Kinesis Video Streams **(Kinesis Video Streams is not an appropriate service here, because it securely streams videos from devices.)**

Q4. A video streaming company wants to analyze its VPC flow logs to build a real-time anomaly detection pipeline. The pipeline must be minimally managed and enable the business to build a near real-time dashboard.

What combination of AWS service and algorithm can the company use for this pipeline?

1. Amazon SageMaker with RandomCutForest **(Amazon SageMaker cannot be used with streamed data.)**
2. Kinesis Data Analytics with RandomCutForest*
3. Amazon QuickSight with ML Insights **(Amazon QuickSight can only be used with structured datasets that need to be stored in Amazon S3 or a database.)**
4. Apache Spark on Amazon EMR with MLLib **(Amazon EMR can be used for this use case, but it doesn't satisfy the minimally managed requirement.)**

Q5. A data and analytics company is expanding its platform on AWS. The company wants to build a serverless product that preprocesses large structured data, while minimizing the cost for data storage and compute. The company also wants to integrate the new product with an existing ML product that uses Amazon EMR with Spark.

What solution should the company use to build this new product?

1. Use AWS Lambda for data preprocessing. Save the data in Amazon S3 in CSV format. **(AWS Lambda has a runtime of 15 minutes, making it less than ideal for this particular situation. Additionally, saving the data in CSV format will not meet the question's cost requirements.)**
2. Use AWS Glue for data preprocessing. Save the data in Amazon S3 in CSV format. **(AWS Glue will work as a solution for data preprocessing, but saving the data in CSV format does not fulfull the company's cost requirements.)**
3. Use AWS Glue for data preprocessing. Save the data in Amazon S3 in Parquet format.*
4. Use AWS Lambda for data preprocessing. Save the data in Amazon S3 in Parquet format. **(Lambda has a runtime of 15 minutes, making it less than ideal for this particular situation.)**

Q6. A financial organization uses multiple ML models to detect irregular patterns in its data to combat fraudulent activity such as money laundering. They use a TensorFlow-based Docker container on GPU-enabled Amazon EC2 instances to concurrently train the multiple models for this workload.

However, they want to automate the batch data preprocessing and ML training aspects of this pipeline, scheduling them to take place automatically every 24 hours.

What AWS service can they use to do this?

1. AWS Glue **(AWS Glue cannot import a Docker container with TensorFlow to be used in the pipeline.)**
2. AWS Batch*
3. Amazon EMR **(Amazon EMR can be used for this use case, but the pipeline will not be automatically scheduled)**
4. Kinesis Data Analytics **(Kinesis Data Analytics can be used only on streaming data.)**

Q7. A real estate startup wants to use ML to predict the value of homes in various cities. To do so, the startup's data science team is joining real estate price data with other variables such as weather, demographic, and standard of living data.

However, the team is having problems with slow model convergence. Additionally, the model includes large weights for some features, which is causing degradation in model performance.

What kind of data preprocessing technique should the team use to more effectively prepare this data?

1. Standard scaler* **(Standard scaler is the best option, because it performs scaling and shifting/centering.)**
2. Normalizer **(This would perform row normalization. This situation requires *column* normalization.)**
3. Max absolute scaler **(This would scale each column by its max value, but would not shift/center the data.)**
4. One hot encoder **(There is no symbolic/string data being mentioned here to perform 1-hot encoding on.)**

Q8. A Data Scientist at a retail company is using Amazon SageMaker to classify social media posts that mention the company into one of two categories: Posts that require a response from the company, and posts that do not. The Data Scientist is using a training dataset of 10,000 posts, which contains the timestamp, author, and full text of each post.

However, the Data Scientist is missing the target labels that are required for training.

Which approach can the Data Scientist take to create valid target label data? (Select TWO.)

1. Ask the social media handling team to review each post using Amazon SageMaker GroundTruth and provide the label*
2. Use the sentiment analysis natural language processing library to determine whether a post requires a response **(Sentiment analysis would not directly create a binary label.)**
3. Use Amazon Mechanical Turk to publish Human Intelligence Tasks that ask Turk workers to label the posts*
4. Use the *a priori* probability distribution of the two classes. Then, use Monte-Carlo simulation to generate the labels **(It's not clear how this approach would assign the binary classification label that is required by this question.)**
5. Use K-Means to cluster posts into various groups, and pick the most frequent word in each group as its label **(This creates labels, but those labels will not align to the required two categories mentioned in this question.)**

Q9. A Data Scientist wants to include “month” as a categorical column in a training dataset for an ML model that is being built. However, the ML algorithm gives an error when the column is added to the training data.

What should the Data Scientist do to add this column?

1. Convert the “month” column to 12 different columns, one for each month, by using one-hot encoding.*
2. Map the “month" column data to the numbers 1 to 12 and use this new numerical mapped column. **(This is not a good option, because the numerical mapping of the months would imply magnitudes of difference between the different months (for instance, April could be looked at as twice as much as February).)**
3. Scale the months using *StandardScaler*. **(*StandardScaler* is used to scale numerical data. It will will not work with categorical data, which is what this question is about.)**
4. Use pandas *fillna()* to convert the column to numerical data. **(This approach, which deals with missing data, is not relevant to this question.)**

Q 10. A Data Scientist for a credit card company is creating a solution to predict credit card fraud at the time of transaction. To that end, the Data Scientist is looking to create an ML model to predict fraud and will do so by training that model on an existing dataset of credit card transactions. That dataset contains 1,000 examples of transactions in total, only 50 of which are labeled as fraud.

How should the Data Scientist deal with this class imbalance?

1. Use the Synthetic Minority Oversampling Technique (SMOTE) to oversample the fraud records* **(Instead of undersampling the major class, SMOTE is oversampling the minor class synthetically, which makes it the best solution for this situation.)**
2. Undersample the non-fraudulent records to improve the class imbalance **(This approach essentially requires throwing away data, which is definitely not a good solution given the small dataset in this question.)**
3. Use K-fold cross validation when training the model **(This is a good evaluation technique, but will not improve the model's capability to differentiate between the classes.)**
4. Drop all the fraud examples, and use a One-Class SVM to classify **(This artificially throws away real data, and one class methods are useful for anomaly detection, not for binary classification as is the case here)**

Q11. An ML Engineer at a real estate startup wants to use a new quantitative feature for an existing ML model that predicts housing prices. Before adding the feature to the cleaned dataset, the Engineer wants to visualize the feature in order to check for outliers and overall distribution and skewness of the feature.

What visualization technique should the ML Engineer use? (Select TWO.)

1. Box Plot*
2. Histogram*
3. Scatterplot **(Scatterplot can help check for outliers, but it won't show the skewness of the data.)**
4. Heatmap **(Heatmaps show relationships between two variables, but is not enough to check for overall distribution or skewness in the data.)**
5. T-SNE **(T-SNE is used to reduce the dimensionality of the data. It is not used to visualize outliers.)**

Q12. A company is using its genomic data to classify how different human DNA affects cell growth, so that they can predict a person's chances of getting cancer. Before creating and preparing the training and validation datasets for the model, the company wants to reduce the high dimensionality of the data.

What technique should the company use to achieve this goal? (Select TWO.)

1. Use seaborn distribution plot (distplot) to visualize the correlated data. Remove the unrelated features. **(This does not show correlation between features, and does not perform dimension reduction.)**
2. Use T-SNE to reduce the dimensionality of the data. Visualize the data using matplotlib.*
3. Use Principle Component Analysis (PCA) to reduce the dimensionality of the data. Visualize the data using matplotlib.*
4. Calculate the eigenvectors. Use a scatter matrix to choose the best features. **(Eigenvectors cannot reduce the dimensionality of the data.)**
5. Use L2 regularization to reduce the features used in the data. Visualize the data using matplotlib. **(L2 regularization is not a feature reduction technique—although for linear models, L1 regularization can act like a feature reduction technique.)**

Q 13. A Data Scientist wants to create a linear regression model to train on a housing dataset to predict home prices. As part of that process, the Data Scientist created a correlation matrix between the dataset's features and the target variable. The correlations between the target and two of the features, feature 3 and feature 7, are 0.64 and -0.85, respectively.

Which feature has a stronger correlation with the target variable?

1. Feature 3*
2. Feature 7 **(Feature 7 has a negative correlation with the target variable, and even though the magnitude is higher than the correlation with Feature 3, the question asks for a *stronger* correlation, which is associated with a positive correlation.)**
3. There is not sufficient enough data to determine which variable has a stronger correlation to the target **(There is sufficient data. Even though the magnitude of the correlation with Feature 7 is higher, "stronger" is associated with a positive correlation.)**
4. Feature 7 and feature 3 both have weak correlations to the target **(This is not true, as 0.64 is not considered very weak.)**

Q 14. A video streaming company is looking to create a personalized experience for its customers on its platform. The company wants to provide recommended videos to stream based on what other similar users watched previously. To this end, it is collecting its platform's clickstream data using an ETL pipeline and storing the logs and syslogs in Amazon S3.

What kind of algorithm should the company use to create the simplest solution in this situation?

1. Regression **(Regression will not provide personalized recommendations to customer, because regression can either predict a number or classify based on historical data.)**
2. Classification **(Classification cannot deliver a personalized recommendation for every user.)**
3. Recommender system* **(A *recommendation system* is a subclass of an information filtering system that seeks to predict the "rating" or "preference" a user would give to an item. It is ideal for the situation in this question.)**
4. Reinforcement learning **(Reinforcement learning is a relatively new field and while there are solutions that use it, it is not the simplest solution, since we already have historical data.)**

Q15. A security and networking company wants to use ML to flag certain IP addresses that have been known to send spam and phishing information. The company wants to build an ML model based on previous user feedback indicating whether specific IP addresses have been connected to a website designed for spam and phishing.

What is the simplest solution that the company can implement?

1. Regression **(Regression needs a historical dataset with a numerical output which fails the requirement of the use case in this question.)**
2. Classification **(Classification can work in this situation, but it's not the simplest solution.)**
3. Natural language processing (NLP) **(ML for natural language processing and text analytics is used to understand the meaning of text documents. It can be one part of the solution in this context, but it is not the simplest solution.)**
4. A rule-based solution should be used instead of ML*

Q 16. What factors lead to the wide adoption of neural networks in the last decade? (Select THREE.)

1.

1. Efficient algorithms*
2. Cheaper GPUs*
3. An orders of magnitude increase in data collected* **(Over the last two decades, the amount of available data of all sorts and the power of our data storing and processing machines (GPUs) have exponentially increased. Combined with better computing capabilities, the massive increases in the amount of available data to train models have allowed the creation of larger, deeper neural networks, which just perform better than smaller ones.)**
4. Cheaper CPUs **(*GPUs* are needed to train neural networks efficiently, so cheaper CPUs don't have much to do with the wide adoption of neural networks in the last decade.)**
5. Wide adoption of cloud-based services **(While cloud-based services made it easy for everyone to do machine learning, they are based on the actual factors like efficient algorithms and cheaper GPUs)**

Q 17. An online news organization wants to expand its reach globally by translating some of its most commonly read articles into different languages using ML. The organization's data science team is gathering all the news articles that they have published in both English and at least one other language. They want to use this data to create one machine learning model for each non-English language that the organization is targeting. The models should only require minimum management.

What approach should the team use to building these models?

1. Use Amazon SageMaker Object2Vec to create a vector. Use the SockEye model in Amazon SageMaker using Building Your Own Containers (BYOC) **(BYOC requires more management of the model training process, because you have to maintain the containers and code.)**
2. Use Amazon SageMaker Object2Vec to create a vector. Use the Amazon SageMaker built-in Sequence to Sequence model (Seq2Seq)* **(This is the best answer, because Amazon SageMaker takes care of the management and heavy lifting of the model training and deployment.)***
3. Use Amazon SageMaker Object2Vec to create a vector. Use Amazon EC2 instances with the Deep Learning Amazon Machine Image (AMI) to create a language encoder-decoder model **(This solution is not ideal, given the situation outlined in the question, because it requires you to manage the model training and deployment yourself.)**
4. Use Amazon SageMaker Object2Vec to create a vector. Then use a Long Short-term Memory (LSTM) model using Building Your Own Containers (BYOC) **(BYOC requires more management of the model training process, because you have to maintain containers and code.)**

Q18. An ad tech company is using an XGBoost model to classify its clickstream data. The company's Data Scientist is asked to explain how the model works to a group of non-technical colleagues.

What is a simple explanation the Data Scientist can provide?

1. XGBoost is an Extreme Gradient Boosting algorithm that is optimized for boosted decision trees*
2. XGBoost is a state-of-the-art algorithm that uses logistic regression to split each feature of the data based on certain conditions **(XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.)**
3. XGBoost is a robust, flexible, scalable algorithm that uses logistic regression to classify data into buckets **(XGBoost uses decision trees to perform both regression and classification.)**
4. XGBoost is an efficient and scalable neural network architecture. **(XGBoost is not a neural network but a tree boosting algorithm.)**

Q19. An ML scientist has built a decision tree model using scikit-learn with 1,000 trees. The training accuracy for the model was 99.2% and the test accuracy was 70.3%.

Should the Scientist use this model in production?

1. Yes, because it is generalizing well on the training set **(This is incorrect, because the model is not generalizing well, as illustrated by the difference in accuracy scores between training and testing.)**
2. No, because it is generalizing well on the training set **(This is incorrect, because the model is not generalizing well, as illustrated by the difference in accuracy scores between training and testing.)**
3. No, because it is not generalizing well on the test set* **(This is correct, because the model is not generalizing well, as illustrated by the difference in accuracy scores between training and testing. Therefore, the model should not be used in production.)**
4. Yes, because it is not generalizing well on the test set **(This is correct in that the model is not generalizing well, but as a result, the scientist shouldn't use the model in production.)**

Q 20. A Machine Learning Engineer wants to use Amazon SageMaker and the built-in XGBoost algorithm for model training. The training data is currently stored in CSV format, with the first 10 columns representing features and the 11th column representing the target label.

What should the ML Engineer do to prepare the data for use in an Amazon SageMaker training job?

1. The target label should be changed to the first column. The data should be split into training, validation, and test sets. Finally, the datasets should be uploaded to Amazon S3.* **(For training data in CSV format, the XGBoost algorithm assumes that the target variable is in the first column and that it does not have a header record. Refer to the following for more information: https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html#InputOutput-XGBoost for more info)**
2. The dataset should be uploaded directly to Amazon S3. Amazon SageMaker can then be used to split the data into training, validation, and test sets. **(You should split the data before you upload the train, test and validation datasets to Amazon S3. Amazon S3 cannot split the data automatically.)**
3. The data should be split into training, validation, and test sets. The datasets should then be uploaded to Amazon S3. **(Splitting the data before uploading to Amazon S3 is right, but Amazon SageMaker expects the first column to be the target variable which has to be done before upload.)**
4. The target label should be changed to the first column. The dataset should then be uploaded to Amazon S3. Finally, Amazon SageMaker can be used to split the data into training, validation, and test sets. **(Amazon SageMaker cannot split the data automatically.)**

Q21. A navigation and transportation company is using satellite images to model weather around the world in order to create optimal routes for its ships and planes. The company is using Amazon SageMaker training jobs to build and train its models.

However, during training, it takes too long to download the company's 100 GB data from Amazon S3 to the training instance before the training starts.

What should the company do to speed up its training jobs while keeping the costs low?

1. Increase the instance size for training **(Increasing instance size may increase the network throughput a little but it wont speed up the training job time since the training job will still have to wait for the whole dataset to download to the instance)**
2.

1. Increase the batch size in the model **(Increasing batch size doesnt help with improving training time)**
2. Change the input mode to Pipe* **(With Pipe input mode, your dataset is streamed directly to your training instances instead of being downloaded first. This means that your training jobs start sooner, finish quicker, and need less disk space.)**
3. Create an Amazon EBS volume with the data on it and attach it to the training job **(Amazon EBS volume will certainly help in speeding up the time taken but you cannot attach an existing EBS volume to a training job)**

Q22. A Data Scientist wants to tune the hyperparameters of a machine learning model to improve the model's F1 score.

What technique can be used to achieve this desired outcome on Amazon SageMaker? (Select TWO)

1. Grid Search**(The traditional way of performing hyperparameter optimization has been *grid search*, or a *parameter sweep*, which is simply an exhaustive searching through a manually specified subset of the hyperparameter space of a learning algorithm. A grid search algorithm must be guided by some performance metric, typically measured by cross-validation on the training set or evaluation on a held-out validation set.)**
2. Random Search* **(Random Search replaces the exhaustive enumeration of all combinations by selecting them randomly. It can outperform Grid search, especially when only a small number of hyperparameters affects the final performance of the machine learning algorithm.)**
3. Breadth First Search **(Breadth First Search is not an algorithm for hyperparameter optimization. Rather, its a graph algorithm.Breadth-first search is an algorithm for traversing or searching tree or graph data structures.)**
4. Bayesian optimization***(Bayesian optimization builds a probabilistic model of the function mapping from hyperparameter values to the objective evaluated on a validation set. In practice, Bayesian optimization has been shown to obtain better results in fewer evaluations compared to grid search and random search, due to the ability to reason about the quality of experiments before they are run. Amazon Sagemaker supports Bayesian hyperparameter optimization.)**
5. Depth first search **(Although a good search method algorithm, this is not used for hyperparameter optimization since it generally deals with array search)**

Q 23. A Data Scientist is using stochastic gradient descent (SGD) as the gradient optimizer to train a machine learning model. However, the model training error is taking longer to converge to the optimal solution than desired.

What optimizer can the Data Scientist use to improve training performance? (Select THREE)

1. Adam* **(Adam stands for adaptive momentum which can help the model converge faster and get out of being stuck in local minima.)**
2. Adagrad* **(Adagrad is an algorithm for gradient-based optimization that adapts the learning rate to the parameters by performing smaller updates and, in turn, helps with convergence.)**
3. Gradient Descent **(Gradient descent will take longer to converge than SGD does since it needs the whole dataset for every step calculation.)**
4. RMSProp* **(RMSProp uses a moving average of squared gradients to normalize the gradient itself, which helps with faster convergence.)**
5. Mini-batch gradient descent **(Mini batch gradient descent will suffer from some of the same problems as SGD.)**
6.

1. Xavier **(Xavier is an initialization technique and not an optimization technique.)**

Q 24. A Data Scientist wants to use the Amazon SageMaker hyperparameter tuning job to automatically tune a random forest model.

What API does the Amazon SageMaker SDK use to create and interact with the Amazon SageMaker hyperparameter tuning jobs?

1. HyperparameterTunerJob()
2. HyperparameterTuner()* **(This is the correct class for creating and interacting with Amazon SageMaker hyperparameter tuning jobs, as well as deploying the resulting model(s). It takes an estimator to obtain configuration information for training jobs that are created as the result of a hyperparameter tuning job. Refer to the following for more information: https://sagemaker.readthedocs.io/en/stable/tuner.html)**
3. HyperparameterTuningJobs()
4. Hyperparameter()

Q 25. A Machine Learning Engineer is creating a regression model for forecasting company revenue based on an internal dataset made up of past sales and other related data.

What metric should the Engineer use to evaluate the ML model?

1. Cross-entropy log loss **(Cross-entropy log loss is generally used for classification.)**
2. Sigmoid **(Sigmoid maps the input value to an output that is between 0 and 1. It is not a good metric to use for this use case)**
3. Root Mean squared error (RMSE)* **(Residuals are a measure of how far from the *regression* line data points are; RMSE is a measure of how spread out these residuals are. The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data, or, put another way, how close the observed data points are to the model's predicted values.)**
4. Precision **(*Precision* means the percentage of your results that are relevant and not used for regression problems.)**

Q 26. A Data Scientist at a credit card company trained a classification model to predict fraud at the time of a transaction. The Data Scientist used a confusion matrix to evaluate the performance of the model.

Using the confusion matrix below, determine the percent of positive records that were classified correctly. Choose the answer that also labels this evaluation metric correctly.

1. True PositiveTrue NegativePredicted Positive10090Predicted Negative2525080%; Recall* **(In this context, *recall* is also referred to as the *true positive rate*. *I*t equals true positive divided by true positive plus false negative. In this situation, that works out to: 100/100+15= 0.8 or 80%.)**
2. 52.6%; Recall **(Recall is the right metric to use, but the answer is wrong.)**
3. 80%; Precision **(Precision is not the right metric to use.)**
4. 52.6%; Precision **(Precision is also referred to as the *positive predictive value* (PPV). Other related measures used in classification include true negative rate and accuracy. True negative rate is also called *specificity*.)**

Q 27. A financial planning company is using the Amazon SageMaker endpoint with an Auto Scaling policy to serve its forecasting model to the company's customers to help them plan for retirement. The team wants to update the endpoint with its latest forecasting model, which has been trained using Amazon SageMaker training jobs. The team wants to do this without any downtime and with minimal change to the code.

What steps should the team take to update this endpoint?

1. Use a new endpoint configuration with the latest model Amazon S3 path in the *UpdateEndpoint* API. **(Using a new endpoint configuration will not have Auto Scaling enabled.)**
2. De-register the endpoint as a scalable target. Update the endpoint using a new endpoint configuration with the latest model Amazon S3 path. Finally, register the endpoint as a scalable target again.*
3. Update the endpoint using a new configuration with the latest model Amazon S3 path. Then, register the endpoint as a scalable target. **(Before you can update the endpoint, you need to deregister the endpoint as a scalable target.)**
4. Create a new endpoint using a new configuration with the latest model. Then, register the endpoint as a scalable target. **(Creating a new endpoint will mean that they have to update the code every time the model changes, which isn't scalable.)**

Q 28. An advertising and analytics company uses machine learning to predict user response to online advertisements using a custom XGBoost model. The company wants to improve its ML pipeline by porting its training and inference code, written in R, to Amazon SageMaker, and do so with minimal changes to the existing code.

How should the company set up this new pipeline?

1. Use the Amazon pre-built R container option and port the existing code over to the container. Register the container in Amazon Elastic Container Registry (Amazon ECR). Finally, run the training and inference jobs using this container. **(Amazon SageMaker's pre-built containers do not include a container supporting code written in R.)**
2. Use Amazon in-built algorithms to run their training and inference jobs. **(Amazon SageMaker built-in algorithms does not support using custom code.)**
3. Use the Build Your Own Container (BYOC) Amazon SageMaker option. Create a new Docker container with the existing code. Register the container in Amazon Elastic Container Registry (ECR). Finally, run the training and inference jobs using this container.*
4. Create a new Amazon SageMaker notebook instance. Copy the existing code into an Amazon SageMaker notebook. Then, run the pipeline from this notebook. **(Amazon SageMaker notebook instances support code written in R, but the notebook won't be scalable or used in a pipeline.)**

Q 29. A multi-national banking organization provides loan services to customers worldwide. Many of its customers still submit loan applications in paper form in one of the bank's branch locations. The bank wants to speed up the loan approval process for this set of customers by using machine learning. More specifically, it wants to create a process in which customers submit the application to the clerk, who scans and uploads it to the system. The system then reads and provides an approval or denial of the application in a matter of minutes.

What can the bank use to read and extract the necessary data from the loan applications without needing to manage the process?

1. A custom CNN model **(You will need to manage the process if you are using a convolutional neural**

1. **network model, which goes against one of the requirements of this question.)**
2. An LSTM model **(A LSTM model is not the right model to use for this use case because it generally works with sequences, rather than images.)**
3. Amazon Textract*
4. Amazon Personalize **(Amazon Personalize cannot read and extract data from images.)**

Q 30. A video streaming company wants to create a searchable video library that provides a personalized searching experience and automated content moderation for its users, so that when the users search for a keyword, they get all the videos that map to that keyword. The company wants to do this with minimal cost and limited need for management.

What approach should the company take to building this solution?

1. Use Amazon SageMaker to create an ML model that extracts metadata from the videos **(Amazon SageMaker requires management of the pipeline, which does not satisfy the company's requirements.)**
2. Use Amazon Rekognition Video to extract metadata from the videos*
3. Use Amazon Kinesis Video Streams to stream the videos to Amazon EMR in order to create an ML model **(Amazon EMR requires management of the pipeline, but the company wanted to avoid management.)**
4. Use AWS Batch to transform a batch of video files into metadata **(AWS Batch does not have a function to transform video files to metadata, which is required in this situation.)**

Q 31. A ride-share company wants to create intelligent conversational chatbots that will serve as first responders to customers who call to report an issue with their ride. The company wants these chatbot-customer calls to mimic natural conversations that provide personalized experiences for the customers.

What combination of AWS services can the company use to create this workflow without a lot of ongoing management?

1. Amazon Lex to parse the utterances and intent of customer comments, Amazon Polly to reply to the customers* **(This is the right blend of services.)**
2. Amazon Polly to parse the utterances and intent of customer comments, Amazon Lex to reply to the customers **(Amazon Polly should be used to reply to customers, rather than to parse utterances and intent. Amazon Lex can't reply to customers.)**
3. Amazon Transcribe to parse the utterances and intent of customer comments, Amazon Lex to reply to the customers **(Amazon Transcribe can transcribe videos with text but not parse utterances. Amazon Lex can't reply to customers.)**
4. Amazon Transcribe to parse the utterances and intent of customer comments, Amazon Polly to reply to the customers **(Amazon Polly is the right choice to reply to customers, but Amazon Transcribe does not parse utterances.)**

Q 32. A healthcare organization has an application that takes in sensitive user data. This data is encrypted at rest and stored in an Amazon S3 bucket using customer-managed encryption with AWS Key Management Service (AWS KMS). A Data Scientist in the organization wants to use this encrypted data as features in a Amazon SageMaker training job. However, the following error continues to occur: "Data download failed."

What should the Data Scientist do to fix this issue?

1. Make sure the AWS Identity and Access Management (IAM) role used for Amazon S3 access has permissions to encrypt and decrypt the data with the AWS KMS key. *
2. Add “S3:*” to the IAM role that is attached to the Amazon SageMaker training job. **(Adding “S3:*” wildcard is not a good security practice. It will not read encrypted AWS KMS data into Amazon SageMaker.)**
3. Specify the “VolumeKmsKeyId” in the Amazon SageMaker training job. **(VolumeKmsKeyId helps in encrypting data on the training job instance storage, not on Amazon S3.)**
4. Add “EnableKMS” to the Amazon SageMaker training job. Then, specify the Amazon S3 bucket that includes the data. **(There is no option to enable AWS KMS that will read at-rest data in an encrypted Amazon S3 bucket.)**

Q 33. A log analytics company wants to provide a history of Amazon SageMaker API calls made on its client’s account for security analysis and operational troubleshooting purposes.

What must be done in the client’s account to ensure that the company can analyze the API calls?

1. Use IAM roles. “logs:*” are added to those IAM roles. **(The “logs:*” option in IAM roles will allow permission for Amazon CloudWatch logs, but not API calls.)**
2. Enable AWS CloudTrail.*
3. Enable CloudWatch logs. **(CloudWatch logs will enable logs from the training instances, but not show API calls.)**
4. Use the Amazon SageMaker SDK to call the ‘sagemaker_history()’ function. **(There is no function that shows the Amazon SageMaker API calls in the Amazon SageMaker SDK.)**

Q 34. A team of Data Scientists wants to use Amazon SageMaker training jobs to run two different versions of the same model in parallel to compare the long-term effectiveness of the different versions in reaching the related business outcome.

How should the team deploy these two model versions with minimum management?

1. Create a Lambda function that preprocesses the incoming data, calls the two Amazon SageMaker endpoints for the two models, and finally returns the prediction. **(Creating a Lambda function and having two Amazon SageMaker endpoints will require more management than an ideal solution.)**
2. Create an endpoint configuration with production variants for the two models with equal weights.*
3. Create an endpoint configuration with production variants for the two models with a weight ratio of 90:10. **(Creating a 90:10 variant will give more prediction power to one model than the other model, which might skew the result.)**
4. Create a Lambda function that downloads the models from Amazon S3 and calculates and returns the predictions of the two models. **(AWS Lambda can be used for prediction as well, but it has a 15-minute runtime limit and does not include GPU instances.)**

Q 35. A Data Scientist at an ad-tech startup wants to update an ML model that uses an Amazon SageMaker endpoint using the canary deployment methodology, in which the production variant 1 is the production model and the production variant 2 is the updated model.

How can the Data Scientist *efficiently* configure this endpoint configuration to deploy the two different versions of the model while monitoring the Amazon CloudWatch invocations?

1. Create an endpoint configuration with production variants for the two models with equal weights. **(Creating equal weights is not how a canary deployment works. Canary deployment adds the new model or new deployment in small iterations.)**
2. Create two Amazon SageMaker endpoints and change the endpoint URL after testing the new endpoint. **(Creating two Amazon SageMaker endpoints will entail manual load to compare metrics.)**
3. Create an endpoint configuration with production variants for the two models with a weight ratio of 0:1. Update the weights periodically.*
4. Create an endpoint configuration with production variants for the two models with a weight ratio of 10:90. **(Creating an endpoint configuration with a weight ration of 10:90 will not satisfy the canary deployment technique, because canary deployment should start with either 100:0 or 90:10. This is done so that the original production model takes most of the load while you test the new model in production to see if there are any errors.)**

Solutions to Domain 1 Quiz Questions

* denotes correct answer

Q 1. A healthcare company using the AWS Cloud has access to a variety of data types, including raw and preprocessed data. The company wants to start using this data for its ML pipeline, but also wants to make sure the data is highly available and located in a centralized repository.

What approach should the company take to achieve the desired outcome?

1. Create a data lake using Amazon S3 as the data storage layer*
2. Store unstructured data in Amazon DynamoDB and structured data in Amazon RDS **(Having two storage layers like this breaks the centralized repository requirement in this question.)**
3. Use Amazon FSx to host the data for training **(Amazon FSx should not be used for workloads and is too costly for a permanent storage solution.)**
4. Use Amazon Elastic Block Store (Amazon EBS) volumes to store the data with data backup **(Amazon EBS data backups are not highly available, which is one of the requirements in this question.)**

Q 2. A Data Scientist wants to implement a near-real-time anomaly detection solution for routine machine maintenance. The data is currently streamed from connected devices by AWS IoT to an Amazon S3 bucket and then sent downstream for further processing in a real-time dashboard.

What service can the Data Scientist use to achieve the desired outcome with minimal change to the pipeline?

1. Amazon CloudWatch **(CloudWatch does not offer an anomaly detection solution.)**
2. Amazon SageMaker **(An Amazon SageMaker training job needs processed data stored in Amazon S3 to train the model. It cannot train the model on streaming data.)**
3. Amazon EMR with Spark **(Amazon EMR would require more than just a minimal change in the pipeline to stream the data to an EMR instance.)**
4. Amazon Kinesis Data Analytics*

Q 3. A transportation company currently uses Amazon EMR with Apache Spark for some of its data transformation workloads. It transforms columns of geographical data (like latitudes and longitudes) and adds columns to segment the data into different clusters per city to attain additional features for the k-nearest neighbors algorithm being used.

The company wants less operational overhead for their transformation pipeline. They want a new solution that does not make significant changes to the current pipeline and only requires minimal management.

What AWS services should the company use to build this new pipeline?

1. Use Amazon EMR to transform files. Use Amazon S3 as the destination. **(Amazon EMR is a good option, but the service still requires management of security settings and other management tasks—so this solution doesn't meet the company's requirements.**
2. Use Lambda to transform files. Use Amazon EMR HDFS as the destination. **(Lambda can transform the data, but that would require changing the code to Python, which increases labor instead of decreasing it.)**
3. Use AWS Glue to transform files. Use Amazon S3 as the destination.*
4. Use AWS Glue to transform files. Use Amazon EMR HDFS as the destination. **(Amazon EMR HDFS requires spinning up and maintaining an Amazon EMR cluster to store the data. But the service still requires management of security settings and other management tasks—so this solution doesn't meet the company's requirements.)**

Solutions to Domain 2 Quiz Questions

* denotes correct answer

Q 1. A Machine Learning Engineer is creating and preparing data for a linear regression model. However, while preparing the data, the Engineer notices that about 20% of the numerical data contains missing values in the same two columns. The shape of the data is 500 rows by 4 columns, including the target column.

How could the Engineer handle the missing values in the data? (Select TWO.)

1. Remove the rows containing the missing values **(The dataset is small enough where removing 20% of the data might result is loss of valuable information inside those rows)**
2. Remove the columns containing the missing values **(This approach causes the lose of data features, and, in this case, there are only three feature columns.)**
3. Fill the missing values with zeros*
4. Impute the missing values using regression*
5. Add regularization to the model **(Adding regularization helps with overfitting, not missing data.)**

Q 2. A social networking organization wants to analyze all the comments and likes from its users to flag offensive language on the site. The organization's data science team wants to use a Long Short-term Memory (LSTM) architecture to classify the raw sentences from the comments into one of two categories: offensive and non-offensive.

What should the team do to prepare the data for the LSTM?

1. Convert the individual sentences into sequences of words. Use those as the input. **(It is more effective to vectorize the sentences to capture relationships across words than it is to convert the sentences into sequences of words.)**
2. Convert the individual sentences into numerical sequences starting from the number 1 for each word in a

1. sentence. Use the sentences as the input. **(Using a numerical sequence for each word in a sentence will retain the placing of the word in the sentence, but will lose the actual word itself, which needs to be coded in.)**
2. Vectorize the sentences. Transform them into numerical sequences. Use the sentences as the input. **(The sentences will need to be padded, because the algorithm expected a fixed vector length and each sentence will not be the same length.)**
3. Vectorize the sentences. Transform them into numerical sequences with a padding. Use the sentences as the input.*

Q 3. A Data Scientist created a correlation matrix between nine variables and the target variable. The correlation coefficient between two of the numerical variables, variable 1 and variable 5, is -0.95.

How should the Data Scientist interpret the correlation coefficient?

1. As variable 1 increases, variable 5 increases **(The question's correlation coefficient indicates a negative correlation between the two variables. This answer option represents a positive correlation.)**
2. As variable 1 increases, variable 5 decreases*
3. Variable 1 does not have any influence on variable 5 **(The question's correlation coefficient indicates a relationship between the two variables: in this case, a negative correlation.)**
4. The data is not sufficient to make a well-informed interpretation **(There is sufficient data to draw a conclusion here, which is that there is a negative correlation between the two variables.)**

Solutions to Domain 3 Quiz Questions

* denotes correct answer

Q 1. A real estate company wants to provide its customers with a more accurate prediction of the final sale price for houses they are considering in various cities. To do this, the company wants to use a fully connected neural network trained on data from the previous ten years of home sales, as well as other features.

What kind of machine learning problem does this situation represent?

1. Regression* **(Regression analysis is the right answer, because the company wants to predict the final house price (independent variable) depending on various cities (dependent variable).)**
2. Classification **(Classification cannot be used for this, because the company wants to predict a number for the sales price rather than a category.)**
3. Recommender system **(A recommender system doesn't fit this use case, because a number needs to be predicted.)**
4. Reinforcement learning **(Reinforcement learning doesn't fit the use case in this questio, because we already have historical data.)**

Q 2. A manufacturing company wants to increase the longevity of its factory machines by predicting when a machine part is about to stop working, jeopardizing the health of the machine. The company's team of Data Scientists will build an ML model to accomplish this goal. The model will be trained on data made up of consumption metrics from similar factory machines, and will span a time frame from one hour before a machine part broke down to five minutes after the part degraded.

What kind of machine learning algorithm should the company use to build this model?

1. Amazon SageMaker DeepAR* **(Amazon Sagemaker DeepAR is a supervised learning algorithm designed for time series forecasting problems. Given the situation laid out in this question, this is the ideal algorithm to use.)**
2. SciKit Learn Regression **(This is a linear regression algorithm, which does not fit well in this question given that it's a time series forecasting problem.)**
3. Convolutional neural network (CNN) **(A CNN is a class of deep neural networks most commonly applied to analyzing visual imagery. It would not be appropriate on its own for a time series forecasting problem.)**
4. Scikit Learn Random Forest **(Random forest is a very popular tree-based algorithm used for either classification or regression problems, but not for time series forecasting.)**

Q 3. A Data Scientist working for an autonomous vehicle company is building an ML model to detect and label people and various objects (for instance, cars and traffic signs) that may be encountered on a street. The Data Scientist has a dataset made up of labeled images, which will be used to train their machine learning model.

What kind of ML algorithm should be used?

1. Image classification **((Image classification will not detect each distinct object in the image. It will only classify one distinct image.)**
2. Instance segmentation*
3. Image localization **(Object or image localization tries to locate the main (or most visible) object in an image, but won't detect each distinct object.)**
4. Semantic segmentation **(*Semantic segmentation* is the process of linking each pixel in an image to a class label. These labels could include a person, car, flower, piece of furniture, etc., just to mention a few. Think of semantic segmentation as image classification at the pixel level. And like image classification, semantic segmentation will not detect each distinct object in the image.)**

Q 4. A Data Scientist is training a convolutional neural network model to detect incoming employees at the company's front gate using a camera so that the system opens for them automatically. However, the model is taking too long to converge and the error oscillates for more than 10 epochs.

What should the Data Scientist do to improve upon this situation? (Select TWO.)

1. Normalize the images before training *
2. Add batch normalization*
3. Add more epochs **(The model already is suffering from convergence issues, so increasing the epochs won't help with convergence.)**
4. Increase batch size **(Increasing the batch size will generally makes the convergence worse.)**
5. Decrease weight decay **(Weight decay is generally used for regularization and overfitting, and, therefore, wont help with convergence issues)**

Q 5. A Data Scientist at a waste recycling company trained a CNN model to classify waste at the company's sites. Incoming waste was classified as either trash, compost, or recyclable to make it easier for the machines to split the incoming waste into the appropriate bins.

During model testing, the F1 score was 0.918. The company's senior leadership originally asked the Data Scientist to reach an F1 score of at least 0.95.

What should the Data Scientist do to improve this score without spending too much time optimizing the model?

1. Use Amazon SageMaker tuning jobs to tune the hyperparameters used*
2. Increase the batch size to improve the score in the Amazon SageMaker training job **(Increasing batch size may or may not help with improving the F1 score.)**
3. Use momentum to improve the training in the Amazon SageMaker training job **(Increasing momentum generally helps with convergence but may not help with increasing your F1 score.)**
4. Run the Amazon SageMaker training job for more epochs **(Running more epochs will overfit the model and will not help with increasing the testing F1 score.)**

Solutions to Domain 4 Quiz Questions

* denotes correct answer

Q 1. A Machine Learning Engineer created a pipeline for training an ML model using an Amazon SageMaker training job. The training job began successfully, but then failed after running for five minutes.

How should the Engineer begin to debug this issue? (Select TWO.)

1. Log into the Amazon SageMaker training job instance and check the job history **(You cannot log into an Amazon SageMaker training job instance.)**
2. Call the *DescribeJob* API to check the FailureReason option*
3. Go to Amazon CloudWatch logs and check the logs for the given training job*
4. Check the error in the given training job directly in the Amazon SageMaker console **(The Amazon SageMaker console doesn't give you insight into what happens with a specific training job.)**
5. Check AWS CloudTrail logs to check the error that caused the training to fail **(AWS CloudTrail logs the API calls for Amazon SageMaker, but will not log the error.)**

Q 2. A news organization wants to extract metadata from its articles and blogs and index that metadata in Amazon Elasticsearch Service (Amazon ES) to enable faster searches.

What AWS service can the organization use to achieve this goal?

1. Amazon Comprehend*
2. Amazon Personalize **(Amazon Personalize is not the right service for this use case, because it is a service that creates recommendations for customers.)**
3. Amazon Textract **(Amazon Textract extracts data, but not metadata, from images and PDFs using optical character recognition (OCR).)**
4. Amazon Rekognition Image **(Amazon Rekognition Image does not extract metadata from articles and blogs, but from images.)**

Q 3. A Machine Learning Specialist is evaluating an ML model using a custom Deep Learning Amazon Machine Image (AMI) with Anaconda installed to run workloads through the terminal. Unfortunately, the ML Specialist does not have any experience with the Deep Learning AMI and wants to log into the instance and create an ipython notebook (*.ipynb), but cannot access the notebook interface.

After creating the AMI instance, what steps should the ML Specialist take to create a notebook?

1. SSH into the Deep Learning AMI instance, start a new Flask interface application, and create a new ipython notebook **(SSH into the Deep Learning AMI will work, but Flask won't create a new notebook.)**
2. SSH into the Deep Learning AMI instance with port forwarding at port 8888, start a Jupyter notebook application, and create a new ipython notebook*
3. SSH into the Deep Learning AMI instance with port forwarding at port 8888 and start a python3.6 application to create a new ipython notebook **(SSH into the Deep Learning AMI with port forwarding is the right option, but the Python application will open a Python terminal instead of a GUI that creates the *.ipynb notebook.)**
4. SSH into the Deep Learning AMI instance with port forwarding at port 8080 and start a Zeppelin application to create a new ipython notebook **(SSH into the deep learning AMI will work, but a Zeppelin application doesn't create a *.ipynb notebook.)**

Q 4. A machine translation company is deploying its language translation models behind an Amazon SageMaker endpoint. The company wants to deploy a solution directly on its website so that users can input text in one language and have it translated into a second language. The company wants to reach a solution with minimal maintenance and latency for spiky traffic times.

How should the company architect this solution?

1. Use Amazon SageMaker *InvokeEndpoint* with API Gateway **(The Amazon SageMaker model cannot be called directly using API Gateway, but needs a compute resource like Lambda in between to call the endpoint.)**
2. Use Lambda to call *InvokeEndpoint*. Use the Amazon API Gateway URL to call the AWS Lambda function.*
3. Create a function on an Amazon EC2 instance that uses CURL to call the *InvokeEndpoint* API. Call the Amazon EC2 instance from the website. **(Using Amazon EC2 will require more maintenance than the requirement in the question states.)**
4. Install the sagemaker-runtime library on the web server. Call *InvokeEndpoint* from the webserver. **(Calling InvokeEndpoint from the web server puts the load of preprocessing language data on the web server, which can slow the website.)**