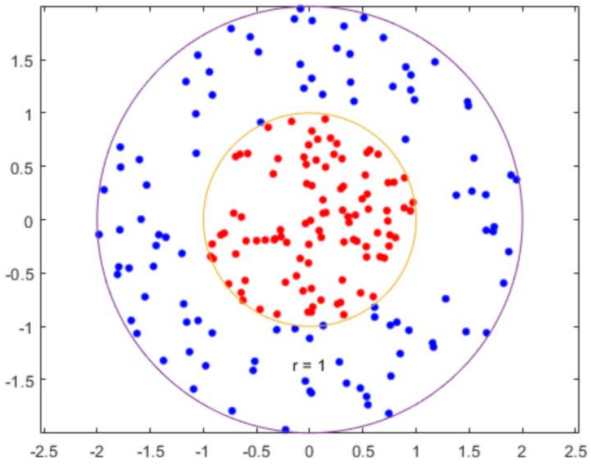
Exploratory Data Analysis:

As part of a clinical study, you have processed millions of medical records with hundreds of features and reduced the feature dimensions to just two using a model based on Principal Component Analysis (PCA). The following graph illustrates the distribution from the PCA model output in the form of two distinct classes shown in red and blue dots.



Which algorithm can you use to accurately classify the above output from the PCA model:

- Logistic Regression
- Linear Regression
- K-means
- Support Vector Machine **(Correct)**

Explanation

This is a classification problem, so Linear Regression can be ruled out. As the decision boundary is non-linear, so Logistic Regression is not the right fit. K-means is used for clustering, so that's also incorrect. SVM can classify for non-linear boundary, so that's the right option.

As an ML Specialist, you observe that one of the features used in a SageMaker Linear Learner model had 30% missing data. You also believe that this specific feature was somehow related to few other features in the data-set. Which technique would you use to address the missing data:

- Use multiple imputations approach via a supervised learning technique that uses other features to figure out the imputed value **(Correct)**
- Replace the missing values with median
- Replace the missing values with mean
- Drop the missing feature as 30% missing data would severely distort the overall training data

Explanation

Dropping the feature is not recommended as we can impute the missing values for the feature via a supervised learning technique that uses other features to figure out the imputed value. Replacing with mean or median can introduce bias into the model.

An insurance company is building a binary classification model to predict insurance claims. The training data contains 1800 instances of the positive class (customers who did claim insurance) and 100 instances of the negative class (customers who did not claim insurance). The final model has 85% accuracy, but poor precision. How can you improve the model performance (Select three):

- [x] Create more samples using algorithms such as SMOTE **(Correct)**
- [ ] Over-sample from the positive class
- [ ] Collect more training data for the positive class
- [x] Over-sample from the negative class **(Correct)**
- [x] Collect more training data for the negative class **(Correct)**

Explanation

In case of a binary classification model with strongly unbalanced classes, we need to over-sample from the minority class, collect more training data for the minority class and create more samples using algorithms such as SMOTE which effectively uses a k-nearest neighbours approach to exclude members of the majority class while in a similar way creating synthetic examples of a minority class. Over-sampling from the positive class or collecting more training data for the positive class would further aggravate the situation. Here are a few good references :

http://www.svds.com/learning-imbalanced-classes/

https://stats.stackexchange.com/questions/235808/binary-classification-with-strongly-unbalanced-classes

A Silicon Valley startup intends to provide real estate recommendations to millions of homebuyers. The data science team is grappling with thousands of features that could go into the model, so the team wants to consider only the most relevant derived features. As an ML Specialist, what solution would you recommend to get them started:

- ○ Use Latent Dirichlet Allocation to identify the most relevant derived features
- ○ Use K-means to identify the most relevant derived features
- ○ Use Neural Topic Model to identify the most relevant derived features
- ◉ Use Principal Component Analysis to identify the most relevant derived features **(Correct)**

Explanation

K-means is used for clustering. LDA and NTM are used for topic modeling. Principal Component Analysis reduces the dimensionality and thus it can help in identifying the most relevant derived features.

You are analyzing the salary trends for Silicon Valley engineers over the last decade. The dataset has information on the age, professional experience in years, skill level from 1 to 10, gender, location (5 distinct locations) and salary. What is the best way to visualize the average annual salary trend at each location over the last decade:

- ○ Pie Chart
- ○ Bar Chart
- ○ Scatter Plot
- ◉ Line Chart **(Correct)**

Explanation

Pie Chart is ruled out for this use-case. While you can use multi bar chart or multi scatter plot, none of these will bring out the annual salary **trend** for each location. Multi series line chart is best suited for this job.

The management of a shopping mall wants to predict the number of footfalls next week. They have data for the average number of footfalls over the last 52 weeks. Which probability distribution should they be using for this scenario planning exercise ?

- ○ Normal distribution
- ◉ Poisson distribution **(Correct)**
- ○ Binomial distribution
- ○ Bernoulli distribution

Explanation

You are expected to have familiarity with the use cases for applying these probability distributions. Here is a good reference : https://medium.com/@srowen/common-probability-distributions-347e6b945ce4

The data science team at an investment bank is analyzing the stock price data for blue chip stocks over the last year. Which visualization tools can be used to spot outliers (Select three):

- [ ] Line chart
- [ ] Heatmap
- [ ] Bar chart
- [x] Histogram **(Correct)**
- [x] Scatter plot **(Correct)**
- [x] Box plot **(Correct)**

Explanation

Box plot, scatter plot and histogram can be used to spot outliers. Please review these reference links for more details:

https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba

https://humansofdata.atlan.com/2017/10/how-to-find-outliers-data-set/

A note for using histograms to detect outliers:

A **histogram represents a frequency distribution**. The vertical axis is a frequency axis whilst the horizontal axis is divided into a range of numeric values (intervals or **bins**) or time intervals.

Consider a scenario where you set only 2 bins for the histogram, then the histogram would not let you see any outliers as everything would be combined in these 2 bins. So depending on the number of bins, **the histogram may not be able to bring out the outliers.**

Box plot and Scatter plot will always show the outlier just by visual inspection. However, a histogram may not show outliers if the user configures too few bins for the histogram.

Tf-idf is a statistical technique frequently used in Machine Learning areas such as text-summarization and classification. Tf-idf measures the relevance of a word in a document compared to the entire corpus of documents. You have a corpus (D) containing the following documents:

Document 1 (d1) : "A quick brown fox jumps over the lazy dog. What a fox!"

Document 2 (d2) : "A quick brown fox jumps over the lazy fox. What a fox!"

Which of the following statements is correct:

- ( ) Using tf-idf, the word "fox" is more relevant for document d1 than document d2
- ( ) Insufficient information has been provided to compute tf-idf
- ( ) Using tf-idf, the word "fox" is more relevant for document d2 than document d1
- (•) Using tf-idf, the word "fox" is equally relevant for both document d1 and document d2 **(Correct)**

Explanation

tf is the frequency of any "term" in a given "document". Using this definition, we can compute the following:

tf("fox", d1) = 2/12 , as the word "fox" appears twice in the first document which has a total of 12 words

tf("fox", d2) = 3/12 , as the word "fox" appears thrice in the second document which has a total of 12 words

An idf is constant per corpus (in this case, the corpus consists of 2 documents) , and accounts for the ratio of documents that include that specific "term". Using this definition, we can compute the following:

idf("fox", D) = log(2/2) = 0 , as the word "fox" appears in both the documents in the corpus

Now,

tf-idf("fox", d1, D) = tf("fox", d1) * idf("fox", D) = (2/12) * 0 = 0

tf-idf("fox", d2, D) = tf("fox", d2) * idf("fox", D) = (3/12) * 0 = 0

Using tf-idf, the word "fox" is equally relevant (or just irrelevant!) for both document d1 and document d2

You can further read about tf-idf on this reference link:

https://en.wikipedia.org/wiki/Tf–idf

The data engineering team at a social media company has handed over the cleaned and prepared dataset to the Machine Learning team. The ML team wants to use this dataset for building a regression model based on the SageMaker Linear Learner algorithm. Which is the first step that the ML team needs to do:

- ○ Standardize the dataset
- ○ Create training set, validation set and test set
- ◉ Shuffle the dataset **(Correct)**
- ○ Normalize the dataset

Explanation

Normalization and Standardization is done on specific features, not on the entire dataset. The first step would be to shuffle the dataset and then do the splits for training set, validation set and test set.

The Machine Learning team at an ecommerce company is analysing the sales data. The data is stored in a highly optimized data compression format and the daily volume of data is around 1TB. The team would like to reduce this volume to one-tenth of its original size without significantly compromising on the quality of data, so that they can complete the classification model training in a much shorter timespan. As an ML Specialist, what solution would you recommend to the team:

- ○ The team needs to use better hardware to run the classification model training so that the original data can be processed in a short timespan
- ○ Use a more efficient compression algorithm
- ○ Use clustering to reduce the data volume and still preserve a significant variance between observations
- ◉ Use dimensionality reduction to reduce the data volume and still preserve a significant variance between observations **(Correct)**

Explanation

Adding another compression layer would not help an already compressed dataset. Clustering does not fit into this use-case. Throwing more hardware at the problem does not reduce the data volume. Correct option is to use dimensionality reduction to reduce the data volume while still preserving a significant variance between observations so that data quality is not compromised.

A data science team is trying to port a legacy binary classification model to Amazon SageMaker. In the legacy workflow, the data engineering component was handled using Apache Spark and scikit-learn based preprocessors. Which feature of Amazon SageMaker can be utilized for seamless integration of the legacy functionality into the new SageMaker model:

- ○ Automatic Scaling
- ○ Batch Transform
- ◉ Inference Pipeline **(Correct)**
- ○ Elastic Inference

Explanation

You can use Inference Pipeline to package Spark and scikit-learn based preprocessors into containers:

https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipeline-mleap-scikit-learn-containers.html

What are the ideal characteristics of a good dataset for Machine Learning problems (Select two):

- ☐ Should have skewed distribution so algorithm can learn the edge cases correctly
- ☑ Should have fair sampling with even distribution of outcomes **(Correct)**
- ☑ Should be representative of the underlying business problem to be solved **(Correct)**
- ☐ Should have some bias so algorithm can learn the edge cases correctly

Explanation

If data has bias or follows a skewed distribution, model will also demonstrate skewed results. Ideally data should be representative of the underlying business problem to be solved and have a fair sampling with even distribution of outcomes.

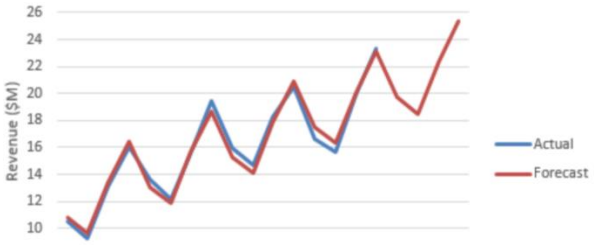
You would like to identify the "global hotspots" for UFO sightings over the last 50 years. The data is available from a globally reputed agency based out of Area 51. Which visualization tool is best suited to bring out these locations on a geographic map:

- Mindmap
- Heatmap **(Correct)**
- Hashmap
- Treemap

Explanation

Hashmap is a programming construct and mindmap is a diagram used to visually organize information. Both are not relevant to this use-case. Treemaps are an alternative way of visualising the hierarchical structure of a tree. Geographic heat maps are an interactive way to identify where something occurs, and demonstrate areas of high and low density.

You are building a forecasting model using the SageMaker DeepAR algorithm. The following graph captures the Actual vs Forecast comparison from the model output for the revenue data for the last few years:



What can you conclude about the graph above:

- The forecast captures both the trend and seasonality well **(Correct)**
- The forecast captures the trend well but misses the seasonality
- The forecast captures neither the trend nor the seasonality
- The forecast captures the seasonality well but misses the trend

Explanation

Trends are continuous increases or decreases in a metric's value. Seasonality, on the other hand, reflects periodic (cyclical) patterns that occur in a system.

A data preprocessing script uses the sklearn and numpy libraries in Python. Which type of Glue job is needed to support this:

- Jupyter Shell
- PySpark
- Zeppelin Shell
- Python Shell **(Correct)**

Explanation

Python Shell supports Glue jobs relying on libraries such as numpy, pandas and sklearn. PySpark supports Glue jobs written primarily using Python API of Apache Spark. There is no such thing as Zeppelin Shell or Jupyter Shell.

Identify the correct statement vis-a-vis the difference between a data lake and data warehouse:

- A data warehouse can only store structured data whereas a data lake can store structured, semi-structured and unstructured data **(Correct)**
- Both data lake and data warehouse can store structured, semi-structured and unstructured data
- Both data lake and data warehouse can only store structured data
- A data lake can only store structured data whereas a data warehouse can store structured, semi-structured and unstructured data

Explanation

A data warehouse can only store structured data whereas a data lake can store structured, semi-structured and unstructured data. Please review this reference link:

https://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html

The data science team at an ecommerce company is working on a training dataset for a forecasting model. The dataset represents the sales data for the last 5 years and has the following features : item description, item price, order date, quantity ordered, shipping address, order amount. The team would like to uncover any cyclical sales patterns such as hourly, daily, weekly, monthly, yearly from this data. As an ML Specialist, what solution would you recommend:

- ○ No need for data preprocessing as the underlying algorithm can detect the cyclical patterns on its own
- ◉ Preprocess the order date to create new features such as hour of the day, day of the week, week of the month, week of the year, date of the month, month of the year and represent these features as (x,y) coordinates on a circle using sin and cos transformations. This transformed data should then be used to train the model. **(Correct)**
- ○ Preprocess the order date to create new features such as hour of the day, day of the week, week of the month, week of the year, date of the month, month of the year and use these features in label encoded format for training the model
- ○ Preprocess the order date to create new features such as hour of the day, day of the week, week of the month, week of the year, date of the month, month of the year and use these features in one-hot encoded format for training the model

Explanation

The best way to engineer the cyclical features is to represent these as (x,y) coordinates on a circle using sin and cos functions. Please review this technique in more detail here -

http://blog.davidkaleko.com/feature-engineering-cyclical-features.html

The data science team at an ecommerce company is working on a training dataset for a forecasting model. The dataset represents the sales data for the last 5 years and has the following features : item description, item price, order date, quantity ordered, shipping address, order amount. The team would like to uncover any cyclical sales patterns such as hourly, daily, weekly, monthly, yearly from this data. As an ML Specialist, what solution would you recommend:

- ○ No need for data preprocessing as the underlying algorithm can detect the cyclical patterns on its own
- ◉ Preprocess the order date to create new features such as hour of the day, day of the week, week of the month, week of the year, date of the month, month of the year and represent these features as (x,y) coordinates on a circle using sin and cos transformations. This transformed data should then be used to train the model. **(Correct)**
- ○ Preprocess the order date to create new features such as hour of the day, day of the week, week of the month, week of the year, date of the month, month of the year and use these features in label encoded format for training the model
- ○ Preprocess the order date to create new features such as hour of the day, day of the week, week of the month, week of the year, date of the month, month of the year and use these features in one-hot encoded format for training the model

Explanation

The best way to engineer the cyclical features is to represent these as (x,y) coordinates on a circle using sin and cos functions. Please review this technique in more detail here -

http://blog.davidkaleko.com/feature-engineering-cyclical-features.html

You are analyzing the income data provided in a case study as part of a data science competition. You observe that the data has several outliers. Which techniques you can use to address outliers in the data (Select two):

- ☑ Standardization **(Correct)**
- ☐ Normalization
- ☐ One-hot encoding
- ☑ Logarithm Transformation **(Correct)**

Explanation

Logarithm transformation and Standardization are the correct techniques to address outliers in data. Please review this reference link:

https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114

The data science team at an ecommerce company wants to analyse and visualize the clickstream data as it arrives into the data lake. What combination of AWS serverless services would you recommend for the team:

- ○ S3, EC2, D3.js
- ◉ S3, Glue, Athena, QuickSight **(Correct)**
- ○ S3, EMR, ElasticSearch, Kibana
- ○ S3, Lambda, ElasticSearch, Kibana

Explanation

ElasticSearch, EMR and EC2 are not "serverless". S3, Glue, Athena, QuickSight is the correct choice of services.

The ML team at an ecommerce company has trained a SageMaker XGBoost model on a large dataset. The evaluation metric looks good for the training job. However, post production deployment, the team observes that the inference results are not correct. As an ML Specialist, what would you recommend to resolve this issue:

- (•) Make sure that the training data distribution is similar to the expected distribution for the production inference data. **(Correct)**
- ( ) Use Elastic Inference to improve the inference results
- ( ) These variations should be acceptable as part of the model performance
- ( ) Replace XGBoost with Linear Learner algorithm

Explanation

There should be no significant variations just for production inference, so that option is ruled out. Elastic Inference is used for low latency and high throughput model inference and it has nothing to do with inference results accuracy. Replacing algorithm will not address the root cause. It seems like data distribution may have been different for the training job, hence model shows major variation on production inference. So the team needs to make sure that the training data distribution is similar to the expected distribution for the production inference data.

A car insurance company wants to automate the claims process. The company wants the customers to upload the video footage of the damaged car. This video footage is then pre-assessed by an Amazon SageMaker model as part of the damage evaluation process. The company has no prior training data to get started on this endeavor. As an ML Specialist, what would you recommend to the company:

- ( ) Use Kinesis Video Streams to create the labels for the training videos. The labeled videos can be used to train the downstream Amazon SageMaker model for the damage evaluation process.
- ( ) Use an unsupervised learning algorithm to label the videos which can be used in the downstream Amazon SageMaker model for the damage evaluation process
- (•) Use Amazon SageMaker Ground Truth to create the labels for the training videos. The labeled videos can be used to train the downstream Amazon SageMaker model for the damage evaluation process. **(Correct)**
- ( ) Use AWS Rekognition to create the labels for the training videos. The labeled videos can be used to train the downstream Amazon SageMaker model for the damage evaluation process.

Explanation

Rekognition or Kinesis Video Streams or an unsupervised learning algorithm cannot be used to create labels for the training videos. Correct option is to use Amazon SageMaker Ground Truth to create the labels for the training videos. The labeled videos can be used to train the downstream Amazon SageMaker model for the damage evaluation process.

You are pre-processing a training dataset to be used on the Amazon SageMaker Linear Learner algorithm. The dataset has hundreds of features and you need to decide which features to drop. Identify the guidelines that you would follow (Select three):

- [x] Drop a feature if it has a low correlation to the target label **(Correct)**
- [ ] Drop a feature if it has high variance
- [x] Drop a feature if it has low variance **(Correct)**
- [ ] Drop a feature if it has a high correlation to the target label
- [x] Drop a feature if it has a lot of missing values **(Correct)**
- [ ] Drop a feature if it has a few missing values

Explanation

The thumb rule is that you drop a feature that will not help a model to learn. Any feature that has low variance or a lot of missing values or has a low correlation to the target label ought to be dropped.

Researchers at NASA are creating a model on Amazon SageMaker to analyze images for detecting strong gravitational lensing, a phenomenon in which an accumulation of matter in space is dense enough that it bends light waves as they travel around it. The training data contains 200K images of the negative class (images with no gravitational lensing) and only 2000 images of the positive class (images with gravitational lensing). The final model has 85% accuracy, but poor recall. How can you improve the model performance (Select two):

- [x] Collect more training data for the positive class **(Correct)**
- [ ] Over-sample from the negative class
- [ ] Collect more training data for the negative class
- [x] Over-sample from the positive class **(Correct)**

Explanation

In case of a binary classification model with strongly unbalanced classes, we need to over-sample from the minority class, collect more training data for the minority class. Here are a few good references :

http://www.svds.com/learning-imbalanced-classes/

https://stats.stackexchange.com/questions/235808/binary-classification-with-strongly-unbalanced-classes

The research team at a University wants to do sentiment analysis of the most famous quotes from the classical english literature over the last 500 years. Some of the sample quotes from the corpus are like so : *"All that glitters is not gold", "Brevity is the soul of wit", "The lady doth protest too much, methinks.", "Love all, trust a few, do wrong to none."* As an ML Specialist, what data pre-processing steps would you recommend before the team starts building the model (Select three):

- [ ] Create n-gram vector for each word
- [x] Remove stop words **(Correct)**
- [x] Lowercase each word **(Correct)**
- [x] Tokenize each word **(Correct)**
- [ ] Remove archaic words such as doth and methinks
- [ ] Create one-hot encoding for each word

Explanation

Removing stop words, tokenizing each word and "lowercase-ing" each word are the recommended pre-processing steps for this use case. N-gram vector and one-hot encoding is not relevant for this use-case. Archaic words should not be removed, since these play a crucial role to determine the sentiment of the sentence. Please review the common text preprocessing steps in more detail on this reference link:

https://medium.com/@annabiancajones/sentiment-analysis-of-reviews-text-pre-processing-6359343784fb

A plumbing company wants to better predict the sales of its flagship copper tubing for the next year. The sales data has copper tubing sizes captured as XS, S, M, L, XL and the retail price of the copper tubing varies with the size. What data preparation steps need to be done for the copper tubing size before it goes into the regression model for prediction:

- ○ Quantile Binning
- ○ One-hot Encoding
- ◉ Categorical Encoding **(Correct)**
- ○ Interval Binning

Explanation

As pricing varies with the size, we need to carry out categorical encoding that is representative of the size of the copper tubing. An example could be like so:

XS -> 2

S -> 4

M -> 7

L -> 10

XL -> 12

One-hot encoding would not capture the price variance with respect to size. Binning (of any type) is not relevant in this case.

The data science team at an analytics company is working on a credit score model using SageMaker Linear Learner algorithm. The training data consists of these fields : name, age, annual salary, gender, employment status and credit score. The model needs to predict the credit score label. Which data preparation steps need to be done before working on the model:

- Drop the age and one-hot-encode name
- Drop the name and one-hot-encode gender and employment status **(Correct)**
- Drop the age and one-hot-encode name and credit score
- Drop the name and one-hot-encode annual salary and credit score

Explanation

As gender and employment status are categorical they need to be one-hot-encoded. Name has no bearing as a useful feature for the model, so it can be discarded. You cannot one-hot encode annual salary as it's not categorical.

A leading technology company offers a fast-track leadership program to the best performing executives at the company. At any point in time, more than a thousand executives are part of this leadership program. Which is the best visualization type to analyze the salary distribution for these executives:

- Bubble Chart
- Histogram **(Correct)**
- Bar Chart
- Pie Chart

Explanation

Histogram is best suited to analyse the underlying distribution of data such as described for this use-case.

Here is a good reference on visualizations in ML:

https://medium.com/data-science-bootcamp/data-visualization-in-machine-learning-beyond-the-basics-baf2cbea8989

An online real estate database company provides information on the housing prices for all states in the US by capturing information such as house size, age, location etc. The company is capturing data for a city where the typical housing prices are around $200K except for some houses that are more than 100 years old with an asking price of about $1 million. These heritage houses will never be listed on the platform. What data processing step would you recommend to address this use-case?

- Normalize the data for all houses in this city and then train the model
- One-hot encode the data for all houses in this city and then train the model
- Drop the heritage houses from the training data and then train the model **(Correct)**
- Standardize the data for all houses in this city and then train the model

Explanation

One-hot encoding is used only for nominal categorical features, so this option is not correct. While normalizing and standardizing is a valid strategy but for this use-case it would end up injecting noise into the model due to the data from the heritage houses. As the heritage houses are clear outliers in terms of price and will never be listed, it is best to drop these from the training data and then train the model.

You want to create an AWS Glue crawler to read the transaction data dumped into an S3 based data lake in the s3://mybucket/myfolder/ location. The transaction data is in CSV format however there are some additional metadata files with .metadata extension in the same location. The metadata needs to be ignored while reading the transaction data via Athena. How would you implement this solution:

- It is not possible to ignore the metadata in crawler. Create a daily ETL job to transfer only the transaction data specific CSV files into a new location and then read this cleansed transaction data into Athena.
- Use exclude pattern .**metadata in the crawler definition to ignore the metadata
- Use exclude pattern **.metadata in the crawler definition to ignore the metadata **(Correct)**
- Use exclude pattern .metadata/** in the crawler definition to ignore the metadata

Explanation

Correct option is to use exclude pattern **.metadata in the crawler definition to ignore the metadata. AWS Glue crawler supports exclude patterns. Please read more details here - https://docs.aws.amazon.com/glue/latest/dg/define-crawler.html#crawler-data-stores-exclude

An analyst is trying to create a box plot for the following data points :

10.2, 14.1, 14.4. 14.4, 14.4, 14.5, 14.5, 14.6, 14.7, 14.7, 14.7, 14.9, 15.1, 15.9, 16.4

Based on these data points, we have the following characteristics :

Q1(25th percentile) = 14.4

Q2(50th percentile) = 14.6

Q3(75th percentile) = 14.9

Identify the data points that would show up as outliers on the box plot (Select three):

- [x] 15.9 **(Correct)**
- [x] 16.4 **(Correct)**
- [x] 10.2 **(Correct)**
- [ ] 14.1
- [ ] 14.4
- [ ] 15.1

Explanation

Interquartile Range (IQR) = Q3-Q1 = 0.5
Minimum outlier cutoff = Q1 - 1.5 * IQR = 14.4 - (1.5*0.5) = 13.65
Maximum outlier cutoff = Q3 + 1.5 * IQR = 14.9 + (1.5*0.5) = 15.65
So the outlier would be anything less than 13.65 or anything more than 15.65. Thus the outliers are 10.2, 15.9, 16.4 for the given problem statement.
More details on the box plot statistical characteristics:
https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

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?

- (•) Athena **(Correct)**
- ( ) EMR Hive
- ( ) Redshift Spectrum
- ( ) EMR Spark

Explanation

With Athena, you can query the data in S3 using SQL. You can either create the table structure in Glue Catalog or let the Glue Crawler collect the metadata and create the table. Athena automatically provisions the resources required for running queries. Redshift Spectrum also provides a capability similar to Athena; however, the query is executed in your Cluster. So, you would need to provision the servers. EMR Hive and Spark are also good options, but it would require provisioning your cluster, and you would also need to figure out how to load the data to the cluster.

You are developing a machine learning model to predict house sale prices based on features of a house. 10% of the houses in your training data are missing the number of square feet in the home. Your training data set is not very large. Which technique would allow you to train your model while achieving the highest accuracy?

- (•) Impute the missing square footage values using kNN **(Correct)**
- ( ) Impute the missing values using deep learning, based on other features such as number of bedrooms
- ( ) Drop all rows that contain missing data
- ( ) Impute the missing values using the mean square footage of all homes

Explanation

Deep learning is better suited to the imputation of categorical data. Square footage is numerical, which is better served by kNN. While simply dropping rows of missing data or using the mean values are a lot easier, they won't result in the best results.

Your company wishes to monitor social media, and perform sentiment analysis on Tweets to classify them as positive or negative sentiment. You are able to obtain a data set of past Tweets about your company to use as training data for a machine learning system, but they are not classified as positive or negative. How would you build such a system?

- ( ) Use RANDOM_CUT_FOREST to automatically identify negative tweets as outliers.
- ( ) Use Amazon Machine Learning with a binary classifier to assign positive or negative sentiments to the past Tweets, and use those labels to train a neural network on an EMR cluster.
- ( ) Stream both old and new tweets into an Amazon Elasticsearch Service cluster, and use Elasticsearch machine learning to classify the tweets.
- (•) Use SageMaker Ground Truth to label past Tweets as positive or negative, and use those labels to train a neural network on SageMaker. **(Correct)**

Explanation

A machine learning system needs labeled data to train itself with; there's no getting around that. Only the Ground Truth answer produces the positive or negative labels we need, by using humans to create that training data initially. Another solution would be to use natural language processing through a service such as Amazon Comprehend.

Data Engineering

The ML solutions team at a leading ecommerce company wants to build a real time fraud detection system. As an ML Specialist, what is the recommended course of action to build such a system with the least number of components and ease of maintenance:

- (●) Ingest the clickstream data into Kinesis Data Streams, which is then written into Kinesis Data Analytics for real time fraud detection using the Random Cut Forest Algorithm. A Lambda function sends an email alert for every fraud detected by the algorithm. The processed stream data is finally sent to Kinesis Data Firehose for subsequent storage in S3. **(Correct)**
- ( ) Ingest the clickstream data into Kinesis Data Streams, which is then written into Kinesis Data Analytics for real time fraud detection and the processed stream data is finally directly written into S3.
- ( ) Ingest the clickstream data into Kinesis Data Analytics for real time Fraud Detection and the data is finally sent to Kinesis Data Firehose for subsequent storage in S3.
- ( ) Ingest the clickstream data into a Spark Streaming application running on EMR cluster to detect fraud records. The application then writes the data into S3.

Explanation

Using EMR cluster would imply managing the underlying infrastructure so it's ruled out. Kinesis Data Analytics cannot directly ingest incoming stream data. Kinesis Data Analytics cannot directly write data into S3. Using the combination of Kinesis Data Streams followed by Kinesis Data Analytics (running RCF algorithm) and then using Kinesis Data Firehose is the correct solution.

A ride-hailing company needs to ingest and store certain attributes of real-time automobile health data which is in JSON format. The company does not want to manage the underlying infrastructure and it wants the data to be available for visualization on a near real time basis. As an ML specialist, what is your recommendation:

- ( ) Ingest the data using a Spark Streaming application running on an EMR cluster. The output data with selected attributes is written in JSON format on S3. Further pipe this data as source into QuickSight for visualizations
- ( ) Ingest the data using Kinesis Data Streams and use a Lambda function to write the stream into S3. Launch a Glue ETL Job every 15 minutes to select specific attributes from the source data and write the output in another S3 location. Further pipe this processed data into QuickSight for visualizations
- ( ) Ingest the data into S3 using Kinesis Firehose. Launch a Glue ETL Job every 15 minutes to write S3 data into RDS. Use RDS Connector to visualize this data in QuickSight
- (●) Ingest the data using Kinesis Firehose that uses a Lambda function to write the selected attributes from the input data stream into an S3 location. Further pipe this processed data into QuickSight for visualizations **(Correct)**

Explanation

Glue ETL is best suited for batch ETL use cases and it's not meant to process near real time data. EMR cluster is not an option as the company does not want to manage the underlying infrastructure. Correct option is to process the streaming JSON data via Kinesis Firehose that uses a Lambda to write the selected attributes as JSON data into an S3 location. This data is then consumed in QuickSight for visualizations.

The data engineering team at an ecommerce company is migrating its data architecture from a data lake to a data warehouse. The current data lake is housed in S3. Which is the fastest way to load this data into Amazon Redshift:

- (•) Use the COPY command to load data from files in S3. **(Correct)**
- ( ) Use the LOAD command to load data from files in S3.
- ( ) Use the INSERT command to load data from files in S3.
- ( ) Use the UNLOAD command to load data from files in S3.

Explanation

There is no such thing as a LOAD command for Redshift. COPY is much faster than insert. UNLOAD is used to write the results of a Redshift query to one or more text files on Amazon S3.

A bioinformatics company wants to automate the secondary analysis of the raw DNA reads into a complete genomic sequence by comparing the multiple overlapping reads and the reference sequence, as well as potentially reduce data errors caused by incorrect alignment between the reference and the sample. Which AWS service can be used to configure and schedule this secondary analysis:

- ( ) Amazon SageMaker
- ( ) AWS Glue
- (•) AWS Batch **(Correct)**
- ( ) Amazon CloudWatch

Explanation

AWS Glue is used to run ETL jobs. SageMaker provides means to train, test and deploy ML services. CloudWatch is used for infrastructure monitoring. AWS Batch can be used to configure and schedule resources:

https://aws.amazon.com/batch/use-cases/

A SageMaker training job with 1TB of training data is taking too long to run. As an ML specialist, which is the most cost effective course of action that requires the least infrastructure management:

- ( ) Since time is the limiting resource, upgrade the SageMaker Training Instance type to the highest possible type so that the job runs quickly
- (•) Convert the training data to recordIO-protobuf file type **(Correct)**
- ( ) Add additional storage to the SageMaker Training Instance
- ( ) Run the training job on an EMR cluster having Amazon SageMaker Spark Library along with the training container image.

Explanation

Converting the data to recordIO-protobuf file type can significantly improve the training time with a marginal increase in cost to store the recordIO-protobuf data on S3. Spinning up EMR clusters would be costly and require complex infrastructure maintenance. Upgrading the existing instance type would incur much higher costs, additionally data would still need to be copied into the instance.

How many shards would a Kinesis Data Stream need if the average record size is 500KB and 2 records per second are being written into the Stream application. Additionally, there are 7 consumer applications using this Kinesis Data Stream Application.

- ( ) 3
- ( ) 7
- ( ) 1
- (•) 4 **(Correct)**

Explanation

number_of_shards = max (incoming_write_bandwidth_in_KB/1000, outgoing_read_bandwidth_in_KB/2000)

where

incoming_write_bandwidth_in_KB = average_data_size_in_KB multiplied by the number_of_records_per_seconds. = 500 * 2 = 1000

outgoing_read_bandwidth_in_KB = incoming_write_bandwidth_in_KB multiplied by the number_of_consumers = 1000 * 7 = 7000

So, number_of_shards = max(1000/1000, 7000/2000) = max(1, 3.5) = 4

So, 4 shards are needed to address this use-case

Some supervised learning algorithms on SageMaker require the training data to be in CSV format. Which constraints should the CSV file meet (Select two):

- [ ] Target variable is in the last column
- [ ] It should have a header record
- [x] Target variable should be in the first column **(Correct)**
- [x] It should not have a header record **(Correct)**

Explanation

For those SageMaker supervised learning algorithms which require the training data to be in CSV format, the target variable should be in the first column and it should not have a header record. You can read more about the common data formats for training:

https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html

AWS Glue jobs can be used to create serverless ETL jobs. Identify the ETL source input type that AWS Glue does not support for ETL jobs:

- ○ DocumentDB
- ◉ Timestream **(Correct)**
- ○ Oracle
- ○ PostgreSQL

Explanation

AWS Glue does not support Timestream as the source input type:

https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-etl-connect.html

An analytics company wants to create a text summarization model based on the SageMaker seq2seq algorithm. However the training data is in the form of 1TB of flat files whereas seq2seq only expects RecordIO-Protobuf format. As an ML Specialist, what solution would you recommend:

- ○ Use AWS Glue to create an ETL job to write the data in RecordIO-Protobuf format on S3. This data can be used by seq2seq based model for training
- ○ Use AWS Step function with Lambda to write the data in RecordIO-Protobuf format on S3. This data can be used by seq2seq based model for training
- ◉ Spin-up an Apache Spark EMR cluster to transform the data from flat files into RecordIO-Protobuf format and save it on S3. This data can be used by seq2seq based model for training **(Correct)**
- ○ Use Kinesis Data Firehose to transform the data into RecordIO-Protobuf format

Explanation

Glue cannot write the output in RecordIO-Protobuf format. Lambda is not suited for long-running processes such as the task of transforming 1TB data into RecordIO-Protobuf format. Kinesis Firehose is not meant to be used for batch processing use cases and it cannot write data in RecorIO-Protobuf format. Apache Spark (running on the EMR cluster in this use-case) can write the output in RecorIO-Protobuf format.

The data science team at an email marketing company has created a data lake with raw and refined zones. The raw zone has the data as it arrives from the source, however, the team wants to de-duplicate the data before it is written into the refined zone. What is the best way to accomplish this with the least amount of development time and infrastructure maintenance effort:

- ○ Create a Lambda function with the code to handle all possible data duplication use cases. Trigger a Lambda function when new files arrive in the S3 raw zone.
- ◉ Invoke an AWS Glue ML Transforms job when new data arrives into raw zone so that de-duplicated records can be written into the refined zone **(Correct)**
- ○ Create an Apache Spark based EMR job and run it once a day to de-duplicate the records from raw zone into refined zone.
- ○ Ingest the raw zone data in Kinesis Data Firehose and process the data using a Lambda function before it is finally dumped into the refined zone.

Explanation

Spark on EMR and Lambda function involve significant development and maintenance effort, so these options are ruled out. Kinesis Data Firehose is used for streaming data scenarios, so it's not the right fit. AWS Glue ML Transforms job can perform deduplication in a serverless fashion and that's the correct choice for this use-case.

https://docs.aws.amazon.com/glue/latest/dg/machine-learning.html

A media company needs to ingest and store a continuous stream of social media data. The source data is in JSON format. The company does not want to manage the underlying infrastructure and it wants the data to be immediately available for ad-hoc analysis. The solution must be cost efficient and scalable. As an ML specialist, what is your recommendation:

- Ingest the data into S3 using Kinesis Firehose. Launch a Glue ETL Job every 15 minutes to write S3 data into RDS. Perform ad-hoc query analysis once data is ingested into RDS.
- Ingest the data using a Spark Streaming application running on an EMR cluster. The output data is written in Parquet format on S3. Use an AWS Glue Crawler to read this data into an Athena table for ad-hoc analysis.
- **Ingest the data using Kinesis Firehose that further transforms the data into Parquet format while writing to S3. Use an AWS Glue Crawler to read this data into an Athena table for ad-hoc analysis. (Correct)**
- Ingest the data using Kinesis Data Streams and use a Lambda function to write the stream into S3. Launch a Glue ETL Job every 15 minutes to convert the data from JSON format to Parquet format. Use an AWS Glue Crawler to read this data into an Athena table for ad-hoc analysis.

Explanation

Kinesis Firehose can transform data to Parquet format and store it on S3 without provisioning any servers. Also this transformed data can be read into an Athena Table via a Glue Crawler and then the underlying data is readily available for ad-hoc analysis. Although Glue ETL Job can transform the source data to Parquet format, it is best suited for batch ETL use cases and it's not meant to process streaming data. EMR cluster is not an option as the company does not want to manage the underlying infrastructure.

An ecommerce company wants to optimize the cost structure for its Redshift data warehouse by moving out some of the infrequently accessed data to S3. What solution would you recommend so that the company can still access this infrequently accessed data from Redshift whenever required:

- Create an EMR based Spark ETL job that writes the data from S3 back into Redshift. The job needs to be triggered every time the data needs to be analysed in Redshift
- Create an AWS Glue ETL job that writes the data from S3 back into Redshift. The job needs to be triggered every time the data needs to be analysed in Redshift
- **Use Redshift Spectrum so that the infrequently accessed data in S3 can be queried from Redshift. (Correct)**
- Create a Glue crawler to read the S3 data into Athena so there is no need to use Redshift

Explanation

EMR and Glue based ETL jobs are not practical as the job needs to be invoked every time data needs to be queried in Redshift. Once the query is done, data needs to be deleted again to save costs. Using Athena is not an option as the query needs to be done in Redshift. Using Redshift Spectrum is the correct choice for this use-case. Please review more details here:

https://docs.aws.amazon.com/redshift/latest/dg/c-using-spectrum.html

Which data set should you use for hyperparameter tuning:

- Test Set
- Training Set
- **Validation Set (Correct)**
- Any of the Training Set, Validation Set or Test Set can be used.

Explanation

Hyperparameters should be tuned against the Validation Set

A financial services company wants to migrate its data architecture from a data warehouse to a data lake. It wants to use a solution that takes the least amount of time and needs no infrastructure management. What options would you recommend to transfer the data from AWS Redshift to S3 (Select two):

- [ ] Apache Spark ETL script running on EMR cluster
- [x] **AWS Glue ETL job (Correct)**
- [x] **AWS Data Pipeline (Correct)**
- [ ] Lambda functions orchestrated by AWS Step Function

Explanation

EMR cluster need to be provisioned and managed, so that option is ruled out. Lambda is not meant to handle ETL workloads, so that is also ruled out. Both AWS Data Pipeline and AWS Glue ETL job are the correct choices for this use-case.

A retail organization ingests 100GB of data into S3 from its global storefronts on a daily basis. This data needs to be cleaned, prepared and analyzed daily so that sales reports can be sent out to the business stakeholders. Which option takes the least effort to make this data available for SQL queries:

- Setup a daily Glue job to write the incremental S3 data into RDS and have it available for SQL queries
- Setup Glue crawlers to initially read the data into Athena tables. Since the data schema does not change, the daily data is readily available for SQL queries in Athena as soon as it arrives **(Correct)**
- Setup a daily Glue job to write the incremental S3 data into DynamoDB and have it available for SQL queries
- Setup a daily Glue job to write the incremental S3 data into Redshift and have it available for SQL queries

Explanation

Using Glue Crawler with Athena is the least effort way to make S3 data available for SQL queries. Using a daily Glue job adds unnecessary complexity into the solution. Also you can't use SQL queries with DynamoDB.

When you create a training job with the SageMaker API, Amazon SageMaker replicates the entire dataset from S3 to ML compute instances by default. What option would you use to replicate only a subset of the data on each ML compute instance:

- Set the S3Uri field to ShardedByS3Key
- Set the S3DataType field to ShardedByS3Key
- Set the S3DataDistributionType field to ShardedByS3Key **(Correct)**
- Set the S3DataDistributionType field to FullyReplicated

Explanation

If you want Amazon SageMaker to replicate a subset of data on each ML compute instance that is launched for model training, specify `ShardedByS3Key` for S3DataDistributionType field.

The data engineering team at a social media company ingests the clickstream data into the Kinesis Data Stream using the PutRecord API in the source system. Now, the team wants to ingest this data into Kinesis Data Firehose instead and they want to use the PutRecord API for Firehose. Identify the differences between the PutRecord API call for Kinesis Data Stream v/s Kinesis Data Firehose:

- Both Kinesis Data Firehose PutRecord API and Kinesis Data Streams PutRecord API use the name of the stream, a partition key and the data blob
- Kinesis Data Streams PutRecord API uses name of the stream, a partition key and the data blob whereas Kinesis Data Firehose PutRecord API uses the name of the delivery stream and the data record **(Correct)**
- Both Kinesis Data Firehose PutRecord API and Kinesis Data Streams PutRecord API use the name of the delivery stream and the data record
- Kinesis Data Firehose PutRecord API uses name of the stream, a partition key and the data blob whereas Kinesis Data Streams PutRecord API uses the name of the delivery stream and the data record

Explanation

Kinesis Data Streams PutRecord API uses name of the stream, a partition key and the data blob whereas Kinesis Data Firehose PutRecord API uses the name of the delivery stream and the data record.

The traffic monitoring authorities at a city want to monitor the traffic at busy intersections and take corrective action at the earliest. An ML solutions company has developed a Proof-of-Concept for processing this video data and now it wants to productionalize it to cover all city intersections. What is the recommended solution stack with the LEAST amount of development effort and ongoing maintenance:

- Process the incoming video data using Spark Streaming on EMR cluster to detect anomalous traffic situations and alert the authorities
- Process the incoming video data using Kinesis Data Analytics to detect anomalies in real time
- Analyse the incoming video streams using Kinesis Video Streams in real time and then send an alert using a downstream EC2 instance **(Correct)**
- Process the incoming video data using Kinesis Data Streams, trigger a Lambda for each stream and then do frame analysis to detect anomalous traffic situations and alert the authorities

Explanation

Using EMR cluster would imply managing the underlying infrastructure so it's ruled out. Kinesis Data Stream and Kinesis Data Analytics cannot directly consume the incoming video data. Kinesis Video Streams is the correct option.

The data engineering team at an ecommerce company wants to ingest the clickstream data from the source system in a reliable way. The solution should provide built-in performance benefits and ease of use on the client side. Which solution would you implement on the source system:

- inesis Client Library
- Kinesis Producer Library **(Correct)**
- Kinesis API
- Spark Streaming

Explanation

Kinesis Product Library provides built-in performance benefits and ease of use advantages. Please review more details here:

https://docs.aws.amazon.com/streams/latest/dev/developing-producers-with-kpl.html#developing-producers-with-kpl-advantage

You are working on a computer vision application (using Convolutional Neural Networks) to recognize an endangered species of tigers. 70% of the incorrectly classified images from the CNN were in a 90-degrees counter-clockwise rotated state. What corrective action you will take to address this issue:

- Perform rigorous hyperparameter tuning to achieve better classification accuracy
- Add more hidden layers to the CNN
- Use Recurrent Neural Network (RNN) for correct classification of all image orientations.
- Procure more training images that are in 90-degrees counter-clockwise rotated state. Retrain the CNN with this new dataset. **(Correct)**

Explanation

RNN is not the right fit for computer vision applications. Adding more hidden layers or hyperparameter tuning are just meant to throw you off. CNN needs to be retrained with more images that are in a 90-degrees counter-clockwise rotated state. A good reference link : https://stats.stackexchange.com/questions/239076/about-cnn-kernels-and-scale-rotation-invariance

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

Explanation

When you create a new IAM user without attaching any policies, the user is not allowed access to any AWS resource. User needs to be granted permissions by assigning policies or by adding them to a Group with necessary permissions. IAM is a global resource – when you create a policy, role, user, or group, they can granted permission to AWS resources in any region.

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?

- Increase timeout
- Increase allocated vCPU
- Use a compute-optimized instance
- Increase allocated memory **(Correct)**

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.

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?

- [x] Parquet **(Correct)**
- [ ] Avro
- [ ] CSV
- [ ] JSON

Explanation

Parquet is a columnar storage format that transparently compresses data. It is a very efficient format for querying a subset of columns across a large number of records. Avro is a suitable binary format that uses row storage and optimized for use cases that need to access the entire row. JSON and CSV are text formats that use Row storage

The marketing analytics team at a financial services company is working on creating a customer loyalty program targeted at specific groups of customers. Which data analysis technique should be used for this goal:

- [ ] Bivariate visualizations
- [x] Clustering **(Correct)**
- [ ] Dimensionality Reduction
- [ ] Multivariate visualizations

Explanation

Clustering is the best way to uncover similar groups. These groups can then be further analyzed to customize the customer loyalty program.

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?

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

Explanation

Glue Crawler is used for automated collection and maintenance of metadata in the Glue Catalog. You would need to configure the Crawler to periodically scan the source data to detect any structural changes in the files and keep the metadata in sync with data. With Athena service, you can query files in S3 using SQL – however, the metadata about the files needs to be created in Glue Catalog first. Lambda function is used for do-it-yourself catalog management. This requires more effort when compared to using Glue Crawler

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)

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

Explanation

You can enable S3 versioning to keep the older version of the objects. You can create life cycle policies to remove the older version after 30 days. Cross-region can help in protecting against accidental deletion and disaster recovery by keeping a copy of data in a different region. But it is more expensive as a full copy of your bucket is maintained in another region. Moving the deleted objects to another bucket is unnecessary and requires other components.

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

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

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 resources in the region

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 **(Correct)**
- ( ) Enforce controls like these at the application level
- ( ) Implement IAM Access Policy to remove delete access or modify access
- ( ) Replicate Data to another read-only bucket

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

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 Glue ETL to run Spark ETL scripts and configure it as a scheduled job **(Correct)**
- ( ) Use Kinesis Firehose for reading the data from S3 and use built-in transformation to store the results in Parquet format
- ( ) Use Kinesis Datastreams for collecting 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

Explanation

Glue ETL provides an easy option to automatically generate ETL scripts and run the script as a scheduled job. Glue ETL provisions required Spark infrastructure to run the job and automatically terminates the environment after the job is completed.

A solution involving Kinesis Firehose requires an additional component to read data from S3 and add it firehose stream. For large files, you would also need to chunk into many messages when adding to the firehose.

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

- ( ) Virtualization infrastructure
- (•) Configuring Access to S3 bucket based on job role **(Correct)**
- ( ) Patching Host Operating System
- ( ) Physical security of hardware

Explanation

Under the shared responsibility model, data security is the responsibility of the customer. AWS provides capabilities to manage data security; however, it is up to the customer to take advantage of security capabilities based on their individual needs. Physical infrastructure, Facilities, Host Computers (underlying physical servers on which virtual instances run), Network infrastructure are all responsibilities of AWS.

Patching Host Operating System is AWS responsibility - here, host refers to the Physical server.

Patching Guest/Instance Operating System is Customer responsibility - here, instance refers to the virtual instance that customer created.

Additional reading and references: https://wa.aws.amazon.com/wat.concept.shared-resp-model.en.html,

Security is Job Zero https://youtu.be/T7MnJOfOVcY

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 a single instance **(Correct)**
- SageMaker Endpoint with multiple instances

Explanation

Each AWS region consists of three or more availability zones. Availability Zones are physically separate infrastructure. Among the choices presented, a SageMaker Endpoint that has only one instance to handle inference requests may be impacted if that instance is running in that Availability Zone. To improve Availability, for production workloads, you need to use at least two instances behind a SageMaker Endpoint – SageMaker will ensure that the instances are deployed in different availability zones. S3 automatically replicates data in three or more availability zones, and S3 can transparently handle availability zone failure. Managed AI Services like Rekognition is also multi-availability zone enabled and can handle availability zone failures automatically

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?

- AMI Provider
- EC2 Support
- Customer **(Correct)**
- AWS

Explanation

For Infrastructure as a Service (IaaS) products like EC2, the customer who launched the instance is responsible for adequately patching the instance. AWS is responsible for keeping AMI up-to-date. Once the EC2 instance is launched, only the customer can patch the instance. Reference: Security is Job Zero https://youtu.be/T7MnJOfOVcY

A large news website needs to produce personalized recommendations for articles to its readers, by training a machine learning model on a daily basis using historical click data. The influx of this data is fairly constant, except during major elections when traffic to the site spikes considerably. Which system would provide the most cost-effective and simplest solution?

- (●) Publish click data into Amazon S3 using Kinesis Firehose, and process the data nightly using Apache Spark and MLLib using spot instances in an EMR cluster. Publish the model's results to DynamoDB for producing recommendations in real-time. **(Correct)**
- ( ) Publish click data into Amazon S3 using Kinesis Firehose, and process the data nightly using Apache Spark and MLLib using reserved instances in an EMR cluster. Publish the model's results to DynamoDB for producing recommendations in real-time.
- ( ) Publish click data into Amazon Elasticsearch using Kinesis Firehose, and query the Elasticsearch data to produce recommendations in real-time.
- ( ) Publish click data into Amazon S3 using Kinesis Streams, and process the data in real time using Splunk on an EMR cluster with spot instances added as needed. Publish the model's results to DynamoDB for producing recommendations in real-time.

Explanation

The use of spot instances in response to anticipated surges in usage is the most cost-effective approach for scaling up an EMR cluster. Kinesis streams is over-engineering because we do not have a real-time streaming requirement. Elasticsearch doesn't make sense because Elasticsearch is not a recommender engine.

You are training an XGBoost model on SageMaker with millions of rows of training data, and you wish to use Apache Spark to pre-process this data at scale. What is the simplest architecture that achieves this?

- ( ) Use Amazon EMR to pre-process your data using Spark, and use the same EMR instances to host your SageMaker notebook.
- ( ) Use Sparkmagic to pre-process your data within a SageMaker notebook, transform the resulting Spark DataFrames into RecordIO format, and then use Spark's XGBoost algorithm to train the model.
- ( ) Use Amazon EMR to pre-process your data using Spark, and then use AWS Data Pipelines to transfer the processed training data to SageMaker
- (●) Use sagemaker_pyspark and XGBoostSageMakerEstimator to use Spark to pre-process, train, and host your model using Spark on SageMaker. **(Correct)**

Explanation

The SageMakerEstimator classes allow tight integration between Spark and SageMaker for several models including XGBoost, and offers the simplest solution. You can't deploy SageMaker to an EMR cluster, and XGBoost actually requires LibSVM or CSV input, not RecordIO.

An Analytics Consulting Firm would like to capture and analyse the real time metrics for a cab hailing service. The Firm would like to identify "demand hotspots" in real time so that additional cabs can be dispatched to meet the sudden spurt in demand. What is the least effort way of building a real time analytics solution for this use case :

- Ingest the data into Kinesis Data Streams and immediately write the stream into Kinesis Data Analytics for SQL based analysis so that appropriate alerts can be sent to the drivers. Once processing is done, the streams are dumped into S3 using Kinesis Data Firehouse. **(Correct)**
- Ingest the data into Kinesis Data Streams that writes the data into a Spark Streaming application running on an EMR cluster. Once the processing is done, the output is written on S3
- Ingest the source data directly into Kinesis Data Analytics so that real time analytics can be done without any processing delay. Once processing is done, the streams are dumped into S3 using Kinesis Data Firehouse.
- Ingest the data into Kinesis Data Firehose and write into S3, which triggers a Lambda that analyses the event data. The Lambda finally writes the output to S3.

Explanation

Kinesis Data Analytics cannot directly ingest source data. Using a combination of Kinesis Firehose with lambda would introduce a buffering delay of at least 1 minute or 1MB of data, so the solution will not be real time. Using EMR would significantly increase the development and maintenance effort, so it's not the right choice. Correct solution is to ingest the data into Kinesis Data Streams and immediately write the stream into Kinesis Data Analytics for SQL based analysis so that appropriate alerts can be sent to the drivers. Once processing is done, the streams are dumped into S3 using Kinesis Data Firehouse.

Modeling

The data science team at a leading Questions and Answers website wants to improve the user experience and therefore would like to identify duplicate questions based on similarity of the text found in a given question. As an ML Specialist, which SageMaker algorithm would you recommend to help solve this problem:

- (●) Object2Vec **(Correct)**
- ( ) XGBoost
- ( ) BlazingText Word2Vec mode
- ( ) Factorization Machines

Explanation

Object2Vec can be used to find semantically similar objects such as questions. BlazingText Word2Vec can only find semantically similar words. Factorization Machines and XGBoost are not fit for this use-case. A good reference read for Object2Vec:

https://aws.amazon.com/blogs/machine-learning/introduction-to-amazon-sagemaker-object2vec/

The data science team at a financial services company has created a multi-class classification model to segment the company's customers into three tiers - Platinum, Gold and Silver. The confusion matrix for the underlying model was reported as follows:

	Actual - Platinum	Actual - Gold	Actual - Silver
Predicted - Platinum	30	20	10
Predicted - Gold	50	60	10
Predicted - Silver	20	20	80

What is the overall precision for this multiclass classification model:

- ( ) 0.67
- ( ) 0.47
- (●) 0.56 **(Correct)**
- ( ) 0.50

Explanation

Precision for Platinum = (True Positives / (True Positives + False Positives))

= (30/(30+ (20+10))) = 30/60 = 0.50

Precision for Gold = (True Positives / (True Positives + False Positives))

= (60/(60+ (50+10))) = 60/120 = 0.50

Precision for Silver = (True Positives / (True Positives + False Positives))

= (80/(80+ (20+20))) = 80/120 = 0.67

Overall Precision = Average of the precision for Platinum, Gold and Silver

= (0.50+0.50+0.67)/3 = 0.56

What technique would you use in SageMaker to train a new model using an expanded dataset that contains an underlying pattern that was not accounted for in the previous training and which resulted in poor model performance:

- [x] Incremental Training **(Correct)**
- [ ] Batch Training
- [ ] Beta Testing
- [ ] Transfer Learning

Explanation

Batch Training and Beta Testing are distractors meant to throw you off track. Transfer Learning is a generic technique not relevant to the SageMaker specific situation described in the given use-case. Incremental Training is the correct choice. Please read more on this reference link -

https://docs.aws.amazon.com/sagemaker/latest/dg/incremental-training.html

The marketing team at an Enterprise SaaS company has determined that the cost of customer churn is much greater than the cost of customer retention for its existing customer base. To address this issue, the team worked on a classification model to predict if a customer is likely to churn and boiled it down to two model variants. Model A had 92% accuracy with 40 False Negatives (FN) and 100 False Positives (FP) whereas model B also had 92% accuracy with 100 FN and 40 FP. Which of the two models is more cost effective for the company :

- [ ] None of the Model A and Model B are cost effective. Company needs to try something different.
- [ ] Model A **(Correct)**
- [ ] Both Model A and Model B are equally cost effective, as accuracy is same
- [x] Model B

Explanation

The classification model predicts if a customer is likely to churn. This implies that a False Negative is very costly for the company because the model predicted that the customer will not churn, however in reality the customer did churn. So the ideal model would focus on reducing the False Negatives. Thus Model A is the right choice.

A company has one year of raw data on demographics and sales for existing customers. The digital marketing executives at the company want to identify potential customers on social media. As the festive season is coming up, the solution needs to be built in the shortest possible time. What is the best course of action for this goal:

- [ ] Use XGBoost to predict the best matching customers from their social media profiles
- [ ] Use Recommendation Systems to build the existing customer profiles and then predict the best matching customers from their social media profiles
- [ ] Use Linear Regression to predict the best matching customers from their social media profiles
- [x] Use K-means to identify groups of customers and then find similar customers on social media. **(Correct)**

Explanation

Given that the solution needs to be built in the shortest possible time, as there is no labeled data (for both customers and non-customers) to be used for supervised learning techniques such as Linear Regression and XGBoost, these two are ruled out. Recommendation Systems also need significant time investment for either collaborative-filtering or content-based approaches. K-means being unsupervised is the right choice.

The data science team at an Email Service Provider has determined that the long term cost of marking a bona fide email as spam is much greater than the cost of marking a spam email as bona fide. To address this issue, the team worked on a classification model to predict if an email is spam and boiled it down to two model variants. Model A had 95% accuracy with 40 False Negatives (FN) and 100 False Positives (FP) whereas model B also had 95% accuracy with 100 FN and 40 FP. Which of the two models is more cost effective for the company :

- [ ] Both Model A and Model B are equally cost effective, as accuracy is same
- [x] Model B **(Correct)**
- [ ] None of the Model A and Model B are cost effective. Company needs to try something different.
- [ ] Model A

Explanation

The classification model predicts if an email is spam. This implies that a False Positive is very costly for the company because the model predicted that the email is spam however in reality the email turned out to be bona fide. So the ideal model would focus on reducing the False Positives. Thus Model B is the right choice.

An FMCG company has 33 shampoo and 37 conditioner variants in its product portfolio. Senior executives are planning to launch a hybrid product with features from its shampoo and conditioner portfolio. Given the lack of reference historical data for this hybrid product, which AWS SageMaker algorithm can help the executives in predicting the product sales over the next financial year:

- ○ Linear Learner
- ○ XGBoost
- ◉ DeepAR **(Correct)**
- ○ Factorization Machines

Explanation

SageMaker DeepAR algorithm specializes in forecasting new product performance. Other algorithms are not a good fit for this use case.

The ML team at a research lab have trained a Deep Neural Network using a huge training dataset. After a series of training runs, the team observes that the training error is much lower than the test error. What steps would you recommend to address this issue (Select three) :

- ☑ Use dropout **(Correct)**
- ☑ Use early stopping while training **(Correct)**
- ☐ Do not use early stopping while training
- ☑ Add parameter regularization **(Correct)**
- ☐ Remove parameter regularization

Explanation

This use-case has the telltale signs of overfitting in a Deep Learning context. Please review this material for a deep-dive on preventing overfitting for Deep Learning based solutions:

https://www.jeremyjordan.me/deep-neural-networks-preventing-overfitting/

An online retail company specializing in fashion wear wants to automate the various categories of fashion wear in their catalog. They have about 50,000 images of their product catalog but none of them have the right labels for the associated categories. As an ML Specialist, what solution will you recommend to build the training data with the correct labels:

- ○ Use SageMaker Image Classification to create the category labels for the training images
- ○ Use SageMaker Semantic Segmentation to create the category labels for the training images
- ○ Use AWS Rekognition to create the category labels for the training images
- ◉ Use SageMaker Ground Truth to create the category labels for the training images **(Correct)**

Explanation

Image Classification, Semantic Segmentation or Rekognition cannot be used to create labels for training data. Ground Truth is the correct service for this use-case.

A bio-technology company has invented a new drug testing procedure that can identify substance abuse from blood samples within a minute. The law stipulates that anyone found indulging in substance abuse faces a steep fine along with a prison term. Identify the metric that the data scientists at the company need to focus on, so that they can analyze the results of the trials for the underlying model (The model's predicted value of 1 implies that the individual was predicted to have consumed drugs):

- ○ F1-score
- ◉ Specificity **(Correct)**
- ○ Recall
- ○ Accuracy

Explanation

Specificity = (True Negatives / (True Negatives + False Positives))

If the model has a high specificity, it implies that all false positives (think of it as false alarms) have been weeded out. In other words, the **specificity** of a test refers to how well the test identifies those who have not indulged in substance abuse.

Please read this excellent reference article for more details:

https://www.statisticshowto.datasciencecentral.com/sensitivity-vs-specificity-statistics/

Executives at a leading smartphone brand are contemplating the launch of a radical new phone model with never-before-seen features. Given the lack of reference historical data for similar phone models, which AWS SageMaker algorithm can help the executives in predicting the product sales over the next quarter for this innovative phone:

- Factorization Machines
- DeepAR **(Correct)**
- Linear Learner
- XGBoost

Explanation

SageMaker DeepAR algorithm specializes in forecasting new product performance. Other algorithms are not a good fit for this use case.

A retail organization wants to forecast the sales of its flagship products for the upcoming festive season. They have the last 5 years of sales data for these products. As an ML specialist, which algorithm would you use to implement the forecasting solution:

- Linear Learner **(Correct)**
- Semantic Segmentation
- Latent Dirichlet Allocation (LDA)
- Random Cut Forest

Explanation

This is a regression problem that can be easily solved using the Linear Learner algorithm. LDA is used for topic modeling, Random Cut Forest is used to detect outliers and Semantic Segmentation is used for image analysis.

A Silicon Valley startup has introduced a new email service that would completely eradicate spam from the inbox. The data scientists at the startup have prepared the following confusion matrix for the underlying model. What is the F1 score for the underlying model:

	Actual : Positive	Actual : Negative
Predicted : Positive	8000	2000
Predicted : Negative	1000	3000

- 0.73
- 0.63
- 0.93
- 0.84 **(Correct)**

Explanation

Precision (P) = (True Positives / (True Positives + False Positives))

= (8000/(8000+2000)) = 0.8

Recall (R) = (True Positives / (True Positives + False Negatives))

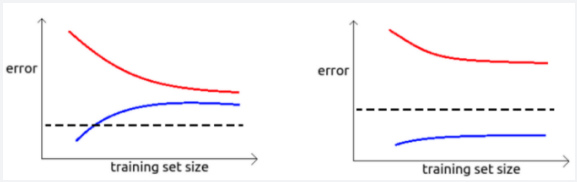
= (8000/(8000+1000)) = 0.89

F1 score = (2 * P * R) / (P+R) = (2 * 0.8 * 0.89) / (0.8+0.89) = 0.84

Please review the reference link for F1-score :

https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Learning curve is a graph that compares the performance of a model on training andx` validation datasets over a varying number of training instances. Following figure illustrates the learning curve for two separate models:



The learning curve on the left was captured for a model running the Amazon SageMaker Linear Learner algorithm and the learning curve on the right was captured for another model running the Amazon SageMaker XGBoost algorithm. The horizontal line with dashes represents the desired performance. The curve in red represents the validation dataset and the curve in blue represents the training dataset. Identify two ways in which you can fix the issues plaguing the model on the left (Select two):

- [x] Decrease regularization parameters **(Correct)**
- [ ] Increase regularization parameters
- [ ] Remove features from the model
- [ ] Get more training data
- [x] Add more features to the model **(Correct)**

Explanation

Learning curve on the left represents a model that is underfitting (has bias) because both the training and validation error are high. You can address this by adding more features or by decreasing regularization parameters.

A Neural Network is tasked with classifying one of the ten breeds of dogs for a scientific experiment. Which activation function should be used in the output layer of this Neural Network:

- (•) Softmax **(Correct)**
- ( ) RELU
- ( ) Sigmoid
- ( ) Tanh

Explanation

Softmax is used for classification examples where multiple classes of results can be computed. You can read more here -

https://medium.com/fintechexplained/neural-network-activation-function-types-a85963035196

After training a SageMaker Linear Learner model over a huge training dataset, the data science team observed that it achieved high accuracy on the training data, but had low accuracy on the test data. As an ML specialist, what are the three techniques that you will recommend to help resolve this problem? (Select three):

- [x] Add regularization to the model **(Correct)**
- [ ] Use less training data
- [x] Use more training data **(Correct)**
- [ ] Use more features in the model
- [x] Use less features in the model **(Correct)**
- [ ] Remove regularization from the model

Explanation

Description : As the model has high accuracy on the training data but low accuracy on the test data, it suggests that the model is overfitting. When a model is overfitting then adding more training data, adding regularization or using less features can help in addressing the underlying problem.

Identify the three step process for training with the Amazon SageMaker k-nearest neighbors (k-NN) algorithm (Select three):

- [x] Sampling **(Correct)**
- [x] Dimension Reduction **(Correct)**
- [ ] One-Hot Encoding
- [ ] Data Engineering
- [x] Index Building **(Correct)**

Explanation

Data Engineering and One-Hot Encoding are made-up options. Training with the k-NN algorithm has three steps: sampling, dimension reduction, and index building:

https://docs.aws.amazon.com/sagemaker/latest/dg/k-nearest-neighbors.html

Which SageMaker algorithm comes in both supervised and unsupervised learning modes:

- ( ) Random Cut Forest
- ( ) Latent Dirichlet Allocation
- (•) Blazing Text **(Correct)**
- ( ) XGBoost

Explanation

Blazing Text algorithm can be used in both supervised and unsupervised learning modes:

https://docs.aws.amazon.com/sagemaker/latest/dg/blazingtext.html

You are developing a multi-class classification model with Sagemaker XGBoost algorithm using the AWS SDK for Python (Boto 3). After calling the create_training_job() method to start the training job, you would like to get a status about the progress of the training job. Which method can be used to get the training job status:

- get_training_job
- describe_training_status
- describe_training_job **(Correct)**
- get_training_status

Explanation

describe_training_job is the correct method to get the training job status. Get_training_job, get_training_status and describe_training_status are made up options. Please read more about the various method calls for the AWS SDK for Python (Boto 3) related to SageMaker training and job description:

https://docs.aws.amazon.com/sagemaker/latest/dg/ex1-train-model.html

In order to run unsupervised algorithms on SageMaker you need to configure content type parameter. How should you specify the number of label columns in the content type:

- application/csv;label_size=None
- application/csv;label_size=0
- text/csv;label_size=0 **(Correct)**
- text/csv;label_size=None

Explanation

To run unsupervised learning algorithms that don't have a target, specify the number of label columns in the content type. In this case 'text/csv;label_size=0' is the correct option. Reference link:

https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html

A medical diagnostics company specializes in cancer detection tests. The data scientists at the company need to focus on which metric for the underlying classification model in order to not miss any cases of cancer (The model's predicted value of 1 implies that the patient is predicted to have cancer):

- Accuracy
- Precision
- Recall **(Correct)**
- F1-score

Explanation

Recall = (True Positives / (True Positives + False Negatives))

The company would like to be extra sure that a patient does not have cancer before they pronounce them healthy. This implies that they want less false negatives. As false negatives decrease, the model would have a higher recall, so recall is the metric to focus on.

While using the K-means SageMaker algorithm, which strategies are available to determine how the initial cluster centers are selected (Select two):

- [x] Random approach **(Correct)**
- [x] k-means++ **(Correct)**
- [ ] Euclidean distance
- [ ] Within Cluster Sum of Squares

Explanation

Within Cluster Sum of Squares is a valid strategy to use in the generic k-means algorithm, however, this is not an option for the SageMaker K-means algorithm. Euclidean distance is a made up option. Random approach and k-means++ are the correct choices:

https://docs.aws.amazon.com/sagemaker/latest/dg/algo-kmeans-tech-notes.html

A sports enthusiast wants to build a mobile app that is able to recognize celebrity athletes. Which AWS service can help him set this up with minimum possible effort:

- [x] Amazon Rekognition **(Correct)**
- [ ] Amazon Predict
- [ ] Amazon Polly
- [ ] Amazon Lex

Explanation

Rekognition has a celebrity recognition feature that can be leveraged for this functionality. Predict is not an Amazon service. Lex and Polly do not fit this use case. You can read more here - https://aws.amazon.com/machine-learning/ai-services/

Researchers at a university claim a breakthrough in early cancer detection. The lab results for a series of trials yield the following confusion matrix. What is the recall of the underlying model:

	Actual : Yes	Actual : No
Predicted : Yes	80	20
Predicted : No	10	30

- [ ] 80%
- [ ] 20%
- [ ] 50%
- [x] 89% **(Correct)**

Explanation

Recall = (True Positives / (True Positives + False Negatives)) = (80/(80+10)) = 0.89 or 89%

Identify the correct definitions for the Latent Dirichlet Allocation algorithm:

- [ ] Observations are referred to as words. The feature set is referred to as vocabulary. A feature is referred to as a document. And the resulting categories are referred to as topics.
- [x] Observations are referred to as documents. The feature set is referred to as vocabulary. A feature is referred to as a word. And the resulting categories are referred to as topics. **(Correct)**
- [ ] Observations are referred to as topics. The feature set is referred to as word. A feature is referred to as a vocabulary. And the resulting categories are referred to as documents.
- [ ] Observations are referred to as vocabulary. The feature set is referred to as documents. A feature is referred to as a topic. And the resulting categories are referred to as words.

Explanation

Observations are referred to as documents. The feature set is referred to as vocabulary. A feature is referred to as a word. And the resulting categories are referred to as topics.

https://docs.aws.amazon.com/sagemaker/latest/dg/lda-how-it-works.html

The marketing analytics team at a leading bank has created a multi-class classification model to segment the bank's customers into three tiers - Platinum, Gold and Silver. The confusion matrix for the underlying model was reported as follows:

	Actual - Platinum	Actual - Gold	Actual - Silver
Predicted - Platinum	30	20	10
Predicted - Gold	50	60	10
Predicted - Silver	20	20	80

What is the overall recall for this multiclass classification model:

- [ ] 0.20
- [ ] 0.46
- [ ] 0.30
- [x] 0.57 **(Correct)**

Explanation

Recall for Platinum Tier = (True Positives / (True Positives + False Negatives)) = (30/(30+ (50+20))) = 30/100 = 0.30

Recall for Gold Tier = (True Positives / (True Positives + False Negatives)) = (60/(60+ (20+20))) = 60/100 = 0.60

Recall for Silver Tier = (True Positives / (True Positives + False Negatives)) = (80/(80+ (10+10))) = 80/100 = 0.80

Overall Recall = Average of the recall for Platinum, Gold and Silver Tiers = (0.30+0.60+0.80)/3 = 0.57

You are creating a computer vision application to recognize truck brands. Your application uses Convolutional Neural Networks (CNN) but you do not have enough data to train the model. However, there are pre-trained third-party image recognition models available for similar tasks. What steps will you take to build your solution in the shortest possible duration:

- Use Kinesis Video Streams to identify the truck brand by using image manipulation algorithms and then do a pixel by pixel comparison
- Use transfer learning with Kinesis Video Streams
- Use transfer learning in your CNN by using the pre-trained third-party image recognition model as the convolutional base. Then remove the original classifier from the pre-trained model and add the new classifier for recognizing truck brands. **(Correct)**
- Use transfer learning by retraining the pre-trained third-party image recognition model with your own data.

Explanation

You cannot use transfer learning with Kinesis Video Streams. Retraining the pre-trained model with your own data is not correct because you do not have enough data to train. The correct option is to use transfer learning in your CNN by using the pre-trained third-party image recognition model as the convolutional base. Then remove the original classifier from the pre-trained model and add the new classifier for recognizing truck brands. Please read this excellent reference on using transfer learning in computer vision applications:

https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751

A financial services company has the goal of reducing fraud transactions by 10% over the next financial year. In order to achieve this goal, which of the following is the most relevant model evaluation metric that the data scientists at the company need to focus on (the model's predicted value of 1 implies that the transaction is predicted to be fraud) :

- Recall
- Precision
- Precision-Recall Area-Under-Curve (PR AUC) **(Correct)**
- F1-score

Explanation

This is an example where the dataset is imbalanced with fewer instances of positive class because of a fewer number of actual fraud records in the dataset. In such scenarios where we care more about the positive class, then using PR AUC is a better choice, which is more sensitive to the improvements for the positive class.

PR AUC is a curve that combines precision (PPV) and Recall (TPR) in a single visualization. For every threshold, you calculate PPV and TPR and plot it. The higher on y-axis your curve is the better your model performance.

Please review these excellent resources for a deep-dive into PR AUC.

https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

https://machinelearningmastery.com/imbalanced-classification-with-the-fraudulent-credit-card-transactions-dataset/

An entrepreneur wants to launch the next unicorn business with futuristic Business Intelligence features. The product would allow business managers to gather business insights using a voice based interface rather than typing tedious SQL commands. Which AWS services would you use to build this product with the least amount of time and resources (Select three):

- [ ] Translate
- [ ] Comprehend
- [x] Lex **(Correct)**
- [x] Transcribe **(Correct)**
- [x] Polly **(Correct)**

Explanation

Transcribe can be used to convert speech to text. This text can be fed into Lex that uses the pre-configured Intents and Entities to come back with the most relevant text response. In the end, this text response would be converted to speech via Polly.

After training a SageMaker XGBoost based model over a huge training dataset, the data science team observed that it has low accuracy on the training data as well as low accuracy on the test data. As an ML specialist, what are the two techniques that you will recommend to help resolve this problem? (Select two):

- [ ] Use less training data
- [ ] Add regularization to the model
- [ ] Use more training data
- [x] Use more features in the model **(Correct)**
- [x] Remove regularization from the model **(Correct)**
- [ ] Use less features in the model

Explanation

As the model has low accuracy on the training data as well as low accuracy on the test data, it suggests that the model is biased. When a model is biased then adding more features to the model or removing regularization can help in addressing the underlying problem. In case of a biased model, adding more training data may or may not help.

Which SageMaker unsupervised learning algorithms can be used for Fraud Detection (Select two):

- [x] Random Cut Forest **(Correct)**
- [ ] Factorization Machines
- [x] IP Insights **(Correct)**
- [ ] Object2Vec

Explanation

Random Cut Forest and IP Insights can be used for Fraud Detection. Please review these reference links:

https://docs.aws.amazon.com/sagemaker/latest/dg/randomcutforest.html

https://docs.aws.amazon.com/sagemaker/latest/dg/ip-insights.html

You are building a Deep Learning based context-sensitive spelling correction functionality for a consumer facing application. For example, consider the following misspelled food description: "low tat milk". A traditional spell checker might correct it to "low tar milk", which is an inappropriate suggestion for the domain of food text, therefore it should be corrected to "low fat milk". What technique would you use to build your context aware model:

- ( ) Levenstein distance
- ( ) Word2Vec
- (•) Seq2seq **(Correct)**
- ( ) N-grams

Explanation

N-grams, Levenstein distance, Word2Vec do not fit into this use-case. The correct solution would be based on a variation of the seq2seq model. Please read this case study for more details:

https://makers.underarmour.com/context-sensitive-spell-correction-with-deep-learning/

A Silicon Valley startup has introduced a new email service that aims to address the problem of email spam. The startup also wants to make sure that genuine emails are not marked as spam. The underlying machine learning model's predicted value of 1 implies that the model predicts a given email to be spam. The data scientists at the startup would like to analyze the results of the underlying model. Identify the most important evaluation metric for this model:

- Recall
- F1-score
- Accuracy
- Precision **(Correct)**

Explanation

Given, Precision = (True Positives / (True Positives + False Positives))

The startup would like to be extra sure that an email is spam before potentially putting in the spam folder. According to the underlying model, a false positive implies that the email was actually genuine, but the model predicted it to be spam. Therefore in the case of a false positive, the user never sees this genuine email as the mail went to the spam folder instead. This implies that the data scientists would want to have less false positives from the model. As false positives decrease, the model would have a higher precision per the formula quoted above. Hence precision is the most important evaluation metric for this model.

To understand more about precision and recall, please read : https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Considering the following ROC curve generated for the Amazon SageMaker XGBoost algorithm for a binary classification use-case

Which of the following statements are correct:

- The model represented by the red ROC curve is best at distinguishing the two classes. The model represented by the blue ROC curve is worst at distinguishing the two classes.
- The model represented by the black ROC curve is best at distinguishing the two classes. The model represented by the red ROC curve is worst at distinguishing the two classes.
- The model represented by the blue ROC curve is best at distinguishing the two classes. The model represented by the black ROC curve is worst at distinguishing the two classes. **(Correct)**
- The model represented by the black ROC curve is best at distinguishing the two classes. The model represented by the blue ROC curve is worst at distinguishing the two classes.

Explanation

Please review the concept of AUC/ROC as applied to binary classification. Here is a good reference:

https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

A marketing analyst wants to group current and prospective customers into 10 groups based on their attributes. He wants to send mailings to prospective customers in the group which has the highest percentage of current customers. As an ML Specialist, which Sagemaker algorithm would you recommend as a solution:

- KNN
- K-means **(Correct)**
- Latent Dirichlet Allocation
- PCA

Explanation

As there is no historic data with labels, so KNN is ruled out. PCA is used for dimensionality reduction and LDA for topic modeling. K-means is the right algorithm to uncover discrete groupings within data.

An analytics company is doing the sentiment analysis of tweets about a leading sports event. The company has prepared the following confusion matrix. What is the precision of the underlying model:

	Actual : Positive	Actual : Negative
Predicted : Positive	800	200
Predicted : Negative	100	300

- 50%
- 20%
- 89%
- 80% **(Correct)**

Explanation

Precision = (True Positives / (True Positives + False Positives))

= (800/(800+200)) = 0.8 or 80%

The data science team at an ecommerce company has been tasked to improve the underlying recommendations engine. The team has tried to use both the random search and bayesian search approaches to hyperparameter tuning, but the results have been inconsistent. As an ML Specialist, what would be your suggestion to the team:

- Increase the number of concurrent hyperparameter tuning jobs to 100
- Increase the maximum run time for a hyperparameter tuning job to 30 days
- Hyperparameter tuning might not improve your model and look at other options such as data engineering or alternate algorithm to improve the model **(Correct)**
- Increase the number of hyperparameters that can be searched to 20

Explanation

Hyperparameter tuning is not a panacea for model under-performance and you may need to look at other options such as data engineering or alternate algorithm for the model. The other options are actually the upper limits for some of the hyperparameter tuning resources, they have no direct bearing on this use-case. You can read more on this reference link :

https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-how-it-works.html

You are working on a click prediction system and you want to capture the underlying click rate patterns when ads from a certain ad-category are placed on pages from a certain page-category. Which SageMaker algorithm can be used to accomplish this:

- Latent Dirichlet Analysis
- XGBoost
- Factorization Machines **(Correct)**
- Blazing Text

Explanation

Factorization Machine can be used to capture click patterns for a click prediction system:

https://docs.aws.amazon.com/sagemaker/latest/dg/fact-machines.html

You are working on a fraud detection model based on SageMaker IP Insights algorithm with a training data set of 1TB in CSV format. Your Sagemaker Notebook instance has only 5GB of space. How would you go about building your model, given these constraints:

- Spin-up an EMR Cluster running Apache Spark to transform the CSV data into recordIO-protobuf format. Read the entire transformed data in recordIO-protobuf format from S3 in your Jupyter Notebook instance while training your model.
- Create an AWS Glue job to transform the training data into recordIO-protobuf format. Read the entire transformed data in recordIO-protobuf format from S3 in your Jupyter Notebook instance while training your model.
- **(selected)** Shuffle the training data and create a 5GB slice of this shuffled data. Build your model on the Jupyter Notebook using this slice of training data. Once the evaluation metric looks good, create a training job on SageMaker infrastructure with the appropriate instance types and instance counts to handle the entire training data. **(Correct)**
- Create an AWS Glue job to compress the training data into parquet format using an appropriate compression codec. This should allow you to use the entire compressed training data on your notebook instance.

Explanation

IP Insights algorithm supports only CSV file type as training data, so other options using parquet or recordIO-protobuf are ruled out. An important aside, AWS Glue job cannot write output in recordIO-protobuf format. The correct option is to shuffle the training data and create a 5GB slice of this shuffled data. Build your model on the Jupyter Notebook using this slice of training data. Once the evaluation metric looks good, create a training job on SageMaker infrastructure with the appropriate instance types and instance counts to handle the entire training data.

An upcoming music streaming service wants to build a Minimum Viable Product and would like to have the underlying music recommendation engine developed at the earliest with the least development effort. As an ML Specialist, which AWS service would you suggest for the music recommendation engine:

- **(selected)** Amazon Personalize **(Correct)**
- Amazon SageMaker Factorization Machines
- Amazon SageMaker XGBoost
- Amazon SageMaker Neural Topic Model

Explanation

Amazon Personalize is a machine learning service that makes it easy for developers to create individualized recommendations for customers using their applications. Other options require significant effort to train and test the models. Please read more:

https://aws.amazon.com/personalize/

An ML specialist is examining the root cause for underperformance of a regression model and has a hunch that it is consistently overestimating the outcome. Which metrics should he track on a chart to help identify any pattern of model overestimation:

- Mean Absolute Error
- Residuals **(Correct)**
- RMSE
- AUC

Explanation

The residuals plot would indicate any trend of underestimation or overestimation. Both Mean Absolute Error and RMSE would only give the error magnitude. AUC is a metric used for classification models.

A marketing manager at an email marketing company is trying to figure out the right content and subject line for emails to be sent to the customers. The manager has access to all the historical responses to mailings from each customer. As an ML Specialist, which SageMaker algorithm would you recommend as a solution:

- Blazing Text
- Linear Learner
- Factorization Machines **(Correct)**
- XGBoost

Explanation

Factorization Machines algorithm specializes in building recommendation systems for such use-cases.

You would like to tune the hyperparameters for the Sagemaker XGBoost algorithm. Identify the correct options for the model validation techniques (Select two):

- [ ] Validation using training set
- [x] K-fold validation **(Correct)**
- [ ] Validation using SageMaker Ground Truth
- [x] Validation using a holdout set **(Correct)**

Explanation

Validation using training set and validation using SageMaker Ground Truth are made up options. You can use Validation using a holdout set or K-fold validation to tune the hyperparameters for the Sagemaker XGBoost algorithm. Please read more on the model validation https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-model-validation.html

A company wants to enhance the existing security procedure at its data center so that only the employees with the required privilege are able to access certain sensitive areas of the facility. The company wants to do a facial match so that only bona fide employees can access these sensitive areas. Which service would you recommend to build this solution with the least possible effort:

- SageMaker Object Detection
- SageMaker Image Classification
- AWS Rekognition **(Correct)**
- SageMaker Semantic Segmentation

Explanation

AWS Rekognition can be used for face recognition with the least possible effort: https://aws.amazon.com/blogs/machine-learning/build-your-own-face-recognition-service-using-amazon-rekognition/

You are developing a computer vision system for a factory assembly line. The system can trigger a robotic arm to correct the orientation of the product components whenever it detects a misalignment. Which SageMaker algorithm should be used in the computer vision system?

- Image Classification
- Object Detection
- Semantic Segmentation **(Correct)**
- Reinforcement Learning (RL)

Explanation

RL is not relevant for this use-case. Image Classification is used to classify images into multiple classes such as cat vs dog. Object Detection is used to detect objects in an image. Semantic Segmentation is used for pixel level analysis of an image and it can be used in this computer vision system to detect misalignment

You are building a feature for a web application such that when a user attempts to log in from an anomalous IP address, a web login server would trigger a multi-factor authentication system. Which SageMaker algorithm would you use for this feature:

- IP Insights **(Correct)**
- Random Cut Forest
- Factorization Machines
- XGBoost

Explanation

Amazon SageMaker IP Insights is an unsupervised learning algorithm that learns the usage patterns for IPv4 addresses and detects any underlying anomalies:

https://docs.aws.amazon.com/sagemaker/latest/dg/ip-insights.html

As an ML Engineer using Amazon SageMaker, which are the user interface options you can use to train a SageMaker model (Select three) :

- [ ] AWS Batch
- [x] Jupyter Notebook **(Correct)**
- [ ] AWS Glue
- [x] SageMaker Console **(Correct)**
- [x] SageMaker SDK **(Correct)**

Explanation

SageMaker Console, SageMaker SDK and Jupyter Notebook can be used to train SageMaker models. AWS Glue is primarily used for ETL jobs utilising Apache Spark as the distributed processing engine. AWS Batch is used to run Batch computing jobs on AWS infrastructure.

Which is the best evaluation metric for a binary classification model:

- AUC/ROC **(Correct)**
- Precision
- Accuracy
- F1-Score

Explanation

AUC/ROC is the correct choice. AUC/ROC metric does not require you to set a classification threshold.

For imbalanced datasets, you are better off using another metric called - **PR AUC** - that is also used in production systems for a highly imbalanced dataset, where the fraction of positive class is small, such as in case of credit card fraud detection.

Please review this excellent reference material for a deep-dive:

https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

After training a SageMaker model over a huge training dataset, the data science team observed that it has low accuracy on the training data as well as low accuracy on the test data. What can you say about the model:

- Model is overfitting
- Model is neither underfitting nor overfitting
- The model needs more training data
- Model is underfitting **(Correct)**

Explanation

When a model underfits, it exhibits low accuracy on both the training and test data

Identify the mandatory hyperparameters for the SageMaker K-means algorithm (Select two):

- [x] k **(Correct)**
- [ ] epochs
- [ ] mini_batch_size
- [x] feature_dim **(Correct)**

Explanation

feature_dim and k are the required hyperparameters for the SageMaker K-means algorithm

https://docs.aws.amazon.com/sagemaker/latest/dg/k-means-api-config.html

The data science team at a SaaS CRM company wants to improve its customer support workflow. The team wants to identify duplicate support tickets or route tickets to the correct support queue based on similarity of the text found in a ticket. As an ML Specialist, which SageMaker algorithm would you recommend to help solve this problem:

- BlazingText Word2Vec mode
- XGBoost
- Object2Vec **(Correct)**
- Factorization Machines

Explanation

Object2Vec can be used to find semantically similar objects such as tickets. BlazingText Word2Vec can only find semantically similar words. Factorization Machines and XGBoost are not fit for this use-case. A good reference read for Object2Vec:

https://aws.amazon.com/blogs/machine-learning/introduction-to-amazon-sagemaker-object2vec/

Identify the mandatory hyperparameter for both the Word2Vec (unsupervised) and Text Classification (supervised) modes of the SageMaker BlazingText algorithm:

- mode **(Correct)**
- learning_rate
- epochs
- buckets

Explanation

mode is the mandatory hyperparameter for both the Word2Vec (unsupervised) and Text Classification (supervised) modes of the SageMaker BlazingText algorithm :

https://docs.aws.amazon.com/sagemaker/latest/dg/blazingtext_hyperparameters.html

Which of the following statements is true for the Sagemaker Linear Learner algorithm:

- When you use automatic model tuning, the linear learner internal tuning mechanism is turned off automatically. This sets the number of parallel models, num_models, to 0
- When you use automatic model tuning, the linear learner internal tuning mechanism is turned off automatically. This sets the number of parallel models, num_models, to 1 **(Correct)**
- When you use automatic model tuning, the linear learner internal tuning mechanism is turned on automatically. This sets the number of parallel models, num_models, to 1
- When you use automatic model tuning, the linear learner internal tuning mechanism is turned on automatically. This sets the number of parallel models, num_models, to 0

Explanation

When you use automatic model tuning, the linear learner internal tuning mechanism is turned off automatically. This sets the number of parallel models, num_models, to 1.

https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner-tuning.html

The data science team at an analytics company is working on a linear regression model and it observes that the training error as well as the test error are high, implying that the model has a bias. Which of the following L1 and L2 regularization optimizations may be done to resolve this issue (Select two):

- [ ] L1 and L2 regularization are not required, just get more training data
- [ ] Increase L1 regularization
- [x] Use L2 regularization and drop L1 regularization **(Correct)**
- [x] Decrease L1 regularization **(Correct)**

Explanation

Getting more training data alone will not address the model bias. You can think of L1 as reducing the number of features in the model altogether. L2 "regulates" the feature weight instead of just dropping them. "Decreasing L1 regularization" and "Using L2 regularization along with dropping L1 regularization" are the correct options. Please review the concept of L1 and L2 regularization in more detail:

https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c

You are doing topic modeling using the SageMaker Latent Dirichlet Allocation (LDA) algorithm. Which of the following are correct:

- LDA is a "bag-of-words" model, which means that the order of words does matter
- LDA is a not a "bag-of-words" model, which means that the order of words does not matter
- LDA is a "bag-of-words" model, which means that the order of words does not matter **(Correct)**
- LDA is a not a "bag-of-words" model, which means that the order of words does matter

Explanation

LDA is a "bag-of-words" model, which means that the order of words does not matter

https://docs.aws.amazon.com/sagemaker/latest/dg/lda-how-it-works.html

A leading ecommerce company is looking to improve the user experience by recommending the related product categories for its catalog of products. As an ML Specialist, which SageMaker algorithms would you use to develop a solution for this use-case (Select two):

- [x] Latent Dirichlet Allocation (LDA) **(Correct)**
- [ ] XGBoost
- [x] Factorization Machines **(Correct)**
- [ ] K-means

Explanation

Use LDA to figure out the right categories for each product. Use Factorization Machines to recommend the right related categories for the given product's categories.

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)

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

Explanation

Since the gap between human-level performance and training error is large, the model is underfitting. When a model underfits, it is not learning from training data. To fix the high training error, you can increase the model complexity, train the model longer (more epochs), and use a different network architecture. However, increasing regularization would reduce the model complexity (by suppressing the importance of features), and it will underfit more.

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?

- ○ 0.6
- ○ 0.5
- ○ 0.3
- ◉ 0.8 **(Correct)**

Explanation

Recall or true positive rate = TP/(TP+FN)

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?

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

Explanation

Principal Component Analysis (PCA) is a dimensionality reduction technique – it works by capturing information contained in the original dataset using far fewer features known as components. The newly generated features (component) can then be used for model training. The one drawback with PCA is the newly generated components cannot be mapped to real-world features as there is no easy way to figure how each feature contributes to a component. You also need to standardize or normalize data before you perform PCA.

The factorization algorithm works very well with large sparse datasets. Since this is a dense dataset, this option is not valid.

GZIP compression merely reduces the storage needed – it does not reduce the number of features.

Parquet format is an efficient binary storage format for columnar access – it is useful for scenarios where you want to extract only some columns from 1000s of columns

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
- ○ L2 Regularization
- ○ Learning Rate
- ◉ L1 Regularization **(Correct)**

Explanation

Regularization is used to control how a feature can influence the outcome.

When the model overfits, you can increase regularization to reduce the relative weight of each feature.

Similarly, when a model underfits, you can reduce regularization to allow features to assist in predicting the outcome more actively.

L1 Regularization works by eliminating features that are not important.

L2 Regularization keeps all the features but simply assigns a very small weight to features that are not important

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

- (•) Ensure there are no duplicates **(Correct)**
- ( ) Ensure all duplicates are in test data
- ( ) Ensure all duplicates are in train data
- ( ) Ensure data is shuffled before creating train and test set

Explanation

Duplicates can accidentally leak into validation and test sets when you split your data. This can cause artificially better performance on validation and test sets. You should clean up the data so that all examples are distinct.

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
- ( ) L2 Regularization
- ( ) Learning Rate
- (•) L1 Regularization **(Correct)**

Explanation

Regularization is used to control how a feature can influence the outcome.

When the model overfits, you can increase regularization to reduce the relative weight of each feature.

Similarly, when a model underfits, you can reduce regularization to allow features to assist in predicting the outcome more actively.

L1 Regularization works by eliminating features that are not important.

L2 Regularization keeps all the features but simply assigns a very small weight to features that are not important

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?

- ( ) Recall
- (•) F1 Score **(Correct)**
- ( ) Precision
- ( ) Accuracy

Explanation

Accuracy is not a useful metric for skewed data sets. Recall on its own is not enough as the model that predicts everything as positive will have a very high Recall. Precision on its own is not enough as the model can have very high precision even if it predicts only one positive correctly and misclassifies everything as negative. F1 Score is a useful metric as it considers both recall and precision. Another metric that you can use is the ROC AUC score.

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?

- (•) Softmax **(Correct)**
- ( ) None
- ( ) ReLU
- ( ) Sigmoid

Explanation

Softmax activation is used for predicting a single label from a set of possible labels. Softmax returns the probability for each label, and the sum of all probabilities adds up to a 1. The class with the highest probability is used as the final class for the example.

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?

- Use Transfer learning and remove the first hidden layer of image classification model 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
- (selected) Use Transfer learning by removing the output layer of the image classification model, reinitialize the weights of last hidden layer and retrain the model **(Correct)**

Explanation

Transfer learning is an approach of reusing a model that works well for a similar problem. With neural networks and deep learning, some domains like speech recognition, image recognition, and so forth require an extensive dataset for the algorithm to learn all patterns. You can reuse these models for more specialized tasks by using transfer learning. Commonly, with transfer learning, you remove the output layer of one model and feed the hidden layer to a different set of neurons that assess performance with the new dataset. The algorithm is now retrained to learn new patterns and adjust the weight. You can start by randomly initializing the weights of the last layer of the existing, and for more complex use cases, you may need to random initialize of weights of the final few layers.

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?

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

Explanation

XGBoost requires all numeric features. Tree-based algorithms can handle features with different scales. It also handles numeric categorical features (does not require one-hot encoding). However, XGBoost also supports One-hot encoded features. For a binary feature like Lawn (yes or no), only label encoding is needed (i.e., convert to 0 and 1). You should not perform one-hot encoding on binary features. For non-binary categorical features, you can test with label encoding first and then optionally, test performance with one-hot encoding.

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
- 8
- 5 **(Correct)**
- 6

Explanation

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.

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%
- Average of the error rates
- 5%
- 1% **(Correct)**

Explanation

1% should be used as the human-level performance and it is a good proxy for Bayes optimal error (theoretical best possible error rate).

Reference: (Starting at 1:24:00)
https://www.youtube.com/watch?v=wjqaz6m42wU

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?

- Replace missing values with 0
- 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)**

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

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?

- Order data by time and set aside first 80% for training and the remaining 20% for testing **(Correct)**
- Split data into 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
- Shuffle data and perform a random split to keep 80% for training and 20% for testing

Explanation

For time-series forecasting, our objective is to predict the values in the future. To get a realistic assessment of model performance, you need to split the dataset based on time. Set aside the first 70-80% for training and keep the most recent data (toward the end) for testing the accuracy of predictions. Random shuffling is not recommended for time series forecasting

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

- Hosting
- Training **(Correct)**
- Testing
- Validation

Explanation

Hyperparameters are used when training the model – These parameters control how a model learns. Once a model is trained, you cannot change the hyperparameters – you need to retrain it

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

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

Explanation

Mini-batch has the effect of making more substantial changes to weight as it uses a smaller set of samples to determine the gradient. In a deep learning network, the loss curve is very complex with multiple local minima. To prevent the optimizer from getting stuck in local minima, you can reduce the batch size to jump over local minima

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

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

Explanation

Weights are adjusted based on error observed in a mini-batch. The training set is divided into mini-batches, and when you increase the number of examples in a mini-batch, you have fewer mini-batches. And this would result in less-frequent weight adjustment.

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 **(Correct)**
- Model 2
- Model 3
- Model 4

Explanation

Recall needs to be at least 60%. Recall = TP/(TP+FN).

Model 1 recall is over 0.6, and Model 2 recall is 0.3; model 3 recall is 1/15 and is a small value and Model 4 recall is 0.6. So, the answer has to be either Model 1 or Model 4.

The cost of misclassifying a positive sample is three times more than misclassifying a negative sample.

Total Cost = 3 * FN + 1 * FP

Model 1 cost = 3*5 + 10 = 25. Model 4 cost is = 3 * 6 + 1 * 15 = 18 + 15 = 33.

So, Model 1 has the lowest cost while meeting 0.6 recall.

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 sets. Optimize the model to improve validation performance. Use 100 hours of noisy data for final testing
- [x] 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)**
- [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)**
- [ ] Use 100 hours of noisy data for training and split the general speech data for validation and testing

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.

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)

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

Explanation

For the unbalanced dataset, accuracy is not a good measure as a model that predicts all instances as normal will be 99% accurate. ROC AUC metric considers True Positive Rate and False positive Rates at all possible cutoff thresholds. It is a useful metric for binary classifiers. However, they are not suitable for highly imbalanced datasets as ROC curve considers only true positive and false positive rates. ROC does not account for negatives and does not measure the performance well. Instead precision-recall curve is used for imbalanced datasets. Oversampling by duplicating data is not going to improve quality as it does not add any new patterns that algorithms can learn. Synthetic Minority Over-sampling Technique (SMOTE) provides a mechanism for artificial data generation that has shown to improve the accuracy of unbalanced classifiers. To generate data similar to an existing instance, SMOTE uses the nearest neighbors, and generate synthetic data along the lines that connect the neighbors. This method ensures data is close to other similar instances. Reference
(ROC Imbalance): https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/ https://towardsdatascience.com/sampling-techniques-for-extremely-imbalanced-data-281cc01da0a8

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

- [x] xgboost **(Correct)**
- [ ] linear regression
- [x] decision tree **(Correct)**
- [ ] neural network

Explanation

Tree-based algorithms like decision tree, random forest, and xgboost have a lower and upper bound it can predict for regression. The lower and upper bound is determined based on the range of values seen during training.

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
- ○ Behavior depends on the number of layers
- ○ The network would automatically learn insights about missing values
- ◉ The network would not learn from missing values **(Correct)**

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

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:

- ○ Underfitting
- ○ Performing close to human-level performance
- ◉ Overfitting **(Correct)**
- ○ Normal

Explanation

Here, the model training error is comparable to human-level performance. So, the model is doing well with training data. However, it is not generalizing well for unseen data, and the test error is much larger. So, it is showing signs of overfitting. (memorized too much about training data)

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

Explanation

This problem is an example of where you don't want to use Machine Learning. Temperature conversion is easily doable with a simple formula, whereas implementing as an ML solution is more complicated and has a lot of overhead in terms of model training, validation, hosting, and ongoing maintenance.

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?

- [x] Positive: 10, Negative: 9 **(Correct)**
- [ ] Positive: 6, Negative: 13
- [ ] Positive: 12, Negative: 7
- [ ] Positive: 13, Negative: 6

Explanation

Positive = True Positive + False Negative

Negative = True Negative + False Positive

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?

- [ ] 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
- [x] 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

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.

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
- [ ] 0.8
- [ ] 0.3
- [x] 0.6 **(Correct)**

Explanation

Precision = TP / (TP + FP)

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?

- [ ] Normalization
- [ ] Remove one feature from every highly correlated feature pairs
- [x] Data transformation is not needed for this dataset **(Correct)**
- [ ] Scaling

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.

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)

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

Explanation

The categorical features need to be one-hot encoded for algorithms like the linear model and neural network.

XGBoost and tree-based algorithms can work with numeric categorical features as well as one-hot encoded categorical features (one-hot encoding is not required; however, XGBoost can handle one-hot encoded data).

For binary feature like holiday, you need to convert to numeric value using label-encoding.

For numeric features, convert them to a similar range and scale.

AirQuality is the label, and the model needs to learn to predict one of two outcomes.

Since the label contains text, it needs to be converted to a numeric value and we can do that using label encoding.

Note: For multi-class problem, neural networks require one-hot encoding of labels

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

Explanation

The training error is low, and the test error is high. So, the model has a variance problem. The model is overfitting the training data, or the data distribution between test and train data sets is different. You can train with more data to handle issues with different data distribution between train and test data sets. More data is also useful if the model is not detecting all the patterns. If you suspect overfitting, you can also increase regularization to simplify the model. Finally, you can also try different neural network architecture (for example reduce number of neurons, change the number of hidden layers and so forth). However, decreasing regularization would cause the model to overfit more. You can reduce the number of epochs to minimize variance. This would prevent the model from memorizing too much about training data.

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

- ○ Train
- ○ Test
- ◉ Validation **(Correct)**
- ○ Use a random sample from train, validation and test sets

Explanation

Tune model using validation data. To prevent the model from overfitting the validation data, you need to plan to do a final check with an unseen test data

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?

- ○ F1 Score
- ◉ ROC AUC Metric **(Correct)**
- ○ Accuracy
- ○ Squared Error

Explanation

Receiver Operating Characteristic (ROC) curve compares the true positive rate and the false positive rate at different thresholds. AUC metrics measure the area formed by such a curve, and the ROC AUC is used for summarizing the model performance with a range of tradeoffs

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

- ○ Recurrent Neural Network
- ○ Try different neural network architectures
- ○ General Purpose Neural Network
- ◉ Convolutional Neural Network **(Correct)**

Explanation

CNN's are ideal for image and video analysis applications – it considers a pixel and surround pixels to identify patterns. RNNs are used for applications where the predicted value depends on previously seen values – time-series forecasting, speech recognition and so forth. The general-purpose neural network considers each pixel as a separate feature and may require a very complex network for image analysis applications – whereas a simple CNN can easily outperform a general-purpose neural network for visual analysis application.

After training a deep neural network over 100 epochs, it achieved high accuracy on your training data, but lower accuracy on your test data, suggesting the resulting model is overfitting. What are TWO techniques that may help resolve this problem?

- ☐ Use more layers in the network
- ☐ Employ gradient checking
- ☑ Use early stopping **(Correct)**
- ☑ Use dropout regularization **(Correct)**
- ☐ Use more features in the training data

Explanation

Early stopping is a simple technique for preventing neural networks from training too far, and learning patterns in the training data that can't be generalized. Dropout regularization forces the learning to be spread out amongst the artificial neurons, further preventing overfitting. Removing layers, rather than adding them, might also help prevent an overly complex model from being created - as would using fewer features, not more.

Your automatic hyperparameter tuning job in SageMaker is consuming more resources than you would like, and coming at a high cost. What are TWO techniques that might reduce this cost?

- [ ] Use inference pipelines
- [x] Use logarithmic scales on your parameter ranges **(Correct)**
- [ ] Use linear scales on your parameter ranges
- [ ] Use more concurrency while tuning
- [x] Use less concurrency while tuning **(Correct)**

Explanation

Since the tuning process learns from each incremental step, too much concurrency can actually hinder that learning. Logarithmic ranges tend to find optimal values more quickly than linear ranges. Inference pipelines are a thing, but have nothing to do with this problem.

You are developing a computer vision system that can classify every pixel in an image based on its image type, such as people, buildings, roadways, signs, and vehicles. Which SageMaker algorithm would provide you with the best starting point for this problem?

- ○ Rekognition
- ◉ Semantic Segmentation **(Correct)**
- ○ Object2Vec
- ○ Object Detection

Explanation

Semantic Segmentation produces segmentation masks that identify classifications for each individual pixel in an image. It uses MXNet and the ResNet architecture to do this.

A system designed to classify financial transactions into fraudulent and non-fraudulent transactions results in the confusion matrix below. What is the recall of this model?

	Actual Positives	Actual Negatives
Predicted Positives	90	45
Predicted Negatives	10	20

- ◉ 90% **(Correct)**
- ○ 50%
- ○ 66.67%
- ○ 74%

Explanation

Recall is defined as true positives / (true positives + false negatives). This works out to 90/(90+10) in this example, or 90%. 66.67% is the precision (true positives / (true postives + false positives.) Recall is an important metric in situations where classifications are highly imbalanced, and the positive case is rare. Accuracy tends to be misleading in these cases.

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?

- ○ Comprehend
- ○ BlazingText
- ◉ Factorization Machines **(Correct)**
- ○ DeepAR

Explanation

Factorization Machines algorithm is used for building recommender systems and for collaborative filtering.

Collaborative filtering algorithms learn the likelihood of a customer purchasing a product based on other customer purchase behavior.

BlazingText is used for text analysis and classification problems.

Comprehend is used for natural language processing and not suitable for this use case.

DeepAR is used for time series forecasting.

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?

- Textract
- Comprehend **(Correct)**
- Rekognition Text Extraction
- Transcribe

Explanation

Comprehend. With Comprehend, you can analyze a document to extract entities and key phrases along with confidence scores. You can use it to provide a summary of the talking points of a document. Seq2Seq algorithm supported by SageMaker is another algorithm that you could for this purpose.

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
- Custom Classification with Comprehend
- Sentiment Analysis with Comprehend **(Correct)**

Explanation

Sentiment Analysis with Comprehend. Sentiment analysis can evaluate text based on the content and provide a confidence score for Positive, Negative, Neutral, Mixed. You could use this to assess reviews based on the sentiment and shortlist strong positive and strong negative reviews.

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?

- Neural Topic Modeling (NTM)
- LDA
- Seq2Seq
- Comprehend **(Correct)**

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

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?

- LDA
- Seq2Seq **(Correct)**
- Random Cut Forest
- K-Means

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

ML Implementation and Operation

You want to secure the API calls made to your published Amazon SageMaker model endpoints from your customer VPC. By default, these API calls traverse the public network to the request router. What measures would you take to address this issue:

- ○ Use AWS-SSE for private connectivity between the customer's VPC and the request router to access hosted model endpoints.
- ○ Use SSH for private connectivity between the customer's VPC and the request router to access hosted model endpoints.
- ○ Amazon SageMaker ensures that machine learning (ML) model artifacts and other system artifacts are encrypted in transit, so no special measures are required
- ◉ Use Amazon Virtual Private Cloud interface endpoints powered by AWS PrivateLink for private connectivity between the customer's VPC and the request router to access hosted model endpoints. **(Correct)**

Explanation

SSH and AWS-SSE are not used for this requirement. Amazon SageMaker supports Amazon Virtual Private Cloud interface endpoints powered by AWS PrivateLink for private connectivity between the customer's VPC and the request router to access hosted model endpoints.

https://docs.aws.amazon.com/sagemaker/latest/dg/inter-network-privacy.html

A research agency is developing a robotic submarine to map the marine life forms in the pacific ocean. The robot should be able to classify the images of marine life forms in an autonomous way with low latency. Which Amazon SageMaker architecture would you recommend for this use-case:

- ○ Use Kinesis Data Streams to process the video stream and invoke a lambda to infer via a classification model to classify the images of the marine life forms
- ○ Use AWS Rekognition to classify the images of the marine life forms
- ○ Use Kinesis Video Streams to classify the images of the marine life forms
- ◉ Use SageMaker Neo to compile and package the classification model on the underlying runtime infrastructure on the robotic device. **(Correct)**

Explanation

Kinesis Video Streams, Rekognition and Kinesis Data Streams are over-the-air options, hence ruled out for this use-case. SageMaker Neo provides the correct local solution with low latency:

https://docs.aws.amazon.com/sagemaker/latest/dg/neo.html

How many containers can a SageMaker Inference Pipeline support (Select two):

- ☐ 7
- ☐ 1
- ☑ 2 **(Correct)**
- ☑ 3 **(Correct)**

Explanation

Inference Pipeline is composed of a linear sequence of two to five containers:

https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipelines.html

You are running a SageMaker training job. You notice that the job has failed but there are no logs on CloudWatch. Identify the possible scenarios behind this issue (Select two):

- ☑ The training job has the wrong training image **(Correct)**
- ☑ S3 location for training data is incorrect **(Correct)**
- ☐ The default IAM role needs to have the correct permissions in order to write logs into Cloudwatch
- ☐ Logs appear on Cloudwatch 15 minutes after the attempted job run

Explanation

There is no 15 minutes gap for logs to appear on Cloudwatch. The default IAM has the required permissions to write logs into Cloudwatch. Possible reasons could be incorrect training image or incorrect S3 location for training data
- https://docs.aws.amazon.com/sagemaker/latest/dg/common-info-all-sagemaker-models-logs.html

Identify the three step process while deploying a model using SageMaker hosted services (Select three)

- [ ] Validate the model
- [x] Create the model **(Correct)**
- [x] Create the HTTPS endpoint **(Correct)**
- [ ] Tune model hyperparameters
- [x] Create the endpoint configuration for HTTPS endpoint **(Correct)**

Explanation

Deploying a model using SageMaker hosting services is a three-step process with the following steps: create a model in SageMaker, create an endpoint configuration for an HTTPS endpoint and create an HTTPS endpoint.

You can understand the process for deploying a model using SageMaker hosted services in more detail here:

https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-deployment.html

Which of the following is correct regarding the model types that Amazon SageMaker Neo supports:

- ○ Neo supports recommendation system models
- ○ Neo supports dimensionality reduction models
- ○ Neo supports regression models
- ◉ Neo supports image classification models. **(Correct)**

Explanation

Neo currently supports image classification models exported as frozen graphs from TensorFlow, MXNet, or PyTorch, and XGBoost models.:

https://docs.aws.amazon.com/sagemaker/latest/dg/neo.html

Identify the correct statement regarding SageMaker Inference Pipelines:

- ◉ Inference Pipelines can be used to make either real-time predictions or to process batch transforms **(Correct)**
- ○ Inference Pipelines can only be used to make real-time predictions but not to process batch transforms
- ○ Inference Pipelines can only be used to process batch transforms but not to make real-time predictions
- ○ Inference Pipelines can neither be used to process batch transforms nor to make real-time predictions

Explanation

Inference Pipeline can be considered as an Amazon SageMaker model that you can use to make either real-time predictions or to process batch transforms directly without any external preprocessing.

https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipelines.html

You are creating a classification model using one of the Amazon SageMaker built-in algorithms and you want to use GPUs for both training and inference. Identify the correct steps (Select two):

- [x] Select the correct instance type that supports GPUs **(Correct)**
- [x] Select a built-in algorithm that supports GPUs for both training and inference **(Correct)**
- [ ] Specify gpu=True as the parameter in the create_training_job boto3 API call on Jupyter Notebook
- [ ] Enable GPU support in the docker image for the SageMaker algorithm

Explanation

create_training_job has no such parameter as gpu, so that option is invalid. You can't enable GPU support in docker image as that's a made up option. You need to select the appropriate built-in Algorithm as well as choose the correct instance type for GPU support.

Identify the correct statements regarding the IAM features available to use with Amazon SageMaker (Select two):

- [x] Amazon SageMaker supports authorization based on resource tags **(Correct)**
- [ ] Amazon SageMaker supports service linked roles
- [ ] Amazon SageMaker supports resource-based policies
- [x] With IAM identity-based policies, you can specify allowed or denied actions and resources as well as the conditions under which actions are allowed or denied for Amazon SageMaker **(Correct)**

Explanation

You can review the IAM features available to use with Amazon SageMaker in detail here:

https://docs.aws.amazon.com/sagemaker/latest/dg/security_iam_service-with-iam.html

The data science team at an Analytics company is working on a dataset with a large number of observations and features. They want to use the SageMaker Principal Component Analysis (PCA) Algorithm to reduce the dimensionality within the dataset. Which mode should be used for using PCA on this dataset:

- (•) Use PCA in randomized mode **(Correct)**
- ( ) Use PCA in either regular or randomized mode
- ( ) PCA is not the right selection for this use-case
- ( ) Use PCA in regular mode

Explanation

Use PCA in randomized mode: For datasets with a large number of observations and features, this mode uses an approximation algorithm.

https://docs.aws.amazon.com/sagemaker/latest/dg/pca.html

Identify the correct statements regarding the Amazon SageMaker logging and monitoring options on CloudWatch and CloudTrail (Select four):

- [x] CloudWatch keeps the SageMaker monitoring statistics for 15 months. However, the Amazon CloudWatch console limits the search to metrics that were updated in the last 2 weeks. **(Correct)**
- [x] CloudTrail does not monitor calls to InvokeEndpoint **(Correct)**
- [x] SageMaker monitoring metrics are available on CloudWatch at a 1-minute frequency **(Correct)**
- [ ] CloudTrail monitors calls to InvokeEndpoint
- [x] AWS CloudTrail provides a record of actions taken by a user, role, or an AWS service in Amazon SageMaker. CloudTrail keeps this record for a period of 90 days **(Correct)**
- [ ] SageMaker monitoring metrics are available on CloudWatch at a 2-minute frequency

Explanation

Please review the best practices for Amazon SageMaker logging and monitoring options on CloudWatch and CloudTrail. Some good reference links:

https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-incident-response.html
https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-cloudwatch.html
https://docs.aws.amazon.com/sagemaker/latest/dg/logging-using-cloudtrail.html
https://docs.aws.amazon.com/awscloudtrail/latest/userguide/view-cloudtrail-events.html

For the upcoming festive season, an ecommerce company is anticipating a major shift in the expected workload for its product recommendation engine hosted on SageMaker. Which solution would you recommend to address this issue:

- ( ) SageMaker hosting has a built-in mechanism to address this issue. Nothing else needs to be done.
- ( ) Use inference pipelines to manage the workload
- ( ) Use elastic inference to manage the workload
- (•) Use automatic scaling for the production variants to manage the workload **(Correct)**

Explanation

Elastic inference is used to improve the inference throughput. Inference Pipelines are used to define and deploy a combination of algorithms in SageMaker. Automatic scaling is the correct option to address any changes in workload. You need to understand concepts such as Target Metric for Automatic Scaling, Minimum and Maximum Capacity for Automatic Scaling, Cooldown Period for Automatic Scaling.

https://docs.aws.amazon.com/sagemaker/latest/dg/endpoint-auto-scaling.html

As a best practice, you should deploy multiple instances across availability zones

https://docs.aws.amazon.com/sagemaker/latest/dg/deployment-best-pratices.html

Identify the SageMaker algorithm that can be used both as a built-in-algorithm as well as a framework such as Tensorflow:

- Linear Learner
- Object2Vec
- Factorization Machines
- XGBoost **(Correct)**

Explanation

The XGBoost algorithm can be used as a built-in algorithm or as a framework such as TensorFlow.

https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html

For production systems based on SageMaker, which version tag should be used in the Registry Paths:

- :latest
- :-1
- :0
- :1 **(Correct)**

Explanation

":1" is the correct version tag for production systems. ":0" and ":-1" are not valid values for version tags. You can find more information regarding the version tag options for registry paths for SageMaker here:

https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-algo-docker-registry-paths.html

To get inference for an entire dataset, you are developing a batch transform job using Amazon SageMaker High-level Python Library. Which method would you call so that the inferences are available for the entire dataset:

- predict
- transform **(Correct)**
- deploy
- fit

Explanation

predict is used for real time inference. Deploy and fit are used for models hosted on SageMaker Hosting Services. You need to call the transform method on the sagemaker.transformer.Transformer object. The constructor argument - output_path - takes the location in S3 where you want to store the inferences

https://docs.aws.amazon.com/sagemaker/latest/dg/ex1-batch-transform.html

A data science intern at an Analytics Company is working on creating a binary classification model. He has created a SageMaker notebook instance using its default IAM role and is trying to read the training data from a S3 bucket with the name "model-training-data". However the data is not accessible in the Jupyter Notebook. How can he resolve this issue (Select two):

- [x] Use a bucket with the word "sagemaker" in its name **(Correct)**
- [ ] Restart the Jupyter notebook instance and that should resolve the data access issue
- [ ] Make the S3 bucket public
- [x] Add a policy to the role that grants the SageMaker service principal S3FullAccess permission **(Correct)**

Explanation

The default Sagemaker IAM role gets permissions to access any bucket that has sagemaker in the name. If you add a policy to the role that grants the SageMaker service principal S3FullAccess permission, the name of the bucket does not need to contain sagemaker. Granting public access to S3 bucket is not recommended. You can read further on this -

https://docs.aws.amazon.com/sagemaker/latest/dg/gs-config-permissions.html

The AI research department at a University is collaborating with a consultancy firm. A research assistant at the department would like to allow developers from the consultancy firm to access some of the SageMaker resources created in the AWS account of the research department. What are the recommended ways this access can be granted (Select two):

- [x] Create a role to delegate access to your resources with the third-party AWS account **(Correct)**
- [ ] Create SageMaker resource based policies to allow this access
- [x] Provide access to externally authenticated users through identity federation **(Correct)**
- [ ] Create a new AWS user account and share the username and password via email

Explanation

Sharing user account credentials via email is wrong. SageMaker does not support resource based policies. You can create a role to delegate access or provide access via identity federation. Please read more :

https://docs.aws.amazon.com/sagemaker/latest/dg/security_iam_troubleshoot.html#security_iam_troubleshoot-cross-account-access

https://docs.aws.amazon.com/IAM/latest/UserGuide/id_roles_common-scenarios_third-party.html

https://docs.aws.amazon.com/IAM/latest/UserGuide/id_roles_common-scenarios_federated-users.html

Identify the criteria on which early stopping works in Amazon SageMaker:

- (•) If the value of the objective metric for the current training job is worse (higher when minimizing or lower when maximizing the objective metric) than the median value of running averages of the objective metric for previous training jobs up to the same epoch, Amazon SageMaker stops the current training job. **(Correct)**
- ( ) If the value of the objective metric for the current training job is worse (higher when minimizing or lower when maximizing the objective metric) than the mean value of running averages of the objective metric for previous training jobs up to the same epoch, Amazon SageMaker stops the current training job.
- ( ) If the value of the objective metric for the current training job is better (higher when minimizing or lower when maximizing the objective metric) than the median value of running averages of the objective metric for previous training jobs up to the same epoch, Amazon SageMaker stops the current training job.
- ( ) If the value of the objective metric for the current training job is better (higher when minimizing or lower when maximizing the objective metric) than the mean value of running averages of the objective metric for previous training jobs up to the same epoch, Amazon SageMaker stops the current training job.

Explanation

If the value of the objective metric for the current training job is worse (higher when minimizing or lower when maximizing the objective metric) than the median value of running averages of the objective metric for previous training jobs up to the same epoch, Amazon SageMaker stops the current training job.

https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-early-stopping.html

You have built a deep learning model and now want to deploy it using the SageMaker Hosting Services. For inference, you want a cost-effective option that guarantees low latency but still comes at a fraction of the cost of using a GPU instance for your endpoint. As an ML Specialist, what would you recommend to use:

- Automatic Scaling
- Elastic Inference **(Correct)**
- SageMaker Neo
- Inference Pipeline

Explanation

By using Amazon Elastic Inference (EI), you can speed up the throughput and decrease the latency of getting real-time inferences from your deep learning models that are deployed as Amazon SageMaker hosted models, but at a fraction of the cost of using a GPU instance for your endpoint.

https://docs.aws.amazon.com/sagemaker/latest/dg/ei.html

An Analytics Consulting Firm wants you to review a Classification Model trained on historical data and deployed about 6 months ago. At the time of deployment the model performance was upto the mark. Post deployment, the model has not been retrained on the incremental data coming in every day. Now the model performance has gone down significantly. As an ML Specialist, what is your recommended course of action:

- Completely retrain the model using only the data for the last 6 months
- Completely retrain the model using the historical data along with the data for the last 6 months. **(Correct)**
- Completely retrain the model again using only the historical data
- Change the algorithm behind the model for better performance.

Explanation

This is an example of model deterioration because the training data has aged. The solution is to retrain the model using the historical data along with the data for the last 6 months

An electronics goods company wants to do sentiment analysis of the customer feedback for its latest product that was recently launched in France. However, the customer feedback is only available in the form of recorded audio snippets in the French language. As an ML specialist, which AWS AI services can you use to build a quick solution with the least effort:

- Transcribe -> Lex -> Comprehend
- Translate -> Transcribe -> Comprehend
- Transcribe -> Lex -> Polly
- Transcribe -> Translate -> Comprehend **(Correct)**

Explanation

The recorded audio would have to be first transcribed into French Language, then translated into English and finally it can be fed into the Comprehend service for sentiment analysis. So the correct order is : Transcribe -> Translate -> Comprehend

Identify the three built-in SageMaker algorithms that support incremental training (Select three):

- [x] Object Detection **(Correct)**
- [x] Semantic Segmentation **(Correct)**
- [ ] Linear Learner
- [ ] XGBoost
- [x] Image Classification **(Correct)**

Explanation

Only three built-in algorithms currently support incremental training: Object Detection Algorithm, Image Classification Algorithm, and Semantic Segmentation Algorithm:

https://docs.aws.amazon.com/sagemaker/latest/dg/incremental-training.html

As a security policy, the data science team at an ecommerce company does not want Amazon SageMaker to provide external network access to the training or inference containers, so network isolation is enabled for all containers. Identify the Amazon SageMaker containers that do not support network isolation, so the data science team does not use them for modeling (Select three):

- [x] Pytorch **(Correct)**
- [x] Amazon SageMaker Reinforcement Learning **(Correct)**
- [x] Scikit-learn **(Correct)**
- [ ] TensorFlow
- [ ] MXNet

Explanation

Network isolation is not supported by the following managed Amazon SageMaker containers as they require access to Amazon S3:

Chainer

PyTorch

Scikit-learn

Amazon SageMaker Reinforcement Learning

https://docs.aws.amazon.com/sagemaker/latest/dg/mkt-algo-model-internet-free.html

Which of the following is correct regarding an inference pipeline on Amazon SageMaker:

- ○ Within an inference pipeline model, Amazon SageMaker handles invocations as a sequence of HTTPS requests.
- ○ Within an inference pipeline model, Amazon SageMaker handles invocations as a sequence of RPC requests.
- ◉ Within an inference pipeline model, Amazon SageMaker handles invocations as a sequence of HTTP requests. **(Correct)**
- ○ Within an inference pipeline model, Amazon SageMaker handles invocations as a sequence of MQTT requests.

Explanation

Within an inference pipeline model, Amazon SageMaker handles invocations as a sequence of HTTP requests.

https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipelines.html

You have launched a new Jupyter Notebook instance and you want to make sure that you don't lose any files and data when the notebook instance restarts. Where should you save your files and data so that they are not overwritten on restart:

- ○ /home/ec2-user/data
- ○ /home/ec2-user/model
- ○ /home/ec2-user/code
- ◉ /home/ec2-user/SageMaker **(Correct)**

Explanation

Only files and data saved within the /home/ec2-user/SageMaker folder persist between notebook instance sessions. Files and data that are saved outside this directory are overwritten when the notebook instance stops and restarts.

https://docs.aws.amazon.com/sagemaker/latest/dg/howitworks-create-ws.html

The Training Image and Inference Image Registry Paths used for SageMaker algorithms are of which type:

- ◉ Region based **(Correct)**
- ○ Country based
- ○ Global
- ○ City based

Explanation

The Training Image and Inference Image Registry Paths are region based. You can further review the registry path options available for SageMaker:

https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-algo-docker-registry-paths.html

A Financial Services company has asked you to finetune its SageMaker model training process. You observe that the company runs the training jobs multiple times in a day with a little tweaking of the training data for each run. Which steps would you recommend to improve the training performance so that the training jobs can complete faster (Select two) :

- [ ] Upgrade the training instance to the highest possible type
- [ ] Spin up an EMR cluster to process the training job
- [x] Change the data format to protobuf recordIO format **(Correct)**
- [x] Use pipe mode to stream data from S3 **(Correct)**

Explanation

Using the training data in protobuf recordIO format along with pipe mode can significantly improve the training job performance. Neither using the EMR cluster nor changing the instance type guarantees improvement in the training performance.

You have disabled direct internet access to your Amazon SageMaker notebook instance while connecting to your VPC in order to prevent unauthorized access to your data. As there is no data access, the notebook instance is not able to train or host models. Which of the following solutions would address this issue when combined together? (Select two):

- [ ] Setup NAT gateway for your VPC along with Security Groups for your VPC that allow inbound connections
- [ ] Setup S3 gateway for your VPC
- [x] Setup NAT gateway for your VPC along with Security Groups for your VPC that allow outbound connections **(Correct)**
- [x] Create a VPC interface endpoint to use PrivateLink for your notebook instance **(Correct)**

Explanation

There is no such thing as S3 gateway for your VPC. If you disable direct internet access, the notebook instance won't be able to train or host models unless your VPC has an interface endpoint (PrivateLink) or a NAT gateway and your security groups allow outbound connections. Please read more:

https://docs.aws.amazon.com/sagemaker/latest/dg/appendix-notebook-and-internet-access.html

The compliance department at a major Financial Services Firm wants to monitor the SageMaker services used by the Data Science team for their ML jobs. Which services can be used to achieve this objective (Select two) :

- [ ] AWS Config
- [ ] Amazon Inspector
- [x] AWS Cloudtrail **(Correct)**
- [x] Amazon Cloudwatch **(Correct)**

Explanation

Cloudwatch and Cloudtrail can be used to monitor the SageMaker services. Further details on monitoring options for SageMaker can be found here:

https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-overview.html

You are training a batch transformation job in Amazon SageMaker. You have protected data at rest by using AWS KMS key on S3. Amazon SageMaker ensures that machine learning (ML) model artifacts and other system artifacts are encrypted in transit and at rest. What measures you would take to make sure that the data is protected in-transit even for inter-node training communications:

- ( ) Use AWS-SSE for inter-node traffic encryption
- (•) There are no inter-node communications for batch processing, so inter-node traffic encryption is not required **(Correct)**
- ( ) Enable inter-container traffic encryption from the console
- ( ) Use SSH for inter-node traffic encryption

Explanation

There are no inter-node communications for batch processing, so inter-node traffic encryption is not required. SSH and AWS-SSE are not used for inter-node traffic encryption.

https://docs.aws.amazon.com/sagemaker/latest/dg/encryption-in-transit.html

A Sports Analytics Company wants to analyse the game-plays for the coming NBA season. They would like to track the movement of each athlete for post-game analysis. Which AWS service can they use to build a solution in the least possible time:

- Kinesis Video Streams
- AWS Rekognition **(Correct)**
- Kinesis Data Stream with Lambda based video frame processing
- SageMaker Image Classification

Explanation

AWS Rekognition has a feature called Pathing that can be used for this use-case:

https://aws.amazon.com/rekognition/

Amazon Sagemaker models are stored in which format:

- model.gzip
- model.zip
- model.tar.gz **(Correct)**
- model.tar.gzip

Explanation

Amazon SageMaker models are stored as model.tar.gz in the S3 bucket specified in OutputDataConfig S3OutputPath parameter of the create_training_job call.

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?

- SageMaker GroundTruth **(Correct)**
- ImageClassification
- Rekognition
- SageMaker Neo

Explanation

SageMaker GroundTruth service provides two capabilities to manage labeling process – Automatic Labeling can learn from examples that you provide and label all instances.

Manual labeling uses Mechanical Turk service to distribute the task across human labelers, and GroundTruth provides the capability to manage the entire workflow.

Neo is used for deploying your Machine Learning algorithm anywhere in the Cloud and at Edge Locations – it is a cross-compilation capability to compile your Machine Learning Algorithm to run on specified hardware.

Rekognition is not used for manual labeling even though you can use this service to train to classify images by providing labeled data.

ImageClassification also expects labeled data as input. This question is about how to create labeled data

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 Firehose, Lambda, and Amazon Rekognition
- Kinesis Data Streams, Amazon Rekognition, Kinesis Video Stream
- Kinesis Video Streams, Amazon Rekognition, Amazon Data Stream **(Correct)**
- Kinesis Firehose, Kinesis Analytics, and Amazon Rekognition

Explanation

Use Kinesis Video Streams to capture the video feed from cameras. Rekognition service can directly consume Kinesis Video Streams, and you can configure Rekognition service to detect celebrities. The output of streaming analysis is stored in a Kinesis Data Stream.

Reference: https://aws.amazon.com/blogs/machine-learning/easily-perform-facial-analysis-on-live-feeds-by-creating-a-serverless-video-analytics-environment-with-amazon-rekognition-video-and-amazon-kinesis-video-streams/

For offline video analysis (video stored in s3), you need to start a job and once the job completes, it will notify using SNS (notification service). You can then pick up the results. Or, you can periodically poll by calling GetCelebrityRecognition.

https://docs.aws.amazon.com/rekognition/latest/dg/video.html

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
- Transcribe, Translate, Polly **(Correct)**
- Translate, Polly

Explanation

Transcribe, Translate, Polly – Translation step requires text data. So, the first step is to transcribe the text from audio and then translate the text. Finally, to convert to speech, use Polly

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?

- 100
- 1 **(Correct)**
- 4
- 32

Explanation

"Running more hyperparameter tuning jobs concurrently gets more work done quickly, but a tuning job improves only through successive rounds of experiments. Typically, running one training job at a time achieves the best results with the least amount of compute time."

https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-considerations.html

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
- General Purpose family
- Accelerated Computing family **(Correct)**
- Memory-Optimized family

Explanation

Accelerated computing family (P and G type instances) come with GPUs, and these are ideal for algorithms that are optimized for GPUs.

General Purpose family are some of the lowest cost instances and offer balanced performance and memory configuration (T and M type instances).

Compute Optimized family comes with the latest generation CPUs and is a higher performance system. These are suitable for CPU intensive model training and hosting (C type instances).

Memory-optimized family are optimized for workloads that process large datasets in memory (R type instances).

Besides, the sagemaker also has Elastic Inference Acceleration (partial GPUs) that provides fractional GPU capacity at a fraction of the cost of accelerated computing family.

Elastic inference Acceleration is suitable for inference workloads that can benefit from GPUs and can be easily added to other instance families.

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)

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

Explanation

"you can simultaneously use up to 20 variables in a hyperparameter tuning job, limiting your search to a much smaller number is likely to give better results"

"a tuning job improves only through successive rounds of experiments. Typically, running one training job at a time achieves the best results with the least amount of compute time"

"Choose logarithmic scaling when you are searching a range that spans several orders of magnitude"

"you specify a range of values between .0001 and 1.0 for the learning_rate hyperparameter, searching uniformly on a logarithmic scale gives you a better sample of the entire range than searching on a linear scale would, because searching on a linear scale would, on average, devote 90 percent of your training budget to only the values between .1 and 1.0, leaving only 10 percent of your training budget for the values between .0001 and .1"

https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-define-ranges.html

https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-considerations.html

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

Explanation

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

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 correct Region and Language
- ( ) Use batch streaming for highest quality outputs
- (●) Use Speech Synthesis Markup Language **(Correct)**
- ( ) Use real-time streaming for highest quality output

Explanation

With Polly, you can use Speech Synthesis Markup Language (SSML) to "control aspects of speech, such as pronunciation, volume, pitch, speed rate, etc." Reference: https://aws.amazon.com/polly/

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?

- ( ) SageMaker Batch Transform **(Correct)**
- ( ) Autoscaling
- ( ) S3 Analytics
- (●) SageMaker Endpoint

Explanation

Batch Transform is a cost-effective way to get large scale inference using SageMaker. Batch transform is ideal for situations where you don't need a persistent real-time endpoint, scenarios where you don't need sub-second latency performance. SageMaker manages all resources required for batch transform. SageMaker Endpoint is used for real-time inference. Autoscaling allows you to maintain capacity, handle instance failures and scale based on workload. S3 Analytics is used for analyzing storage access patterns, which in turn can help you to transition data to the right storage class in S3.

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?

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

Explanation

Translate to German with Source Language set to auto-detect. Translate service can auto-detect source language and convert it to a target language. However, both source and target languages must be on the supported list. https://docs.aws.amazon.com/translate/latest/dg/how-it-works.html

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?

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

Explanation

SageMaker provides pre-built TensorFlow docker images that you can use to train, and host models on SageMaker managed infrastructure efficiently.

Deep Learning AMIs are another option, and you can use it to launch desired EC2 instances pre-configured with necessary tools. However, this requires you to manage and patch EC2 instances.

Launching desired EC2 instances and installing machine learning frameworks is another option – however, you need to manage and patch EC2 instances, and besides, you need to validate and patch ML framework.

Custom Docker images are required when we need to deploy custom models or use a machine learning framework not supported by SageMaker

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?

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

Explanation

In Pipe Mode, training job streams data from S3 to your training instance.

Streaming can provide faster start times and better throughput. It also reduces the storage needed on your training instances as you need storage only for the final model artifacts.

In File mode, training job copies entire data from S3 to your training instance volumes.

So, you would need to allocate enough disk space in your training instances to store your full training dataset and for the final model artifacts

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.

- 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
- 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 **(Correct)**

Explanation

To streamline access to the backend services, you can API Gateway. You can use API Gateway as a Gatekeeper to ensure only authorized users and applications have access to the services. Configure API Gateway to invoke the Lambda function. Lambda function, in turn, invokes SageMaker endpoint. You need to configure Autoscaling to ensure SageMaker endpoint scales on demand. This approach also makes it easy to try different versions of Machine Learning Models without requiring code changes in the client application.

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.

- 9,000
- (•) 4,500 **(Correct)**
- 2,250
- 18,000

Explanation

SageMakerVariantInvocationsPerInstance is a per minute metric that you can monitor with CloudWatch to trigger Auto Scaling actions. When this value exceeds 4500, Autoscaling needs to add a server to handle the increased workload.

SageMakerVariantInvocationsPerInstance = (MAX_RPS * SAFETY_FACTOR) * 60 = 150 * 0.5 * 60 = 4500

You wish to use a SageMaker notebook within a VPC. SageMaker notebook instances are Internet-enabled, creating a potential security hole in your VPC. How would you use SageMaker within a VPC without opening up Internet access?

- (•) Disable direct Internet access when specifying the VPC for your notebook instance, and use VPC interface endpoints (PrivateLink) to allow the connections needed to train and host your model. Modify your instance's security group to allow outbound connections for training and hosting. **(Correct)**
- No action is required, the VPC will block the notebook instances from accessing the Internet.
- Use IAM to restrict Internet access from the notebook instance.
- Uncheck the option for Internet access when creating your notebook instance, and it will handle the rest automatically.

Explanation

This is covered under "Infrastructure Security" in the SageMaker developer guide. You really do need to read all 1,000+ pages of it and study it in order to ace this certification.

You are developing an autonomous vehicle that must classify images of street signs with extremely low latency, processing thousands of images per second. What AWS-based architecture would best meet this need?

- (•) Develop your classifier with TensorFlow, and compile it for an NVIDIA Jetson edge device using SageMaker Neo, and run it on the edge with IoT GreenGrass. **(Correct)**
- Use Amazon Rekognition on AWS DeepLens to identify specific street signs in a self-contained manner.
- Develop your classifier using SageMaker Object Detection, and use Elastic Inference to accelerate the model's endpoints called over the air from the vehicle.
- Use Amazon Rekognition in edge mode

Explanation

SageMaker Neo is designed for compiling models using TensorFlow and other frameworks to edge devices such as Nvidia Jetson. The low latency requirement requires an edge solution, where the classification is being done within the vehicle itself and not over the air. Rekognition (which doesn't have an "edge mode," but does integrate with DeepLens) can't handle the very specific classification task of identifying different street signs and what they mean.