← Back to the **Course**

Level: Advanced

AWS Certified Machine Learning Specialty

Data Engineering

Completed on **Mon, 27 Jun 2022**

1st	6/11	54.55%
Attempt	Marks Obtained	Your Score

0h 12m 48s	FAIL
Time Taken	Result

Domain wise Quiz Performance Report

Join us on Slack community

No.	Domain	Total Question	Correct	Incorrect	Unattempte
1	Data Engineering	11	6	4	1
Total	**All Domains**	**11**	**6**	**4**	**1**

Review the Answers

Filter By All Questions

Question 1

Incorrect

Domain: Data Engineering

Your machine learning team is responsible for processing video clips posted to your company's Twitter social media account to understand the sentiment of the video clips. Your team takes these video clips and labels them with the appropriate sentiment so that your marketing department can use them in their advertising campaigns. You are now expanding into Spanish and Portuguese speaking regions of the world. So you now need to translate video clip audio as a part of your sentiment labeling process.

What AWS services and SageMaker built-in algorithms allow your team to label the foreign language video clips in the most efficient manner?

A. Transcribe -> Comprehend -> Translate right

B. Transcribe -> Comprehend -> SageMaker seq2seq

C. Transcribe -> Translate -> SageMaker Neural Topic Model (NTM) wrong

D. Transcribe -> Translate -> SageMaker BlazingText

E. Translate -> Transcribe -> Comprehend

Explanation:

Answer: A

Option A is CORRECT. This is the most efficient option given. It is accomplished completely using AWS managed services. Comprehend can identify sentiment in Spanish and Portuguese language transcriptions. You don't have to translate to English before using Comprehend.

Option B is incorrect. Using the SageMaker seq2seq built-in algorithm overcomplicates the solution. With this option, your team would have to build a machine learning model to translate the Spanish or Portuguese language to English when using the Translate service for this step.

Option C is incorrect. Using the SageMaker Neural Topic Model built-in algorithm overcomplicates the solution. With this option, your team would have to build a machine learning model to identify the sentiment in the video clips when you could just use the Comprehend service for this step.

Option D is incorrect. Using the SageMaker BalzingText built-in algorithm overcomplicates the solution. With this option, your team would have to build a machine learning model to identify the sentiment in the video clips when you could just use the Comprehend service for this step.

Option E is incorrect. The Amazon Translate service requires the input of a text document. In this option, you are attempting to translate directly from the audio of your video clip.

Reference:

Please see the **Amazon Comprehend developer guide** titled **Languages Supported in Amazon Comprehend**.

Please refer to the **Amazon Translate developer guide** titled **What Is Amazon Translate?.**

Please review the **Amazon Translate developer guide** titled **How Amazon Translate Works.**

Ask our Experts

View Queries

Question 2

Correct

Domain: Data Engineering

You work for a machine learning team at a global retail auto parts chain. Your team ingests purchasing data from its 100,000 global auto parts stores to S3 using Kinesis Data Firehose. You are now ready to start training an improved machine learning model that will be used to predict purchasing patterns by global region. The training data requires additional simple transformations. Also, you will need to combine some data attributes. Finally, your team expects to train the model on a daily basis.

Based on a large number of stores plus changing data ingestion, which of the following options will require the least amount of administration and development effort?

A. Have the stores capture their purchasing data locally on Storage Gateway and then load the data into S3. Transform the data using Glue.

B. Create an EMR cluster with Apache Spark installed to perform the transformation logic. Run the cluster each day on the incremental records in S3, and write the transformed records to S3.

C. Create a fleet of EC2 instances that run the transformation logic which transforms the incremental data records on S3, and write the transformed records to S3.

D. Create a Kinesis Data Analytics stream and use it as the destination of the Kinesis Data Firehose stream. Use Kinesis Data Analytics to transform the raw purchasing data attributes into transformed values using SQL and write the transformed data to S3. right

Explanation:

Answer: D

Option A is incorrect. Having 100,000 stores use Storage Gateway to move the data to S3 would require a tremendous administrative effort.

Option B is incorrect. Using EMR for this solution would require administrative cost to build and maintain EMR. Also, your development team would have to write the Apache Spark code to perform the transformations.

Option C is incorrect. Using a fleet of EC2 instances would require the administrative cost of creating and maintaining the EC2 instances. Also, your development team would have to write the transformation logic that runs on the EC2 instances.

Option D is CORRECT. Kinesis Data Analytics can receive your data from Kinesis Data Firehose, transform it, and then write it to S3. The code needed to perform the transformations in Kinesis Data Analytics would be much simpler than the coding suggested in the other options. Your machine learning model can then use the transformed data in S3 for training.

Reference:

Please see the **Amazon Kinesis Data Analytics developer guide** titled **Example: Writing to an Amazon S3 Bucket.**

Please refer to the **Amazon Kinesis Data Firehose developer guide** titled **Using Amazon Kinesis Data Analytics.**

Ask our Experts

View Queries

Question 3

Correct

Domain: Data Engineering

You are a machine learning specialist working for an oil and gas company. Your company's oil and gas drilling sites around the world are equipped with sensors that stream site equipment status and external conditions like weather. You are responsible for building a machine learning model that predicts equipment failures at the sites. The streaming data from the sites needs to be ingested, transformed and stored in Apache Parquet files for exploration and analysis before you use the data in your model.

Which of the following options would ingest, transform, and store your data in the parquet format with the least amount of effort on your part?

A. Kinesis Data Streams

B. Kinesis Data Analytics

C. Kinesis Data Firehose right

D. Managed Streaming for Apache Kafka (MSK)

Explanation:

Answer: C

Option A is incorrect. While you could use Kinesis Data Streams to ingest your sensor data, you would have to write a Kinesis Client Library application or Lambda function to transform the sensor data to the parquet format. This involves more work than using Kinesis Data Firehose.

Option B is incorrect. You cannot stream data directly into Kinesis Data Analytics. You would have to stream your sensor data into either Kinesis Data Streams or Kinesis Data Firehose first and then send your data downstream to Kinesis Data Analytics. This involves more work than using Kinesis Data Firehose.

Option C is CORRECT. With Kinesis Data Firehose, you can stream your sensor data directly to Kinesis Data Firehose, use its built-in parquet transform, then write the parquet files to S3. This approach requires the least amount of work on your part.

Option D is incorrect. While Kafka could be used to stream your sensor data and transform it, this option requires creating an MSK cluster, creating a client machine, creating a topic, and other effort-consuming tasks. This involves more work than using Kinesis Data Firehose.

Reference:

Please see the **Amazon Kinesis Data Streams developer guide** titled **What Is Amazon Kinesis Data Streams?.**

Please refer to the **Amazon Kinesis Data Firehose developer guide** titled **What Is Amazon Kinesis Data Firehose?.**

Please refer to the **Amazon Kinesis Data Analytics for SQL Applications developer guide** titled **What Is Amazon Kinesis Data Analytics for SQL Applications?**

Ask our Experts

View Queries

Question 4

Correct

Domain: Data Engineering

You are a member of a machine learning team at a large online retailer. Your team is responsible for retail competitor analysis. You have a competitive product data streaming source that you need to ingest into your data lake. You need to use that streaming competitor product data to match the corresponding product data in your catalog of products. Using this matching, your data scientists can produce competitive analysis dashboards in a BI tool.

Which of the following options gives you the best data ingestion and most efficient product comparison solution?

A. Kinesis Data Streams -> S3 -> Lake Formation -> QuickSight

B. Kinesis Data Streams -> Elasticsearch -> Kibana

C. Kinesis Data Firehose -> S3 -> Lake Formation -> QuickSight right

D. Kinesis Data Firehose -> Elasticsearch -> Kibana

Explanation:

Answer: C

Option A is incorrect. Kinesis Data Streams cannot write directly to S3. It needs a Kinesis Consumer Library application to receive the data and then write it to S3.

Option B is incorrect. Kinesis Data Streams cannot write directly to Elasticsearch. Also, Elasticsearch has no built-in data matching capability.

Option C is CORRECT. Kinesis Data Firehose can stream directly to S3. Lake Formation has a FindMatches transform that enables you to identify matching records in your dataset, even when the records do not have a common unique identifier. No fields match exactly. In this scenario, you will match products in your product catalog with your competitive product sources even though

the product entries are structured differently. QuickSight allows your data scientists to produce their analysis dashboards.

Option D is incorrect. Kinesis Data Firehose can write directly to Elasticsearch. But Elasticsearch has no built-in data matching capability.

Reference:

Please see the **AWS Lake Formation developer guide** titled **What Is AWS Lake Formation?.**

Please see the **AWS Glue developer guide** titled **Matching Records with AWS Lake Formation FindMatches.**

Ask our Experts

View Queries

Question 5

Correct

Domain: Data Engineering

You are a machine learning specialist working for the digital banking division of a global banking firm. Your bank is in the process of introducing a conversational user interface for its digital banking service. The service will receive streaming audio from the conversational user interface and converse with the user in real-time. Your machine learning team lead has decided to use the Amazon Transcribe service to convert the streaming audio to streaming text.

To handle issues in the network connection when users are on mobile phones, how can you leverage the features of Amazon Transcribe to keep your solution as cost-effective as possible?

A. Use the Transcribe JSON streaming client.

B. Use the Transcribe HTTP/2 streaming client. right

C. Use the Transcribe WebSocket protocol.

D. Use the Transcribe HTTP streaming client.

Explanation:

Answer: B

Option A is incorrect. The streaming client available with Transcribe is an HTTP/2 streaming client.

Option B is CORRECT. You can use the Transcribe HTTP/2 streaming client to handle retrying the connection when there are intermittent problems on the network.

Option C is incorrect. The Transcribe WebSocket protocol does not provide retry logic to handle retrying the connection when there are intermittent problems on the network. With this option, you would have to code the retry logic yourself.

Option D is incorrect. The streaming client available with Transcribe is an HTTP/2 streaming client.

Reference:

Please see the **Amazon Transcribe developer guide** titled **Streaming Transcription.**

Please see the **Amazon Transcribe developer guide** titled **HTTP/2 Streaming Retry Client,**

Please see the **Amazon Transcribe developer guide** titled **Using Amazon Transcribe Streaming with WebSockets**.

Ask our Experts

View Queries

Question 6

Incorrect

Domain: Data Engineering

You are a machine learning specialist working at a language translation software vendor. Your mobile app team would like to add live transcriptions to the company's mobile language conversation application.

Which AWS services should you use to provide the live transcriptions feature to your mobile app?

A. Stream your audio data to Amazon Transcribe Streaming and use the StartTranscriptionJob API call to start a bidirectional HTTP stream that streams your audio to Amazon Transcribe. Amazon Transcribe then streams the transcription results to your application and your app code produces the live transcription.

B. Stream your audio data to Amazon Transcribe Streaming and use the StartStreamTranscription API call to start a bidirectional HTTP stream that streams your audio to Amazon Transcribe. Amazon Transcribe then streams the transcription results to your application and your app code produces the live transcription.

C. Stream your audio data to Amazon Transcribe Streaming and use the StartStreamTranscription API call to start a bidirectional HTTP/2 stream that streams your audio to Amazon Transcribe. Amazon Transcribe then streams the transcription results to your application and your app code produces the live transcription. right

D. Stream your audio data to Amazon Connect Streaming and use the StartStreamTranscription API call to start a bidirectional HTTP/2 stream that streams your audio to Amazon Connect. Amazon Connect then streams the transcription results to your application and your app code produces the live transcription. wrong

Explanation:

Answer: C

Option A is incorrect. The StartTranscriptionJob API is used for Amazon Transcribe in batch translation mode. You are building a real-time transcription service. Also, Amazon Transcribe operates using HTTP/2, not HTTP.

Option B is incorrect. Amazon Transcribe operates using HTTP/2, not HTTP.

Option C is CORRECT. Use the Amazon Transcribe Streaming service and its StartStreamTranscription API. It starts a bidirectional HTTP/2 stream that streams your audio to Amazon Transcribe. Then transcribe streams the transcription results to your application and your app code produces the live transcription to be displayed in your app user interface.

Option D is incorrect. Amazon Connect does not have a transcription feature.

Reference:

Please see the **Amazon Transcribe developer guide** titled **Streaming Transcription.**

Please see the **Amazon Transcribe developer guide** titled **StartTranscriptionJob,**

Please see the **Amazon Transcribe developer guide** titled **StartStreamTranscription,**

Please see the **GitHub repository** titled **Amazon Connect Real-time Transcription Lambda**.

Ask our Experts

View Queries

Question 7

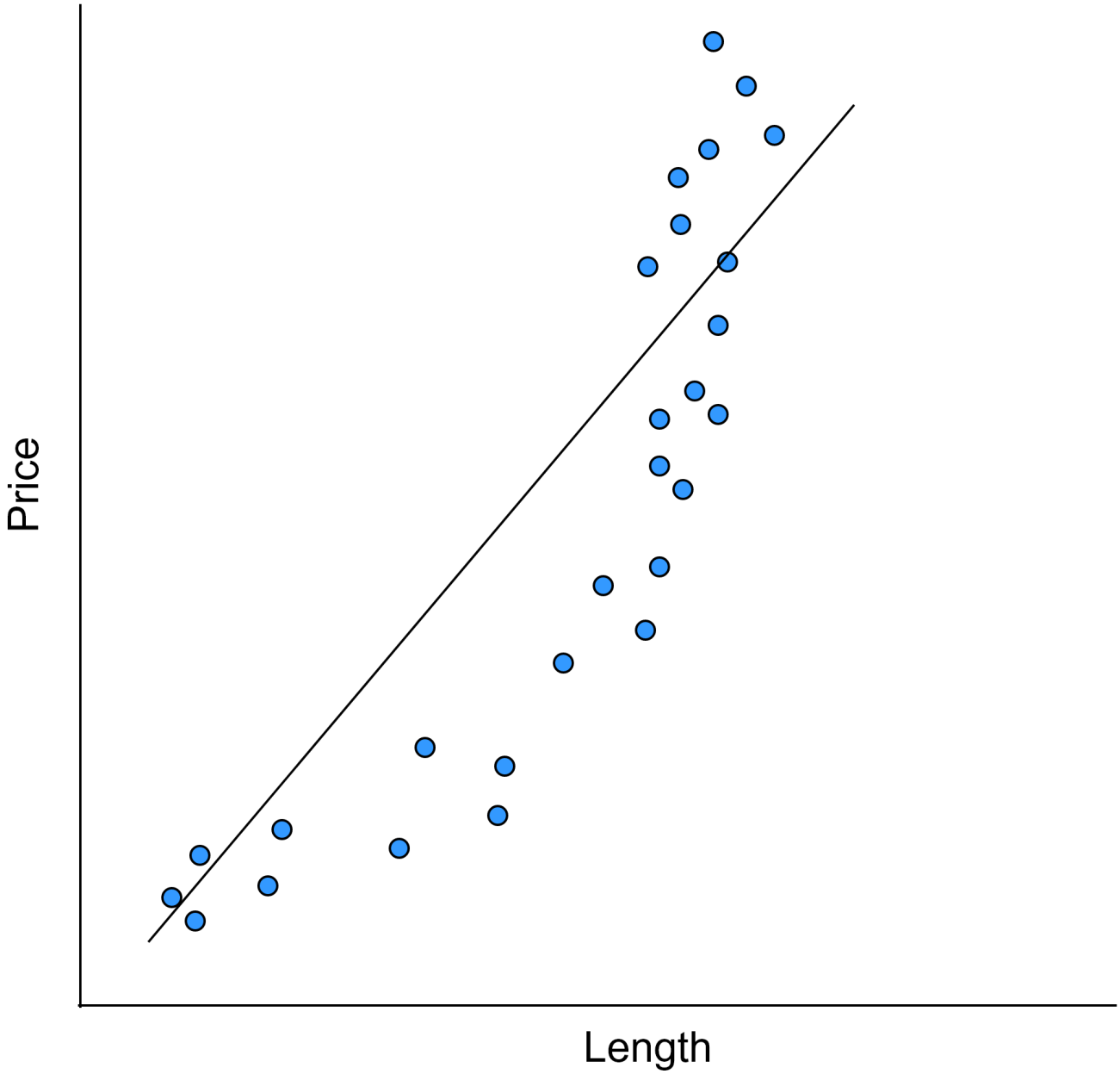
Incorrect

Domain: Data Engineering

0

You are a machine learning specialist working for a farming corporation. Your team is working on a machine learning problem where you are trying to determine the price of a given square plot of corn. You have many features in your data set that you can use in your model, for example: length of the plot in meters, corn type, corn height, longitude and latitude of the plot, etc.

You are trying to understand how you can use the length-of-plot (in meters) feature in your model. You have chosen to use a Linear Learner algorithm. When you fit the model to the length-of-plot feature you get results that are not optimal. As shown by the plot of length to price, you don't get a linear relationship:



How can you transform the length-of-plot feature to make it usable in your Linear Learner based model?

A. Transform the length-of-plot feature by converting the length from meters to feet

B. Transform the length-of-plot feature by determining the Mutual Information (MI) score between the price and the length-of-plot feature wrong

C. Transform the length-of-plot feature by squaring the length-of-plot for each observation to create an area feature and add it to your data set right

D. Transform the length-of-plot feature by dividing the length-of-plot by the corn height for each observation to create a corn-height_to_length-of-plot ratio feature and add it to your data set

Explanation:

Correct Answer: C

Option A is incorrect. Converting the length-of-plot to a different unit of measure, from meters to feet, will not change its relationship to the price. You'll still have a non-linear relationship.

Option B is incorrect. The Mutual Information (MI) score is a feature metric that you can use to measure the association of a feature to your target. It is used to determine which features are most relevant features for your model based on their MI scores (higher is better). This technique will tell you which features carry promise for making your model perform better, but it won't help you transform your length-of-plot feature.

Option C is correct. Squaring the length-of-plot to create an area feature will give you a more linear relationship between area and price. The Linear Learner algorithm works when your features have a linear relationship to the target.

Option D is incorrect. A length-of-plot_to_corn-height transformation generated feature is unlikely to create a linear relationship to the price target.

Reference:

Please see the Kaggle article titled **What Is Feature Engineering** (https://www.kaggle.com/ryanholbrook/what-is-feature-engineering), the Amazon SageMaker developer guide titled **Linear Learner Algorithm** (https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html), the the Towards Data Science article titled **Select Features for Machine Learning Model with Mutual Information** (https://towardsdatascience.com/select-features-for-machine-learning-model-with-mutual-information-534fe387d5c8)

Ask our Experts

View Queries

Question 8

Correct

Domain: Data Engineering

0

You are a machine learning specialist working for a research lab that is studying the crime rates in various cities across Europe. Your task is to project if the crime rate, or number of crimes per day, in the cities in your study group will rise in the next month. You have crime per day data for the cities in your study going back 100 years. You have calculated the mean for each city and most cities present a distribution that looks like the following when you visualize your data in a plot:

In order to create a proper repository for your machine learning model, you need to understand the distribution of your data. Based on your data source and your visualization, which type of distribution do you have?

A. Binomial distribution

B. Poisson distribution right

C. Normal distribution

D. Gaussian distribution

Explanation:

Correct Answer: B

Option A is incorrect. A binomial distribution has a binary outcome, either success or failure. Your data describes the count of events (crimes) in a given period of time (per day).

Option B is correct. A poisson distribution is represented as the count of events in a given period of time. This fits your data: number of crimes per day.

Option C is incorrect. A normal distribution has data that distributes evenly on either side of the mean. A normal distribution forms a “bell-shaped curve.” Your data does not center around the

mean, your data skews towards 0.

Option D is incorrect. A gaussian distribution is another name for a normal distribution.

Reference:

Please see the Towards Data Science article titled **The Poisson Distribution** (https://towardsdatascience.com/the-poisson-distribution-103abfddc312), the Analysis Factor article titled **Differences Between the Normal and Poisson Distributions** (https://www.theanalysisfactor.com/differences-between-normal-and-poisson-distributions/), the the Make Me An Analyst article titled **NORMAL DISTRIBUTION, BINOMIAL DISTRIBUTION & POISSON DISTRIBUTION** (http://makemeanalyst.com/normal-distribution-binomial-distribution-poisson-distribution/)

Ask our Experts

View Queries

Question 9

Unattempted

Domain: Data Engineering

You are a machine learning specialist working for a financial services firm. Your machine learning team is responsible for producing security master data for your quantitative analysis group. Your quant group uses your security master data to feed their quantitative analysis machine learning models to drive their stock selection for their active management portfolios. You need to stream security master data from various sources into your security master data store in near-real time. Which solutions meet your requirements in the most efficient manner? (Select TWO)

A. Stream your security master data using Kafka; extract the data into your online feature store using the PutRecord API call in small batch sets. right

B. Stream your security master data using Spark Streaming; extract the data into your online feature store using the PutRecords API call in large batch sets.

C. Stream your security master data using Kinesis; extract the data into your online feature store using the PutRecord API call in small batch sets. right

D. Stream your security master data using Apache Storm; extract the data into your online feature store using the PutRecords API call in small batch sets.

E. Stream your security master data using Kinesis; extract the data into your online feature store using the PutRecords API call in large batch sets.

Explanation:

Correct Answers: A and C

Option A is correct. You can use SageMaker Feature Store to house your security master data. You can ingest data into SageMaker Feature Store using the PutRecord API call using small batches of data. Kafka is a common service used to stream data into SageMaker Feature Store.

Option B is incorrect. The API call to put your data in small batches into SageMaker Feature Store is the PutRecord API call, not the PutRecords API call. Also, to meet your near-real time requirement, you should use small batches, not large batches of data.

Option C is correct. You can use SageMaker Feature Store to house your security master data. You can ingest data into SageMaker Feature Store using the PutRecord API call using small batches of data. Kinesis is a common service used to stream data into SageMaker Feature Store.

Option D is incorrect. The API call to put your data in small batches into SageMaker Feature Store is the PutRecord API call, not the PutRecords API call.

Option E is incorrect. The API call to put your data in small batches into SageMaker Feature Store is the PutRecord API call, not the PutRecords API call. Also, to meet your near-real time requirement, you should use small batches, not large batches of data.

Reference:

Please see the Amazon SageMaker developer guide titled **Data Sources and Ingestion** (https://docs.amazonaws.cn/en_us/sagemaker/latest/dg/feature-store-ingest-data.html), the Amazon SageMaker developer guide titled **Get started with Amazon SageMaker Feature Store** (https://docs.amazonaws.cn/en_us/sagemaker/latest/dg/feature-store-getting-started.html)

Ask our Experts

View Queries

Question 10

Incorrect

Domain: Data Engineering

You are a machine learning consultant who has been contracted to evaluate and correct a model built by a client's machine learning team. The team's model performs poorly against their selected metric when the team tries to fit the model. The model is built using scikit-learn and uses the RandomForest algorithm. After doing some investigation you see that the data source has missing values in both numeric and categorical features and the machine learning team chose the strategy of dropping the features with missing data. You want to use the scikit-learn ColumnTransformer class to transform the missing data, specifically replacing the missing data in the numeric and categorical data features using imputation. You have decided to replace the numeric missing values with predictions for the missing values, and the categorical missing values with the most frequent value in the feature. Which of the following are the best approaches to achieving your goal? (Select TWO)

A. Create a one-step preprocessing transformer for the numerical missing values that uses a SimpleImputer using an ExtraTreesRegressor estimator

- [x] B. Create a one-step preprocessing transformer for the numerical missing values that uses a SimpleImputer using the constant strategy wrong

- [x] C. Create a two-step preprocessing transformer for the categorical missing values that uses a SimpleImputer using the most_frequent strategy then uses the OneHotEncoder in step two to encode the categorical data right

- [ ] D. Create a two-step preprocessing transformer for the categorical missing values that uses an IterativeImputer using the most_frequent strategy then uses the OneHotEncoder in step two to encode the categorical data

- [ ] E. Create a one-step preprocessing transformer for the numerical missing values that uses a KNNImputer right

Explanation:

Correct Answers: C and E

Option A is incorrect. The scikit-learn SimpleImputer does not have an ExtraTreesRegressor estimator.

Option B is incorrect. This approach will replace all of your missing numerical values with 0, the default for the constant strategy. This will not achieve your goal of replacing the missing numeric values with predictions for the missing values.

Option C is correct. Using the scikit-learn SimpleImputer with the most_frequent strategy and then using the OneHotEncoder class will get the results you are trying to achieve.

Option D is incorrect. You could use the IterativeImputer class with the most_frequent strategy, but the IterativeImputer class is released as experimental. The SimpleImputer class using the most_frequent strategy is a better choice.

Option E is correct. Using the scikit-learn KNNImputer class to impute your numeric missing values will give you the results you need of replacing the missing values with estimates.

Reference:

Please see the Scikit-learn modules page titled **sklearn.impute.SimpleImputer** (https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html), the Scikit-learn modules page titled **sklearn.impute.IterativeImputer** (https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html#sklearn.impute.IterativeImputer), the Scikit-learn modules page titled **sklearn.impute.KNNImputer** (https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html#sklearn.impute.KNNImputer), the Scikit-learn modules page titled **6.4. Imputation of missing values** (https://scikit-learn.org/stable/modules/impute.html), and the Kaggle page titled **Pipelines** (https://www.kaggle.com/alexisbcook/pipelines)

Ask our Experts

View Queries

Question 11

Correct

Domain: Data Engineering

You work as a machine learning specialist for a software company that is building a bird identifier mobile app. The app will allow users to take a photo of a bird and then use the app to process the image and identify the bird. You are building your data repository for your machine learning model to use for training. You need to label the very large set of images you plan to use for training. You have decided to use SageMaker Ground Truth to label your images. You are creating the Ground Truth worker tasks for your workers to use when performing their labeling tasks. You need to identify the birds in your images at the pixel level for accuracy. Which SageMaker Ground Truth built-in task should you use?

A. Bounding Box

B. Image Classification (Single Label)

C. Image Classification (Multi-Label)

D. Image Semantic Segmentation right

Explanation:

Correct Answer: D

Option A is incorrect. The Bounding Box SageMaker Ground Truth built-in task is used to allow your workers to use a box that bounds various subjects in the image. You need your workers to identify the subjects (birds) at the pixel level.

Option B is incorrect. The Image Classification (Single Label) SageMaker Ground Truth built-in task is used to allow your workers to label subjects in your images using a predefined set of labels provided by you. You need your workers to identify the subjects (birds) at the pixel level.

Option C is incorrect. The Image Classification (Multi-Label) SageMaker Ground Truth built-in task is used to allow your workers to label multiple subjects in your images. You need your workers to identify the subjects (birds) at the pixel level.

Option D is correct. The Semantic Segmentation Ground Truth built-in task is used to allow your workers to label subjects in your image at the pixel level.

Reference:

Please see the Amazon SageMaker developer guide titled **Label Data** (https://docs.aws.amazon.com/sagemaker/latest/dg/data-label.html), the Amazon SageMaker developer guide titled **Built-in Task Types** (https://docs.aws.amazon.com/sagemaker/latest/dg/sms-task-types.html), the Amazon SageMaker developer guide titled **Image Classification (Single Label)** (https://docs.aws.amazon.com/sagemaker/latest/dg/sms-image-classification.html), the Amazon SageMaker developer guide titled **Image Classification (Multi-**

label) (https://docs.aws.amazon.com/sagemaker/latest/dg/sms-image-classification-multilabel.html), and the Amazon SageMaker developer guide titled **Image Semantic Segmentation** (https://docs.aws.amazon.com/sagemaker/latest/dg/sms-semantic-segmentation.html), and the Amazon SageMaker developer guide titled **Bounding Box** (https://docs.aws.amazon.com/sagemaker/latest/dg/sms-bounding-box.html)

Ask our Experts

View Queries

Finish Review