
Question: 11

Case study

An ML engineer is developing a fraud detection model on AWS. The training dataset includes transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.

The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.

After the data is aggregated, the ML engineer must implement a solution to automatically detect anomalies in the data and to visualize the result.

Which solution will meet these requirements?

- A. Use Amazon Athena to automatically detect the anomalies and to visualize the result.
- B. Use Amazon Redshift Spectrum to automatically detect the anomalies. Use Amazon QuickSight to visualize the result.
- C. Use Amazon SageMaker Data Wrangler to automatically detect the anomalies and to visualize the result.
- D. Use AWS Batch to automatically detect the anomalies. Use Amazon QuickSight to visualize the result.

Answer: C

Explanation:

Question: 12

Case study

An ML engineer is developing a fraud detection model on AWS. The training dataset includes transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.

The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.

The training dataset includes categorical data and numerical data.

a. The ML engineer must prepare the training dataset to maximize the accuracy of the model.

Which action will meet this requirement with the LEAST operational overhead?

- A. Use AWS Glue to transform the categorical data into numerical data.
- B. Use AWS Glue to transform the numerical data into categorical data.
- C. Use Amazon SageMaker Data Wrangler to transform the categorical data into numerical data.
- D. Use Amazon SageMaker Data Wrangler to transform the numerical data into categorical data.

Answer: C

Explanation:

Question: 13

Case study

An ML engineer is developing a fraud detection model on AWS. The training dataset includes transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.

The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.

Before the ML engineer trains the model, the ML engineer must resolve the issue of the imbalanced data.

Which solution will meet this requirement with the LEAST operational effort?

- A. Use Amazon Athena to identify patterns that contribute to the imbalance. Adjust the dataset accordingly.
- B. Use Amazon SageMaker Studio Classic built-in algorithms to process the imbalanced dataset.
- C. Use AWS Glue DataBrew built-in features to oversample the minority class.
- D. Use the Amazon SageMaker Data Wrangler balance data operation to oversample the minority class.

Answer: D

Explanation:

Question: 14

Case study

An ML engineer is developing a fraud detection model on AWS. The training dataset includes transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.

The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.

The ML engineer needs to use an Amazon SageMaker built-in algorithm to train the model.

Which algorithm should the ML engineer use to meet this requirement?

- A. LightGBM
- B. Linear learner
- C. K-means clustering
- D. Neural Topic Model (NTM)

Answer: B

Explanation:

Question: 15

A company has deployed an XGBoost prediction model in production to predict if a customer is likely to cancel a subscription. The company uses Amazon SageMaker Model Monitor to detect deviations in the F1 score.

During a baseline analysis of model quality, the company recorded a threshold for the F1 score. After several months of no change, the model's F1 score decreases significantly.

What could be the reason for the reduced F1 score?

- A. Concept drift occurred in the underlying customer data that was used for predictions.
- B. The model was not sufficiently complex to capture all the patterns in the original baseline data.
- C. The original baseline data had a data quality issue of missing values.
- D. Incorrect ground truth labels were provided to Model Monitor during the calculation of the baseline.

Answer: A

Explanation:

Question: 16

A company has a team of data scientists who use Amazon SageMaker notebook instances to test ML models. When the data scientists need new permissions, the company attaches the permissions to each individual role that was created during the creation of the SageMaker notebook instance.

The company needs to centralize management of the team's permissions.

Which solution will meet this requirement?

- A. Create a single IAM role that has the necessary permissions. Attach the role to each notebook instance that the team uses.
- B. Create a single IAM group. Add the data scientists to the group. Associate the group with each notebook instance that the team uses.
- C. Create a single IAM user. Attach the AdministratorAccess AWS managed IAM policy to the user. Configure each notebook instance to use the IAM user.
- D. Create a single IAM group. Add the data scientists to the group. Create an IAM role. Attach the AdministratorAccess AWS managed IAM policy to the role. Associate the role with the group. Associate the group with each notebook instance that the team uses.

Answer: A

Explanation:

Question: 17

An ML engineer needs to use an ML model to predict the price of apartments in a specific location. Which metric should the ML engineer use to evaluate the model's performance?

- A. Accuracy
- B. Area Under the ROC Curve (AUC)
- C. F1 score
- D. Mean absolute error (MAE)

Answer: D

Explanation:

Question: 18

An ML engineer has trained a neural network by using stochastic gradient descent (SGD). The neural network performs poorly on the test set. The values for training loss and validation loss remain high and show an oscillating pattern. The values decrease for a few epochs and then increase for a few epochs before repeating the same cycle.

What should the ML engineer do to improve the training process?

- A. Introduce early stopping.
- B. Increase the size of the test set.
- C. Increase the learning rate.
- D. Decrease the learning rate.

Answer: D

Explanation:

Question: 19

An ML engineer needs to process thousands of existing CSV objects and new CSV objects that are uploaded. The CSV objects are stored in a central Amazon S3 bucket and have the same number of columns. One of the columns is a transaction date. The ML engineer must query the data based on the transaction date.

Which solution will meet these requirements with the LEAST operational overhead?

- A. Use an Amazon Athena CREATE TABLE AS SELECT (CTAS) statement to create a table based on the transaction date from data in the central S3 bucket. Query the objects from the table.
- B. Create a new S3 bucket for processed data. Set up S3 replication from the central S3 bucket to the new S3 bucket. Use S3 Object Lambda to query the objects based on transaction date.
- C. Create a new S3 bucket for processed data. Use AWS Glue for Apache Spark to create a job to query the CSV objects based on transaction date. Configure the job to store the results in the new S3 bucket. Query the objects from the new S3 bucket.
- D. Create a new S3 bucket for processed data. Use Amazon Data Firehose to transfer the data from

the central S3 bucket to the new S3 bucket. Configure Firehose to run an AWS Lambda function to query the data based on transaction date.

Answer: A

Explanation:

Question: 20

A company has a large, unstructured dataset. The dataset includes many duplicate records across several key attributes.

Which solution on AWS will detect duplicates in the dataset with the LEAST code development?

- A. Use Amazon Mechanical Turk jobs to detect duplicates.
- B. Use Amazon QuickSight ML Insights to build a custom deduplication model.
- C. Use Amazon SageMaker Data Wrangler to pre-process and detect duplicates.
- D. Use the AWS Glue FindMatches transform to detect duplicates.

Answer: D

Explanation:

Question: 21

A company needs to run a batch data-processing job on Amazon EC2 instances. The job will run during the weekend and will take 90 minutes to finish running. The processing can handle interruptions. The company will run the job every weekend for the next 6 months.

Which EC2 instance purchasing option will meet these requirements MOST cost-effectively?

- A. Spot Instances
- B. Reserved Instances
- C. On-Demand Instances
- D. Dedicated Instances

Answer: A

Explanation:

Question: 22

An ML engineer has an Amazon Comprehend custom model in Account A in the us-east-1 Region. The ML engineer needs to copy the model to Account B in the same Region.

Which solution will meet this requirement with the LEAST development effort?

- A. Use Amazon S3 to make a copy of the model. Transfer the copy to Account B.
- B. Create a resource-based IAM policy. Use the Amazon Comprehend ImportModel API operation to copy the model to Account B.
- C. Use AWS DataSync to replicate the model from Account A to Account B.
- D. Create an AWS Site-to-Site VPN connection between Account A and Account B to transfer the model.

Answer: B

Explanation:

Question: 23

An ML engineer is training a simple neural network model. The ML engineer tracks the performance of the model over time on a validation dataset. The model's performance improves substantially at first and then degrades after a specific number of epochs. Which solutions will mitigate this problem? (Choose two.)

- A. Enable early stopping on the model.
- B. Increase dropout in the layers.
- C. Increase the number of layers.
- D. Increase the number of neurons.
- E. Investigate and reduce the sources of model bias.

Answer: A,B

Explanation:

Question: 24

A company has a Retrieval Augmented Generation (RAG) application that uses a vector database to store embeddings of documents. The company must migrate the application to AWS and must implement a solution that provides semantic search of text files. The company has already migrated the text repository to an Amazon S3 bucket. Which solution will meet these requirements?

- A. Use an AWS Batch job to process the files and generate embeddings. Use AWS Glue to store the embeddings. Use SQL queries to perform the semantic searches.
- B. Use a custom Amazon SageMaker notebook to run a custom script to generate embeddings. Use SageMaker Feature Store to store the embeddings. Use SQL queries to perform the semantic searches.
- C. Use the Amazon Kendra S3 connector to ingest the documents from the S3 bucket into Amazon Kendra. Query Amazon Kendra to perform the semantic searches.
- D. Use an Amazon Textract asynchronous job to ingest the documents from the S3 bucket. Query