← Back to the **Course**

Level: Advanced

AWS Certified Machine Learning Specialty

Practice Test I

Completed on **Sat, 02 Jul 2022**

1st	59/65	90.77%
Attempt	Marks Obtained	Your Score

0h 43m 50s	PASS
Time Taken	Result

Domain wise Quiz Performance Report

Join us on Slack community

No.	Domain	Total Question	Correct	Incorrect	Unattempte
1	Data Engineering	12	12	0	0
2	Exploratory Data Analysis	15	12	3	0
3	Modeling	24	23	1	0
4	ML Implementation and Operations	14	12	2	0
Total	**All Domains**	**65**	**59**	**6**	**0**

Review the Answers

Filter By All Questions

Question 1

Correct

Domain: Data Engineering

You are a machine learning expert working for a marketing firm. You are supporting a team of data scientists and marketing managers who are running a marketing campaign. Your data scientists and marketing managers need to answer the question, “Will this user subscribe to my campaign?” You have been given a dataset in the form of a CSV file which is formatted as such:

UserId, jobId, jobDescription, educationLevel, campaign, duration, willRespondToCampaign

When you perform feature engineering on this dataset, which of the following data types would you use to define the willRespondToCampaign attribute?

A. CATEGORICAL

B. TEXT

C. BINARY right

D. Numeric

Explanation:

Answer: C

Option A is incorrect because you choose the CATEGORICAL data type for an attribute that holds a limited set of unique strings. For example, a user name, the region, and a product code are categorical values. The willRespondToCampaign attribute takes on either ‘yes’ or ‘no’ values, which are binary in nature.

Option B is incorrect because for each user observation you are trying to discern “Will this user subscribe to my campaign?” You are solving for a "yes" or "no" answer, which is binary data type, not a text data type.

Option C is correct because you choose the BINARY data type for an attribute that only has two possible values, such as yes or no, or true or false. The attribute willRespondToCampaign has only two possible answers: yes or no.

Option D is incorrect because the willRespondToCampaign feature holds a "yes" or "no" value, you should define it as a binary data type, not a numeric data type.

Reference:

Please see the Machine Learning Mastery article titled **Discover Feature Engineering, How to Engineer Features and How to Get Good at It**

https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

for a complete description of the schema attributes.

Ask our Experts

View Queries

Question 2

Correct

Domain: Data Engineering

You work for an energy company that buys and sells energy to customers. To get the best prices for their energy customers, your company trades financial energy derivative futures contracts. The trading of these future contracts requires accurate forecasting of energy prices. You need to build a model that compares spot prices (current commodity price) to future commodity prices (the price that a commodity can be bought or sold in the future). Your model needs to assist your company's futures traders in hedging against future energy price changes based on current price predictions. To source the model with appropriate data, you need to gather and process the energy price data automatically.

The data pipeline requires two sources of data:

1. Historic energy spot prices
2. Energy consumption and production rates

Based on the company analysts' requirements, you have decided you need multiple years of historical data. You also realize you'll need to update the data feed daily as the market prices change. You can gather the required data through APIs from data provider vendor systems. Your company's traders require a forecast from your model multiple times per day to help them form their trading strategy. So your pipeline needs to call the data provider APIs multiple times per day. Your data-ingestion pipeline needs to take the data from the API calls, perform preprocessing, and then store the data in an S3 data lake from which your forecasting model will access the data.

Your data-ingestion pipeline has three main steps:

1. Data ingestion
2. Data storage
3. Inference generation

Assuming you have written a lambda function that interacts with the data provider APIs and stores the data in CSV format, which of the following python libraries are the best option to perform the data preprocessing to transform the data by changing raw feature vectors into a format best suited for a SageMaker batch transform job to generate your forecast?

A. matplotlib and plotly

B. boto3 and moto

C. pandas and scikit-learn right

D. NLTK and scrapy

Explanation:

Answer: C

Option A is incorrect because matplotlib and plotly are data visualization python libraries that contain no data transformation functions (see https://matplotlib.org and https://plot.ly/python/).

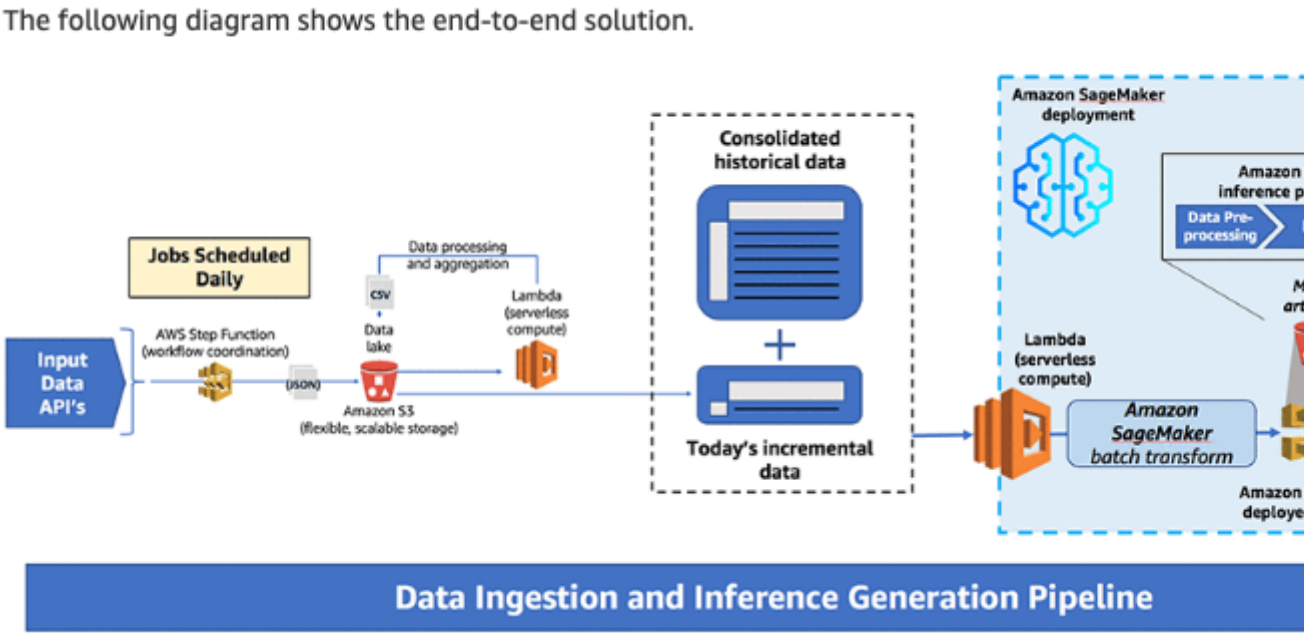
Option B is incorrect because boto3 is a python library used to interface with AWS services such as S3, DynamoDB, SQS, etc. Boto3 has no data transformation functions (see https://aws.amazon.com/sdk-for-python/). Moto is a python library used to mock interfaces to AWS services such as S3, DynamoDB, SQS, etc. The moto library also contains no data transformation functions (see https://pypi.org/project/moto/).

Option C is correct because pandas are the best choice for data wrangling and manipulation of tabular data such as CSV formatted data (see https://pypi.org/project/pandas/). Scikit-learn is the best python package to transform raw feature vectors into a format suited to downstream estimators (see https://scikit-learn.org/stable/modules/preprocessing.html).

Option D is incorrect because Natural Language Toolkit (NLTK) is best suited to text tagging, classification, and tokenizing, not manipulating tabular data (see https://www.nltk.org). Scrapy is best suited to crawling functionality used to gather structured data from websites, not manipulating tabular data (see https://scrapy.org).

Diagram:

Here is a screenshot from the AWS Machine Learning blog depicting the solution:



Reference:

Please see the scikit-learn preprocessing data documentation: https://scikit-learn.org/stable/modules/preprocessing.html, and a detailed pandas example: https://towardsdatascience.com/why-and-how-to-use-pandas-with-large-data-9594dda2ea4c

Ask our Experts

View Queries

Question 3 Correct

Domain: Modeling

You work for a retail firm that wishes to conduct a direct mail campaign to attract new customers. Your marketing manager wishes to get answers to questions that can be put into discrete categories, such as "using historical customer email campaign responses. Should this customer receive an email from our current campaign?" You decide to use the SageMaker Linear Learner algorithm to build your model. Which hyperparameter setting would you use to get the algorithm to produce discrete results?

A. set the objective hyperparameter to reg:logistic.

B. set the predictor_type hyperparameter to binary_classifier. right

C. set the predictor_type hyperparameter to regressor.

D. set the objective hyperparameter to reg:linear.

Explanation:

Answer: B

Option A is incorrect because the objective hyperparameter is set to reg:logistic when using the XGBoost algorithm (See the AWS SageMaker developer documentation: https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html).

Option B is correct because the AWS SageMaker documentation states that you set the predictor type hyperparameter to binary_classifier when using the Linear Learner algorithm for this type of discrete classification problem. (See the AWS SageMaker documentation: https://sagemaker.readthedocs.io/en/stable/linear_learner.html).

Option C is incorrect because the predictor_type hyperparameter is set to regressor when you are using the Linear Learner algorithm for answers that are quantitative, not discrete (See the AWS SageMaker documentation: https://sagemaker.readthedocs.io/en/stable/linear_learner.html).

Option D is incorrect because the objective hyperparameter is set to reg:linear when you are using the XGBoost algorithm for quantitative answers (See the AWS SageMaker developer documentation: https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html).

Reference:

Please see the AWS SageMaker developer guide titled **Using Amazon SageMaker Built-in Algorithms**: https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html) for a complete

description of the SageMaker hyperparameter settings.

Ask our Experts

View Queries

Question 4

Correct

Domain: Modeling

You work for the information security department of a major corporation. You have been asked to build a solution that detects web application log anomalies to protect your organization from fraudulent activity. The system needs to have near-real-time updates to the model where log entry data points dynamically change the underlying model as the log files are updated. Which AWS service component do you use to implement the best algorithm based on these requirements?

A. SageMaker Random Cut Forest

B. Kinesis Data Streams Naive Bayes Classifier

C. Kinesis Data Analytics Random Cut Forest right

D. Kinesis Data Analytics Nearest Neighbor

Explanation:

Answer: C

Option A is incorrect because SageMaker Random Cut Forest is best used for large batch data sets where you don't need to update the model frequently (See AWS Kinesis Data Analytics documentation: https://docs.aws.amazon.com/kinesisanalytics/latest/sqlref/sqlrf-random-cut-forest.html).

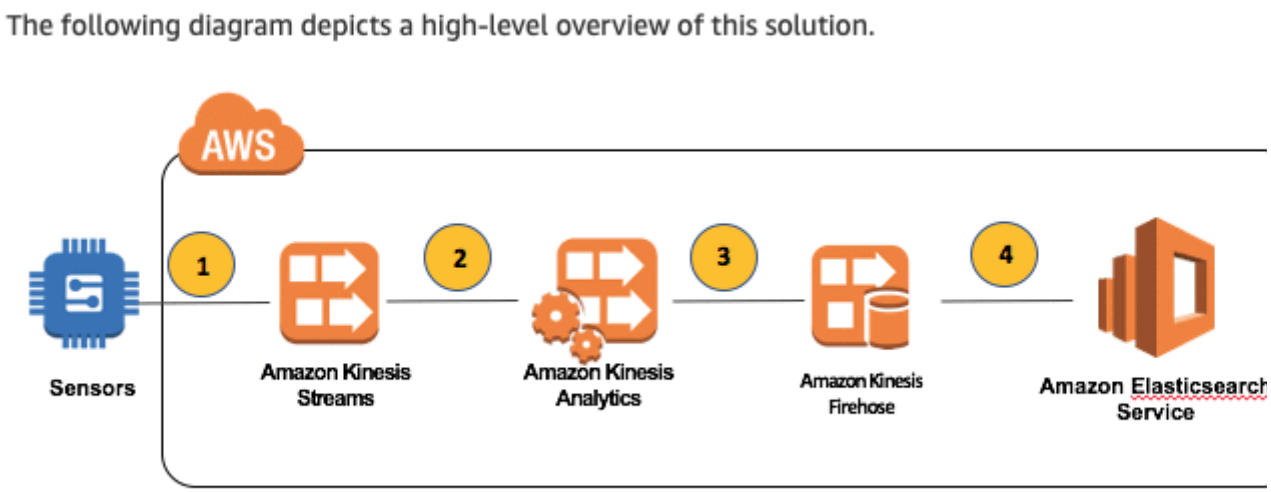
Answer B is incorrect because the Naive Bayes Classifier is used to find independent data points. The Kinesis Data Streams service does not have machine learning algorithm capabilities (See the AWS Kinesis Streams developer documentation: https://docs.aws.amazon.com/streams/latest/dev/introduction.html).

Option C is correct. The Kinesis Data Analytics Random Cut Forest algorithm works really well for near-real-time updates to your model (See the AWS Kinesis Data Analytics documentation: https://docs.aws.amazon.com/kinesisanalytics/latest/sqlref/sqlrf-random-cut-forest.html).

Option D is incorrect because Kinesis Data Analytics provides a hotspots function that detects higher than normal activity using the distance between a hotspot and its nearest neighbor. But it does not provide ML model update capabilities (See AWS Kinesis Data Analytics documentation: https://docs.aws.amazon.com/kinesisanalytics/latest/sqlref/sqlrf-hotspots.html).

Diagram:

Here is a screenshot from the AWS Big Data blog:



Reference:

For an example, please see the AWS Big Data blog post titled **Perform Near Real-time Analytics on Streaming Data with Amazon Kinesis and Amazon Elasticsearch Service**: https://aws.amazon.com/blogs/big-data/perform-near-real-time-analytics-on-streaming-data-with-amazon-kinesis-and-amazon-elasticsearch-service/) for a complete description of the use of Kinesis Data Analytics and the random cut forest algorithm.

Ask our Experts

View Queries

Question 5

Correct

Domain: Exploratory Data Analysis

You work in the data analytics department of a ride sharing software company. You need to use the K-means machine learning algorithm to separate your company's optimized ride data into clusters based on ride coordinates. How would you use AWS Glue in the best way to build the data schema needed to classify the ride data?

A. Use Glue crawlers to crawl your ride share data. right

B. Use Glue FindMatches to find and remove duplicate records in your data.

C. Use Glue to automatically generate code to classify the ride data based on coordinates.

D. Use Glue to transform and flatten your data so that you can classify the ride data based on coordinates.

Explanation:

Answer: A

Option A is correct. The best way to build the schema for your data is to use a Glue crawler that leverages a classifier or multiple classifiers. (See the AWS Glue crawler https://docs.aws.amazon.com/glue/latest/dg/add-crawler.html documentation).

Answer B is incorrect because there is no stated need to remove duplicates from the data.

Option C is incorrect because you don't need to automatically generate code since Glue will generate your schema for your data based on a prioritized list of classifiers without custom code (See the AWS Glue developers guide: (https://docs.aws.amazon.com/glue/latest/dg/add-classifier.html).

Option D is incorrect because there is no stated requirement to flatten the ride data.

Reference:

For an example, please see the AWS Machine Learning blog post titled Serverless unsupervised machine learning with AWS Glue and Amazon Athena: https://aws.amazon.com/blogs/machine-learning/serverless-unsupervised-machine-learning-with-aws-glue-and-amazon-athena/.

Ask our Experts

View Queries

Question 6

Correct

Domain: ML Implementation and Operations

You work in the security department of your company's IT division. Your company has decided to try to use facial recognition to improve security on their campus. You have been asked to design a system that augments your company's building access security by scanning the faces of people entering their buildings and recognizing the person as either an employee/contractor/consultant, who is in the company's database, or visitor, who is not in their database.

Your company has over 750,000 employees and over 250,000 contractors and consultants across their many campus locations worldwide. These workers are all registered in their HR database. Each of these workers has an image of their face stored in the HR database. You have decided to use Amazon Rekognition for your facial recognition solution. On occasion, the Rekognition model fails to recognize visitors to the buildings. What could be the source of the problem?

A. Face landmarks filters are set to a max sharpness.

B. Bounding box and confidence score for face comparison threshold tolerances are set to max values.

C. Confidence threshold tolerance is set to the default.

D. Face collection contents right

Explanation:

Answer: D

Option A is incorrect. From the Amazon Rekognition FAQs: "Face landmarks are a set of salient points, usually located on the corners, tips or midpoints of key facial components such as the eyes, nose, and mouth. Amazon Rekognition DetectFaces API returns a set of face landmarks that can be used to crop faces, morph one face into another, overlay custom masks to create custom filters, and more." Face landmarks don't have a sharpness parameter.

Option B is incorrect. The bounding box and confidence score are used to determine confidence in the Rekognition comparison result. A maximum confidence score tolerance would not cause failures to recognize faces. A low confidence score tolerance would do.

Option C is incorrect. Similar to option C, the default threshold would not be a common source of recognition failure. A confidence threshold tolerance that is set too low would cause a failure in recognition.

Option D is correct. A suboptimal face collection can be the source of recognition failure. Our face collection has only one image per person. The recommendation from the Amazon Rekognition FAQs is "Besides video resolution, the quality and representative faces part of the face collections to search has a major impact. Using multiple face instances per person with variations like beard, glasses, poses (profile and frontal) will significantly improve the performance."

Reference:

Please see the Amazon Recognition developer guide titled Detecting and Analyzing Faces.

Ask our Experts

View Queries

Question 7

Correct

Domain: ML Implementation and Operations

Your marketing department wishes to understand how their products are being represented in the various social media services in which they have active content streams. They would like insights into the reception of a current product line in order to plan for the roll-out of a new product in the line in the future. You have been tasked with creating a service that organizes the social media content by sentiment across all languages so that your marketing department can determine how best to introduce the new product.

How would you quickly and most efficiently design and build a service for your marketing team that gives insight into the social media sentiment?

A. Use the scikit-learn python library to build a sentiment analysis service to provide insight data to the marketing team's internal application platform. Build a dashboard into the application platform using React or Angular.

B. Use the DetectSentiment Amazon Comprehend API as a service to provide insight data to the marketing team's internal application platform. Build a dashboard into the application platform using React or Angular.

C. Use the Amazon Lex API as a service to implement the solution to provide insight data to the marketing team's internal application platform. Build a dashboard into the application platform using React or Angular.

D. Use Amazon Kinesis to stream the data to S3, Amazon Translate to translate it, Amazon Comprehend to extract the sentiment, Amazon Athena with Amazon QuickSight to build a natural-language-processing (NLP)-powered social media dashboard. right

Explanation:

Answer: D

Option A is incorrect since this option is not the quickest to implement nor is it the most efficient, since developers will have to code a react UI and build an end-point to connect the sentiment service to the React or Angular UI.

Option B is incorrect since this option is also not the quickest to implement nor is it the most efficient, since developers will have to code a react UI and build an end-point to connect the sentiment service to the React or Angular UI.

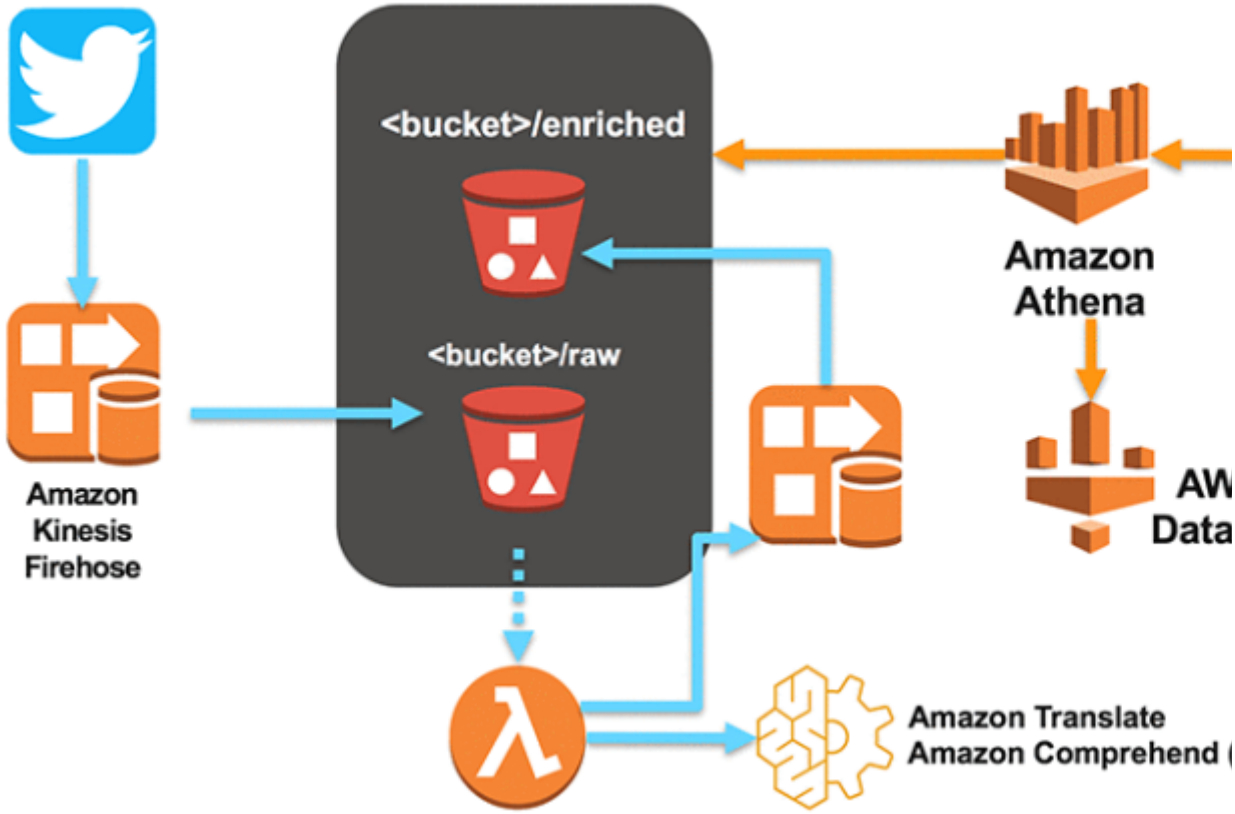
Option C is incorrect since Amazon Lex is used primarily for building conversational interfaces into any application using voice or text. This would not give you the most efficient solution to the problem.

Option D is correct since it is the most efficient and quickest way to implement the solution. Amazon Kinesis Data Firehose is used to capture and prepare the social media content. A lambda function can be used to analyze the social media content using Amazon Translate and Amazon Comprehend. Amazon Athena to query the data produced by the lambda function. Use Amazon QuickSight to produce the dashboard. (See the AWS Machine Learning blog post titled: **Build a social media dashboard using machine learning and BI services:**https://aws.amazon.com/blogs/machine-learning/build-a-social-media-dashboard-using-machine-learning-and-bi-services/)

Diagram:

Here is a screenshot from the AWS Machine Learning blog that shows the desired solution:

The following diagram shows both the ingest (blue) and query (orange) flows.



Reference:

Please see the Amazon Comprehend documentation: https://aws.amazon.com/comprehend/.

Ask our Experts

View Queries

Question 8

Correct

Domain: Exploratory Data Analysis

You work for a financial services firm that wishes to enhance its fraud detection capabilities further. The firm has implemented fine-grained transaction logging for all transactions their customers make using their credit cards. The fraud prevention department would like to use this data to produce dashboards to give them insight into their customer's transaction activity and provide real-time fraud prediction.

You plan to build a fraud detection model using the transaction observation data with Amazon SageMaker. Each transaction observation has a date-time stamp. In its raw form, the date-time stamp is not very useful in your prediction model since it is unique. Can you make use of the date-time stamp in your fraud prediction model, and if so, how?

A. No, you cannot use the date-time stamp since this data point will never occur again. Unique features like this will not help identify patterns in your data.

B. Yes, you can use the date-time stamp data point. You can just use feature selection to deselect the date-time stamp data point, thus dropping it from the learning process.

C. Yes, you can use the date-time stamp data point. You can transform the date-time stamp into features for the hour of the day, the day of the week, and the month. right

D. No, you cannot use the date-time feature since there is no way to transform it into a unique data point.

Explanation:

Answer: C

Option A is incorrect since you can use the date-time stamp if you use feature engineering to transform the data point into a useful form.

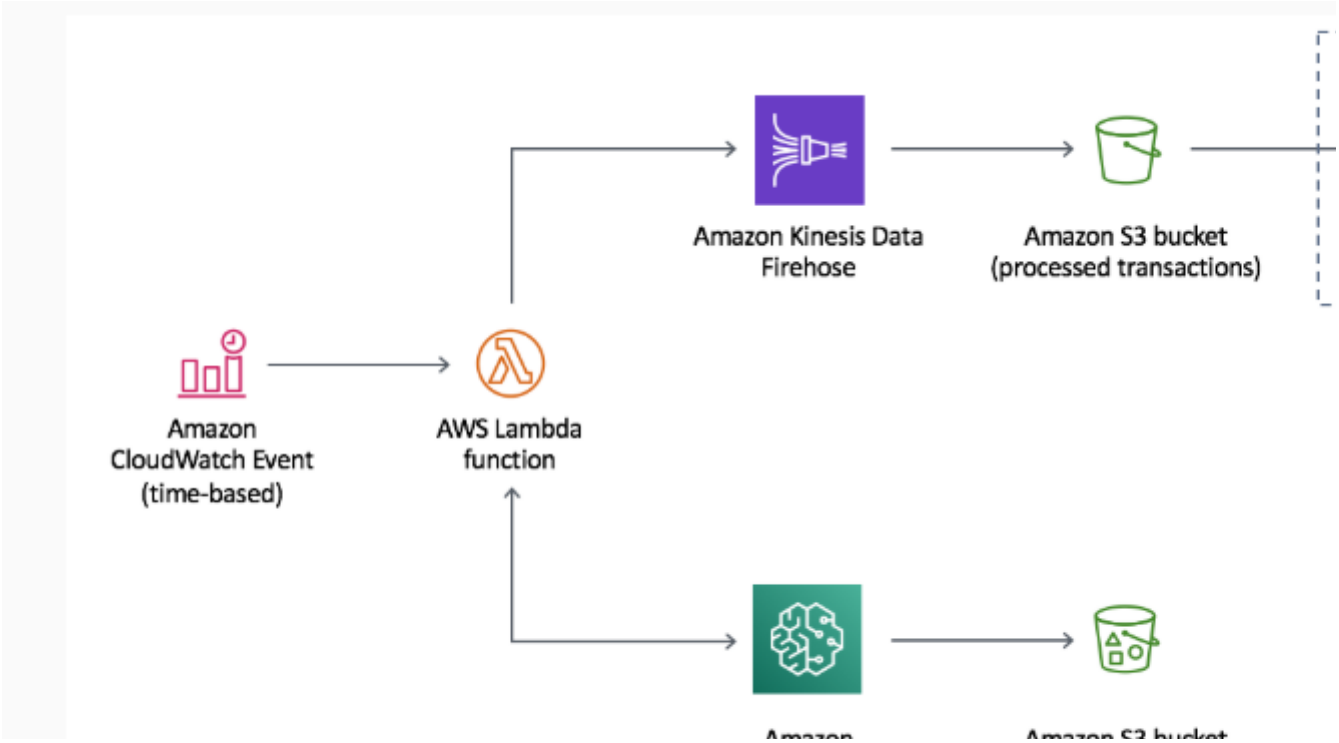
Option B is incorrect since this option is really just another way of ignoring, thus not using, the date-time stamp data point.

Option C is correct. You can transform the data point using feature engineering and thus gain value from it for the learning process of your model. (See the AWS Machine Learning blog post: **Simplify machine learning with XGBoost and Amazon SageMaker:** https://aws.amazon.com/blogs/machine-learning/simplify-machine-learning-with-xgboost-and-amazon-sagemaker/)

Option D is incorrect since we can transform the data point into unique features that represent the hour of the day, the day of the week, and the month. These variables could be useful to learn if the fraudulent activity tends to happen at a particular hour, day of the week, or month.

Diagram:

Here is a screenshot from the AWS Machine Learning documentation depicting a typical fraud detection machine learning solution:



SageMaker (example dataset)

Reference:

Please see the Amazon Machine Learning developer documentation:

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

Ask our Experts

View Queries

Question 9

Correct

Domain: Modeling

You work for a real estate company where you are building a machine learning model to predict the prices of houses. You are using a regression decision tree. As you train your model, you see that it is overfitted to your training data, and it doesn't generalize well to unseen data. How can you improve your situation and get better training results most efficiently?

- A. Use a random forest by building multiple randomized decision trees and averaging their outputs to get the predictions of the housing prices. right
- B. Gather additional training data that gives a more diverse representation of the housing price data.
- C. Use the "dropout" technique to penalize large weights and prevent overfitting.
- D. Use feature selection to eliminate irrelevant features and iteratively train your model until you eliminate the overfitting.

Explanation:

Answer: A

Option A is correct because the random forest algorithm is well known to increase the prediction accuracy and prevent overfitting that occurs with a single decision tree. (See these articles comparing the decision tree and random forest algorithms: https://medium.com/datadriveninvestor/decision-tree-and-random-forest-e174686dd9eb and https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991)

Option B is incorrect since gathering additional data will not necessarily improve the overfitting problem, especially if the additional data has the same noise level of the original data.

Option C is incorrect since while the "dropout" technique improves models that are overfitted, it is a technique used with neural networks, not decision trees.

Option D is incorrect since it requires significantly more effort than using the random forest algorithm approach.

Reference:

Please see this overview of the random forest machine learning algorithm:

https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb

Ask our Experts

View Queries

Question 10

Correct

Domain: Exploratory Data Analysis

You work for the data analytics department of your company, where you have been asked to build a visualization of the company's corporate performance for the company's annual report. The visualization needs to demonstrate company performance by showing how likely it is for a customer to recommend your company's products, and how much profit a customer brings to the business net acquisition and retention costs. Which types of charts would you use to create this visualization? (Select TWO)

A. Use a distribution scatter chart to show the customer recommendation rate.

B. Use a Conversion Rate KPI chart to show the conversion rate of customers.

C. Use a Relative Market Share KPI chart to show competitive market share.

- [x] D. Use a Net Promoter Score KPI chart to graph customer recommendations. right
- [x] E. Use a Customer Profitability Score KPI chart to show customer profitability. right

Explanation:

Answers: D and E

Option A is incorrect because a distribution scatter chart would show the size of recommendations, but not the likelihood of a recommendation.

Option B is incorrect since a Conversion Rate KPI shows how many leads were converted to customers, not the likelihood of recommendation or customer profitability.

Option C is incorrect since the Relative Market Share KPI shows how much market share your company owns versus your company's competitors.

Option D is correct since the Net Promoter Score KPI shows how likely a current customer would recommend your company's products.

Option E is correct since the Customer Profitability Score KPI shows how much profit a customer contributes to your company's profits after the expenses of acquiring the customer and the expenses associated with retaining the customer.

Reference:

Please see this AWS overview of analyzing and visualizing your data for business analytics:

https://aws.amazon.com/data-visualization/

Ask our Experts

View Queries

Question 11

Correct

Domain: Modeling

You work for a manufacturing company that produces retail apparel, such as shoes, dresses, blouses, etc. Your head of manufacturing has asked you to use your data science skills to determine which product, among a list of potential next products, your company should invest its resources to produce. You decide that you need to predict the sales levels of each of the potential next products and select the one with the highest predicted purchase rate. Which type of machine learning approach should you use?

A. You are trying to solve for the greatest number of sales across the potential next products. Therefore, you are solving a multiclass classification problem, and you should use multinomial logistic regression.

B. You are trying to solve for the greatest number of sales across the potential next products. Therefore, you are solving a classification problem, and you should use the random cut forest model.

C. You are trying to solve for the greatest number of sales across the potential next products. Therefore, you are solving a regression problem, and you should use a linear regression model. right

D. You are trying to solve for the greatest number of sales across the potential next products. Therefore, you are solving a binary classification problem, and you should use a logistic regression model.

Explanation:

Answer: C

Option A is incorrect. This is not a multiple classification problem that you are trying to solve for more than two outcomes. So, a multinomial logistic regression would be the wrong choice for your machine learning model.

Option B is incorrect. You are trying to solve a numeric result: the number of purchases customers will make for each next potential product. From the Amazon SageMaker developer guide titled How RCF Works: "Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a dataset."

Option C is correct. You are trying to solve a numeric result: the number of purchases customers will make for each next potential product. This numeric result case calls for the use of a regression model such as the linear regression model.

Option D is incorrect. This is not a classification problem where you're solving a binary (yes or no) result. So, a logistic regression model would be the wrong choice for your machine learning.

Reference:

Please see this AWS overview of machine learning concepts: https://docs.aws.amazon.com/machine-learning/latest/dg/machine-learning-concepts.html, and the Amazon Machine Learning developer guide titled: Types of ML Models.

Ask our Experts

View Queries

Question 12

Correct

Domain: ML Implementation and Operations

You work as a data scientist manager at a large financial services firm where your team is responsible for building machine learning solutions such as price prediction of equities, futures, and options. You need petabytes of data from dozens of sources internal and external to your organization. All external data sources are contractually constrained as to where the data is used and who has access to the data. Your machine learning models require storage of these data in a data lake to allow quick retrieval of data to fuel your ML models. You have chosen to use S3 to house your data lake. How will you most efficiently protect this data lake, your machine learning data source, against internal threats to data confidentiality and security?

A. Create IAM resource-based policies for each data lake S3 bucket resource. Use bucket policies and Access Control Lists (ACLs) to control the resources at the bucket level and the object level.

B. Create IAM user policies so that permissions to access your S3 data lake assets are linked to user roles and permissions. Place your data scientists into IAM groups and assign the user policies to those groups. These policies and permissions will define access to the data processing and analytics services which your data scientists will use. right

C. Create an access key ID and a secret access key for each internal user of your S3 data lake. Your internal users will then only be able to gain access to your data lake using these keys.

D. Use the AWS CloudHSM cloud-based hardware security module (HSM) to secure your S3 data lake. Internal users of your data lake will use the encryption keys generated by the CloudHSM

module to gain access to the data needed for their machine learning models.

Explanation:

Answer: B

Option A is incorrect because this is a very inefficient approach to the problem of securing a data lake. Most large data lakes contain large numbers of buckets and objects. Using resource-based policies would mean creating an extensive set of policies to secure the data lake.

Option B is correct. Per the AWS white paper Building Big Data Storage Solutions (Data Lakes) for Maximum Flexibility - Securing, Protecting, and Managing Data, "for most data lake environments, we recommend using user policies so that permissions to access data assets can also be tied to user roles and permissions for the data processing and analytics services and tools that your data lake users will use."

Option C is incorrect because access keys are used primarily for applications running outside the AWS environment. Resources running inside AWS, as is the case in this scenario, the best practice is to use IAM roles and policies. (See AWS Security blog entry **Guidelines for protecting your AWS account while using programmatic access:** https://aws.amazon.com/blogs/security/guidelines-for-protecting-your-aws-account-while-using-programmatic-access/)

Option D is incorrect. The CloudHSM module is used to generate encrypted access keys. However, since we're dealing with users internal to AWS, IAM role-based security is the best practice for this scenario.

Reference:

Please see the AWS white paper Building Big Data Storage Solutions (Data Lakes) for Maximum Flexibility - Securing, Protecting, and Managing Data.

Ask our Experts

View Queries

Question 13

Correct

Domain: Exploratory Data Analysis

You work for a financial services company where you are building a model to analyze equity futures prices to predict price movement for your firm's hedging strategy. You receive several data feeds, some of which contain missing values for some data points. The missing data points in your data feeds are of the categorical type, such as the expiration month or the exchange on which the futures contract is traded. Which strategy should you employ to deal with the missing data point values while attempting to maximize the accuracy of your model without introducing bias into the model?

A. Remove the observations that have the missing data.

B. Impute the missing values using the Mean/Median strategy.

C. Impute the missing values using the Most Frequent strategy.

D. Impute the missing values using a Deep Learning strategy. right

Explanation:

Answer: D

Option A is incorrect because this approach will lead to the loss of data points with potentially useful information.

Option B is incorrect because, by definition, it can only be used with numeric data. It is not advisable to use the Mean/Median approach with categorical data points.

Option C is incorrect because while working with categorical data, this method can introduce bias into your data.

Option D is correct. Using a library such as the datawig python library, a deep learning approach uses deep neural networks to impute missing data values. This is the most accurate strategy, in the list of given options, at imputing categorical values. (See the datawig documentation: https://github.com/awslabs/datawig)

Reference:

Please see the article 6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples)

Ask our Experts

View Queries

Question 14

Correct

Domain: Modeling

You are a data scientist working for a cancer screening center. The center has gathered data on many patients that have been screened over the years. The data is obviously skewed toward true negative results, as most screened patients don't have cancer. You evaluate several machine learning models to decide which model best predicts true positives when using your cancer screening data. You have split your data into a 70/30 ratio of the training set to the test set. You now need to decide which metric to use to evaluate your models.

Which metric will most accurately determine the model best suited to solve your classification problem?

A. ROC Curve

B. Precision

C. Recall

D. PR Curve right

Explanation:

Answer: D

Option A is incorrect because it is best used when both outcomes have equal importance. Due to the importance of true negative in this equation, it will not differentiate models well for the cancer screening problem, since this data set is skewed to true negatives. The true negative cases are heavily weighted in the equation, thus amplifying the impact of the imbalance.

Option B is incorrect because it only takes into account the percentage of positive cases out of the total predicted positive.

Option C is incorrect because it only takes into account the percentage of positive cases out of the total actual positive.

Option D is correct because the PR Curve is best used to evaluate models on data sets where most of the cases are negative, as in the cancer screening data set. The true negative cases are not weighted heavily in the equation, thus reducing the impact of the imbalance.

Reference:

Please see the article Various ways to evaluate a machine learning model's performance.

Ask our Experts

View Queries

Question 15

Correct

Domain: ML Implementation and Operations

You work for a web retailer where you need to analyze data produced for your company by an outside market data provider. You need to produce recommendations based on patterns in user preferences by demographic found in the supplied data. You have stored the data in one of your company's S3 buckets. You have created a Glue crawler that you have configured to crawl the data on S3 and you have written a custom classifier. Unfortunately, the crawler failed to create a schema. Why might the Glue crawler have failed in this way?

A. You did not add an exclude pattern when you configured the data store.

B. The IAM role you assigned to the crawler has the AWSGlueServiceRole managed policy attached plus an inline policy that allows read access to your S3 bucket.

C. All the classifiers returned a certainty of 0.0 right

D. You chose to create a single schema for each S3 path.

Explanation:

Answer: C

Option A is incorrect. This configuration option is used to exclude objects from the crawler. From the help text on the Add a Data Store screen in the Add Crawler console flow: "The exclude pattern is relative to the include path. Objects that match the exclude pattern are not crawled. For example, with include path **s3://mybucket/** and exclude pattern, **mydir/**.** Then all objects in the include path below the **mydir** directory are skipped. In this example, any object whose path matches **s3://mybucket/mydir/**** is not crawled. For more information about patterns, see Cataloging Tables with a Crawler"

Option B is incorrect. The IAM role assigned to your crawler needs exactly this managed policy and S3 bucket access. From the Choose an IAM Role screen on the Add Crawler console flow: "Create an IAM role named 'AWSGlueServiceRole-rolename' and attach the AWS managed policy, AWSGlueServiceRole, plus an inline policy that allows read access to: s3://yourbucketname"

Option C is correct. The data from the market data provider did not match with certainty any of the built-in classifiers that are part of Glue or your custom classifier. Therefore, Glue returned the default classification string of UNKNOWN. (See the Amazon Glue doc Adding Classifiers to a Crawler)

Option D is incorrect. This setting allows you to group compatible schemas. Choosing this option would not prevent the crawler from producing the schema. From the Configure the Crawler's Output screen in the Add Crawler console flow: "This crawler configuration groups compatible schemas into a single table definition across all S3 objects under the provided include path. Other criteria will still be considered to determine proper grouping."

Reference:

Please see the AWS developer guides AWS Glue: How It Works and AWS Glue Concepts.

Ask our Experts

View Queries

Question 16

Correct

Domain: Exploratory Data Analysis

You work for a mining company where you are responsible for the data science behind identifying the origin of mineral samples. Your data origins are Canada, Mexico, and the US. Your training data set is imbalanced as such:

Canada | Mexico | US |

1,210 | 120 | 68 |

You run a Random Forest classifier on the training data and get the following results for your test data set (your test data set is balanced):

Confusion matrix:

	Predicted			
Observed	Canada	Mexico	US	Accuracy
Canada	45	3	0	94%
Mexico	5	38	5	79%
US	19	8	21	44%

In order to address the imbalance in your training data, you will need to use a preprocessing step before you create your SageMaker training job. Which technique should you use to address the imbalance?

- A. Run your training data through a preprocessing script that uses the SMOTE (Synthetic Minority Over-sampling Technique) approach right
- B. Run your training data through a Spark pipeline in AWS Glue to one-hot encode the features
- C. Run your training data through a preprocessing script that uses the feature-split technique.
- D. Run your training data through a preprocessing script that uses the min-max normalization technique.

Explanation:

Answer: A

Option A is correct. The SMOTE sampling technique uses the k-nearest neighbors algorithm to create synthetic observations to balance a training data set. (See the article SMOTE Explained for Noobs)

Option B is incorrect because the Spark pipeline creates one-hot encoded columns in your data. One-hot encoding is a process for converting categorical data points into numeric form. This

won't do anything to address the imbalance in your training data. (See this explanation of one-hot encoding)

Option C is incorrect because it splits a feature (data point) in your observations into multiple features per observation. This also will have no impact on your imbalanced training data. (See the article Fundamental Techniques of Feature Engineering for Machine Learning)

Option D is incorrect because the min-max normalization technique is used to normalize data points into a range of 0 to 1, for example. (See the Wikipedia article Feature Scaling)

Reference:

Please see the article How to Handle Imbalanced Classification Problems in machine learning.

Ask our Experts

View Queries

Question 17

Correct

Domain: ML Implementation and Operations

You work for a scientific research company where you need to gather data on tree specimens. You have scientist peers who go out in the field across the globe and photograph tree species. The images that they gather need to be classified and labeled to use them in your training datasets in your machine learning models. What is the best way to label your image data most accurately and in the most cost-efficient manner?

A. Hire human image labelers to process all of your images and label them.

B. Use Amazon Rekognition to analyze all of your images. For the ones that the Rekognition cannot label, have human labelers that you hire attempt to label them.

C. Use an open-source labeling tool such as BBox-Label-Tool to process all of your images. For the ones that the tool cannot label, have human labelers that you hire attempt to label them.

D. Use AWS SageMaker Ground Truth to automatically label your images and use the AWS Ground Truth human labelers to label the images that the automatic labeling cannot label. right

Explanation:

Answer: D

Option A is correct. Human labelers may be able to label all of your images correctly. But they will be slow and expensive.

Option B is incorrect. While the Amazon Rekognition service analyzes image data, it does not have the human labeler to active learning model loop that trains an automatic labeling model that Amazon SageMaker Ground Truth has. Therefore, a labeling process based on Rekognition will be

more costly and less accurate than a process based on Amazon SageMaker Ground Truth. (See the Amazon Rekognition overview and the Amazon SageMaker Ground Truth overview)

Option C is incorrect. An open-source image labeling solution may label some images automatically, and a human labeling team that you hire can label the ones the open-source software cannot label. This process lacks the human labeler to active learning model loop that trains an automatic labeling model that Amazon SageMaker Ground Truth has. Therefore, a labeling process based on an open-source image labeling solution will be less accurate than a process based on Amazon SageMaker Ground Truth.

Option D is correct. As documented in the Amazon SageMaker Ground Truth overview: "Amazon SageMaker Ground Truth uses a process that starts with an active learning model that is trained from human-labeled data. Any image that it understands is automatically labeled. Ambiguous data is sent to human labelers for annotation. Then the human-labeled images are sent back to the active learning model to retrain the model to improve its accuracy incrementally. (See the Amazon SageMaker Ground Truth service overview)

Reference:

See the Amazon SageMaker Ground Truth service overview) and the Amazon Rekognition overview

Ask our Experts

View Queries

Question 18

Correct

Domain: Data Engineering

You need to use machine learning to produce real-time analysis of streaming data from IoT devices out in the field. These devices monitor oil well rigs for malfunction. Due to the safety and security nature of these IoT events, the events must be analyzed by your safety engineers in real-time. You also have an audit requirement to retain your IoT device events for 7 days since you cannot fail to process any of the events. Which approach would give you the best solution for processing your streaming data?

A. Use Amazon Kinesis Data Streams and its Kinesis Producer Library to pass your events from your producers to your Kinesis stream.

B. Use Amazon Kinesis Data Streams and its Kinesis API PutRecords call to pass your events from your producers to your Kinesis stream. right

C. Use Amazon Kinesis Data Streams and its Kinesis Client Library to pass your events from your producers to your Kinesis stream.

D. Use Amazon Kinesis Data Firehose to pass your events directly to your S3 bucket where you store your machine learning data.

Explanation:

Answer: B

Option A is incorrect. The Amazon Kinesis Data Streams Producer Library is not meant to be used for real-time processing of event data since, according to the AWS developer documentation, "it can incur an additional processing delay of up to RecordMaxBufferedTime within the library". Therefore, it is not the best solution for a real-time analytics solution. (See the AWS developer documentation titled Developing Producers Using the Amazon Kinesis Producer Library)

Option B is correct. The Amazon Kinesis Data Streams API PutRecords call is the best choice for processing in real-time since it sends its data synchronously and does not have the processing delay of the Producer Library. Therefore, it is better suited to real-time applications. (See the AWS developer documentation titled Developing Producers Using the Amazon Kinesis Data Streams API with the AWS SDK for Java)

Option C is incorrect. The Amazon Kinesis Data Streams Client Library interacts with the Kinesis Producer Library to process its event data. Therefore, you'll have the same processing delay problem with this option. (See the AWS developer documentation titled Developing Consumers Using the Kinesis Client Library 1.x)

Option D is incorrect. The Amazon Kinesis Data Firehose service directly streams your event data to your S3 bucket for use in your real-time analytics model. However, Amazon Kinesis Data Firehose retries to send your data for a maximum of 24 hours, but you have a 7-day retention requirement. (See the Amazon Kinesis Data Firehose FAQs)

Reference:

Please see the Amazon Kinesis Data Streams documentation.

Ask our Experts

View Queries

Question 19

Correct

Domain: Data Engineering

You work as a machine learning specialist for the department of defense in the NSA (National Security Agency). The NSA is responsible for security in the ports of entry around the United States. You need to process real-time video streams from airports around the country to identify questionable activity within the airport facilities and send the streaming data to SageMaker to be used as training data for your model. Your model needs to trigger an alert system when a security event is detected. What AWS services would you use to create this system most accurately and cost-effectively?

A. Use AWS Rekognition to process your video streams and send the processed data to your SageMaker model. When the model detects a security event, a lambda function is triggered to publish an SNS message to the alert system.

B. Use AWS Elastic Transcoder to process the video streams and send the processed data to your SageMaker model. When the model detects a security event, a lambda function is triggered to publish an SNS message to the alert system.

C. Use Amazon Kinesis Video Streams to stream the video to a set of processing workers running in ECS Fargate. The workers send the video data to your SageMaker machine learning model which identifies alert situations. These alerts are processed by Kinesis Data Streams which uses a lambda function to trigger the alert system. right

D. Use Amazon Kinesis Data Streams to process your video data using lambda functions which push out an SNS notification to the alert system when a security event is detected.

Explanation:

Answer: C

Option A is incorrect. The AWS Rekognition service is not meant to process streams. It works with Kinesis Video Streams to provide this capability. Also, it needs another component to send its output to your SageMaker model. This part of the solution is missing.

Option B is incorrect. The Amazon Elastic Transcoder service is used to convert video files from one format to another. It would not be useful to stream video to a processing service. (See the AWS documentation titled Amazon Elastic Transcoder)

Option C is correct. The Amazon Kinesis Video Streams service will stream your videos to a processing service that feeds your machine learning model running in SageMaker. Kinesis Streams using lambda to trigger event consumption. (See the AWS machine learning blog titled Analyze live video at scale in real-time using Amazon Kinesis Video Streams and Amazon SageMaker)

Option D is incorrect. This option lacks the machine learning component of the solution.

Reference:

Please see the Amazon Kinesis Video Streams documentation.

See a depiction of the proposed solution (in the AWS machine Learning blog titled: Analyze live video at scale in real-time using Amazon Kinesis Video Streams and Amazon SageMaker)

Ask our Experts

View Queries

Question 20

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist at a ride sharing software company. You need to analyze the streaming ride data of your firm's drivers. First, you need to clean, organize, and transform the drive

data and load it into your firm's data lake. So you can then use the data in your machine learning models in SageMaker. Which AWS services would give you the simplest solution?

- A. Use Amazon Kinesis Data Streams to capture the streaming ride data. Use Amazon Kinesis Data Analytics to clean, organize, and transform the drive data and then output the data to your S3 data lake using a Lambda function. right
- B. Use Amazon Kinesis Data Streams to capture the streaming ride data. Have Amazon Kinesis Data Streams trigger a lambda function to clean, organize, and transform the drive data and then output the data to your S3 data lake.
- C. Use Amazon Kinesis Data Streams to capture the streaming ride data. Have Kinesis Data Streams stream the data to a set of processing workers running in ECS Fargate. The workers send the data to your S3 data lake.
- D. Use Amazon Kinesis Data Firehose to stream the data directly to your S3 data lake.

Explanation:

Answer: A

Option A is correct. Amazon Kinesis Data Analytics is a very efficient service for taking streams from Amazon Kinesis Data Streams and transforming them with SQL or Apache Flink. (See the Amazon Kinesis Data Analytics overview)

Option B is incorrect. Using Lambda to retrieve your ride data from your Kinesis Data Stream and process the data records would require more effort on your part as compared to using Kinesis Data Analytics to do the transformation work.

Option C is incorrect. Using ECS Fargate as an intermediary between Amazon Kinesis Data Streams and your data lake would require you to write the transformation logic in your ECS workers. This would not be the simplest solution to the options given.

Option D is incorrect. This option lacks the transformation aspect of the solution.

Reference:

Please see the Amazon Kinesis Data Analytics documentation, and the AWS Lambda developer guide titled Tutorial: Using AWS Lambda with Amazon Kinesis (https://docs.aws.amazon.com/lambda/latest/dg/with-kinesis-example.html), and the Amazon Kinesis Data Analytics for SQL Applications Developer Guide SQL developer guide titled Using a Lambda Function as Output (https://docs.aws.amazon.com/kinesisanalytics/latest/dev/how-it-works-output-lambda.html)

Ask our Experts

View Queries

Question 21

Correct

Domain: Data Engineering

You work as a machine learning specialist at a marketing company. Your team has gathered market data about your users into an S3 bucket. You have been tasked to write an AWS Glue job to convert the files from json to a format that will be used to store Hive data. Which data format is the most efficient to convert the data for use with Hive?

A. ion

B. grokLog

C. xml

D. orc right

Explanation:

Answer: D

Option A is incorrect. Currently, AWS Glue does not support ion for output. (See the AWS developer guide documentation titled Format Options for ETL Inputs and Outputs in AWS Glue)

Option B is incorrect. Currently, AWS Glue does not support grokLog for output. (See the AWS developer guide documentation titled Format Options for ETL Inputs and Outputs in AWS Glue)

Option C is incorrect. Currently, AWS Glue does not support xml for output. (See the AWS developer guide documentation titled Format Options for ETL Inputs and Outputs in AWS Glue)

Option D is correct. From the Apache Hive Language Manual: "The *Optimized Row Columnar* (ORC) file format provides a highly efficient way to store Hive data. It was designed to overcome the limitations of the other Hive file formats. Using ORC files improves performance when Hive is reading, writing, and processing data." Also, AWS Glue supports orc for output. (See the Apache Hive Language Manual and the AWS developer guide documentation titled Format Options for ETL Inputs and Outputs in AWS Glue)

Reference:

Please see the AWS developer guide documentation titled General Information about Programming AWS Glue ETL Scripts.

Ask our Experts

View Queries

Question 22

Correct

Domain: Data Engineering

You work for a software company that has developed a popular mobile gaming app that has a large, active user base. You want to run a predictive model on real-time data generated by the app users to see how to structure an upcoming marketing campaign. The data you need for the model is the user's age, location, and level of activity in the game as measured by playing time. You need to filter the data for users who have not yet signed up for your company's premium service. You'll also need to deliver your data in json format, convert the playing time into a string format, and finally put the data onto an S3 bucket.

Which of the following is the simplest, most cost-effective, performant, and scalable way to architect this data pipeline?

A. Create a Kinesis Data Streams application running on an EC2 instance that gathers the mobile user data from its log files; use Kinesis Analytics to transform the log data into the subset you need; connect the Kinesis Data Stream to a Kinesis Firehose which puts the data onto your S3 bucket.

B. Create a Kinesis Data Streams application running on EC2 instances in an Auto Scaling Group that gathers the mobile user data from its log files; use Kinesis Analytics to transform the log data into the subset you need; connect Kinesis Data Analytics to a Kinesis Firehose which uses a lambda function to convert the playing time; Kinesis Firehose then puts the data onto your S3 bucket. right

C. Create a Kinesis Firehose which gathers the data and puts it onto your S3 bucket.

D. Create a Kinesis Data Streams application running on EC2 instances in an Auto Scaling Group that gathers the mobile user data from its log files and puts it onto your S3 bucket.

Explanation:

Answer: B

Option A is incorrect. This option has a bottleneck at the single EC2 instance used to gather the log data from the application log files. This solution would not be the most scalable.

Option B is correct. This option scales well at the Kinesis Data Streams application level because of the Auto Scaling Group. It also uses Kinesis Data Analytics to transform the data into the subset you need. It uses the Kinesis Firehose lambda option to convert the playing time to the proper format.

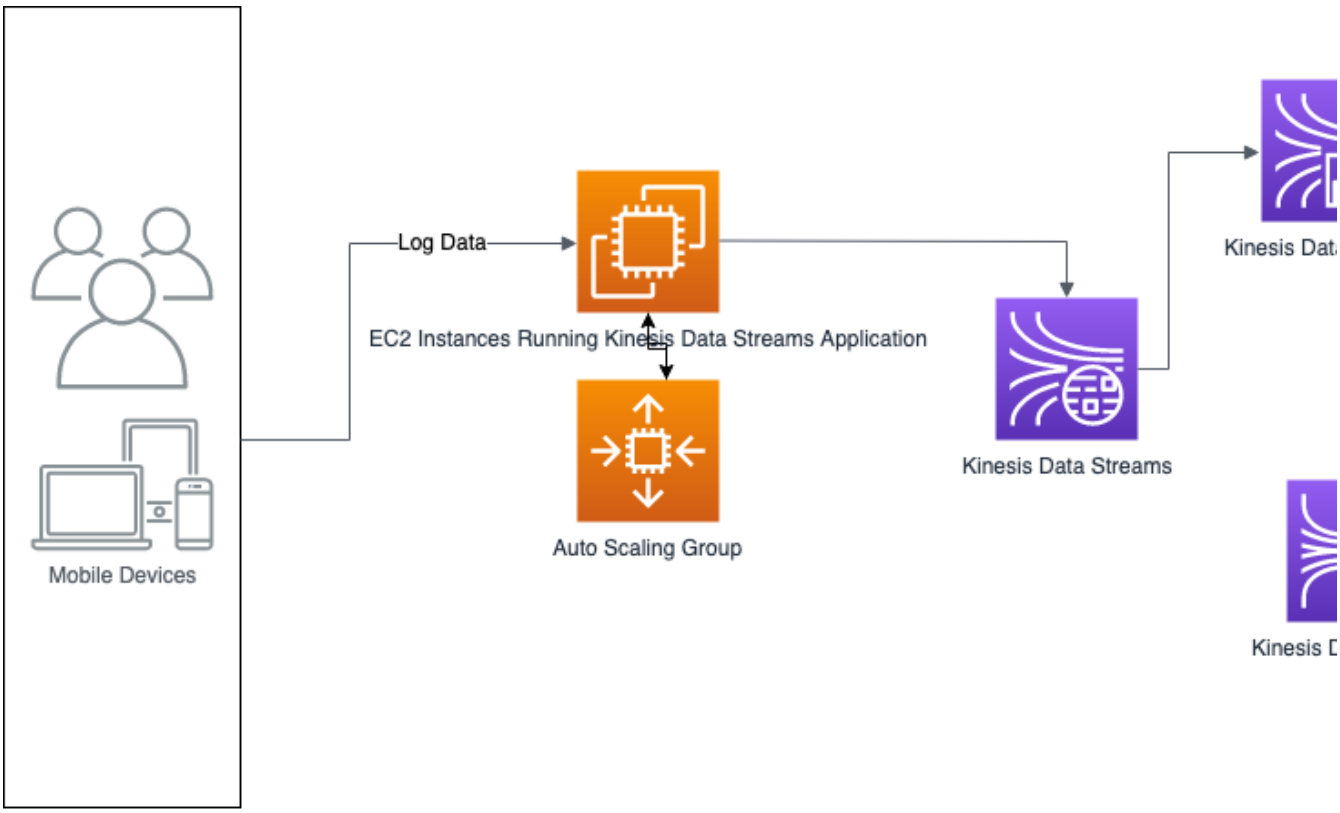
Option C is incorrect. This option does not transform the log data gathered by the Kinesis Firehose before writing the data to the S3 bucket.

Option D is incorrect. This option does not transform the log data gathered by the Kinesis Data Streams application before writing the data to the S3 bucket.

Reference:

Please see the AWS developer guide documentation titled What is Kinesis Data Streams, the AWS Auto Scaling documentation, the Amazon Kinesis Data Firehose documentation, and the Amazon Kinesis Data Analytics documentation.

Here is a diagram of the solution:



Ask our Experts

View Queries

Question 23

Correct

Domain: ML Implementation and Operations

You are deploying your data streaming pipeline for your machine learning environment. Your cloud formation stack has a Kinesis Data Firehose using the Data Transformation feature where you have configured Firehose to write to your S3 data lake. When you stream data through your Kinesis Firehose, you notice that no data is arriving your S3 bucket. What might be the problem that is causing the failure?

A. Your lambda memory setting is set to the maximum value allowed.

B. Your S3 bucket is in the same region as your Kinesis Data Firehose.

C. Your Kinesis Data Firehose buffer setting is set to the default value.

D. Your lambda timeout value is set to the default value. right

Explanation:

Answer: D

Option A is incorrect. The maximum memory setting for lambda is 3 GB. Using the maximum memory would not cause Firehose to fail to write to S3. It will increase the cost of your solution. However, since per the AWS documentation, "Lambda allocates CPU power linearly in proportion to the amount of memory configured."

Option B is incorrect. Your S3 bucket used by Kinesis Data Firehose to output your data must be in the same region as your Firehose. Since they are in the same region, this would not cause a failure to write to the S3 bucket.

Option C is incorrect. The Kinesis Data Firehose documentation states that "Kinesis Data Firehose buffers incoming data before delivering it to Amazon S3. You can choose a buffer size (1–128 MBs) or buffer interval (60–900 seconds). The condition that is satisfied first triggers data delivery to Amazon S3." Using the default setting would not prevent Firehose from writing to S3.

Option D is correct. The lambda timeout value default is 3 seconds. For many Kinesis Data Firehose implementations, 3 seconds is not enough time to execute the transformation function.

Reference:

Please see the Amazon Kinesis Data Firehose developer guide documentation titled Configure Settings, the Amazon Kinesis Data Firehose developer guide documentation titled Amazon Kinesis Data Firehose Data Transformation, and the AWS Lambda developer guide documentation titled AWS Lambda Function Configuration.

Ask our Experts

View Queries

Question 24

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist at a credit card transaction processing company. You have built a data streaming pipeline using Kinesis Data Firehose and S3. Due to the personally identifiable information contained in your data stream, your data must be encrypted in flight and at rest. How should you configure your solution to achieve encryption at rest?

A. Encrypt the data at the data consumer application level.

B. Encrypt the data by configuring Firehose to use S3-managed encryption keys (SSE-S3).

C. Encrypt the data by configuring Firehose to use S3 server-side encryption with AWS Key Management Service (SSE-KMS). right

D. Encrypt the data by configuring Firehose to use S3 server-side encryption with 256-bit AES-GCM with HKDF.

Explanation:

Answer: C

Option A is incorrect. Encrypting the data at the Kinesis consumer application level does not allow for encryption at the S3 bucket. Once the data has reached the consumer application, it has already been stored in S3 without being encrypted.

Option B is incorrect. Kinesis Data Firehose does not use SSE-S3. It uses SSE-KMS. (See the Amazon Kinesis Data Firehose developer documentation titled Configure Settings)

Option C is correct. The Kinesis Data Firehose documentation states that "Kinesis Data Firehose supports Amazon S3 server-side encryption with AWS Key Management Service (AWS KMS) for encrypting delivered data in Amazon S3. You can choose not to encrypt the data or encrypt with a key from the list of AWS KMS keys that you own. For more information, see Protecting Data Using Server-Side Encryption with AWS KMS–Managed Keys (SSE-KMS)."

Option D is incorrect. Kinesis Data Firehose does not use 256-bit AES-GCM with HKDF. It uses SSE-KMS. (See the Amazon Kinesis Data Firehose developer documentation titled Configure Settings)

Reference:

Please see the Amazon Kinesis Data Firehose developer guide documentation titled Creating an Amazon Kinesis Data Firehose Delivery Stream and the Amazon Kinesis Data Streams developer guide documentation titled What is Server-Side Encryption for Kinesis Data Streams.

Ask our Experts

View Queries

Question 25

Correct

Domain: ML Implementation and Operations

You are working as a machine learning specialist at a medical research facility. You have set up a data pipeline delivery stream using Amazon Kinesis Data Firehose as your data streaming service and Amazon Redshift as your data warehouse. Your researchers have set up the S3 bucket in their own account that you have used for your Kinesis Data Firehose. Your researchers need to access the data using BI tools such as Amazon QuickSight to build dashboards and use metrics in their research. However, when you implement your solution, you notice that your streaming data does not load into your Redshift data warehouse. What could be a reason why this is happening? Choose 2 answers.

- [x] **A. You have not created an IAM role for your Kinesis Firehose to access the S3 bucket.** right
- [ ] **B. You defined a cluster security group and associated it with your Redshift cluster.**

C. The access policy associated with your Kinesis Firehose does not have lambda:InvokeFunction specified in the Allow Action section of the Lambda actions.

D. The access policy associated with your Kinesis Firehose does not have kms:GenerateDataKey specified in the Allow Action section of the KMS actions.

- [x] E. The access policy associated with your Kinesis Firehose does not have S3:PutObjectAcl specified in the Allow Action section of the S3 actions. right

Explanation:

Answers: A and E

Option A is correct. As documented in the Amazon Kinesis Data Firehose developer guide, "Kinesis Data Firehose uses the specified Amazon Redshift user name and password to access your cluster and uses an IAM role to access the specified bucket, key, CloudWatch log group, and streams. You are required to have an IAM role when creating a delivery stream."

Option B is incorrect. The cluster security group is used to grant users inbound access to the Redshift cluster. Defining a cluster security group would not prevent Kinesis Firehose from accessing your Redshift cluster. (See the Amazon Redshift database developer guide titled Amazon Redshift Security Overview)

Option C is incorrect. Since you are not using the Lambda function feature of Kinesis Data Firehose, this Lambda action is not needed in the access policy.

Option D is incorrect. Since you are not using the data encryption feature of Kinesis Data Firehose, this KMS action is not needed in the access policy.

Option E is correct. Since you are not the owner of the S3 bucket used by Kinesis Data Firehose, you need to specify the S3:PutObjectAcl in the S3 actions of the access policy. (See the Amazon Kinesis Data Firehose developers guide titled Grant Kinesis Data Firehose Access to Amazon Redshift Destination)

Reference:

Please see the Amazon Kinesis Data Firehose developers guide titled Grant Kinesis Data Firehose Access to Amazon Redshift Destination, and the Amazon Kinesis Data Firehose overview page, and the Amazon Redshift database developer guide titled Amazon Redshift Security Overview.

Ask our Experts

View Queries

Question 26

Correct

Domain: Data Engineering

You work as a machine learning specialist at a hedge fund. You are working on a time-series price prediction model for the firm, and you have set up a data delivery stream using Amazon Kinesis Data Streams. You are creating the data producer application code to take trade data from your trade system and send the trade records to your Kinesis Data Stream. Your python code is structured as follows:

```
import boto3

import requests

import json

client =  boto3.client('kinesis', region_name='us-east1')

while True:

    r = requests.get('https://trading-applicatio-url')

    data = json.dumps(r.json())

    client.put_record(
        parameters needed for put_record api call

    )

    ...
```

Which of the following options are valid put_record request parameters? Select 3.

- [x] A. Data right
- [ ] B. ImplicitHashKey
- [x] C. ExplicitHashKey right
- [ ] D. PartitionKeys
- [x] E. SequenceNumberForOrdering right
- [ ] F. ShardId

Explanation:

Answers: A, C, and E

Options A, C and E are correct. As documented in the Amazon Kinesis Data Streams API reference guide titled PutRecord, "The request accepts the following data in JSON format: Data, ExplicitHashKey, PartitionKey, SequenceNumberForOrdering, and StreamName".

Option B is incorrect. There is no ImplicitHashKey request parameter. (See the Amazon Kinesis Data Streams API reference guide titled PutRecord)

Option D is incorrect. There is no PartitionKeys request parameter. However, there is a PartitionKey request parameter. (See the Amazon Kinesis Data Streams API reference guide titled PutRecord)

Option F is incorrect. There is no ShardId request parameter. However, there is a ShardId response element. (See the Amazon Kinesis Data Streams API reference guide titled PutRecord)

Reference:

Please see the Amazon Kinesis Data Streams developers guide titled Kinesis Data Stream Producers and the Amazon Kinesis Data Streams API reference guide titled PutRecord.

Ask our Experts

View Queries

Question 27

Correct Marked for review

Domain: Exploratory Data Analysis

You work as a machine learning specialist at a firm that runs a web application that allows users to research and compare real estate properties worldwide. You are working on a property foreclosure model to predict potential price drops. You have decided to use the SageMaker Linear Learner algorithm. Here is a small sample of the data you'll have to work with:

Type	Bedrooms	Area	Solar_Rating	Price	Foreclosed
condo	2	2549	H	125400	N
house	4	4124	M	250250	Y
house	3	3250		200000	N
condo	1	900	N	90250	N
condo	2	?	L	125400	Y

In order to feed this data into your model, you will first need to clean and format your data.

Which of the following SageMaker built-in scikit-learn library transformers would you use to clean and format your data? Select 4.

A. StandardScaler to encode the Solar_Rating feature

B. OneHotEncoder to encode the Area feature

- [x] C. SimpleImputer to complete the missing values in the Solar_Rating and Area features right
- [x] D. OneHotEncoder to encode the Type feature right

E. OrdinalEncoder to complete the missing values in the Solar_Rating and Area features

- [x] F. OrdinalEncoder to encode the Solar_Rating feature right
- [x] G. LabelBinarizer to encode the Foreclosed feature right
- [ ] H. MinMaxScaler to encode the Foreclosed feature

Explanation:

Answers: C, D, F, and G

Options A is incorrect. From the scikit-learn API Reference, the StandardScaler transformer is used to Standardize features by removing the mean and scaling to unit variance. The OrdinalEncoder transformer would be the better choice for this feature since H > M > L > N. Therefore, this feature has ordinal values.

Option B is incorrect. The OneHotEncoder transforms nominal categorical features and creates new binary columns for each observation. The Area feature holds numerical or quantitative data, which does not need to be transformed.

Option C is correct. The Solar_Rating and Area features have missing data in some observations. From the scikit-learn API Reference: the SimpleImputer transformer is used to complete missing values.

Option D is correct. The Type feature is a good candidate for the OneHotEncoder transformer since the Type feature holds a limited number of categorical types. The OneHotEncoder transforms nominal categorical features and creates new binary columns for each observation.

Option E is incorrect. From the scikit-learn API Reference: the OrdinalEncoder transformer encodes categorical features as an integer array. This encoder does not complete missing values.

Option F is correct. From the scikit-learn API Reference: the OrdinalEncoder transformer encodes categorical features as an integer array that maintains the ordinal nature of the data. Since H > M > L > N, this feature has ordinal values.

Option G is correct. The Foreclosed feature holds one of two choices, either a 'Y' or 'N'. Therefore, this feature is a good candidate for the LabelBinarizer. From the scikit-learn API Reference: the LabelBinarizer transformer binarizes label in a one-versus-all fashion.

Option H is incorrect. From the scikit-learn API Reference: the MinMaxScaler transformer transforms features by scaling each feature to a given range. The Foreclosed feature has binary data: either 'Y' or 'N'. So it is better suited to the LabelBinarizer transformer.

Reference:

Please see the Amazon SageMaker developer guide titled Use Scikit-learn with Amazon SageMaker, and the scikit-learn API Reference.

Ask our Experts

View Queries

Question 28

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist for a polling company. For the upcoming election, you need to classify the over 500,000 registered voters in your voter database by age for a campaign your team is about to launch. Your data is structured as such:

| voter_id | voter_age | voter_occupation | voter_income | ...

| 1 | 21 | student | 0 | ...

| 2 | 35 | nurse | 25000 | ...

| 3 | 49 | manager | 150000 | ...

| 4 | 63 | truck driver | 45000 | ...

| 5 | 55 | teacher | 65000 | ...

...

Because you have continuous data for your voter age feature, classifying your observations by age would result in too many classifications, i.e., one for every possible voter age from 21 though probably over 90. You need to have uniform classifications that are limited in number to make the best use of your data in your machine learning model.

What numerical feature engineering technique will give you the best distribution of classifications?

A. Cartesian Product Transformation

B. N-Gram Transformation

C. Orthogonal Sparse Bigram (OSB) Transformation

D. Normalization Transformation

E. Quantile Binning Transformation right

Explanation:

Answer: E

Options A is incorrect. From the Amazon Machine Learning developer guide titled Data Transformations Reference, "The Cartesian product transformation takes categorical variables or text as input, and produces new features that capture the interaction between these input variables." Because this transformation is for transforming text, it would not give you uniform age classifications that are limited in number.

Option B is incorrect. From the Amazon Machine Learning developer guide titled Data Transformations Reference, "The n-gram transformation takes a text variable as input and

produces strings corresponding to sliding a window of (user-configurable) n words, generating outputs in the process." Because this transformation is also for transforming text, it would not give you uniform age classifications that are limited in number.

Option C is incorrect. From the Amazon Machine Learning developer guide titled Data Transformations Reference, "The OSB transformation is intended to aid in text string analysis and is an alternative to the bi-gram transformation (n-gram with window size 2). OSBs are generated by sliding the window of size n over the text and outputting every pair of words that includes the first word in the window." Because this transformation is also for transforming text, it would not give you uniform age classifications that are limited in number.

Option D is incorrect. From the Amazon Machine Learning developer guide titled Data Transformations Reference, "The normalization transformer normalizes numeric variables to have a mean of zero and variance of one. Normalization of numeric variables can help the learning process if there are very large range differences between numeric variables because variables with the highest magnitude could dominate the ML model, no matter if the feature is informative with respect to the target or not." Because this transformation is for normalizing continuous data, it would not give you uniform age classifications that are limited in number.

Option E is correct. From the Amazon Machine Learning developer guide titled Data Transformations Reference, "The quantile binning processor takes two inputs, a numerical variable and a parameter called *bin number*, and outputs a categorical variable. The purpose is to discover non-linearity in the variable's distribution by grouping observed values together." Because Quantile binning is used to create uniform bins of classifications, it would be the right choice to give you uniform age classifications that are limited in number. For example, you could create classification bins such as: Under 30, 30 to 50, Over 50. Or even better: Millennial, Generation X, Baby Boomer, etc.

Reference:

Please see the Amazon Machine Learning developer guide titled Data Transformations for Machine Learning and the article Feature Engineering in Machine Learning (Part 1) Handling Numeric Data with Binning.

Ask our Experts

View Queries

Question 29

Incorrect

Domain: Exploratory Data Analysis

You work as a machine learning specialist for a consulting firm where you analyze data about the consultants who work there in preparation for using the data in your machine learning models. The features you have in your data are things like employee id, specialty, practice, job description, billing hours, and principle. The principle attribute is represented as 'yes' or 'no', whether the consultant has

made principle level or not. For your initial analysis, you need to identify the distribution of consultants and their billing hours for the given period. What visualization best describes this relationship?

A. Scatter plot

B. Histogram right

C. Line chart

D. Box plot

E. Bubble chart wrong

Explanation:

Answer: B

Options A is incorrect. You are looking for distribution on a single dimension: the consultants billing hours. From the Amazon QuickSite User Guide titled Working with Visual Types in Amazon QuickSight, "A scatter chart shows multiple distributions, i.e., two or three measures for a dimension."

Option B is correct. You are looking for a distribution of a single dimension: the consultants billing hours. From the Wikipedia article titled Histogram, "A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable." The continuous variable in this question: the billing hours, binned into ranges (x-axis), at a frequency: the number of consultants at a billing hour range (y-axis).

Option C is incorrect. From the Amazon QuickSite User Guide titled Working with Visual Types in Amazon QuickSight, "Use line charts to compare changes in measured values over a period of time." You are looking for distribution, not a comparison of changes over a period of time.

Option D is incorrect. From the Statistics How To article titled Types of Graphs Used in Math and Statistics, "A boxplot, also called a box and whisker plot, is a way to show the spread and centers of a data set. Measures of spread include the interquartile range and the mean of the data set. Measures of the center include the mean or average and median (the middle of a data set)." A Box Plot shows the distribution of multiple dimensions of the data. Once again, you are looking for a distribution of a single dimension, not a distribution on multiple dimensions.

Option E is incorrect. From the Wikipedia article titled Bubble Chart, "A bubble chart is a type of chart that displays three dimensions of data. Each entity with its triplet ($v1$, $v2$, $v3$) of associated data is plotted as a disk that expresses two of the vi values through the disk's xy location and the third through its size." Once again, you are looking for a distribution of a single dimension, not a distribution on three dimensions.

Reference:

Please see the Amazon QuickSight user guide titled Working with Amazon QuickSight Visuals and the Statistics How To article titled Types of Graphs Used in Math and Statistics.

Ask our Experts

View Queries

Question 30

Correct

Domain: Modeling

You work as a machine learning specialist for a robotics manufacturer where you are attempting to use unsupervised learning to train your robots to perform their prescribed tasks. You have engineered your data and produced a CSV file and placed it on S3.

Which of the following input data channel specifications are correct for your data?

A. Metadata Content-Type is identified as text/csv

B. Metadata Content-Type is identified as application/x-recordio-protobuf;boundary=1

C. Metadata Content-Type is identified as application/x-recordio-protobuf;label_size=1

D. Metadata Content-Type is identified as text/csv;label_size=0 right

Explanation:

Answer: D

Option A is incorrect. The Content-Type of text/csv without specifying a label_size is used when you have target data, usually in column one, since the default value for label_size is 1, meaning you have one target column. (See the Amazon SageMaker developer guide titled Common Data Formats for Training)

Option B is incorrect. The boundary content type is not relevant to CSV files. It is used for multipart form data.

Option C is incorrect. For unsupervised learning, the label_size should equal 0, indicating the absence of a target. (See the Amazon SageMaker developer guide titled Common Data Formats for Training)

Option D is correct. For unsupervised learning, the label_size equals 0, indicating the absence of a target. (See the Amazon SageMaker developer guide titled Common Data Formats for Training)

Reference:

Please see the Amazon SageMaker developer guide, specifically Common Data Formats for Built-in Algorithms and Common Data Formats for Training.

Ask our Experts

View Queries

Question 31

Correct

Domain: Modeling

You work as a machine learning specialist for a manufacturing plant where you are attempting to use supervised learning to train assembly line image recognition to categorize malformed parts. You have engineered your data and produced a CSV file and placed it on S3.

Which of the following input data channel specifications are correct for your data? (Select TWO)

- [x] A. Metadata Content-Type is identified as text/csv right
- [ ] B. Metadata Content-Type is identified as text/csv;label_size=0
- [x] C. Target value should be in the first column with no header right
- [ ] D. Target value should be in the last column with no header
- [ ] E. Target value should be in the last column with a header
- [ ] F. Target value should be in the first column with a header

Explanation:

Answers: A and C

Option A is correct. The Content-Type of text/csv without specifying a label_size is used when you have target data, usually in column one, since the default value for label_size is 1, meaning you have one target column. (See the Amazon SageMaker developer guide titled Common Data Formats for Training)

Option B is incorrect. The Content-Type of text/csv specifying a label_size of 0 is used when you do not have target data. You usually choose this setting when using unsupervised learning. (See the Amazon SageMaker developer guide titled Common Data Formats for Training)

Option C is correct. From the Amazon SageMaker developer guide titled Common Data Formats for Training, "Amazon SageMaker requires that a CSV file doesn't have a header record and that the target variable is in the first column".

Option D is incorrect. From the Amazon SageMaker developer guide titled Common Data Formats for Training, "Amazon SageMaker requires that a CSV file doesn't have a header record and that the target variable is in the first column".

Option E is incorrect. From the Amazon SageMaker developer guide titled Common Data Formats for Training, "Amazon SageMaker requires that a CSV file doesn't have a header record and that the target variable is in the first column".

Option F is incorrect. From the Amazon SageMaker developer guide titled Common Data Formats for Training, "Amazon SageMaker requires that a CSV file doesn't have a header record and that the target variable is in the first column".

Reference:

Please see the Amazon SageMaker developer guide, specifically Common Data Formats for Built-in Algorithms and Common Data Formats for Training.

Ask our Experts

View Queries

Question 32 Correct

Domain: Modeling

You work as a machine learning specialist for a marketing firm. Your firm wishes to determine which customers in a dataset of its registered users will respond to a new proposed marketing campaign. You plan to use the XGBoost algorithm on the binary classification problem. In order to find the optimal model, you plan to run many hyperparameter tuning jobs to reach the best hyperparameter values. Which of the following hyperparameters must you use in your tuning jobs if your objective is set to multi:softprob? (Select TWO)

A. alpha

B. base_score

C. eta

☐ D. num_round right

E. gamma

☐ F. num_class right

Explanation:

Answers: D and F

Option A is incorrect. The alpha hyperparameter is used to adjust the L1 regulation term on weights. This term is optional. (See the Amazon SageMaker developer guide titled XGBoost Hyperparameters)

Option B is incorrect. The base_score hyperparameter is used to set the initial prediction score of all instances. This term is optional. (See the Amazon SageMaker developer guide titled XGBoost Hyperparameters)

Option C is incorrect. The eta hyperparameter is used to prevent overfitting. This term is optional. (See the Amazon SageMaker developer guide titled XGBoost Hyperparameters)

Option D is correct. The num_round hyperparameter is used to set the number of rounds to run in your hyperparameter tuning jobs. This term is required. (See the Amazon SageMaker developer

guide titled XGBoost Hyperparameters)

Option E is incorrect. The gamma hyperparameter is used to set the minimum loss reduction required to make a further partition on a leaf node of the tree. This term is optional. (See the Amazon SageMaker developer guide titled XGBoost Hyperparameters)

Option F is correct. This hyperparameter is used to set the number of classes. This term is required if the objective is set to multi:softmax or multi:softprob. (See the Amazon SageMaker developer guide titled XGBoost Hyperparameters)

Reference:

Please see the Amazon SageMaker developer guide titled Automatic Model Tuning and the Amazon SageMaker developer guide titled How Hyperparameter Tuning Works.

Ask our Experts

View Queries

Question 33

Correct

Domain: Modeling

You work as a machine learning specialist for a healthcare insurance company. Your company wishes to determine which registered plan participants will choose a new health care option your company plans to release. The roll-out plan for the new option is compressed, so you need to produce results quickly. You plan to use a binary classification algorithm on this problem. In order to find the optimal model quickly, you plan to run the maximum number of concurrent hyperparameter training jobs to reach the best hyperparameter values. Which of the following types of hyperparameters tuning techniques will best suit your needs?

A. Bayesian Search

B. Hidden Markov Models

C. Conditional Random Fields

D. Random Search right

Explanation:

Answer: D

Option A is incorrect. Bayesian Search uses regression to choose sets of hyperparameters to test iteratively. Due to this iterative approach, this method cannot run the maximum number of concurrent training jobs without impacting the performance of the search. Therefore, this method will take longer than the Random Search method.

Option B is incorrect. The Hidden Markov Model is a class of probabilistic graphical model. It is not used by SageMaker for hyperparameter tuning.

Option C is incorrect. Conditional Random Fields is a type of discriminative classifier. It is not used by SageMaker for hyperparameter tuning.

Option D is correct. The Random Search technique allows you to run the maximum number of concurrent training jobs without impacting the performance of the search. Therefore, getting you to your optimized hyperparameters quickly.

Reference:

Please see the Amazon SageMaker developer guide titled How Hyperparameter Tuning Works.

Ask our Experts

View Queries

Question 34

Incorrect

Domain: Modeling

You work as a machine learning specialist for a financial services company. You are building a machine learning model to perform futures price prediction. You have trained your model, and you now want to evaluate it to make sure it is not overtrained and can generalize.

Which of the following techniques is the appropriate method to cross-validate your machine learning model?

A. Leave One Out Cross Validation (LOOCV)

B. K-Fold Cross Validation wrong

C. Stratified Cross Validation

D. Time Series Cross Validation right

Explanation:

Answer: D

Option A is incorrect. Since we are trying to validate a time series set of data, we need to use a method that uses a rolling origin with day n as training data and day n+1 as test data. The LOOCV approach doesn't give us this option. (See the article K-Fold and Other Cross-Validation Techniques)

Option B is incorrect. The K-Fold cross validation technique randomizes the test dataset. We cannot randomize our test dataset since we try to validate a time series set of data. Randomized time series data loses its time-related value.

Option C is incorrect. We are trying to cross-validate time series data. We cannot randomize the test data because it will lose its time-related value.

Option D is correct. The Time Series Cross Validation technique is the correct choice for cross-validating a time series dataset. Time series cross validation uses forward chaining, where the origin of the forecast moves forward in time. Day n is training data and day n+1 is test data.

Reference:

Please see the Amazon Machine Learning developer guide titled Cross Validation, and the article K-Fold and Other Cross-Validation Techniques.

Ask our Experts

View Queries

Question 35

Correct

Domain: Modeling

You work as a machine learning specialist for a bank. Your bank management team is concerned about a recent increase in fraudulent transactions. You need to build a machine learning model to recognize fraudulent transactions in real-time. The following is a sample of the dataset you are using to train your model:

| Transaction ID | Account ID | type | amount | source | ... | fraud |

| 12576477 | 37564378 | debit | 350.00 | ATM | ... | not_fraud |

| 39844569 | 74897544 | credit | 756.23 | ATM | ... | not_fraud |

| 54986984 | 55656753 | credit | 243.90 | ATM | ... | undetermined |

| 34567863 | 27564378 | debit | 1250.00 | ATM | ... | fraud |

You are using the fraudulent feature as your label. You have decided to use the linear learner built-in algorithm for your model. Which predictor type should you use for your linear learner?

A. binary_classifier

B. regressor

C. cross_entropy_loss

D. multiclass_classifier right

Explanation:

Answer: D

Option A is incorrect. The binary_classifier predictor type is used when your target feature is binary, either yes or no, 1 or 0, etc. Your target feature has three potential values. Therefore it is classified across multiple classes.

Option B is incorrect. The regressor predictor type is used for regression models where your target feature is a continuous numeric value. Your target feature has three potential values. Therefore it is classified across multiple classes.

Option C is incorrect. The cross_entropy_loss is an option for the binary_classifier_model_selection_criteria parameter of the linear learner SageMaker algorithm. This parameter is used when you are using a binary classifier as your predictor type.

Option D is correct. The multiclass_classifier predictor type is used when your target feature has more than two potential values. Your target feature has three potential values. Therefore it is classified across multiple classes.

Reference:

Please see the Amazon SageMaker developer guide titled Linear Learner Algorithm, the Amazon SageMaker read the docs API doc titled LinearLearner, and the Amazon Machine Learning developer guide titled Multiclass Classification.

Ask our Experts

View Queries

Question 36

Correct

Domain: Modeling

You work as a machine learning specialist for a robotics product manufacturer. Your company is trying to use machine learning to help its automatic vacuuming robot determine the most efficient path across the floor of a room. You need to build a machine learning model to accomplish this problem.

Which modeling approach best fits your problem?

A. Multi-Class Classification

B. Simulation-based Reinforcement Learning right

C. Binary Classification

D. Heuristic Approach

Explanation:

Answer: B

Option A is incorrect. Multi-Class Classification is used when your model needs to has many class outcomes from which to choose, as in a car model classification image recognition problem. In this strategy determination problem, we need to learn a strategy that optimizes an objective. A Multi-Class Classification approach wouldn't give you this result.

Option B is correct. Simulation-Based Reinforcement Learning is used in problems where your model needs to learn through trial and error. This is how a robot would best learn the optimal path through a given environment.

Option C is incorrect. Binary Classification is the approach you use when you are trying to predict a binary outcome. This strategy determination problem would not fit a binary classification model.

Option D is incorrect. The Heuristic Approach is used when a machine learning approach is not necessary. An example is the rate of acceleration of a particle through space. There are well known formulas for speed, inertia, and friction that can solve a problem such as this.

Reference:

Please see the Amazon SageMaker developer guide titled Linear Learner Algorithm, the Amazon SageMaker developer guide titled Reinforcement Learningwith Amazon SageMaker RL, the Amazon Machine Learning developer guide titled Multiclass Classification, and the article titled What is the difference between a machine learning algorithm and a heuristic, and when to use each?

Ask our Experts

View Queries

Question 37

Correct

Domain: Modeling

You work as a machine learning specialist for a state highway administration department. Your department is trying to use machine learning to help determine the make and model of cars as they pass a camera on the state highways. You need to build a machine learning model to accomplish this problem.

Which modeling approach best fits your problem?

- A. Multi-Class Classification right
- B. Simulation-based Reinforcement Learning
- C. Binary Classification
- D. Heuristic Approach

Explanation:

Answer: A

Option A is correct. Multi-Class Classification is used when your model needs to choose from a finite set of outcomes, such as this car make and model classification image recognition problem.

Option B is incorrect. Simulation-Based Reinforcement Learning is used in problems where your model needs to learn through trial and error. An image recognition problem with a finite set of outcomes is better suited to a multi-class classification model.

Option C is incorrect. Binary Classification is the approach you use when you are trying to predict a binary outcome. This strategy determination problem would not fit a binary classification model since you have a finite set from which to choose that is greater than 2.

Option D is incorrect. The Heuristic Approach is used when a machine learning approach is not necessary. An example is the rate of acceleration of a particle through space. There are well known formulas for speed, inertia, and friction that can solve a problem such as this.

Reference:

Please see the Amazon SageMaker developer guide titled Linear Learner Algorithm, the Amazon SageMaker developer guide titled Reinforcement Learningwith Amazon SageMaker RL, the Amazon Machine Learning developer guide titled Multiclass Classification, and the article titled What is the difference between a machine learning algorithm and a heuristic, and when to use each?

Ask our Experts

View Queries

Question 38

Correct

Domain: Modeling

You work as a machine learning specialist for a retail pet products chain. Your company is trying to use machine learning to help determine the breed of dogs in the photos your customers tag on Instagram and Twitter. You need to build a machine learning model to accomplish this problem.

Which SageMaker model would you use to fit your machine learning problem best?

A. K-Means

B. Image Classification right

C. Sequence-to-Sequence

D. Neural Topic Model

Explanation:

Answer: B

Option A is incorrect. K-Means is used to find discrete groupings in data. It is mostly used on numeric data that is continuous. Image data is not numeric and is not continuous, so K-Means would not be a good model for your dog image classification problem. (See the Amazon SageMaker developer guide titled K-Means Algorithm)

Option B is correct. The Image Classification model is used to solve classification problems such as image classification. (See the Amazon SageMaker developer guide titled Image Classification Algorithm)

Option C is incorrect. The Sequence-to-Sequence model is used to take a sequence of tokens and produces another sequence of tokens. It is used for problems like language translation, text summarization, and speech-to-text. (See the Amazon SageMaker developer guide titled Sequence-to-Sequence Algorithm)

Option D is incorrect. The Neural Topic Model algorithm is used to organize documents into topics. This type of model is not suited to image classification. (See the Amazon SageMaker developer guide titled Neural Topic Model (NTM) Algorithm)

Reference:

Please see the Amazon SageMaker developer guide titled Use Amazon SageMaker Built-in Algorithms.

Ask our Experts

View Queries

Question 39

Correct

Domain: ML Implementation and Operations

You are building a machine learning model for your user behavior prediction problem using your company's user interaction data stored in DynamoDB. You want to get your data into CSV format and load it into an S3 bucket so you can use it for your machine learning algorithm to give personalized recommendations to your users. Your data set needs to be updated automatically to produce real-time recommendations. Your business analysts also want to have the ability to run ad hoc queries on your data.

Which of the following architectures will be the most efficient way to achieve this?

A. Use AWS Data Pipeline to coordinate the following set of tasks: export DynamoDB data to S3 as JSON; Convert JSON to CSV; SageMaker model uses the data to produce real-time predictions; analysts use Amazon Athena to perform ad hoc queries against the CSV data in S3. right

B. Create a custom classifier in an AWS Glue ETL job that extracts the DynamoDB data to CSV format on your S3 bucket; run your SageMaker model on the new data to produce real-time

recommendations; analysts use Amazon Athena to perform ad hoc queries against the CSV data in S3.

C. Use AWS DMS to connect to your DynamoDB database and export the data to S3 in CSV format; run your SageMaker model on the new data to produce real-time recommendations; analysts use Amazon Athena to perform ad hoc queries against the CSV data in S3.

D. Use Kinesis Data Streams to receive the data from DynamoDB; use an ETL job running on an EC2 instance to consume the data and produce the CSV representation; run your SageMaker model on the new data to produce real-time recommendations; analysts use Amazon Athena to perform ad hoc queries against the CSV data in S3.

Explanation:

Answer: A

Option A is correct. AWS Data Pipeline is used here to schedule frequent runs of the described workflow: DynamoDB export, transformation, and running the model to give real-time predictions.

Option B is incorrect. This approach lacks the pipeline coordination described in Option A.

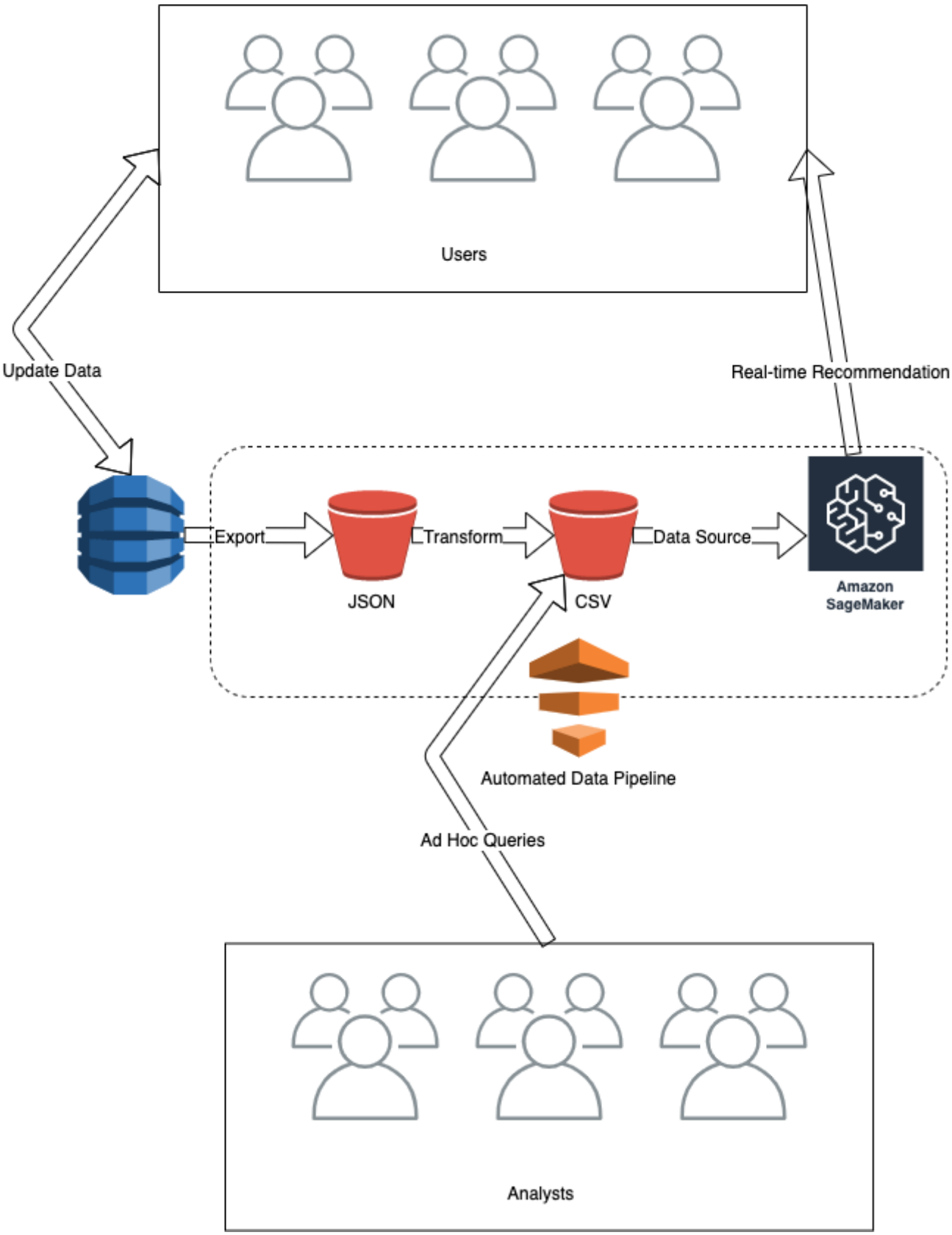
Option C is incorrect. AWS DMS does not support DynamoDB as a data source. Also, this approach lacks the pipeline coordination described in Option A.

Option D is incorrect. You would have to write more code to make this option work when compared to option A. You would need to write an extraction job to make the DynamoDB data into a Kinesis producer. You would also have to write the consumer ETL job. Also, this approach lacks the pipeline coordination described in Option A.

Reference:

Please see the AWS Data Pipeline developer guide titled What is AWS Data Pipeline, and AWS Database Migration Service user guide titled How AWS Data Migration Service Works specifically the section on sources, the Amazon Kinesis Data Streams developer guide titled Amazon Kinesis Data Streams Terminology and Concepts.

Here is a diagram of the best option:



Ask our Experts

View Queries

Question 40 Correct

Domain: ML Implementation and Operations

You are building a machine learning model to use your web server logs to predict which users are most likely to buy a given product. Using your company's unstructured web server log data stored in S3, you want to get your data into CSV format and load it into another S3 bucket so that you can use it for your machine learning algorithm.

Which of the following architectures will be the most efficient way to achieve this?

A. Load the log data into a Redshift cluster; use the UNLOAD Redshift command with a select statement to unload the data in CSV format to S3; SageMaker model uses the data to produce product purchase predictions.

B. Use a built-in classifier in an AWS Glue crawler that crawls the web server logs and outputs the log data to CSV format on your ML S3 bucket; SageMaker model uses the data to produce product purchase predictions. right

C. Use AWS Schema Conversion tool to convert your web log data to CSV format and output it to your ML S3 bucket; run your SageMaker model on the new data to produce product purchase predictions.

D. Use AWS Snowball Edge and its lambda function capability to convert and then move the web log to S3 in CSV format; run your SageMaker model on the new data to produce product purchase predictions.

Explanation:

Answer: B

Option A is incorrect. Using Redshift as an intermediary step in this architecture is an expensive, in terms of implementation effort, an extraneous design decision that makes this option less efficient than Option B.

Option B is correct. AWS Glue has built-in classifiers designed specifically for web server log crawling. The crawler will generate CSV formatted data and output it to your ML S3 bucket. This option is the simplest to implement, and therefore the most efficient.

Option C is incorrect. The AWS Schema Conversion tool is used to convert a database from one database engine to another database engine, such as from PostgreSQL to MySQL. The AWS Schema Conversion tool will not work with unstructured web log data.

Option D is incorrect. AWS Snowball Edge is used to move data into and out of AWS. It would not be the most efficient way to transform your web log data to CSV and store it in your ML S3 bucket.

Reference:

Please see the Amazon Redshift Database developer guide titled Unloading Data, and Amazon Machine Learning developer guide titled Creating an Amazon ML Datasource from Data in Amazon Redshift, the AWS Schema Conversion Tool user guide titled What is the AWS Schema

Conversion Tool?, and the Cloud Data Migration Guide, specifically the section on AWS Snowball Edge, and the AWS Glue developer guide titled Adding Classifiers to a Crawler.

Ask our Experts

View Queries

Question 41

Incorrect

Domain: Exploratory Data Analysis

You work for a language translation software company. Your company needs to move from traditional translation software to a machine learning model-based approach that produces the translations accurately. One of your first tasks is to take text given in the form of a document and use a histogram to measure the occurrence of individual words in the document for use in document classification.

Which of the following text feature engineering techniques is the best solution for this task?

A. Orthogonal Sparse Bigram (OSB)

B. Term Frequency-Inverse Document Frequency (tf-idf) wrong

C. Bag-of-Words right

D. N-Gram

Explanation:

Answer: C

Option A is incorrect. The Orthogonal Sparse Bigram natural language processing algorithm creates groups of words and outputs the pairs of words that include the first word. You are trying to measure the occurrence of individual words.

Option B is incorrect. Term Frequency-Inverse Document Frequency determines how important a word is in a document by giving weights to words that are common and less common in the document. You are not trying to determine the importance of the words in your document, just the count of the individual words.

Option C is correct. The Bag-of-Words natural language processing algorithm creates tokens of the input document text and outputs a statistical depiction of the text. The statistical depiction, such as a histogram, shows the count of each word in the document.

Option D is incorrect. The N-Gram natural language processing algorithm is used to find multi-word phrases in the text of a document. You are not trying to find multi-word phrases. You are just trying to find the count of the individual words.

Reference:

Please see the article titled Introduction to Natural Language Processing for Text.

Ask our Experts

View Queries

Question 42

Incorrect

Domain: Exploratory Data Analysis

You work for a marketing firm that wants to analyze Twitter user stream data to find popular subjects among users who buy products produced by the firm's clients. You need to analyze the streamed text to find important or relevant repeated common words and phrases and correlate this data to client products. You'll then include these topics in your client product marketing material.

Which of the following text feature engineering techniques is the best solution for this task?

A. Orthogonal Sparse Bigram (OSB)

B. Term Frequency-Inverse Document Frequency (tf-idf) right

C. Bag-of-Words

D. N-Gram wrong

Explanation:

Answer: B

Option A is incorrect. The Orthogonal Sparse Bigram natural language processing algorithm creates groups of words and outputs the pairs of words that include the first word. You are trying to determine how important a word is in a document by finding relevant repeated common words.

Option B is correct. Term Frequency-Inverse Document Frequency determines how important a word is in a document by giving weights to words that are common and less common in the document. You can use this information to select the most important repeated phrases in the user's tweets in your client marketing material.

Option C is incorrect. The Bag-of-Words natural language processing algorithm creates tokens of the input document text and outputs a statistical depiction of the text. The statistical depiction, such as a histogram, shows the count of each word in the document. You are looking for relevant common repeated phrases, not individual words.

Option D is incorrect. The N-Gram natural language processing algorithm is used to find multi-word phrases in the text of a document. However, it does not weigh common words or phrases. You need the weighting aspect of the tf-idf algorithm to find the relevant, important repeated phrases used in the tweets.

Reference:

Please see the article titled Introduction to Natural Language Processing for Text.

Ask our Experts

View Queries

Question 43

Correct

Domain: Modeling

You work as a machine learning specialist for major phone and internet providers. Your customer support department needs to upgrade its phone response systems to reduce the number of human service representatives needed to handle dramatically increasing call volume. Your senior management team has leveraged off-shore call center services to reduce costs. Still, they now want to take advantage of voice recognition to automate many of the most frequent support call types, such as “I forgot my password” or “my internet is down.”

Your management team has assigned you to the team to implement the machine learning model behind the voice recognition system. Which SageMaker built-in algorithm is the best choice for this problem?

- A. Sequence-to-Sequence right
- B. K-Means
- C. Semantic Segmentation
- D. Neural Topic Model (NTM)

Explanation:

Answer: A

Option A is correct. The Sequence-to-Sequence algorithm takes audio as input data and generates a sequence of tokens, such as the words in the audio. This can then be used to provide automated responses to users’ requests.

Option B is incorrect. The K-Means algorithm is used to find groups within data where the group members are similar to each other but different from members of other groups. This algorithm will not help you encode speech audio streams.

Option C is incorrect. The semantic segmentation algorithm is used to develop computer vision applications. You are trying to solve a speech recognition problem. So this algorithm would not work for this problem.

Option D is incorrect. The Neural Topic Model algorithm is used to group documents into topics using the statistical distribution of words within the documents. You are trying to solve a speech recognition problem. So this algorithm would not work for this problem.

Reference:

Please see the SageMaker developer guide titled Using Amazon SageMaker Built-in Algorithms.

Ask our Experts

View Queries

Question 44

Incorrect Marked for review

Domain: ML Implementation and Operations

You work as a machine learning specialist for an eyewear manufacturing plant. There you have used XGBoost to train a model that uses assembly line image data to categorize contact lenses as malformed or correctly formed. You have engineered your data and used CSV as your Training ContentType. You are now ready to deploy your model using the Amazon SageMaker hosting service.

Assuming you used the default configuration settings, which of the following are true statements about your hosted model? (Select THREE)

A. The training instance class is multiple-instance GPU.

- [x] B. The algorithm is not parallelizable for distributed training. wrong
- [x] C. The training data target value should be in the first column of the CSV with no header. right

D. The training data target value should be in the last column of the CSV with no header.

E. The inference data target value should be in the first column of the CSV with no header.

- [x] F. The inference CSV data has no label column. right

G. The training instance class is CPU. right

Explanation:

Answers: C, F, G

Option A is incorrect. The SageMaker XGBoost currently only supports a CPU instance or a single-instance GPU instance type for training. (See the Amazon SageMaker developer guide titled XGBoost Algorithm, particularly the EC2 Instance Recommendation for the XGBoost Algorithm section)

Option B is incorrect. The XGBoost algorithm is parallelizable and therefore can be deployed on multiple CPU instances for distributed training. (See the Amazon SageMaker developer guide titled

Common Parameters for Built-in Algorithms)

Option C is correct. From the Amazon SageMaker developer guide titled XGBoost Algorithm, "For CSV training, the algorithm assumes that the target variable is in the first column and that the CSV does not have a header record".

Option D is incorrect. From the Amazon SageMaker developer guide titled Common Data Formats for Training, "Amazon SageMaker requires that a CSV file doesn't have a header record and that the target variable is in the first column".

Option E is incorrect. From the Amazon SageMaker developer guide titled XGBoost Algorithm, "For CSV inference, the algorithm assumes that CSV input does not have the label column".

Option F is correct. From the Amazon SageMaker developer guide titled XGBoost Algorithm, "For CSV inference, the algorithm assumes that CSV input does not have the label column".

Option G is correct. The SageMaker XGBoost currently only supports a CPU instance type for training. (See the Amazon SageMaker developer guide titled XGBoost Algorithm, particularly the EC2 Instance Recommendation for the XGBoost Algorithm section)

Reference:

Please see the Amazon SageMaker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services for an overview of the deployment of a SageMaker model.

Ask our Experts

View Queries

Question 45

Correct

Domain: Modeling

You work as a machine learning specialist for the highway toll collection division of the regional state area. The toll collection division uses cameras to identify car license plates as the cars pass through the various toll gates on the state highways. You are on the team that is using SageMaker Image Classification machine learning to read and classify license plates by state and then identify the actual license plate number.

Very rarely, cars pass through the toll gates with plates from foreign countries, for example, Great Britain or Mexico. The outliers must not adversely affect your model's predictions.

Which hyperparameter should you set, and to what value, to ensure these outliers do not adversely impact your model?

A. feature_dim set to 5

B. feature_dim set to 1

C. sample_size set to 10

D. sample_size set to 100

E. learning_rate set to 0.1 right

F. learning_rate set to 0.75

Explanation:

Answer: E

Option A is incorrect. The feature_dim hyperparameter is a setting on the K-Means and K-Nearest Neighbors algorithms, not the Image Classification algorithm.

Option B is incorrect. The feature_dim hyperparameter is a setting on the K-Means and K-Nearest Neighbors algorithms, not the Image Classification algorithm.

Option C is incorrect. The sample_size hyperparameter is a setting on the K-Nearest Neighbors algorithm, not the Image Classification algorithm.

Option D is incorrect. The sample_size hyperparameter is a setting on the K-Nearest Neighbors algorithm, not the Image Classification algorithm.

Option E is correct. The learning_rate hyperparameter governs how quickly the model adapts to new or changing data. Valid values range from 0.0 to 1.0. Setting this hyperparameter to a low value, such as 0.1, will make the model learn more slowly and be less sensitive to outliers. This is what you want. You want your model not to be adversely impacted by outlier data.

Option F is incorrect. The learning_rate hyperparameter governs how quickly the model adapts to new or changing data. Valid values range from 0.0 to 1.0. Setting this hyperparameter to a high value, such as 0.75, will make the model learn more quickly but be sensitive to outliers. This is not what you want. You want your model not to be adversely impacted by outlier data.

Reference:

Please see the Amazon SageMaker developer guide titled Image Classification Hyperparameters, and the Amazon SageMaker developer guide titled Use Amazon SageMaker Built-in Algorithms.

Ask our Experts

View Queries

Question 46

Incorrect

Domain: ML Implementation and Operations

You work as a machine learning specialist for a major oil refinery company. Your company needs to do complex analysis on its crude and oil chemical compound structures. You have selected an algorithm for your machine learning model that is not one of the SageMaker built-in algorithms. You have created your model using CreateModel, and you have created your HTTPS endpoint. Your docker

container running your model is now ready to receive inference requests for real-time inferences. When SageMaker returns the inference result from a client's request, which of the following are true? (Select THREE)

- [x] A. To receive inference requests your inference container must have a web server running on port 8080. right
- [x] B. Your inference container must accept GET requests to the /invocations endpoint. wrong
- [ ] C. Your inference container must accept PUT requests to the /inferences endpoint.
- [ ] D. Amazon SageMaker strips all POST headers except those supported by InvokeEndpoint. Amazon SageMaker might add additional headers. Your inference container must be able to ignore these additional headers safely. right
- [x] E. Your inference container must accept POST requests to the /inferences endpoint. wrong
- [ ] F. Your inference container must accept POST requests to the /invocations endpoint. right

Explanation:

Answers: A, D, F

Option A is correct. From the Amazon SageMaker developer guide titled Use You Own Inference Code with Hosting Services, "To receive inference requests, the container must have a web server listening on port 8080".

Option B is incorrect. The inference container must accept POST requests to the /invocations endpoint. (See the Amazon SageMaker developer guide titled Use You Own Inference Code with Hosting Services)

Option C is incorrect. The inference container must accept POST requests to the /invocations endpoint. (See the Amazon SageMaker developer guide titled Use You Own Inference Code with Hosting Services)Maker developer guide titled Deploy a Model on Amazon SageMaker Hosting Services)

Option D is correct. From the Amazon SageMaker developer guide titled Use You Own Inference Code with Hosting Services, "Amazon SageMaker strips all POST headers except those supported by InvokeEndpoint. Amazon SageMaker might add additional headers. Inference containers must be able to ignore these additional headers safely."

Option E is incorrect. The inference container must accept POST requests to the /invocations endpoint. (See the Amazon SageMaker developer guide titled Use You Own Inference Code with Hosting Services)

Option F is correct. The inference container must accept POST requests to the /invocations endpoint. (See the Amazon SageMaker developer guide titled Use You Own Inference Code with Hosting Services)

Reference:

Please see the Amazon SageMaker developer guide titled Deploy a Model.

Ask our Experts

View Queries

Question 47

Correct

Domain: Modeling

You work as a machine learning specialist for a personal care product manufacturer. You are creating a binary classification model that you want to use to predict whether a customer is likely to positively respond to toothbrush and toothpaste samples mailed to their house. Since your company incurs expenses for the products and the shipping when sending samples, you only want to send your samples to customers who, you believe, have a high probability of buying your products. When analyzing if a customer will follow up with a purchase, which outcome do you want to minimize in your confusion matrix to save costs?

A. True Negative

B. False Negative

C. False Affirmative

D. True Positive

E. False Positive right

Explanation:

Answer: E

Option A is incorrect. True Negatives are definitely not an outcome you want to minimize because you definitely don't want to send samples to customers who will not respond.

Option B is incorrect. You don't need to limit False Negatives as much as false positives, since False Negatives only omit customers with a higher probability of following up. Not sending a sample to these customers won't save costs.

Option C is incorrect. The terms used in a confusion matrix are: True Positive, False Negative, True Negative, and False Positive.

Option D is incorrect. True Positives are the ones to which you want to send your samples.

Option E is correct. You use a confusion matrix, or table, to describe the performance of a classification model on a set of test data when you know the true values. It's called a confusion matrix because it shows when one class is mislabeled (or confused) as another. For example, when the observation is negative, the model prediction is positive (a False Positive). To reduce the number of mailings to customers who probably won't follow up with a purchase, you want to limit False Positives.

Reference:

Please see the Wikipedia article titled Confusion Matrix.

Ask our Experts

View Queries

Question 48

Correct

Domain: Modeling

You work as a machine learning specialist for a clothing manufacturer. You have built a linear regression model using SageMaker's built-in linear learner algorithm to predict sales for a given year. Your training dataset observations are based on several features such as marketing dollars spent, number of active stores, traffic per store, online traffic to the company website, overall market indicators, etc. You have decided to use the k-fold method of cross-validation to assess how the results of your model will generalize beyond your training data.

Which of these will indicate that you don't have biased training data?

A. The variance of the estimate increases as you increase k.

B. You shouldn't have to worry about bias because your error function removes bias in the data.

C. Every k-fold cross-validation round increases the training error rate.

D. Every k-fold cross-validation round has a very similar error rate to the rate of all the other rounds. right

E. You would not normally use k-fold with linear regression models.

Explanation:

Answer: D

Option A is incorrect. When using k-fold for cross-validation, the variance of the estimate is reduced as you increase k. So 10-fold cross-validation should have a lower variance than 5-fold cross-validation.

Option B is incorrect. The k-fold error function just gives you the error rate of the cross-validation round. It doesn't resolve bias.

Option C is incorrect. The goal of k-fold cross-validation is to produce relatively equal error rates for each round (indicating proper randomization of the data), not to reduce the error rate for each round.

Option D is correct. If you have relatively equal error rates for all k-fold rounds, it is an indication that you have properly randomized your test data, therefore reducing the chance of bias.

Option E is incorrect. The k-fold cross-validation technique is commonly used with linear regression analysis.

Reference:

Please see the Amazon Machine Learning developer guide titled Evaluating ML Models, and the Amazon Machine Learning developer guide titled Cross-Validation

Ask our Experts

View Queries

Question 49

Correct

Domain: Modeling

You work as a machine learning specialist for the National Oceanic and Atmospheric Administration (NOAA Research). NOAA has developed a great white shark detection program to help warn shore populations when the sharks are in the area of a populated beach. You have the assignment to use your machine learning expertise to decide where to place 10 high-tech shark detection sensors on the oceanic floor as part of a pilot to determine if the NOAA invests broadly in these very expensive sensors. You have great white sightings data from around the globe gathered over the past several years to use your model training and test data. The model dataset contains several useful features, such as the longitude and latitude of each sighting.

You have decided to use an unsupervised learning algorithm that attempts to find discrete groupings within the data. Specifically, you want to find similarities in the longitude and latitude and find groupings of these. You need to produce 10 longitude and latitude pairs to determine where to place the sensors.

Which algorithm can you use in SageMaker that best suits this task?

A. Linear Learner

B. Neural Topic Model

C. K-Means right

D. Random Cut Forest

E. Semantic Segmentation

F. XGBoost

Explanation:

Answer: C

Option A is incorrect. From the Amazon SageMaker developer guide titled Linear Learner Algorithm, "*Linear models* are supervised learning algorithms used for solving either classification or regression problems." But you are trying to solve a data clustering problem so that you can find the ten best clustered sightings to determine where to place your shark detection sensors.

Option B is incorrect. From the Amazon SageMaker developer guide titled Neural Topic Model (NTM) Algorithm, "Amazon SageMaker NTM is an unsupervised learning algorithm that is used to organize a corpus of documents into *topics* that contain word groupings based on their statistical distribution." So this algorithm is used for natural language processing, not data clustering.

Option C is correct. The k-means algorithm is a clustering algorithm. From the Amazon SageMaker developer guide titled K-Means Algorithm, "K-means is an unsupervised learning algorithm. It attempts to find discrete groupings within data, where members of a group are as similar as possible to one another and as different as possible from members of other groups." By setting the k hyperparameter to 10, this algorithm will allow you to find the 10 best groupings of shark sightings worldwide.

Option D is incorrect. From the Amazon SageMaker developer guide titled Random Cut Forest (RCF) Algorithm, "Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a data set." But you are trying to solve a data clustering problem so you can find the ten best clustered sightings to determine where to place your shark detection sensors.

Option E is incorrect. From the Amazon SageMaker developer guide titled Semantic Segmentation Algorithm, "The Amazon SageMaker semantic segmentation algorithm provides a fine-grained, pixel-level approach to developing computer vision applications." So the Semantic Segmentation algorithm is used for computer vision applications, but you are trying to solve a data clustering problem.

Option F is incorrect. The XGBoost algorithm is a gradient boosting algorithm. From the Amazon SageMaker developer guide titled XGBoost Algorithm, "gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable by combining an ensemble of estimates from a set of simpler, weaker models." You are not trying to predict a target value; you are trying to find discrete groupings in your dataset.

Reference:

Please see the Amazon SageMaker developer guide titled Use Amazon SageMaker Built-in Algorithms.

Ask our Experts

View Queries

Question 50

Correct

Domain: Modeling

You work as a machine learning specialist for a sports analytics company. The Major League Baseball Association has contracted your company to perform real-time analytics on baseball statistics as baseball plays unfold live on national television. Your first assignment is to predict the outcome of situational set plays (such as stolen bases or pitch results) as they are about to unfold. Therefore, your model must deliver its predictions in close to real-time.

You have decided to use a SageMaker built-in algorithm. You have looked at classical forecasting methods like autoregressive integrated moving average (ARIMA) and exponential smoothing (ETS) which use one model for each time series in your data. However, you have many time series over which to train.

Based on your performance requirements and your training requirements, which SageMaker built-in algorithm should you use?

A. Linear Learner

B. Neural Topic Model

C. K-Means

D. Random Cut Forest

E. DeepAR Forecasting right

F. XGBoost

Explanation:

Answer: E

Option A is incorrect. From the Amazon SageMaker developer guide titled Linear Learner Algorithm "*Linear models* are supervised learning algorithms used for solving either classification or regression problems." But you are trying to solve one-dimensional time series problem so that you can extrapolate the baseball playtime series into the future.

Option B is incorrect. From the Amazon SageMaker developer guide titled Neural Topic Model (NTM) Algorithm "Amazon SageMaker NTM is an unsupervised learning algorithm used to organize a corpus of documents into *topics* that contain word groupings based on their statistical distribution." So this algorithm is used for natural language processing, not time series problems.

Option C is incorrect. The k-means algorithm is a clustering algorithm. From the Amazon SageMaker developer guide titled K-Means Algorithm "K-means is an unsupervised learning algorithm. It attempts to find discrete groupings within data, where members of a group are as similar as possible to one another and as different as possible from members of other groups." You are trying to solve one-dimensional time series problems to extrapolate playtime series into the future, not a data clustering problem.

Option D is incorrect. From the Amazon SageMaker developer guide titled Random Cut Forest (RCF) Algorithm "Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for

detecting anomalous data points within a data set." But you are trying to solve a one-dimensional time series problem to extrapolate baseball playtime series into the future.

Option E is correct. From the Amazon SageMaker developer guide titled DeepAR Forecasting Algorithm "... you have many similar time series across a set of cross-sectional units. For example, you might have time series groupings for demand for different products, server loads, and requests for webpages. For this type of application, you can benefit from training a single model jointly over all of the time series. DeepAR takes this approach. When your dataset contains hundreds of related time series, DeepAR outperforms the standard ARIMA and ETS methods. You can also use the trained model to generate forecasts for new time series that are similar to the ones it has been trained on." Also, from the same developer guide, "The training input for the DeepAR algorithm is one or, preferably, more target time series that the same process or similar processes have generated. Based on this input dataset, the algorithm trains a model that learns an approximation of this process/processes and uses it to predict how the target time series evolves." So the DeepAR algorithm is used for one-dimensional time series problems for complex analysis like baseball play prediction.

Option F is incorrect. The XGBoost algorithm is a gradient boosting algorithm. From the Amazon SageMaker developer guide titled XGBoost Algorithm, "gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable by combining an ensemble of estimates from a set of simpler, weaker models." You are not trying to predict a target value; you are trying to solve a one-dimensional time series problem.

Reference:

Please see the Amazon SageMaker developer guide titled Use Amazon SageMaker Built-in Algorithms, the AWS Machine Learning Blog titled Now Available in Amazon SageMaker: DeepAR algorithm for more accurate time series forecasting, and the AWS StatCast AI page titled See how AI on AWS gives baseball fans new insights into the game.

Ask our Experts

View Queries

Question 51

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a flight data company. Your company has a contract with the US National Defence to produce real-time detection capabilities for fighter jet flight assist software. Due to the nature of the use case, the implementation of the algorithm you choose for your machine learning model must be able to perform detections as close to real-time as possible.

You are in the development stages and have chosen to use the Image Classification SageMaker built-in deep learning model. You are setting up your jupyter notebook instance in SageMaker. Which of the following jupyter notebook settings will allow you to test and evaluate production performance when you are building your models?

A. Notebook instance type

B. Lifecycle configuration

C. Volume size

D. Elastic inference right

E. Primary container

Explanation:

Answer: D

Option A is incorrect. This is the type of EC2 instance on which your notebook will run. This won't help you understand production performance.

Option B is incorrect. The lifecycle configuration allows you to customize your notebook environment with default scripts and plugins. Default jupyter notebook scripts and plugins won't give you an insight into production performance.

Option C is incorrect. The volume size is just the size of the jupyter instance in GBs. This won't give you an insight into production performance.

Option D is correct. From the Amazon SageMaker developer guide titled Amazon SageMaker Elastic Inference (EI) "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 ... You can also add an EI accelerator to an Amazon SageMaker notebook instance so that you can test and evaluate inference performance when you are building your models". Therefore, while you are in the development stage using jupyter notebooks, Elastic Inference allows you to gain insight into the production performance of your model once it is deployed.

Option E is incorrect. From the Amazon SageMaker developer guide titled CreateModel "... you name the model and describe a primary container. For the primary container, you specify the docker image containing inference code, artifacts (from prior training), and custom environment map that the inference code uses when you deploy the model for predictions.

Use this API to create a model if you want to use Amazon SageMaker hosting services or run a batch transform job." So the primary container is a parameter used in the CreateModel request when you are creating a model in SageMaker. It is not used when setting up your jupyter notebook.

Reference:

Please see the Amazon SageMaker developer guide titled Amazon SageMaker Elastic Inference (EI), the AWS FAQ titled Amazon Elastic Inference FAQs, and the AWS Machine Learning blog titled Optimizing costs in Amazon Elastic Inference with TensorFlow.

Ask our Experts

View Queries

Question 52

Correct

Domain: ML Implementation and Operations

You work as a machine learning specialist for a polling research company. You have national polling data for the last 10 presidential elections that you have engineered, randomized, partitioned into various training and test datasets, and stored on S3. You have selected a SageMaker built-in algorithm to use for your model. Your training datasets are very large. As you repeatedly run your training job with different large datasets, you find your training takes a very long time.

How can you improve the performance of your training runs? (Select TWO)

- [x] **A. Use the protobuf recordIO format.** right
- [ ] **B. Convert your data to XML and use file mode to load your data to the EBS training instance volumes.**
- [x] **C. Use pipe mode to stream the training data directly to your EBS training instance volumes.** right
- [ ] **D. Convert your data to CSV and use file mode to load your data to the EBS training instance volumes.**
- [ ] **E. Change your Elastic Inference accelerator type to a larger instance type.**

Explanation:

Answers: A, C

Option A is correct. The protobuf recordIO format, used for training data, is the optimal way to load data into your model for training. (See the Amazon SageMaker developer guide titled Common Data Formats for Training)

Option B is incorrect. XML is not a supported data format for training in SageMaker. (See the Amazon SageMaker developer guide titled Common Data Formats for Training)

Option C is correct. When you use the protobuf recordIO format, you can also take advantage of pipe mode when training your model. Pipe mode, used together with the protobuf recordIO format, gives you the best data load performance by streaming your data directly from S3 to your EBS volumes used by your training instance. (See the Amazon SageMaker developer guide titled Common Data Formats for Training)

Option D is incorrect. When you use the CSV format and file mode, all of your data is loaded from S3 to the EBS volumes used by your training instance. This is much less efficient from a performance perspective than streaming the training data directly from S3 to your EBS volumes used by your training instance. (See the Amazon SageMaker developer guide titled Common Data Formats for Training)

Option E is incorrect. Elastic Inference is used to speed up the throughput of retrieving real-time inferences from models deployed as SageMaker hosted models. Elastic Inference accelerators

accelerate your inference calls; they aren't used while training. (See the Amazon SageMaker developer guide titled Amazon SageMaker Elastic Inference (EI))

Reference:

Please see the Amazon SageMaker developer guide titled Common Data Formats for Built-in Algorithms and the AWS FAQ titled Amazon Elastic Inference FAQs.

Ask our Experts

View Queries

Question 53

Correct

Domain: Data Engineering

You work for a financial services company where you have a large Hadoop cluster hosting a data lake in your on-premises data center. Your department has loaded your data lake with financial services operational data from your corporate actions, order management, cash management, reconciliations, and trade management systems. Your investment management operations team now wants to use data from the data lake to build financial prediction models. You want to use data from the Hadoop cluster in your machine learning training jobs. Your Hadoop cluster has Hive, Spark, Sqoop, and Flume installed.

How can you most effectively load data from your Hadoop cluster into your SageMaker model for training?

A. Use the distcp utility to copy your dataset from your hadoop platform to the S3 bucket where your SageMaker training job can use it.

B. Use the HadoopActivity command with AWS Data Pipeline to move your dataset from your hadoop platform to the S3 bucket where your SageMaker training job can use it.

C. Use the SageMaker Spark library using the data frames in your Spark clusters to train your model. right

D. Use the Sqoop export command to export your dataset from your Hadoop cluster to the S3 bucket where your SageMaker training job can use it.

Explanation:

Answer: C

Option A is incorrect. The Hadoop distcp utility is used for inter/intra cluster data movement. It is not an efficient method to get data into your SageMaker training instance. (See the Apache Hadoop distcp guide)

Option B is incorrect. The HadoopActivity command is used to run a job on a cluster. You would have to write the job to extract and load the data onto S3. This would not be the most efficient method of the options listed. (See AWS Data Pipeline developer guide titled HadoopActivity)

Option C is correct. The SageMaker Spark library makes it; so you can easily train models using data frames in your Spark clusters. This is the most efficient method of the options listed. (See the Amazon SageMaker developer guide titled Use Apache Spark with Amazon SageMaker)

Option D is incorrect. The Sqoop export command is used for exporting files from HDFS to an RDBMS. This would not help you load your data into your SageMaker training instance. (See the Sqoop User Guide)

Reference:

Please see the Amazon SageMaker developer guide titled Use Machine Learning Frameworks with Amazon SageMaker.

Ask our Experts

View Queries

Question 54

Correct

Domain: Exploratory Data Analysis

You are working for a consulting firm in its machine learning practice. Your current client is a sports equipment manufacturer. You are building a linear regression model to predict ski and snowboard sales based on the daily snowfall in various regions around the country.

After you have cleaned and performed feature engineering on your CSV data, which of the following tasks would you perform next?

A. Use the scikit-learn cross_validate method to evaluate the estimation precision of your model.

B. Load your data into a pandas DataFrame and remove header rows and any superfluous features.

C. Use one-hot encoding to convert categorical values, such as 'region of the country' to numerical values.

D. Shuffle your data using a shuffling technique. right

Explanation:

Answer: D

Option A is incorrect. The scikit-learn cross_validate method is used to evaluate your model's precision while tuning the model's hyperparameters. (See Scikit-Learn user guide titled cross_validate)

Option B is incorrect. Using a Pandas DataFrame to remove superfluous rows and features is part of cleaning and doing feature engineering of your data, which you have already done.

Option C is incorrect. One-hot encoding is another way to do feature engineering on your data in preparation for training. You have already completed the cleaning and feature engineering of your data.

Option D is correct. Once you have cleaned and engineered your data for a linear regression model, you need to shuffle the data to prevent overfitting and reduce variance. (See Amazon Machine Learning developer guide titled The Amazon Machine Learning Process)

Reference:

Please see the Amazon Machine Learning developer guide titled Machine Learning Concepts, and the Amazon Machine Learning developer guide titled The Amazon Machine Learning Process.

Ask our Experts

View Queries

Question 55

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist at a retail shoe manufacturer. Your marketing department wants to do a promotion for a new running shoe they are about to release into their product pipeline. They need a model to predict sales of the new shoe using the purchase history of their registered customers based on past releases of new shoes.

You have decided to use a linear regression algorithm for your model. Your data has thousands of observations and 35 numeric features. While doing analysis to understand your data better, you find 25 observations that have what looks like outlier data points. After speaking to your marketing department, you learn that these values are valid. You also find several hundred observations that have some blank feature values.

How should you correct the outlier and blank feature problems?

A. Remove the observations with the outlier data points and replace the blank values with the null value.

B. Remove the outlier and blank value observations.

C. Remove the observations with the outlier data points and replace the blank values with the mean value. right

D. Remove the observations with the outlier data points and replace the blank values with the value 0.

Explanation:

Answer: C

Option A is incorrect. Null values in observation should be replaced since linear regression calculations will have a problem with null values. Therefore, you would not replace empty fields with null.

Option B is incorrect. Removing the observations with blank values will reduce the accuracy of your model's predictions since you have removed many features from the training dataset.

Option C is correct. You should remove the outlier observations. You should also replace the blank values with a meaningful value. The mean value is the best option of those listed.

Option D is incorrect. You should remove the outlier observations. You should also replace the blank values with a meaningful value. The 0 value is not the best option of those listed because the mean is invariably a better approximation than 0 for a continuous numeric value.

Reference:

Please see the Amazon Machine Learning developer guide titled Feature Processing.

Ask our Experts

View Queries

Question 56

Correct

Domain: Data Engineering

You work as a machine learning specialist at a hedge fund firm. Your firm is working on a new quant algorithm to predict when to enter and exit holdings in their portfolio. You are building a machine learning model to predict these entry and exit points in time. You have cleaned your data, and you are now ready to split the data into training and test datasets.

Which splitting technique is best suited to your model's requirements?

A. Use k-fold cross validation to split the data.

B. Sequentially splitting the data right

C. Randomly splitting the data

D. Categorically splitting the data by holding

Explanation:

Answer: B

Option A is incorrect. Using k-fold cross validation will randomly split your data. But you need to consider the time-series nature of your data when splitting. So randomizing the data would eliminate the time element of your observations, making the datasets unusable for predicting price changes over time.

Option B is correct. By sequentially splitting the data, you preserve the time element of your observations.

Option C is incorrect. Randomly splitting the data would eliminate the time element of your observations, making the datasets unusable for predicting price changes over time.

Option D is incorrect. If you split the data by a category such as the holding attribute, you would create imbalanced training and test dataset since some holdings would only be in the training dataset and others would only be in the test dataset.

Reference:

Please see the Amazon Machine Learning developer guide titled Splitting Your Data.

Ask our Experts

View Queries

Question 57

Correct

Domain: Data Engineering

You work as a machine learning specialist for a software company developing a movie rating social media site where users can rate movies. You want to use your companies data to predict the rating distribution of a movie based on the genre of the movie. Your training data contains a genre feature with a set of categories such as documentary, romance, etc. You have sorted your data by the genre feature and then used the Amazon ML sequential split option to split your data into training and test datasets.

When using your test dataset to verify your genre-prediction model, you discover that the accuracy rate is very low. What could be the underlying problem?

A. You should have sorted by a different feature before you used the sequential split option.

B. You should have split your data categorically by genre.

C. You should have split your data sequentially by year.

D. You should not have used the sequential split option. right

Explanation:

Answer: D

Option A is incorrect. Sorting the data by a different feature wouldn't solve the problem. You used the sequential option when splitting the data. So you have not properly randomized your data.

Option B is incorrect. By categorically splitting the data by definition, you will have some genre movies only in the training dataset and others only in the test dataset. This reduces the genre feature to a meaningless datapoint.

Option C is incorrect. Sequentially splitting the data by year wouldn't solve the problem. You used the sequential option when splitting the data. So you have not properly randomized your data.

Option D is correct. You should not have used the sequential option when splitting your data. To get proper generalization from your data, you need to randomize it for this type of problem.

Reference:

Please see the Amazon Machine Learning developer guide titled Splitting Your Data.

Ask our Experts

View Queries

Question 58

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist for a real estate company. You are using the Kaggle housing prices data as your experimentation data to optimize your model before using your model on the real estate data for your area of the country. You have a hypothesis that you can predict the price of a real estate property based on the foundation type. You have your data from Kaggle. But you want to make sure your model is not overly influenced by outliers.

What is the quickest way to identify outliers in your data?

A. Arrange your data points from lowest to highest, calculate the median of the data set and use a qualitative assessment to determine whether to remove outliers.

B. Calculate the Z-Score for your data points.

C. Visualize your data using scatter plots and/or box plots. right

D. Visualize your data using network and correlation matrices.

Explanation:

Answer: C

Option A is incorrect. You can find your outliers using a quantitative assessment. But it will involve more effort and, therefore, more time than visualizing your data.

Option B is incorrect. The z-score of a data point shows how many standard deviations the data point is from the mean. This would help you find your outliers. But it will involve more effort and, therefore, more time than visualizing your data.

Option C is correct. With large datasets, such as the real estate data you are using in this problem, the quickest way to find outliers is to visualize your data. The best plots for this task are the scatter plot and the box plot. (See the article titled How to Make your Machine Learning Models Robust to Outliers)

Option D is incorrect. Visualization is the quickest and easiest way to find outliers, but the network and/or correlation matrix charting choices will not show outliers. They are used to represent relations between data points as nodes. These relationships would not give you any information about the extremity of a data point.

Reference:

Please see the article titled How to Make your Machine Learning Models Robust to Outliers and the article titled A Brief Overview of Outlier Detection Techniques.

Ask our Experts

View Queries

Question 59

Correct

Domain: Modeling

You work as a machine learning specialist for a company that runs a car rating website. Your company wants to build a price prediction model that is more accurate than their current model, which is a linear regression model using the age of the car as the single independent variable in the regression to predict the price. You have decided to add the horsepower, fuel type, city mpg (miles per gallon), drive wheels, and a number of doors as independent variables in your model. You believe that adding these additional independent variables will give you a more accurate prediction of price.

Which type of algorithm will you now use for your prediction?

A. Logistic Regression

B. Decision Tree

C. Naive Bayes

D. Multivariate Regression right

Explanation:

Answer: D

Option A is incorrect. Logistic regression is used for problems where you are trying to classify and estimate a discrete value (on or off, 1 or 0) based on a set of independent variables. In your problem, you are trying to estimate a continuous numerical value: price, not a binary classification.

Option B is incorrect. A decision tree can be used as a classification algorithm or a regression algorithm, however, this problem involves multiple independent variables which leads us to the more relevant answer: Multivariate Regression.

Option C is incorrect. Naive Bayes is another classification algorithm, so it is not a good fit for your continuous numerical value prediction problem.

Option D is correct. You are trying to predict the price of a car (dependent variable) based on a number of independent variables (horsepower, fuel type, city mpg, drive wheels, and a number of doors, etc.) The Multivariate Regression algorithm is the best choice for this type of problem. (See the article Data Science Simplified Part 5: Multivariate Regression Models)

Reference:

Please see the Amazon Machine Learning developer guide titled Regression Model Insights, and the article titled Commonly Used Machine Learning Algorithms (with Python and R codes)

Ask our Experts

View Queries

Question 60

Correct

Domain: Exploratory Data Analysis

You work as a machine learning specialist for a company that produces polling data and uses it for predictive modeling. Your company wants to build an election prediction model that uses multiple independent variables such as age of the voter, religion, sex, registered affiliation, etc. to predict the candidate for which each observed voter will vote in the upcoming election.

Which type of algorithm is NOT a good choice to use for your prediction? (Select FOUR)

- [x] A. Ordinary Least Squares Regression (OLSR) right
- [x] B. Local Outlier Factor (LOF) right
- [ ] C. Naive Bayes
- [x] D. Least-Angle Regression (LARS) right
- [x] E. K-Means right

Explanation:

Answers: A, B, D, E

Option A is correct. Ordinary Least Squares Regression (OLSR) is a regression technique that predicts a dependent variable using one or more independent variables. You are trying to solve a classification problem, which candidate, from a discrete list of candidates, will a voter choose.

Option B is correct. The Local Outlier Factor (LOF) algorithm is used to discover outlier data points. So this would NOT be a good choice for your algorithm where you are trying to solve a classification problem, which candidate, from a discrete list of candidates, will a voter choose.

Option C is incorrect. The Naive Bayes algorithm can be used as a classifier. You are trying to solve a classification problem, which candidate, from a discrete list of candidates, will a voter choose.

Option D is correct. Least-Angle Regression (LARS) is also a regression technique that predicts a dependent variable using one or more independent variables. You are trying to solve a classification problem, which candidate, from a discrete list of candidates, will a voter choose.

Option E is correct. The K-Means algorithm is used as a clustering algorithm, so it would NOT be a good choice for your algorithm where you are trying to solve for a dependent variable based on multiple independent variables.

Reference:

Please see the Amazon Machine Learning developer guide titled Regression Model Insights, and the article titled A Tour of the Most Popular Machine Learning Algorithms.

Ask our Experts

View Queries

Question 61

Correct

Domain: Data Engineering

You are a machine learning specialist for a research firm. Your team uses Amazon SageMaker and its built-in scikit-learn library for feature transformation in your machine learning process. When using the SimpleImputer transformer to replace missing values in your observations, which strategy is the default strategy that your SageMaker scikit-learn code will use if you don't explicitly pass a strategy parameter?

A. constant

B. most_frequent

C. median

D. mean right

E. mode

Explanation:

Answer: D

Option A is incorrect. The default strategy is mean. The constant strategy replaces the missing values with a constant you supply.

Option B is incorrect. The default strategy is mean. The most_frequent strategy replaces the missing values with the most frequent value along each column.

Option C is incorrect. The default strategy is mean. The median strategy replaces the missing values with the median along each column.

Option D is correct. The default strategy is mean. The mean strategy replaces the missing values with the mean along each column.

Option E is incorrect. There is no mode strategy in the SimpleImputer scikit-learn transformer.

Reference:

Please see the Amazon Machine Learning blog titled Preprocess input data before making predictions using Amazon SageMaker inference pipelines and Scikit-learn.

Ask our Experts

View Queries

Question 62

Correct

Domain: Data Engineering

You are a machine learning specialist for a gaming software startup. Your company is investigating ways to use machine learning to enhance its game software platform. The team has selected the Amazon SageMaker platform for their machine learning efforts. You are participating in the feature transformation process in preparation to create your machine learning models. Instead of transforming your data before you use it in your SageMaker models, you and your team have decided to use the built-in transformations of SageMaker. Specifically, you and your team have decided to use the built-in OneHotEncoder transformer to transform your categorical data.

You have decided to drop one of the categories per feature because you suspect you may have perfectly collinear features. Which of the following is NOT a drop methodology used in the OnHotEncoder transformer?

A. None

B. Last right

C. Array

D. First

Explanation:

Answer: B

Option A is incorrect. The OneHotEncoder transformer has the following methodologies you can use to drop one of the categories per feature: None, first, array. None is the default methodology.

Option B is correct. The OneHotEncoder transformer has the following methodologies you can use to drop one of the categories per feature: None, first, array. None is the default methodology. The OneHotEncoder transformer drop parameter does not offer the last methodology.

Option C is incorrect. The OneHotEncoder transformer has the following methodologies you can use to drop one of the categories per feature: None, first, array. None is the default methodology.

Option D is incorrect. The OneHotEncoder transformer has the following methodologies you can use to drop one of the categories per feature: None, first, array. None is the default methodology.

Reference:

Please see the Amazon Machine Learning blog titled Preprocess input data before making predictions using Amazon SageMaker inference pipelines and Scikit-learn, and the Scikit-learn API documentation OneHotEncoder

Ask our Experts

View Queries

Question 63

Correct

Domain: Modeling

You work as a machine learning specialist for a consulting firm that has the NFL as a client. You are working on the passer completion probability model using statistics from in-play metrics. You are running your linear learner model in Amazon SageMaker using a CSV file representation of your passer completion probability statistics. You are now running your inference.

Some of the features and their data types are listed below.

Feature Name	Data Type
Passer age	Numeric
Length of pass	Numeric

| Complete (yes/no) | Categorical |

| Feature Name | Data Type |

| Distance between receiver and nearest defender | Numeric |

| Play called (post, crossing, screen, etc.) | Categorical |

You are using the Complete feature as your prediction response feature. You are now making predictions on new data. When you interrogate the response of your model, which of the following do you expect to find?

A. score: the prediction produced by the model

B. score: the prediction produced by the model AND predicted_class which is an integer from 0 to num_classes-1

C. score: single floating point number measuring the strength of the prediction AND predicted_label which is 0 or 1 right

D. score: the prediction produced by the model OR predicted_label which is 0 or 1

Explanation:

Answer: C

Option A is incorrect. For a binary classification (complete yes or no), the model produces a score denoting the strength of the prediction AND a predicted_label denoting complete or not complete.

Option B is incorrect. This option describes the response for multiclass classification, but you are working with binary classification.

Option C is correct. For a binary classification (complete yes or no), the model produces a score denoting the strength of the prediction AND a predicted_label denoting complete or not complete.

Option D is incorrect. For a binary classification (complete yes or no), the model produces a score denoting the strength of the prediction AND a predicted_label denoting complete or not complete.

Reference:

Please see the Amazon SageMaker developer guide titled Linear Learner Algorithm.

Ask our Experts

View Queries

Question 64

Correct

Domain: Modeling

You work in the machine learning department of a major retail company. Your team is working on a model to predict the region with the highest sales for a given quarter. You have selected your observations from past sales cycles for all regions and split your data into training and evaluation datasets. You are now training your linear learner model in Amazon SageMaker. You are trying to select the model hyperparameters that give your team the best predictions.

You have set the predictor_type hyperparameter to binary_classifier. Which loss function hyperparameter setting is NOT one of your options?

A. auto

B. logistic

C. hinge_loss

D. softmax_loss right

Explanation:

Answer: D

Option A is incorrect. The three hyperparameters values that you can set for the loss function are auto, logistic, and hinge_loss. The default for auto is logistic.

Option B is incorrect. The three hyperparameters values that you can set for the loss function are auto, logistic, and hinge_loss. The default for auto is logistic.

Option C is incorrect. The three hyperparameters values that you can set for the loss function are auto, logistic, and hinge_loss. The default for auto is logistic.

Option D is correct. The three hyperparameters values that you can set for the loss function are auto, logistic, and hinge_loss. The default for auto is logistic. The softmax_loss setting is an option if your predictor_type is set to multiclass_classifier.

Reference:

Please see the Amazon SageMaker developer guide titled Linear Learner Hyperparameters.

Ask our Experts

View Queries

Question 65

Correct

Domain: Modeling

You work in the machine learning department of a major retail company. Your team is working on a model to classify customers by purchase history. Your marketing department wants to use the results of your model predictions to determine which customers should receive a new campaign offer. You have selected your observations and cleaned your data. You have also split your data into training and evaluation datasets. You are now training your k-means model in Amazon SageMaker, and you are trying to select the model hyperparameters that give your marketing team the best predictions.

You have set the feature_dim hyperparameter to equal the number of features in your input data. You have set the k hyperparameter to 10. The number of clusters you estimate is appropriate for your model. You have set the epochs hyperparameter to 1 so that the model performs one pass over your data.

You need to report a score for your model. Which k-means hyperparameter allows you to select the metric types to report this scoring, and what are the available metric options?

A. extra_center_factor with msd, ssd, or [msd, ssd] as the available metric type values

B. score_metrics with mse, ssd, or [mse, ssd] as the available metric type values

C. eval_method with mse, ssd, or [mse, ssd] as the available metric type values

D. eval_metrics with msd, ssd, or [msd, ssd] as the available metric type values right

Explanation:

Answer: D

Option A is incorrect. The hyperparameter you chose to report a score for your model is the eval_metrics hyperparameter. The eval_metrics hyperparameter has the allowed values of msd for Mean Square Error, ssd for Sum of Square Distance, and the option of both msd and ssd. The extra_center_factor is used to control the number of clusters.

Option B is incorrect. The hyperparameter you chose to report a score for your model is the eval_metrics hyperparameter. The eval_metrics hyperparameter has the allowed values of msd for Mean Square Error, ssd for Sum of Square Distance, and the option of both msd and ssd. The Amazon SageMaker k-means algorithm does not have a score_metrics hyperparameter.

Option C is incorrect. The hyperparameter you chose to report a score for your model is the eval_metrics hyperparameter. The eval_metrics hyperparameter has the allowed values of msd for Mean Square Error, ssd for Sum of Square Distance, and the option of both msd and ssd. The Amazon SageMaker k-means algorithm does not have an eval_method hyperparameter.

Option D is correct. The hyperparameter you chose to report a score for your model is the eval_metrics hyperparameter. The eval_metrics hyperparameter has the allowed values of msd for Mean Square Error, ssd for Sum of Square Distance, and the option of both msd and ssd.

Reference:

Please see the Amazon SageMaker developer guide titled K-Means Hyperparameters.

Ask our Experts

View Queries

Finish Review

Certification

Cloud Certification

Java Certification

PM Certification

Big Data Certification

Company

Become Our Instructor

Support

Discussions

Blog

Business

Support

Contact Us

Help Topics

Join us on Slack!

Join our open **Slack community** and get your queries answered instantly! Our experts are online to answer your questions!