DUMPSARENA

Senior Data Scientist

DASCA SDS

Version Demo

Total Demo Questions: 10

Total Premium Questions: 106

Buy Premium PDF

https://dumpsarena.co

sales@dumpsarena.co

sales@dumpsarena.co
dumpsarena.co

QUESTION NO: 1

A data-science team is preparing a binary classifier for a medical-screening workflow in which missing a true case is substantially more harmful than referring a non-case for follow-up. Select two actions that are appropriate when selecting the operating threshold.

A. Select a threshold using the relative costs or utilities of false negatives and false positives

B. Compare recall and false-negative rates across candidate thresholds

C. Use 0.50 because it is the standard classification threshold

D. Select the threshold that maximizes overall accuracy

E. Select the threshold that maximizes precision regardless of recall

ANSWER: A B

Explanation:

The threshold should be selected using the consequences of false negatives and false positives, rather than by treating 0.50 as inherently optimal. A cost-sensitive or utility-based evaluation makes the stated clinical trade-off explicit. The team should also examine recall and false-negative behaviour at candidate thresholds, because the principal concern is failing to identify true cases.

Overall accuracy can be misleading, especially when the positive class is uncommon, because a model may achieve high accuracy while missing many true cases. Choosing the threshold that maximizes precision alone would favour fewer positive referrals and can worsen the costly false-negative outcome.

QUESTION NO: 2

Which of the following is used to summarize a dataset by showing the median, quantiles, and min/max values for each of the variables?

A. Box Plots

B. Pie Charts

C. Histogram

D. Scatter Chart

E. Bar Charts

ANSWER: A

Explanation:

Box Plots are used to provide a compact visual summary of a variable’s distribution. A standard box plot marks the median as the central line in the box and uses the lower and upper edges of the box to represent the first and third quartiles, which are commonly called quantiles. The box therefore displays the interquartile range: the middle half of the observed values. Lines extending from the box, often called whiskers, show the spread toward the low and high ends of the data according to the plotting convention; points beyond those limits can be displayed individually as potential outliers. This makes box plots particularly useful during exploratory data analysis, because they allow a data scientist to compare the center, variability, skew, and unusual values across several variables or groups without examining every observation. The exact whisker definition can vary by software and convention, so whiskers do not always equal the literal minimum and maximum; nevertheless, box plots are the recognized chart for summarizing data with the median, quartiles, and range-related extremes. See the NIST Engineering Statistics Handbook’s box plot overview and Seaborn boxplot documentation.

QUESTION NO: 3

What is DevOps?

A. Software Development

B. Software Operations

C. Quality Assurance

D. All

ANSWER: D

Explanation:

DevOps is correctly represented by *All* because it is an organizational and technical approach that brings software development, software operations, and quality practices together across the software delivery lifecycle. Its purpose is to replace isolated handoffs with shared responsibility for planning, building, testing, releasing, deploying, operating, monitoring, and continually improving software. Development contributes application design and implementation; operations contributes reliable infrastructure, deployment, observability, and service management; and quality assurance contributes automated testing and continuous feedback that help teams deliver dependable changes quickly.

In mature DevOps practice, these areas are not merely adjacent functions. They collaborate through common processes and automation such as continuous integration, continuous delivery, infrastructure as code, automated quality checks, monitoring, and incident learning. This integration improves delivery speed while preserving reliability, security, and customer value. Microsoft describes DevOps as the union of people, process, and technology to continually provide value, while AWS emphasizes collaboration between development and operations teams and automation throughout delivery. These definitions support the inclusive scope captured by *All*. See Microsoft Learn: What is DevOps? and AWS: What is DevOps?.

QUESTION NO: 4

Data wrangling is the process of getting the data from:

A. Its raw format into something suitable for more conventional analytics

B. Its modified meaning format into something suitable for more conventional analytics

C. Both A and B

D. None of the above

ANSWER: A

Explanation:

“Its raw format into something suitable for more conventional analytics” correctly describes data wrangling. In practical data-science work, source data commonly arrives in forms that cannot be reliably analyzed as-is: values may be missing, fields can use inconsistent formats, records can be duplicated, and information may be spread across multiple files or systems. Data wrangling is the set of activities used to prepare those raw inputs for meaningful downstream use. It includes acquiring and profiling the data, cleaning quality issues, reshaping tables, standardizing data types and representations, integrating relevant sources, and producing an analysis-ready dataset. The goal is not merely to move data, but to make it structured, consistent, and usable for analytics, visualization, statistical methods, and machine-learning workflows. This preparation is foundational to the data-engineering and data-preprocessing capabilities expected of a senior data scientist, since the quality and usability of the prepared data directly affect the validity of subsequent analysis. DASCA positions these practical data-science capabilities within its Senior Data Scientist certification curriculum; see the DASCA Senior Data Scientist certification overview. For a complementary industry definition of the process of transforming raw data into an analysis-ready form, see IBM’s data-wrangling overview.

QUESTION NO: 5

A senior data scientist must make a deployed model reproducible six months after release, even if the source data and software packages have changed. Select two artefacts that are most important to retain with the release record.

A. A versioned training-data reference and feature-definition logic

B. The trained model artefact and versioned software environment specification

C. Only a presentation showing aggregate validation performance

D. Informal notes describing what team members remember about training

E. A fresh extract of the current production data

ANSWER: A B

Explanation:

A versioned training-data snapshot, or a reproducible immutable reference to it, together with the feature-definition logic allows the organization to determine exactly what observations and transformations informed the model. The trained model artefact and a versioned execution environment or dependency specification allow the approved model to be rerun or inspected with the intended software context. A presentation of aggregate performance is useful documentation, but it cannot recreate the training process. Informal recollection and a current production extract are not reliable substitutes because both can differ from the release state.

QUESTION NO: 6

Which of the following standardizes scores similar to a percentile rank but preserves equal interval properties of a Z-score?

A. Trend analysis

B. Normal Curve Equivalent (NCE)

C. High Curve Equivalent (HCE)

D. Medium Curve Equivalent (MCE)

E. None of the above

ANSWER: B

Explanation:

Normal Curve Equivalent (NCE) is the correct standardized-score scale because it is designed to retain the intuitive interpretability of percentile-like reporting while providing equal-interval measurement characteristics derived from a normalized distribution. Percentile ranks indicate the percentage of scores at or below a given score, but the numerical distance between adjacent percentile ranks is not consistent across the distribution. NCE scores address this limitation by applying a normal-curve transformation and scaling the result, conventionally to a 1–99 range with a mean of 50 and a standard deviation of approximately 21.06. Consequently, equal differences in NCE values represent equal differences in relative standing on the underlying normalized scale, supporting statistical operations and comparisons that require interval-level measurements. This makes NCE scores useful when results must be communicated in a familiar rank-like scale but also analyzed with methods that benefit from the equal-interval behavior of z-scores. The National Center for Education Statistics describes normal curve equivalent scores and their relationship to normalized score distributions in its technical documentation. Background on z-score standardization and normal distributions is also available from NIST/SEMATECH.

QUESTION NO: 7

The Big Data Vision Workshop process is ideal for organizations who:

A. Have a desire to leverage Big Data to transform their business but do not know where and how to start

B. Have a wealth of data that they do not know how to monetize

C. Have a desire to leverage the Big Data Vision Workshop to identify where and how to leverage data and analytics to power their business models

D. Both A and B

E. All of the above

ANSWER: E

Explanation:

All of the above is correct because a Big Data Vision Workshop is intended to help an organization move from broad interest in data-driven transformation to a practical, business-aligned direction. It is particularly valuable when leaders recognize the potential of Big Data but have not yet determined the most valuable starting points, priorities, or implementation path. The workshop provides a structured setting for identifying high-value business opportunities, clarifying how available data can be converted into business value, and connecting analytics initiatives to the organization's operating and revenue models.

This scope also includes organizations that possess substantial data assets but have not established a clear monetization or value-realization strategy. Rather than treating data as a purely technical resource, the vision process focuses on the business outcomes it can support, such as improving decisions, creating data-enabled products and services, enhancing customer experiences, or optimizing operations. It therefore supports both early-stage strategic exploration and the identification of concrete ways to apply data and analytics to power business models. This business-first orientation is consistent with DASCA's focus on translating data-science and Big Data capabilities into organizational value; see DASCA and its Senior Data Scientist certification.

QUESTION NO: 8

A neural-network image classifier has excellent training accuracy but substantially worse validation accuracy. The training set is modest in size, and the model contains many layers and parameters. Select two changes that are most appropriate for reducing overfitting.

A. Apply label-preserving data augmentation to the training images

B. Introduce regularization such as dropout or weight decay

C. Add substantially more layers and parameters without increasing data

D. Continue training until training accuracy reaches 100 percent

E. Use the test set repeatedly to select the number of epochs

ANSWER: A B

Explanation:

Data augmentation can increase the effective diversity of the training examples through label-preserving transformations such as appropriate crops, flips, or small rotations. Regularization methods such as dropout or weight decay constrain the model and reduce its tendency to memorize idiosyncrasies in the training data. Both actions address the gap between training and validation performance.

Adding substantially more parameters generally increases capacity and can worsen overfitting when data is limited. Training until the model reaches perfect training accuracy optimizes the wrong objective; validation performance should guide stopping and model selection. Evaluating repeatedly on the test set contaminates the final unbiased assessment rather than improving generalization.

QUESTION NO: 9

A demand-forecasting team must predict weekly sales four weeks ahead. Historical data include a promotion field that is recorded only after a promotion has been approved, and some approvals occur after the forecast is issued. Select two practices that are appropriate for producing an honest evaluation of the forecasting model.

A. Use rolling-origin backtesting in which training data precede each forecast period

B. Exclude or reconstruct promotion information unavailable at the forecast-issue time

C. Randomly shuffle all weekly observations before creating training and test sets

D. Compute every feature using the complete historical period, including future weeks

E. Select the model with the smallest in-sample training error

ANSWER: A B

Explanation:

Evaluation should use rolling-origin or other time-ordered backtesting so that every training period precedes the period being forecast. The team must also restrict features to information genuinely available at the forecast-issue time. A promotion approval recorded after the forecast date is future information and would create leakage if used in training or evaluation.

Randomly shuffling weekly observations breaks the temporal forecasting setting and lets future observations influence training folds. Computing features with full-history information likewise leaks future knowledge. Choosing a model solely because it fits historical training sales most closely ignores performance on unseen future periods.

QUESTION NO: 10

An organization is moving operational records, clickstream events, documents, and images into a data lake for future BI and machine-learning use. Select two controls that are most important for making the lake a governed analytical asset rather than an unmanaged data repository.

A. Maintain a searchable catalog containing metadata and lineage

B. Enforce role-based access controls for sensitive datasets

C. Allow unrestricted write access so that ingestion is never delayed

D. Rely on schema-on-read instead of documenting data definitions

E. Use folder names as the only mechanism for classifying datasets

ANSWER: A B

Explanation:

A searchable catalog with metadata and lineage helps users discover data, understand its origin, and judge whether it is appropriate for an analytical use case. Role-based access controls protect sensitive data by limiting access according to approved business responsibilities. Schema-on-read permits flexible interpretation of data at analysis time, but it does not eliminate the need for metadata, quality controls, or governance. Allowing unrestricted writes and relying on folder names alone would reduce trust and increase the risk of duplicated, poorly understood, or improperly exposed data.