Vendor: Snowflake

Exam Code:DSA-C02

Exam Name: SnowPro Advanced: Data Scientist Certification

Version: Demo

QUESTION 1

Data providers add Snowflake objects (databases, schemas, tables, secure views, etc.) to a share us-ing. Which of the following options? Choose 2.

A. Grant privileges on objects to a share via Account role.

B. Grant privileges on objects directly to a share.

C. Grant privileges on objects to a share via a database role.

D. Grant privileges on objects to a share via a third-party role.

Correct Answer: BC

Explanation:What is a Share?

Shares are named Snowflake objects that encapsulate all of the information required to share a database.

Data providers add Snowflake objects (databases, schemas, tables, secure views, etc.) to a share using either or both of the following options:

Option 1: Grant privileges on objects to a share via a database role. Option 2: Grant privileges on objects directly to a share. You choose which accounts can consume data from the share by adding the accounts to the share.

After a database is created (in a consumer account) from a share, all the shared objects are accessible to users in the consumer account. Shares are secure, configurable, and controlled completely by the provider account:

New objects added to a share become immediately available to all consumers, providing real-time access to shared data.

Access to a share (or any of the objects in a share) can be revoked at any time.

QUESTION 2

Which one is not the feature engineering techniques used in ML data science world?

A. Imputation

B. Binning

C. One hot encoding

D. Statistical

Correct Answer: D

Explanation:

Feature engineering is the pre-processing step of machine learning, which is used to transform raw data into features that can be used for creating a predictive model using Machine learning or statistical Modelling.

What is a feature?

Generally, all machine learning algorithms take input data to generate the output. The input data re-mains in a tabular form consisting of rows (instances or observations) and columns (variable or at-tributes), and these attributes are often

known as features. For example, an image is an instance in computer vision, but a line in the image could be the feature. Similarly, in NLP, a document can be an observation, and the word count could be the feature. So, we can say a feature

is an attribute that impacts a problem or is useful for the problem.

What is Feature Engineering?

Feature engineering is the pre-processing step of machine learning, which extracts features from raw data. It helps to represent an underlying problem to predictive models in a better way, which as a result, improve the accuracy of the model

for unseen data. The predictive model contains predictor variables and an outcome variable, and while the feature engineering process selects the most useful predictor variables for the model. Some of the popular feature engineering

techniques include:

1.

Imputation Feature engineering deals with inappropriate data, missing values,human interruption, general errors, insufficient data sources, etc. Missing values within the dataset highly affect the performance of the algorithm, and to deal with them "Imputation" technique is used. Imputation is responsible for handling irregularities within the dataset. For example, removing the missing values from the complete row or complete column by a huge percentage of missing values. But at the same time, to maintain the data size, it is required to impute the missing data, which can be done as:

For numerical data imputation, a default value can be imputed in a column, and missing values can be filled with means or medians of the columns. For categorical data imputation, missing values can be interchanged with the maximum occurred value in a column.

2.

Handling Outliers

Outliers are the deviated values or data points that are observed too away from other data points in such a way that they badly affect the performance of the model. Outliers can be handled with this feature engineering technique. This

technique first identifies the outliers and then remove them out.

Standard deviation can be used to identify the outliers. For example, each value within a space has a definite to an average distance, but if a value is greater distant than acertain value, it can be considered as an outlier. Z-score can also be

used to detect outliers.

3.

Log transform

Logarithm transformation or log transform is one of the commonly used mathematical techniques in machine learning. Log transform helps in handling the skewed data, and it makes the distribution more approximate to normal after

transformation. It also reduces the effects of outliers on the data, as because of the normalization of magnitude differences, a model becomes much robust.

4.

Binning

In machine learning, overfitting is one of the main issues that degrade the performance of the model and which occurs due to a greater number of parameters and noisydata. However, one of the popular techniques of feature engineering,

"binning", can be used to normalize the noisy data. This process involves segmenting different features into bins.

5.

Feature Split

As the name suggests, feature split is the process of splitting features intimately into two or more parts and performing to make new features. This technique helps the algorithms to better understand and learn the patterns in the dataset. The

feature splitting process enables the new features to be clustered and binned, which results in extracting useful information and improving the performance of the data models.

6.

One hot encoding

One hot encoding is the popular encoding technique in machine learning. It is a technique that converts the categorical data in a form so that they can be easily understood by machine learning algorithms and hence can make a good

prediction. It enables group theof categorical data without losing any information.

QUESTION 3

To return the contents of a DataFrame as a Pandas DataFrame, Which of the following method can be used in SnowPark API?

A. REPLACE_TO_PANDAS

B. SNOWPARK_TO_PANDAS

C. CONVERT_TO_PANDAS

D. TO_PANDAS

Correct Answer: D

Explanation:

To return the contents of a DataFrame as a Pandas DataFrame, use the to_pandas method.

For example:

1.>>> python_df = session.create_dataframe(["a", "b", "c"]) 2.>>> pandas_df = python_df.to_pandas()

QUESTION 4

Skewness of Normal distribution is ___________

A. Negative

B. Positive

C. 0

D. Undefined

Correct Answer: C

Explanation:

Since the normal curve is symmetric about its mean, its skewness is zero. This is a theoretical explanation for mathematical proofs, you can refer to books or websites that speak on the same in detail.

QUESTION 5

Which ones are the known limitations of using External function? Choose all apply.

A. Currently, external functions cannot be shared with data consumers via Secure Data Sharing.

B. Currently, external functions must be scalar functions. A scalar external function re-turns a single value for each input row.

C. External functions have more overhead than internal functions (both built-in functions and internal UDFs) and usually execute more slowly

D. An external function accessed through an AWS API Gateway private endpoint can be accessed only from a Snowflake VPC (Virtual Private Cloud) on AWS and in the same AWS region.

Correct Answer: ABCD

QUESTION 6

There are a couple of different types of classification tasks in machine learning, Choose the Correct Classification which best categorized the below Application Tasks in Machine learning?

To detect whether email is spam or not

To determine whether or not a patient has a certain disease in medicine.

To determine whether or not quality specifications were met when it comes to QA (Quality Assurance).

A. Multi-Label Classification

B. Multi-Class Classification

C. Binary Classification

D. Logistic Regression

Correct Answer: C

Explanation: The Supervised Machine Learning algorithm can be broadly classified into Regression and Classification Algorithms. In Regression algorithms, we have predicted the output for continuous values, but to predict the categorical

values, we need Classification algorithms.

What is the Classification Algorithm?

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then

classifies new observation into a number of classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called as targets/labels or categories. Unlike regression, the output variable of Classification is a

category, not a value, such as "Green or Blue", "fruit or animal", etc. Since the Classification algorithm is a Supervised learning technique, hence it takes labeled input data, which means it contains input with the corresponding output.

In classification algorithm, a discrete output function(y) is mapped to input variable(x).

y=f(x), where y = categorical output

The best example of an ML classification algorithm is Email Spam Detector. The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the

categorical data. The algorithm which implements the classification on a dataset is known as a classifier.

There are two types of Classifications:

Binary Classifier: If the classification problem has only two possible outcomes, then it is called as Binary Classifier.

Examples: YES or NO, MALE or FEMALE, SPAM or NOT SPAM, CAT or DOG, etc. Multi-class Classifier: If a classification problem has more than two outcomes, then it is called as Multi-class Classifier.

Example: Classifications of types of crops, Classification of types of music. Binary classification in deep learning refers to the type of classification where we have two class labels - one normal and one abnormal. Some examples of binary

classification use:

To detect whether email is spam or not

To determine whether or not a patient has a certain disease in medicine.

To determine whether or not quality specifications were met when it comes to QA (Quality Assurance).

For example, the normal class label would be that a patient has the disease, and the abnormal class label would be that they do not, or vice-versa. As is with every other type of classification, it is only as good as the binary classification

dataset that it has ?or, in other words, the more training and data it has, the better it is.

QUESTION 7

In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?

A. by 1

B. no change

C. by intercept

D. by its slope

Correct Answer: D

Explanation:

What is linear regression?

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable\\'s value is

called the independent variable.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatoryvariable, and the other is considered to be a dependent variable. For

example, a modeler might want to relate the weights of individuals to their heights using a linear regression model. A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent

variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

For linear regression Y=a+bx+error.

If neglect error then Y=a+bx. If x increases by 1, then Y = a+b(x+1) which implies Y=a+bx+b. So Y increases by its slope.

For linear regression Y=a+bx+error. If neglect error then Y=a+bx. If x increases by 1, then Y = a+b(x+1) which implies Y=a+bx+b. So Y increases by its slope.

QUESTION 8

Which object records data manipulation language (DML) changes made to tables, including inserts, updates, and deletes, as well as metadata about each change, so that actions can be taken using the changed data of Data Science Pipelines?

A. Task

B. Dynamic tables

C. Stream

D. Tags

E. Delta

F. OFFSET

Correct Answer: C

Explanation: A stream object records data manipulation language (DML) changes made to tables, including inserts, updates, and deletes, as well as metadata about each change,so that actions can be taken using the changed data. This process is referred to as change data capture (CDC). An individual table stream tracks the changes made to rows in a source table. A table stream (also referred to as simply a "stream") makes a "change table" available of what changed, at therow level, between two transactional points of time in a table. This allows querying and consuming a sequence of change records in a transactional fashion. Streams can be created to query change data on the following objects: Standard tables, including shared tables. Views, including secure views Directory tables Event tables

---

QUESTION 9

The most widely used metrics and tools to assess a classification model are:

A. Confusion matrix

B. Cost-sensitive accuracy

C. Area under the ROC curve

D. All of the above

Correct Answer: D

QUESTION 10

Select the Data Science Tools which are known to provide native connectivity to Snowflake?

A. Denodo

B. DvSUM

C. DiYotta

D. HEX

Correct Answer: D

Explanation:

Hex -- collaborative data science and analytics platform Denodo -- data virtualization and federation platform DvSum -- data catalog and data intelligence platform Diyotta -- data integration and migration

---

QUESTION 11

Which metric is not used for evaluating classification models?

A. Recall

B. Accuracy

C. Mean absolute error

D. Precision

Correct Answer: C

Explanation:

The four commonly used metrics for evaluating classifier performance are:

1.

Accuracy: The proportion of correct predictions out of the total predictions.

2.

Precision: The proportion of true positive predictions out of the total positive predictions (precision = true positives / (true positives + false positives)).

3.

Recall (Sensitivity or True Positive Rate): The proportion of true positive predictions out of the total actual positive instances (recall = true positives / (true positives + false negatives)).

4.

F1 Score: The harmonic mean of precision and recall, providing a balance between the two metrics (F1 score = 2 * ((precision * recall) / (precision + recall))). Root Mean Squared Error (RMSE)and Mean Absolute Error (MAE) are metrics used to evaluate a Regression Model. These metrics tell us how accurate our predictions are and, what is the amount of deviation from the actual values.

QUESTION 12

Which command manually triggers a single run of a scheduled task (either a standalone task or the root task in a DAG) independent of the schedule defined for the task?

A. RUN TASK

B. CALL TASK

C. EXECUTE TASK

D. RUN ROOT TASK

Correct Answer: C

Explanation: The EXECUTE TASK command manually triggers a single run of a scheduled task (either a standalone task or the root task in a DAG) independent of the schedule defined for the task. A successful run of a roottask triggers a cascading run of child tasks in the DAG as their precedent task completes, as though the root task had run on its defined schedule. This SQL command is useful for testing new or modified standalone tasks and DAGs before you enable them to execute SQL code in production. Call this SQL command directly in scripts or in stored procedures. In addition, this command sup-ports integrating tasks in external data pipelines. Any third-party services that can authenticate into your Snowflake account and authorize SQL actions can execute the EXECUTE TASK command to run tasks.