DSA-C02$^{Q\&As}$

SnowPro Advanced: Data Scientist Certification

Pass Snowflake DSA-C02 Exam with 100% Guarantee

Following Questions and Answers are all new published by Snowflake Official Exam Center

QUESTION 1

Which tools helps data scientist to manage ML lifecycle and Model versioning? Choose 2.

A. MLFlow

B. Pachyderm

C. Albert

D. CRUX

Correct Answer: AB

Explanation:

Model versioning in a way involves tracking the changes made toan ML model that has been previously built. Put differently, it is the process of making changes to the configurations of an ML Model. From another perspective, we can see

model versioning as a feature that helps Machine Learning Engineers, Data Scientists, and related personnel create and keep multiple versions of the same model. Think of it as a way of taking notes of the changes you make to the model

through tweaking hyperparameters, retraining the model with more data, and so on. In model versioning, a number of things need to be versioned, to help us keep track of important changes. I\\'ll list and explain them below:

Implementation code: From the early days of model building to optimization stages, code or in this case source code of the model plays an important role. This code experiences significant changes during optimization stages which can easily

be lost if not tracked properly. Because of this, code is one of the things that are taken into consideration during the model versioning process.

Data: In some cases, training data does improve significantly from its initial state during model op-timization phases. This can be as a result of engineering new features from existing ones to train our model on. Also there is metadata (data

about your training data and model) to consider versioning. Metadata can change different times over without the training data actually changing. We need to be able to track these changes through versioning

Model: The model is a product of the two previous entities and as stated in their explanations, an ML model changes at different points of the optimization phases through hyperparameter setting, model artifacts and learning coefficients.

Versioning helps take record of the different versions of a Machine Learning model. MLFlow and Pachyderm are the tools used to manage ML lifecycle and Model versioning.

QUESTION 2

What is the risk with tuning hyper-parameters using a test dataset?

A. Model will overfit the test set

B. Model will underfit the test set

C. Model will overfit the training set

D. Model will perform balanced

Correct Answer: A

Explanation:

The model will not generalize well to unseen data because it overfits the test set. Tuning model hyper-parameters to a test set means that the hyper-parameters may overfit to that test set. If the same test set is used to estimate performance,

it will produce an overestimate. The test set should be used only for testing, not for parameter tuning. Using a separate validation set for tuning and test set for measuring performance provides unbiased, realistic measurement of

performance.

What are hyper-parameters?

Hyper-parameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. We can\\'t calculate their values from the data.

Example: Number of clusters in clustering, number of hidden layers in a neural network, and depth of a tree are some of the examples of hyper-parameters.

What is the hyper-parameter tuning?

Hyper-parameter tuning is the process of choosing the right combination of hyper- parameters that maximizes the model performance. It works by running multiple trials in a single training process. Each trial is a complete execution of your

training application with values for your chosen hyper-parameters, set within the limits you specify. This process once finished will give you the set of hyper-parameter values that are best suited for the model to give optimal results.

QUESTION 3

Which ones are the type of visualization used for Data exploration in Data Science? Choose 3.

A. Heat Maps

B. Newton AI

C. Feature Distribution by Class

D. 2D-Density Plots

E. Sand Visualization

Correct Answer: ADE

Explanation:

Type of visualization used for exploration:

Correlation heatmap

Class distributions by feature

Two-Dimensional density plots.

All the visualizations are interactive, as is standard for Plotly.

For More details, please refer the below link:

https://towardsdatascience.com/data-exploration-understanding-and-visualization- 72657f5eac41

QUESTION 4

What Can Snowflake Data Scientist do in the Snowflake Marketplace as Provider? Choose all apply.

A. Publish listings for free-to-use datasets to generate interest and new opportunities among the Snowflake customer base.

B. Publish listings for datasets that can be customized for the consumer.

C. Share live datasets securely and in real-time without creating copies of the data or im- posing data integration tasks on the consumer.

D. Eliminate the costs of building and maintaining APIs and data pipelines to deliver data to customers.

Correct Answer: ABCD

Explanation:

All are correct!

About the Snowflake Marketplace

You can use the Snowflake Marketplace to discover and access third-party data and services, as well as market your own data products across the Snowflake Data Cloud. As a data provider, you can use listings on the Snowflake

Marketplace to share curated data offer-ings with many consumers simultaneously, rather than maintain sharing relationships with each indi-vidual consumer. With Paid Listings, you can also charge for your data products.

As a consumer, you might use the data provided on the Snowflake Marketplace to explore and ac-cess the following:

Historical data for research, forecasting, and machine learning. Up-to-date streaming data, such as current weather and traffic conditions. Specialized identity data for understanding subscribers and audience targets.

New insights from unexpected sources of data.

The Snowflake Marketplace is available globally to all non-VPS Snowflake accounts hosted on Amazon Web Services, Google Cloud Platform, and Microsoft Azure, with the exception of Mi-crosoft Azure Government. Support for Microsoft

Azure Government is planned.

QUESTION 5

You are training a binary classification model to support admission approval decisions for a college degree program.

How can you evaluate if the model is fair, and doesn\\'t discriminate based on ethnicity?

A. Evaluate each trained model with a validation datasetand use the model with the highest accuracy score.

B. Remove the ethnicity feature from the training dataset.

C. Compare disparity between selection rates and performance metrics across ethnicities.

D. None of the above.

Correct Answer: C

Explanation:

By using ethnicity as a sensitive field, and comparing disparity between selection rates and performance metrics for each ethnicity value, you can evaluate the fairness of the model.

QUESTION 6

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 7

Which of the following is a Python-based web application framework for visualizing data and analyzing results in a more efficient and flexible way?

A. StreamBI

B. Streamlit

C. Streamsets

D. Rapter

Correct Answer: B

Explanation:

Streamlit is a Python-based web application framework for visualizing data and analyzing results in a more efficient and flexible way. It is an open source library that assists data scientists and academics to develop Machine Learning (ML)

visualization dashboards in a short period of time. We can build and deploy powerful data applications with just a few lines of code.

Why Streamlit?

Currently, real-world applications are in high demand and developers are developing new libraries and frameworks to make on-the-go dashboards easier to build and deploy. Streamlit is a library that reduces your dashboard development

time from days to hours. Following are some reasons to choose the Streamlit:

It is a free and open-source library.

Installing Streamlit is as simple as installing any other python package It is easy to learn because you won\\'t need any web development experience, only a basic under-standing of Python is enough to build a data application. It is compatible

with almost all machine learning frameworks, including Tensorflow and Pytorch, Scikit-learn, and visualization libraries such as Seaborn, Altair, Plotly, and many others.

QUESTION 8

Which command is used to install Jupyter Notebook?

A. pip install jupyter

B. pip install notebook

C. pip install jupyter-notebook

D. pip install nbconvert

Correct Answer: A

Explanation:

Jupyter Notebook is a web-based interactive computational environment. The command used to install Jupyter Notebook is pip install jupyter. The command used to start Jupyter Notebook is jupyter notebook.

QUESTION 9

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 10

Select the Correct Statements regarding Normalization? Choose 2.

A. Normalization technique uses minimum and max values for scaling of model.

B. Normalization technique uses mean and standard deviation for scaling of model.

C. Scikit-Learn provides a transformer RecommendedScaler for Normalization.

D. Normalization got affected by outliers.

Correct Answer: AD

Explanation:

Normalization is a scaling technique in Machine Learning applied during data preparation to change the values of numeric columns in the dataset to use a common scale.It is not necessary for all datasets in a model. It is required only when

features of machine learning models have different ranges.

Scikit-Learn provides a transformer called MinMaxScaler for Normalization. This technique uses minimum and max values for scaling of model.Itis useful when feature distribution is unknown.It got affected by outliers.

QUESTION 11

Which of the following process best covers all of the following characteristics?

Collecting descriptive statistics like min, max, count and sum.

Collecting data types, length and recurring patterns. ?Tagging data with keywords, descriptions or categories.

Performing data quality assessment, risk of performing joins on the data.

Discovering metadata and assessing its accuracy.

Identifying distributions, key candidates, foreign-key candidates,functional dependencies, embedded value dependencies, and performing inter-table analysis.

A. Data Visualization

B. Data Virtualization

C. Data Profiling

D. Data Collection

Correct Answer: C

Explanation:

Data processing and analysis cannot happen without data profiling--reviewing source data for con-tent and quality. As data gets bigger and infrastructure moves to the cloud, data profiling is increasingly important.

What is data profiling?

Data profiling is the process of reviewing source data, understanding structure, content and interrelationships, and identifying potential for data projects.

Data profiling is a crucial part of:

Data warehouse and business intelligence (DW/BI) projects--dataprofiling can uncover data quality issues in data sources, and what needs to be corrected in ETL.

Data conversion and migration projects--data profiling can identify data quality issues, which you can handle in scripts and data integration tools copying data from source to target. It can also un-cover new requirements for the target system.

Source system data quality projects--data profiling can highlight data which suffers from serious or numerous quality issues, and the source of the issues (e.g. user inputs, errors in interfaces, data corruption).

Data profiling involves:

Collecting descriptive statistics like min, max, count and sum.

Collecting data types, length and recurring patterns.

Tagging data with keywords, descriptions or categories.

Performing data quality assessment, risk of performing joins on the data.

Discovering metadata and assessing its accuracy.

Identifying distributions, key candidates, foreign-key candidates, functional dependencies, embedded value dependencies, and performing inter-table analysis.

QUESTION 12

Which of the following method is used for multiclass classification?

A. one vs rest

B. loocv

C. all vs one

D. one vs another

Correct Answer: A

Explanation: Binary vs. Multi-Class Classification Classification problems are common in machine learning. In most cases, developers prefer using a supervised machine-learning approach to predict class tables for a given dataset. Unlike regression, classification involves designing the classifier model and training it to input and categorize the test dataset. For that, you can divide the dataset into either binary or multi-class modules. As the name suggests, binary classification involves solving a problem with only two class labels. This makes it easy to filter the data, apply classification algorithms, and train the model to predict outcomes. On the other hand, multi-class classification is applicable when there are more than two class labels in the input train data. The technique enables developers to categorize the test data into multiple binary class labels. That said, while binary classification requires only one classifier model, the one used in the multi-class approach depends on the classification technique. Below are the two models of the multi-class classification algorithm. One-Vs-Rest Classification Model for Multi-Class Classification Also known as one-vs-all, the one-vs-rest model is a defined heuristic method that leverages a binary classification algorithm for multi-class classifications. The technique involves splitting a multi-class dataset into multiple sets of binary problems. Following this, a binary classifier is trained to handle each binary classification model with the most confident one making predictions. For instance, with a multi-class classification problem with red, green, and blue datasets, binary classification can be categorized as follows: Problem one: red vs. green/blue Problem two: blue vs. green/red Problem three: green vs. blue/red The only challenge of using this model is that you should create a model for every class. The three classes require three models from the above datasets, which can be challenging for large sets of data with million rows, slow models, such as neural networks and datasets with a significant number of classes. The one-vs-rest approach requires individual models to prognosticate the probability-like score. The class index with the largest score is then used to predict a class. As such, it is commonly used forclassification algorithms that can naturally predict scores or numerical class membership such as perceptron and logistic regression.

QUESTION 13

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 14

Mark the Incorrect understanding of Data Scientist about Streams? Choose 2.

A. Streams on views support both local views and views shared using Snowflake Secure Data Sharing, including secure views.

B. Streams can track changes in materialized views.

C. Streams itself does not contain any table data.

D. Streams do not support repeatable read isolation.

Correct Answer: BD

Explanation: Streams on views support both local views and views shared using Snowflake Secure Data Sharing, including secure views. Currently, streams cannot track changes in materialized views. stream itself does not contain any table data. A stream only stores an offset for the source object and returns CDC records by leveraging the versioning history for the source object. When the first stream for a table is created, several hidden columns are added to the source table and begin storing change tracking metadata. These columns consume a small amount of storage. The CDC records returned when querying a stream rely on a combination of the offset stored in the stream and the change tracking metadata stored in the table. Note that for streams on views, change tracking must be enabled explicitly for the view and underlying tables to add the hidden columns to these tables. Streams support repeatable read isolation. In repeatable read mode, multiple SQL statements within a transaction see the same set of records in a stream. This differs from the read committed mode supported for tables, in which statements see any changes made by previous statements executed within the same transaction, even though those changes are not yet committed. The delta records returned by streams in a transaction is the range from the current position of the stream until the transaction start time. The stream position advances to the transaction start time if the transaction commits; otherwise it stays at the same position.

QUESTION 15

Consider a data frame df with 10 rows and index [\\'r1\\', \\'r2\\', \\'r3\\', \\'row4\\', \\'row5\\', \\'row6\\', \\'r7\\', \\'r8\\', \\'r9\\', \\'row10\\']. What does the expression g = df.groupby(df.index.str.len()) do?

A. Groups df based on index values

B. Groups df based on length of each index value

C. Groups df based on index strings

D. Data frames cannot be grouped by index values. Hence it results in Error.

Correct Answer: D

Explanation: Data frames cannot be grouped by index values. Hence it results in Error.