DATABRICKS-MACHINE-LEARNING-ASSOCIATE Q&As

Databricks Certified Machine Learning Associate Exam

Pass Databricks DATABRICKS-MACHINE-LEARNING-ASSOCIATE Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.geekcert.com/databricks-machine-learning-associate.html

100% Passing Guarantee
100% Money Back Assurance

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

QUESTION 1

A machine learning engineer has created a Feature Table new_table using Feature Store Client fs. When creating the table, they specified a metadata description with key information about the Feature Table. They now want to retrieve that metadata programmatically.

Which of the following lines of code will return the metadata description?

A. There is no way to return the metadata description programmatically.

B. fs.create_training_set("new_table")

C. fs.get_table("new_table").description

D. fs.get_table("new_table").load_df()

E. fs.get_table("new_table")

Correct Answer: C

To retrieve the metadata description of a feature table created using the Feature Store Client (referred here asfs), the correct method involves callingget_tableon thefsclient with the table name as an argument, followed by accessing

thedescription attribute of the returned object. The code snippetfs.get_table("new_table").description correctly achieves this by fetching the table object for "new_table" and then accessing its description attribute, where the metadata is stored.

The other options do not correctly focus on retrieving the metadata description.References:

Databricks Feature Store documentation (Accessing Feature Table Metadata).

QUESTION 2

A machine learning engineer wants to parallelize the inference of group-specific models using the Pandas Function API. They have developed theapply_modelfunction that will look up and load the correct model for each group, and they want to apply it to each group of DataFramedf.

They have written the following incomplete code block:

```
prediction_df = (df
    .groupby("device_id")
    ._____(apply_model, schema=apply_return_schema)
)
```

Which piece of code can be used to fill in the above blank to complete the task?

A. applyInPandas

B. groupedApplyInPandas

C. mapInPandas

D. predict

Correct Answer: A

To parallelize the inference of group-specific models using the Pandas Function API in PySpark, you can use theapplyInPandasfunction. This function allows you to apply a Python function on each group of a DataFrame and return a

DataFrame, leveraging the power of pandas UDFs (user-defined functions) for better performance.

prediction_df = (df.groupby("device_id") .applyInPandas(apply_model, schema=apply_return_schema))

In this code:

groupby("device_id"): Groups the DataFrame by the "device_id" column. applyInPandas(apply_model, schema=apply_return_schema): Applies theapply_modelfunction to each group and specifies the schema of the return DataFrame.

References:

PySpark Pandas UDFs Documentation

QUESTION 3

A data scientist has created two linear regression models. The first model uses price as a label variable and the second model uses log(price) as a label variable. When evaluating the RMSE of each model bycomparing the label predictions to the actual price values, the data scientist notices that the RMSE for the second model is much larger than the RMSE of the first model.

Which of the following possible explanations for this difference is invalid?

A. The second model is much more accurate than the first model

B. The data scientist failed to exponentiate the predictions in the second model prior tocomputingthe RMSE

C. The datascientist failed to take the logof the predictions in the first model prior to computingthe RMSE

D. The first model is much more accurate than the second model

E. The RMSE is an invalid evaluation metric for regression problems

Correct Answer: E

The Root Mean Squared Error (RMSE) is a standard and widely used metric for evaluating the accuracy of regression models. The statement that it is invalid is incorrect. Here\\'s a breakdown of why the other statements are or are not valid:

Transformations and RMSE Calculation:If the model predictions were transformed (e.g., using log), they should be converted back to their original scale before calculating RMSE to ensure accuracy in the evaluation. Missteps in this

conversion process can lead to misleading RMSE values. Accuracy of Models:Without additional information, we can\\'t definitively say which model is more accurate without considering their RMSE values properly scaled back to the original

price scale.

Appropriateness of RMSE:RMSE is entirely valid for regression problems as it provides a measure of how accurately a model predicts the outcome, expressed in the same units as the dependent variable.

References:

"Applied Predictive Modeling" by Max Kuhn and Kjell Johnson (Springer, 2013), particularly the chapters discussing model evaluation metrics.

QUESTION 4

A data scientist is wanting to explore the Spark DataFrame spark_df. The data scientist wants visual histograms displaying the distribution of numeric features to be included in the exploration.

Which of the following lines of code can the data scientist run to accomplish the task?

A. spark_df.describe()

B. dbutils.data(spark_df).summarize()

C. This task cannot be accomplished in a single line of code.

D. spark_df.summary()

E. dbutils.data.summarize (spark_df)

Correct Answer: E

To display visual histograms and summaries of the numeric features in a Spark DataFrame, the Databricks utility functiondbutils.data.summarizecan be used. This function provides a comprehensive summary, including visual histograms.

Correct code:

dbutils.data.summarize(spark_df)

Other options likespark_df.describe()andspark_df.summary()provide textual statistical summaries but do not include visual histograms.

References:

Databricks Utilities Documentation

QUESTION 5

What is the name of the method that transforms categorical features into a series of binary indicator feature variables?

A. Leave-one-out encoding

B. Target encoding

C. One-hot encoding

D. Categorical

E. String indexing

Correct Answer: C

The method that transforms categorical features into a series of binary indicator variables is known as one-hot encoding. This technique converts each categorical value into a new binary column, which is essential for models that require

numerical input. One-hot encoding is widely used because it helps to handle categorical data without introducing a false ordinal relationship among categories.References:

Feature Engineering Techniques (One-Hot Encoding).