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

Which of the following hyperparameter optimization methods automatically makes informed selections of hyperparameter values based on previous trials for each iterative model evaluation?

A. Random Search

B. Halving Random Search

C. Tree of Parzen Estimators

D. Grid Search

Correct Answer: C

Tree of Parzen Estimators (TPE) is a sequential model-based optimization algorithm that selects hyperparameter values based on the outcomes of previous trials. It models the probability density of good and bad hyperparameter values and

makes informed decisions about which hyperparameters to try next. This approach contrasts with methods like random search and grid search, which do not use information from previous trials to guide the search process.

References:

Hyperopt and TPE

QUESTION 2

A machine learning engineer wants to parallelize the training of group-specific models using the Pandas Function API. They have developed thetrain_modelfunction, and they want to apply it to each group of DataFramedf.

They have written the following incomplete code block:

```
model_directories_df = (df
    .withColumn("run_id", f.lit(run_id))
    .groupby("device_id")
    .____(train_model, schema=train_return_schema)
)
```

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

A. applyInPandas

B. mapInPandas

C. predict

D. train_model

E. groupedApplyIn

Correct Answer: B

The functionmapInPandasin the PySpark DataFrame API allows for applying a function to each partition of the DataFrame. When working with grouped data,groupbyfollowed by applyInPandasis the correct approach to apply a function to each group as a separate Pandas DataFrame. However, if the function should apply across each partition of the grouped data rather than on each individual group,mapInPandaswould be utilized. Since the code snippet indicates the use ofgroupby, the intent seems to be to applytrain_model on each group specifically, which aligns withapplyInPandas. Thus,applyInPandasis a better fit to ensure that each group generated bygroupbyis processed through the train_modelfunction, preserving the partitioning and grouping integrity. References: PySpark Documentation on applying functions to grouped data:https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.GroupedData.applyInPandas.html

QUESTION 3

In which of the following situations is it preferable to impute missing feature values with their median value over the mean value?

A. When the features are of the categorical type

B. When the features are of the boolean type

C. When the features contain a lot of extreme outliers

D. When the features contain no outliers

E. When the features contain no missingno values

Correct Answer: C

Imputing missing values with the median is often preferred over the mean in scenarios where the data contains a lot of extreme outliers. The median is a more robust measure of central tendency in such cases, as it is not as heavily

influenced by outliers as the mean. Using the median ensures that the imputed values are more representative of the typical data point, thus preserving the integrity of the dataset\\'s distribution. The other options are not specifically relevant to

the question of handling outliers in numerical data.

References:

Data Imputation Techniques (Dealing with Outliers).

QUESTION 4

A data scientist has replaced missing values in their feature set with each respective feature variable\\'s median value. A colleague suggests that the data scientist is throwing away valuable information by doing this.

Which of the following approaches can they take to include as much information as possible in the feature set?

A. Impute the missing values using each respective feature variable\\'s mean value instead of the median value

B. Refrain from imputing the missing values in favor of letting the machine learning algorithm determine how to handle them

C. Remove all feature variables that originally contained missing values from the feature set

D. Create a binary feature variable for each feature that contained missing values indicating whether each row\\'s value has been imputed

E. Create a constant feature variable for each feature that contained missing values indicating the percentage of rows from the feature that was originally missing

Correct Answer: D

By creating a binary feature variable for each feature with missing values to indicate whether a value has been imputed, the data scientist can preserve information about the original state of the data. This approach maintains the integrity of

the dataset by marking which values are original and which are synthetic (imputed). Here are the steps to implement this approach:

Identify Missing Values:Determine which features contain missing values. Impute Missing Values:Continue with median imputation or choose another method (mean, mode, regression, etc.) to fill missing values. Create Indicator Variables:For

each feature that had missing values, add a new binary feature. This feature should be \\'1\\' if the original value was missing and imputed, and \\'0\\' otherwise.

Data Integration:Integrate these new binary features into the existing dataset. This maintains a record of where data imputation occurred, allowing models to potentially weight these observations differently. Model Adjustment:Adjust machine

learning models to account for these new features, which might involve considering interactions between these binary indicators and other features.

References:

"Feature Engineering for Machine Learning" by Alice Zheng and Amanda Casari (O\\'Reilly Media, 2018), especially the sections on handling missing data. Scikit-learn documentation on imputing missing values: https://scikit-learn.org/stable/

modules/impute.html

QUESTION 5

A machine learning engineer is trying to scale a machine learning pipeline by distributing its feature engineering process.

Which of the following feature engineering tasks will be the least efficient to distribute?

A. One-hot encoding categorical features

B. Target encoding categorical features

C. Imputing missing feature values with the mean

D. Imputing missing feature values with the true median

E. Creating binary indicator features for missing values

Correct Answer: D

Among the options listed, calculating the true median for imputing missing feature values is the least efficient to distribute. This is because the true median requires knowledge of the entire data distribution, which can be computationally

expensive in a distributed environment. Unlike mean or mode, finding the median requires sorting the data or maintaining a full distribution, which is more intensive and often requires shuffling the data across partitions.

References:

Challenges in parallel processing and distributed computing for data aggregation like median calculation:https://www.apache.org