Question #1 *Topic 1*

Which of the following describes concept drift?

A. Concept drift is when there is a change in the distribution of an input variable

B. Concept drift is when there is a change in the distribution of a target variable

C. Concept drift is when there is a change in the relationship between input variables and target variables

D. Concept drift is when there is a change in the distribution of the predicted target given by the model

E. None of these describe Concept drift

Suggested Answer: *D*

Community vote distribution

C (100%)

BokNinja Highly Voted 2 months ago

C. Concept drift is when there is a change in the relationship between input variables and target variables.

Concept drift refers to the situation when the statistical properties of the target variable, which the model is trying to predict, change over time in unforeseen ways. This causes the model to become less accurate as time passes.

upvoted 6 times

Oreocake Most Recent 4 months, 1 week ago

Selected Answer: C

C is correct

upvoted 1 times

zafarsohaib 5 months, 1 week ago

Answer is C. Concept drift is a change in the relationship between the input data and the model target.

upvoted 1 times

sindhu_gowda 5 months, 2 weeks ago

Selected Answer: C

Answer is C

upvoted 1 times

c4b65b5 5 months, 2 weeks ago

Selected Answer: C

C is correct

upvoted 1 times

nnn_666 8 months, 3 weeks ago

Selected Answer: C

concept drift or drift is an evolution of data that invalidates the data model

upvoted 2 times

ADVIT 9 months ago

Selected Answer: C

Answer is C

upvoted 2 times

hugodscarvalho 10 months ago

Selected Answer: C

C

upvoted 2 times

random_data_guy 11 months ago

Selected Answer: C

agree with BokNinja - C should be correct

upvoted 2 times

Question #2 *Topic 1*

A machine learning engineer is monitoring categorical input variables for a production machine learning application. The engineer believes that missing values are becoming more prevalent in more recent data for a particular value in one of the categorical input variables.
Which of the following tools can the machine learning engineer use to assess their theory?

A. Kolmogorov-Smirnov (KS) test

B. One-way Chi-squared Test

C. Two-way Chi-squared Test

D. Jenson-Shannon distance

E. None of these

Suggested Answer: *B*

Community vote distribution

B (100%)

fshb4ztm 4 weeks ago

It seems like we want to determine if there's a statistical difference in one direction. Ex: between last month's data and new data, particularly observing more blanks for a specific categorical variable. It seems like a one-way Chi-squared Test is the most appropriate, what do you say?

upvoted 1 times

e12ec59 3 months, 3 weeks ago

Selected Answer: B

As the question is about one input variable, so my answer would be B.

upvoted 1 times

ThoBustos 6 months, 3 weeks ago

It seems like we want to determine if there's a statistical difference in one direction. Ex: between last month's data and new data, particularly observing more blanks for a specific categorical variable. It seems like a one-way Chi-squared Test is the most appropriate, what do you say?

upvoted 1 times

Alishahab70 9 months, 2 weeks ago

C. Two-way Chi-squared Test

This tool can help the engineer determine if there is a statistically significant association between the time period and the presence of missing values in the categorical variable.

upvoted 1 times

ldoyle3332 10 months ago

Would the answer not be C?

One-Way Chi-Squared Tests will pick up changes in the overall count of null values as a shift in distribution, even if the nulls are distributed evenly among the category values.

The question specifies that they are looking for a shift in a particular class value distribution, which would be better for Two-Way Chi-Squared Test

upvoted 1 times

hugodscarvalho 10 months ago

Selected Answer: B

Since it's just one categorical input variable over time, the one-way Chi-squared test would be the more appropriate choice.

upvoted 1 times

BokNinja 11 months, 1 week ago

The correct answer is B. One-way Chi-squared Test.

The Chi-squared test is a statistical hypothesis test that is used to determine whether there is a significant association between two categorical variables. In this case, the two variables could be the presence (or absence) of a value and the time period (old data vs. new data).

upvoted 1 times

Question #3 *Topic 1*

A data scientist is using MLflow to track their machine learning experiment. As a part of each MLflow run, they are performing hyperparameter tuning. The data scientist would like to have one parent run for the tuning process with a child run for each unique combination of hyperparameter values.

They are using the following code block:

```
with mlflow.start_run(run_name="Parent run") as run:
    print("Start parent run")
with mlflow.start_run(run_name="Child 1", nested=True):
    mlflow.log_param("run_name", "child_1")
with mlflow.start_run(run_name="Child 2", nested=True):
    mlflow.log_param("run_name", "child_2")
```

The code block is not nesting the runs in MLflow as they expected.

Which of the following changes does the data scientist need to make to the above code block so that it successfully nests the child runs under the parent run in MLflow?

A. Indent the child run blocks within the parent run block

B. Add the nested=True argument to the parent run

C. Remove the nested=True argument from the child runs

D. Provide the same name to the run_name parameter for all three run blocks

E. Add the nested=True argument to the parent run and remove the nested=True arguments from the child runs

Suggested Answer: *E*

Community vote distribution

zafarsohaib 5 months, 1 week ago

A should be the correct answer.

upvoted 1 times

sindhu_gowda 5 months, 2 weeks ago

A is correct

upvoted 1 times

c4b65b5 5 months, 2 weeks ago

Selected Answer: A

A is correct

upvoted 1 times

hugodscarvalho 10 months ago

Selected Answer: A

A

upvoted 1 times

random_data_guy 11 months ago

Selected Answer: A

A should be correct - see https://mlflow.org/docs/latest/python_api/mlflow.html#mlflow.start_run for an example (where nested=True is added to children runs + they have different names)

upvoted 1 times

mozuca 11 months ago

Selected Answer: A

The answer is A since the correct sintax is:

with mlflow.start_run(run_name="Nested Example") as run:
Create nested run with nested=True argument
with mlflow.start_run(run_name="Child 1", nested=True):
mlflow.log_param("run_name", "child_1")

with mlflow.start_run(run_name="Child 2", nested=True):
mlflow.log_param("run_name", "child_2")

upvoted 1 times

BokNinja 11 months, 1 week ago

A. Indent the child run blocks within the parent run block

upvoted 1 times

Question #4 *Topic 1*

A machine learning engineer wants to log feature importance data from a CSV file at path importance_path with an MLflow run for model model.
Which of the following code blocks will accomplish this task inside of an existing MLflow run block?

A.
```
mlflow.log_model_and_data(
    model,
    importance_path,
    "feature-importance.csv"
)
```

B.
```
mlflow.log_model(
    model,
    importance_path,
    "feature-importance.csv"
)
```

C. mlflow.log_data(importance_path, "feature-importance.csv")

D. mlflow.log_artifact(importance_path, "feature-importance.csv")

E. None of these code blocks tan accomplish the task.

Suggested Answer: *C*

Community vote distribution

64934ca 4 months, 3 weeks ago

Selected Answer: D

By using the mlflow.log_artifact function, you can log the feature importance CSV file as an artifact within an existing MLflow run. Additionally, you can log the model using the appropriate MLflow flavor and optionally log the feature importance data as metrics for easier access and analysis. This approach ensures that all relevant information is logged and tracked within the same MLflow run.

upvoted 1 times

sindhu_gowda 5 months, 2 weeks ago

Answer is D

upvoted 1 times

c4b65b5 5 months, 2 weeks ago

Selected Answer: D

mlflow does not have log_data method

upvoted 1 times

srikanth923 7 months ago

Selected Answer: D

Answer is D

upvoted 2 times

hugodscarvalho 10 months ago

Selected Answer: D

D

upvoted 3 times

mozuca 11 months ago

Selected Answer: C

Agree!

upvoted 2 times

trendy01 11 months ago

Selected Answer: D

D. mlflow.log_artifact(importance_path, "feature-importance.csv")

upvoted 2 times

dplyr 11 months ago

Selected Answer: D

D. mlflow.log_artifact(importance_path, "feature-importance.csv")

upvoted 2 times

BokNinja 11 months, 1 week ago

D. mlflow.log_artifact(importance_path, "feature-importance.csv")

upvoted 2 times

Question #5 *Topic 1*

Which of the following is a simple, low-cost method of monitoring numeric feature drift?

A. Jensen-Shannon test

B. Summary statistics trends

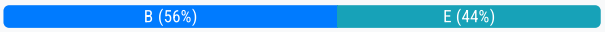
C. Chi-squared test

D. None of these can be used to monitor feature drift

E. Kolmogorov-Smirnov (KS) test

Suggested Answer: *B*

Community vote distribution



karo3 3 months, 1 week ago

It would be E, but low cost is B

upvoted 1 times

64934ca 4 months, 3 weeks ago

"Low cost" method is the key here, computationally simple on a spark/delta back end, as summary stats are compiled/updated automatically...won't require a re-index/shuffle.

Answer is B.

upvoted 1 times

03355a2 5 months ago

Selected Answer: B

Monitoring summary statistics involves tracking basic statistical measures such as mean, median, variance, and standard deviation over time. By comparing these statistics between different time periods or datasets, you can detect significant changes that may indicate feature drift. Summary statistics trends also meets the simple and low-cost method requirements.

upvoted 1 times

sindhu_gowda 5 months, 2 weeks ago

Selected Answer: E

E is answer

upvoted 2 times

hugodscarvalho 10 months ago

Selected Answer: B

Monitoring summary statistics trends over time is a simple and low-cost method of monitoring numeric feature drift. It involves tracking basic statistical metrics such as mean, median, standard deviation, etc., and observing how they change over time. This method provides insights into whether the distribution of the feature values is shifting, which could indicate drift.

upvoted 4 times

GVR76 10 months, 3 weeks ago

B. Summary statistics trends

Monitoring changes in summary statistics such as mean, median, standard deviation, and other relevant metrics over time can provide valuable insights into numeric feature drift. This method is simple, easy to implement, and does not require sophisticated statistical tests.

upvoted 3 times

StevenTan 10 months, 3 weeks ago

Selected Answer: E

E

Is it this?

upvoted 2 times

ThoBustos 6 months, 1 week ago

you're right, idk if it is consider "simple" and "low-cost" though

upvoted 1 times

Question #6 *Topic 1*

A data scientist has developed a model to predict ice cream sales using the expected temperature and expected number of hours of sun in the day. However, the expected temperature is dropping beneath the range of the input variable on which the model was trained.
Which of the following types of drift is present in the above scenario?

A. Label drift

B. None of these

C. Concept drift

D. Prediction drift

E. Feature drift

Suggested Answer: *E*

Community vote distribution

E (100%)

sindhu_gowda 5 months, 2 weeks ago

Selected Answer: E

E is answer

upvoted 1 times

hugodscarvalho 10 months ago

Selected Answer: E

Feature drift occurs when there is a change in the distribution or properties of the input features used by the model. In this case, the expected temperature, one of the input features, is dropping beneath the range of values seen during model training, indicating a shift in the feature distribution.

upvoted 1 times

trendy01 11 months ago

Selected Answer: E

E. feature drift

upvoted 1 times

BokNinja 11 months, 1 week ago

A data scientist has developed a model to predict ice cream sales using the expected temperature and expected number of hours of sun in the day. However, the expected temperature is dropping beneath the range of the input variable on which the model was trained.
Which of the following types of drift is present in the above scenario?

A. Label drift
B. None of these
C. Concept drift
D. Prediction drift
E. Feature drift

upvoted 1 times

BokNinja 11 months, 1 week ago

E. is the answer. Apologies

upvoted 1 times

Question #7 *Topic 1*

A data scientist wants to remove the star_rating column from the Delta table at the location path. To do this, they need to load in data and drop the star_rating column.
Which of the following code blocks accomplishes this task?

A. spark.read.format("delta").load(path).drop("star_rating")

B. spark.read.format("delta").table(path).drop("star_rating")

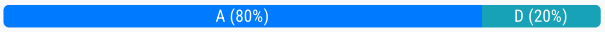
C. Delta tables cannot be modified

D. spark.read.table(path).drop("star_rating")

E. spark.sql("SELECT * EXCEPT star_rating FROM path")

Suggested Answer: *D*

Community vote distribution



BokNinja Highly Voted 11 months, 1 week ago

A. spark.read.format("delta").load(path).drop("star_rating")

upvoted 7 times

sindhu_gowda Most Recent 5 months, 2 weeks ago

Selected Answer: A

A is correct

upvoted 1 times

Alishahab70 9 months, 3 weeks ago

A is correct, if it was table_name instead of table_path then D would be correct

upvoted 2 times

spaceexplorer 9 months, 3 weeks ago

Selected Answer: D

D is correct:

upvoted 1 times

hugodscarvalho 10 months ago

Selected Answer: A

This code reads the Delta table located at path, loads it into a DataFrame, and then drops the "star_rating" column from the DataFrame.

upvoted 3 times

Question #8 *Topic 1*

Which of the following operations in Feature Store Client fs can be used to return a Spark DataFrame of a data set associated with a Feature Store table?

A. fs.create_table

B. fs.write_table

C. fs.get_table

D. There is no way to accomplish this task with fs

E. fs.read_table

Suggested Answer: *A*

Community vote distribution

E (100%)

zafarsohaib 5 months, 1 week ago

Selected Answer: E

fs.read_table returns Spark df

upvoted 1 times

sindhu_gowda 5 months, 2 weeks ago

Selected Answer: E

E is correct

upvoted 1 times

hugodscarvalho 10 months ago

Selected Answer: E

E

upvoted 2 times

random_data_guy 11 months ago

Selected Answer: E

E

upvoted 1 times

trendy01 11 months ago

Selected Answer: E

E. fs.read_table

upvoted 1 times

BokNinja 11 months, 1 week ago

E. fs.read_table

upvoted 1 times

Question #9 *Topic 1*

A machine learning engineer is in the process of implementing a concept drift monitoring solution. They are planning to use the following steps:
1. Deploy a model to production and compute predicted values
2. Obtain the observed (actual) label values
3. ____
4. Run a statistical test to determine if there are changes over time
Which of the following should be completed as Step #3?

A. Obtain the observed values (actual) feature values

B. Measure the latency of the prediction time

C. Retrain the model

D. None of these should be completed as Step #3

E. Compute the evaluation metric using the observed and predicted values

Suggested Answer: *D*

Community vote distribution

sindhu_gowda 5 months, 2 weeks ago

Selected Answer: E

Answer is E

upvoted 1 times

nnn_666 8 months, 3 weeks ago

Selected Answer: E

concept drift can only be detected by observing model performance metrics.

upvoted 2 times

Alishahab70 9 months, 3 weeks ago

Selected Answer: E

It should be E

upvoted 1 times

spaceexplorer 9 months, 3 weeks ago

Selected Answer: A

A is correct

upvoted 1 times

hugodscarvalho 10 months ago

Selected Answer: E

By computing evaluation metrics such as accuracy, precision, recall, or F1 score using these values, the engineer can assess the model's performance over time and detect any potential concept drift.

upvoted 2 times

trendy01 11 months ago

Selected Answer: E

E. Compute the evaluation metric using the observed and predicted values

upvoted 1 times

BokNinja 11 months, 1 week ago

E. Compute the evaluation metric using the observed and predicted values

upvoted 1 times

Question #10 *Topic 1*

Which of the following is a reason for using Jensen-Shannon (JS) distance over a Kolmogorov-Smirnov (KS) test for numeric feature drift detection?

A. All of these reasons

B. JS is not normalized or smoothed

C. None of these reasons

D. JS is more robust when working with large datasets

E. JS does not require any manual threshold or cutoff determinations

Suggested Answer: *D*

Community vote distribution

Jackeyquan 5 months ago

Selected Answer: D

D is the answer. For E, it's also need set a threshold

upvoted 1 times

james_donquixote 6 months ago

Selected Answer: E

This is a key advantage of using Jensen-Shannon divergence. It produces a value between 0 and 1, which represents the divergence between two distributions. This value can be interpreted without needing to set arbitrary thresholds or cutoffs. In contrast, the KS test involves comparing the test statistic to a critical value, which can depend on the significance level chosen.

upvoted 1 times

ThoBustos 6 months, 3 weeks ago

not sure about this one...

upvoted 1 times

Alishahab70 9 months, 3 weeks ago

Selected Answer: D

D is correct

upvoted 1 times

Question #11 *Topic 1*

A data scientist is utilizing MLflow to track their machine learning experiments. After completing a series of runs for the experiment with experiment ID exp_id, the data scientist wants to programmatically work with the experiment run data in a Spark DataFrame. They have an active MLflow Client client and an active Spark session spark.
Which of the following lines of code can be used to obtain run-level results for exp_id in a Spark DataFrame?

A. client.list_run_infos(exp_id)

B. spark.read.format("delta").load(exp_id)

C. There is no way to programmatically return row-level results from an MLflow Experiment.

D. mlflow.search_runs(exp_id)

E. spark.read.format("mlflow-experiment").load(exp_id)

Suggested Answer: *B*

Community vote distribution

E (100%)

hugodscarvalho 10 months ago

Selected Answer: E

Doc: https://docs.databricks.com/en/query/formats/mlflow-experiment.html#load-data-using-experiment-ids

upvoted 4 times

random_data_guy 11 months ago

Selected Answer: E

https://docs.databricks.com/en/query/formats/mlflow-experiment.html#load-data-using-experiment-ids

upvoted 1 times

mozuca 11 months ago

Agree with BookNinja. Correct Answer is E - Doc: https://docs.databricks.com/en/query/formats/mlflow-experiment.html

upvoted 1 times

BokNinja 11 months, 1 week ago

Correct Answer is E. if you want to work with the experiment run data in a Spark DataFrame, you can use E. spark.read.format(“mlflow-experiment”).load(exp_id)

upvoted 4 times

Question #12 *Topic 1*

A data scientist has developed and logged a scikit-learn random forest model model, and then they ended their Spark session and terminated their cluster. After starting a new cluster, they want to review the feature_importances_ of the original model object.
Which of the following lines of code can be used to restore the model object so that feature_importances_ is available?

A. mlflow.load_model(model_uri)

B. client.list_artifacts(run_id)["feature-importances.csv"]

C. mlflow.sklearn.load_model(model_uri)

D. This can only be viewed in the MLflow Experiments UI

E. client.pyfunc.load_model(model_uri)

Suggested Answer: *A*

Community vote distribution

C (100%)

Alishahab70 9 months, 3 weeks ago

Selected Answer: C

C is correct

upvoted 1 times

hugodscarvalho 10 months ago

Selected Answer: C

This line of code loads the scikit-learn model from the specified model URI, allowing you to access its attributes such as feature_importances.
Doc: https://mlflow.org/docs/latest/python_api/mlflow.sklearn.html#mlflow.sklearn.load_model

upvoted 1 times

random_data_guy 11 months ago

Selected Answer: C

https://mlflow.org/docs/latest/python_api/mlflow.sklearn.html#mlflow.sklearn.load_model

upvoted 1 times

trendy01 11 months ago

Selected Answer: C

C. mlflow.sklearn.load_model(model_uri)
for two different case, write different code
1. sklean : mlflow.sklearn.load_model(model_uri)
2. xgboost : mlflow.xgboost.load_model(model_uri)

upvoted 2 times

BokNinja 11 months, 1 week ago

The correct answer is C. mlflow.sklearn.load_model(model_uri)

upvoted 2 times

Question #13 *Topic 1*

Which of the following is a simple statistic to monitor for categorical feature drift?

A. Mode

B. None of these

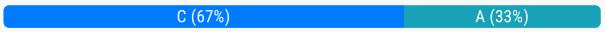
C. Mode, number of unique values, and percentage of missing values

D. Percentage of missing values

E. Number of unique values

Suggested Answer: *C*

Community vote distribution



jlocke 3 months ago

Selected Answer: C

Categorical Features
->Summary Statistics:
Mode, Number of unique values, Number of missing values

upvoted 1 times

e12ec59 3 months, 2 weeks ago

Selected Answer: C

C is the correct answer

upvoted 1 times

03355a2 5 months ago

Selected Answer: A

Among the options, the most appropriate simple statistic to monitor for categorical feature drift is mode.

The others aren't right due to the below reasons:
Number of unique values - while useful for understanding the diversity of categories, it does not directory indicate drive unless the number of categories changes significantly.

Percentage of missing values - while important for data quality, it does not directly indicate drift in the distribution of categorical values.

upvoted 1 times

ThoBustos 6 months, 3 weeks ago

would go with C

upvoted 1 times

Question #14 *Topic 1*

Which of the following is a probable response to identifying drift in a machine learning application?

A. None of these responses

B. Retraining and deploying a model on more recent data

C. All of these responses

D. Rebuilding the machine learning application with a new label variable

E. Sunsetting the machine learning application

Suggested Answer: *A*

Community vote distribution

B (100%)

hugodscarvalho 10 months ago

Selected Answer: B

This response involves updating the model by retraining it on more recent data to adapt to changes in the underlying data distribution and maintain its performance.

upvoted 1 times

random_data_guy 11 months ago

Selected Answer: B

"... it can trigger a notification or an action to recreate a new model using newer data."
https://www.databricks.com/blog/2019/09/18/productionizing-machine-learning-from-deployment-to-drift-detection.html

upvoted 1 times

BokNinja 11 months, 1 week ago

The correct answer is B. Retraining and deploying a model on more recent data12. When drift is identified in a machine learning application, one common response is to retrain the model on more recent data to account for the changes12. This can help to maintain the performance and accuracy of the model

upvoted 1 times

Question #15 *Topic 1*

A data scientist has computed updated feature values for all primary key values stored in the Feature Store table features. In addition, feature values for some new primary key values have also been computed. The updated feature values are stored in the DataFrame features_df. They want to replace all data in features with the newly computed data.
Which of the following code blocks can they use to perform this task using the Feature Store Client fs?

A.
```
fs.create_table(
    name="features",
    df=features_df,
    mode="overwrite"
)
```

B.
```
fs.write_table(
    name="features",
    df=features_df,
)
```

C.
```
fs.write_table(
    name="features",
    df=features_df,
    mode="merge"
)
```

D.
```
fs.write_table(
    name="features",
    df=features_df,
    mode="overwrite"
)
```

E.
```
fs.create_table(
    name="features",
    df=features_df,
    mode="merge"
)
```

Suggested Answer: *E*

Community vote distribution

D (100%)

hugodscarvalho 10 months ago

Selected Answer: D

The data scientist already has the table created, so the method should be "write_table". Since he wants to replace all data in features with the newly computed data the "mode" overwrite should be used.
Doc: https://docs.databricks.com/en/machine-learning/feature-store/workspace-feature-store/feature-tables.html#create-a-feature-table-in-databricks-feature-store

upvoted 4 times

mozuca 11 months ago

Selected Answer: D

Alternatively, you can create_table with schema only (without df), and populate data to the feature table with fs.write_table, fs.write_table has both overwrite and merge mode.

Example:

fs.create_table(
name=table_name,
primary_keys=["index"],
schema=airbnb_df.schema,
description="Original Airbnb data"

)

fs.write_table(
name=table_name,
df=airbnb_df,
mode="overwrite"
)

Is this case the answer is D
upvoted 2 times

BokNinja 11 months, 1 week ago
Answer is D. The mode='overwrite' argument ensures that the existing data in the feature table is replaced with the new data from features_df1.
upvoted 2 times

Question #16 *Topic 1*

After a data scientist noticed that a column was missing from a production feature set stored as a Delta table, the machine learning engineering team has been tasked with determining when the column was dropped from the feature set.
Which of the following SQL commands can be used to accomplish this task?

A. VERSION

B. DESCRIBE

C. HISTORY

D. DESCRIBE HISTORY

E. TIMESTAMP

Suggested Answer: *D*

Community vote distribution

D (100%)

hugodscarvalho 10 months ago

Selected Answer: D

DESCRIBE HISTORY command in SQL provides a history of all changes made to a Delta table, including column additions and deletions.
Doc: https://docs.databricks.com/en/sql/language-manual/delta-describe-history.html

upvoted 2 times

StevenTan 10 months, 3 weeks ago

Selected Answer: D

D is correct.
https://docs.databricks.com/en/sql/language-manual/delta-describe-history.html

upvoted 2 times

Question #17 *Topic 1*

Which of the following describes label drift?

A. Label drift is when there is a change in the distribution of the predicted target given by the model

B. None of these describe label drift

C. Label drift is when there is a change in the distribution of an input variable

D. Label drift is when there is a change in the relationship between input variables and target variables

E. Label drift is when there is a change in the distribution of a target variable

Suggested Answer: *C*

Community vote distribution

E (100%)

hugodscarvalho 10 months ago

Selected Answer: E

Label drift is when drift occurs in the label itself.

Doc: https://docs.deepchecks.com/stable/tabular/auto_checks/train_test_validation/plot_label_drift.html

upvoted 3 times

trendy01 11 months ago

Selected Answer: E

Label drift is caused by change of target variables

upvoted 1 times

BokNinja 11 months, 1 week ago

The correct answer is E. Label drift is when there is a change in the distribution of a target variable

upvoted 1 times

Question #18 *Topic 1*

Which of the following machine learning model deployment paradigms is the most common for machine learning projects?

A. On-device

B. Streaming

C. Real-time

D. Batch

E. None of these deployments

Suggested Answer: *B*

Community vote distribution

D (100%)

hugodscarvalho 10 months ago

Selected Answer: D

The most widely used deployment paradigm for machine learning projects is typically batch.

upvoted 1 times

random_data_guy 11 months ago

Selected Answer: D

D is correct.

According to the training provided by Databricks "80-90% of deployments are Batch"

upvoted 1 times

Question #19 *Topic 1*

A data scientist would like to enable MLflow Autologging for all machine learning libraries used in a notebook. They want to ensure that MLflow Autologging is used no matter what version of the Databricks Runtime for Machine Learning is used to run the notebook and no matter what workspace-wide configurations are selected in the Admin Console.
Which of the following lines of code can they use to accomplish this task?

A. mlflow.sklearn.autolog()

B. mlflow.spark.autolog()

C. spark.conf.set("autologging", True)

D. It is not possible to automatically log MLflow runs.

E. mlflow.autolog()

Suggested Answer: *C*

Community vote distribution

E (100%)

hugodscarvalho 10 months ago

Selected Answer: E

mlflow.autolog(): MLflow Autologging is enabled for all supported libraries without explicitly specifying each library.
Doc: https://mlflow.org/docs/latest/python_api/mlflow.html#mlflow.autolog

upvoted 3 times

Gursethi 10 months, 3 weeks ago

Correct Answer: E
https://mlflow.org/blog/2023/11/30/using-autolog/index.html#:~:text=MLflow's%20automatic%20logging%20functionality%20offers,specify%20what%20to%20capture%20manually.

upvoted 2 times

BokNinja 11 months, 1 week ago

The correct answer is E. mlflow.autolog()1. This function enables autologging for each supported library you have installed as soon as you import it1. It allows you to log metrics, parameters, and models without the need for explicit log statements

upvoted 4 times

Question #20 *Topic 1*

A data scientist has developed a model model and computed the RMSE of the model on the test set. They have assigned this value to the variable rmse. They now want to manually store the RMSE value with the MLflow run.
They write the following incomplete code block:

```
with mlflow.start_run(experiment_id=exp_id, run_name=run_name) as run:
    # Log rmse
    mlflow._____("rmse", rmse)
```

Which of the following lines of code can be used to fill in the blank so the code block can successfully complete the task?

A. log_artifact

B. log_model

C. log_metric

D. log_param

E. There is no way to store values like this.

Suggested Answer: *A*

Community vote distribution

C (100%)

hugodscarvalho 10 months ago

Selected Answer: C

RMSE is a metric so we should use the inbuilt mlflow.log_metric().
Doc: https://mlflow.org/docs/latest/python_api/mlflow.html#mlflow.log_metric
upvoted 2 times

random_data_guy 11 months ago

Selected Answer: C

https://mlflow.org/docs/latest/python_api/mlflow.html#mlflow.log_metric

```
import mlflow

# Log a metric
with mlflow.start_run():
mlflow.log_metric("mse", 2500.00)
```

upvoted 2 times

BokNinja 11 months, 1 week ago

C. import numpy as np

```
from sklearn.metrics import mean_squared_error
import mlflow

# Assuming 'actual' is your array of actual values and 'pred' is your array of predicted values
actual = ...
pred = ...

# Calculate RMSE
rmse = np.sqrt(mean_squared_error(actual, pred))

# Log RMSE metric in MLflow
mlflow.log_metric("rmse", rmse)
```

upvoted 2 times

Question #21 *Topic 1*

Which of the following MLflow operations can be used to automatically calculate and log a Shapley feature importance plot?

A. mlflow.shap.log_explanation

B. None of these operations can accomplish the task.

C. mlflow.shap

D. mlflow.log_figure

E. client.log_artifact

Suggested Answer: *C*

Community vote distribution

A (100%)

hugodscarvalho 10 months ago

Selected Answer: A

C is partially correct. However, going with A since it contains the method itself. Using mlflow.shap.log_explanation, you can automatically calculate and log Shapley feature importance plots, providing insights into the importance of different features in the model's predictions.

upvoted 1 times

random_data_guy 11 months ago

Selected Answer: A

Answer A.

https://mlflow.org/docs/latest/python_api/mlflow.shap.html#mlflow.shap.log_explanation

"... computes and logs explanations of an ML model's output. Explanations are logged as a directory of artifacts containing the following items generated by SHAP (SHapley Additive exPlanations).
- Base values
- SHAP values (computed using shap.KernelExplainer)
- Summary bar plot (shows the average impact of each feature on model output)"

upvoted 2 times

mozuca 11 months ago

Selected Answer: A

mlflow.shap: Automatically calculates and logs Shapley feature importance plots.

```
# Generate and log SHAP plot for first 5 records
mlflow.shap.log_explanation(rf.predict, X_train[:5])
```

upvoted 1 times

BokNinja 11 months, 1 week ago

C. mlflow.shap

upvoted 1 times

Question #22 *Topic 1*

A data scientist has developed a scikit-learn random forest model model, but they have not yet logged model with MLflow. They want to obtain the input schema and the output schema of the model so they can document what type of data is expected as input.
Which of the following MLflow operations can be used to perform this task?

A. mlflow.models.schema.infer_schema

B. mlflow.models.signature.infer_signature

C. mlflow.models.Model.get_input_schema

D. mlflow.models.Model.signature

E. There is no way to obtain the input schema and the output schema of an unlogged model.

Suggested Answer: *E*

Community vote distribution

B (100%)

hugodscarvalho 10 months ago

Selected Answer: B

The correct import statement would be: 'from mlflow.models import infer_signature.' I'm choosing B since it contains the actual method name infer_signature.

upvoted 1 times

GVR76 11 months ago

The correct answer is B. mlflow.models.signature.infer_signature.

Here's why the other options are incorrect:

A. mlflow.models.schema.infer_schema: This function is for inferring schema from data, not from models.
C. mlflow.models.Model.get_input_schema: This method only works for models that have already been logged with MLflow.
D. mlflow.models.Model.signature: This method also requires a logged model to access the signature.
E. There is no way to obtain the input schema and the output schema of an unlogged model: This is incorrect; mlflow.models.signature.infer_signature can be used for unlogged models.

upvoted 2 times

mozuca 11 months ago

I think it is: There is no way to obtain the input schema and the output schema of an unlogged model

upvoted 1 times

akssha74 11 months ago

Any questions so far from this guide?

upvoted 1 times

BokNinja 11 months, 1 week ago

The correct answer is B. mlflow.models.signature.infer_signature.

The mlflow.models.signature.infer_signature function can be used to infer an MLflow model signature, which includes the input schema and the output schema of the model. This function takes as input the model's training data (or a sample of it) and optionally the model's labels, and returns a ModelSignature object that contains the input and output schema of the model.

upvoted 1 times

Question #23 *Topic 1*

A machine learning engineer and data scientist are working together to convert a batch deployment to an always-on streaming deployment. The machine learning engineer has expressed that rigorous data tests must be put in place as a part of their conversion to account for potential changes in data formats.
Which of the following describes why these types of data type tests and checks are particularly important for streaming deployments?

A. Because the streaming deployment is always on, all types of data must be handled without producing an error

B. All of these statements

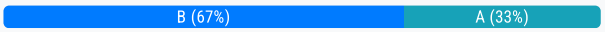
C. Because the streaming deployment is always on, there is no practitioner to debug poor model performance

D. Because the streaming deployment is always on, there is a need to confirm that the deployment can autoscale

E. None of these statements

Suggested Answer: *D*

Community vote distribution



PincoPallinoQualunque 2 months, 2 weeks ago

All of these statements is not logically possible. Otherwise it would include also "non of these statements".

upvoted 1 times

james_donquixote 6 months ago

Selected Answer: B

Agree with Bokninja

upvoted 1 times

ThoBustos 6 months ago

Selected Answer: B

agree with BokNinja

upvoted 1 times

spaceexplorer 9 months, 3 weeks ago

Selected Answer: A

A is correct

upvoted 1 times

BokNinja 11 months, 1 week ago

The correct answer is B. All of these statements.

In a streaming deployment, the system is always on and continuously processing data. Therefore, it's crucial to handle all types of data without producing an error. Rigorous data tests and checks can help ensure that the system can handle changes in data formats and continue to operate smoothly.

Additionally, because the system is always on, there may not be a practitioner available to debug poor model performance in real-time. Data tests and checks can help catch potential issues before they impact model performance.

Finally, because the system is always on, it's important to confirm that the deployment can autoscale to handle varying data volumes. Data tests and checks can help validate the system's ability to scale.

upvoted 3 times

ThoBustos 6 months, 3 weeks ago

makes more sense to me

upvoted 1 times

Question #24 *Topic 1*

Which of the following deployment paradigms can centrally compute predictions for a single record with exceedingly fast results?

A. Streaming

B. Batch

C. Edge/on-device

D. None of these strategies will accomplish the task.

E. Real-time

Suggested Answer: *D*

Community vote distribution

E (100%)

hugodscarvalho 10 months ago

Selected Answer: E

Real-time deployment is designed for low-latency, high-throughput inference, making it suitable for scenarios where predictions need to be computed quickly for individual records. This paradigm ensures rapid responses to requests, allowing for fast results even for single records.

upvoted 1 times

trendy01 11 months ago

Selected Answer: E

E. Real-time.

upvoted 1 times

BokNinja 11 months, 1 week ago

The correct answer is E. Real-time.

Real-time deployment is designed to compute predictions for individual records with very low latency. This makes it ideal for applications that require immediate predictions, such as recommendation systems, fraud detection systems, and more. In a real-time deployment, the model is typically hosted on a server, and predictions are made on-demand for incoming data.

upvoted 1 times

Question #25 *Topic 1*

A machine learning engineering team wants to build a continuous pipeline for data preparation of a machine learning application. The team would like the data to be fully processed and made ready for inference in a series of equal-sized batches.
Which of the following tools can be used to provide this type of continuous processing?

A. Spark UDFs

B. Structured Streaming

C. MLflow

D. Delta Lake

E. AutoML

Suggested Answer: *A*

Community vote distribution

B (100%)

hugodscarvalho 10 months ago

Selected Answer: B

Structured Streaming in Spark allows for continuous processing of data streams, where data can be processed in real-time and output in equal-sized batches. This makes it suitable for building continuous pipelines for data preparation in machine learning applications.

upvoted 2 times

trendy01 11 months ago

Selected Answer: B

B. Structured streaming

upvoted 1 times

BokNinja 11 months, 1 week ago

B. Structured streaming

upvoted 1 times

Question #26 *Topic 1*

A machine learning engineer wants to deploy a model for real-time serving using MLflow Model Serving. For the model, the machine learning engineer currently has one model version in each of the stages in the MLflow Model Registry. The engineer wants to know which model versions can be queried once Model Serving is enabled for the model.
Which of the following lists all of the MLflow Model Registry stages whose model versions are automatically deployed with Model Serving?

A. Staging, Production, Archived

B. Production

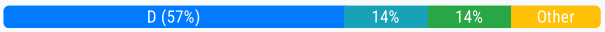
C. None, Staging, Production, Archived

D. Staging, Production

E. None, Staging, Production

Suggested Answer: *D*

Community vote distribution



james_donquixote 6 months ago

Selected Answer: D

https://www.databricks.com/blog/2020/06/25/announcing-mlflow-model-serving-on-databricks.html

upvoted 1 times

len 6 months, 1 week ago

Selected Answer: A

A is correct.

Over the course of the model's lifecycle, a model evolves—from development to staging to production. You can transition a registered model to one of the stages: Staging, Production or Archived.

https://mlflow.org/docs/latest/model-registry.html#deprecated-using-model-stages

upvoted 1 times

Alishahab70 9 months, 3 weeks ago

Selected Answer: D

D. Staging, Production

Model versions in the Staging and Production stages are automatically deployed with Model Serving. When Model Serving is enabled for a model, the latest version in the Staging stage is deployed for testing, and the latest version in the Production stage is deployed for serving predictions in production environments.

upvoted 1 times

spaceexplorer 9 months, 3 weeks ago

Selected Answer: E

E is correct

upvoted 1 times

ldoyle3332 10 months ago

Selected Answer: D

Correct answer is D.

See https://www.databricks.com/blog/2020/06/25/announcing-mlflow-model-serving-on-databricks.html

"Note the URL for each model: you can query either by the version number (1 or 2) or by the stage (Production or Staging)"

upvoted 2 times

trendy01 11 months ago

Selected Answer: C

C. None, Staging, Production, Archived

upvoted 1 times

BokNinja 11 months, 1 week ago

C. Correct

upvoted 1 times

Question #27 *Topic 1*

A data scientist has written a function to track the runs of their random forest model. The data scientist is changing the number of trees in the forest across each run.
Which of the following MLflow operations is designed to log single values like the number of trees in a random forest?

A. mlflow.log_artifact

B. mlflow.log_model

C. mlflow.log_metric

D. mlflow.log_param

E. There is no way to store values like this.

Suggested Answer: *C*

Community vote distribution

D (100%)

hugodscarvalho 10 months ago

Selected Answer: D

Using mlflow.log_param, the data scientist can easily track and record the number of trees in the random forest model across different runs.
Doc: https://mlflow.org/docs/latest/python_api/mlflow.html#mlflow.log_param

upvoted 1 times

random_data_guy 11 months ago

Selected Answer: D

Log a parameter (e.g. model hyperparameter) under the current run.
https://mlflow.org/docs/latest/python_api/mlflow.html#mlflow.log_param

upvoted 1 times

trendy01 11 months ago

Selected Answer: D

mlflow.log_param("num_trees", num_trees)

upvoted 1 times

BokNinja 11 months, 1 week ago

```
import mlflow

# Start a new MLflow run
with mlflow.start_run():
# Log the parameter
mlflow.log_param("num_trees", num_trees)
```

upvoted 1 times

BokNinja 11 months, 1 week ago

Which makes it D

upvoted 1 times

Question #28 *Topic 1*

A machine learning engineer is converting a Hyperopt-based hyperparameter tuning process from manual MLflow logging to MLflow Autologging. They are trying to determine how to manage nested Hyperopt runs with MLflow Autologging.
Which of the following approaches will create a single parent run for the process and a child run for each unique combination of hyperparameter values when using Hyperopt and MLflow Autologging?

A. Starting a manual parent run before calling fmin

B. Ensuring that a built-in model flavor is used for the model logging

C. Starting a manual child run within the objective_function

D. There is no way to accomplish nested runs with MLflow Autologging and Hyperopt

E. MLflow Autologging will automatically accomplish this task with Hyperopt

Suggested Answer: *A*

Community vote distribution

A (100%)

random_data_guy 11 months ago

Selected Answer: A

A

the notebook linked here https://learn.microsoft.com/en-us/azure/databricks/machine-learning/automl-hyperparam-tuning/hyperopt-spark-mlflow-integration#parallelize-hyperparameter-tuning-with-automated-mlflow-tracking-notebook shoes that a parent run is started before calling fmin

upvoted 1 times

BokNinja 11 months, 1 week ago

A Correct. The approach that will create a single parent run for the process and a child run for each unique combination of hyperparameter values when using Hyperopt and MLflow Autologging is starting a manual parent run before calling fmin1.

upvoted 1 times

Question #29 *Topic 1*

A data scientist has created a Python function compute_features that returns a Spark DataFrame with the following schema:

```
customer_id STRING,
spend DOUBLE,
units INT,
loyal INT,
region STRING
```

The resulting DataFrame is assigned to the features_df variable. The data scientist wants to create a Feature Store table using features_df. Which of the following code blocks can they use to create and populate the Feature Store table using the Feature Store Client fs?

A.
```
fs.create_table(
    name="new_table",
    primary_keys="customer_id",
    df=features_df,
    description="Customer features"
)
```

B.
```
fs.create_table(
    name="new_table",
    primary_keys="customer_id",
    description="Customer features"
)
```

C. features_df.write.mode("fs").path("new_table")

D.
```
fs.create_table(
    name="new_table",
    primary_keys="customer_id",
    function=compute_features,
    description="Customer features"
)
```

E. features_df.write.mode("feature").path("new_table")

Suggested Answer: *D*

Community vote distribution

A (100%)

Joy999 5 months ago

create_table(name: str, primary_keys: Union[str, List[str]], df: Optional[pyspark.sql.dataframe.DataFrame] = None, *, timestamp_keys: Union[str, List[str], None] = None, partition_columns: Union[str, List[str], None] = None, schema: Optional[pyspark.sql.types.StructType] = None, description: Optional[str] = None, tags: Optional[Dict[str, str]] = None, **kwargs) → databricks.feature_store.entities.feature_table.FeatureTable

Right Answer A

upvoted 1 times

victorcolome 10 months, 1 week ago

Selected Answer: A

A is correct according to documentation https://docs.databricks.com/en/machine-learning/feature-store/workspace-feature-store/feature-tables.html#create-a-feature-table-in-databricks-feature-store.

upvoted 1 times

BokNinja 11 months, 1 week ago

A is correct

upvoted 3 times