



- Expert Verified, Online, **Free**.

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

Community vote distribution

C (100%)

🗨️ **NarenKA** 2 months ago

Selected Answer: C

I took the exam today (4th January 2025) and noticed that not a single question was from the provided set of 74 questions (74 questions as of this date). It seems this set is completely outdated. I have gone through all 74 questions 4 times expecting them in the exam but didn't help me, so, failed the exam.

upvoted 4 times

🗨️ **olivia_sky** 5 days, 13 hours ago

Have you tried again for the exam ?

I need to take it but i'm struggling to find up to date mock exams :(

upvoted 2 times

🗨️ **yobamo** 3 months, 2 weeks ago

C is the correct answer.

upvoted 1 times

🗨️ **Deuterium44** 4 months ago

Selected Answer: C

C is the correct answer.

upvoted 1 times

🗨️ **NagaoShingo** 9 months, 1 week ago

Selected Answer: C

C is correct answer.

upvoted 1 times

A data scientist has a Spark DataFrame `spark_df`. They want to create a new Spark DataFrame that contains only the rows from `spark_df` where the value in column `price` is greater than 0.

Which of the following code blocks will accomplish this task?

- A. `spark_df[spark_df["price"] > 0]`
- B. `spark_df.filter(col("price") > 0)`
- C. `SELECT * FROM spark_df WHERE price > 0`
- D. `spark_df.loc[spark_df["price"] > 0,:]`
- E. `spark_df.loc[:,spark_df["price"] > 0]`

Correct Answer: B

Community vote distribution

B (60%)

A (40%)

Deuterium44 4 months ago

Selected Answer: B

B, given answer is correct

upvoted 1 times

Shubhamdh1 6 months, 2 weeks ago

Selected Answer: B

`spark_df.filter(col("price") > 0)` this is correct answer

upvoted 3 times

Spark_Knight 9 months ago

B is correct

upvoted 3 times

[Removed] 9 months, 2 weeks ago

Selected Answer: A

Both A and B are valid ways to filter a Spark DataFrame. You could argue that A is slightly "more" correct since option B requires you to import "pyspark.sql.functions.col"

upvoted 2 times

A health organization is developing a classification model to determine whether or not a patient currently has a specific type of infection. The organization's leaders want to maximize the number of positive cases identified by the model. Which of the following classification metrics should be used to evaluate the model?

- A. RMSE
- B. Precision
- C. Area under the residual operating curve
- D. Accuracy
- E. Recall

Correct Answer: E

Community vote distribution

E (100%)

Deuterium44 4 months ago

Selected Answer: E

E is the correct answer, the recall = the nb of items correctly classed in class i / the nb of items in class i
upvoted 1 times

pbyjny 4 months, 2 weeks ago

E is correct answer.
upvoted 1 times

NagaoShingo 9 months, 1 week ago

Selected Answer: E

E is correct answer.
upvoted 1 times

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 missing values

Correct Answer: C

Community vote distribution

C (100%)

Deuterium44 4 months ago

Selected Answer: C

C because if there are a lot of outliers, the mean would be biased and therefore median is a better choice to replace missing values
upvoted 1 times

NagaoShingo 9 months, 1 week ago

Selected Answer: C

C is correct answer.
upvoted 1 times

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

 **Deuterium44** 4 months ago

Selected Answer: D

correct answer is D

upvoted 1 times

 **8605246** 8 months, 3 weeks ago

the answer is D

Creating a binary feature variable (also known as a missing indicator) for each feature that contained missing values is a common technique to retain information about the missingness itself. This approach allows the model to potentially learn patterns related to the missingness of data, which can be informative.

Benefits of Creating Binary Indicators

Retains Information: By adding a binary indicator, you preserve the information about which values were originally missing. This can be useful if the fact that a value is missing carries predictive power.

Improves Model Performance: In some cases, the pattern of missing data can be correlated with the target variable. Including this information can help improve the model's performance.

Flexibility: This method allows you to impute missing values (e.g., with the median) while still providing the model with additional context about the data.

upvoted 2 times

A data scientist is wanting to explore summary statistics for Spark DataFrame `spark_df`. The data scientist wants to see the count, mean, standard deviation, minimum, maximum, and interquartile range (IQR) for each numerical feature. Which of the following lines of code can the data scientist run to accomplish the task?

- A. `spark_df.summary ()`
- B. `spark_df.stats()`
- C. `spark_df.describe().head()`
- D. `spark_df.printSchema()`
- E. `spark_df.toPandas()`

Correct Answer: A

 **Deuterium44** 4 months ago

Selected Answer: A

A but there is no space before the parenthesis :

upvoted 1 times

 **lontaixanh97** 5 months, 3 weeks ago

A is correct answer

upvoted 3 times

An organization is developing a feature repository and is electing to one-hot encode all categorical feature variables. A data scientist suggests that the categorical feature variables should not be one-hot encoded within the feature repository. Which of the following explanations justifies this suggestion?

- A. One-hot encoding is not supported by most machine learning libraries.
- B. One-hot encoding is dependent on the target variable's values which differ for each application.
- C. One-hot encoding is computationally intensive and should only be performed on small samples of training sets for individual machine learning problems.
- D. One-hot encoding is not a common strategy for representing categorical feature variables numerically.
- E. One-hot encoding is a potentially problematic categorical variable strategy for some machine learning algorithms.

Correct Answer: E

🗨️ 👤 **oliver29** 3 months ago

Selected Answer: C

One-hot encoding introduces a significant computational overhead, particularly when there are many categories in the feature variables. This is because it creates a new binary feature for each category, resulting in a high-dimensional representation. Such transformations are typically more efficient when performed on smaller, task-specific training datasets rather than on a global feature repository that serves multiple applications. Encoding in the feature repository also risks being rigid and unable to adapt to new categories, which further supports your argument.

upvoted 2 times

🗨️ 👤 **jaydip650** 4 months, 2 weeks ago

. One-hot encoding is dependent on the target variable's values which differ for each application.

This one doesn't hold up because one-hot encoding isn't dependent on the target variable at all. It only represents categorical features and their individual categories as binary vectors.

E. One-hot encoding is a potentially problematic categorical variable strategy for some machine learning algorithms.

Certain algorithms, like tree-based models, are less sensitive to the one-hot encoding and might perform better with other encoding techniques. However, the primary issue often boils down to the high dimensionality that one-hot encoding can introduce, which can affect algorithms that don't handle sparse data well.

So, you're still left with C as the best justification.

upvoted 2 times

🗨️ 👤 **8605246** 8 months, 3 weeks ago

It might actually be E.

According to these docs, this is the reason why the change was introduced was to allow algorithms that expect continuous features, such as logistic regression to use categorical features

upvoted 3 times

🗨️ 👤 **EricP99** 9 months ago

Correct answer B

upvoted 1 times

A data scientist has created two linear regression models. The first model uses price as a label variable and the second model uses $\log(\text{price})$ as a label variable. When evaluating the RMSE of each model by comparing 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 to computing the RMSE
- C. The data scientist failed to take the log of the predictions in the first model prior to computing the 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: B

Community vote distribution


B (100%)

 **piwardox** 1 month, 2 weeks ago

Selected Answer: D


I think the correct one should be D! As stated in the exercise prompt, in the first model I considered the RMSE to be the RMSE between the price (target variable) and the predictions, BUT the second to be the RMSE between the $\log(\text{price})$ (target variable) and the predictions. In that case, you don't need to exponentiate anything!

upvoted 1 times

 **piwardox** 1 month, 2 weeks ago

Rethinking it for one minute, I was wrong. Because in the case I imagined, there should not be so much difference between the RMSEs

upvoted 1 times

 **Akber81** 3 months, 1 week ago

Selected Answer: B

B is the right answer.

upvoted 1 times

 **Deuterium44** 4 months ago

A is the correct answer, A and D

upvoted 2 times

 **Deuterium44** 4 months ago

A and D are cannot be true in the same time, and the less the RMSD is, the more the model is accurate. Thus, as model 1 as a lower RMSD than model 2, the invalid proposition is A

upvoted 1 times

 **jaydip650** 4 months, 2 weeks ago

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

This explanation seems contradictory because if the second model were more accurate, its RMSE would be expected to be smaller, not larger.

upvoted 1 times

 **mj_infomotion** 4 months, 4 weeks ago

Selected Answer: B

Needs to bi expon due to the logarithmic transformation before

upvoted 1 times

 **SusanDeeg** 5 months ago

The question is Which of the following possible explanations for this difference is INVALID? It would have to be E since RMSE is used for regression frequently. That cannot be the explanation.



upvoted 3 times

 **Soumyashree** 9 months, 2 weeks ago

Selected Answer: B

The second model uses $\log(\text{price})$ as the label variable. If the data scientist directly computes the RMSE on the predicted log values without exponentiating them back to the original price scale, the errors will be much larger, leading to a higher RMSE.

upvoted 2 times

  **vikasgautam** 9 months, 3 weeks ago

Selected Answer: B

the current answer is wrong.. RMSE is used for regression all the time

upvoted 3 times

A data scientist uses 3-fold cross-validation when optimizing model hyperparameters for a regression problem. The following root-mean-squared-error values are calculated on each of the validation folds:

- 10.0
- 12.0
- 17.0

Which of the following values represents the overall cross-validation root-mean-squared error?

- A. 13.0
- B. 17.0
- C. 12.0
- D. 39.0
- E. 10.0

Correct Answer: A

  **Deuterium44** 4 months ago

Selected Answer: A

answer A because $RMSD = (10 + 12 + 17) / 3$; $RMSD = 39 / 3$; $RMSD = 13$
upvoted 1 times

A machine learning engineer is trying to scale a machine learning pipeline pipeline that contains multiple feature engineering stages and a modeling stage. As part of the cross-validation process, they are using the following code block:

```
cv = CrossValidator(  
    estimator=pipeline,  
    evaluator=evaluator,  
    estimatorParamMaps=param_grid,  
    numFolds=3,  
    parallelism=2,  
    seed=42  
)
```

A colleague suggests that the code block can be changed to speed up the tuning process by passing the model object to the estimator parameter and then placing the updated cv object as the final stage of the pipeline in place of the original model.

Which of the following is a negative consequence of the approach suggested by the colleague?

- A. The model will take longer to train for each unique combination of hyperparameter values
- B. The feature engineering stages will be computed using validation data
- C. The cross-validation process will no longer be parallelizable
- D. The cross-validation process will no longer be reproducible
- E. The model will be refit one more per cross-validation fold

Correct Answer: B

 **Deuterium44** 4 months ago

Selected Answer: B

- A. This is incorrect because the training time is generally determined by the pipeline complexity and the data size, not by the cross-validation object placement.
 - C. This is incorrect. The parallelism=2 parameter still allows parallel processing, and the placement of the cross-validation object doesn't impact parallelization.
 - D. This is incorrect as reproducibility depends on the seed parameter, which is set to 42 here. Moving the cross-validator in the pipeline does not inherently affect reproducibility.
 - E. This is incorrect. Moving the cross-validator object does not affect the number of times the model is refit in each fold.
- upvoted 1 times

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 embeddings
- E. String indexing

Correct Answer: *C*

  **Deuterium44** 4 months ago

Selected Answer: C

obvious answer, C

upvoted 1 times

A data scientist wants to parallelize the training of trees in a gradient boosted tree to speed up the training process. A colleague suggests that parallelizing a boosted tree algorithm can be difficult.

Which of the following describes why?

- A. Gradient boosting is not a linear algebra-based algorithm which is required for parallelization.
- B. Gradient boosting requires access to all data at once which cannot happen during parallelization.
- C. Gradient boosting calculates gradients in evaluation metrics using all cores which prevents parallelization.
- D. Gradient boosting is an iterative algorithm that requires information from the previous iteration to perform the next step.
- E. Gradient boosting uses decision trees in each iteration which cannot be parallelized.

Correct Answer: *D*

  **Deuterium44** 4 months ago

Selected Answer: *D*

D : Gradient boosting is an iterative, sequential algorithm where each tree is trained to correct the errors of the previous trees. This dependency on prior iterations means that each step relies on the output of the previous step

upvoted 1 times

A data scientist wants to efficiently tune the hyperparameters of a scikit-learn model. They elect to use the Hyperopt library's fmin operation to facilitate this process. Unfortunately, the final model is not very accurate. The data scientist suspects that there is an issue with the objective_function being passed as an argument to fmin.

They use the following code block to create the objective_function:

```
def objective_function(params):
    max_depth = params["max_depth"]
    max_features = params["max_features"]
    regressor = RandomForestRegressor(
        max_depth=max_depth,
        max_features=max_features
    )
    r2 = mean(cross_val_score(regressor, X_train, y_train, cv=3))
    return r2
```

Which of the following changes does the data scientist need to make to their objective_function in order to produce a more accurate model?

- A. Add test set validation process
- B. Add a random_state argument to the RandomForestRegressor operation
- C. Remove the mean operation that is wrapping the cross_val_score operation
- D. Replace the r2 return value with -r2
- E. Replace the fmin operation with the fmax operation

Correct Answer: D


 **Deuterium44** 4 months ago

Selected Answer: D

In Hyperopt, the fmin function is designed to minimize the objective function. In this code, the objective_function is returning the R² score directly, which typically ranges from 0 to 1 (or possibly negative if the model is poor). Since higher R² values indicate better model performance, the fmin function would mistakenly aim to minimize it, selecting configurations with lower R² values.

To correct this, the data scientist should return the negative R² value (i.e., -r2). By minimizing -r2, fmin will effectively maximize the R² score, leading to better model accuracy

upvoted 1 times

 **8605246** 8 months, 3 weeks ago

the answer is D:

The fmin function in Hyperopt is designed to minimize the objective function. In the provided code, the objective function returns the R-squared value (r2), which is a measure of how well the model explains the variance in the target variable. Since higher R-squared values indicate better model performance, the goal is to maximize this value. However, fmin minimizes the objective function, so you need to return the negative R-squared value to effectively maximize it.

upvoted 3 times

A data scientist is attempting to tune a logistic regression model using scikit-learn. They want to specify a search space for two hyperparameters and let the tuning process randomly select values for each evaluation.

They attempt to run the following code block, but it does not accomplish the desired task:

```
distributions = dict(C=uniform(loc=0, scale=4), penalty=['l2', 'l1'])
clf = GridSearchCV(logistic, distributions, random_state=0)
search = clf.fit(feature_data, target_data)
```

Which of the following changes can the data scientist make to accomplish the task?

- A. Replace the GridSearchCV operation with RandomizedSearchCV
- B. Replace the GridSearchCV operation with cross_validate
- C. Replace the GridSearchCV operation with ParameterGrid
- D. Replace the random_state=0 argument with random_state=1
- E. Replace the penalty= ['l2', 'l1'] argument with penalty=uniform ('l2', 'l1')

Correct Answer: A

 **Deuterium44** 4 months ago

Selected Answer: A

A: The data scientist wants to perform random hyperparameter sampling for tuning the logistic regression model. However, GridSearchCV performs an exhaustive search over all possible combinations of the provided parameter values, not a random selection. To accomplish random sampling of hyperparameters, they should use RandomizedSearchCV, which samples a specified number of random combinations from the defined search space.

With RandomizedSearchCV, they can specify a distribution (e.g., uniform) for continuous hyperparameters (like C) and use a list of discrete options for categorical hyperparameters (like penalty), enabling the desired random sampling approach.

upvoted 2 times

Which of the following tools can be used to parallelize the hyperparameter tuning process for single-node machine learning models using a Spark cluster?

- A. MLflow Experiment Tracking
- B. Spark ML
- C. Autoscaling clusters
- D. Hyperopt
- E. Delta Lake

Correct Answer: *D*

 **Deuterium44** 4 months ago



Selected Answer: *D*

Hyperopt is used to perform autoML and hyperparameter tuning
upvoted 2 times

Which of the following describes the relationship between native Spark DataFrames and pandas API on Spark DataFrames?

- A. pandas API on Spark DataFrames are single-node versions of Spark DataFrames with additional metadata
- B. pandas API on Spark DataFrames are more performant than Spark DataFrames
- C. pandas API on Spark DataFrames are made up of Spark DataFrames and additional metadata
- D. pandas API on Spark DataFrames are less mutable versions of Spark DataFrames
- E. pandas API on Spark DataFrames are unrelated to Spark DataFrames

Correct Answer: C

  **smonov** 3 months, 4 weeks ago

Why not A?

upvoted 1 times

  **Deuterium44** 4 months ago

Selected Answer: C

C: The pandas API on Spark (pandas-on-Spark) provides a pandas-like interface for distributed data processing in Apache Spark. Internally, a pandas API on Spark DataFrame is built on top of a Spark DataFrame with additional metadata that allows it to mimic the pandas DataFrame interface. This design enables users to use familiar pandas syntax while leveraging Spark's distributed computing capabilities

upvoted 2 times

A data scientist has written a data cleaning notebook that utilizes the pandas library, but their colleague has suggested that they refactor their notebook to scale with big data.

Which of the following approaches can the data scientist take to spend the least amount of time refactoring their notebook to scale with big data?

- A. They can refactor their notebook to process the data in parallel.
- B. They can refactor their notebook to use the PySpark DataFrame API.
- C. They can refactor their notebook to use the Scala Dataset API.
- D. They can refactor their notebook to use Spark SQL.
- E. They can refactor their notebook to utilize the pandas API on Spark.

Correct Answer: *E*

 **oliver29** 3 months ago

Selected Answer: *E*

The pandas API on Spark (pyspark.pandas) is the most efficient path for minimal disruption, scalability, and productivity for a data scientist familiar with pandas.

upvoted 1 times

A data scientist has defined a Pandas UDF function predict to parallelize the inference process for a single-node model:

```
@pandas_udf("double")
def predict(iterator: Iterator[pd.DataFrame]) -> Iterator[pd.Series]:
    model_path = f"runs://{run.info.run_id}/model"
    model = mlflow.sklearn.load_model(model_path)
    for features in iterator:
        pdf = pd.concat(features, axis=1)
        yield pd.Series(model.predict(pdf))
```

They have written the following incomplete code block to use predict to score each record of Spark DataFrame spark_df:

```
prediction_df = spark_df.withColumn(
    "prediction",
    _____
)
```

Which of the following lines of code can be used to complete the code block to successfully complete the task?

- A. predict(*spark_df.columns)
- B. mapInPandas(predict)
- C. predict(Iterator(spark_df))
- D. mapInPandas(predict(spark_df.columns))
- E. predict(spark_df.columns)

Correct Answer: A

Community vote distribution

A (100%)

 **souarav** Highly Voted 9 months, 1 week ago

Selected Answer: A

mapInPandas is used in pandas api functions and syntax is mapInPandas(predict,schema)

upvoted 5 times

 **ricorosol** Most Recent 5 months, 1 week ago

B. mapInPandas(predict): This is the correct choice. mapInPandas is used to apply a Pandas UDF to a Spark DataFrame. This function expects the UDF to take an iterator of Pandas DataFrames and return an iterator of Pandas Series or DataFrames, which matches the signature of the predict function defined.

upvoted 2 times

 **rajneesharora** 8 months, 1 week ago

correct answer is A, Scalar Pandas UDFs work with column names or expressions and return a column that gets added to the DataFrame. In this particular case, the use of


*spark_df.columns unpacks the column names, which allows the UDF to operate on all these columns. No other option provides all column names

upvoted 1 times

Which of the Spark operations can be used to randomly split a Spark DataFrame into a training DataFrame and a test DataFrame for downstream use?

- A. TrainValidationSplit
- B. DataFrame.where
- C. CrossValidator
- D. TrainValidationSplitModel
- E. DataFrame.randomSplit

Correct Answer: E

  **oliver29** 3 months ago

Selected Answer: E

DataFrame.randomSplit is specifically designed to randomly split a Spark DataFrame into multiple subsets based on specified proportions.
upvoted 1 times

A data scientist is using Spark ML to engineer features for an exploratory machine learning project.

They decide they want to standardize their features using the following code block:

```
scaler = StandardScaler(  
    withMean=True,  
    inputCol="input_features",  
    outputCol="output_features"  
)  
scaler_model = scaler.fit(features_df)  
scaled_df = scaler_model.transform(features_df)  
train_df, test_df = scaled_df.randomSplit([.8, .2], seed=42)
```

Upon code review, a colleague expressed concern with the features being standardized prior to splitting the data into a training set and a test set.

Which of the following changes can the data scientist make to address the concern?

- A. Utilize the MinMaxScaler object to standardize the training data according to global minimum and maximum values
- B. Utilize the MinMaxScaler object to standardize the test data according to global minimum and maximum values
- C. Utilize a cross-validation process rather than a train-test split process to remove the need for standardizing data
- D. Utilize the Pipeline API to standardize the training data according to the test data's summary statistics
- E. Utilize the Pipeline API to standardize the test data according to the training data's summary statistics

Correct Answer: E

  **2d84e25** 8 months ago


The concern raised by the colleague is valid. Standardizing the entire dataset before splitting into training and test sets can cause data leakage, where information from the test set influences the training process. To avoid this, the data should be standardized based on the training set statistics only, and then those statistics should be applied to the test set.

upvoted 1 times

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

 **wesleylc** 3 months, 1 week ago

Selected Answer: D


D. Calculating the median is computationally expensive in a distributed system because it requires sorting, a global operation involving data shuffling, and node coordination. In contrast, calculating the mean is efficient as it only requires summing and aggregating results across partitions.

upvoted 1 times

 **ricorosol** 5 months, 1 week ago

B. Target encoding involves replacing each category of a categorical variable with a statistic related to the target variable (like the mean of the target for that category).

upvoted 1 times

 **EricP99** 8 months, 4 weeks ago



would argue that the answer is b - Target encoding (also known as mean encoding) involves replacing each category in a categorical feature with the mean of the target variable for that category. This process is more complex and challenging to distribute efficiently because it requires calculating and applying the mean target value for each category.

upvoted 3 times

Which of the following is a benefit of using vectorized pandas UDFs instead of standard PySpark UDFs?

- A. The vectorized pandas UDFs allow for the use of type hints
- B. The vectorized pandas UDFs process data in batches rather than one row at a time
- C. The vectorized pandas UDFs allow for pandas API use inside of the function
- D. The vectorized pandas UDFs work on distributed DataFrames
- E. The vectorized pandas UDFs process data in memory rather than spilling to disk

Correct Answer: *B*

  **Leigh857** 4 months, 4 weeks ago

I think the answer should be C

upvoted 1 times

A data scientist wants to tune a set of hyperparameters for a machine learning model. They have wrapped a Spark ML model in the objective function `objective_function` and they have defined the search space `search_space`.

As a result, they have the following code block:

```
num_evals = 100
trials = SparkTrials()
best_hyperparam = fmin(
    fn=objective_function,
    space=search_space,
    algo=tpe.suggest,
    max_evals=num_evals,
    trials=trials
)
```


Which of the following changes do they need to make to the above code block in order to accomplish the task?

- A. Change `SparkTrials()` to `Trials()`
- B. Reduce `num_evals` to be less than 10
- C. Change `fmin()` to `fmax()`
- D. Remove the `trials=trials` argument
- E. Remove the `algo=tpe.suggest` argument

Correct Answer: A

Community vote distribution

A (100%)

 **[Removed]**  3 months, 1 week ago

Selected Answer: A

Option A is correct. Trying to use `SparkTrials` when the objective function itself uses Spark ML would give errors. From Hyperopt documentation "Since `SparkTrials` fits and evaluates each model on one Spark worker, it is limited to tuning single-machine ML models and workflows, such as scikit-learn or single-machine TensorFlow. For distributed ML algorithms such as Apache Spark MLlib or Horovod, you can use Hyperopt's default `Trials` class." <https://hyperopt.github.io/hyperopt/scaleout/spark/>
upvoted 5 times

A machine learning engineer would like to develop a linear regression model with Spark ML to predict the price of a hotel room. They are using the Spark DataFrame `train_df` to train the model.

The Spark DataFrame `train_df` has the following schema:

```
hotel_room_id STRING,  
price DOUBLE,  
features UDT
```

The machine learning engineer shares the following code block:

```
lr = LinearRegression(featuresCol="features", labelCol="price")  
lr_model = lr.fit(train_df)
```

Which of the following changes does the machine learning engineer need to make to complete the task?

- A. They need to call the `transform` method on `train_df`
- B. They need to convert the `features` column to be a vector
- C. They do not need to make any changes
- D. They need to utilize a Pipeline to fit the model
- E. They need to split the `features` column out into one column for each feature

Correct Answer: B

  **nizare** 5 months, 4 weeks ago

Correct Answer: B

The `features` column in Spark ML should be a vector type for the linear regression model to work. If the `features` column is not already a vector, it needs to be converted.

This can be done using a `VectorAssembler` to combine the feature columns into a single vector column, which can then be used as input to the model.

upvoted 2 times

Which of the following tools can be used to distribute large-scale feature engineering without the use of a UDF or pandas Function API for machine learning pipelines?

- A. Keras
- B. pandas
- C. PyTorch
- D. Spark ML
- E. Scikit-learn

Correct Answer: *D*

Currently there are no comments in this discussion, be the first to comment!

A data scientist has developed a linear regression model using Spark ML and computed the predictions in a Spark DataFrame `preds_df` with the following schema: prediction DOUBLE actual DOUBLE

Which of the following code blocks can be used to compute the root mean-squared-error of the model according to the data in `preds_df` and assign it to the `rmse` variable?

- ```
rmse = BinaryClassificationEvaluator(
 predictionCol="prediction",
 labelCol="actual",
 metricName="rmse"
)
A.
rmse = RegressionEvaluator(
 predictionCol="prediction",
 labelCol="actual",
 metricName="rmse"
)
B.
rmse = Summarizer(
 predictionCol="prediction",
 labelCol="actual",
 metricName="rmse"
)
C.
classification_evaluator = BinaryClassificationEvaluator(
 predictionCol="prediction",
 labelCol="actual",
 metricName="rmse"
)
D.
rmse = classification_evaluator.evaluate(preds_df)
regression_evaluator = RegressionEvaluator(
 predictionCol="prediction",
 labelCol="actual",
 metricName="rmse"
)
E.
```

**Correct Answer:** B

Currently there are no comments in this discussion, be the first to comment!

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

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:** A

Currently there are no comments in this discussion, be the first to comment!

Which of the following statements describes a Spark ML estimator?

- A. An estimator is a hyperparameter grid that can be used to train a model
- B. An estimator chains multiple algorithms together to specify an ML workflow
- C. An estimator is a trained ML model which turns a DataFrame with features into a DataFrame with predictions
- D. An estimator is an algorithm which can be fit on a DataFrame to produce a Transformer
- E. An estimator is an evaluation tool to assess to the quality of a model

**Correct Answer:** *D*

  **8605246** 8 months, 3 weeks ago

The correct statement is D. An estimator is an algorithm which can be fit on a DataFrame to produce a Transformer

In Spark MLlib, an estimator is a machine learning algorithm or pipeline stage that is used to fit a model to data. When you call the fit method on an estimator with a DataFrame, it produces a transformer. The transformer can then be used to transform a DataFrame, typically by adding predictions or other derived values.

upvoted 2 times

A data scientist has been given an incomplete notebook from the data engineering team. The notebook uses a Spark DataFrame `spark_df` on which the data scientist needs to perform further feature engineering. Unfortunately, the data scientist has not yet learned the PySpark DataFrame API.

Which of the following blocks of code can the data scientist run to be able to use the pandas API on Spark?

- A. 

```
import pyspark.pandas as ps
df = ps.DataFrame(spark_df)
```
- B. 


```
import pyspark.pandas as ps
df = ps.to_pandas(spark_df)
```
- C. 

```
spark_df.to_sql()
```
- D. 

```
import pandas as pd
df = pd.DataFrame(spark_df)
```
- E. 

```
spark_df.to_pandas()
```


**Correct Answer: A**

 **smonov** 3 months, 3 weeks ago

**Selected Answer: E**

It's E

upvoted 1 times

 **ricorosol** 5 months, 1 week ago

E. is the closest answer, the correct method name is `toPandas()`.

`pyspark.sql.DataFrame.toPandas`


`DataFrame.toPandas()` → `PandasDataFrameLike`

upvoted 2 times

 **rajneesharora** 8 months, 1 week ago

A is correct

upvoted 1 times

 **68c6a4b** 8 months, 3 weeks ago

It's not A.

E. `spark_df.to_pandas()`

Here's why:

The `to_pandas()` method is a built-in method of the PySpark DataFrame API. It converts a Spark DataFrame to a pandas DataFrame.

By calling `spark_df.to_pandas()`, the data scientist can convert the Spark DataFrame `spark_df` to a pandas DataFrame, allowing them to use the familiar pandas API for further feature engineering.

The resulting pandas DataFrame will be stored in memory on the driver node, so this approach is suitable when the data size is relatively small and can fit in the memory of the driver.

upvoted 3 times

 **rajneesharora** 8 months, 1 week ago

E is not correct as `to_pandas` would convert into pandas DF, while what is given is a Spark DF

upvoted 2 times

A data scientist has produced two models for a single machine learning problem. One of the models performs well when one of the features has a value of less than 5, and the other model performs well when the value of that feature is greater than or equal to 5. The data scientist decides to combine the two models into a single machine learning solution.

Which of the following terms is used to describe this combination of models?

- A. Bootstrap aggregation
- B. Support vector machines
- C. Bucketing
- D. Ensemble learning
- E. Stacking

**Correct Answer:** *D*

Currently there are no comments in this discussion, be the first to comment!



Which of the following machine learning algorithms typically uses bagging?

- A. Gradient boosted trees
- B. K-means
- C. Random forest
- D. Linear regression
- E. Decision tree

**Correct Answer:** *C*

Currently there are no comments in this discussion, be the first to comment!

The implementation of linear regression in Spark ML first attempts to solve the linear regression problem using matrix decomposition, but this method does not scale well to large datasets with a large number of variables.

Which of the following approaches does Spark ML use to distribute the training of a linear regression model for large data?

- A. Logistic regression
- B. Spark ML cannot distribute linear regression training
- C. Iterative optimization
- D. Least-squares method
- E. Singular value decomposition

**Correct Answer:** C

Currently there are no comments in this discussion, be the first to comment!

A machine learning engineer is converting a decision tree from sklearn to Spark ML. They notice that they are receiving different results despite all of their data and manually specified hyperparameter values being identical.

Which of the following describes a reason that the single-node sklearn decision tree and the Spark ML decision tree can differ?

- A. Spark ML decision trees test every feature variable in the splitting algorithm
- B. Spark ML decision trees automatically prune overfit trees
- C. Spark ML decision trees test more split candidates in the splitting algorithm
- D. Spark ML decision trees test a random sample of feature variables in the splitting algorithm
- E. Spark ML decision trees test binned features values as representative split candidates

**Correct Answer:** E

  **8605246** 8 months, 3 weeks ago

A machine learning engineer is converting a decision tree from sklearn to Spark ML. They notice that they are receiving different results despite all of their data and manually specified hyperparameter values being identical.

Which of the following describes a reason that the single-node sklearn decision tree and the Spark ML decision tree can differ?

- A. Spark ML decision trees test every feature variable in the splitting algorithm
- B. Spark ML decision trees automatically prune overfit trees
- C. Spark ML decision trees test more split candidates in the splitting algorithm
- D. Spark ML decision trees test a random sample of feature variables in the splitting algorithm
- E. Spark ML decision trees test binned features values as representative split candidates

upvoted 1 times

A data scientist is using MLflow to track their machine learning experiment. As a part of each of their MLflow runs, 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. All parent and child runs are being manually started with `mlflow.start_run()`. Which of the following approaches can the data scientist use to accomplish this MLflow run organization?

- A. They can turn on Databricks Autologging
- B. They can specify `nested=True` when starting the child run for each unique combination of hyperparameter values
- C. They can start each child run inside the parent run's indented code block using `mlflow.start_run()`
- D. They can start each child run with the same experiment ID as the parent run
- E. They can specify `nested=True` when starting the parent run for the tuning process

**Correct Answer:** *B*

Currently there are no comments in this discussion, be the first to comment!

Which of the following approaches can be used to view the notebook that was run to create an MLflow run?

- A. Open the MLmodel artifact in the MLflow run page
- B. Click the "Models" link in the row corresponding to the run in the MLflow experiment page
- C. Click the "Source" link in the row corresponding to the run in the MLflow experiment page
- D. Click the "Start Time" link in the row corresponding to the run in the MLflow experiment page



**Correct Answer:** C

Currently there are no comments in this discussion, be the first to comment!

A data scientist is developing a machine learning pipeline using AutoML on Databricks Machine Learning. Which of the following steps will the data scientist need to perform outside of their AutoML experiment?

- A. Model tuning
- B. Model evaluation
- C. Model deployment
- D. Exploratory data analysis

**Correct Answer:** C

  **b8dadce** 8 months, 3 weeks ago

Exploratory Data Analysis is the correct answer. The data operator is trying to get a feel for the data. If they were doing this in an automated manner, they would not be exploring the data.

upvoted 3 times