Databricks

Exam Questions Databricks-Machine-Learning-Associate

Databricks Certified Machine Learning Associate Exam

About Exambible

Your Partner of IT Exam

Found in 1998

Exambible is a company specialized on providing high quality IT exam practice study materials, especially Cisco CCNA, CCDA, CCNP, CCIE, Checkpoint CCSE, CompTIA A+, Network+ certification practice exams and so on. We guarantee that the candidates will not only pass any IT exam at the first attempt but also get profound understanding about the certificates they have got. There are so many alike companies in this industry, however, Exambible has its unique advantages that other companies could not achieve.

Our Advances

* 99.9% Uptime

All examinations will be up to date.

* 24/7 Quality Support

We will provide service round the clock.

* 100% Pass Rate

Our guarantee that you will pass the exam.

* Unique Gurantee

If you do not pass the exam at the first time, we will not only arrange FULL REFUND for you, but also provide you another exam of your claim, ABSOLUTELY FREE!

NEW 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

Answer: C

Explanation:
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

NEW QUESTION 2
A data scientist has created two linear regression models. The first model uses price as a label variable and the second model uses log(price) as a label variable. When evaluating the RMSE of each model bycomparing the label predictions to the actual price values, the data scientist notices that the RMSE for the second model is much larger than the RMSE of the first model.
Which of the following possible explanations for this difference is invalid?

A. The second model is much more accurate than the first model
B. The data scientist failed to exponentiate the predictions in the second model prior tocomputingthe RMSE
C. The datascientist failed to take the logof the predictions in the first model prior to computingthe RMSE
D. The first model is much more accurate than the second model
E. The RMSE is an invalid evaluation metric for regression problems

Answer: E

Explanation:
The Root Mean Squared Error (RMSE) is a standard and widely used metric for evaluating the accuracy of regression models. The statement that it is invalid is incorrect. Here??s a breakdown of why the other statements are or are not valid:
? Transformations and RMSE Calculation:If the model predictions were transformed (e.g., using log), they should be converted back to their original scale before calculating RMSE to ensure accuracy in the evaluation. Missteps in this conversion process can lead to misleading RMSE values.
? Accuracy of Models:Without additional information, we can't definitively say which model is more accurate without considering their RMSE values properly scaled back to the original price scale.
? Appropriateness of RMSE:RMSE is entirely valid for regression problems as it provides a measure of how accurately a model predicts the outcome, expressed in the same units as the dependent variable.
References
? "Applied Predictive Modeling" by Max Kuhn and Kjell Johnson (Springer, 2013), particularly the chapters discussing model evaluation metrics.

NEW QUESTION 3
A data scientist has developed a random forest regressor rfr and included it as the final stage in a Spark MLPipeline pipeline. They then set up a cross-validation process with pipeline as the estimator in the following code block:

```
pipeline = [string_indexer, vector_assembler, rfr]
cv = CrossValidator(
    estimator=pipeline,
    evaluator=evaluator,
    estimatorParamMaps=param_grid,
    numFolds=3,
    seed=42
)
cv_model = cv.fit(train_df)
```

Which of the following is a negative consequence of includingpipelineas the estimator in the cross-validation process rather thanrfras the estimator?

A. The process will have a longer runtime because all stages of pipeline need to be refit or retransformed with each mode
B. The process will leak data from the training set to the test set during the evaluation phase
C. The process will be unable to parallelize tuning due to the distributed nature of pipeline
D. The process will leak data prep information from the validation sets to the training sets for each model

Answer: A

Explanation:
Including the entire pipeline as the estimator in the cross-validation process means that all stages of the pipeline, including data preprocessing steps like string indexing and vector assembling, will be refit or retransformed for each fold of the cross-validation. This results in a longer runtime because each fold requires re-execution of these preprocessing steps, which can be computationally expensive.
If only the random forest regressor (rfr) were included as the estimator, the preprocessing steps would be performed once, and only the model fitting would be repeated for each fold, significantly reducing the computational overhead.
References:
? Databricks documentation on cross-validation: Cross Validation

NEW QUESTION 4
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 withadditional 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

Answer: C

Explanation:
The pandas API on Spark DataFrames are made up of Spark DataFrames with additional metadata. The pandas API on Spark aims to provide the pandas-like experience with the scalability and distributed nature of Spark. It allows users to work with pandas functions on large datasets by leveraging Spark??s underlying capabilities. References:
? Databricks documentation on pandas API on Spark: pandas API on Spark

NEW QUESTION 5
A data scientist is working with a feature set with the following schema:

```
customer_id STRING,
spend DOUBLE,
units INTEGER,
loyalty_tier STRING
```

Thecustomer_idcolumn is the primary key in the feature set. Each of the columns in the feature set has missing values. They want to replace the missing values by imputing a common value for each feature.
Which of the following lists all of the columns in the feature set that need to be imputed using the most common value of the column?

A. customer_id, loyalty_tier
B. loyalty_tier
C. units
D. spend
E. customer_id

Answer: B

Explanation:
For the feature set schema provided, the columns that need to be imputed using the most common value (mode) are typically the categorical columns. In this case, loyalty_tieris the only categorical column that should be imputed using the most common value.customer_idis a unique identifier and should not be imputed, whilespendandunits are numerical columns that should typically be imputed using the mean or median values, not the mode.
References:
? Databricks documentation on missing value imputation: Handling Missing Data
If you need any further clarification or additional questions answered, please let me know!

NEW QUESTION 6
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

Answer: A

Explanation:
To calculate the overall cross-validation root-mean-squared error (RMSE), you average the RMSE values obtained from each validation fold. Given the RMSE values of 10.0, 12.0, and 17.0 for the three folds, the overall cross-validation RMSE is calculated as the average of these three values:
Overall CV RMSE=10.0+12.0+17.03=39.03=13.0Overall CV RMSE=310.0+12.0+17.0 =339.0=13.0
Thus, the correct answer is 13.0, which accurately represents the average RMSE across all folds. References:
? Cross-validation in Regression (Understanding Cross-Validation Metrics).

NEW QUESTION 7
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()

Answer: A

Explanation:
Thesummary()function in PySpark's DataFrame API provides descriptive statistics which include count, mean, standard deviation, min, max, and quantiles for numeric columns. Here are the steps on how it can be used:
? Import PySpark:Ensure PySpark is installed and correctly configured in the Databricks environment.
? Load Data:Load the data into a Spark DataFrame.
? Apply Summary:Usespark_df.summary()to generate summary statistics.
? View Results:The output from thesummary()function includes the statistics specified in the query (count, mean, standard deviation, min, max, and potentially quartiles which approximate the interquartile range).
References
? PySpark Documentation:https://spark.apache.org/docs/latest/api/python/reference/api/pysp ark.sql.DataFrame.summary.html

NEW QUESTION 8
A data scientist learned during their training to always use 5-fold cross-validation in their model development workflow. A colleague suggests that there are cases where a train- validation split could be preferred over k-fold cross-validation when k > 2.
Which of the following describes a potential benefit of using a train-validation split over k- fold cross-validation in this scenario?

A. A holdout set is not necessary when using a train-validation split
B. Reproducibility is achievable when using a train-validation split
C. Fewer hyperparameter values need to be tested when usinga train-validation split
D. Bias is avoidable when using a train-validation split
E. Fewer models need to be trained when using a train-validation split

Answer: E

Explanation:
A train-validation split is often preferred over k-fold cross-validation (with k > 2) when computational efficiency is a concern. With a train-validation split, only two models (one on the training set and one on the validation set) are trained, whereas k-fold cross- validation requires training k models (one for each fold).
This reduction in the number of models trained can save significant computational resources and time, especially when dealing with large datasets or complex models. References:
? Model Evaluation with Train-Test Split

NEW QUESTION 9
A data scientist wants to use Spark ML to impute missing values in their PySpark DataFrame features_df. They want to replace missing values in all numeric columns in features_df with each respective numeric column??s median value.
They have developed the following code block to accomplish this task:

```
imputer = Imputer(
    strategy="median",
    inputCols=input_columns,
    outputCols=output_columns
)
imputed_features_df = imputer.transform(features_df)
```

The code block is not accomplishing the task.
Which reasons describes why the code block is not accomplishing the imputation task?

A. It does not impute both the training and test data sets.
B. The inputCols and outputCols need to be exactly the same.
C. The fit method needs to be called instead of transform.
D. It does not fit the imputer on the data to create an ImputerModel.

Answer: D

Explanation:
In the provided code block, theImputerobject is created but not fitted on the data to generate anImputerModel. Thetransformmethod is being called directly on the Imputerobject, which does not yet contain the fitted median values needed for imputation. The correct approach is to fit the imputer on the dataset first.
Corrected code:
imputer = Imputer(strategy="median", inputCols=input_columns, outputCols=output_columns) imputer_model = imputer.fit(features_df)# Fit the imputer to the dataimputed_features_df = imputer_model.transform(features_df)# Transform the data using the fitted imputer
References:
? PySpark ML Documentation

NEW QUESTION 10
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?
A)
```
rmse = BinaryClassificationEvaluator(
    predictionCol="prediction",
    labelCol="actual",
    metricName="rmse"
)
```
B)
```
rmse = RegressionEvaluator(
    predictionCol="prediction",
    labelCol="actual",
    metricName="rmse"
)
```
C)
```
rmse = RegressionEvaluator(
    predictionCol="prediction",
    labelCol="actual",
    metricName="rmse"
)
```
D)
```
classification_evaluator = BinaryClassificationEvaluator(
    predictionCol="prediction",
    labelCol="actual",
    metricName="rmse"
)
```
E)
```
rmse = classification_evaluator.evaluate(preds_df)
regression_evaluator = RegressionEvaluator(
    predictionCol="prediction",
    labelCol="actual",
    metricName="rmse"
)
```

A. Option A
B. Option B
C. Option C
D. Option D
E. Option E

Answer: C

Explanation:
The code block to compute the root mean-squared error (RMSE) for a linear regression model in Spark ML should use theRegressionEvaluatorclass withmetricNameset to "rmse". Given the schema ofpreds_dfwith columnspredictionandactual, the correct evaluator setup will specifypredictionCol="prediction"andlabelCol="actual". Thus, the appropriate code block (Option C in your list) that usesRegressionEvaluatorto compute the RMSE is the correct choice. This setup correctly measures the performance of the regression model using the predictions and actual outcomes from the DataFrame.
References:
? Spark ML documentation (Using RegressionEvaluator to Compute RMSE).

NEW QUESTION 10
A machine learning engineer is trying to scale a machine learning pipelinepipelinethat 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 theestimatorparameter and then placing the updated cv object as the final stage of thepipelinein 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 longerto train for each unique combination of hvperparameter values
B. The feature engineering stages will be computed using validation data
C. The cross-validation process will no longer be
D. The cross-validation process will no longer be reproducible
E. The model will be refit one more per cross-validation fold

Answer: B

Explanation:
If the model object is passed to the estimator parameter ofCrossValidator and the cross-validation object itself is placed as a stage in the pipeline, the feature engineering stages within the pipeline would be applied separately to each training and validation fold during cross-validation. This leads to a significant issue: the feature engineering stages would be computed using validation data, thereby leaking information from the validation set into the training process. This would potentially invalidate the cross- validation results by giving an overly optimistic performance estimate.References:
? Cross-validation and Pipeline Integration in MLlib (Avoiding Data Leakage in Pipelines).

NEW QUESTION 12
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

Answer: D

Explanation:
Ensemble learning is a machine learning technique that involves combining several models to solve a particular problem. The scenario described fits the concept of ensemble learning, where two models, each performing well under different conditions, are combined to create a more robust model. This approach often leads to better performance as it combines the strengths of multiple models.
References
? Introduction to Ensemble Learning: https://machinelearningmastery.com/ensemble-machine-learning-algorithms- python-scikit-learn/

NEW QUESTION 13
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 paqe
B. Click the "Models" link in the row corresponding to the run in the MLflow experiment paqe
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

Answer: C

Explanation:
To view the notebook that was run to create an MLflow run, you can click the "Source" link in the row corresponding to the run in the MLflow experiment page. The

"Source" link provides a direct reference to the source notebook or script that initiated the run, allowing you to review the code and methodology used in the experiment. This feature is particularly useful for reproducibility and for understanding the context of the experiment. References:
? MLflow Documentation (Viewing Run Sources and Notebooks).

NEW QUESTION 15
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. Theycan turn on Databricks Autologging
B. Theycan specify nested=True when startingthe child run for each unique combination of hyperparameter values
C. Theycan start each child run inside the parentrun's indented code block usingmlflow.start runO
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 tuningprocess

Answer: B

Explanation:
To organize MLflow runs with one parent run for the tuning process and a child run for each unique combination of hyperparameter values, the data scientist can specifynested=Truewhen starting the child run. This approach ensures that each child run is properly nested under the parent run, maintaining a clear hierarchical structure for the experiment. This nesting helps in tracking and comparing different hyperparameter combinations within the same tuning process.References:
? MLflow Documentation (Managing Nested Runs).

NEW QUESTION 20
A data scientist is using the following code block to tune hyperparameters for a machine learning model:

```
num_evals = 4
trials = SparkTrials()
best_hyperparam = fmin(
    fn=objective_function,
    space=search_space,
    algo=tpe.suggest,
    max_evals=num_evals,
    trials=trials
)
```

Which change can they make the above code block to improve the likelihood of a more accurate model?

A. Increase num_evals to 100
B. Change fmin() to fmax()
C. Change sparkTrials() to Trials()
D. Change tpe.suggest to random.suggest

Answer: A

Explanation:
To improve the likelihood of a more accurate model, the data scientist can increasenum_evalsto 100. Increasing the number of evaluations allows the hyperparameter tuning process to explore a larger search space and evaluate more combinations of hyperparameters, which increases the chance of finding a more optimal set of hyperparameters for the model.
References:
? Databricks documentation on hyperparameter tuning: Hyperparameter Tuning

NEW QUESTION 22
A data scientist has written a feature engineering notebook that utilizes the pandas library. As the size of the data processed by the notebook increases, the notebook's runtime is drastically increasing, but it is processing slowly as the size of the data included in the process increases.
Which of the following tools can the data scientist use to spend the least amount of time refactoring their notebook to scale with big data?

A. PySpark DataFrame API

B. pandas API on Spark
C. Spark SQL
D. Feature Store

Answer: B

Explanation:
The pandas API on Spark provides a way to scale pandas operations to big data while minimizing the need for refactoring existing pandas code. It allows users to run pandas operations on Spark DataFrames, leveraging Spark??s distributed computing capabilities to handle large datasets more efficiently. This approach requires minimal changes to the existing code, making it a convenient option for scaling pandas-based feature engineering notebooks.
References:
? Databricks documentation on pandas API on Spark: pandas API on Spark

NEW QUESTION 25
A machine learning engineer is trying to perform batch model inference. They want to get predictions using the linear regression model saved at the pathmodel_urifor the DataFramebatch_df.
batch_dfhas the following schema: customer_id STRING
The machine learning engineer runs the following code block to perform inference onbatch_dfusing the linear regression model atmodel_uri:

```
predictions = fs.score_batch(
    model_uri,
    batch_df
)
```

In which situation will the machine learning engineer??s code block perform the desired inference?

A. When the Feature Store feature set was logged with the model at model_uri
B. When all of the features used by the model at model_uri are in a Spark DataFrame in the PySpark
C. When the model at model_uri only uses customer_id as a feature
D. This code block will not perform the desired inference in any situation.
E. When all of the features used by the model at model_uri are in a single Feature Store table

Answer: A

Explanation:
The code block provided by the machine learning engineer will perform the desired inference when the Feature Store feature set was logged with the model at model_uri. This ensures that all necessary feature transformations and metadata are available for the model to make predictions. The Feature Store in Databricks allows for seamless integration of features and models, ensuring that the required features are correctly used during inference.
References:
? Databricks documentation on Feature Store: Feature Store in Databricks

NEW QUESTION 27
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.

Answer: E

Explanation:
One-hot encoding transforms categorical variables into a format that can be provided to machine learning algorithms to better predict the output. However, when done prematurely or universally within a feature repository, it can be problematic:
? Dimensionality Increase:One-hot encoding significantly increases the feature
space, especially with high cardinality features, which can lead to high memory consumption and slower computation.
? Model Specificity:Some models handle categorical variables natively (like decision
trees and boosting algorithms), and premature one-hot encoding can lead to inefficiency and loss of information (e.g., ordinal relationships).
? Sparse Matrix Issue:It often results in a sparse matrix where most values are zero,
which can be inefficient in both storage and computation for some algorithms.
? Generalization vs. Specificity:Encoding should ideally be tailored to specific models and use cases rather than applied generally in a feature repository.
References
? "Feature Engineering and Selection: A Practical Approach for Predictive Models" by Max Kuhn and Kjell Johnson (CRC Press, 2019).

NEW QUESTION 30
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. Scikit-learn
C. PyTorch
D. Spark ML

Answer: D

Explanation:
Spark MLlib is a machine learning library within Apache Spark that provides scalable and distributed machine learning algorithms. It is designed to work with Spark DataFrames and leverages Spark??s distributed computing capabilities to perform large- scale feature engineering and model training without the need for user-defined functions (UDFs) or the pandas Function API. Spark MLlib provides built-in transformations and algorithms that can be applied directly to large datasets.
References:
? Databricks documentation on Spark MLlib: Spark MLlib

NEW QUESTION 34
.....

Relate Links

100% Pass Your Databricks-Machine-Learning-Associate Exam with Exambible Prep Materials

https://www.exambible.com/Databricks-Machine-Learning-Associate-exam/

Contact us

We are proud of our high-quality customer service, which serves you around the clock 24/7.

Viste - https://www.exambible.com/