databricks

Databricks試験ガイド

Databricks認定 機械学習アソシエート

この試験ガイドの目的

この試験ガイドでは、試験の概要と試験範囲を説明し、試験の準備状況を判断するのに役立ちます。試験内容が変更された場合（およびこれらの変更が試験に反映される際）には、本書の内容も更新されますので、それに合わせて準備を行ってください。このバージョンは、**2024**年**10**月**28**日時点で現在公開されているバージョンをカバーしています。試験を受ける**2**週間前にもう一度確認して、最新バージョンであることを確認してください。

対象者の説明

Databricks Certified Machine Learning Associate 認定試験では、Databricks を使用して、基本的な機械学習タスクを行う能力を評価します。これには、Databricksとその機械学習機能、例えばAutoML、Unity Catalog、MLflowの一部の機能を理解し使用する能力が含まれます。また、データを探索し、特徴エンジニアリングを実行する能力も評価します。さらに、この試験では、トレーニング、チューニング、そしてモデルの評価と選択を通じてモデル構築を評価します。最後に、機械学習モデルをデプロイする能力が評価されます。この認定試験に合格すると、Databricks とその関連ツールを使用して、基本的な機械学習タスクを完了できると期待されます。

試験について

- アイテム数: 48問の多肢選択式または複数選択式の採点問題
- 制限時間： 90 分
- 登録料: $200
- 配送方法: オンライン（監督付き）
- テスト用の補助ツール: 不可
- 前提条件: 必須ではありませんが、コースへの参加と、以下の試験概要に記載されているタスクを実行する6か月の実務経験が強く推奨されます。また、このドキュメントの「推奨される準備」も参照してください。
- 有効期間： 2 年間。
- 再認定: 認定資格を維持するには、2 年ごとに再認定を受ける必要があります。再認定を受けるには、その時点で実施されている完全な試験を受ける必要があります。試験のウェブページの「試験の準備」セクションを確認して、再度試験を受ける準備をしてください。
- 採点対象外の内容: 今後使用するための統計情報を収集する目的で、試験には採点されない項目が含まれている場合があります。これらの項目は、フォームでは特定されず、得点には影響しません。この内容については、追加の時間が考慮されています。

推奨される準備

- インストラクター主導: Databricksによる機械学習
- セルフペース (Databricks Academy で提供): Databricksによる機械学習
- Pythonと、scikit-learnやSparkMLなどの機械学習をサポートする主要なライブラリに関する実用的な知識
- Unity CatalogおよびDelta Live Tablesなどのその他のDatabricksデータ管理機能に関する実用的な知識
- Databricksドキュメントの機械学習の主要トピックに精通していること

試験概要

セクション **1: Databricks Machine Learning**

- MLOps戦略のベストプラクティスを特定する
- MLランタイムを使用する利点を特定する
- AutoMLがモデル/特徴の選択をどのように容易にするかを特定する
- AutoMLがモデル開発プロセスにもたらす利点を特定する
- DatabricksのUnity Catalogでアカウント レベルとワークスペース レベルで特徴量ストア テーブルを作成する利点を特定する
- Unity Catalogで特徴量ストア テーブルを作成する
- 特徴量ストアテーブルにデータを書き込む
- 特徴量ストアテーブルからの特徴量でモデルをトレーニングする。
- 特徴量ストアテーブルからの特徴量でモデルをスコア付けする。
- オンラインとオフラインの特徴量テーブルの違いを説明する
- MLflow クライアント API を使用して、最適な実行を見極める。
- MLflow 実行中のメトリクス、アーティファクト、モデルを手動でロギングする。
- MLFlow UIで利用可能な情報を特定する
- MLflowクライアントAPIを使用してUnityカタログ レジストリにモデルを登録する
- ワークスペースレジストリではなくUnityカタログレジストリにモデルを登録する利点を特定する
- コードのプロモーションがモデルのプロモーションよりも優先されるシナリオを特定する(逆の場合も同様)
- モデルのタグを設定または削除する
- エイリアスを使用してチャレンジャーモデルをチャンピオンモデルに昇格する

セクション **2:** データ処理

- Spark DataFrameのサマリー統計を`.summary()`またはdbutilsデータサマリーを使用して計算する
- 標準偏差またはIQRに基づいてSpark DataFrameから外れ値を削除する

- カテゴリまたは連続的な特徴量の視覚化を作成する
- 適切な方法を使用して2つのカテゴリ特徴量または2つの連続特徴量を比較する
- 欠損値の平均値、中央値、最頻値による補完を比較する
- 欠損値を最頻値、平均値、中央値で補完する
- カテゴリ特徴量にはワンホットエンコーディングを使用する
- ワンホットエンコーディングが適切であるか適切でないかのモデルタイプまたはデータセットを識別し、説明する
- 対数スケール変換が適切なシナリオを特定する

セクション **3:** モデル開発
- MLの基礎を使用して、特定のモデル シナリオに適したアルゴリズムを選択する
- トレーニングデータにおけるデータの不均衡を軽減する方法を特定する
- 推定器と変換器を比較する
- トレーニングパイプラインを開発する
- Hyperoptの `fmin` 操作を使用してモデルのハイパーパラメータを調整する
- ハイパーパラメータを調整する方法として、ランダム検索、グリッド検索、またはベイズ検索を実行する
- ハイパーパラメータ調整のために単一ノードモデルを並列化する
- トレーニングと検証の分割ではなく、交差検証を使用することの利点と欠点を説明する
- モデルフィッティングの一環として交差検証を行う
- グリッド検索と交差検証プロセスによって、訓練するモデル数を特定する
- 一般的な分類メトリック、F1、Log Loss、ROC/AUCなどを使用する
- 一般的な回帰メトリック、RMSE、MAE、R二乗などを使用する
- 特定のシナリオ目標に最も適切な指標を選択する
- 評価メトリクスを計算したり予測を解釈する前に、対数変換された変数を指数化する必要性を認識する
- モデルの複雑さとバイアス分散のトレードオフがモデルのパフォーマンスに与える影響を評価する

セクション **4:** モデルのデプロイメント
- モデル サービング アプローチの違いと利点を特定する。バッチ、リアルタイム処理、ストリーミング
- モデルエンドポイントにカスタムモデルをデプロイする
- Pandasを使用してバッチ推論を実行する
- Delta Live Tablesでストリーミング推論がどのように実行されるかを確認する
- リアルタイム推論のためのモデルをデプロイしてクエリする
- リアルタイム推論のためにエンドポイント間でデータを分割する

サンプルの質問

これらの質問は、受験者がこの試験で出題される質問の全般的なスタイルや形式を理解できるよう、実際の問題に似せて作成されています。 試験ガイドに記載されている試験の目的が含まれており、目的に沿った

サンプル問題も提供されます。この試験ガイドには、試験の出題対象であるすべての目的が記載されています。認定試験の準備を行う際は、この試験ガイドの「試験の概要」を確認することをお勧めします。

問題 **1**

目的*: Unity Catalog*に特徴量ストア テーブルを作成します。

データ サイエンティストは、モデルで使用する特徴テーブルを作成したいと考えています。彼らはUnity Catalogが有効になっているワークスペースで作業しており、この特徴量テーブルをそれによって保存および管理したいと考えています。

この特徴量テーブルを作成する正しい方法は何ですか?

A. 通常通り、データを含むDeltaテーブルを作成し、Pythonで **`FeatureStoreClient`** の **`register_table`** メソッドから これをUnity Catalogの特徴量テーブルとして登録する。
B. SQL経由で **`AS feature store`** 句を使用してUnity Catalogに空のDeltaテーブルを作成し、そこにデータを書き込む。
C. Pythonで **`FeatureEngineeringClient`** の **`create_table`** メソッドを使用してテーブルを作成し、そこにデータを書き込む。
D. Unity Catalogにデータを含むDeltaテーブルを作成し、SQLで **`ALTER TABLE`** コマンドを使用して、**`SET AS feature store`** 句を含む特徴量テーブルとして構成する。

問題 **2**

目的*:* 欠損値をモード値、平均値、中央値で補完する。

データ サイエンティストは、連続的な特徴における欠損値を補完する必要があります。彼らは、最小限の労力で正しい結果を得られるよう、これを実行したいと考えています。

どの戦略でこれが実現できるでしょうか?

A. sklearn **`SimpleImputer`** を使用すると、特徴分布に基づいて最適な方法論が自動的に選択される
B. 値の分布を調べ、確認してから適切な代入を選択する
C. 連続列に対する最も適切な代入法である **`.mean()`** を使用する
D. カテゴリ列に最も適切な代入である **`.mode()`** を使用する

問題 **3**

目的*:* トレーニング データ内のデータの不均衡を軽減する方法を特定します。

データ サイエンティストは、顧客がサブスクリプション サービスから離脱するかどうかを予測するモデルを開発するための機械学習プロジェクトに取り組んでいます。データセットは非常に不均衡で、解約した顧客を表す事例はわずか10% です。あなたは、モデルが多数派に偏ることなく、少数派を効果的に識別することを保証したいと考えています。

クラスの不均衡による非解約顧客へのモデルの偏りを直接緩和する戦略はどれですか?

A. 特徴を正規化して同じスケールになるようにし、モデルのパフォーマンスを向上させる。
B. モデルのトレーニング中に少数クラスに高い誤分類コストを割り当てることで、コストに敏感な学習を使用する。
C. 解約していない顧客に関するデータをさらに収集して、トレーニング データセットのサイズを増やす。
D. より単純なモデルを使用して過剰適合を減らし、少数クラスへの一般化をより適切に実現します。

問題 **4**

目的: グリッド検索および交差検証プロセスと組み合わせてトレーニングされるモデルの数を特定します。

データ サイエンティストは、scikit-learnで5分割交差検証と `GridSearchCV` を使用して、サポート ベクター マシン (SVM) モデルを調整しています。パラメータ グリッドには、最適化する3つのハイパーパラメータが含まれています。値 `[0.1、 1、 10]`を持つC、選択肢 `['linear'、 'rbf']`を持つカーネル、値 `[0.01、 0.1、 1]`を持つガンマ。

合計でいくつの異なるモデルがトレーニングされるのでしょうか?

A. 90
B. 18
C. 1
D. 上記のどれでもない。

問題 **5**

目的: *Delta Live Tables*を使用してストリーミング推論を実行する方法を確認します。

ある企業には何千人ものユーザーがいるポッドキャスト プラットフォームがあります。同社は、ポッドキャストの視聴、一時停止、終了などのユーザーイベントの10分間の実行ウィンドウに基づいてポッドキャストのエンゲージメントが低いことを検出する異常検出アルゴリズムを実装しました。機械学習エンジニアは、1秒あたり最大数万件のイベントを処理する必要がある本番運用データパイプラインにこのモデルをデプロイしたいと考えています。イベントの量は一日を通して変動するため、エンジニアはパイプラインのコンピューティングを動的にサイズ変更する必要があります。

どのパイプライン設計アプローチがこれらの要件を満たしていますか?

A. アルゴリズムをSpark UDFとして適用するDelta Live Tablesパイプラインを作成する。
B. アルゴリズムをSpark UDFとして適用する構造化ストリーミング ジョブを作成する。
C. モデルサービングエンドポイントを作成し、エンドポイントを呼び出すカスタムUDFを呼び出すDelta Live Tablesパイプラインを作成する。
D. モデルサービングエンドポイントを作成し、エンドポイントを呼び出すカスタムUDFを呼び出す構造化ストリーミング ジョブを作成する。

答え
問題 1: C
問題 2: B
問題 3: B
問題 4: A
問題 5: A