Appendix

©SKILLUP AI, LTD.

Azure ML における 自動ML の概要

©SKILLUP AI, LTD.

Azure ML における 自動ML の概要

① 機能

② 対応領域

- 対応タスク
- 対応データ

③ Azure ML における 自動ML の全体像

©SKILLUP AI, LTD.

①機能

- Azure ML における 自動ML では主に以下の３つを自動的に実行
  - 前処理
  - モデル選択
  - ハイパーパラメータのチューニング
- 上記３つを合わせて「パイプライン」と呼ぶ

©SKILLUP AI, LTD.

②対応領域

■ 適用タスク

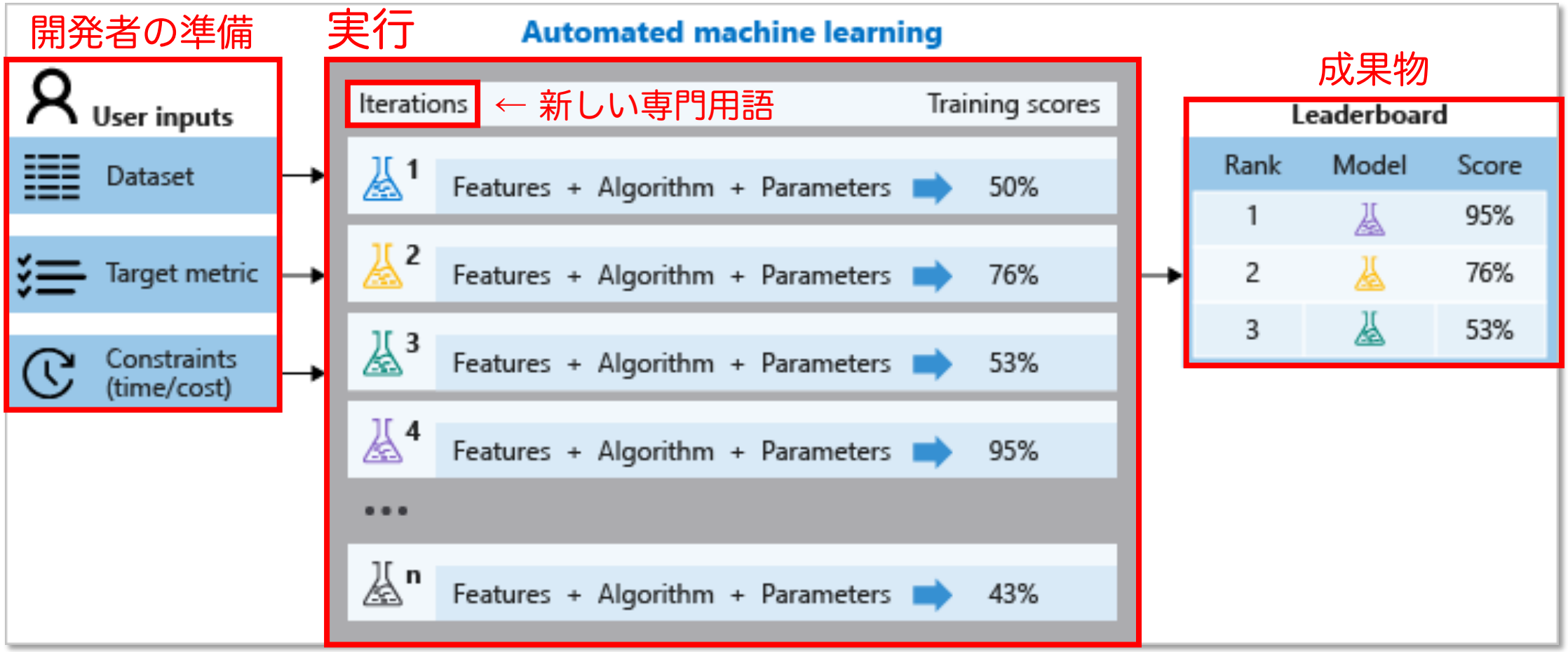
- 回帰
- 分類
- 時系列予測
  - 時系列データに特化したモデルを利用可能

■ 対応データ

- 表形式データ（Tabular Data）のみに対応（2020年10月時点）
  - 画像データに対する分類と物体検出は Cognitive Services の Custom Vision で可能

©SKILLUP AI, LTD.

③ Azure ML における 自動ML の全体像



引用元：https://docs.microsoft.com/ja-jp/azure/machine-learning/concept-automated-ml#how-automl-works

©SKILLUP AI, LTD.

③ Azure ML における 自動ML の全体像｜実行とイテレーション

■ 実行

- イテレーション（子の実行）という処理の集まり
- １回の自動MLで１回の実行

■ イテレーション

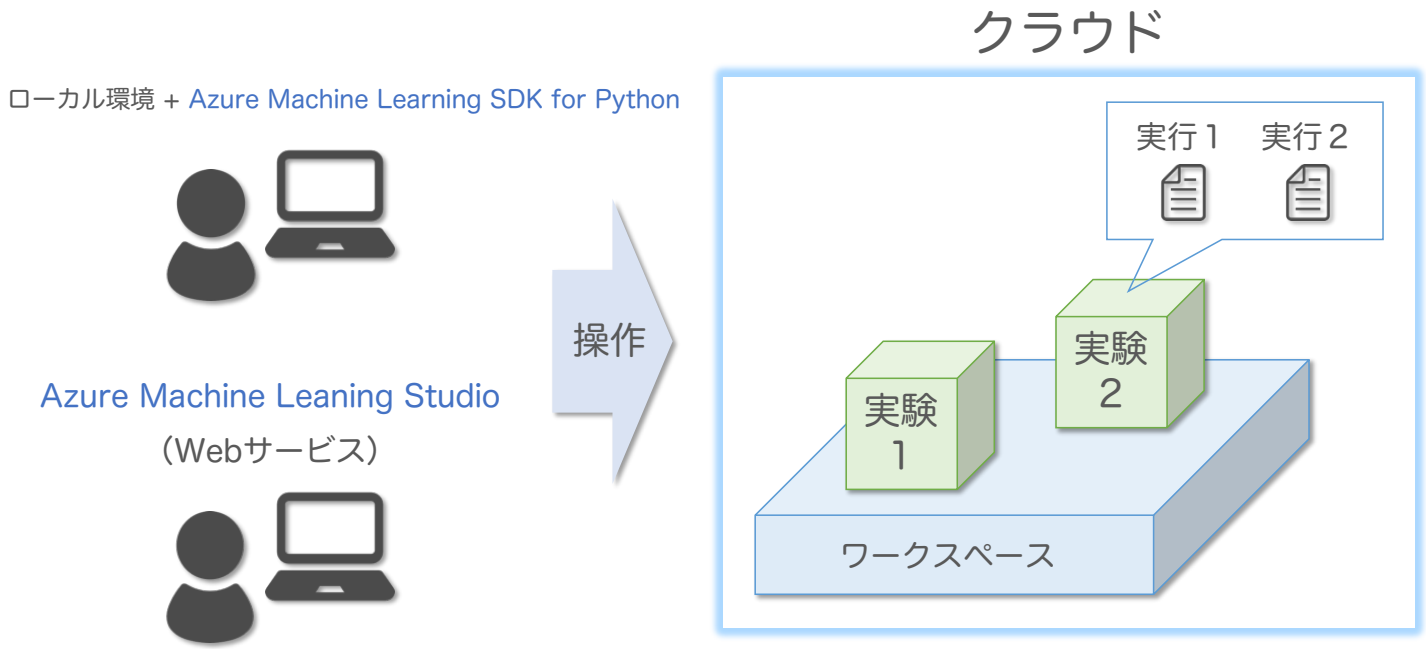
- パイプラインを動かすこと
- １つの実行の中で、数多くのイテレーションを試す

ITERATION	PIPELINE	DURATION	METRIC	BEST
0	MaxAbsScaler LightGBM	0:00:29	0.8090	0.8090
1	MaxAbsScaler XGBoostClassifier	0:00:32	0.8174	0.8174
2	MaxAbsScaler RandomForest	0:00:23	0.7949	0.8174

©SKILLUP AI, LTD.

③ Azure ML における 自動ML の全体像｜ワークスペース・実験・実行

- ワークスペースは１つだけ存在
- ワークスペースには複数の実験を作成可能
- １つの実験で複数の実行を実施



©SKILLUP AI, LTD.

使用するデータセット

©SKILLUP AI, LTD.

使用するデータセット

■ タイタニックデータセット

- タイタニック号の乗客の生存を予測するタスクに用いられる
- titanic.csv
  - https://www.kaggle.com/c/titanic

©SKILLUP AI, LTD.

タイタニックデータセット

データ列名	意味	変数の種類	備考
Survived	生存の是非	カテゴリ	１は生存、０は死亡を意味する
PassengerId	乗客の通し番号	カテゴリ	
Pclass	乗客の客室の等級	カテゴリ	１が最高級
Name	氏名	カテゴリ	敬称が付いている
Sex	性別	カテゴリ	
Age	年齢	量的	
SibSp	乗船した兄弟・配偶者の数	量的	
Parch	乗船した親・子の数	量的	
Ticket	チケット番号	カテゴリ	
Fare	乗船代金	量的	
Cabin	部屋番号	カテゴリ	
Embarked	乗船した港	カテゴリ	S・C・Qの３種類

©SKILLUP AI, LTD.

ハンズオン｜ Azure ML における 自動ML

©SKILLUP AI, LTD.

Azure ML における 自動ML の流れ

- ノートブックファイルの作成
- ライブラリのインポート
- ワークスペースへの接続
- データセットの読み込み
- train/test split
- 自動MLの設定
- 自動MLの実行
- 結果の確認

©SKILLUP AI, LTD.

ノートブックファイルの作成

- 新規でノートブックファイルを作成
  - 名前：好きな名前でOK（例：automl）

©SKILLUP AI, LTD.

ライブラリのインポート

■ ライブラリのインポート

- import pandas as pd
- from azureml.core import Workspace
- from azureml.core import Experiment
- from azureml.core import Dataset

©SKILLUP AI, LTD.

ワークスペースへの接続とデータセットの読み込み

- ワークスペースに接続
  - ws = Workspace.from_config()
- titanicデータセットの読み込み
  - df = pd.read_csv (“titanic.csv”)

©SKILLUP AI, LTD.

train/test split

- データセットを学習用とテスト用に分割
  - from sklearn.model_selection import train_test_split
    train, test = train_test_split(df, test_size=0.2, random_state=123)

©SKILLUP AI, LTD.

自動ML の設定

- ライブラリをインポート
  - import logging # python標準のロギングモジュール
  - from azureml.train.automl import AutoMLConfig
- 設定値を辞書型で定義（次のページにコードを記載）
  - automl_settings = {... 設定値 ...}
  - 設定値の詳細
- 自動MLの設定オブジェクトを作成
  - automl_config = AutoMLConfig(**automl_settings)
  - このオブジェクトを実験に送信する

©SKILLUP AI, LTD.

自動ML の設定 | 設定値を辞書で定義

■ 以下の辞書を作成し、AutoMLConfig クラスに渡す

```
automl_settings = {
    'task': 'classification',  分類を行う
    'primary_metric': 'accuracy',  accuracyが最良なモデルを探す
    'training_data': train,  学習データ
    'label_column_name': 'Survived',  教師ラベルの列
    'n_cross_validations': 5,  交差検証における分割数
    'featurization': 'auto',  特徴に潜む問題（欠損値など）に対する処置を実施
    'iteration_timeout_minutes': 2,  イテレーションの最大実施時間（分）
    'experiment_timeout_hours': 0.25,  実験の最大実施時間（時）
    'debug_log': 'titanic.log',  ログファイル名
    'verbosity': logging.INFO,  ロギングレベル
}
```


©SKILLUP AI, LTD.

自動ML の実行

- 設定オブジェクトを実験に送信
  - ex = Experiment(ws, 'titanic-automl')
  - run = ex.submit(automl_config, show_output=True)
- submitメソッドの実行で以下のようなエラーが発生する場合がある
  - ［エラー概要］
    - zipp パッケージのバージョンは 3.1.0 以下のみサポートされている
      インストールされているバージョンは 3.2.0 である
  - ［解決策］
    - !pip install zipp==3.1.0

©SKILLUP AI, LTD.

自動ML の実行 | 自動ML実行時の出力（抜粋）

■ データガードレール（データの問題を検出するAzureの機能）

- クラスが不均衡であることを検出している

```
DATA GUARDRAILS:

TYPE:         Class balancing detection
STATUS:       PASSED
DESCRIPTION:  Your inputs were analyzed, and all classes are balanced in your training data.
              Learn more about imbalanced data: https://aka.ms/AutomatedMLImbalancedData
```

■ 各イテレーションにおけるスコア

ITERATION	PIPELINE	DURATION	METRIC	BEST
0	MaxAbsScaler LightGBM	0:00:22	0.8090	0.8090
1	MaxAbsScaler XGBoostClassifier	0:00:28	0.8174	0.8174
2	MaxAbsScaler RandomForest	0:00:21	0.7879	0.8174
3	MaxAbsScaler RandomForest	0:00:23	0.6348	0.8174
4	MaxAbsScaler SGD	0:00:21	0.8047	0.8174
5	MaxAbsScaler SGD	0:00:30	0.7921	0.8174

©SKILLUP AI, LTD.

結果の確認 | 最良モデルの取得

■ 実験の概要および最良のモデルを取得

- run_info, best_model = run.get_output()
- print(run_info)
- print(best_model)

©SKILLUP AI, LTD.

結果の確認 | 最良モデルの評価

■ 未知のデータに対して予測を実行

- y_true = test['Survived']
- x_test = test.drop('Survived', axis = 1)
- y_pred = best_model.predict(x_test)

■ 混同行列を表示

- from sklearn.metrics import confusion_matrix
- cm = confusion_matrix(y_true, y_pred)
- print(cm)

©SKILLUP AI, LTD.

結果の確認 | 実験の詳細の確認

- 画面左のメニューより「アセット」＞「実験」をクリック
- 自動ML 実行時に作成した実験をクリック
- 実行の詳細を確認できる

©SKILLUP AI, LTD.

結果の確認 | 実行の詳細の確認

- 最適なモデルは Voting Ensemble であったことがわかる

- 各イテレーションの結果は「子の実行」にある

©SKILLUP AI, LTD.