入門講座
インフォマティクス入門－12

AI／機械学習プラットフォームDataRobotの使用方法

How to Use the AI/Machine Learning Platform DataRobot

伊地知晋平 DataRobot Inc.
リージョナルディレクター、データサイエンス
Shinpei Ijichi

1 緒言

従来、品質管理分野では、プロセスから得られたデータを使って統計解析によりプロセスパラメータとアウトカムとの間の関係性をモデル化し、品質向上や最適化を目的としたモデルの活用が一般的に行われてきた。近年のIoTおよびコンピュータ技術の発展の恩恵により、プロセスからますます多種大量のデータを安価に獲得できるようになったため、製造業の現場でもモデル化の手段として古典的な統計解析手法だけでなく機械学習が用いられるようになってきており、日本の産業界においてその流れは今後ますます加速するものと思われる。

一方で、統計解析や機械学習を使いこなして成果を上げるためには、Problem – Plan – Data – Analysis – Conclusionの頭文字をとったPPDACサイクル[1]（図1）を何度も回す必要がある。

この「何度も回す」が重要で、一般的にはデータを一回分析すればそれでビジネス目標を達成できるだけの成果となるケースは少なく、仮説立案～仮説検証を何度も繰り返しながら、プロセスパラメータとアウトカム間の関係性を帰納的に詰めていく泥臭い作業が必要となることが多い。

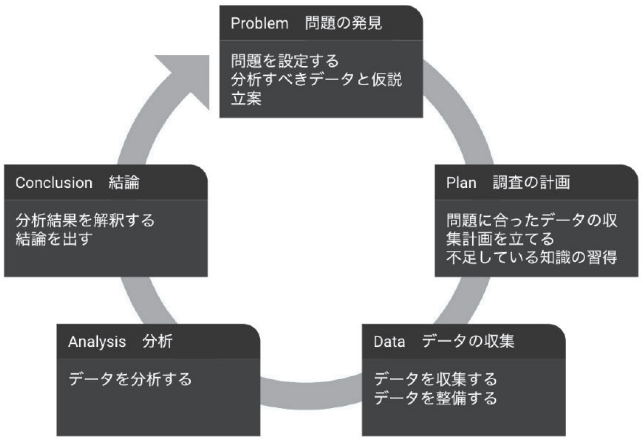

図1　PPDACサイクル

しかしながら、統計学やデータサイエンスの非専門家にとって、RやPythonなどのプログラミング言語を用いてコーディングを行いながら、何度もPPDACサイクルを回す作業は非常に負荷の重いものであるし、本来彼らの専門領域でのドメイン知識を駆使して仮説を立案したり分析結果を解釈したりしなければいけないのに、そのような本質的な業務に十分な時間を割けない矛盾を感じることが少なくない（筆者は過去製造業系企業でのデータ分析業務に従事していたが、一つのモデルを実務で使えるレベルに持っていくために数週間かかることも稀ではなかった）。

この問題に対する有力なソリューションが機械学習プロセスの自動化技術である。筆者が所属するDataRobot Inc.は、2012年米国ボストンに設立以来機械学習自動化技術に特化したソリューション開発に取り組み、2015年にこの分野では世界で最も早くエンタープライズAIプラットフォーム「DataRobot Automated Machine Learning」を世に出した。以降エンタープライズＡＩのグローバルリーダーとして「企業のAI化を実現」を旗印に精力的な新技術開発に努め、現在は全世界数百名のデータサイエンティストやソフトウェア技術者が日夜技術開発に携わっている。またAIプラットフォームをご提供するだけではなく、2019年からはお客様をAIで成功に導くためのカスタムサービスプログラム「AIサクセス®」を通して、日本でも約2年間で数百を超えるAI開発ユースケース（事例）に携わってきた経験・知見に基づくノウハウを惜しみなくユーザー様に提供し、お客様の成功をご支援させていただいている。

2 目的

本稿では、「DataRobot Automated Machine Learning」(以下DataRobot) を例に、データの準備からモデルの業務実装までの一連のプロセスが高度に自動化されたエンタープライズAIプラットフォームの使用方法を紹介する。さらに、モデルを配備 (デプロイ) し運用を開始してからの監視・管理も同プラットフォーム上で一元的に行えることを示す。なお、本稿は2020年3月時点のDataRobot Ver.5.3 (クラウド版) を元に書かれていることをお断りしておく。また、図2～図22は全てDataRobotの製品画面よりキャプチャーした。

3 DataRobotについて

3.1 機械学習プロセス俯瞰

データ分析フレームワークとして世界的に認知されているPPDACサイクルは機械学習プロセスにも適用できる。

a) Problem: 機械学習で解くべき問題の発見・定義とビジネスインパクトの試算
b) Plan: 問題に合ったデータの収集計画を立て、これまでに分かっているドメイン知識を習得
c) Data: データを準備し、データ形式を整える
d) Analyze：機械学習を用いたモデリングを行い、同時にインサイトを得る
e) Conclusion: 機械学習モデルを解釈し実運用可否を判断する

機械学習プロセスにおいては、上記a)－e) を何回か回した後、さらに成果物である機械学習モデルを実使用環境に配備 (デプロイ) し、実運用へ進むまでのステップがある。

f) Deployment: 実運用に進めると判断された機械学習モデルを開発環境から運用環境に展開する
g) Operationalize: 運用環境に実装された機械学習モデルを試験的に運用する
h) Production & Monitoring: 機械学習モデルを本運用に切り替え、そのパフォーマンスをトラックする

3.2 DataRobotの守備範囲

前節で俯瞰した機械学習プロセスは、「データを価値に変換するプロセス」とも言い換えることができる。

2020年3月現在、DataRobotは、c) の一部からh) までのステップをカバーしている。また2019年12月に発表されたPaxata社の買収により、今後データ準備プロセスも高度な省力化が進み、データが価値に変換されるEnd-to-Endプロセスの自動化・省力化はさらに広範囲に実現される。

DataRobotはデータ投入からベストモデルを作成・配備するまでには合計数クリックを必要とするだけである。また、モデルから様々なインサイトを得ることができるため、製造業や医療業界では要因分析を主目的としてご使用いただいているケースもある。

4 DataRobotの使用法

では本章から具体的にDataRobotの使用方法と主要な機能を説明する。

4.1 データの読み込み～探索的データ解析 (EDA)

図2にDataRobotのメイン画面を示す。DataRobotが推奨するブラウザはGoogle Chromeである。

DataRobotにデータを投入する方法はいくつかあり、メイン画面に表示される方法のいずれかを使用してデータをインポートできるが、最も簡便かつよく使われているのはPC内のローカルファイルをドラッグ＆ドロップする方法である。データの読み込み時点で探索的データ解析 (EDA) も同時に行われる (図3)。

データの読み込みが終わると特徴量名がリスト表示され、任意の特徴量のヒストグラム (数値データ) や棒グラフ (カテゴリデータ)、基礎統計量などを表示・確認できる (図4)。

モデリングを始めるには、先に予測ターゲットをユーザーが指定する必要があり、ターゲット特徴量の名前 (予測するデータセットの列) を入力するか、特徴量セットの名前の横にある「ターゲットとして使う」をクリックする。こうして画面上部の予測ターゲットを指定する欄に特徴量名が入力されると右側の開始ボタンがアクティブになる (図5)。

機械学習を実行する際には読み込んだデータを学習データと検定データに分割し、交差検定のフォールド数やホールドアウトデータの割合などを設定しなければならないが、DataRobotはそれらの条件を自動で設定する。そのままの条件でモデリングを行っても問題ないが、もしユーザーが細か

図2 DataRobotのメイン画面

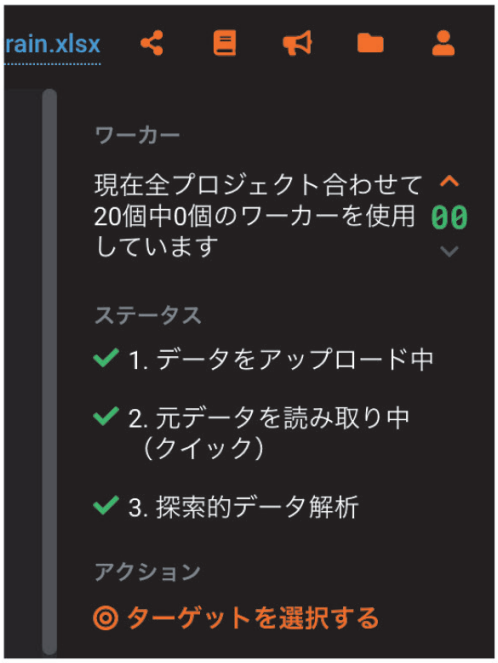

図3　探索的データ解析

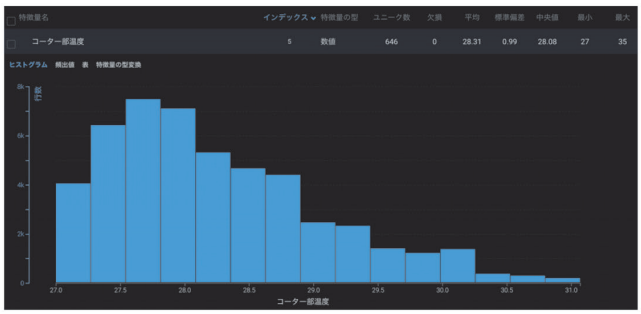

図4　数値データ特徴量の分布と基礎統計量

く調整したい場合には「高度なオプション」メニューから様々な条件をマニュアルで設定できる（図6）。

4.2　機械学習モデルの自動作成

「開始」ボタンを押すとDataRobotは機械学習モデリングを始め、実行中のモデルの進捗インジケーターが画面右側のワーカーキューに表示される（図7）。

モデリング結果は（選択した最適化指標に基づいて）精度パフォーマンスが最も優れているモデルがリストの上位にランク付けされた状態でモデルのリーダーボード（順位表）に表示される。

DataRobotは読み込まれたデータの統計的な性質に基づいてブループリントと呼ばれる「前処理 & アルゴリズム」のフローを概ね30-40種ほど作成し（「オートパイロット」モードの場合）、各ブループリントから高精度な機械学習モデルを自動作成する。選択されるアルゴリズムにはDataRobot社で独自に開発したものに加えてPython scikit-learnやRなど様々

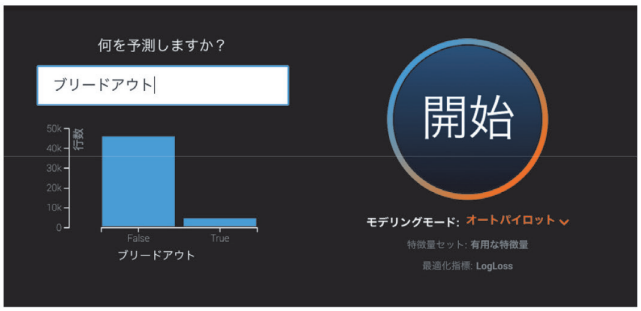

図5　予測ターゲットを指定したところ

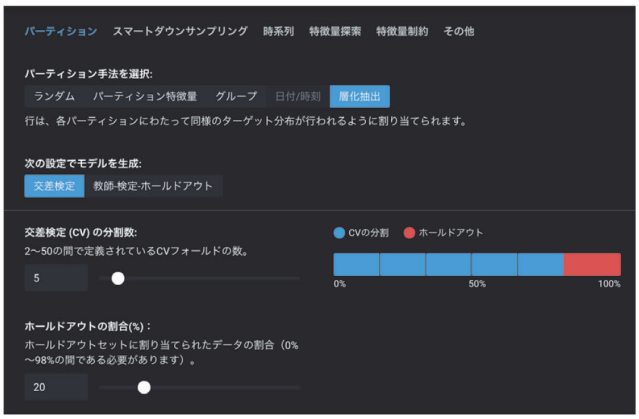

図6　「高度なオプション」画面

図7　モデリング進捗インジケーター

なオープンソースの機械学習ライブラリも網羅されている。

なぜ多様なモデルを作成するのかといえば大きく2つの理由がある。一つは世の中に万能な機械学習アルゴリズムはなく、与えられたデータに対してどのアルゴリズムで作られたモ

デルの精度パフォーマンスが最も良いかはやってみないとわからないという事情による。もう一つは、精度の良い単体のモデルを複数集めて組み合わせるともっと精度の良いモデル（アンサンブルモデルと呼ばれる）になる場合が多いからである。DataRobotは様々な組み合わせ方法を用いたアンサンブルモデルまでも自動作成して、リーダーボード上に表示する。(図8でBlenderという単語が含まれるモデルがアンサンブルモデル)

4.3　機械学習モデルの解釈とインサイト獲得

DataRobotではリーダーボードに表示されたモデル毎に統一のフォーマットで様々なインサイト（知見）を得られる。代表的な機能を以下にまとめる。

a）モデル精度の表示
　（ア）分類モデル：ROC曲線、混同行列など
　（イ）回帰モデル:予測値/実測値プロット、残差プロットなど
b）アウトカム（目的変数）と関係性の高い特徴量の表示:特徴量のインパクト[2] (図9)
c）アウトカムと各特徴量との関係性表示：特徴量ごとの作用[3] (図10)
d）予測結果に強く影響を与えた特徴量とその値を表示：予測の説明 (図11)
e）テキスト変数の単語とアウトカムとの関係の強さ：ワードクラウド (図12)

4.4　ベストモデルの選択

前述のように、リーダーボードには精度指標が良い順に多数のモデルが表示されるので、それらのモデルからユーザーにとってベストなモデルを選択する。もしユーザーが精度パフォーマンスだけに関心があるのであれば、精度が最も良いモデルを選択すれば良い。多くの場合、複数の単独モデルを組み合わせたBlenderモデルの精度が最も良くなる。

モデルの実運用形態からリアルタイムに予測を行う必要があるなどの理由で、予測スコアリング速度も重要な場合には、予測精度と予測速度の両方が良いパフォーマンスを示すバランスの取れたモデルを選択する。一般的にはBlenderモデルよりも単独モデルの予測スコアリング速度が速いので、精度と速度のバランスを重視する場合には単独モデルが選択されることが多い。DataRobotでは速度と精度のトレードオフをプロットしたチャート (図13) でバランスの取れたモ

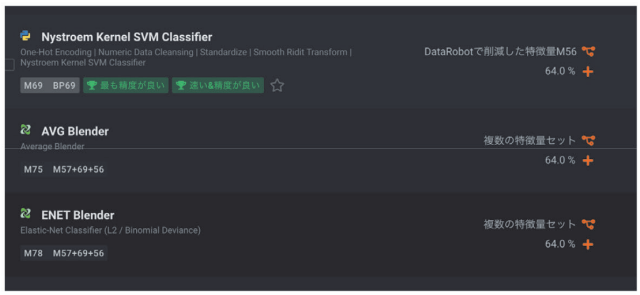

図8　リーダーボードの一部

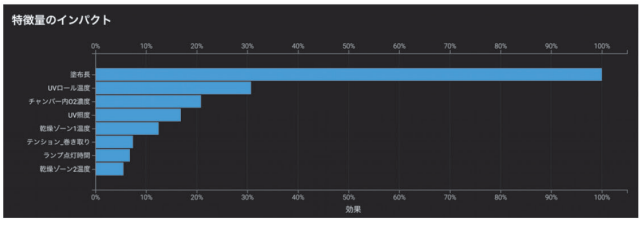

図9　特徴量のインパクト

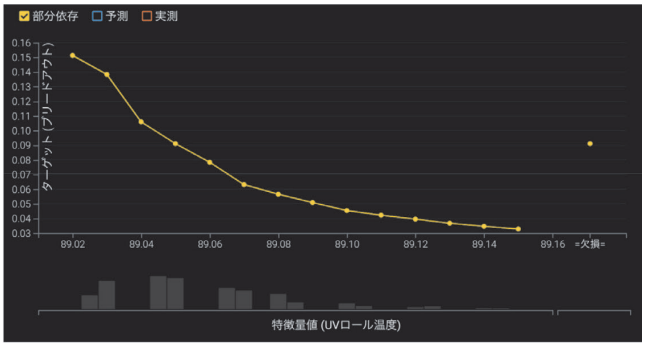

図10　特徴量ごとの作用

図11　予測の説明

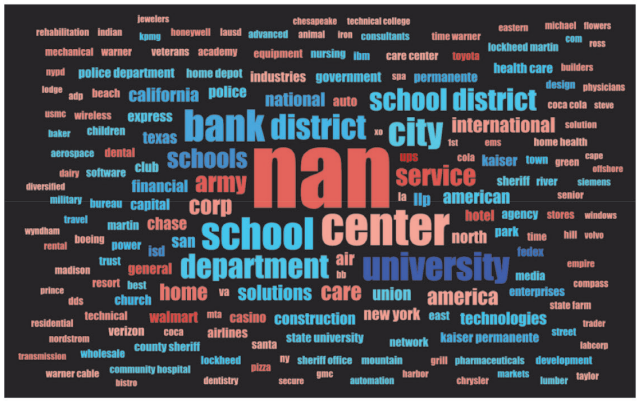

図12　ワードクラウド（日本語にも対応）

デルを確認できる。チャート中の点の一つ一つが精度（縦軸）上位のモデルであり、それらの予測時間（横軸）の推定値に従ってプロットされている（精度上位のモデル10個が右側にリストアップされている）。精度指標はリーダーボードから任意に指定できる。

リーダーボードではホールドアウトデータに対する精度はデフォルトで非表示となっている。まず交差検定の精度指標や「速度vs精度チャート」などを確認してユーザーにとってベストなモデルを選定したら、その後に「ホールドアウトの解除」を行う。ホールドアウトデータの精度も交差検定の精度と比べて大きく乖離していなければ、学習時に使われていない全く未知のデータに対しての精度パフォーマンスも信用できるとして、モデルの配備（デプロイ）に進む。

このとき、ホールドアウトデータの精度を見て改めてモデルを選択しなおしてはいけない（モデルの再選択を行なってしまうとホールドアウトデータに過学習したモデルを選ぶ危険性があるため）。

その後、例えば学習データの行数が小さい場合には、選択したモデルだけに対して全データを使って再学習させることも簡単にできる（図14）。このように学習データ数を増やすとさらなる精度の向上が期待できる。

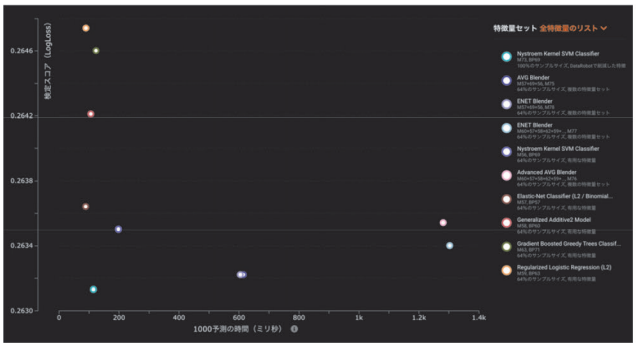

図13　予測速度 vs 予測精度

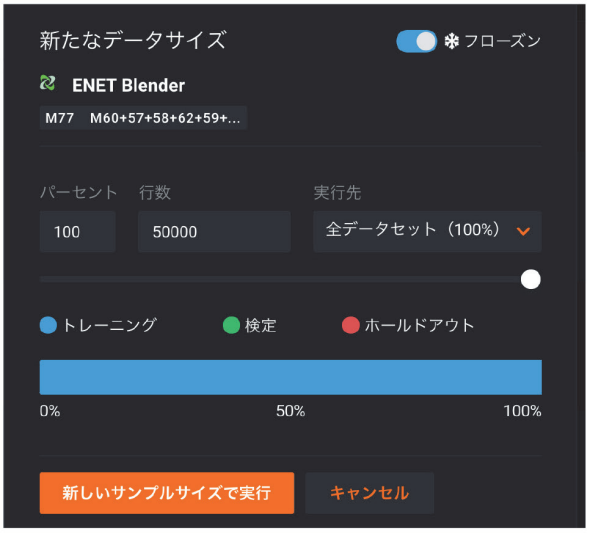

図14　全データを使用してのモデル再学習

4.5　モデルの配備（デプロイ）と予測の実施

モデルを運用する形態は予測を行う頻度や1回の予測量、セキュリティ要件などによって変わるが、DataRobotではベストモデルを選択した後に、モデルを実運用環境へ様々な形態で柔軟に配備できる。本稿ではDataRobot内の実運用環境にモデルを配備する方法を紹介する（なお、モデルをJAVAスクリプトなどの形式で出力するオプション機能を利用すれば、ユーザーのシステムにそのモデルを組み込んで、システム内に閉じた形で予測スコアリングを行うことができる）。

方法は非常に簡単で、選ばれたベストモデルの「予測＞デプロイ」タブを選択し（図15）、表示されたオレンジ色のボタンを2回クリックするだけで短時間に完了する（図16）。

配備モデルでAPI（Application Programming Interface）を利用して予測スコアリングを実施するには、DataRobotが自動作成したPythonスクリプト（図17）を実行すれば良い。実際にスクリプトをユーザーのシステムに組み込んで、定期的に予測スコアリングを自動で実施させている例もよく見られる。

図15　モデルをデプロイ

図16　実運用環境に配備（デプロイ）されたモデル

5 モデルの管理

4章ではDataRobotでモデルを自動作成し実運用環境へ配備（デプロイ）するまでの方法と機能を紹介した。本章では、実運用を開始したモデルの監視と管理もDataRobot上で一元的に行えることを紹介する。

5.1 実運用化されたモデルの管理

DataRobotで作成した機械学習モデルをいくつも業務プロセスで実運用するようになると、それらのモデルのパフォーマンスを監視して、適切なタイミングでモデルの更新などの介入を行う必要があるが、それらの監視・管理もデプロイインベントリタブ（図18）から行える。

インベントリの上部には、信号機の色（緑、黄、赤）で色分けされた正常性インジケーターが表示され、全てのアクティブな配備モデルの使用状況とステータスが要約されている。そのサマリーの下には各配備モデルのステータスが下記項目について表示され、a）～c）は同じく信号機の色で表示される。

a）サービスの正常性：過去24時間におけるエラーの有無
b）データドリフト
c）精度：正解データが与えられると計算される
d）アクティビティ：過去7日間の予測数パターン
e）平均予測数/日
f）最新の予測：モデルに対して最後の予測を行ってからの時間

図17　自動作成されたPythonスクリプトの一部

5.2 監視項目の詳細

インベントリの任意の配備モデル名をクリックすると、その詳細情報（図16）が表示される。そこで「サービスの正常性」、「データドリフト」、「精度」のタブを選択するとそのモデルの過去のステータスがそれぞれ示される。

5.2.1 サービスの正常性

サービスの正常性タブ（図19）では、予測の合計数、予測実行時間、レスポンス時間、システムエラーの割合、などがダッシュボード表示される。

5.2.2 データドリフト

データドリフトタブ（図20、図21）は、配備したモデルの特定の時間間隔における正常性を識別するのに役立つ、3つのインタラクティブな視覚化機能を提供する。

図20左側の「特徴量ドリフトと特徴量の有用性チャート」においては、チャート上部のスライダ（水色の横線）で設定できる任意の期間で、有用性上位の特徴量を「実際に特徴量値の分布がどれだけ変化したか（縦軸）」と「特徴量の有用性（横軸）」の2軸に基づきプロットしている（各特徴量が一つ一つの点）。

ドリフトの大きさによって特徴量を表すポイントの色は信号機の色で表示される。任意のポイントをクリックすると図20右側の「特徴量の詳細チャート」に、学習データで選択された特徴量のレコードのスコアリングデータに対するパーセンテージ（分布）を表示させて確認することができる（薄い青色の棒グラフが学習データの分布、濃い青色の棒グラフが予測スコアリングに使われたデータの分布を示している）。

図18　デプロイインベントリタブ

図19　サービスの正常性

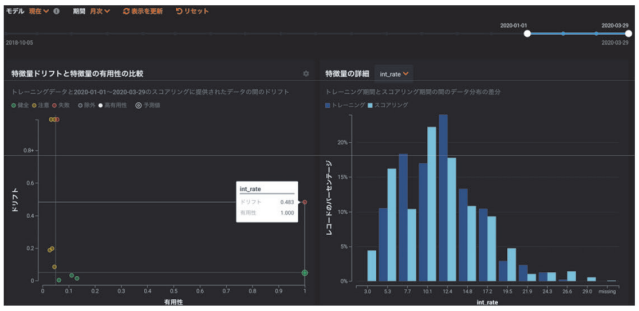

図20　データドリフト

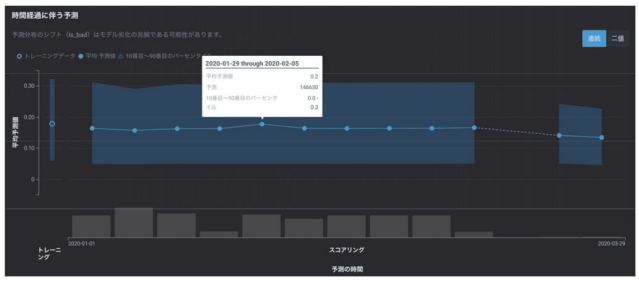

図21　時間経過に伴う予測

図20は左側のチャートで有用性が高くデータドリフトが発生していると表示されたある特徴量のポイント（赤色になっている）をクリックした際の表示例である。右側のチャートでは薄い色の棒グラフの分布と濃い色の棒グラフの分布の形状が一部で大きく異なっている。

図21はデータドリフトチャートの下部に表示される「時間経過に伴う予測」チャートである。縦軸が平均予測値、横軸が時間を表し、回帰モデルの場合には予測値の中央から±40%の範囲を表す視覚的インジケーターもプロットされる。一番左側の1ポイントは、モデル学習時のホールドアウトデータで作成された予測値である。

各ポイントにカーソルを合わせるとそのポイントにおける期間の開始点-終点、平均予測値、予測回数、10パーセンタイル点-90パーセンタイル点の値などを確認できる。もしある期間の予測平均値が大きくシフトしていた場合には、その期間に絞って原因を探るなどのアクションを取れる。

5.2.3　精度

精度タブ（図22）は、標準的な統計的手法と可視化によって、時間の経過に伴うモデルデプロイのパフォーマンスを表示する。精度タブを使用するには、DataRobot外で収集された予測値および実測値を含むデータセットをアップロードする必要がある。

データドリフトチャートと同様に、精度チャートも上部のスライダ（水色の横線）で過去の任意の期間を設定できる。上の1本（緑色）の折れ線グラフは選択した精度指標値（この例ではLogLoss）の時間経過に伴う変化を示している。これ

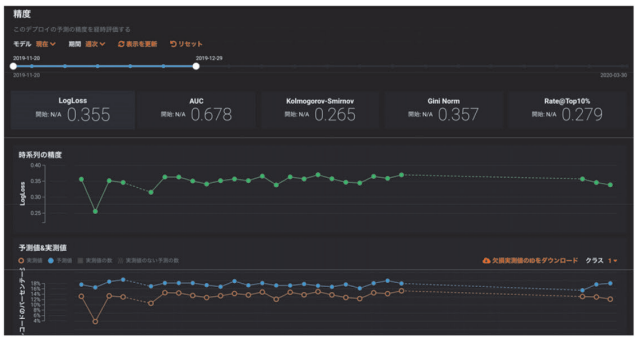

図22　精度

らの指標は、配備（デプロイ）前にモデルの評価に使用される指標と同一である。グラフの上にあるメトリックス・タイルをクリックすると、表示が変更される。

下の2本（水色＆オレンジ色）の折れ線グラフには、データセットの全体のタイムラインに沿って予測値と実測値が表示される。グラフ上のポイントは、データドリフトタブに表示されるポイントに一致している。したがって、いずれかのプロットのポイントの上にマウスを置くと同じ値が表示される（時間スライダが同じように設定されている場合）。

6　結語

以上、エンタープライズＡＩプラットフォームDataRobotの使用方法と主な機能を紹介した。DataRobotにデータを投入してからベストモデルを実運用環境へ配備するには、予測ターゲットの入力、開始ボタンのクリック、モデルデプロイボタンのクリック（2回）とわずか数回のクリックで済む。コーディングや機械学習アルゴリズムの知識が乏しい人でも、DataRobotを使えばデータサイエンティストが数週間かけなければできないようなタスクを数時間で行うことができるため、本稿1章で述べたように様々な仮説を素早く検証し、本当に実運用に堪えるAIアプリケーションを短期間に開発することができる。

参考文献

1）総務省統計局, 統計データ利活用センター：http://www.stat.go.jp/dstart/point/seminar1/01.html, (accessed 2021-7-28).
2）DataRobotブログ, (2019). https://blog.datarobot.com/jp/permutation-importance
3）DataRobotブログ, (2019). https://blog.datarobot.com/jp/2018/02/15/modelxray

(2020年4月1日受付)