未来予測における民主化実現の鍵
―マルチクラウド環境におけるデータセット作成の重要性―

BI（ビジネスインテリジェンス）製品の成熟が進み、過去・現在の情報の可視化がビジネスにおける重要な判断を支えている。今後はさらに「未来の予測」について、高い精度が求められるようになる。これを行えるデータサイエンティストの不足に対して、誰もが未来の予測を行える製品が注目されている。本稿では、その真の価値を得るために必要となる鍵について考察する。

野村総合研究所 マルチクラウドインテグレーション事業本部
マルチクラウドインテグレーション事業部 上級テクニカルエンジニア NRI認定ITアーキテクト
大塚（おおつか） 紳一郎（しんいちろう）
専門はOracle Databaseを中心としたRDBMS技術（Oracle ACE Associate）、AI、データエンジニアリング、ブロックチェーン

機械学習の民主化に必要なこと

データ活用が企業にとって生命線となっていくことに、疑いの余地はない。従来のシステムにおいては、情報系もしくは情報分析システムといった名称でデータ活用が行われている。流通小売業のPOS分析をはじめとして、その歴史は長く、これらはビジネスにおける重要な意思決定を支援している。BI製品の成熟が進み、過去から現在の状態を理解することができている。今後は「未来の予測」の精度が焦点となる。これまで人が経験で判断してきた領域を、機械により精度を高めていく取り組みが試行されている。

このような未来の予測は機械学習により実現されるが、それを遂行できるデータサイエンティストは人材不足の状況である。この課題に対し、誰もが未来の予測を行える製品が近年注目されている。これらは機械学習の「民主化」製品と呼ばれており、統計学やプログラミングなどの膨大な学習コストを肩代わりしてくれる。そして私たちの未来予測能力を高め、問題解決を手助けする。こうした民主化製品が、時代をけん引し始めている。

「DataRobot」に代表される自動機械学習（Automated Machine Learning）製品は、データセットを準備し、データセットの中から未来を知りたい項目を決め、解析を開始すれば、あとは全自動で最適な分析モデルが出力され、これから何が起こるかを高い精度で予測することができるものである。

自動機械学習は、その手軽さと高い予測精度から、トレンドとなっている。これを用いれば、高精度の未来予測を内製化できる。機密性の高いデータが用いられるため、自社で予測できるのは非常に価値がある。それゆえに機械学習の内製化への取り組みに着手したCIOも多いはずである。だが、その真の価値を享受するには必要なことがある。

鍵はデータプレパレーション

自動機械学習製品の真の価値を享受するために必要なのは、質の高いデータセットの準備である。未来スコアの予測精度を高めるには、欠損した値や異常な値などを補正する作

業が必要となる。加えて、入力可能なデータセットは1つに限られるという製品側の制約がある。1つのCSVファイルもしくはデータベーステーブルに、分析に必要な情報をまとめる作業が必要だが、マルチクラウド環境が当たり前の今、データはパブリッククラウド、プライベートクラウド、オンプレミスなど複雑かつさまざまな環境に分散していることが多く、データをまとめる難易度をさらに引き上げている。あらゆるスキルやツールを用いて、この分散化したデータを収集・加工し、未来スコアの予測に必要な質の高いデータセットを作成するデータエンジニアリングが、今後のデータ活用における重要な要素である。

機械学習用のデータセットを作成する作業は、一般にデータプレパレーションと呼ばれており、機械学習そのものよりも時間がかかることが多くの文献で紹介されている。この領域においても誰もができるような「民主化」が、機械学習の内製化実現における鍵である。

内製化のポイント

データエンジニアの立場から、データ活用の内製化に向けた民主化製品におけるデータプレパレーションのポイントを5点述べたい。

(1) マルチクラウドコネクタ

Amazon Web Services/Microsoft Azure／Google Cloud Platform/Oracle Cloud／オンプレミスなど、どの環境にも、またそこに格納されているどのデータソースにも、シームレスに接続できるマルチクラウドコネクタを保有していること。マルチクラウド環境下においてデータ活用を推進するには必須である。

(2) 大規模データへの対応

テラバイト級のデータを収集し、迅速に加工できること。クラウド環境におけるデータは肥大化する傾向がある。そのデータをプレパレーションする製品は、性能拡張性に富む分散処理基盤（Apache Hadoopなど）の上で動作できることが必須である。

(3) 対象データの可視化、加工支援機能

読み込んだデータを解析し、データ分布、欠損値情報、異常値情報などを分かりやすく可視化できること。加えて、豊富な加工機能で探索的なデータ解析を支援できること。

(4) 処理結果の連携機能

データプレパレーションの結果を自動機械学習製品に直接インプットするための、連携機能を保有していること。

(5) 元データの保護

元データに手を加えずに、データセットの作成を行えること。内製化を考える上で最も重要な点である。この点が担保されてこそ、ITシステム部門は安心してユーザーにデータプレパレーション環境を払いだせる。

野村総合研究所（NRI）はこのような要件を満たすデータプレパレーションの民主化製品としてTrifactaに着目し、SIパートナー契約を締結した。そして販売代理店契約を締結済のDataRobot（自動機械学習製品）と組み合わせて、未来予測の民主化にフォーカスしたコンサルティングおよびSIサービスを展開している。本稿が各企業における機械学習の内製化を促進していく上での一助になれば幸いである。 ■