ホワイトペーパー

Analyitcs 123: 大規模なエンタープライズ AI の実現

パイプラインのデカップリングで AI/ML の ROI を実現する方法

著者：テラデータ・コーポレーション テクノロジー（EMEA）担当バイスプレジデント Martin Wilcox、およびプリンシパルデータサイエンティスト（博士）Chris Hillman

04.22 / DATA ANALYTICS / ホワイトペーパー

teradata.

目次

ビジネスリーダーは、AI、マシンラーニングがまもなくユビキタス化し、各業界における競争優位の基礎となることを認識しています。マッキンゼーは、2030 年までに 70% の企業が少なくとも 1 つの形態の AI を導入すると示唆しています。[1] そのため、AI/ML 技術への投資は急速に増加し、同市場はさらに大きく成長すると予測されています。KPMG は、投資が現在の 124 億ドルから 2025 年には 1,500 億ドル近くに急増すると示唆しています。[2]

こうした投資や希望、期待にもかかわらず、多くの企業が AI/ML プロジェクトからリターンを得ることに苦労しています。世界中の経営幹部の 65% が、AI 投資に対する利益がまだ実現されていないと回答しています。[3]Brynjolfsson らは、その評価が高く、数多く参照されている学術論文において、「現代の生産性パラドックス」を指摘しており、AI/ML の潜在能力が十分に発揮されていないのは、「実装の遅れ」が原因であると結論付けています。[4]

1 https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy

2 https://home.kpmg/lu/en/home/insights/2019/04/khube-mag/intelligent-automation-edition/the-vast-world-of-intelligent-automation.html

3 https://www.forbes.com/sites/gilpress/2019/10/17/ai-stats-news-65-of-companies-have-not-seen-business-gains-from-their-ai-investments/#447ffcba19f4

4 Brynjolfsson E, Rock D, Syverson C, (2017) Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics; The National Bureau of Economic Research https://www.nber.org/papers/w24001.pdf

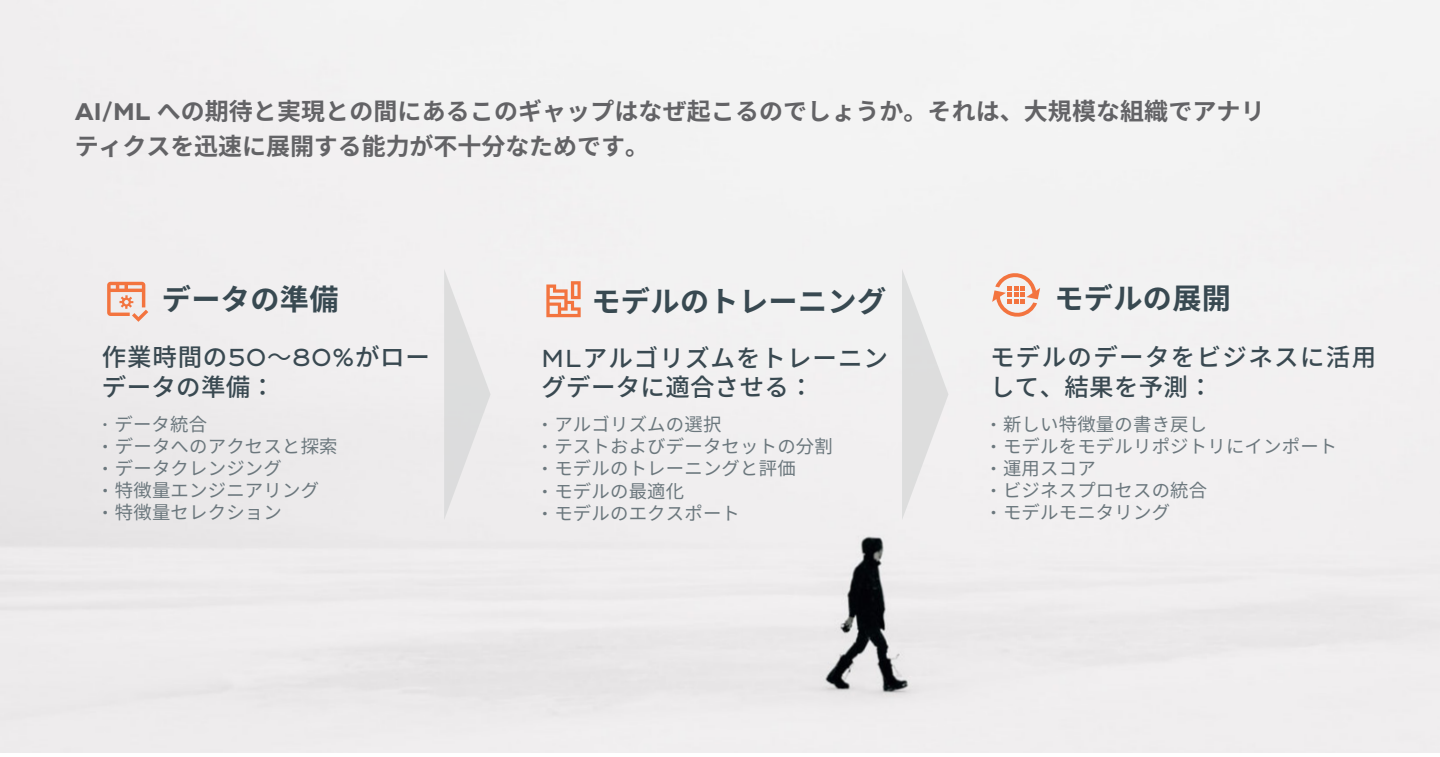
teradata.

期待と現実のギャップ

AI/ML への期待と実現との間にあるこのギャップはなぜ起こるのでしょうか。それは、大規模な組織でアナリティクスを迅速に展開する能力が不十分なためです。AI/ML には、何よりもまずデータが重要です。整理されたデータ[5]や分析データセットの構造と処理の標準化の重要性は以前から認識されていました。しかし、ツールやテクノロジー、データサイロ、「1 パイプライン・1 プロセス」という固定観念がこの分野の進歩を妨げてきました。また、大規模な組織で AI/ML の導入を成功させるために必要な専門知識は、十分ではありません。多くの企業が概念実証レベルのソリューションを実現できる一方で、本番規模でアナリティクスを展開することは数段難しく、成功した企業はほとんどありません。

この問題の根本原因は、非効率的で最適化されていないプロセスにあり、そのため企業はデータ資産、テクノロジー、スキルセットに投資を行ってもメリットが得られません。

今日の競争を勝ち抜いて、生き残ることはもちろん、アナリティクス主導の未来型企業となるために、組織は今すぐ行動を起こして、AI の強固な基盤となる柔軟で反復可能、説明可能なデータプロセスを構築する必要があります。テラデータの Analytics 123 戦略は、ビジネスとアナリティクスの両リーダーのためのわかりやすいロードマップを確立し、AI/ML プロジェクトが期待に応え、真のビジネス価値を提供できるよう、堅牢で効率的かつ容易に導入できるプロセスを構築します。Analytics 123 は、アナリティクスプロセスのさまざまな要素を切り離し、それぞれに適切なウェイトを与えます。ステップ 1 では、再利用を中心とした特徴量エンジニアリングです。ステップ 2 では、データサイエンティストが使い慣れたツールを使用して、ビジネス価値をもたらす予測モデルを作成します。ステップ 3 では、これらのモデルを展開して、ライブデータをスコアリングします。



5 Wickham H (2014) Tidy Data, Journal of Statistical Software https://www.jstatsoft.org/article/view/v059i10

パイプラインの終焉

今日、ほとんどの組織は、緊密に統合された「パイプライン」アプローチによってアナリティクスプロジェクトを進めています。パイプラインとは一般に、プロジェクトごとの問題解決を目的として設計および構築されたエンドツーエンドのプロセスです。そのためにまずソースデータを使ってコードを作成し、特徴量エンジニアリング（別名、データラングリング）を行います。このアプローチは、小規模なテストや研究・実験プロジェクトであればうまく機能します。リソースは集中して効率的に使用されます。このアプローチは、パイプライン全体をバージョン管理リポジトリ（git や svn など）にコードとして格納して再現性を確保できます。これにより、実験や分析を行うたびに、いつでも同じ結果を安定して再現することができます。

しかしこのアプローチを企業レベルにまで拡大すると、非効率なプロセスがすぐに発生し、データやコードのサイロ化を招きます。個々のチームは、同じデータからほとんど同じ特徴量を得ながらも、それぞれのパイプラインの中でサイロ化され、サポートする予測モデルと密接に関連しているため、重複した作業を行うことが多くなります。そのため、データサイエンティストの生産性が低下し、予測モデルの構築ではなくデータラングリングに 50 ～ 80%の時間を費やしているのが現状です。[6] 特徴量は無秩序に作成、利用されています。他のチームが関連する問題に取り組んでいて、必要なデータラングリングの多くがすでに完了している可能性があってもそれに気付きません。予測的アナリティクス、処方的アナリティクスの需要が高まる中、このような「データラングリングのオーバーヘッド」は、持続することができません。

このような生産性の低さとコストの高さは、アナリティクス投資に対する利益が期待できず、データ主導型の組織への進展が停滞する一因となっています。重複した作業によって費用がかさむだけでなく、プロジェクトに要する期間が長くなって市場投入まで時間がかかります。結果、企業全体に対してマシンラーニングの価値が及ぼす影響が小さくなるとともに、マシンラーニングの価値に対する信頼が損なわれます。本番システムにおけるアナリティクスは対象者によってさまざまな意味を持ちますが、ビジネスにとっては、次善のオファー、解約削減戦略、小売価格の変更などの意思決定を行うにあたって、モデルの結果が信頼され、繰り返し使用されて価値を創造することを意味します。高度なアナリティクスモデルを含むすべての本番システムは、スケーラブルで、パフォーマンスが高く、堅牢で、保守が容易で、安全でなければなりません。しかし、残念なことに、そのようなシステムはほとんどありません。アナリティクスアプリケーションの市場投入までの時間は数か月に及び、多くの場合、本番稼動に至りません。Gartner は、アナリティクスイニシアチブの失敗率が 80% を超えると推計しています。[7]

パイプラインジャングルにおけるデータ負債

予測的アナリティクス、処方的アナリティクスの需要が高まる中、このような「データラングリングのオーバーヘッド」は、持続することができません。

アナリティクスプロジェクトが成功し、企業に展開されても、パイプラインのアプローチは将来的に問題を引き起こす可能性があります。コードとして保存されたパイプラインは、元の作者以外にはすぐに解読不能になる可能性があるからです。Google はこれらを「パイプラインジャングル」と表現し、[8] データ依存はマシンラーニングシステムの技術的負債の主な要因の 1 つであり、データ依存はコード依存よりも高コストであることを指摘しています。データサイエンティストの平均在職期間が 1 年未満という高い離職率を考えれば、[9] 新入社員が前任者の作成したモデルやモデルトレーニングを使用、適応、理解できることは必須です。

この実現を阻害する障壁の 1 つは、データサイエンティストが予測モデルの作成に数多くの言語を好んで使用することです。データサイエンティストは好奇心旺盛で探究心が強く、常に新しいツールやプロセスを研究しながら、最新のスキルや技術を網羅しようと知識を広げているイメージがあります。注目を集めるマシンラーニングや AI の分野では、新しいアナリティクスツールや言語、フレームワークが次々と登場し、データサイエンティストたちがそれを貪欲に試そうとします。

6 https://www.information-age.com/productivity-in-data-science-123482699/ and Dasu, T, & Johnson, T (2003) Exploratory, Data Mining and Data Cleaning (Vol. 479), John Wiley & Sons.

7 https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/

8 https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

9 https://www.indeed.co.uk/salaries/data-scientist-Salaries#:~:text=The%20typical%20tenure%20for%20a%20Data%20Scientist%20is%20less%20than%201%20year.

teradata.

一方で、データサイエンスのコミュニティは非常に多様で、さまざまなグループが自分の持っているスキルに固執し、その分野の専門家になることに集中する傾向があります。Python のコーダーは Pythonic のプログラミング手法に特別な忠誠心を持ち続け、R のユーザーは長い時間をかけて開発したスクリプトのライブラリだけで仕事をしています。チームは、使い慣れたツールを使用する方が、効率的で快適、かつ創造的に作業を進めることができます。異なるコーディング言語や方法を導入しようとすると、生産性が低下したり、反対されたりする可能性があります。

市場が成熟するにつれ、現在「人気の」ツールが支持を失うこともあります。そのような状況の中で「勝者」を選ぶことは、リスクと不確実性を伴います。

モデルトレーニング作業を単一のテクノロジーに限定することも、あまり望ましいことではありません。モデルトレーニングツール、言語、フレームワーク化の中で、市場で支配的な地位にあるものは存在しません。市場が成熟するにつれ、現在「人気の」ツールが支持を失うこともあれば、ニッチなニーズを満たすツールが登場したり、多機能製品にツールが統合される可能性もあります。そのような状況の中で「勝者」を選ぶことは、リスクと不確実性を伴います。さらに、大規模で多様な組織では、アナリティクスに対してさまざまな要件があります。それらの要件をすべて満たし、客観的に見て「最適」な単一の技術は存在せず、多くの場合、複数のライブラリ、メソッド、言語を併用することで、適切な結果を得ることができます。

ビジネスチームのリーダーとアナリティクスチームのリーダーは、ジレンマに直面しています。パイプラインによるアナリティクスへのアプローチは、組織が今日の競争に打ち勝ち、明日のデジタル経済のニーズに応じて変革する能力をますます脅かします。モデルの構築とトレーニングを偏重した従来の作業方法、そしてデータ準備とアナリティクスの本番運用に対するアプローチの不備が、ビジネス部門全体への AI/ML 導入が遅々として進まない要因です。

同時に、すべての業界で監視が強化されているため、AI/ML のあらゆるプロセスが監査可能かつ透過的であることが求められます。特に規制の厳しい業界では、AI/ML システムが将来の任意の時点から特定の予測を行った理由と方法を組織が理解し実証できることがますます重要になっています。はるか昔に退職したデータサイエンティストによって作成された機能や予測モデルがコードにまとめて保存されている場合、後からそれを解明することは不可能です。

それに加えて、企業はデータと世界の変化に機敏に対応し、また最新のツールとトップクラスのデータサイエンティストが持つスキルセットを柔軟に使用できる必要があります。最新のディープラーニングライブラリや最新の分析言語とライブラリを、現在の課題に対応したスイートスポットに素早く取り入れるためには、動的なアプローチが必要です。プロジェクトごとにパイプラインを開発し、単一の分析ツール、言語、フレームワークで標準化することができないという状況に直面している企業は、堅牢でエンタープライズグレードの分析の展開をどのように拡大できるでしょうか？

1モデルにつき1パイプラインというアプローチ

teradata.

意図的なデカップリングによるリソースの集中

その答えは、プロセスのさまざまな部分を「デカップリング」し、代わりに特徴量エンジニアリング、モデルトレーニング、デプロイの 3 つの重要なコンポーネントに集中的に取り組むことです。それぞれの要素について簡単に説明すると、Analytics 123 は、特徴量エンジニアリング、再利用、デプロイのための最適なテクノロジーと、モデルトレーニングのための幅広いツールの選択肢を組み合わせることによって、自由とガバナンスを両立させようとしています。

ちょうど現代の IT アーキテクチャでコンピューティングからストレージが分離されているのと同じように、アナリティクスプロセスの個々の部分を分離することでより効率的なシステムが実現し、「Polyglot プログラミング」の原則がサポートされ、[10] 適切なツール、言語、フレームワークが、最も適したタスクに適用されます。

興味深いことに、現在アナリティクスの活用に成功している組織は、単一のアナリティクスツールの標準化よりも、アナリティクスバリューチェーンの両端を占める活動に特に重点を置いて、エンドツーエンドのアナリティクスプロセスを適正化することに注力しています。特徴量エンジニアリングプロセスを再構築し、複数のモデルのトレーニングやスコアリングが可能な再利用可能な特徴量の作成に集中的に取り組むことで、データラングリングにかかる総時間を大幅に削減します。データサイエンティストがモデルの作成とトレーニングに適したツールを選択してシームレスにインポートできるため、大規模なライブデータの展開とスコアリングが向上します。この戦略により、重複が排除され、異なるプラットフォーム間でのデータ移動が不要になって、モデルの強固な監査、監視、更新が大規模に実行できます。

AI/ML を組織内で検討、実装、展開する場合、最近の傾向として、アナリティクスプロジェクトの中核をなす予測モデルの作成とトレーニングに過剰な焦点が当てられています。予測モデルはデータサイエンティストの役割の中でもエキサイティングで「セクシー」な部分と考えられていますが、実際にはプロジェクト全体のほんの一部でしかありません。Analytics 123 の基本的な考え方は、組織が AI/ML イニシアチブを成功裏に拡大するには、モデルトレーニングの両脇にある重要な要素、すなわち特徴量の再利用とモデルの展開にもっと注意を払う必要があるというものです。そのため、組織全体の AI/ML の基盤として、Enterprise Feature Store（EFS）の構築と維持に大きな力を入れる必要があります。

組織が AI/ML イニシアチブを成功裏に拡大するには、モデルトレーニングの両脇にある重要な要素、すなわち特徴量の再利用とモデルの展開にもっと注意を払う必要があありります。

Enterprise Feature Store

EFS は、予測価値が証明された変数の厳選されたコレクションであり、分析用 RDBMS のテーブルとして実体化されています。データ準備、データ統合、特徴量エンジニアリングという厄介で時間のかかる作業を一度行って特徴量を作成し、複数のさまざまなモデルのトレーニングやスコアリングに再利用できます。予測値だけでなく実用性のある特徴量を作成し、それぞれの特徴量をカタログ化するには時間と注意が必要ですが、この初期投資はすぐに報われ、後続のプロジェクトでは、十分に文書化された既存の特徴量を簡単に再利用できます。これは非常に重要です。AI/ML が期待された価値を提供するためにはユビキタス化が不可欠であり、現在データ準備と管理に費やされている 80% のコストと労力を大幅に削減することを意味します。

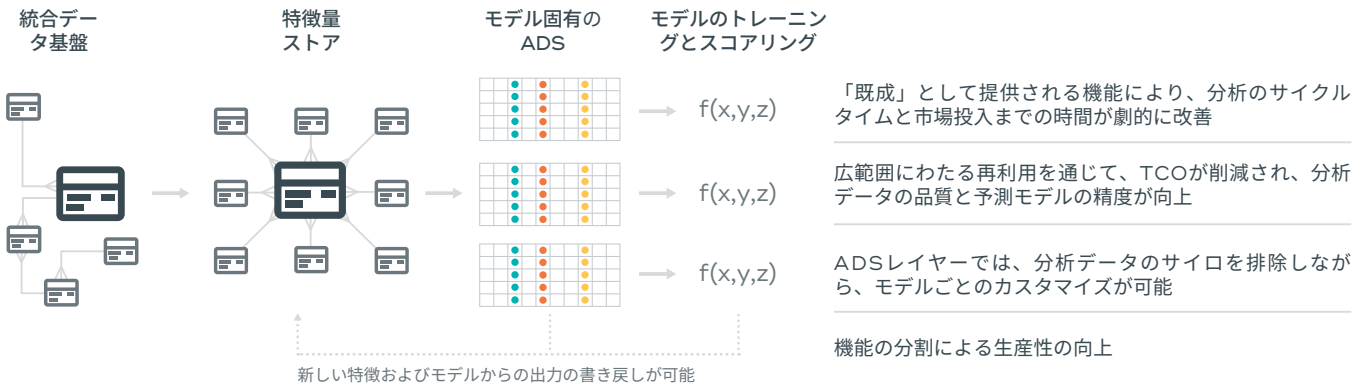
これらの特徴量の正確なカタログ化と、EFS 内での継続的なメンテナンスも一貫性に貢献します。共通の特徴量を使用したさまざまなモデルは、予測の妥当性に確信を持ってスコアリングできます。定期的なテストと更新により、モデルのドリフトを管理できます。

Enterprise Feature Store は、すでに大手企業において、データサイエンティストの生産性と新しい分析の価値創造までの時間を劇的に向上させています。EFS は、大規模な分析データセットの操作と処理において業界をリードするテラデータのパフォーマンスと拡張性を活用し、データ移動とデータの重複を回避して総所有コスト（TCO）とレイテンシーの両方を削減することにより、ビジネスとコスト面でのメリットももたらします。

10 http://nealford.com/memeagora/2006/12/05/Polyglot_Programming.html

teradata.

特徴量ストアのアプローチ



第 2 ステップであるモデルのトレーニングは、データサイエンティストの本領であり、それゆえに非常に注目されています。実際、モデルの作成に過度に重点を置いてきたことが、データサイエンティストの生産性を低下させる一因となっています。大規模な本番アナリティクスが予測モデルで始まるわけでも終わるわけでもないのに、この段階の両側にある重要なプロセスを軽視しているために、広範な展開が遅れる要因であることは明らかです。

モデルトレーニングの幅広い選択肢

モデル作成に関しては、当面の間、企業は複数の異なるテクノロジーを使用する必要があると思われます。賢明で機敏な企業は、価値を生み出す可能性のある特定のツールやテクノロジーを排除したり、義務付けたりする前に、よく考えることをお勧めします。データサイエンティストが確実で定量化可能な ROI を実現するには、データとアルゴリズムを自由に探索して、堅牢で正確なモデルを構築できる必要があります。この自由には、さまざまなツールを使用できる自由も含まれます。前述のように、データサイエンティストにはそれぞれの好みがあり、最適なツールは状況によって異なります。また新しいツール、言語、アプローチが日々考案されています。Analytics 123 では、データサイエンティストが特定のユースケースに最適なアプローチを選択できるように、複数のベンダーの複数のツールや言語の使用が明示的に許可されています。ストアおよび作成された新しい機能は、他のユーザーが再利用できるように EFS に追加する必要があります。ただし、モデルのトレーニングに必要なデータは、特徴量ストアに保存されている変数を再利用する必要があります。また新たに作成された特徴量は、他の人が再利用できるように EFS に追加します。

モデルの作成を別の作業として扱うことで、外部システムでトレーニングされたモデルを、データベース自体で作成されたモデルと一緒にシームレスに取り込むことができます。

データサイエンティストの間では、「モデル」とはアルゴリズムをデータでトレーニングした結果なのか、トレーニングしたアルゴリズムとトレーニングデータを作成した特徴量で構成されるものなのかについて議論があります。Analytics 123 では、特徴量エンジニアリングとモデルトレーニングは、2 つの別個の活動として扱われます。モデル作成の反復的な性質から、発見と評価の段階ではこの2つの活動は本質的に結びついています。しかし、モデルが作成され、正確であることが示された後は、特徴量エンジニアリングコードを特徴量ストアに移行し、特定のモデルと無関係のコードとして扱う必要があります。

モデルの作成を別の作業として扱うことで、外部システムでトレーニングされたモデルを、データベース自体で作成されたモデルと一緒にシームレスに取り込むことができます。データサイエンティストは、予測価値が証明された特徴量を文書化し、モデリング言語とトレーニングプラットフォームを選択するだけでなく、業界をリードするパフォーマンスと拡張性を備えたテラデータ・プラットフォームに簡単に移植できるため、両者の長所を活かすことができます。

この「Bring Your Own Model (BYOM)」のアプローチにより、Analytics 123 の重要な第 3 ステップを開始します。究極的に

言って、アナリティクスプロジェクトの価値は、実際のデータに対して予測を行い、タイムリーで実用的なビジネスインサイトを提供することで初めて実現します。モデルトレーニングの作業は、通常、過去のデータから慎重に選択したサンプルを使用して行われます。これとは対照的に、モデルスコアリングプロセスでは、完全で最新のデータセットにアクセスする必要があります。このプロセスは、多くの場合、ミッションクリティカルであり、予測値を運用エンドポイントで提供する必要があり、ほぼリアルタイムで実行されることが増えています。モデルトレーニングからモデルスコアリングに移行する際の課題は過小評価されることが多く、その結果、アナリティクスプロジェクトの失敗の主な原因となっています。Teradata Vantage の高可用性と業界をリードする混合ワークロード管理機能は、アナリティクスの運用で組織が直面する多くの課題を回避します。

BYOM により、データサイエンティストは最適なツールを使用して予測モデルをトレーニングすることができ、Enterprise Feature Store の本番データに対して直接、大規模なスコアリングを行うことができます。緊密な統合と、PMML、SQL 変換、データベース内で実行されるネイティブ・コードなどのさまざまな方法によって、外部でトレーニングされたモデルをデータベース内の本番でスコアリングし、大規模に展開することができます。

本番データによる価値創造

ライブ本番データのスコアリングは、ビジネスで定期的に使用される予測を作成します。これが、AI/ML が生み出す真の価値であり ROI です。プロセスの本番段階は、シンプルで堅牢であることが重要です。EFS とトレーニング済みモデルがあれば、必要なものはすべてデータベース内に存在しているため、外部システムとの間でのデータ移動は必要ありません。さらに、テラデータのシステムは、通常、複数チャネルで運用中のエンドポイントに直接接続されており、数十〜数百ミリ秒単位の応答時間でほぼリアルタイムのモデルスコアリングを特徴づける「戦術的な」クエリに対応している点が強みとなっています。テラデータの「常時並列」アーキテクチャでは、バッチスコアリングのワークロードが高性能と拡張性の両方を備えています。そのため、必要に応じて何度でも本番データ上に新しい予測値を作成できます。

EFS とトレーニング済みモデルがあれば、必要なものはすべてデータベース内に存在しているため、外部システムとの間でのデータ移動は必要ありません。

モデルスコアリングは、通常、「驚異的並列」プロセスの一例です。また、テラデータでは論理的なハッシュベースのファイルシステムが採用されているため、テラデータでのほぼリアルタイムのスコアリング操作は基本的に「単一 AMP、単一（論理）IO」操作であり、CPU と IO リソースをほとんど消費しません。完全自動化システムのコアには、モデルドリフトの定期テスト、再トレーニング、新たなモデルの本番リリースに使用するチャンピオン / チャレンジャー方式などが含まれており、これらを中心としたシステムを構築することができます。

近い将来、組織は競争力を維持するために、膨大な数の予測モデルを本番に展開し、ユビキタスなマシンラーニングをサポートすることが必要になります。そのためには、AI/ML に対する戦略的アプローチによって、現在のリソースとコスト要件を大幅に削減し、展開に要する時間を短縮する必要があります。ROI の最大化だけでなく、「データ負債」と監査の悪夢を生み出す「パイプラインジャングル」を回避するために、データサイロ、断片化されたシステム、重複作業をすべて回避しなければなりません。これらはすべて、拡張性とパフォーマンスに優れたデータプラットフォームを基盤としている必要があります。

teradata.

テラデータの常時並列アーキテクチャと処理モデルは、多くのデータ準備、モデルトレーニング、モデルスコアリング作業の特徴である大規模で複雑なデータセットの高性能処理に最適です。テラデータは、世界有数の規模と極めて洗練されたアナリティクスシステムを誇る複数のお客様の、非常に要求の厳しい本番環境において、マシンラーニングを垂直方向 (100 万以上のオブザベーションでモデルをトレーニングし、1 日に複数回、250 万以上のオブザベーションに対してスコアリングを実施)、および水平方向 (何百万もの予測モデルをトレーニングしていわゆる「ハイパーセグメンテーション」と呼ばれるユースケースをサポートし、毎日スコアリングを継続) に拡大する能力を実証しています。

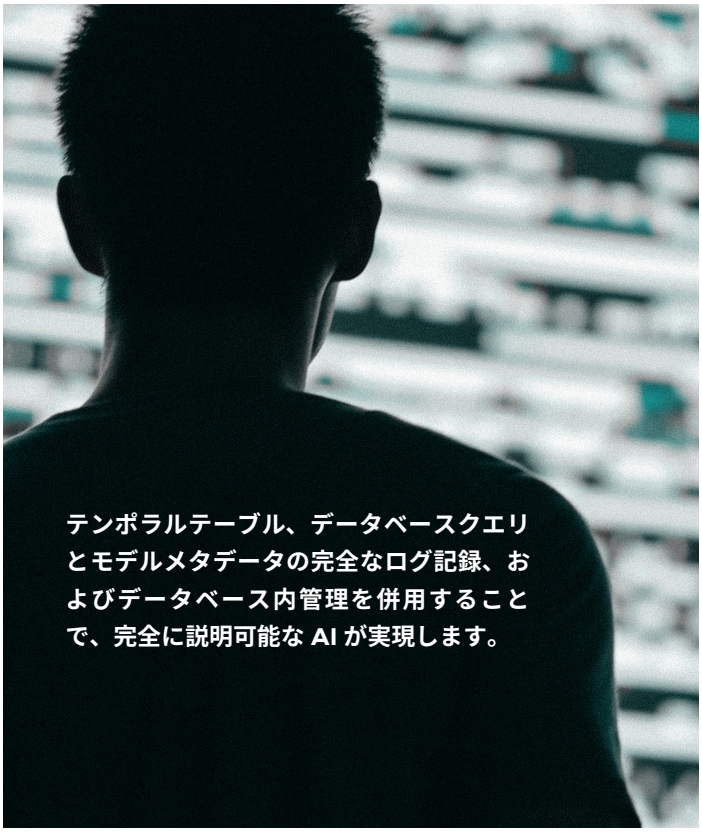
近い将来、組織は競争力を維持するために、膨大な数の予測モデルを本番に展開し、ユビキタスなマシンラーニングをサポートすることが必要になります。

テラデータはローカライズされたデータへの O(1) アクセスを可能にし、ほぼリアルタイムでのモデルスコアリングなどの戦術的なクエリにも対応できる非常に高いスループットと低レイテンシーを実現しています。業界をリードする混合ワークロード管理機能により、これらのミッションクリティカルな運用ワークロードを、複雑でリソース集約的な処理（データ準備やモデルトレーニングなど）と共存させることができるため、複数の冗長な重複するサイロに同じデータを置く必要がなくなります。テラデータの QueryGrid 仮想化フレームワークと Incremental Planning and Execution (IPE) テクノロジーにより、データレイクや分析エコシステム全体に永続化されたデータを透過的かつ実行的にクエリし、テラデータプラットフォームで管理されているデータと組み合わせることで、迅速かつ柔軟なデータ探索を実現します。

Analytics 123 のすべての要素をデータベース内で実行することができます。また、データサイエンティストがよく使うアナリティクス言語やライブラリがデータベース内で処理されるため、外部ツールで開発したモデルを再実装する必要がありません。テラデータの AnalyticOps アクセラレータは、CI/CD パイプラインの利用による Enterprise Feature Stores の維持やモデル管理の大規模化など、分析プロセスの完全自動化を可能にします。これにより、アナリティクスに要する時間が大幅に短縮されます。

テラデータは、オンプレミス、仮想プライベートクラウド、パブリッククラウドプラットフォームなどの展開オプションを組み合わせて選択できるコネクテッド・マルチクラウド / ハイブリッド展開機能を提供しています。テラデータは、基盤となるインフラストラクチャプラットフォームに関係なく、まったく同じ製品を提供しているため、既存のアプリケーションを再設計する必要がありません。この移植の容易さにより、組織が新しいプラットフォームモデルを採用する際の障壁が低くなるとともに、将来の出口戦略が提供されます。

前述のように、規制当局の監視やビジネスリスク管理により、AI/ML の透明性、文書化、一貫性、監査可能性の向上が求められています。テラデータでは、特定の瞬間の正確な状態でデータを表示できるテンポラルテーブル、データベースクエリやモデルメタデータの完全なログ記録、およびデータベース内での管理を組み合わせて使用することで、完全に説明可能な AI を実現できます。この情報を使用して、特定のモデルが特定の日付に特定の予測を行った理由を正確に調査できるため、監査可能性と再現性が確保されます。



テンポラルテーブル、データベースクエリとモデルメタデータの完全なログ記録、およびデータベース内管理を併用することで、完全に説明可能な AI が実現します。

もはや許容できない 80% の失敗率

パイプラインは、データサイエンティストにとって確立された作業方法であり、多くの組織で AI/ML 機能の構築に重要な役割を果たしてきたことは間違いありません。しかし、この手法がもはや目的にそぐわず、将来のデータ駆動型ビジネスのニーズに対応しないことは確かです。特徴量エンジニアリングプロセスの一環として、重要なデータ準備と統合に関する無駄、重複、膨大なリソースの浪費に対処する必要があります。80%の失敗率と数か月に及ぶプロジェクト期間は到底受け入れられません。これでは、真の価値を創出できる規模で AI/ML を実装しようという意欲がくじけてしまいます。主要企業は、今後数年間で数千万、数億の予測モデルの展開とスコアリングが必要になると予測されています。今日、わずか数千の予測モデルの展開ですら苦労しているアプローチがこれに対応できるはずもありません。

Analytics 123 は、その代替案を提示します。プロセスを 3 段階に分離することで、それぞれに必要な重みを配分し、集中的に取り組むことができます。特徴量は Enterprise Feature Store で再利用、文書化、カタログ化できるように設計されているため、重複が減り、効率と一貫性が高まります。データサイエンティストは、特定のタスクに最適と思われるツールや言語を自由に使用することができ、トレーニングされたモデルは、Enterprise Feature Store でライブデータをスコアリングするために簡単に企業に取り込むことができます。そして、そのスコアリングは、テラデータの超並列で高性能なエンタープライズ規模の機能を活用して、実際のビジネスクリティカルなアナリティクスを推進することで組織を変革できます。

テラデータについて

テラデータは、コネクテッド・マルチクラウド・データプラットフォームを提供する企業です。当社のエンタープライズアナリティクスは、あらゆる規模のビジネス課題を解決します。将来の大規模かつ混在するデータワークロードを今すぐに対処できる柔軟性を備えているのは、テラデータだけです。

Teradata Vantage のアーキテクチャは、クラウドネイティブであり、サービスとして提供され、オープンなエコシステムの上に構築されています。これらの設計上の特徴により、Vantage はマルチクラウド環境において価格パフォーマンスを最適化するための理想的なプラットフォームとなっています。詳しくは、Teradata.jp をご覧ください。

著者紹介

Martin Willcox は、テラデータ EMEA 地域のテクノロジー担当 VP です。ヨーロッパ、中東、アフリカ全域における、テラデータソリューション / サービスの販売促進の共同責任者です。

Chris Hillman は、テラデータの国際的な高度な分析チームに属するプリンシパルデータサイエンティストで、ロンドンを拠点としています。これまで 20 年以上、多くの業界でアナリティクスに携わってきた経験があります。

17095 Via Del Campo, San Diego, CA 92127 Teradata.jp