企業における データサイエンス実装のポイント

田村　初

CONTENTS

要　約

1　データサイエンスの技術を自社ビジネスに応用したいという機運は高まっているものの、本格的に運用して新たな利益を生み出すレベルにまで活用できている企業はごく一部である。

2　データサイエンスのプロジェクトには独特の難しさがある。「問題設定フェーズ」「問題解決フェーズ」「実装フェーズ」「フェーズ全体の進め方」のそれぞれにおける難しさを説明する。

3　データサイエンス独特の難しさを克服し、円滑にプロジェクトを遂行するには、「ハイブリッド型人材を育てる」「現地現物にこだわる」「自己研鑽を奨励する」「知恵を絞り困難に挑戦する」の４つがポイントである。

4　日本企業は、本来、データサイエンスが得意である。擦り合わせを重ねて精度を高めるアプローチや、貴重なクローズドデータの蓄積を武器とすれば、GAFAに続くような革新的成果も生み出せるであろう。

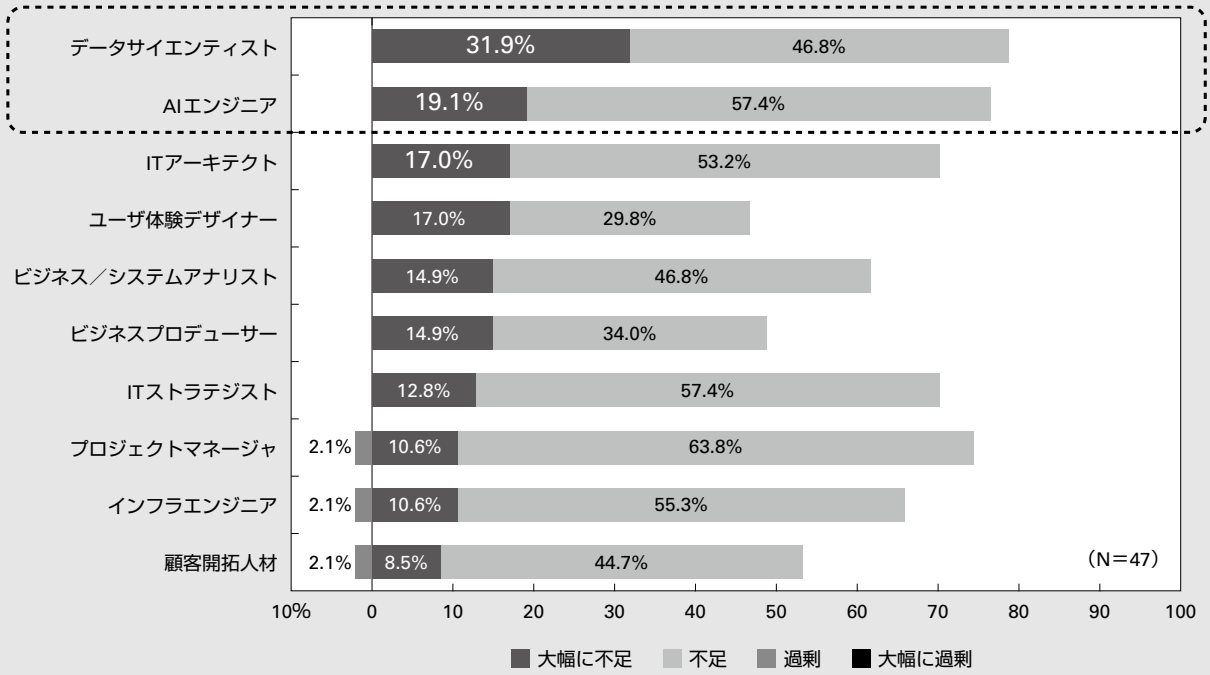
Ⅰ データサイエンスの重要性の高まり

2016年３月、DeepMind社が開発した囲碁AIである「AlphaGo（アルファ碁）」がトップ囲碁棋士であったイ・セドル九段（当時）に勝利したニュースは、世界に衝撃を与えた。あれから５年、GAFA（Google、Apple、Facebook、Amazon）はビッグデータを使ったビジネスで躍進し、利益を伸ばしている。追随するように、世界中の多くの企業でデータサイエンスの技術を自社ビジネスに応用したいという機運が高まっている。

国内企業も、1990年代以降のIT化の進行もあり、既に自社内には多くのデータが蓄積されており、それを活用する余地は大きいと考えられる。2021年３月に、野村総合研究所（NRI）が国内の情報・デジタル子会社を対象として実施したアンケート調査によると、「大幅に不足」している職種の上位２つがデータサイエンティストとAIエンジニアであった（図１）。この結果からも分かるように、国内企業の多くは、データサイエンスの活用に全く未着手というより、既に何らか動き始めているというのが現状だろう。

しかし、データサイエンスの技術を本格的に運用し、新たな利益を生み出すレベルにまで活用できている企業は、まだごく一部にとどまっているのではないか。総務省の「令和２年版情報通信白書」においても、「デジタルデータの活用状況（個人データ除く）」という設問で、既に活用していると回答した企業の割合は米国55％、ドイツ53％に対し、日本は23％と低い[注1]。本稿では、多くの企業が注力しているにもかかわらず、データサイエンスがインパクトのある成果をなかなかもたらさない原因を分析し、その克服と、実装のためのポイントを考察したい。

図1　人材種類別の過不足感



※「過不足感はない」「必要としていない」はグラフから除外
出所）野村総合研究所「情報・デジタル子会社における今後の方向性と課題に関する調査」（2021年3月）より作成

Ⅱ データサイエンスプロジェクトの難しさ

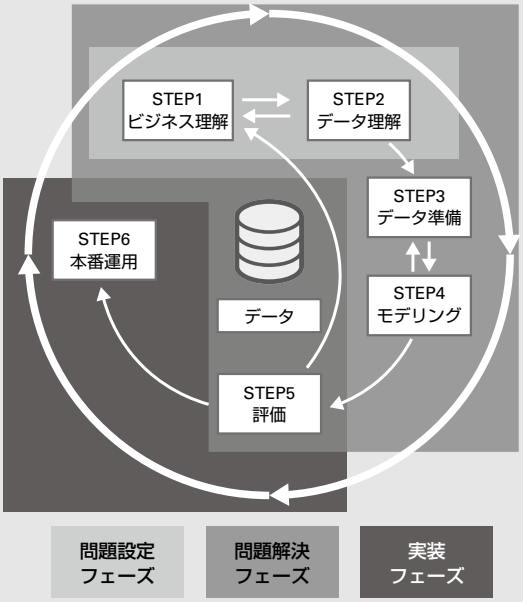
まず、本稿で述べる「データサイエンスの実装」とは何であるかを定義したい。データから何らかの価値を引き出すものを「データサイエンス」とすると、その意味は幅広い。ここでは意味を限定し、「データサイエンス技術を用いて、持続的にビジネス的価値を生む仕組みを構築すること」とする。単なるアドホックな分析業務は含まず、それがビジネスとして利益を生み出しており、かつ、自動化されて業務に組み込まれている状態を「実装」と定義したい。また、このようなデータサイエンスの実装を目指すプロジェクトを「データサイエンスプロジェクト」と呼ぶことにする。

では、このデータサイエンスプロジェクトは具体的にどのようなプロセスで進めるべきものか。さまざまなフレームワークが提唱されている中で、代表的な一つに「CRISP-DM（CRoss-Industry Standard Process for Data Mining）」と呼ばれるものがある。CRISP-DMは、1999年に、旧ダイムラー・ベンツ、旧SPSS、NCR、OHRAがメンバーとなっているコンソーシアムで開発された、データマイニングのためのプロセス論であるが、現在でもさまざまなデータサイエンスプロジェクトで活用できる指針である。CRISP-DMでは、プロジェクトのプロセスを6つのSTEPに分けており、各STEPで行うべき内容が定義されている（図2）。

本章では、データサイエンスプロジェクトの難しさを論じるにあたり、CRISP-DMを大きく3つのフェーズに分ける。1つ目は、STEP1（ビジネス理解）とSTEP2（データ理解）を行き来する「問題設定フェーズ」、2つ目は、STEP1（ビジネス理解）からSTEP5（評価）までの一連の流れを繰り返す「問題解決フェーズ」、3つ目は、STEP5

図2　CRISP-DM

CRISP-DMで定義される6つのSTEP



STEP1〜6でやるべきこと

STEP	やるべきこと
STEP1 ビジネス理解	• ビジネス目標の決定 • 状況の評価 • プロジェクト計画の策定
STEP2 データ理解	• 初期データの収集 • データの探索 • データ品質の検証
STEP3 データ準備	• データの選択、前処理 • データの新規作成 • データの統合
STEP4 モデリング	• モデリング手法の選択 • 評価設計 • モデルの作成、評価
STEP5 評価	• 結果の評価 • プロセスの見直し • 次ステップの決定
STEP6 本番運用	• 展開の計画 • 監視と保守の計画 • 最終レポート

（評価）を終えてSTEP6（本番運用）までを完遂する「実装フェーズ」とする。

データサイエンスプロジェクトは多くの企業にとってなじみのないものであり、独特の難しさがある。この特性に留意して進めないと、後続フェーズに進めない、あるいは実装まで至っても当初の目論見が外れてビジネス価値を生まないという事態が起きる。この3つのフェーズそれぞれで、具体的にどのようなケースでつまづいてしまうのかを論じたい。

1　問題設定フェーズにおける難しさ

第一に、問題設定フェーズにおける難しさについて、代表的なものとして次の3つを挙げる。

（1）事前のROI算定が困難
（2）「多産多死」への誤解
（3）ビッグピクチャー症候群

(1) 事前のROI算定が困難

データサイエンスプロジェクトはほかの投資プロジェクトと同様、最終的な評価を利益貢献（ROI）で測らなければならない。しかし、データサイエンスプロジェクトでは、事前にROIを見極めるのは難しい。一般的なIT投資プロジェクトは、主に収益面はビジネススキルで見極め、投資面はエンジニアリングスキルで見極める。しかし、データサイエンスプロジェクトの場合、収益面を見極めるには、ビジネススキルに加えてデータサイエンススキルが必要であり、投資面を見極めるときも同様に、エンジニアリングスキルに加えてデータサイエンススキルが求められる。つまりROIを適切に算定するには、この3つのスキルをバランス良く備える必要がある。しかし、これらのスキルをハイブリッドに兼ね備えた人材は希少であり、大半の企業では十分に確保できていない。

さらに、利益への意識が希薄なデータサイエンスプロジェクトが生まれてしまうもう一つの理由として、AIを使うこと自体が目的化しているケースがある。経営層の立場からすると、競合他社が機械学習で目覚ましい成果を上げたという事例が報道されれば、自社でも同様の取り組みをしたいと考えるのは自然であろう。しかし、その会社が具体的にどの程度のコストをかけて、どのようなデータをどのような頻度で集めたか、あるいは作り上げたモデルの更新をどのように行っているかなどは開示されていないケースが多い。

実際のデータサイエンスプロジェクトにおいては、競合他社で導入されたものが必ずしも自社でなじむとは限らない。自社でも同じようなデータが準備でき、同じような運用が可能であるかどうかは、やはりデータサイエンススキルを持たないと見極められない。ROIの算定をおろそかにしたままで進行しても、途中で投資効果が得られないという判断がなされ、プロジェクトが頓挫してしまう。あらかじめ、データサイエンスと利益がどう結び付くかをしっかり考えることが重要である。

(2) 「多産多死」への誤解

データサイエンスのような新しい分野のプロジェクトでは、よく「多産多死であるべきだ」などといわれる。しかしこの「多産多死」という言葉を、案件の見極めや技術理解を放棄する方便として使っているケースが見

られる。

データサイエンス技術を用いれば、やりたいことを魔法のようにできると思ってしまいがちであるが、実態は異なる。確かに、データサイエンス技術は応用範囲が非常に広く、やろうと思えば何でも案件化できる。一方で、データサイエンス技術は決して万能ではない。最新の機械学習技術を用い、さまざまな設計の工夫を行ったとしても、基本的には与えられたデータの特定の訓練ロジックに当てはめ、最適な関数を導出するものにすぎず、特に与えられるデータの良し悪しでほぼ達成可能な精度の上限は規定される。

たとえば、業務上99％の精度が要求されるものに対し、現状で入手可能なデータではせいぜい90％程度の精度が限界であるようなケースにおいて、残りの９％を埋めるのは現実的に困難で、解決不可能である。企業は、現在のデータと技術で十分に解決可能な案件を探し出し､そちらに注力しなければならない。

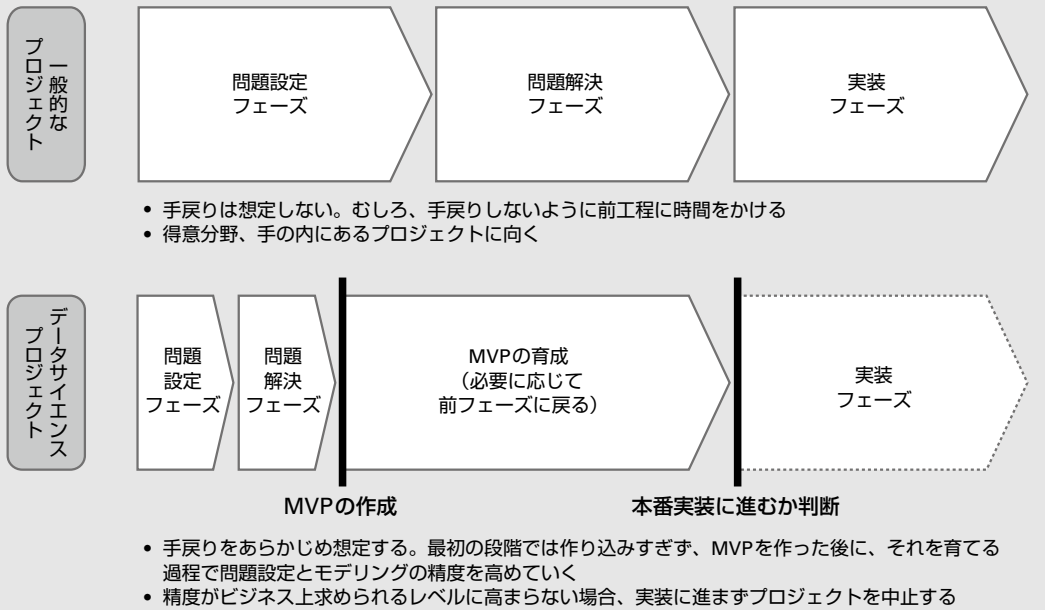
データサイエンスプロジェクトは、幅広く案件化できる一方で、このデータと技術の制約により、本当に解決可能な問題は実は少ない。つまり、分母は大きく分子が小さいため、結果として成功確率は極めて低く､「多産多死」アプローチは非効率である。また、データサイエンスプロジェクトは、後述するが、一般的なプロジェクトと異なり、何度も手戻りを繰り返しながら精度を向上させていくものであり、そこに多くの時間とコストがかかる。案件の見極めをせず闇雲に広げてしまうと、せっかく「当たり」の案件を引いても、十分な精度を得られるようになる前にプロジェクトが中止となってしまうことがある。最初の段階で筋の良い案件を見極め、ある程度まとまった資源を投入することが求められる。

(3) ビッグピクチャー症候群

多産多死とは逆に、案件の見極めに時間をかけすぎてしまうという問題も見られる。たとえば、上流工程で自社のデータ活用のあるべき姿を定義し、それにしたがって後続工程を進めようとするケースがある。これはプロジェクトの進め方として、一般的なウォーターフォールモデルに慣れている企業で起きがちである。ウォーターフォールモデルは、上流工程でしっかりと全体構想を描き、タスクを細分化して、後続工程を順序立てて進めていこうとする進め方である。上流工程に時間をかけ、課題を潰し込み、やるべきことをしっかり整理することがプロジェクトの成否を分ける。

しかし、データサイエンスプロジェクトでは上流工程での整理に時間をかけると、後続でモデリングを繰り返す時間が不足する。最初に大きな構想を描き、それを少しずつブレイクダウンして進めていくアプローチは、自社にある程度のノウハウがあり、勝ち筋が見えているケースにおいて有効なアプローチである。逆に、ノウハウを持たないプロジェクトでこのような進め方をすると、途中の試行錯誤プロセスで必要とされる工数が確保できずに、結果として実用可能な精度のあるモデルを作ることができないという事態が起きる。特に大企業になるほど、さまざまな部署、さまざまな業務プロセスが存在するため、上流工程の整理に時間がかかる。結果として上流工程で工数を多く消費してしまい、十分に時間をかければ出せたはずのモデル精

図3　データサイエンスプロジェクトの進め方の特性



度が出せず、プロジェクトが頓挫してしまう。

精度を高める工程に時間をかける重要性を考えると、MVP（Minimum Viable Product：実用最小限のプロダクト）と呼ばれる最小モデルを最速で作ることがポイントになる。MVPを安く速く作り、これを育てていく。MVPを育てる中で、当初設定した問題が真の課題と合致していなかったと気付くこともあり得るため、必要であればあらためて問題設定フェーズに立ち返る。これもMVPを評価したからこそ分かるケースも多い。データサイエンスプロジェクトを成功させるには、上流工程に時間をかけすぎず、MVPの検証をしながら問題設定を見直していくアプローチが求められる（図3）。

2　問題解決フェーズにおける難しさ

第二に、問題解決フェーズにおける難しさについて、代表的なものとして次の3つを挙げる。

（1）訓練データへの過剰適合
（2）ビジネスロジックと合わないモデリング
（3）不適切なデータサイエンス手法の採用

（1） 訓練データへの過剰適合

よくある問題解決フェーズの失敗として、構築モデルが訓練データに過剰適合（オーバーフィッティング）し、本番データの予測精度が著しく悪くなるというケースがある。機械学習技術では、あらかじめ定義した誤差関数に対し、その誤差を最小化するように繰り返し飽和するまで、内部パラメータを最適な値に更新する。データに対する適合力が極めて強い特徴があり、伝統的な統計手法と比べると相対的に過剰適合が起きやすい。過剰適

合を防ぐ標準的な手法として、訓練データと検証データの適切な分割（クロスバリデーション）を行うが、このクロスバリデーションにもさまざまな手法があり、選定した手法が不適切であると過剰適合してしまう。

また適切なクロスバリデーションを設定していたとしても、本番時には使えないデータを訓練時に使用してしまうリーク（データ漏れ）と呼ばれる問題により、過剰適合が起きることもある。たとえば、小売業の需要予測モデルを作るときに、その日の気象情報を特徴量として入れるケースを考える。気象情報は、過去については100％正しいものが得られるが、未来においては「予報」データでしかない。この場合、モデリング時には天気予報が100％当たる前提での精度が出るが、逆に本番運用時には天気予報が外れる分だけ、モデリング時に達成した精度は期待できない。

このような場合、過去の天気予報の外れやすさを事前に分析し、天気予報が外れやすい時季（季節の変わり目など）には天気予報の影響力を弱める、あるいは天気予報を確率分布で与える（晴れ60％、曇り30％、雨10％など）といったモデリングの工夫が必要となる。いずれにせよ、訓練時と本番運用時の業務の違いを理解した上で、それを適切にモデリングに反映しないと過剰適合が起きてしまう（図４）。

(2) ビジネスロジックと合わないモデリング

現在の機械学習技術は、訓練データに付与された正解ラベルから帰納的にモデルを構築する仕組みであるため、存在する訓練データに相関関係があれば、因果関係があると見なしてモデル構築が行われる。しかし現実には、相関関係があっても因果関係があるケースとないケース（疑似相関と呼ばれる）がある。相関関係と因果関係の違いはデータ上からは判断できない。そのため、さまざまなデータを投入した際に、①因果関係も相関関係もある、②因果関係があるが相関関係は弱い、③因果関係はないが相関関係は強い、④因果関係も相関関係もない、という４種類のデータが投入される可能性がある。欲しいモデルは①②の因果関係を強く効かせたモデルであるが、実際に生成されるモデルは③の効きが強くなることがある。これは機械学習技

図4　過剰適合を起こしたモデルにおける訓練時と推論時の精度差

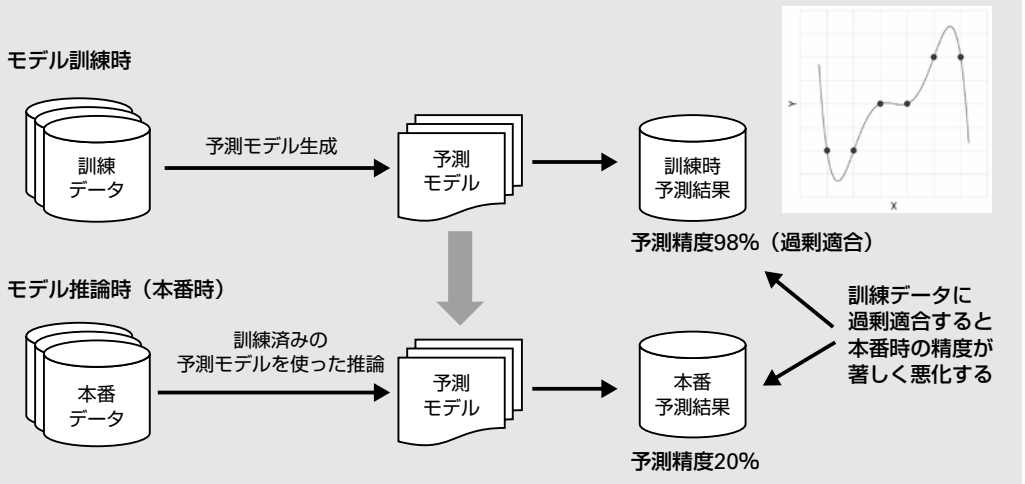
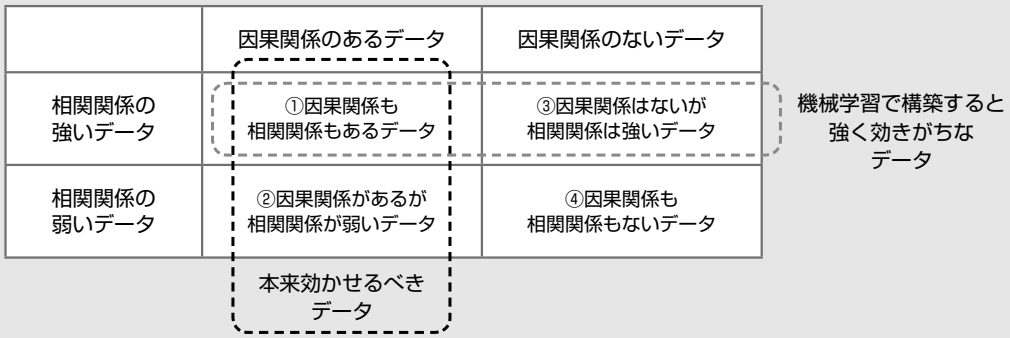


図5　機械学習で効きがちなデータと本来効かせるべきデータとの違い



術では因果関係と相関関係の区別が付かないからである。

たとえばある機械学習モデルが、Xという商品は過去10月によく売れるということを発見しモデルに組み込んだとする。しかし実は、Xという商品は台風が来るときによく売れるものであったが、過去データにおいて、たまたま10月に台風が多かったためにこのようなモデリングが行われただけであった。このモデルは、ある年、9月に台風が多く来る年に大きく精度を落とすことになる。

標準的なアプローチとしては、因果関係がないことが分かっているデータは除外し、因果関係があるデータに絞ってモデリングをする、あるいはこのようなことが起きないぐらい十分なデータを用意する、などが考えられる。しかし、前者について、どのデータを採用すべきであるかはビジネスロジックに依存するし、後者についても、どのようなデータを追加可能であるかはビジネス面の理解が求められる（図5）。

(3) 不適切なデータサイエンス手法の採用

一口にデータサイエンス技術といっても、その手法にはいろいろなものがある。実は精度が高い手法が既に考案されているにもかかわらず、課題と合っていない手法を用いていたが故に精度が出ないケースがある。

データサイエンティストにも分野による得意・不得意があり、表形式データ、時系列データ、自然言語データ、画像データなど、データの種類によって適用すべき手法は異なる。さらに、どの手法が最も適切かは、モデリングしないと分からないことも多い。時代によっても異なり、たとえば自然言語処理分野では、従来主流だったTF-IDFやWord2Vecと呼ばれる技術から、現在ではディープラーニング技術に基づいたBERTやGPT-3などの新しい技術が生まれ、急速に置き換わりつつある。一方で、タスクによっては伝統的な手法の方が高い精度を出すこともある。

このようにさまざまな手法やその違いについてある程度広く理解し、時代の変化に対して適切なデータサイエンス技術が適用できないと、本来、期待される精度のモデルを作ることができない。データサイエンススキルが低い場合にこのような問題が起きる（図6）。

3　実装フェーズにおける難しさ

第三に、実装フェーズにおける難しさにつ

図6　データサイエンス手法のジャンルと進化

	古くからある手法の例 ⟶		最近考案された手法の例
表形式データ	• SVM • ランダムフォレスト	• XGBoost • LightGBM	• TabNet • TabTransformer
時系列予測	• ARIMA • 状態空間モデル	• LSTM • Prophet	• DeepAR • Transformer
画像処理	• AlexNet • VGG	• ResNet • EfficientNet	• BiT • ViT
自然言語処理	• TF-IDF • Word2Vec	• BERT • XLNet	• GPT（GPT-2、GPT-3） • RoBERTa

いて、代表的なものとして次の3つを挙げる。

（1）ITシステムの実装を軽視したモデリング

（2）継続的なデータ取得コストの考慮不足

（3）訓練時の想定外データが実運用時に発生

（1） ITシステムの実装を軽視したモデリング

実装においては、運用可能な技術で構築することが求められる。問題解決フェーズでは、精度向上を優先するあまり、最先端のライブラリを使ったり、訓練時間や推論時間に制約なく構築したりしてしまうことがある。実装時の運用制約を踏まえた最適な技術の選定が必要である。

特に、データサイエンスプロジェクトにおいて一般的に用いられるPythonでは、ライブラリのアップデートのサイクルも早く、優秀なデータサイエンティストであるほど、最新のモデルを試してみたくなるものであろう。しかし、現実にはOSで正式サポートされるPythonのバージョンはこれより古いことが多く、対応可能なライブラリにも制約を受けることがある。コンテナ技術などを用いれば制約を回避することもできるが、そもそもコンテナ技術を用いた実装が認められていないケースもあるだろう。モデリングに集中する前に、実装時の技術制約について事前に本番システムを担当するITエンジニアと擦り合わせた上でモデリングを行う必要がある。

（2） 継続的なデータ取得コストの考慮不足

データサイエンティストが、新たなデータの取得を現場部門に依頼することがある。ここで、アドホックなデータ取得であれば現場部門も頑張ってデータを準備するが、本番運用時にも継続的に可能かという問題がある。現場部門側は、モデリング時に一度データを渡せば、以後、データを渡さなくて済むと誤

解し、一方で、データサイエンティストはモデリング時に入手できたデータは実装後も当然入手できるだろうと考える。

この齟齬があるままプロジェクトを進めると、モデリング時にさまざまな種類のデータを用意して高い精度を達成できたのに、いざ実装するタイミングになって必要なデータが入手できないことが判明し、代替データを用いた再モデリングが必要となるケースがある。理想的にはデータサイエンティストが現場部門のビジネス要件を踏まえた上で、低コストで効果的なデータを抽出し、高コストだが効果の低いデータを除外することが望ましい。

データサイエンティストは現場部門からの精度向上に対する期待に応えるために、できるだけ多くのデータを用いて精度を上げようとする。それ故に、本番運用時を想定しておかないと、多くのデータを使いすぎ、データ取得コストが膨れてしまい実用化できないということが起きがちである。

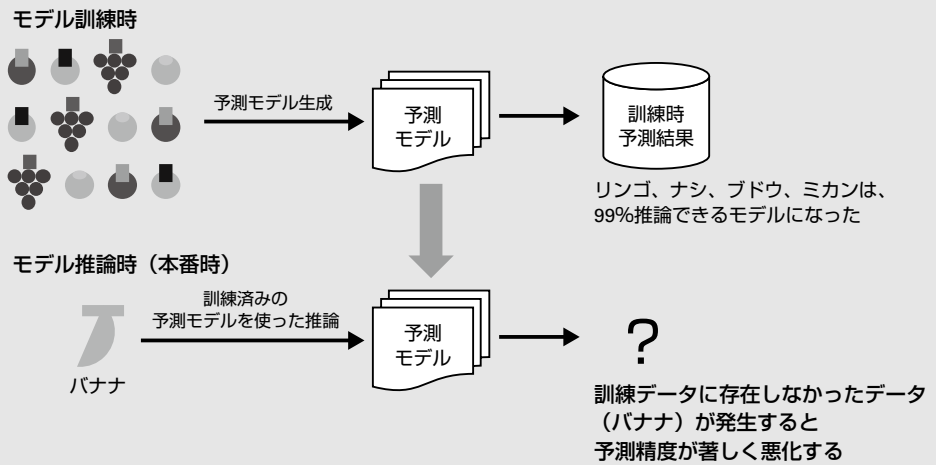
(3) 訓練時の想定外データが実運用時に発生

実装は構築して終わりではなく、実はそこからがスタートである。陥りがちな失敗例として、モデルを構築したまま精度モニタリングが行われず、精度悪化に気付かず価値を失うことがある。利用開始後も現場部門との対話を重ね、改善点がないかどうかをヒアリングし、課題を整理し、必要に応じて継続的改善をしていかなければならない。データサイエンスモデルは、何もせずに高い精度を永遠に保てるわけではない。当初期待していた精度を保てているかを継続的に監視することが必要である。

精度が保てなくなるケースで一番多いのが、訓練時に想定していなかったデータが実運用時に発生する場合である。機械学習技術は、訓練時に存在しないデータの予測を苦手とする。経験のあるデータサイエンティストは、これを回避するため、訓練時に存在しないデータが発生しないよう、設計の工夫を行うことがある。たとえば、商品の需要予測モデリングにおいて、商品コードを特徴量に入れると、存在しない商品（新商品）の予測ができなくなる。そこで新商品については、商品コードを入れなくても過去の類似商品の販売傾向と商品属性から需要を予測するモデルを別で用意し、それぞれ分けて予測をすることで、存在しないデータを発生させずに予測を行うことができる。訓練時に使うデータと本番運用時に発生するデータの違いに注意してモデリングをすることは、実運用時のモデル精度を考えると、非常に重要な視点である。

特に、自然現象を扱うモデルと比べて、社会現象を扱うモデルは精度が劣化しやすいとされる。社会現象は自然現象と比べて、訓練時に想定していない事象が起きやすいからである。たとえば、コロナ禍前のデータに過剰適合した需要予測モデルは、コロナ禍において正しい予測ができなくなる。「想定外の事象だったので精度が落ちました」では、実ビジネスで使えるモデルとはいえない。社会現象は、さまざまな要因において「変わり得るもの」という前提を置いて、過剰適合が起こらないようなモデリングの工夫を行わなければならない。このようなモデリングが行えるようになるにも知識と経験が必要とされる（図７）。

図7　訓練時の想定外データの発生



4　工程全体における難しさ

最後に、工程全体における難しさを１つ挙げる。

イテレーション（繰り返し）工程への理解不足

前述した図２のCRISP-DMを見ても分かるように、データサイエンスプロジェクトの特徴の一つは、一方通行ではないということだ。STEP1（ビジネス理解）とSTEP2（データ理解）は互いに行き来する関係にあり、STEP3（データ準備）とSTEP4（モデリング）も同様の関係であり、さらにSTEP5（評価）後にあらためてSTEP1（ビジネス理解）に戻るという関係もある。

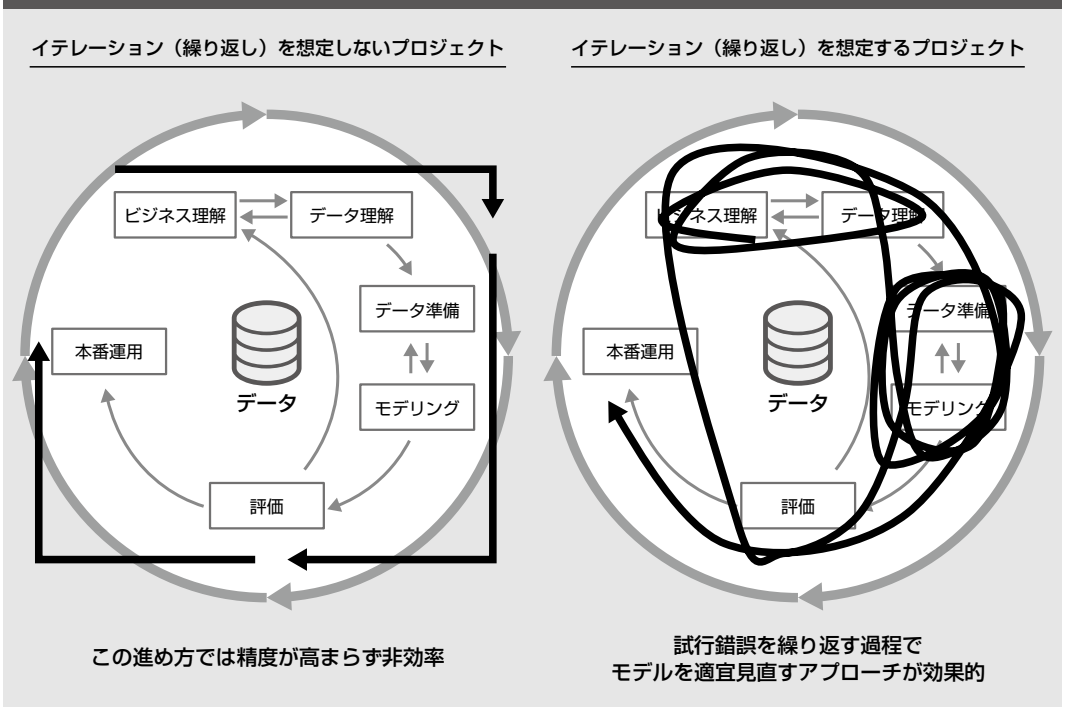
CRISP-DMは手戻りを前提としたフレームワークであり、ウォーターフォールモデルとは大きく異なる。ステップを戻ることは決してネガティブではない必要なプロセスであり、むしろ何度も手戻りを繰り返して精度を高めていくことが重要である。これを理解していないと、単純に「問題設定フェーズ」「問題解決フェーズ」「実装フェーズ」とスケジュールを立ててしまいがちだが、この場合、「問題解決フェーズ」に進んだ後で「問題設定フェーズ」に戻ることが憚られたり、ROIが見極められていないにもかかわらず「実装フェーズ」に突入したりする。

CRISP-DMに従うと、「問題設定フェーズ」と「問題解決フェーズ」は、互いに行ったり来たりしながら徐々に精度を高めてROIを見極めていく、最終的に「実装フェーズ」に突入するかどうかは、必ずROIを判断した後という流れを経る。特に「問題設定フェーズ」と「問題解決フェーズ」を何度も繰り返すことをイテレーション（繰り返し）工程などと呼んだりするが、このような工程が重要であることを理解し、必要な工数を確保しなければならない（図８）。

Ⅲ 実装するための４つのポイント

前章で、データサイエンスプロジェクトの難しさについて、フェーズごとに陥りがちな

図8　イテレーション（繰り返し）の有無によるプロジェクトの違い



失敗例を論じた。データサイエンスは比較的急速に注目された概念であるが故に、多くの企業にとってなじみがなく、さらに技術革新が早いために、これらについての深い知識や経験を持つ人材が不足している。そのため、多くの企業がデータサイエンスのケイパビリティを持たないのは当然である。

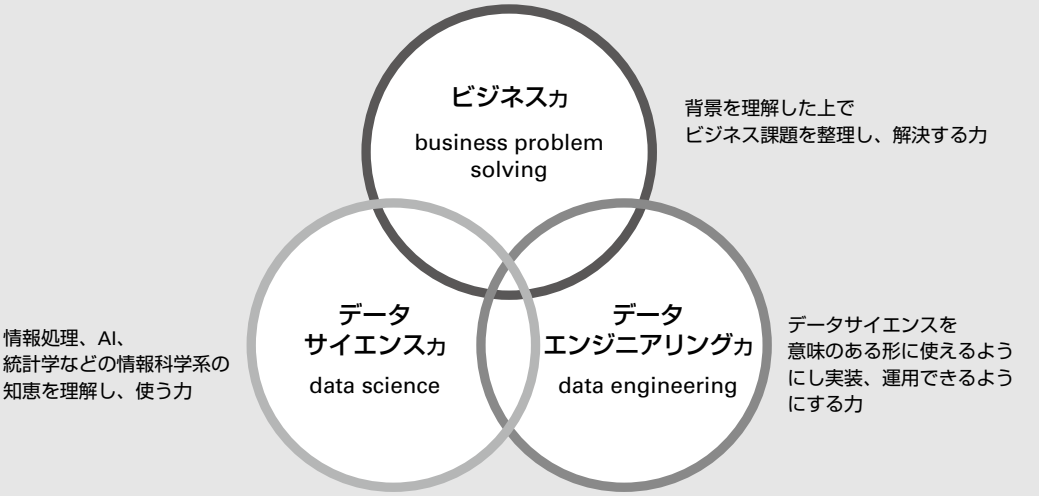
では、各企業がデータサイエンスプロジェクトを円滑に推進するケイパビリティを獲得するにはどうすればよいのか。具体的には次の4つがポイントであると考える。

ポイント1　ハイブリッド型人材を育てる

データサイエンティスト協会の定義によれば、データサイエンティストのスキルセットには「ビジネス力」「データサイエンス力」「データエンジニアリング力」の3種類がある（図9）。一般にデータサイエンティストというとデータサイエンス力に秀でた人をイメージしがちだが、これからデータサイエンスを活用しようとする企業にとっては、たとえば、データサイエンススキルが10点でほかが0点の人よりも、すべてのスキルが3点ずつあるようなハイブリッド型の人の方がよいだろう。初期においては、3つのスキルが必ずしも高いレベルではなくとも、バランスが取れていることが重要だと考える。

特に外部からデータサイエンティストを募集して組織を作ると、データサイエンススキルに強みを持つ人材が集まりやすい。しかし、これではビジネス面が弱く、ビジネスとデータサイエンスをつなぐことができない。さらに、エンジニア面が弱いと、単なるデータ分析やPoCで終わってしまい、それを実装するに至らなくなる。2020年にデータサイエンティスト協会が実施した企業向けアンケー

図9　データサイエンティストの3つのスキルセット



出所）データサイエンティスト協会資料より作成

トにおいても、今後増員したいデータサイエンティストのタイプはエンジニアスキルに強いタイプ（43%）、ビジネススキルに強いタイプ（34%）、データサイエンススキルに強いタイプ（23%）という順であった[注2]。

ハイブリッド型人材は希少であり、特に外部調達は難しい。つまり、自社で育成するのが最も現実的であるということだ。もちろん、企業のデータサイエンスレベルが高まってくれば、バランス型人材だけでなく、それぞれの専門性に秀でた人材を集めて分業する体制に移行していくことになるだろうが、一足飛びにそこに向かうのは、基礎問題が解けないのにいきなり応用問題を解き始めるようなものである。

ビジネスニーズを満たすことができなければ組織の規模は大きくできないし、ビジネスが大きくならないことには分業体制も築けない。まずはハイブリッド人材を集めてニーズを開拓し、具体的な成果を生み出しつつ、ある程度、データサイエンス文化がなじんだ後で、徐々に専門型人材を受け入れていくという順序で進めるのが肝要であろう。また、複数人でデータサイエンス組織を形成した場合、マネジメント層にもハイブリッド型人材を当てるべきであろう。

ポイント2　現地現物にこだわる

データサイエンティストもエンジニアも、実ビジネスの現地現物に当たることが極めて重要である。企業が利益を生み出す源泉、特に競争力の高い企業のビジネスモデルの根幹は、現地現物にあることが多い。また、現地現物を理解しないでモデリングすると、たとえば現場では当たり前の事実を推論するだけのモデルを作るなど、技術に偏って現場から必要とされないものを作ってしまう恐れがある。そうならないためにも、しっかりと現地現物に当たり、理解することが必要となろう。優秀なデータサイエンティストほど、まず現地現物を見てモデリングに当たることの重要性を理解している。

しかし、特に大企業になるほど、現地現物に当たることが難しくなる。大企業は高度な分業化が進んでいるが故に、中堅クラスの社員であっても担当部署以外の実務については意外と知らないことも多い。組織が大きくなり、いわゆるサイロ化と呼ばれる現象で分断されていることもある。その中で、データサイエンスを活用しようと考えたら、ある程度、企業のビジネス全体を俯瞰し理解することが求められる。

希望するデータサイエンティストに現場業務を数カ月担当させるなど、必要な調整を柔軟に行えるかが重要である。データサイエンスと現地現物は一見異質であり、異質なものを組み合わせようとしても並大抵の組織では分離してしまうものである。分離を回避するには経営層のコミットメントが重要となる。経営層がデータサイエンスと現地現物の重要性をしっかりと理解した上で、それを支援する組織を作るなど、適切な枠組みをいかに用意できるかがカギとなる。

ポイント3　自己研鑽を奨励する

データサイエンスの分野は極めて技術進化のスピードが速く、特にアルゴリズムなどは毎月のように新たなものが発表され、世界中で試されている。Kaggleなどのデータ分析コンペにおいても、主流とされるアルゴリズムは時代の変遷とともに変化する。故に、一度知識を得てそのレベルに満足するのではなく、常に自己研鑽を怠らず、自らの知識レベルをアップデートし続けられるような人材が求められる。研修制度の整備、資格取得の支援、リカレント教育の機会創出、インターネットやSNSを活用した情報収集のサポートなど、意欲ある社員がデータサイエンスのリテラシーを自主的に高められるような環境作りが必要だろう。

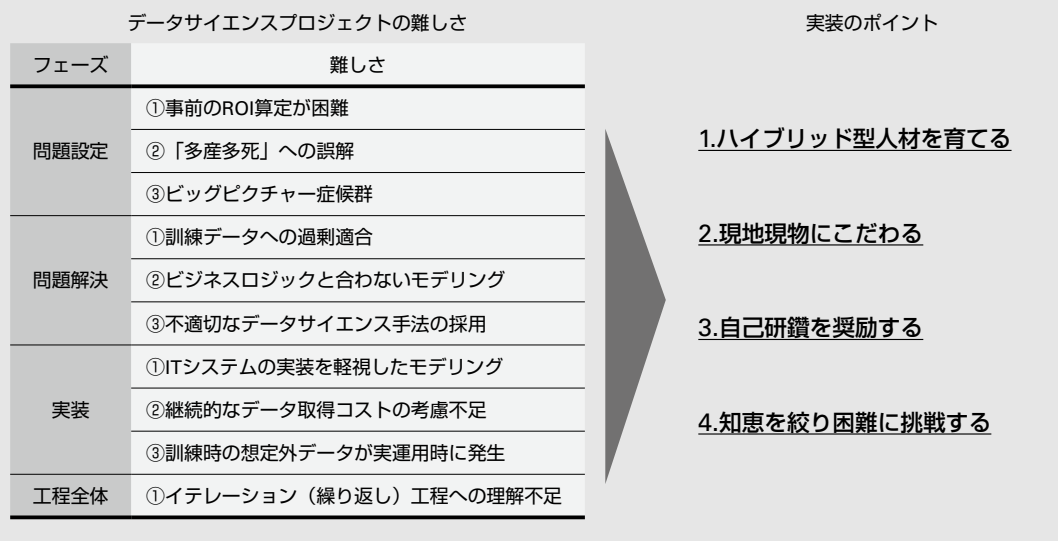
また、データサイエンススキルに関しては、Kaggleに代表される社外のデータ分析コンペティションへの参加を奨励してリテラシーを高めるということもできるだろう。NRIでも2021年6月に、社内で第一回Kaggler座談会（Kaggleに参加している人、および参加に興味がある人の座談会）を開催したところ、60人以上が参加するなど好評であった。企業側だけでなく、実は社員にもデータサイエンス技術に高い関心を持っている人がいる。自己研鑽を奨励する中で、そういったポテンシャルの高い社員を発掘することも期待できるのではないか。

ポイント4　知恵を絞り困難に挑戦する

データサイエンスプロジェクトは、自由度が高く、試行錯誤しながらビジネス価値を創造していくものである。プロジェクトの開始時点で、最終的にどの程度の精度を達成できるかを予見することは難しいし、逆にプロジェクト開始当初は難しいと考えていたことが、革新的なアイデアにより実現できるようになることもある。ゴールとなる価値創造の形を見据えながらも、常に状況変化に応じて柔軟に軌道修正をし続ける必要がある。

さらに、データサイエンスプロジェクトは、非常にクリエイティブな取り組みである。いろいろなアイデアを産み出し、設計の工夫を凝らし、幅広く新しい技術を調べ、知恵を絞り出して、困難に挑戦していくものである。今、大企業の多くは、かつて成功したビジネスモデルの維持に慣れてしまい、この

図10　データサイエンスプロジェクトの難しさと実装のポイント



ようなクリエイティブなマインドを失いつつはないか。七転び八起きの精神で苦労を重ね、何度失敗しても次のモデルを作り、新しい価値創造にチャレンジする。コモディティではなくスペシャリティを作り上げる。そのような文化を今一度作り出すことが求められるだろう。そして、そういったチャレンジを積極的に支え、育てていく支援体制が求められるのではないか。

ここまで述べてきたデータサイエンスプロジェクトの難しさと、それを克服するための４つのポイントをまとめたものが図10である。

Ⅳ 今後の展開

最後に、今後の展開について考えたい。これまで説明してきたようなデータサイエンスプロジェクトの特徴は、実は本来、日本企業が得意とする領域なのではないかと考える。その理由は、次の２点にある。

まず１点目は、日本企業の多能工文化である。これまで述べてきたように、データサイエンスの高度活用には、複数領域がハイブリッドに重なり合うことが重要である。データサイエンスは、トップダウンよりもむしろ現場の細かい業務を理解して、それを解決するソリューションを泥臭く構築するものである。このようなアプローチは、「カイゼン」に象徴されるような日本企業の得意分野に通じるものであり、多くの日本企業が強みとしてカルチャーに持っているものではないだろうか。確かに、最先端の機械学習ライブラリの多くは米国で開発されている。しかし、米国企業が得意とする高度な分業は、実ビジネスにデータサイエンスを適用する上ではなじまない。企業にデータサイエンスを適用するには、必ずその企業のビジネスに対する深い理解が求められる。トップが現場の実態を徹底的に理解していることも強みだろう。多能工社員による擦り合わせを重ねて精度を高めるカルチャーが、データサイエンスプロジェ

クトと相性がよいと感じている。

もう1点は、貴重なクローズドデータの蓄積である。日本には依然として世界を代表するような大企業が多い。冒頭のGAFAはいずれも、インターネット上のデータ（オープンデータ）の取り扱いに強みがある。しかし、企業内データ（クローズドデータ）はまだ発掘されておらず、だからこそ活用のチャンスは大きい。データサイエンスの世界は「データが集まるほど精度が高く、利用価値も高い」という勝ち組企業をさらに強くする性質がある。データサイエンスを得意とする新興企業も、データを持たなければ価値の高いモデルは作れない。むしろ伝統的な大企業こそが、データサイエンスのケイパビリティを獲得し、未発掘のクローズドデータを適切に精錬できれば、新たな利益に変換できる余地が大きいのではないだろうか。

5年前、世界に衝撃を与えたDeepMind社を傘下に持つGoogleは、この5年間、AI分野への投資を積極的に進め、多くの変革を生み出してきた。たとえば、翻訳精度を大幅に向上させ、医療分野や検索分野などでさまざまな進化を実現してきた。機械学習を用いることで解決可能な課題は幅広く、世界はまだ変わり始めたばかりである。データサイエンスはIT企業の専売特許ではなく、すべての企業が武器にできるものである。多くの企業がデータサイエンスのケイパビリティを高め、実践的な取り組みを進められれば、近い未来に世界中のさまざまな企業からGAFAに続く革新的な成果が生まれてくることも十分期待できるだろう。

注

1 総務省「令和2年版情報通信白書」P.227 図3-2-2-3
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r02/pdf/02honpen.pdf

2 データサイエンティスト協会「データサイエンティストの採用に関するアンケート（2020年8～9月実施）」P.7
https://www.datascientist.or.jp/common/docs/corporate_research2020.pdf

参考文献

1 有賀友紀、大橋俊介『RとPythonで学ぶ［実践的］データサイエンス＆機械学習【増補改訂版】』技術評論社、2021年

2 有賀康顕、中山心太、西林孝『仕事ではじめる機械学習』オライリー・ジャパン、2018年

3 野村総合研究所、NRIセキュアテクノロジーズ『ITロードマップ2021年版』東洋経済新報社、2021年

著者

田村 初（たむらはじめ）
野村総合研究所（NRI）データサイエンスラボ上級データサイエンティスト
専門は小売・流通・消費財業界のマーケティング分析、デジタル化戦略、データアナリティクスによる事業開発など