IBM Analytics
ホワイト・ペーパー

# データ・サイエンスのための基本的方法論

IBM

データ・サイエンスの領域では、データ分析による問題の解決や疑問への回答が普通に行なわれています。データ・サイエンティストは洞察を得ることを目指し、結果を予測したり隠れたパターンを発見したりするためのモデルを構築します。企業はこうした洞察を生かし、将来の結果を向上させるための行動を取ることができます。

データ分析やモデル構築のためのあらゆる技術が急速に発展しています。こうした技術は、驚くほど短い時間に、デスクトップから膨大なデータ容量を持つ超並列ウェアハウスならびにリレーショナル・データベースおよびApache Hadoop のインデータベース分析機能へと進歩してきました。非構造化データや半構造化データのテキスト分析は、センチメントやその他有用な情報をテキストから予測モデルに取り込む方法として、ますます重要になってきており、これがしばしばモデルの質や正確性を大きく向上させることにつながっています。

新たに現れてきたさまざまな分析アプローチは、モデル構築および適用の手順を自動化して、機械学習技術を深い計数的スキルのない人々にもより使いやすくしようと努めています。また、まずビジネス上の課題を定義し、次にデータを分析して解決法を探るといった「トップダウン」のアプローチに対して、一部のデータ・サイエンティストは「ボトムアップ」アプローチを使うことがあります。後者のアプローチでは、まずデータ・サイエンティストは大量のデータを観察してそのデータからどのような事業目標が導き出されるかを調べ、次に問題に取り組みます。多くの問題はトップダウン方式で扱われるので、この資料で扱う方法論もそうした視点を反映しています。

## 技術からアプローチまでにわたる、10 段階のデータ・サイエンスの方法論

データ分析機能がより利用しやすく、より普及してきている中で、データ・サイエンティストが必要としているのは技術、データ容量、関係するアプローチなどにかかわらず指針となる戦略を提供できる、基本的な方法論です (図 1 を参照)。この方法論はデータ・マイニングで定評のある方法論[1-5]といくつかの類似点がありますが、非常に大容量のデータの利用、予測モデリングへのテキスト分析の取り入れ、いくつかのプロセスの自動化などといった、データ・サイエンスにおける新たな手法のいくつかを重視していしています。

この方法論は、新たな洞察を得るためにデータを使用する反復的なプロセスの 10 段階により構成されます。それぞれの段階が、方法全体の文脈で重要な役目を果たしています。

### 方法論について

**方法論とは、ある領域におけるプロセスや活動を導くための一般的な戦略のことです。方法論は特定の技術やツールに依存せず、また技術や方法を集めたものでもありません。むしろ方法論とは、答えや結果を得るためにどのような方法、プロセスおよび経験則が利用される場合でも、どのように物事を進めていけばよいのかという枠組みをデータ・サイエンティストに対して提供するものです。**

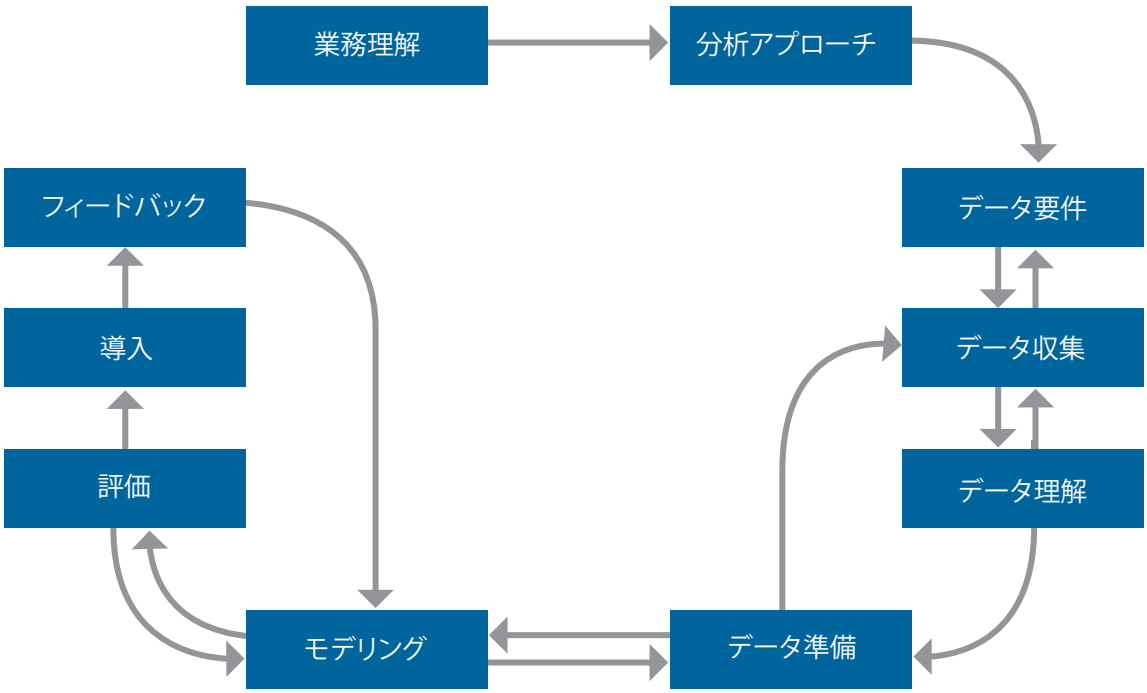


図 1: データ・サイエンスのための基本的方法論

**第 1 段階: 業務理解**

すべてのプロジェクトは、業務理解から始まります。この段階では、分析ソリューションを必要とするビジネス・スポンサーが、最も重要な役目を努めます。ここでのビジネス・スポンサーの役割は、ビジネスの観点から問題、プロジェクトの目標、ソリューションの要件を定義することです。この最初の段階は、ビジネス上の課題を成功裏に解決するための基盤を築くものです。プロジェクトの成功をより確実にするには、スポンサーは専門家の提供や中間的発見の評価などを行なうためプロジェクト全体にわたって関与し、目指す解決にたどり着くために、作業が確実に軌道に乗るようにすべきです。

**第 2 段階: 分析アプローチ**

ビジネス上の課題を明確にした後、データ・サイエンティストは問題解決のための分析アプローチを定義します。この段階では、企業が望む目的のために最適なアプローチを特定できるよう、統計的手法や機械学習技術の文脈で問題を表現することになります。たとえば、目的が「はい」「いいえ」といった回答を予測するものである場合、分析アプローチを、分類モデルの構築、テスト、実装として定義できるかもしれません。

### 第 3 段階: データ要件

選択した分析アプローチが、データ要件を決定します。具体的にいえば、利用する分析手法には、専門知識によって導き出される特定のデータ内容、フォーマット、表現が必要となります。

### 第 4 段階: データ収集

データ収集段階の初期において、データ・サイエンティストは問題の領域に関係する、構造化されたデータ、一部構造化されたデータ、構造化されないデータといった利用可能データ・リソースを特定し、収集します。データ・サイエンティストは、より入手しにくいデータ要素を得るために追加投資するかどうかを選択する必要があります。そのため、データやモデルについてより多くのことが分かるまで、投資の決定を遅らせることが最善かもしれません。また、データ収集に欠損がある場合、データ・サイエンティストは状況に応じてデータ要件を修正し、新しいデータ、またはより多くのデータを収集する必要があるかもしれません。

データ・サンプリングやサブセッティングはなおも重要ですが、今日の高性能のプラットフォームやインデータベース分析機能により、データ・サイエンティストは利用可能な大部分や、前部分さえ含めた、より大きなデータ・セットを扱うようになりました。より多くのデータを取り入れることで、疾病の発生やシステム障害といった極まれな出来事でも、予測モデルがよりよく表現できるようになるかもしれません。

### 第 5 段階: データ理解

オリジナル・データの収集後、データ内容を理解し、データ品質を評価し、データについての最初の洞察を発見するため、データ・サイエンティストは記述統計や可視化技術を用います。データの隙間を埋めるために、追加のデータ収集が必要になることもあります。

### 第 6 段階: データ準備

この段階は、続くモデリング段階で利用されるデータ・セットを構築するためのすべての活動を含みます。データ準備活動には、データ・クリーニング (失った値、無効の値についての対処、重複の削除、正しいフォーマッティング)、複数のソース (ファイル、テーブル、プラットフォーム) からのデータの組み合わせ、およびデータをより役立つ変数に転換することが含まれます。

「フィーチャー・エンジニアリング」と呼ばれるプロセスにおいて、専門知識と既存の構造変数とを組み合わせることにより、データ・サイエンティストは予測変数または特徴変数とも呼ばれる追加の説明変数を作成できます。カスタマー・コール・センターのログや、医師のノートなど、非構造化フォームや半構造化フォームでテキスト・データが利用可能な場合、新たな構造変数をもたらし、予測変数のセットを強化し、モデルの精度を高めるために、テキスト分析が役立ちます。

通常、データ準備はデータ・サイエンス・プロジェクトの中で最も時間のかかる段階です。多くの領域において、データ準備段階のいくつかはさまざまな問題にわたって共通しています。特定のデータ準備段階を事前に自動化しておけば、準備にかかる時間を最小化でき、この段階を早めることができるかもしれません。データ保存場所における今日の高性能化や超並列システム、分析機能により、データ・サイエンティストは非常に大きなデータ・セットを利用して、より簡単かつ素早くデータを準備できます。

### 第 7 段階: モデリング

準備したデータ・セットの初回版を皮切りに、前の段階で定義した分析アプローチに従って、モデリングの段階では予測モデルまたは記述モデルに焦点を合わせます。予測モデルでは、データ・サイエンティストはモデル構

築のためにトレーニング・セット (関心を持つ結果の、既知である履歴データ) を使用します。企業が中間的な洞察を得るにつれ、データの準備やモデルの仕様を洗練させていくことになるため、モデリングのプロセスは、非常に繰り返しが多くなります。既知の技術として、利用可能な変数に対して最適なモデルを発見するため、データ・サイエンティストそれぞれのパラメーターに複数のアルゴリズムを試すことがあります。

### 第 8 段階: 評価

モデル開発中および展開前に、データ・サイエンティストはモデルを評価して、モデルの品質を理解し、それが正しくかつ十分にビジネス上の課題を扱っているかを判断します。モデル評価には、診断基準や表やグラフなど、他の出力の計算が伴います。それによりデータ・サイエンティストがモデルの品質や、問題解決にあたってのモデルの有効性を解釈できるようになります。予測モデルでは、データ・サイエンティストはテスト・セットを使用します。これはトレーニング・セットとは独立したものですが、同じ確率分布と既知の結果を持っています。テスト・セットはモデルの評価に使用され、モデルを必要に応じて洗練できるようにします。最終評価のために、検証セットに対して最終モデルも適用される場合もあります。

加えて、データ・サイエンティストは統計的有意性テストをモデルに行ない、モデルの品質をさらに証明することがあります。この追加的な証明は、高価な補足的医療プロトコルや重要な航空機飛行システムなど、リスクが高いケースでモデル実装を正当化したり、行動を取るために役立つ場合があります。

### 第 9 段階: 導入

満足できるモデルが開発され、ビジネス・スポンサーに承認されると、そのモデルは本番環境または比較テスト環境に導入されます。通常、その性能が十分評価されるまで、モデルは限られた形で導入されます。導入は、推奨事項を記載したレポートを生成するといった単純な場合もあれば、カスタム・アプリケーションに管理された複雑なワークフローやスコア付けプロセスにモデルを組み込むといった複雑な場合もあります。モデルをオペレーション的ビジネスプロセスに導入する際は、通常、社内のグループ、スキルおよび技術が追加で関係することになります。たとえば、営業グループは開発チームによって作成され、営業グループによって管理されるキャンペーン管理プロセスに応答傾向モデルを導入する場合があります。

### 第 10 段階: フィードバック

実装したモデルから結果を収集することにより、導入された環境におけるモデルのパフォーマンスやモデルの影響についてのフィードバックを得ます。そのフィードバックは、「モデルによって反応する可能性が高い」と識別された顧客グループにターゲットを合わせた販促キャンペーンの反応率といった形を取ることができます。フィードバックの分析により、データ・サイエンティストはモデルを洗練し、その正確性と有用性を向上させることができます。データ・サイエンティストは、フィードバック収集、モデル評価、洗練および再導入の段階の一部またはすべてを自動化して、よりよい結果を出すためのモデル更新を早めることができます。

## 企業に継続的な価値を提供

方法論のフローは、問題解決プロセスの反復的な性質を表しています。データ・サイエンティストは、データやモデリングに関してより多くのことを学ぶにつれて、より頻繁に前の段階に戻って調整を行なうようになります。モデルは、一度作成して、導入した状態のままにしておくといったことはありません。むしろフィードバック、洗練化、再導入といったサイクルを通して、新たに発生する状況に合うよう継続的に向上させるものです。解決が必要とされている限り、モデルとその背後にある業務の双方が、企業に継続的な価値を提供できます。

## 詳細情報

基本的なデータ・サイエンスの方法論に関する新しいコースが、Big Data University から受講できます。この無料のオンライン・コースを受講するには、以下を訪問してください。http://bigdatauniversity.com/bdu-wp/bdu-course/data-science-methodology

## 謝辞

有用なコメントをくれた Michael Haide 氏、Michael Wurst 博士、Brandon MacKenzie 氏、Gregory Rodd 氏に感謝いたします。また、Jo A. Ramos 氏の役割にも感謝いたします。彼とは何年もの間共同でこの方法論を築き上げてきました。

## 執筆者プロフィール

John B. Rollins 博士は、IBM Analytics のデータ・サイエンティストです。業界の多岐に及ぶ、エンジニアリング、データ・マイニングおよび経済学を背景としています。7 つの特許を保持し、エンジニアリング・テキストのベストセラーを著しています。また、多くの技術論文も発表しています。石油工学および経済学の分野で、Texas A&M University から博士号を取得しています。



日本アイ・ビー・エム
株式会社〒103-8510
東京都中央区日本橋箱崎町 19-21

Produced in Japan
June 2015



[1] Brachman, R. & Anand, T., “The process of knowledge discovery in databases,” in Fayyad, U. et al., eds., Advances in knowledge discovery and data mining, AAAI Press, 1996 (pp. 37-57)

[2] SAS Institute, http://en.wikipedia.org/wiki/SEMMA, www.sas.com/en_us/software/analytics/enterprise-miner.html, www.sas.com/en_gb/software/small-midsize-business/desktop-data-mining.html

[3] Wikipedia, “Cross Industry Standard Process for Data Mining,” http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining, http://the-modeling-agency.com/crisp-dm.pdf

[4] Ballard, C., Rollins, J., Ramos, J., Perkins, A., Hale, R., Dorneich, A., Milner, E., and Chodagam, J.:Dynamic Warehousing:Data Mining Made Easy, IBM Redbook SG24-7418-00 (2007 年 9 月), pp. 9-26.

[5] Gregory Piatetsky, CRISP-DM, still the top methodology for analytics, data mining, or data science projects, 2014 年 10 月 28 日, www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html



IMW14828-JPJA-00