技術レポート

データマイニング・プロジェクト・ライフサイクル
―データマイニングとCRISP-DM方法論の紹介―

EST コンサルティング本部　IT コンサルティング部　**熊野　匡**

１．はじめに

現在、私はIT コンサルティング部で、ビジネス・インテリジェンスという分野を専門に担当しており、データマイニングについても積極的に情報収集や調査を行っている。本稿では、これまでの調査の成果として、データマイニング技術の概要と、データマイニングの方法論として海外で注目されている「CRISP-DM」を紹介する。少しでも多くの読者に、データマイニングという技術を理解し、興味を持ってもらえれば幸いである。

２．データマイニングの概説

データマイニングという言葉は、ほとんどの人が聞いたことがあることと思う。しかし、実際にどんな技術であるかを理解している人は少ないのではないだろうか。本章では、データマイニングという技術について具体的なイメージを持ってもらうと同時に、データマイニングの技術を利用して、どのようなことができるのかを解説する。

2.1　データマイニングの典型的な事例－ビールと紙オムツ

ある海外の大手スーパーマーケットでは、レジで入力されるPOS（Point Of Sales）情報を基にして、顧客がどのような組み合わせで商品を購入しているのか、という分析を行った。その結果、「ビールと紙オムツ」が一緒に買われているという、思いがけない分析結果が出た。普通に考えると、ビールと紙オムツという商品の間には何の関係も無さそうだが、はたしてコンピュータは正しい分析結果を出しているのだろうか。そこで、さらに詳しく分析を進めたところ、どうやら若い男性の顧客が金曜日の夕方にビールと紙オムツを一緒に買っていくケースが多いようだ。この詳細な分析結果を見たスーパーマーケットのマーケティング担当者は、その理由を以下のように想像した。

「ビールと紙オムツという組み合わせで商品を買っているのは、赤ん坊のいる若い夫婦である。夫婦は、子供がまだ幼いために、週末は自宅でゆっくり過ごそうと考えている。そこで、週末に紙オムツが無くなってしまわないように買い置きしておきたい。でも、奥さんは子供に手がかかりきりのため、優しい旦那が仕事帰りに買い物をして帰ることになる。旦那は紙オムツを買うついでに、自分が飲むビールも買い置きしておこうと考える。」

この想像を基に、マーケティング担当者はコンピュータの分析結果を意味があるものと判断した。そして、実際にビールと紙オムツの売場を並べてみたところ、売上が大幅に増加したのである。既に耳にされていた方も多いであろう、世界で一番有名なデータマイニングの成功事例である。

2.2　データマイニングの定義

データマイニングという言葉には、誰もが認めるような公式の定義はなく、書籍やツールベンダーによって定義はまちまち、というのが現状である。しかし、だいたいにおいて、どの定義を見ても同じようなキーワードが利用されている。前述の事例と重ね合わせて、それぞれのキーワードを解説することで、以下にデータマイニングという言葉を定義してみる。

⑴　大量データ

データマイニングは、大量のデータを分析対象とする。データマイニングという言葉が「大量のデータに埋もれてしまっている情報を発掘（マイニング）する」というイメージに由来するように、大量のデータを分析できることは、非常に重要な特徴である。先の事例では、POS データを

分析対象としているが、POSのデータは毎日膨大な量が発生するため、人手による分析は、ほぼ不可能である。この大量のPOSデータから有用な情報を得られたという点が、この事例のポイントの１つである。

⑵ 新たな知識の発見

データマイニングは、「以前には知られていなかった」情報を得ることを目的として分析が行われる。これまでの統計解析やOLAP（On-Line Analytical Processing）に見られる「仮説検証型」の分析ではなく、データマイニングは「知識発見型」の分析である。ここでいう「仮説検証型」の分析方法は、何らかの仮説を設定した上で分析を行い、設定した仮説が正しいことを証明するという手順をとる。一方、「知識発見型」の分析では、分析を開始する前に仮説を立てる必要がない。

ビールと紙オムツの事例に重ねると、「仮説検証型」であれば、分析者は「ビールと紙オムツという商品は同時に買われているに違いない」という想像に基づいて分析を始める必要がある。そのうえで、ビールと紙オムツの相関係数（関係の深さを数値化する解析手法）を計算したり、他の商品の売上と比較することにより、ビールと紙オムツという商品の組み合わせの有意性を証明していく。

しかし、データマイニングという「知識発見型」の分析では仮説を立てる必要はなく、「商品の中には、他の特定の商品と一緒に買われるものがある」ことだけを知っていればよい。何の先入観もなしに分析にとりかかることで、ビールと紙オムツのように、普通に考えると想像もつかない商品の組み合わせを「発見」できるのである。

⑶ 有用な情報

データマイニングで得られる情報は、実務上「有用な」ものでなければならない。すなわち、データマイニングの分析結果は、ビジネス等の局面で何らかの効果をもたらすべきである。事例では、ビールと紙オムツを並べて売ることで売上が増加したことが、ここでの効果を指している。逆に効果がないのであれば、分析する意味がなくなるため、分析結果の活用方法を意識しておくことが非常に重要となる。

⑷ アルゴリズム主導の分析

ガートナーグループは、「データマイニングはコンピュータ・アルゴリズム主導の分析である」と定義している。もともと、データマイニングという言葉は、データベースから知識を発見するプロセス全体を指すKDD（Knowledge Discovery in Databases：データベースの知識発見）という、1980年代に流行した用語の一要素である。その中では、コンピュータを利用してマイニング・アルゴリズムを適用するという、限られたステップを指すために用いられていた。そのため、このガートナーグループによる定義が、最も的確にデータマイニングを表しているのかもしれない。しかし、データマイニングに対する関心が高まった現在では、もっと広い意味で用いられることが一般的となっている。そこで、言葉の定義の一部分として、「大量のデータを効率的に扱うことのできるマイニング用のアルゴリズムが用いられる」という表現を含める程度にしておくことが適当だと思える。事例では、対象となる全てのデータの相関性の強さを比較することができる、アソシエーションルールというアルゴリズムが用いられている。

2.3 データマイニング適用方法の分類

前項では、ビールと紙オムツという１つの事例を中心に、データマイニングの定義を解説した。しかし、データマイニングができることは、前述の事例のような分析だけではない。以下に、一般的なデータマイニングの適用方法を、６つのタイプに分類して紹介する。なお、分類方法は、本稿の後半で説明する「CRISP-DM」方法論の中で、「Data Mining Problem Type」として解説されている内容に基づく。

⑴ タイプ１：データの解説と集計（Data description and summarization）

「データの解説と集計」では、ユーザーがデータの内容を理解できるように、集計などを用いてデータの特徴を簡潔に説明する。図１に顧客分析における分析イメージを示す。実際の顧客分析では、データの対象がはるかに多くなるが、図中では簡略化して記述する。

「データの解説と集計」は、他の分析と組み合わせて使われることが多く、通常は分析の初期にデータを理解するために行われる。ここでは、データ件数や平均値、分布状況の把握など、分析を進めるうえでの基礎知識を得ることが主な分析内容になる。

「データの解説と集計」だけがプロジェクトの目的であ

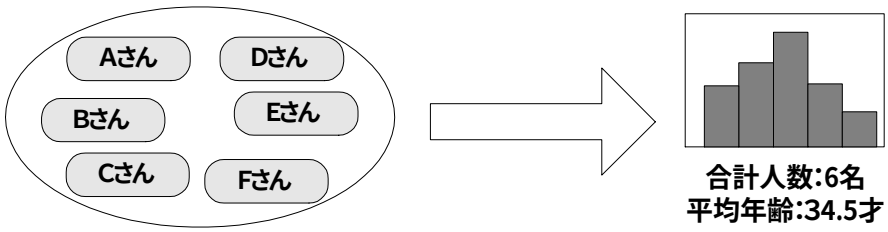

図１ 「データの解説と集計（Data Description and summarization）」のイメージ

る場合は、データマイニング・ツールを利用することなく、統計解析ツールやOLAPツール、EIS（Executive Information System）などでも分析が可能である。

⑵ タイプ２：セグメンテーション（Segmentation）

「セグメンテーション」では、データを何らかの意味のあるグループ（セグメント）に分割する。図２は全ての顧客を分析した結果、顧客を２つのセグメントに分割している例である。

ここでは、セグメントの発見そのものがデータマイニングの通常の目的となる。しかし、POSデータ内のレコードを「セグメンテーション」によってグループ分けしてから「従属性分析」を行うなど、まれに他の問題を解決するために途中のステップとして利用されることもある。

「セグメンテーション」という言葉は「クラスタリング」や「分類」と混同されることが多いが、以下のように整理できる。

- セグメンテーション：データをグループに分けること
- クラスタリング：セグメンテーションを実施するための手法の１つ
- 分類：データが、どのグループに属するかを決定するための予測モデルを作成すること

⑶ タイプ３：コンセプトの解説（Concept Description）

「コンセプトの解説」では、あるグループの特徴を理解できる形で説明する。ここでは、予測を行うような厳密なモデルを定義するのではなく、グループに関する洞察を得ることが主な目的となる。図３は、セグメンテーションによって分けた顧客のセグメントの特徴を分析しているイメージである。

「コンセプトの解説」は、「セグメンテーション」や「分類」と密接に関連している。これらの用語は、以下のように整理できる。

- セグメンテーション：データをグループに分けること
- コンセプトの解説：分けられたグループの特徴を説明すること
- 分類：データがどのグループに属するかを決定するための予測モデルを作成すること

⑷ タイプ４：分類（Classification*1）

「分類」では、データが、どのグループに属するかを決定するための予測モデルを作成する。この予測モデルを利用することで、新しいデータが、どのグループに属するかを判定することができる。分類するためのグループは、事前にユーザーが定義するが、このグループを定義するために「セグメンテーション」を利用することも多い。

図４に分類（Classification）のイメージを示す。既存の顧客データから「優良顧客」と「一般顧客」を分類するための顧客分類モデルを構築し、そのモデルを利用して、新たな顧客である「G さん」が、どちらのセグメントに属するかを判定している例である。

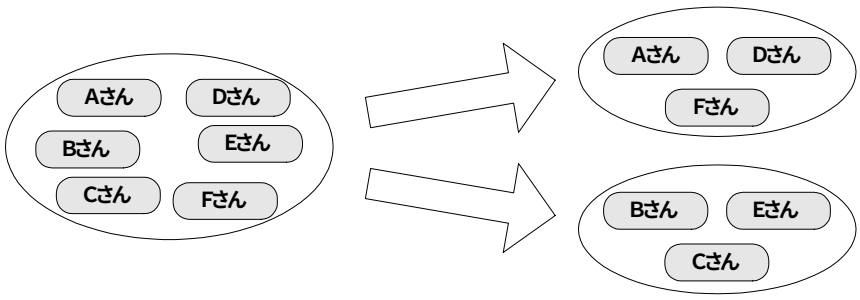

図２　「セグメンテーション（Segmentation）」のイメージ

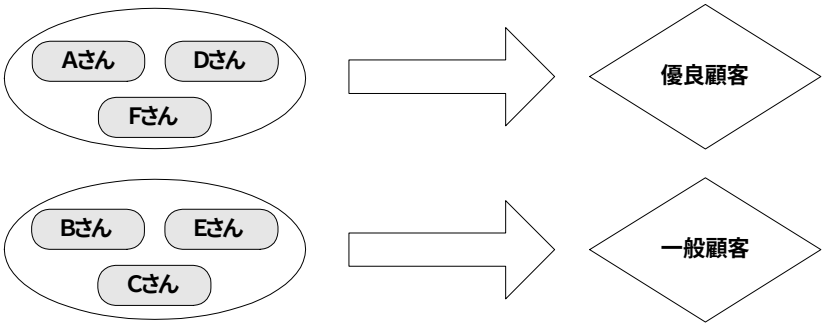

図３　「コンセプトの解説（Concept Description）」のイメージ

＊１）Classificationは「クラス分類」あるいは「クラス判別」と邦訳されることもある。

⑸ タイプ５：予測（Prediction）

「予測」では、「分類」と同様に予測モデルを作成するが、予測する値（目標変数）のタイプが異なる。「分類」がグループ、つまり離散したカテゴリー値を予測するのに対して、「予測」では数値などの連続した値を予測する。一般的に、この区別は厳密なものではなく、両者を包含して「予測（Prediction）」という言葉を利用するケースがよくみられる。また、「予測（Prediction）」のうち、時系列データを利用したものを「時系列予測（Forecasting）」と呼ぶことがある。

図５に、予測（Prediction）のイメージを示す。前年度の顧客購買実績を基に、顧客が１年間でどれだけのお金を自店に費やしてくれるかを予測するモデルを構築し、そのモデルを利用して、新規顧客の「G さん」の年間購買金額を予測している例である。

⑹ タイプ６：従属性分析（Dependency analysis）

「従属性分析」では、データやイベント間の従属性（あるいは関連）をモデルによって表現する。これにより、与えられたデータ・アイテムの値から、関連する、もう一方のデータ・アイテムの値を予測することもできる。しかし、一般的には「従属性分析」は予測ではなく、データからの知識発見に利用されることが多い。

「アソシエーション（関連）」は従属性分析の一種で、近頃では非常にポピュラーな分析手法となっている。これは、データ間の相性、すなわちデータやイベントが同時に発生する頻度を分析するもので、典型的な適用例として「マーケット・バスケット分析」を挙げることができる。図６は、前述のビールと紙オムツ事例の分析イメージである。

「順序パターン（Sequential patterns[*2]）」も従属性の一種で、ここではイベント間の順序が考慮される。マーケット・バスケットの例で考えると、「アソシエーション」は同時に購入される商品を、「順序パターン」は顧客の購買パターンを表現する。

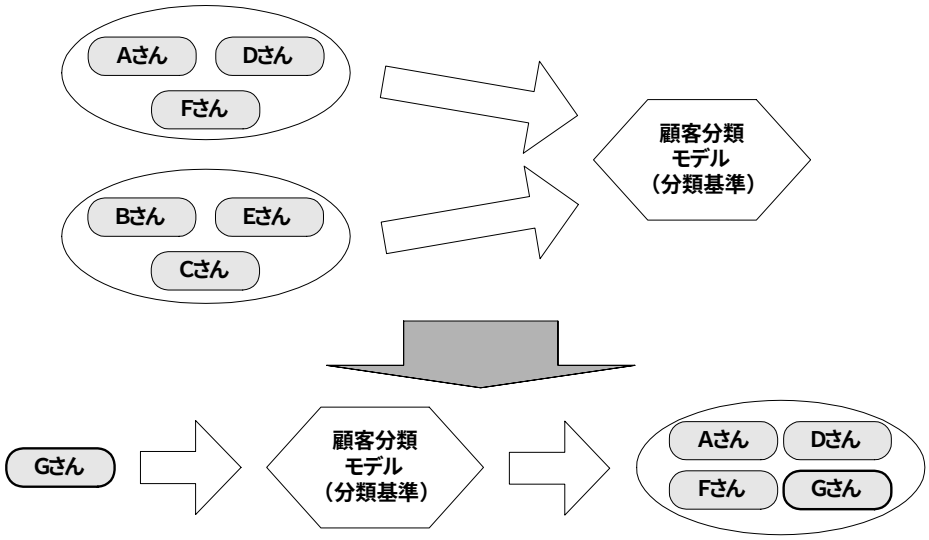

図４　「分類（Classification）」のイメージ

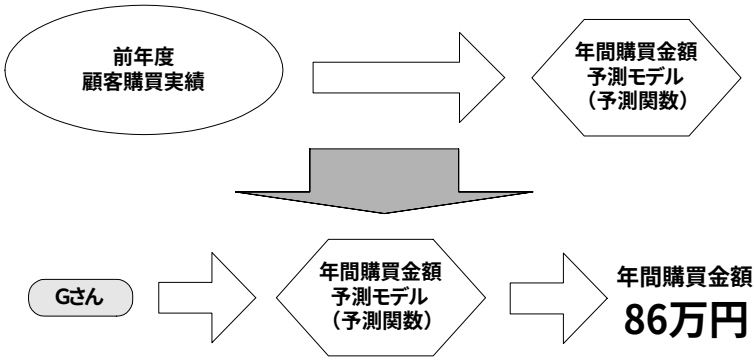

図５　「予測（Prediction）」のイメージ

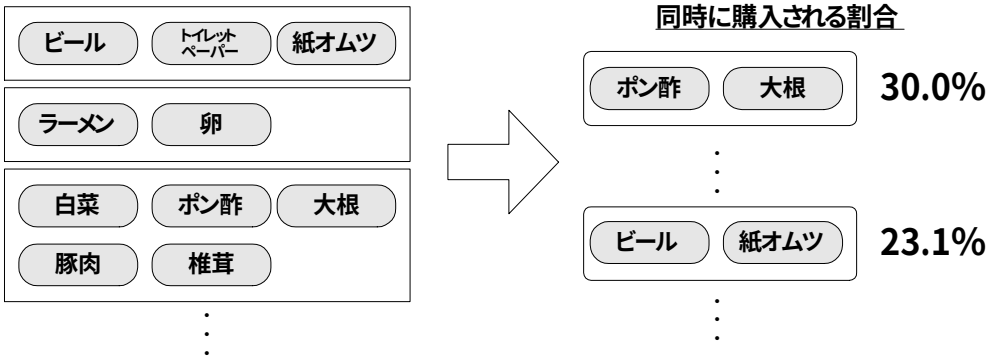

図６　「従属性分析（Dependency analysis）」のイメージ－アソシエーション（関連）の例－

３．データマイニング用方法論 CRISP-DM の紹介

他に先がけてデータマイニングを実践している各社[*3]が共同して作成した、CRISP-DM というデータマイニング用の方法論がある。これは、無償で提供されているにも関わらず、かなり完成度の高い方法論であり、データマイニングを実践するユーザーにとって、非常に有益な情報が提供されている。データマイニング・プロジェクトが、どのような手順で進められるかを理解してもらうために、この方法論を紹介する。

3.1　CRISP-DM とは

CRISP-DM とは、「CRoss-Industry Standard Process for Data Mining」の略語である。日本語に訳すと「業界の枠を越えたデータマイニングの標準プロセス」という意味になり、業界、ツール、業務分野のそれぞれに中立なデータマイニング用の「方法論」として注目されている。

方法論とは、手法や手順を体系的に整理し、明文化して誰でも使えるようにしたものである。方法論を使用するメリットとして、以下のような点が挙げられる。

- 誰もが同じプロセスで作業することが可能（作業を計画／実行する上での手本として利用することもできる）。
- 最短期間、かつ適正なコストで最大の成果を引き出すことが可能。
- 作業プロセスの正当性を評価できる。

CAC 社内でも、EST（Enterprise Systems Transformation）方法論として、「方法論」の利用を積極的に進めている。

3.2　CRISP-DM 開発の目的

CRISP-DM は、新たにデータマイニングを採用しようとするユーザーにノウハウを提供することで、データマイニングを正しく実践してもらい、データマイニングの価値を正当に評価してもらうことを目的に開発された。

近年、ベンダーなどが紹介しているデータマイニングの事例には、センセーショナルなものが多くみられる。その結果、データマイニングへの期待は高まり、データマイニングの採用を検討している企業は、データマイニングを導入することにより、すばらしく画期的な効果を簡単に得ることができると誤解してしまう。しかし、実際にデータマイニングを導入することは容易ではなく、間違った方法でデータマイニングを実践してしまうと、費用ばかりがかさんで何の効果も生み出すことができない。もし、そうした失敗が積み重なってしまった場合、データマイニングそのものの価値や存在意義まで否定されることになりかねない。

そうした事態を防ぐために生み出されたのが、CRISP－DM という方法論である。この方法論により、他に先がけてデータマイニングを実践してきた企業が試行錯誤によって得たノウハウを、誰もが活用することができるようになる。

3.3　CRISP-DM として提供されているもの

CRISP-DM 方法論は、CRISP-DM の公式 Web サイト[*4]から無償で入手できる。ここでは「CRISP-DM 1.0 Step-by-step data mining guide」という方法論（プロセスモデル）のドキュメントが提供されており、PDF 形式のファイルとしてダウンロードできる。残念ながら、現時点では英文のみであり、今後の日本語化が期待される。

＊２）Sequential Patterns は時系列パターン分析と邦訳されることもある。

＊３）SPSS Inc., NCR Corporation, Daimler Chrysler., OHRA Inc.

＊４）CRISP-DM 公式サイト：http : //www.crisp-dm.org/

3.4 CRISP-DM の構成

CRISP-DM プロセスモデルは、以下の５つの要素から構成されている。

Ⅰ．イントロダクション
CRISP-DM の紹介と、プロセスモデルを実際に適用する際の一般的なガイドライン。

Ⅱ．CRISP-DM リファレンスモデル
プロセスモデルを、おおまかに理解するためのクイックリファレンス。それぞれのフェーズで、どんなタスクを行い、どんなアウトプットを作成するかを説明。

Ⅲ．CRISP-DM ユーザーズガイド
プロセスモデル本体。TIPS やヒント、チェックリストなどを含む各フェーズ、タスクで行うべきことを詳細に説明。

Ⅳ．CRISP-DM アウトプット
プロジェクト実施中、および実施後に作成される報告書の詳細な説明と書くべき内容を説明。ジェネリックタスクとアウトプットのクロスリファレンスも掲載。

Ⅴ．付録
用語集、データマイニング問題のタイプの説明。

3.5 プロセスモデルの構造

CRISP-DM プロセスモデルでは、作業をグルーピングするために、フェーズ、ジェネリックタスク、スペシャライズドタスク、プロセスインスタンスの４つのレベルを設定している。これらは図７のような階層構造となっている。

階層を上がるほど抽象度が高くなり、下がるほど具体的な内容となる。このうち、「フェーズ」と「ジェネリックタスク」のレベルで構成されるジェネリック[*5]なプロセスモデルは、どのようなデータマイニング・プロジェクトでも利用できるように構成されており、CRISP-DM はこの部分を提供している。ユーザーは、利用者固有の状況に応じてカスタマイズして利用することになる。この「利用者固有の状況」をプロセスモデルでは「データマイニング・コンテキスト」と名付けており、具体的には以下のような内容となる。表１に例を示すので、合わせて参照されたい。

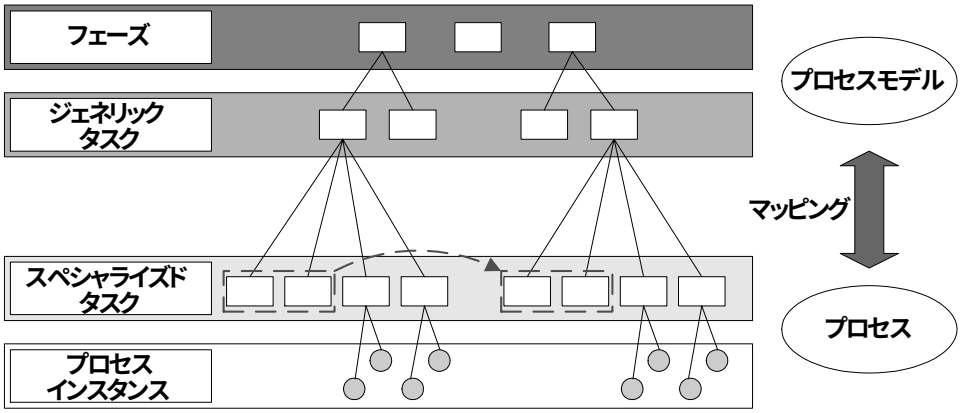

図７　CRISP-DM の階層構造

表１　データマイニング・コンテキスト

	データマイニング・コンテキスト			
次元	アプリケーション領域	データマイニング問題のタイプ	技術的側面	ツールと手法
例	ダイレクトメールの反応予測	データの解説と集計	欠損値	Enterprise Miner
	クレジット解約予測	セグメンテーション	外れ値	Intelligent Miner
		コンセプトの解説		Clementine
		分類		決定木
		予測		ニューラルネット
		従属性分析		

＊５）ジェネリックは「一般的な」「総称的な」という意味。ここではそのまま利用する。

・アプリケーション領域：データマイニング・プロジェクトを実施する業務領域
・データマイニング問題のタイプ：データマイニング適用方法の分類
・技術的側面：データマイニングを行う際の技術上の課題
・ツールと手法：プロジェクトで利用されるデータマイニング・ツールや手法

4．データマイニング・プロジェクトのライフサイクル

ここでは、CRISP-DM プロセスモデルのフェーズにしたがって、データマイニング・プロジェクトのライフサイクルを紹介する。

4.1 ライフサイクルの概観

CRISP-DM で提供しているデータマイニング・プロジェクトのライフサイクルを図８に示す。

四角い箱はフェーズを、矢印はフェーズの流れを示している。ライフサイクルは、６つのフェーズから構成されており、「ビジネスの理解」から始まって、最終フェーズの「展開」に進むまでのフェーズ間の移動を矢印で示している。しかし、必ずしもこの順序で作業を進めなければならない訳ではなく、前のフェーズに戻ったり、最初からやり直したり、ということが当たり前に発生する。なお、外円の矢印はデータマイニング・プロジェクトが本質的に繰り返して行われることを表現している。

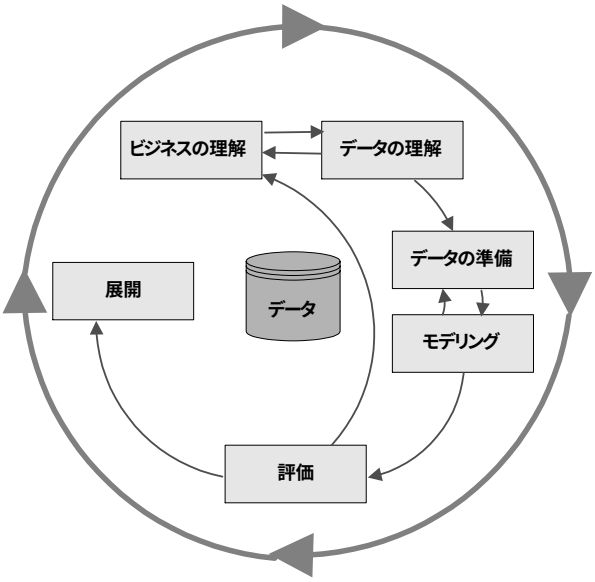

図８　CRISP-DM のライフサイクル

4.2 フェーズの解説

各フェーズで行う作業の内容を、以下に簡単に説明する。

(1) フェーズ１：ビジネスの理解（Business Understanding）

最初のフェーズでは、実施しようとしているデータマイニング・プロジェクトが、自社のビジネスにとってどのような役割を果たすのか、どのような効果をもたらすのかということを明確にする。この際、技術的視点ではなく、ビジネスの視点で考えることが重要であり、これを誤ると分析結果が正しくとも、実践では役に立たないという事態に陥ってしまう。その後、ビジネス上の目標を達成するにはどのような分析が必要か、という観点からデータマイニングの目標設定を行い、プロジェクトの実施計画を策定する。

(2) フェーズ２：データの理解（Data Understanding）

このフェーズでは、分析に利用するデータを収集し、整理と簡単な分析を行う。この中で発生する、データの意味を理解したり、データ品質の問題を発見したりという作業を通じて、どんな知識が発見できそうか、というデータマイニング実施のヒントを見つけることができる。

(3) フェーズ３：データの準備（Data Preparation）

このフェーズでは、モデリングに利用するデータを準備する。収集されたデータの内、どの情報（属性）がデータマイニングに役立ちそうかを判断し、さらに利用しようと考えているデータマイニング・アルゴリズムに合わせた形で属性値の加工、欠損値や外れ値のあるデータのクリーニングを行う。そして、最終的にはデータマイニング・ツールが利用できる形式にデータを編成する。

なお、次フェーズで実施するモデリング（データマイニングの実行）が、１回の実行で終わるケースはほとんどない。そのため、モデリングを実施するたびに、この「データの準備」作業を繰り返し行う必要がある。

(4) フェーズ４：モデリング（Modeling）

このフェーズでは、データマイニング処理を実行してモデルを作成する。利用可能なモデリング手法とアルゴリズムから適切なものを選択し、処理を実行しながらパラメータを最適化していく。そうして作成されたモデルが、正しくて意味のあるものかどうかを評価しながら、繰り返しマイニング処理を実施する。

(5) フェーズ５：評価（Evaluation）

モデリングが成功して正しく意味のあるモデルができたとしても、それがビジネスに効果をもたらすとは限らない。このフェーズでは、できあがったモデルをビジネスの視点から評価して、ビジネス上の目標を達成できるかどうかを確認する。モデルがビジネスに役立つものだと判断できた場合は、データマイニングを実施してきたプロセスをレビューして、ミスや間違いが無かったかを検証する。そして、さらに分析を繰り返すか、終了して分析の成果をビジ

ネスに展開するかなど、その後のアクションを決定する。

⑹ フェーズ６：展開（Deployment）

最後のフェーズでは、データマイニングの分析結果を、何らかの形でビジネスに展開する。実際にどんな作業を行うかは、プロジェクトの目的や要件によって異なってくる。例えば、データの分析そのものが目的であった場合には、簡単な報告書を作成するだけで済むかも知れない。また、Web ページのパーソナリゼーションやマーケティング・データベースのスコアリングのように、マイニングの結果をシステムに組み込む必要がある場合もある。いずれにしても、このフェーズの中で分析結果の展開に関する戦略を立案し、どのようにして実現させるのかを決めることになる。

変化の激しい分野でプロジェクトが実施された場合、作成されたモデルがいつまでも有効なものであるとは限らない。そのため、モデルの有効性を常にチェックするための計画も必要となる。そして、最後にプロジェクト全体のレビューを実施し、そこで得た教訓を文書化することによって、知識の再利用を促進することができる。

表２に、それぞれのフェーズ内で実施するタスクと作成する成果物を、一覧形式にまとめたので参照されたい。

５．CRISP-DM によるデータマイニングの実践

実際のデータマイニング・プロジェクトは、どのように進められるのだろうか。本章では、架空の通信販売業 A 社における、CRISP-DM に沿って進められる分析の過程を紹介する。

5.1 CRISP-DM を利用した A 社事例の紹介

A 社はカタログ通販という事業形態をとっており、３カ月に１度、年４回のサイクルでカタログを作成している。カタログ作成の後、雑誌や TV などの媒体に広告を掲載することでカタログの請求を募集し、顧客に対して無料でカタログを配布している。顧客はカタログの中から好みの商品を選んで、電話、郵便、Fax などで商品を注文する。

以上のような業務を行う A 社におけるデータマイニング・プロジェクトの実施過程を、フェーズごとに説明していこう。

⑴ フェーズ１：ビジネスの理解

近年、通信販売の利用者は大幅に増加しており、A 社は着実に業績を伸ばしている。しかし、競合する企業が多く、生き残るためには経営の効率化は避けて通れない課題である。

これまで、A 社ではカタログを請求してきた顧客だけでなく、過去１年以内に購買実績のある、既存の全ての顧客に対して最新のカタログを送付してきた。しかし、既存の顧客のリピータはわずかであった。さらに、競争によって商品価格は低下してきており、「買ってくれない顧客」に対してカタログを送付するコストが利益を圧迫していた。そこで、A 社では「カタログ送付先の選別による送付コストの削減」を、データマイニング実施にあたってのビジネス上の目標として設定することにした。さらに、「売上の減少を５％以内に抑えながら、カタログの送付コストを30％削減する」という数値目標を、データマイニングが成功したかどうかを判断する基準として設定した。

表２　CRISP-DM のタスクとアウトプット

フェーズ	1.ビジネスの理解	2.データの理解	3.データの準備 ・データセット ・データセット解説書	4.モデリング	5.評価	6.展開
ジェネリック・タスク	1.1 ビジネス目標の決定 ・バックグラウンド ・ビジネス上の目標 ・ビジネス上の成功基準	2.1 初期データの収集 ・初期データ収集レポート	3.1 データの選択 ・利用／非利用の根拠	4.1 モデリング手法の選択 ・モデリング手法 ・モデリングの前提	5.1 結果の評価 ・ビジネス上の成功基準に関するデータマイニング結果の評価	6.1 展開計画の策定 ・展開計画
	1.2 状況の評価 ・リソース一覧 ・要件、仮定、制約 ・リスクと偶発事象 ・用語集（ビジネス、マイニング） ・費用対効果	2.2 データの記述 ・データ解説書	3.2 データのクリーニング ・データクリーニングレポート	4.2 テストデザインの生成 ・テストデザイン	5.2 プロセスのレビュー ・プロセスレビュー結果	6.2 モニタリングとメンテナンスの計画 ・モニタリングとメンテナンス計画
	1.3 データマイニング目標の決定 ・データマイニング上の目標 ・データマイニング上の成功基準	2.3 データの調査 ・データ調査レポート	3.3 データの作成 ・導出属性 ・生成されたデータ	4.3 モデルの構築 ・設定パラメータ ・モデル ・モデルの説明	5.3 次ステップの決定 ・可能なアクションのリスト ・進め方に関する決定	6.3 最終レポートの作成 ・最終レポート ・最終プレゼンテーション
	1.4 プロジェクト計画の策定 ・プロジェクト計画 ・ツールと技術の初期評価	2.4 データ品質の検証 ・データ品質レポート	3.4 データの統合 ・統合されたデータ 3.5 データの編成（フォーマット） ・再編成されたデータ	4.4 モデルの評価 ・モデルの評価 ・設定パラメータの改訂		6.4 プロジェクトのレビュー ・経験の文書化

目標が定まったら、次はデータマイニングを行う際に、どのようなリソースが利用できるかを調査する。特に自社で保有しているデータについては、正確に把握する必要がある。A 社では､顧客に関する過去のデータを全て蓄積しており、以下のデータを簡単に利用できることが分かった。

・顧客情報（氏名、生年月日、性別等）
・顧客のカタログ請求実績
・顧客の購買実績

利用可能なデータの中から､どのような知識を得れば｢カタログ送付先の選別」に結び付けることができるだろうか。ここからは、データマイニングという技術的な視点で、問題を分析しなければならない｡そこで､顧客の購買実績データに基づいて、顧客を図９のように分類した。

ここでの「新規の顧客」とは、最新のカタログで初めて商品を購入した顧客である。この顧客が継続して商品購入してくれるかを予測できれば、次のカタログを送る価値があるか判断することができる。そこで、継続購入したかどうかを把握している「既存の顧客」の情報を利用することにした。どのようなタイプの「既存の顧客」が商品を複数回購入するかを判定するためのモデルを構築し、これを利用して新規顧客が「複数回商品を購入しそうな顧客」なのか、「初回しか商品を購入しそうにない顧客」かを判定することをデータマイニング上の目標に設定した。A 社のデータマイニング目標のイメージを図10に示す。

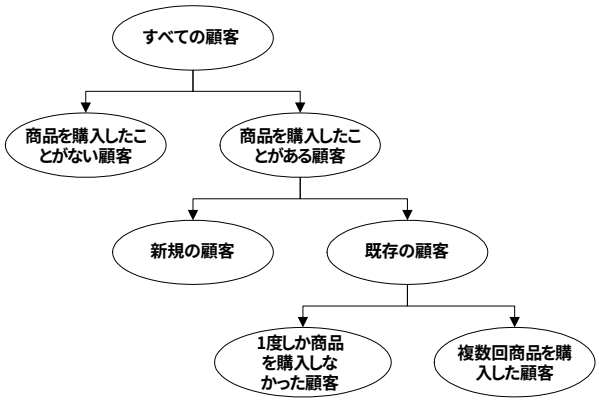

図９　A 社顧客の分類

⑵ フェーズ２：データの理解

データマイニングの目標を設定して詳細なスケジュールを作成した後、分析にとりかかるための準備を進めていく。まず、前フェーズでリスティングされたデータを収集して分析用の環境に取り込む。ここでは、データ利用時の利便性を考えて、リレーショナル・データベースに収集したデータを取り込むことにした。この準備作業を効率的に行うためには、データ収集の際にできるだけデータの量を絞り込んでおく必要がある。そこで、分析要件を勘案しながら、以下のような抽出条件でデータを絞り込んだ。

・過去２年間のデータに限定（２年前に顧客 DB のレイアウトが変更されたため）

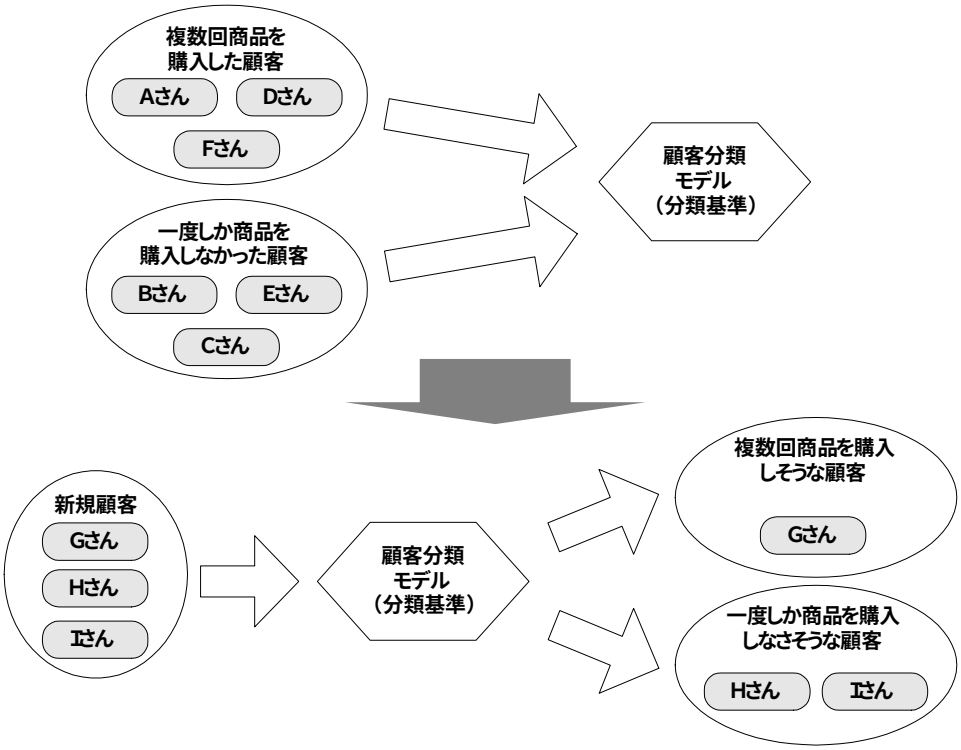

図10　A 社のデータマイニング目標

・一度でも商品を購入した事がある顧客に限定

次に、データ閲覧用のソフトウェアを利用して、実際にデータの中身を調査していく。グラフ化などを利用してデータを照会してみると、以下のような情報を発見できた。このような情報は、後にモデリングを行う際の貴重なヒントとなるため、確実に文書化して記録しておく。

・女性と男性の比率が約３対１と、女性顧客が圧倒的に多い。
・顧客の年齢は20代前半～60代くらいまでで、30代～40代が特に多い。
・複数回（２回以上）購入した顧客の中には、何度も繰り返し購入してくれる顧客が相当数存在した。ほぼ毎月、何らかの商品を購入してくれる顧客もみられた。

また、以下のようなデータ品質の問題も明らかになった。これらの問題は、分析前に解決しておく必要があるため、このような情報についても確実に文書化しておく。

・３才や150才といった年齢の顧客が存在する（外れ値）。
・性別の欄に値が入っていない（欠損値）。

分析者は、データをよく理解している技術者や業務担当者へ質問を繰り返し行うなど、データの意味や特徴を理解しておくことが必要である。ある程度、データの理解ができた時点で、いよいよマイニングの準備にとりかかる。

⑶ フェーズ３：データの準備

前フェーズで理解したデータの意味を吟味しながら、顧客に関する、どの情報を分析に利用すれば精度の高いモデルが構築できるかを考える。今回のケースでは「複数回買う顧客」と「一度しか買わない顧客」を分けるポイントとなる項目を見つけ出さなければならない。住所という切り口ではどうか。都市よりも地方に住む人の方が、通信販売を利用するメリットが高いのではないだろうか。さらに、一度メリットを感じてもらえれば、リピータとなる確率は高くなるだろう。このような想像から、住所に関する情報を分析に利用することは、充分に意味があると考えられる。しかし、都道府県や市区町村のレベルで情報を分析したとしても、理解しにくいモデルになってしまう。そこで、「関東／近畿」といった地方名と、「都市部／都市近郊／郊外」といった、おおまかな分類で分析に利用することにした。他の全てのデータ項目についても、利用／非利用を検討した上で、モデリングに利用する情報を決定した。

また、このフェーズではデータのクリーニングも行う必要がある。これは、前フェーズで発見されたデータ品質に関する問題を解決する作業である。幸い、異常な年齢が入っているデータや性別に値が入っていないようなデータは、それほど件数が多くなかったため、いずれも分析対象から除外するという対応をとった。

データの種類や件数によっては、以下のような対応をとる場合もある。

・最も出現頻度の高い値を設定する（圧倒的に多い「女性」を性別に設定する）。
・平均値を設定する（全顧客の平均年齢を利用する）。
・欠損データの予測モデルを構築して、欠損している値自体を予測する。
・「Unknown」など、データに異常があったことが分かるような値を設定する。

さらに、より効率的に分析を行うために、存在するデータ項目を利用して、新たな情報を作り出すことも必要となる。実は、今回の分析で最も重要なのは、予測の対象となる「複数回買った顧客」と「一度しか買わなかった顧客」を識別するための項目なのだが、通常のデータベースでは、このような項目を顧客の属性として管理することはない。そこで、顧客ごとに購買実績レコードを集計することによって、この項目を新たに作成した。先に挙げた市区町村を「都市部／都市近郊／郊外」といった分類に集約することも、同様の作業である。

このように、さまざまなデータ加工を行った後、データマイニング・アルゴリズムが取り扱える形でデータを再編成した。ここでは、マイニングに利用する「決定木（デシジョンツリー）」というアルゴリズムが取り扱えるように、顧客１人について１レコードの形にデータを統合／編成した。図11に分析に利用するデータのレイアウトを示す。

⑷ フェーズ４：モデリング

準備したデータをツールに取り込み、いよいよデータマイニングの核心部分であるモデリングを実施する。今回のケースでは、決定木（デシジョンツリー）という手法を利用してモデリングを行った。デシジョンツリー手法を利用する場合は、モデルを構築するため、およびモデルを検証するために２種類の入力データが必要となる。これは、作成したモデルが真に正しい予測をするのかどうか、既に予測結果の分かっている別データで検証するためである。ここでは、既存顧客のデータを２つに分けて利用した。

全ての準備が整ったので、実際にモデリング処理を実行した。「どの項目を予測に利用するか」や「何階層まで分割を進めるか」など、マイニング・アルゴリズムにわたすパラメータを設定して処理を実行すると、コンピュータが自動的に顧客を最適な形で分類する。パラメータを調整しながら何度か処理を繰り返した結果、最終的に図12のような分析結果を得ることができた。

デシジョンツリーでは効率的に予測が行えるように、属性値を利用してデータを階層的に分割する。今回のケースでは、「複数回購入する顧客」の比率が高いグループと、そうでないグループに顧客が分割されているはずである。

それでは、実際に図12を参照しながら、モデルの中身（分析結果）を見てみよう。一番上のボックスは、グループに分割される前の分析対象とした全顧客の情報である。これ

顧客を固有に識別するID （顧客コード）	顧客を分類するために利用する属性 （年齢、性別、職業、居住地方、居住地域分類、決済方法、 カタログ請求回数、初回請求日、初回商品購入日、 初回購入品目、初回購入金額 等）	予測する対象となる項目 （複数回購入者フラグ）

図11　分析に利用するデータのレイアウト

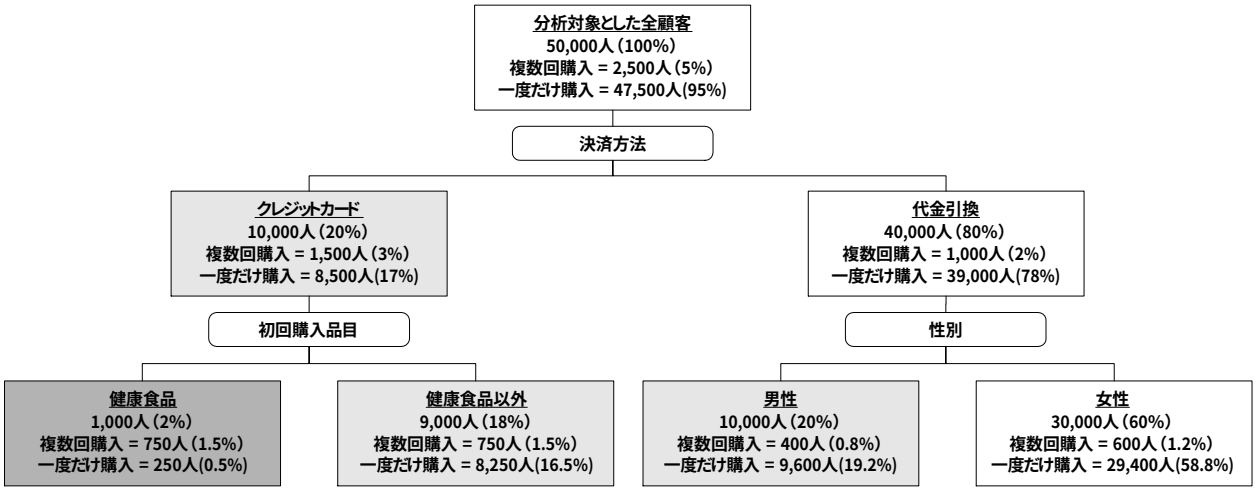

図12　決定木（デシジョンツリー）による分析結果

によると、分析対象とした顧客の総数は50,000人で、その内の５％にあたる2,500人が複数回、商品を購入したことが分かる。これらの顧客を、まず「決済方法」によって２つのグループに分割したのが２番目の階層である。この階層をみると、以下のように「クレジットカード」で決済しているグループの方が「代金引換」で決済するグループよりも、複数回購入する顧客の割合が高いことが解る。

・クレジットカード：10,000人のうち1,500人＝15%の顧客が複数回購入
・代金引換：40,000人のうち1,000人＝2.5%の顧客が複数回購入

ここで、「クレジットカード」で決済している顧客のボックスの方が濃い色なのは、「複数回購入する顧客の割合」の高さを、ボックスの色の濃さで表現しているためである。

３番目の階層では、「クレジットカード」で決済している顧客のグループが「初回購入品目」で、「代金引換」で決済している顧客のグループが「性別」で分割されている。その結果、「複数回購入する顧客」の割合が最も高いのは、左下にあるボックスが示す「決済方法がクレジットカードで、初回購入品目が健康食品」の顧客グループとなった。以下のとおり、カタログを送付した後で商品を買ってくれる割合だけで考えると、全ての顧客に対してカタログを送付した場合と比べて、15倍も効率がよくなるということになる。

・分析対象とした全ての顧客：50,000人のうち2,500人＝５％の顧客が複数回購入
・クレジット決済で初回に健康食品を購入した顧客：1,000人のうち750人＝75%の顧客が複数回購入

しかし、ここで効率だけを重視してしまうと、多くの顧客を失ってしまうことになりかねない。このグループだけにカタログを継続送付した場合、カタログの送付コストは前顧客50,000人に送付する場合と比べると98%も削減できる。しかし、その一方で再購入してくれる顧客は、わずか30%にまで減少してしまう。そこで、「送付コストの削減」と「再購入見込顧客の減少」という２つの面でバランスをとりながら、どのグループに対してカタログ送付を行うのかを見きわめることが重要である。なお、説明を簡潔にするために、図12には２階層４つのグループまでしか記載しなかった。実際には、さらに深い階層、かつ多くのグループに分割されていると考えてほしい。

⑸　フェーズ５：評価

その後も、何度かデータの準備とモデリングを繰り返した結果、予測のために最適と考えられるモデルが構築できた。ここで、このモデルを利用することによって、最初に設定したビジネス上の目標を達成することができるかどうか、再確認しておく必要がある。当初に設定したビジネス上の目標は「カタログ送付先の選別による、カタログ送付コストの削減」であり、その成否を判定する基準は「売上の減少を５％以内に抑えながら、カタログの送付コストを30%削減すること」であった。構築できたモデル中のいくつかのグループを選定してカタログを送付することで、理論上はビジネス目標をクリアすることができる。そこで、

データマイニングを実施してきたプロセスにも問題がないことを確認し、データマイニングの結果をビジネスに展開していくことを決定した。

⑹ フェーズ6：展開

A社ではマイニングの結果を受け、新規の顧客を分類して「複数回商品を購入する」可能性の高い顧客グループだけにカタログを発送することにした。この決定は、直接的に送付コストの削減につながるため、ビジネス上の成功基準の片方である「カタログ送付コストの30%削減」は実現したことになる。もう一方の「売上の減少を５%以内に抑える」については、新規顧客の今後の購買行動にかかっているが、A社ではこの目標も達成されるだろうと自信を持っている。

顧客の購買行動は時間が経てば変化してしまうため、A社では今回作成したモデルを常にアップデートしていく計画が立てられた。新しいカタログを発行するサイクルと同じ３カ月周期で分析を行い、常に最新の顧客に対応できるモデルを構築する計画である。

さらに、分析継続にあたって、データマイニング用のデータマートを構築することになった。「複数回買った顧客」と「一度しか買わなかった顧客」を識別する項目のように、分析を実行する前に作成する必要がある項目をデータマート内で管理することにより、分析スピードの向上を狙っている。また、最終的に顧客がどのグループに分類されたかなどの結果情報をデータマートに登録／管理して、任意のグループを指定して顧客を抽出する、といった分析結果の再利用と有効活用も容易に実現できるようになる。

⑺ 新たなテーマへ

分析結果のビジネスへの展開計画を立て、その適用を行うことでデータマイニング・プロジェクトは完了する。

今回の分析により、顧客の購買行動を左右する重要な要素に「商品の決済方法」があることが分かった。そして、特に「クレジットカード」で決済を行う顧客がリピータとなる傾向にあることも判明した。それならば、クレジットカード決済の顧客を増やす戦略を立てれば、業績の向上につながるはずである。

そこで、A社では「クレジット決済の顧客を増やす」ことを目標に、新たなデータマイニング・プロジェクトを開始することを決定した。こうして新たなテーマを発見しながら、データマイニングは繰り返し実施されていく。

６．CRISP-DMを適用した感想

前章では、できるだけ実際に近い状況を設定して、CRISP-DMを利用したデータマイニングの手順を仮想的に紹介した。現在、私の参加しているデータウェアハウス関連技術の研究会では、実際の企業から顧客データを提供していただいて、データマイニングのケーススタディを行っている。データマイニングを実践できる貴重な機会であり、CRISP-DMを適用しながら作業を進めている。まだ始まったばかりで、わずかな経験しかないが、実際にCRISP-DMを利用した感想を以下に述べる。

まず、CRISP-DMを読んでみて、その完成度の高さに驚かされた。無料の方法論で、しかも初期バージョンということもあり、あまり期待していなかったせいなのかも知れない。しかし、実際に読んでみると、有償で提供されている方法論と比較しても遜色なく、先人のノウハウが充分に伝わる内容となっていた。データマイニングを実践するユーザーにとっては、一読するだけでも非常に価値がある。特に初めてデータマイニングを行うユーザーにとっては、データマイニングを理解するうえで、他の資料にはない貴重な情報を得ることができる。

そして、実際にプロジェクトで利用してみると、その有効性をいっそう理解することができる。今回のケーススタディも、最初はCRISP-DMを意識せずに開始したのだが、提供されたデータの方に興味が向いてしまい、何のために、何を分析しているのかということを見失いがちであった。しかし、机上にCRISP-DMタスクの一覧表を置いて参照しながら分析を進めるだけで、自分たちがどのフェーズの作業をしていて、何をしなければならないかを常に意識できるようになった。現在では、作業を行うメンバー全員の意識をまとめるうえでも、CRISP-DMは非常に有効なツールとなっている。

その一方で、英語版しか提供されていないことは最も大きな障害となっている。CRISP-DMを実践するためには、少なくとも各自がそれぞれプロセスモデルに目を通し、書かれている内容を理解しなければならない。プロセスモデル本体は30ページ程度と大した量ではないのだが、それが全て英語で書かれているとなると、読むだけでも大きな負担となる。やはり、早期の日本語化を期待したい。

７．おわりに

本稿では、データマイニング技術の概要とCRISP-DM方法論について紹介してきた。データマイニング技術は、システム開発者が利用するというより、むしろユーザーが自社のデータを有効に活用するために利用する技術といえる。そういう意味では、開発者側であるCACが、この技術をどのような形で応用していくかを考えていく必要がある。だが、ネットビジネスやCRM（Customer Relationship Management）など、データマイニングが必要とされる領域は確実に広がってきている。今後もこの技術に注目していきたい。

〈参考文献〉

1．The CRISP-DM consortium 著：『CRISP-DM 1.0 Step-by-step data mining guide』, http : //www.crisp-dm.org/

2．ピーター・キャベナ他著，河村佳洋他訳：『データマイニング活用ガイド』，トッパン（1999）

3．マイケル J. A. ベリー、ゴードン・リノフ著，SAS インスティチュートジャパン他訳：『データマイニング手法』，海文堂出版（1999）