IBM SPSS Modeler
CRISP-DM ガイド

注記

本書および本書で紹介する製品をご使用になる前に、 39 ページの『特記事項』に記載されている情報をお読みください。

製品情報

本書は、IBM SPSS Modeler バージョン 17 リリース 1 モディフィケーション 0、および新しい版で明記されていない限り、以降のすべてのリリースおよびモディフィケーションに適用されます。

お客様の環境によっては、資料中の円記号がバックスラッシュと表示されたり、バックスラッシュが円記号と表示されたりする場合があります。

原典： IBM SPSS Modeler CRISP-DM Guide

発行： 日本アイ・ビー・エム株式会社

担当： トランスレーション・サービス・センター

目次

前書き

IBM® SPSS® Modeler は、IBM Corp. が提供するエンタープライズ対応のデータ・マイニング・ワークベンチです。 SPSS Modeler は、データを詳細に理解して顧客や市民との関係を改善するのに役立ちます。組織は、SPSS Modeler から得られた洞察を使用して、収益性のある顧客の維持、抱き合わせ販売の機会の識別、新規顧客の引き寄せ、詐欺の検出、リスクの低減、および行政サービス提供の改善を行います。

SPSS Modeler は、ビジュアル・インターフェースを備えているため、ユーザーの特定のビジネス技能が適用しやすくなっています。これにより、より強力な予測モデルを作成し、解決までの時間を短縮することが可能です。 SPSS Modeler では、多数のモデリング手法、例えば、予測、分類、セグメンテーション、関連検出などのアルゴリズムを利用できます。モデルを作成したら、IBM SPSS Modeler Solution Publisher により、そのモデルを全社的に意思決定者またはデータベースに配布できます。

IBM Business Analytics について

IBM Business Analytics ソフトウェアは、一貫性のある詳細かつ正確な情報を提供します。意思決定者は、この情報を信頼して企業業績を改善します。 ビジネス・インテリジェンス、予測分析、財務実績および戦略管理、ならびに分析アプリケーションからなる包括的なポートフォリオにより、現在の業績にすぐに適用できる明確な洞察と、将来の結果を予測する能力が得られます。豊富な業界ソリューション、実績のある手法、および専門のサービスと併用することにより、あらゆる規模の組織が、高い生産性を実現し、意思決定を確実に自動化し、よりよい結果を達成できます。

このポートフォリオの一部である IBM SPSS Predictive Analytics ソフトウェアは、将来のイベントを予測し、その洞察に基づいてプロアクティブに行動し、より優れたビジネス結果を得るために役立ちます。世界中の商業、政府、および学術にかかわるお客様が、競争を優位に進めるために IBM SPSS テクノロジーを使用して、顧客を引き寄せ、維持、成長させ、詐欺を削減し、リスクを軽減しています。 IBM SPSS ソフトウェアを日常業務に取り入れることにより、組織は、予測的な企業、つまりビジネス目標を達成するための意思決定を導いて自動化し、大きな競争優位性を達成できる企業になります。詳細な情報の確認や担当者への連絡については、http://www.ibm.com/spss をご覧ください。

技術サポート

お客様はテクニカル・サポートをご利用いただけます。技術サポートに連絡すると、IBM Corp. 製品の使用に関する支援を受けたり、サポート対象ハードウェア環境についてインストール操作のヘルプを受けたりすることができます。技術サポートに連絡するには、IBM Corp. Web サイト (http://www.ibm.com/support) を参照してください。支援を要求するときは、ご本人、組織、およびサポート契約を確認できるものをご用意ください。

第 1 章 CRISP-DM の概要

CRISP-DM ヘルプの概要

CRISP-DM (Cross-Industry Standard Process for Data Mining) は、業界で実証されている、データ・マイニングの指針となる手法です。

- CRISP-DM には、**方法論**として、プロジェクトの典型的なフェーズの記述、各フェーズに含まれるタスク、およびこれらのタスク間の関係の説明が含まれています。
- CRISP-DM は、**プロセス・モデル**として、データ・マイニングのライフサイクルの概要を提供します。

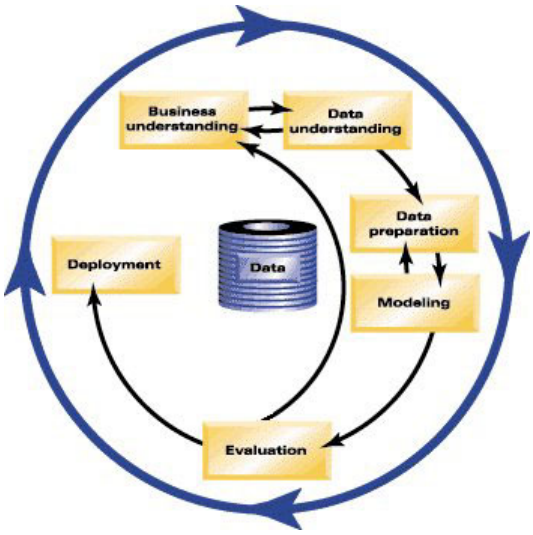

図 1. データ・マイニングのライフサイクル

ライフサイクル・モデルは、6 つのフェーズから構成されています。矢印は、最も重要かつ高い頻度で発生するフェーズ間の依存関係を示しています。各フェーズの順序は厳密ではありません。実際、大半のプロジェクトでは、必要に応じてフェーズ間を行き来して作業を行います。

CRISP-DM モデルは、柔軟性を備えており、簡単にカスタマイズできます。例えば、組織がマネー・ロンダリングの検出を目標にしている場合は、特定のモデリング目標を定めずに大量のデータを取捨選択するでしょう。この場合は、モデリングを行うのではなく、財務データ中の疑わしいパターンを検出するためのデータ検討および視覚化に焦点が当てられます。 CRISP-DM では、特定のニーズに適合したデータ・マイニング・モデルを作成できます。

このような状況の場合、モデリング、評価、および展開の各フェーズは、データの理解フェーズとデータの準備フェーズよりも関連性が低くなることがあります。しかし、その場合でも、これらの後のフェーズで生じた問題のいくつかを検討することは、長期計画や将来のデータ・マイニングの目標を達成するために重要です。

IBM SPSS Modeler における CRISP-DM

IBM SPSS Modeler では、効果的なデータ・マイニングを独自にサポートするために CRISP-DM 方法論が 2 つの方法で組み込まれています。

- CRISP-DM プロジェクト・ツール。典型的なデータ・マイニング・プロジェクトのフェーズに従ってプロジェクトのストリーム、出力、および注釈を編成するのに役立ちます。ストリームおよび CRISP-DM フェーズのメモに基づいて、プロジェクトの任意の時点でレポートを作成できます。
- CRISP-DM のヘルプ。データ・マイニング・プロジェクトの実施プロセスの指針となります。このヘルプ・システムには、各ステップのタスク・リストや、実際の CRISP-DM の利用方法の例が含まれています。 CRISP-DM ヘルプにアクセスするには、メイン・ウィンドウの「ヘルプ」メニューから「**CRISP-DM ヘルプ**」を選択します。

CRISP-DM プロジェクト・ツール

CRISP-DM プロジェクト・ツールは、データ・マイニングに対する構造化されたアプローチを提供するものであり、プロジェクトを確実に成功させるために役立ちます。これは基本的に、標準の IBM SPSS Modeler プロジェクト・ツールを拡張したものです。実際には、CRISP-DM ビューと標準のクラス・ビューを切り替えて、CRISP-DM のタイプ別またはフェーズ別に編成されたストリームおよび出力を表示できます。

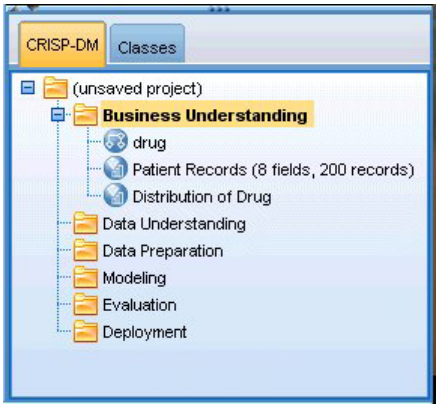

図 2. CRISP-DM プロジェクト・ツール

プロジェクト・ツールの CRISP-DM ビューを使用すると、以下の作業を行うことができます。

- データ・マイニング・フェーズに応じたプロジェクトのストリームおよび出力の編成
- 各フェーズに対する組織の目標の記録
- 各フェーズのカスタム・ツールチップの作成
- 特定のグラフまたはモデルから導き出された結論の記録
- プロジェクト・チームに配布する HTML レポートまたは更新の生成

CRISP-DM のヘルプ

IBM SPSS Modeler は、独自仕様ではない CRISP-DM プロセス・モデルのオンライン・ガイドを提供しています。このガイドは、プロジェクトのフェーズ別に編成されており、以下のサポートを提供しています。

- CRISP-DM の各フェーズの概要およびタスク・リスト
- 各種のマイルストーンのレポートの作成に関するヘルプ
- プロジェクト・チームが CRISP-DM を使用してデータ・マイニング作業を進める方法の実例
- CRISP-DM に関する他のリソースへのリンク

CRISP-DM ヘルプにアクセスするには、メイン・ウィンドウの「ヘルプ」メニューから「**CRISP-DM ヘルプ**」を選択します。

その他のリソース

データ・マイニング・プロセスの理解を深めるには、CRISP-DM に関する IBM SPSS Modeler のサポートに加えて、いくつかの方法があります。

- CRISP-DM コンソーシアム Web サイト (www.crisp-dm.org) にアクセスします。
- CRISP-DM コンソーシアムが作成した、このリリースに付属する CRISP-DM マニュアルを参照します。
- 「マーケティングのためのデータマイニング入門 *データから隠れたパターンを発見する*」(copyright 2002 by SPSS Inc., ISBN 9784492554388) を参照します。

第 2 章 ビジネスの理解

ビジネスの理解の概要

IBM SPSS Modeler で作業する前であっても、組織がデータ・マイニングから得ることを期待している内容を時間をとって検討する必要があります。この議論にはできるだけ多くの重要人物を参加させ、その結果を文書化することを試みてください。この CRISP-DM フェーズの最終段階では、ここで収集した情報を使用してプロジェクト計画を作成する方法を検討します。

この調査は不要に思われるかもしれませんが、そうではありません。データ・マイニングを行うビジネス上の理由を理解することにより、貴重なリソースを使用する前に、全員が同じ考えを持つことができるようになります。

ビジネス目標の確認

最初のタスクは、データ・マイニングのビジネス目標に対して、できる限りの洞察を得ようとすることです。これは思ったほど簡単ではないかもしれません。しかし、問題、目標、およびリソースを明確にすることにより、将来のリスクを最小限に抑えることができます。

CRISP-DM 方法論は、これを達成するための系統立てられた手法です。

タスク・リスト

- 現在のビジネス状況に関する背景情報の収集を開始します。
- 主要な意思決定者が決定した特定のビジネス目標を文書化します。
- ビジネスの観点からデータ・マイニングの成功を判別するために使用される基準に合意します。

e-Commerce の例 -- ビジネス目標の確認

CRISP-DM を使用した Web マイニングのシナリオ

Web 販売に移行する企業が増加するにつれて、ある既存のコンピューター/電機 e-Commerce 業者は、新興サイトとますます激しい競争を繰り広げています。顧客が Web に移行する以上の速さで Web ストアが出現している現実に直面し、同社は、顧客獲得コストが増加している中で、以前と同じ収益性を維持する方法を探す必要があります。提案される一つの解決策は、現在の同社の各顧客の価値を最大化するために、既存の顧客関係を深化させることです。

したがって、以下を目的とした調査が依頼されます。

- より適切なお勧めによる抱き合わせ販売の増進。
- よりパーソナライズされたサービスの提供による顧客ロイヤリティーの向上。

調査が暫定的に成功したと判断されるのは、以下の場合です。

- 抱き合わせ販売が 10 % 増加する。
- 顧客が 1 回あたりのサイト訪問で費やす時間および参照するページ数が増加する。
- 調査が時間どおりに予算内で終了する。

ビジネス背景情報の収集

組織のビジネス状況を把握すると、以下の観点で作業対象を理解するのに役立ちます。

- 使用可能なリソース (要員および資材)
- 問題
- 目標

質問事項に対する実際の回答のうち、データ・マイニング・プロジェクトの結果に影響を与える可能性のあるものを見つけるためには、現在のビジネス状況に関して若干の調査を行う必要があります。

タスク 1 -- 組織構造の確認

- 企業の部門、部署、およびプロジェクト・グループ示す組織図を作成します。管理者の名前と責任を必ず含めてください。
- 組織における重要な個人を確認します。
- 財務的な支援やドメインの専門知識を提供する内部のスポンサーを確認します。
- 運営委員会が存在するかどうかを確認し、メンバーのリストを入手します。
- データ・マイニング・プロジェクトの影響を受ける事業単位を確認します。

タスク 2 -- 問題領域の記述

- マーケティング、カスタマー・ケア、ビジネス開発などの問題領域を確認します。
- 問題を一般的な用語で記述します。
- プロジェクトの前提条件を明確にします。プロジェクトの背後にある動機は何ですか? ビジネスで既にデータ・マイニングが使用されていますか?
- ビジネス・グループ内のデータ・マイニング・プロジェクトの状況を確認します。作業は承認されましたか? データ・マイニングをビジネス・グループの重要なテクノロジーとして「宣伝」する必要はありますか?
- 必要な場合は、組織向けにデータ・マイニングの情報プレゼンテーションを用意します。

タスク 3 -- 現在の解決策の記述

- ビジネス問題に対処するために現在使用されている解決策をすべて記述します。
- 現在の解決策の利点と欠点を記述します。また、この解決策が組織内でどの程度受け入れられているかを検討します。

ビジネス目標の定義

この段階で目標を具体的に定めます。調査およびミーティングの結果として、具体的な主目標を作成する必要があります。この目標は、それらの結果の影響を受けるプロジェクト・スポンサーおよびその他の事業単位との間で合意されたものとなります。この目標は最終的に、「顧客のチャーンの削減」のような漠然とした目標から、分析の指針となる具体的なデータ・マイニングの目標へと翻訳されます。

タスク・リスト

以下のポイントを書き留め、後でプロジェクト計画に取り入れられるようにしてください。目標は現実的なものにすることを忘れないでください。

- データ・マイニングを使用して解決する問題を記述します。
- できる限り正確にビジネス上のすべての問題を明記します。

- その他のビジネス要件があれば、それを確認します (例: 既存の顧客を失わずにクロスセルの機会を増加する)。
- 期待される利益をビジネス用語で明記します (例: 価値の高い顧客のチャーンを 10% 削減する)。

ビジネスの成功基準

将来の目標は明確に決まったかもしれません。しかし、目標に到達したときにそのことがわかるでしょうか? 作業を進める前に、まずデータ・マイニング・プロジェクトのビジネスの成功基準を定義しておくことが大切です。成功基準は、以下の 2 つのカテゴリーに分類できます。

- **客観的。** これらの基準は、監査精度の一定の増加や取り決められたチャーンの削減のような簡単なもので大丈夫です。
- **主観的。** 主観的基準 (効果的な治療法群の発見など) はなかなか明確にできませんが、最終決定者について合意します。

タスク・リスト

- できる限り厳密に、このプロジェクトの成功基準を文書化します。
- 各ビジネス目標に、成功のための相関基準があることを確認します。
- 主観的な成功基準を判断するための責任者を決めておきます。可能であれば、各責任者が何を求めているかを記録しておいてください。

状況の評価

目標を明確に定義したので、次に現在の状況を評価します。このステップでは、以下のような質問事項を確認する必要があります。

- どのような種類のデータを分析に使用できますか?
- プロジェクトを完了するために必要な要員はいますか?
- 関係する最大のリスク因子は何ですか?
- 各リスクに対する代替計画はありますか?

e-Commerce の例 -- 状況の評価

CRISP-DM を使用した Web マイニングのシナリオ

これは、電機 e-Commerce 業者による最初の Web マイニングの試みです。同社は、作業を開始するにあたり、データ・マイニングの専門家の支援を得ることを決定しました。専門家が最初に取り組むタスクの一つが、同社のデータ・マイニング対象のリソースを評価することです。

要員。明らかなことですが、サーバー・ログや製品データベース、購入データベースの管理に関する社内経験はありますが、分析用のデータウェアハウジングやデータ・クリーニングの経験はほとんどありません。したがって、データベースの専門家に助言を求めることにもなるでしょう。同社は、調査結果が継続的な Web マイニング・プロセスの一部になることを期待しています。そのため、経営陣は、現在の作業で生じるポジションがあれば、それが永続的なものになるかどうかを検討する必要もあります。

データ。 これは既存の企業であるため、抽出する Web ログおよび購入データは多数存在します。実際、この初期調査において、同社は、分析対象をサイトの「登録」顧客に限定します。これが成功した場合は、プログラムを展開できます。

リスク。 コンサルタントに対する金銭的な支出、および従業員が調査に費やす時間を除き、この事業に差し迫ったリスクはさほどありません。ただし、時間は常に重要です。そのため、この初期プロジェクトは、単一の財政四半期にスケジュールされています。

また、当面は余分なキャッシュ・フローがあまりないので、調査が予算内に収まることが必須です。ビジネス管理者は、これらの目標のいずれかの達成が危うくなった場合、プロジェクトの範囲を縮小すべきであると提案しました。

リソースのインベントリー

リソースのインベントリーを正確に調べることは不可欠です。ハードウェア、データ・ソース、および人員の問題を実際に調べることにより、時間を大幅に節約し、問題をなくすことができます。

タスク 1 -- ハードウェア・リソースの調査

- どのようなハードウェアをサポートする必要がありますか?

タスク 2 -- データ・ソースおよび知識情報の確認

- どのデータ・ソースをデータ・マイニングに利用できますか? データ型およびデータ形式を記録します。
- データはどのように保管されますか? データウェアハウスまたは運用データベースへのライブ・アクセスが可能ですか?
- デモグラフィック情報などの外部データの購入を計画していますか?
- 必要なデータにアクセスできなくなるセキュリティー上の問題がありますか?

タスク 3 -- 要員リソースの確認

- ビジネスおよびデータの専門家を利用できますか?
- 必要なデータベース管理者やその他のサポート・スタッフを確認しましたか?

これらの質問事項を確認したら、連絡先とリソースのリストをフェーズ・レポートに記入します。

要件、前提、および制約

プロジェクトに関する問題点を率直に評価すれば、作業の効果がより得られやすくなります。これらの問題点をできる限り明確にすることが、将来の問題を回避するのに役立ちます。

タスク 1 -- 要件の確認

基本的な要件は、先に検討したビジネス目標ですが、以下を検討してください。

- データまたはプロジェクトの結果にセキュリティー上の制約および法的な制約はありますか?
- 全員がプロジェクトのスケジュール要件に基づいて適合していますか?
- 結果の展開に関する要件はありますか? (例: Web への公開、データベースへのスコアの読み込み)

タスク 2 -- 前提の明確化

- プロジェクトに影響する可能性のある経済的要素はありますか? (例: コンサルティング料金、競合製品)
- データ品質に関する前提はありますか?
- プロジェクトのスポンサーや管理チームは、どのような結果を期待しているのか？つまり、作成したモデルを理解したいのか、それとも単に結果だけを知りたいのか？

タスク 3 -- 制約の検証

- データ・アクセスに必要なパスワードがすべてありますか?
- データの使用に対する法的な制約をすべて検証しましたか?
- 財務的な制約はすべてプロジェクトの予算でまかなわれていますか?

リスクおよび不測の事態

プロジェクトの間に発生する可能性があるリスクも検討しておくことをお勧めします。リスクのタイプを以下に示します。

- スケジュール (プロジェクトが予想よりも延びた場合はどうなりますか?)
- 財務 (プロジェクトのスポンサーに予算上の問題が発生した場合はどうなりますか?)
- データ (データの品質が悪かったり、必要なデータがなかった場合はどうなりますか?)
- 結果 (最初の結果が予想していたほどでなかった場合はどうなりますか?)

さまざまなリスクを検討したら、事故を避けるための代替計画を策定します。

タスク・リスト

- 考えられる各リスクを文書化します。
- リスクごとに代替計画を文書化します。

用語

ビジネス・チームとデータ・マイニング・チームとが「同じ言葉で話す」ために、技術用語や専門用語の用語集を編集して、それらの意味を明確にすることを検討する必要があります。例えば、ご自分の業務で「チャーン」という用語が特定かつ独自の意味を持つ場合は、チーム全体の作業を円滑に行うためにも、その用語を明確に定義しておく価値があります。同様に、ゲイン・グラフの使用法を明確にしておくことで利益が得られる場合もあります。

タスク・リスト

- チーム・メンバーにとって混乱のもとになる用語または専門用語のリストを作成します。ビジネス用語とデータ・マイニング用語の両方を含めてください。
- 作成したリストをイントラネット上で公開するか、または他のプロジェクト文書に記載することを検討します。

費用対効果の分析

このステップでは、**最終的な収益はいくらですか?** という質問事項に回答します。最終評価の一環として、プロジェクトのコストを、成功した場合の利益と比較することが重要です。

タスク・リスト

分析に以下の推定コストを含めます。

- データ収集、および使用する外部データ (ある場合)
- 結果の展開
- 運用コスト

次に、以下の利益を考慮します。

- 達成される第一目標
- データの検討から得られるその他の洞察

- データの理解を深めることから得られる可能性のある利益

データ・マイニングの目標の決定

ビジネス目標が明確になったので、次はそれをデータ・マイニングの現実に翻訳します。例えば、「チャーンの削減」というビジネス目標は、以下のようなデータ・マイニングの目標に翻訳できます。

- 最近の購入データに基づき、価値の高い顧客を識別する。
- 利用可能な顧客データを使用してモデルを構築し、各顧客のチャーンの可能性を予測する。
- チャーンの傾向および顧客の価値の両方に基づいて、各顧客にランクを割り当てる。

これらのデータ・マイニングの目標が達成された場合は、それらをビジネスで使用して、最も価値の高い顧客のチャーンを削減できます。

おわかりのように、データ・マイニングを効果的に行うには、ビジネスとテクノロジーが協力して作業を行う必要があります。データ・マイニングの目標を決定する方法のヒントについては、本書を読み進めてください。

データ・マイニングの目標

ビジネス上の問題に対する技術的な解決策を定義するためにビジネス・アナリストやデータ・アナリストと共に作業するときは、常に物事を具体的にすることを忘れないでください。

タスク・リスト

- データ・マイニングの問題の**タイプ**を記述します (クラスタリング、予測、種別など)。
- 特定の時間単位を使用して、技術的な目標を文書化します (有効期間が 3 カ月の予測など)。
- 可能な場合は、望まれる結果の数値を実際に記述します (既存顧客の 80% に対してチャーン・スコアを作成するなど)。

e-Commerce の例 -- データ・マイニングの目標

CRISP-DM を使用した Web マイニングのシナリオ

データ・マイニング・コンサルタントの支援を得ることにより、e-Commerce 業者は、同社のビジネス目標をデータ・マイニングの用語に翻訳できるようになりました。この四半期で完了する初期調査の目標は、以下のとおりです。

- 以前の購入に関する履歴情報を使用して、「関連」項目をリンクしたモデルを生成します。ユーザーが項目の説明を参照したときに、関連グループ内の他の項目へのリンクを表示します (**マーケット・バスケット分析**)。
- Web ログを使用して、さまざまな顧客が探しているものを判別し、それらの項目を強調するようにサイトを再設計します。顧客の「タイプ」ごとに、表示されるサイトのメイン・ページは異なります (**プロファイル**)。
- Web ログを使用して、ユーザーがどこから来たのか、およびユーザーがサイトのどこにいるのかを確認した上で、ユーザーの次の移動先の予測を試みます (**シーケンス分析**)。

データ・マイニングの成功基準

データ・マイニング作業を順調に進めるために、成功を技術用語で定義する必要もあります。以前決定したデータ・マイニングの目標を使用して、成功のベンチマークを策定します。 IBM SPSS Modeler は、結果の正確性と妥当性の分析に役立つ評価ノードや分析ノードなどのツールを備えています。

タスク・リスト

- モデル評価手法を記述します (精度、パフォーマンスなど)。
- 成功を評価するためのベンチマークを定義します。特定の数値を指定します。
- できる限り主観的な測定基準を定義し、成功の判定者を決定します。
- モデルの結果の正常な展開をデータ・マイニングの成功の一部とするかどうかを検討します。展開の計画を直ちに開始します。

プロジェクト計画の作成

この時点で、データ・マイニング・プロジェクトの計画を作成する準備が整いました。今までに調査したこと、および策定したビジネス目標とデータ・マイニングの目標が、このロードマップの基盤になります。

プロジェクト計画の作成

プロジェクト計画は、すべてのデータ・マイニング作業のマスター文書となります。適切に作成すれば、データ・マイニングのすべてのフェーズの目標、リソース、リスク、およびスケジュールを、プロジェクトに関わる全員に知らせることができます。計画、およびこのフェーズで収集した文書は、社内イントラネットに公開することもできます。

タスク・リスト

計画を作成するときは、以下の質問事項に回答済みであることを確認してください。

- プロジェクト・タスクおよび提案された計画をプロジェクトに関わる全員と検討しましたか?
- すべてのフェーズまたはタスクの時間見積もりが含まれていますか?
- 結果またはビジネス・ソリューションの展開に必要な作業およびリソースが含まれていますか?
- 決定点およびレビュー要求が計画で強調されていますか?
- モデリングのように一般に何度も反復されるフェーズをマークしましたか?

サンプル・プロジェクト計画

調査の計画の概要を以下の表に示します。

表 1. サンプル・プロジェクト計画の概要

フェーズ	時間	リソース	リスク
ビジネスの理解	1 週間	すべてのアナリスト	経済情勢の変動
データの理解	3 週間	すべてのアナリスト	データの問題、テクノロジーの問題
データの準備	5 週間	データ・マイニングのコンサルタント、1 人以上のデータベース・アナリストの時間	データの問題、テクノロジーの問題
モデリング	2 週間	データ・マイニングのコンサルタント、1 人以上のデータベース・アナリストの時間	テクノロジーの問題、適切なモデルが見つからない
評価	1 週間	すべてのアナリスト	経済情勢の変動、結果を実装できない

表 1. サンプル・プロジェクト計画の概要 (続き)

フェーズ	時間	リソース	リスク
展開	1 週間	データ・マイニングのコンサルタント、1 人以上のデータベース・アナリストの時間	経済情勢の変動、結果を実装できない

ツールおよび手法の評価

データ・マイニングを成功させるための手段として IBM SPSS Modeler を使用することを既に選択しているので、このステップを使用して、ビジネス・ニーズに最適なデータ・マイニング手法を調査することができます。 IBM SPSS Modeler は、データ・マイニングの各フェーズに対応した包括的なツールを備えています。さまざまな手法をいつ使用するのかを判断するには、オンライン・ヘルプのモデリングに関するセクションを参照してください。

次のステップの準備

IBM SPSS Modeler でデータを検討して作業を開始する前に、以下の質問事項に回答済みであることを確認してください。

ビジネスの観点からの質問事項

- ビジネスでは、このプロジェクトから何を得ることを期待していますか?
- どのように作業の正常な完了を定義しますか?
- 目標達成に必要な予算およびリソースはありますか?
- このプロジェクトに必要なすべてのデータへのアクセス権はありますか?
- このプロジェクトに関連するリスクや不測の事態についてチームと検討しましたか?
- 費用対効果の分析の結果、このプロジェクトを実施する価値はありますか?

上記の質問事項に回答した後、それらの回答をデータ・マイニングの目標に翻訳しましたか?

データ・マイニングの観点からの質問事項

- データ・マイニングはビジネス目標を達成するために具体的にどのように役立ちますか?
- どのデータ・マイニング手法を使用すると最良の結果が得られるかについての知識はありますか?
- どのようにして結果が十分に正確または有効であったと判断しますか? (*データ マイニングを成功と判断するための基準を設定しましたか?*)
- モデリングの結果をどのように展開しますか? プロジェクト計画で展開を検討しましたか?
- プロジェクト計画に CRISP-DM のすべてのフェーズが含まれていますか?
- リスクと依存関係は計画に表記されていますか?

上記の質問事項に「はい」と答えることができた場合は、データをより詳細に調査する準備ができています。

第 3 章 データの理解

データの理解の概要

CRISP-DM のデータの理解フェーズでは、マイニング対象のデータをより詳細に調べる必要があります。このステップは、次のフェーズで予期しない問題が発生することを避けるために重要です (次のフェーズはデータの準備であり、通常はプロジェクトの最長部分となります)。

データの理解では、データにアクセスし、表とグラフィックを使用してデータを検討する必要があります。表とグラフィックは、IBM SPSS Modeler で CRISP-DM プロジェクト・ツールを使用して編成できます。これにより、データの品質を判別し、これらのステップの結果をプロジェクト文書に記述できます。

初期データの収集

CRISP-DM のこの時点で、データにアクセスして IBM SPSS Modeler に取り込む準備が整いました。データは、以下のようなさまざまなソースから提供されます。

- **既存のデータ。** これには、トランザクション・データ、調査データ、Web ログなど、幅広いデータが含まれます。既存データでニーズが十分に満たされるかどうかを検討してください。
- **購入履歴データ。**デモグラフィックなどの補足データを組織で使用していますか? 使用していない場合は、そのようなデータが必要かどうかを検討してください。
- **追加データ。** 上記のソースがご自分のニーズを満たさない場合は、既存のデータ・ストアを補足するために、調査の実施、または追加トラッキングの開始が必要な場合があります。

タスク・リスト

IBM SPSS Modeler でデータを参照して、以下の質問事項を検討します。発見は必ず記録してください。詳しくは、トピック 14 ページの『データ収集レポートの作成』を参照してください。

- データベースのどの属性 (列) が最も期待できるように思われますか?
- 無関係のように見え、除外できる属性はどれですか?
- 一般化できる結論を導き出す、または正確な予測を行うのに十分なデータがありますか?
- 選択したモデリング手法に対して属性数が多すぎませんか?
- 各種のデータ・ソースを結合しますか? 結合する場合、そのときに問題が起きる可能性のある領域はありますか?
- 各データ・ソースで欠損値を処理する方法を検討しましたか?

e-Commerce の例 -- 初期データの収集

CRISP-DM を使用した Web マイニングのシナリオ

この例では、e-Commerce 業者は、以下を含むいくつかの重要なデータ・ソースを使用します。

Web ログ。未加工のアクセス・ログには、顧客が Web サイトをどのようにナビゲートしたかについての情報がすべて含まれます。 Web ログ内のイメージ・ファイルおよびその他の非情報エントリーへの参照は、データの準備プロセスの一環として削除する必要があります。

購入データ。 顧客が注文を送信すると、その注文に関するすべての情報が保存されます。購入データベースの注文は、Web ログの対応するセッションにマップする必要があります。

製品データベース。 「関連」製品を判別するときは、製品属性が役立つ場合があります。製品情報は、対応する注文にマップする必要があります。

顧客データベース。 このデータベースには、登録顧客から収集した追加の情報が含まれています。レコードは完全なものではありません。多くの顧客が調査票を記入しないからです。顧客情報は、Web ログの対応する購入履歴およびセッションにマップする必要があります。

現在のところ、同社は、外部データベースを購入する予定も、調査の実施にお金を使う予定もありません。同社のアナリストが、現在所有しているデータの管理に手が離せないためです。ただし、ある時点になったら、データ・マイニング結果のさらなる展開を検討するのがよいでしょう。この場合、未登録顧客のデモグラフィック・データを追加で購入すると、きわめて有益な場合があります。また、デモグラフィック情報を所有すると、e-Commerce 業者の顧客ベースが Web の平均的な買い物客とどのように違うのかを理解するのに役立つ場合があります。

データ収集レポートの作成

前のステップで収集した情報を使用することにより、データ収集レポートの作成を開始できます。レポートが完成したら、プロジェクトの Web サイトに追加したり、チームに配布したりできます。レポートは、後続のステップ (データの記述、検討、および品質検査) で作成されるレポートと結合することもできます。これらのレポートは、データの準備フェーズ全体における作業の指針になります。

データの記述

データを記述するには多くの方法がありますが、ほとんどの記述では、データの数量と品質 (使用可能なデータの量およびデータの状況) に焦点が当てられます。データの記述時に検討すべき重要な特性をいくつか以下に示します。

- **データの量。**ほとんどのモデリング手法では、データ・サイズに関するトレードオフがあります。データ・セットを大きくするとモデルの正確性が向上しますが、その一方で、処理時間が長くなります。データのサブセットを使用できるかどうかを検討してください。最終レポートについて記録するときは、必ず、すべてのデータ・セットのサイズ統計を含めてください。また、データの記述時にレコード数とフィールド (属性) 数の両方を忘れずに検討してください。
- **値のタイプ。** データの形式には、**数値**、**カテゴリー** (ストリング)、**ブール値** (true/false) など、さまざまなものがあります。値のタイプに注意すると、後でモデリングを行うときに問題の発生を防ぐことができます。
- **コード体系。** データベース内の値が特性 (性別や製品タイプなど) を表していることがよくあります。例えば、*男性* と *女性* を表すために、*M* と *F* を使用するデータ・セットもあれば、数値 *1* と *2* を使用するデータ・セットもあります。データ・レポートに矛盾する方式がないか注意してください。

この知識を理解することにより、データ記述レポートを作成して、発見をより多くの読者と共有できるようになります。

e-Commerce の例 -- データの記述

CRISP-DM を使用した Web マイニングのシナリオ

Web マイニング・アプリケーションで処理するレコードおよび属性は多数あります。このデータ・マイニング・プロジェクトを実施している e-Commerce 業者が、初期調査を約 30,000 人のサイト登録顧客に限定したとしても、Web ログには、依然として何百万件ものレコードが存在しています。

これらのデータ・ソースの値のタイプの大半はシンボルです (値のタイプが日時、アクセスされた Web ページ数、登録調査票の複数選択式の質問事項に対する回答であるかどうかにかかわらず)。これらの変数のいくつかは、数値である新しい変数 (アクセスされた Web ページ数や Web サイトで費やされた時間など) を作成するために使用されます。データ・ソース内にある少数の既存の数値変数としては、各製品の注文数、購入時に費やされた金額、製品データベースからの製品の重量と寸法の仕様などがあります。

各種のデータ・ソースのコード体系に重複はほとんどありません。データ・ソースには、非常に異なる属性が含まれているからです。重複する唯一の変数は、「キー」です (顧客 ID や製品コードなど)。これらの変数のコード体系は、データ・ソース間で同一でなければなりません。そうでないと、データ・ソースを結合できなくなります。結合のためにこれらのキー・フィールドを再コーディングするには、追加データをいくつか準備する必要があります。

データ記述レポートの作成

データ・マイニング・プロジェクトを効率的に進めるには、以下のメトリックを使用して、正確なデータ記述レポートを作成する価値を検討します。

データの量

- データの形式。
- データ収集に使用する方法の識別 (ODBC など)。
- データベースのサイズ (行数および列数)。

データ品質

- ビジネス上の質問事項に関連のある特性がデータに含まれるか。
- どのような種類の値がありますか (シンボル値、数値など)??
- キー属性の基本統計を計算したか。これにより、ビジネス上の質問事項に対してどのような洞察が得られたか。
- 関連属性を優先順位付けできるか。できない場合、ビジネス・アナリストがさらなる洞察を提供できるか。

データの検討

CRISP-DM のこのフェーズを使用して、IBM SPSS Modeler で使用可能な表、グラフ、およびその他の視覚化ツールでデータを検討します。このような分析は、ビジネスの理解フェーズで作成したデータ・マイニングの目標に取り組むのに役立ちます。また、仮説を立て、データの準備時に発生するデータ変換タスクを作成するのにも役立ちます。

e-Commerce の例 -- データの検討

CRISP-DM を使用した Web マイニングのシナリオ

CRISP-DM は、この時点で初期検討を行うことを提案していますが、e-Commerce 業者が気付いているように、未加工の Web ログに対するデータの検討は、不可能ではないとしても困難です。通常、Web ログ・データは、データの準備フェーズで最初に処理する必要があります。これは、意味のある検討ができるデータを作成するためです。こうした CRISP-DM からの逸脱は、プロセスは特定のデータ・マイニング・ニー

ズに応じてカスタマイズできるし、またする必要があるという事実を浮き彫りにしています。 CRISP-DM は循環する作業であり、通常、データ・マイナーは、各フェーズ間を行き来しながら作業します。

Web ログは検討前に処理する必要がありますが、e-Commerce 業者が使用可能なその他のデータ・ソースは、これよりも簡単に検討できます。購入データベースを使用して検討すると、顧客に関する興味深い要約情報が得られます。例えば、購入金額、1 回の商品購入数、どこからサイトを訪れたかなどです。顧客データベースの要約情報からは、登録調査票の各項目に対する回答の分布がわかります。

検討は、データのエラーを探すのにも役立ちます。大部分のデータ・ソースは自動的に生成されますが、製品データベースの情報は手動で入力されています。製品寸法をリストした簡単な要約により、「119 インチ」モニター (「19 インチ」ではない) のようなタイプミスを容易に発見できます。

データ検討レポートの作成

提供されているデータに対してグラフを作成し、統計を実行しながら、これらのデータが技術目標とビジネス目標にどのように応えられるかについて、仮説の作成を開始します。

タスク・リスト

データ検討レポートに含める発見を記録します。以下の質問事項に必ず回答してください。

- データに関してどのような種類の仮説を作成しましたか?
- どの属性がさらなる分析の対象として期待できると思われますか?
- データに関する新しい特性が検討で明らかになりましたか?
- これらの検討により、最初の仮説はどのように変化しましたか?
- 後で使用するデータの特定のサブセットを識別できますか?
- データ・マイニングの目標を見直してください。この検討によって、目標は変わりましたか?

データ品質の検証

データが完全であることはめったにありません。実際、大部分のデータには、分析を妨げることのあるコーディング・エラー、欠損値、または他のタイプの不整合が含まれています。このような潜在的な問題を回避する一つの方法は、モデリングの前に、利用できるデータに対して徹底的な品質分析を行うことです。

IBM SPSS Modeler のレポート作成ツール (データ検査ノード、表ノード、その他の出力ノードなど) は、以下のタイプの問題を探すのに役立ちます。

- **データの欠落**。これには、ブランク値、または無応答としてコーディングされた値 (*$null$*、*?*、*999* など) が含まれます。
- **データ・エラー**。通常、これは、データの入力時に発生したタイプミス・エラーです。
- **測定エラー**。これには、正しく入力されたが不適切な測定方式に基づくデータが含まれます。
- **コーディングの不整合**。通常、これには、非標準の測定単位や、値の不整合 (例えば、性別に *M* と *male* の両方を使用する) が含まれます。
- **メタデータの不良**。これには、フィールドの明白な意味と、フィールドの名前または定義で示された意味との不一致が含まれます。

このような品質の問題は、必ず記録してください。詳しくは、トピック 17 ページの『データ品質レポートの作成』を参照してください。

e-Commerce の例 -- データ品質の検証

CRISP-DM を使用した Web マイニングのシナリオ

多くの場合、データ品質の検証は、記述プロセスおよび検討プロセスの過程で行われます。 e-Commerce 業者が遭遇する問題のいくつかを以下に示します。

データの欠落。 知られているデータの欠落には、一部の登録ユーザーによる調査票の未回答などがあります。調査票で提供される追加の情報がない場合、これらの顧客は、以降のいくつかのモデルから除外することが必要な場合があります。

データ・エラー。 大部分のデータ・ソースは自動的に生成されるので、これはそれほど心配する必要はありません。製品データベースのタイプミス・エラーは、検討プロセスで見つけることができます。

測定エラー。 測定エラーの最大の原因として考えられるのは、調査票です。よく考えられていない項目や言葉足らずの項目があると、e-Commerce 業者が望む情報が得られない可能性があります。繰り返しますが、検討プロセスでは、回答の分布が異常である項目に特に注意することが重要です。

データ品質レポートの作成

データ品質の検討および検証に基づいて、CRISP-DM の次のフェーズをガイドするレポートを作成する準備が整いました。詳しくは、トピック 16 ページの『データ品質の検証』を参照してください。

タスク・リスト

前述したように、データ品質の問題には、いくつかのタイプがあります。次のステップに進む前に、以下の品質問題について検討し、解決策を計画します。データ品質レポートにすべての回答を文書化します。

- 欠落している属性およびブランク・フィールドを識別しましたか? 識別した場合、そのような欠損値の背後に意味はありますか?
- 以降の結合または変換で問題になるようなスペリングの不整合はありますか?
- 偏差を検討して、それが「ノイズ」であるのか、それとも詳細分析に値する現象であるのかを判別しましたか?
- 値の妥当性検査を実施しましたか? 明らかな矛盾が見つかった場合は (例えば、ティーンエージャーなのに高収入)、それを記録します。
- 仮説への影響がないデータの除外を検討しましたか?
- データはフラット・ファイルに保管されていますか? フラット・ファイルに保管されている場合、区切り文字はファイル間で一貫していますか? 各レコードに含まれるフィールド数は同じですか?

次のステップの準備

IBM SPSS Modeler でモデリングするデータを準備する前に、以下の点を検討します。

データをどの程度理解していますか?

- すべてのデータ・ソースが明確に識別およびアクセスされていますか? 問題または制限をすべて認識していますか?
- 使用可能なデータからキー属性を識別しましたか?
- これらの属性は、仮説を立てるのに役立ちましたか?
- すべてのデータ・ソースのサイズを記録しましたか?
- 必要に応じてデータのサブセットを使用できますか?

- 関心のある属性ごとに基本統計を計算しましたか? 有意な情報が現れましたか?
- 探索グラフィックを使用して、キー属性へのより深い洞察が得られましたか? この洞察により、いずれかの仮説が再構築されましたか?
- このプロジェクトのデータ品質の問題は何ですか? これらの問題に対処する計画はありますか?
- データの準備ステップは明確ですか? 例えば、結合するデータ・ソースや、フィルターに掛けるまたは選択する属性がわかっていますか?

ビジネスとデータの両方を理解したので、次は IBM SPSS Modeler を使用して、モデリングするデータを準備します。

第 4 章 データの準備

データの準備の概要

データの準備は、データ・マイニングの最も重要かつ時間がかかることの多い側面の一つです。実際、データの準備には、プロジェクトの時間および作業の 50% から 70% が通常かかると見積もられます。初期フェーズであるビジネスの理解およびデータの理解に十分なエネルギーを投入すれば、このオーバーヘッドを最小限に抑えることができますが、それでも、マイニング対象データの準備およびパッケージングには、相当な労力を費やす必要があります。

組織とその目標にもよりますが、通常、データの準備では、以下のタスクを実行する必要があります。

- データ・セットおよびレコード (またはそのいずれか) の結合
- サンプルのデータ・サブセットの選択
- レコードの集計
- 新規属性の派生
- モデリング対象データのソート
- ブランクまたは欠損値の削除または置換
- 学習データ・セットおよびテスト・データ・セットへの分割

データの選択

直前の CRISP-DM フェーズで実施した初期データの収集に基づき、データ・マイニングの目標に関連したデータの選択を開始する準備が整っています。通常、データを選択するには、以下の 2 つの方法があります。

- **項目 (行) の選択**。どのアカウント、製品、または顧客を含めるかなどの決定を行います。
- **属性または特性 (列) の選択**。取り引き額や世帯収入などの特性の使用に関する決定を行います。

e-Commerce の例 -- データの選択

CRISP-DM を使用した Web マイニングのシナリオ

データの選択に関する決定の多くは、データ・マイニング・プロセスの初期フェーズで e-Commerce 業者により既に行われています。

項目の選択。 初期調査の対象は、サイトに登録済みの (約) 30,000 人の顧客に限定されます。そのため、未登録顧客の購入履歴および Web ログを除外するフィルターを設定する必要があります。また、Web ログ内のイメージ・ファイルおよびその他の非情報エントリーの呼び出しを削除する別のフィルターを設定する必要があります。

属性の選択。 購入データベースには、e-Commerce 業者の顧客に関する機密情報が格納されます。そのため、顧客名、住所、電話番号、クレジット・カード番号などの属性をフィルターに掛けることが重要です。

データの組み込みまたは除外

組み込むまたは除外するデータのサブセットを決定したら、その決定の根拠を必ず文書化します。

検討する質問事項

- 選択した属性は、データ・マイニングの目標に関係がありますか?
- 特定のデータ・セットまたは属性の品質は、結果の妥当性を損なうものですか?
- そのようなデータを修復できますか?
- *性別* や *人種* などの特定フィールドの使用に制限はありますか?

ここでの決定は、データの理解フェーズで立てた仮説と異なりますか? 異なる場合は、必ずプロジェクト・レポートにその理由を文書化します。

データのクリーニング

データをクリーニングするには、分析に含めることを選択したデータの問題について、より詳細に調べる必要があります。 IBM SPSS Modeler でレコード操作ノードおよびフィールド操作ノードを使用してデータをクリーニングするには、いくつかの方法があります。

表 2. データのクリーニング

データの問題	考えられる解決策
データの欠落	行または特性を除外します。または、ブランクを推定値で埋めます。
データ・エラー	ロジックを使用してエラーを手動で発見し、置換します。または、特性を除外します。
コーディングの不整合	単一のコード体系を決定して、値を変換および置換します。
メタデータの欠落または不良	疑わしいフィールドを手動で検査して、正しい意味を特定します。

データの理解フェーズで作成されたデータ品質レポートには、ご使用のデータに特有のタイプの問題に関する詳細情報が含まれています。これは、IBM SPSS Modeler でデータを操作する際の出発点として使用できます。

e-Commerce の例 -- データのクリーニング

CRISP-DM を使用した Web マイニングのシナリオ

e-Commerce 業者は、データ品質レポートに記載されている問題に対処するために、データ・クリーニング・プロセスを使用します。

データの欠落。 オンラインの調査票の一部に回答しなかった顧客は、以降のモデルの一部から除外する必要が生じる場合があります。これらの顧客は、再び質問を受けて調査票に記入することができますが、時間とコストがかかるため、e-Commerce 業者にそれを行う余裕はありません。 e-Commerce 業者ができることは、調査票に回答した顧客と回答しなかった顧客の購入の差をモデリングすることです。これらの 2 群の顧客の購入傾向が類似している場合、調査票がないことはさほど問題ではありません。

データ・エラー。 検討プロセスで見つかったエラーは、ここで訂正できます。ただし、ほとんどの Web サイトでは、顧客がページをバックエンド・データベースに送信する前に、データを正しく入力するように強制されます。

測定エラー。 調査票に対して項目を不適切に入力すると、データの品質に大きな影響が及びます。これは、調査票がない場合と同様の難しい問題です。新しい代わりの質問事項に対する回答を収集する時間やコ

ストがない場合があるからです。項目に問題がある場合、最良の解決策は、選択プロセスに戻り、それらの項目をフィルターに掛けて、今後の分析に含まれないようにすることです。

データ・クリーニング・レポートの作成

データ・クリーニング作業のレポートを作成することは、データの変更を追跡するために極めて重要です。作業の詳細をすぐに参照できるようにしておくと、将来のデータ・マイニング・プロジェクトで役立ちます。

タスク・リスト

レポートを作成するときは、以下の質問事項を検討することをお勧めします。

- データにどのようなタイプのノイズがありましたか?
- そのノイズを除去するために、どのような方法を使用しましたか? どの手法がうまくいきましたか?
- 修復できないケースまたは属性がありましたか? ノイズのために除外されたデータは必ず記録してください。

新規データの作成

新規データの作成が必要になる場合が高い頻度であります。例えば、新規列を作成して、トランザクションごとに延長保証の購入をフラグ付けすると便利な場合があります。この新規フィールド *purchased_warranty* は、IBM SPSS Modeler でフラグ設定ノードを使用して簡単に生成できます。

新規データを作成するには、以下の 2 つの方法があります。

- 属性 (列または特性) の派生
- レコード (行) の生成

IBM SPSS Modeler では、レコード操作ノードとフィールド操作ノードを使用して、さまざまな方法でデータを作成できます。

e-Commerce の例 -- データの作成

CRISP-DM を使用した Web マイニングのシナリオ

Web ログを処理すると、新しい属性が多数作成される場合があります。ログに記録されるイベントに対して、e-Commerce 業者は、タイム・スタンプを作成し、訪問者およびセッションを識別し、アクセスされたページおよびイベントが表すアクティビティーのタイプを記録する必要があります。これらの変数のいくつかは、より多くの属性 (セッション内のイベント間の時間など) を作成するために使用されます。

結合または他のデータ再構築の結果としてさらに属性が作成される場合があります。例えば、各行がセッションになるように行あたりのイベント Web ログが「ロールアップ」されると、セッション中のアクション総数、費やされた合計時間、および合計購入数を記録する新しい属性が作成されます。各行が顧客になるように Web ログが顧客データベースと結合されると、セッション数、アクション総数、使用された合計時間、および各顧客の合計購入数を記録する新しい属性が作成されます。

新しいデータを作成した後、e-Commerce 業者は、検討プロセスを実施して、データが正常に作成されたかどうかを確認します。

属性の派生

IBM SPSS Modeler では、以下のフィールド操作ノードを使用して、新規属性を派生させることができます。

- **フィールド作成ノード**を使用して、既存のフィールドから派生された新規フィールドを作成します。
- **フラグ設定ノード**を使用して、フラグ・フィールドを作成します。

タスク・リスト

- 属性を派生させるときは、モデリングのデータ要件を検討します。モデリング・アルゴリズムで特定のデータ型 (数値など) が期待されますか? 期待される場合は、必要な変換を実行します。
- モデリングを行う前にデータを正規化する必要がありますか?
- 集計、平均、または帰納法を使用して、欠落している属性を作成できますか?
- 背景知識に基づく、既存のフィールドから派生可能な重要な情報 (Web サイトで費やされた時間の長さなど) はありますか?

データの統合

同じ一連のビジネスの質問事項に対して複数のデータ・ソースを使用することは、珍しいことではありません。例えば、同じ一連のクライアントについて、抵当融資と購入デモグラフィック・データの両方にアクセスする場合があります。これらのデータ・セットに同じ固有の ID (社会保障番号など) が含まれている場合は、このキー・フィールドを使用して、IBM SPSS Modeler でデータ・セットを結合できます。

データを統合する基本的な方法として、以下の 2 つがあります。

- データの**結合**。レコードは同じだが属性が異なる 2 つのデータ・セットを結合します。データは、各レコードの同じキー ID (顧客 ID など) を使用して結合されます。生成されるデータは、列または特性が増加します。
- データの**追加**。属性は同じだがレコードが異なる 2 つ以上のデータ・セットを統合します。データは、同じフィールド (製品名や契約期間など) に基づいて統合されます。

e-Commerce の例 -- データの統合

CRISP-DM を使用した Web マイニングのシナリオ

複数のデータ・ソースが存在する場合、e-Commerce 業者は、以下のようなさまざまな方法でデータを統合できます。

- **顧客および製品の属性をイベント・データに追加する。** 他のデータベースからの属性を使用して Web ログ・イベントをモデリングするには、各イベントに関連付けられたすべての顧客 ID、製品番号、および注文書番号を正しく指定し、対応する属性を、処理される Web ログに結合する必要があります。結合されたファイルでは、顧客または製品がイベントに関連付けられるたびに、顧客および製品の情報が複製されることに注意してください。
- **購入履歴情報および Web ログ情報を顧客データに追加する。** 顧客の価値をモデリングするには、顧客の購入履歴情報とセッション情報を当該データベースから抽出し、総計して、顧客データベースに結合する必要があります。このためには、データの作成プロセスの説明に従って、新しい属性を作成する必要があります。

データベースを統合した後、e-Commerce 業者は、検討プロセスを実施して、データの結合が正常に行われたことを確認します。

統合タスク

データの作成と理解に十分な時間を費やさないと、データの統合が複雑になる場合があります。データ・マイニングの目標に最も関連すると思われる項目および属性を検討してから、データの統合を開始します。

タスク・リスト

- IBM SPSS Modeler で結合ノードまたは追加のノードを使用して、モデリングに役立つと考えられるデータ・セットを統合します。
- モデリングに進む前に、結果の出力を保存することを検討します。
- 結合の後、値を**集計**してデータを簡素化します。集計とは、複数のレコードおよび表 (またはそのいずれか) からの情報を要約して、新しい値を計算することです。
- 新規レコードの生成が必要な場合もあります (複数年に渡って結合された納税申告書からの平均控除など)。

データのフォーマット

モデルを構築する前の最終ステップとして、特定の手法において、特定のデータ・フォーマットやデータ順序が必要かどうかを確認すると役立ちます。例えば、シーケンス・アルゴリズムでは、モデルを実行する前にデータを事前にソートする必要があることが珍しくありません。モデルで自動的にソートを実行できる場合でも、モデリングの前にソート・ノードを使用すると、処理時間を節約できる場合があります。

タスク・リスト

データをフォーマットするときは、以下の質問事項を検討してください。

- 使用するモデルは何ですか?
- 使用するモデルでは、特定のデータ・フォーマットまたはデータ順序が必要ですか?

変更が推奨される場合、IBM SPSS Modeler の処理ツールは、必要なデータ操作を適用するのに役立ちます。

モデリングの準備

IBM SPSS Modeler でモデルを構築する前に、以下の質問事項に回答済みであることを確認してください。

- すべてのデータに IBM SPSS Modeler 内からアクセスできますか?
- 初期の検討および理解に基づいて、関連するデータのサブセットを選択できましたか?
- データを効果的にクリーニングしましたか? または、修復不可能な項目を削除しましたか? 最終レポートにすべての決定を文書化してください。
- 複数のデータ・セットが正しく統合されましたか? 文書化する必要がある結合上の問題がありましたか?
- 使用するモデリング・ツールの要件を調べましたか?
- モデリングの前に対処できるフォーマット設定の問題はありますか? これには、必要なフォーマット設定と、フォーマット設定に関連するタスクの両方が含まれます。これにより、モデリングの時間が削減される場合があります。

上記の質問事項に答えることができる場合は、データ・マイニングの最も重要な作業であるモデリングの準備が整っています。

第 5 章 モデリング

モデリングの概要

この時点で、今までの作業の成果が現れ始めます。準備に時間を費やしたデータが IBM SPSS Modeler の分析ツールに取り込まれると、その結果により、ビジネスの理解で示されたビジネス上の問題が明らかになり始めます。

通常、モデリングでは、何度も繰り返しながら実施されます。一般にデータ・マイナーは、デフォルト・パラメーターを使用して複数のモデルを実行した後、パラメーターを微調整するか、またはデータの準備フェーズに戻って、選択したモデルに必要な操作を行います。 1 つのモデル、および 1 回の実行で組織のデータ・マイニングの質問事項に対する回答が十分に得られることはめったにありません。このことが、データ・マイニングを非常に興味深いものにしています。与えられた問題を調べる方法はたくさんありますが、IBM SPSS Modeler は、調査に役立つ多種多様なツールを備えています。

モデリング手法の選択

どのタイプのモデリングがご自分の組織のニーズに適しているかについての考えが既におありかもしれませんが、ここではどのモデルを使用すべきかについての確固たる決定を行います。通常、最適なモデルは、以下の検討事項に基づいて決定されます。

- **マイニングで使用可能なデータ型。** 例えば、関心のあるフィールドは、カテゴリー (シンボル) ですか?
- **データ・マイニングの目標。** 単にトランザクション・データ・ストアに対する洞察を得て、興味深い購入パターンを発掘したいですか? それとも、奨学金の未払い傾向などを示すスコアを作成する必要がありますか?
- **特定のモデリング要件。**モデルでは、特定のデータ・サイズまたはデータ型が必要ですか? プレゼンテーションで結果を容易に使用できるモデルが必要ですか?

IBM SPSS Modeler のモデル・タイプおよびその要件について詳しくは、IBM SPSS Modeler の資料またはオンライン・ヘルプを参照してください。

e-Commerce の例 -- モデリング手法

e-Commerce 業者が採用するモデリング手法は、同社のデータ・マイニングの目標によって決まります。

お勧めの改善。 端的に言えば、この作業では、注文書をクラスタリングして、どの製品同士が最も多く結び付いているかを判別します。より価値のある結果を得るために、顧客データやサイト訪問レコードも追加することができます。 2 ステップ・クラスタリング手法または Kohonen ネットワーク・クラスタリング手法は、このタイプのモデリングに適しています。後で C5.0 ルール・セットを使用してクラスターをプロファイルすることにより、顧客の訪問時に任意の時点で最適なお勧めを判別できます。

サイト・ナビゲーションの改善。 差し当たり、e-Commerce 業者は、しばしば使用されるもののユーザーが見つけるのに数回のクリックが必要なページの特定に重点的に取り組みます。これには、Web ログにシーケンス・アルゴリズムを適用して、顧客が Web サイトを移動する「固有のパス」を生成し、次に、ページ訪問数は多いがアクションは行われないセッションを具体的に探す作業が伴います。後でより詳細な分析でクラスタリング手法を使用することにより、さまざまな「タイプ」の訪問および訪問者を識別し、タイプに応じてサイト・コンテンツを編成および表示することができます。

適切なモデリング手法の選択

IBM SPSS Modeler では、多数のモデリング手法が用意されています。多くの場合、データ・マイナーは、問題をさまざまな方向から検討するために複数の手法を使用します。

タスク・リスト

使用するモデルを決定するときは、以下の問題がその選択に影響を与えるかどうかを検討します。

- モデルではデータをテスト・セットと学習セットに分割する必要がありますか?
- 所定のモデルで信頼できる結果を生成できるだけの十分なデータがありますか?
- モデルでは一定レベルのデータ品質が必要ですか? 現在のデータでこのレベルを達成することはできますか?
- データの型は特定のモデルに適していますか? 適していない場合、データ操作ノードを使用して必要な変換を行うことはできますか?

IBM SPSS Modeler のモデル・タイプおよびその要件について詳しくは、IBM SPSS Modeler の資料またはオンライン・ヘルプを参照してください。

モデリングの前提

最適なモデリング・ツールの絞り込みを開始するときは、意思決定処理を記録してください。データの前提、およびモデルの要件を満たすために行ったデータ操作をすべて文書化します。

例えば、ロジスティック回帰ノードでも、ニューラル・ネット・ノードでも、実行前にデータ型を完全に**インスタンス化** (データ型が既知) する必要があります。つまり、モデルを構築して実行する前に、データ型ノードをストリームに追加して実行し、データを調査する必要があります。同様に、C5.0 のような予測モデルでは、稀なイベント用のルールを予測するときに、データを再バランス化するとよい結果が得られる場合があります。このタイプの予測を行う際には、バランス・ノードをストリームに挿入し、よりバランス化されたサブセットをモデルに適用すると、よりよい結果が得られることが多くなります。

これらのタイプの決定は、必ず文書化してください。

テスト設計の生成

実際にモデルを構築する前の最後のステップとして、モデルの結果をどのようにテストするかをもう一度検討する必要があります。広範なテスト設計の作成作業は、以下の 2 つの部分に分かれています。

- モデルの「適合度」基準の記述
- これらの基準をテストするデータの定義

モデルの**適合度**は、いくつかの方法で測定できます。通常、C5.0 や C&R Tree などの監視モデルの場合、適合度の測定では、特定モデルのエラー率を推定します。 Kohonen クラスター・ネットなどの非監視モデルの場合、測定には展開、解釈のしやすさ、または必要な処理時間などの基準が含まれることがあります。

モデル構築は、反復プロセスであることを忘れないでください。つまり、通常はいくつかのモデルの結果をテストしてから、使用および展開するモデルを決定します。

テスト設計の作成

テスト設計は、作成したモデルをテストするために行われるステップを記述したものです。モデリングは反復的なプロセスであるため、どのようなタイミングでパラメーターの調整を中止して、他の手法やモデルを試してみるかを判断することが重要になります。

タスク・リスト

テスト設計を作成するときは、以下の質問事項を検討してください。

- モデルのテストにどのようなデータを使用しますか? データを学習セットとテスト・セットに分割しましたか? (これはモデリングでよく使われる手法です。)
- 監視モデル (C5.0 など) の適合度をどのように測定しますか?
- 非監視モデル (Kohonen クラスター・ネットなど) の適合度をどのように測定しますか?
- 他のタイプのモデルを試す前に、何回設定を調整してモデルを再実行しますか?

e-Commerce の例 -- テスト設計

CRISP-DM を使用した Web マイニングのシナリオ

モデルの評価基準は、検討中のモデルおよびデータ・マイニングの目標によって決まります。

お勧めの改善。 改善したお勧めは、実際の顧客に提示するまで、完全に客観的に評価する方法はありません。しかし、e-Commerce 業者は、ビジネス上の観点から意味をなす非常にシンプルなお勧めを生成するルールを必要とするでしょう。さらに、ルールは、顧客やセッションごとに異なるお勧めを生成するように十分に複雑にする必要があります。

サイト・ナビゲーションの改善。 顧客が Web サイトのどのページにアクセスしているかという情報に基づいて、e-Commerce 業者は、重要なページへのアクセスのしやすさという観点で、更新されたサイト設計を客観的に評価できます。しかし、お勧めと同様に、再編成されたサイトに顧客がどの程度慣れるかを事前に評価することは困難です。時間と資金に余裕があれば、ユーザビリティー・テストを実施することが望ましい場合があります。

モデルの構築

この時点で、長期間に渡って検討してきたモデルを構築する準備が十分に整ったはずです。最終的な結論を出す前に、時間と場所を確保して、複数のモデルを試してください。通常、ほとんどのデータ・マイナーは、複数のモデルを構築し、結果を比較してから、それらのモデルを展開または統合します。

各種のモデルの進行状況を追跡するために、各モデルで使用される設定およびデータを必ず記録してください。この作業は、他のメンバーと結果を検討し、必要に応じてステップを再追跡するのに役立ちます。モデル構築プロセスの最後の時点で、データ・マイニングの決定で使用する以下の 3 つの情報が得られます。

- **パラメーター設定**。最良の結果を生み出すパラメーターの記録が含まれます。
- 生成された実際の**モデル**。
- **モデルの結果に関する説明**。モデルの実行時とその結果の検討時に発生したパフォーマンスおよびデータに関する問題が含まれます。

e-Commerce の例 -- モデルの構築

CRISP-DM を使用した Web マイニングのシナリオ

お勧めの改善。 クラスタリングは、さまざまなレベルのデータ統合に応じて作成されます。最初は購入データベースのみで、その後、関連する顧客情報およびセッション情報が含められます。統合のレベルごとにクラスターが作成されますが、これらのクラスターは、2 ステップ・ネットワーク・アルゴリズムと Kohonen ネットワーク・アルゴリズム用のさまざまなパラメーター設定で作成されます。これらのクラスタリングごとに、異なるパラメーター設定を持つ C5.0 ルール・セットがいくつか生成されます。

サイト・ナビゲーションの改善。 顧客パスを生成するためにシーケンス・モデリング・ノードが使用されます。このアルゴリズムでは、最も一般的な顧客パスに焦点を当てるのに役立つ最小サポート基準を指定できます。パラメーターに対して各種の設定が試行されます。

パラメーターの設定

大半のモデリング手法では、各種のパラメーターや設定が使用されます。これらを調整することにより、モデリング・プロセスを制御できます。例えば、デシジョン・ツリーは、ツリーの深さ、分岐、およびその他の多数の設定を調整することにより制御できます。通常、最初はデフォルト・オプションを使用してモデルを構築し、以降のセッションでパラメーターを洗練させていきます。

最も正確な結果が得られるパラメーターが決まったら、ストリームおよび生成されたモデル・ノードを必ず保存してください。また、最適な設定を記録しておくと、自動化を決定するときや、新しいデータでモデルを再構築するときに役立ちます。

モデルの実行

IBM SPSS Modeler では、モデルの実行は簡単なタスクです。モデル・ノードをストリームに挿入してパラメーターを編集し、モデルを実行すれば、表示可能な結果が作成されます。結果は、ワークスペースの右側にある「生成されたモデル」ナビゲーターに表示されます。モデルを右クリックすると、結果を参照できます。大部分のモデルでは、生成されたモデルをストリームに挿入して詳細な評価を行い、結果を展開することができます。モデルは IBM SPSS Modeler に保存することもできるので、簡単に再利用することが可能です。

モデルの記述

モデルの結果を調査するときは、モデリングの経験を必ず記録してください。モデル自体に関するメモは、ノードの注釈ダイアログ・ボックス、またはプロジェクト・ツールを使用して保管できます。

タスク・リスト

各モデルに対して、以下のような情報を記録します。

- このモデルから意味のある結論を導き出せるか?
- モデルにより明らかになった新しい洞察または見慣れないパターンがあるか?
- モデルの実行に関する問題はあったか? 処理時間はどの程度妥当であったか?
- モデルに欠損値が多いなどのデータ品質上の問題があったか?
- 記録する必要のある計算の矛盾があったか?

モデルの評価

これで一連の初期モデルを用意できました。これらのモデルを詳細に調べて、最終モデルにするのに十分な正確性や有効性を備えているものを判別します。「最終」には、「展開可能」や「興味深いパターンを示す」など、いくつかの意味があります。前に作成したテスト計画を調べると、この評価を組織の観点から実施するのに役立ちます。

総合的なモデル評価

テスト計画で策定した基準に基づいて、検討中の各モデルについて秩序だった評価を行うことをお勧めします。ここでは、生成したモデルをストリームに追加し、評価グラフ・ノードまたは分析ノードを使用して、

結果の有効性を分析します。また、結果に論理的な意味があるのかどうかや、ビジネス目標に対して結果が単純すぎないかどうかも検討する必要があります (例えば、ワイン > ワイン > ワインのような購入順序)。

評価を行ったら、客観的基準 (モデルの正確性) および主観的基準 (結果の使いやすさや解釈のしやすさ) の両方に基づく順序で、モデルをランク付けします。

タスク・リスト

- IBM SPSS Modeler のデータ・マイニング・ツール (評価グラフ、分析ノード、交差検証グラフなど) を使用して、モデルの結果を評価します。
- ビジネス上の問題の理解に基づいて結果のレビューを実施します。特定の結果の関連性に対する洞察が可能なデータ・アナリストや他の専門家に助言を求めます。
- モデルの結果を簡単に展開できるかどうかを検討します。組織は、結果を Web に展開する必要がありますか? または、結果をデータウェアハウスに戻す必要がありますか?
- 成功基準に基づいて、結果の影響を分析します。ビジネスの理解フェーズで作成した目標を達成していますか?

上記の問題に正しく対処することができて、また現在のモデルが目標を達成できると思われる場合は、モデルの詳細評価および最終展開に進みます。そうでない場合は、学んだことを考慮し、パラメーター設定を調整してモデルを再実行します。

e-Commerce の例 -- モデルの評価

CRISP-DM を使用した Web マイニングのシナリオ

お勧めの改善。 Kohonen ネットワークの一つと、2 ステップ・クラスタリングは、それぞれが妥当な結果を提供します。e-Commerce 業者は、それらの結果からどちらかを選択することが難しいことを理解します。やがて同社は、これらの両方を使用しようと考えました。両方の手法で一致したお勧めを受け入れ、お勧めが異なる状況をより詳細に調査しようとしたのです。わずかな労力とビジネス・ナレッジの適用により、e-Commerce 業者は、2 つの手法の差異を解決するルールをさらに開発できます。

e-Commerce 業者は、セッション情報を含む結果が驚くほどよいことも発見します。お勧めをサイト・ナビゲーションに結びつけることが可能であることを示唆する証拠があります。顧客が次に移動しそうな場所を定義したルール・セットをリアルタイムで使用することにより、顧客がサイトをブラウズしているときにサイト・コンテンツに直接影響を与えることが可能です。

サイト・ナビゲーションの改善。 シーケンス・モデルにより、e-Commerce 業者は、高い確実性で特定の顧客パスを予測できます。これにより、サイト設計に対する変更が比較的少なくて済みます。

変更したパラメーターの追跡

モデル評価時に学んだ内容に基づいて、モデルを見直します。ここでは、以下の 2 つの選択肢があります。

- 既存のモデルのパラメーターを調整します。
- データ・マイニングの問題に対処するために別のモデルを選択します。

どちらの場合でも、モデル構築タスクに戻って、よい結果が得られるまで作業を繰り返します。このステップを繰り返すことについての心配は不要です。ニーズを満たすモデルが見つかるまでデータ・マイナーがモデルの評価と再実行を繰り返すことは、きわめて一般的なことです。これは、一度に複数のモデルを構築し、それらの結果を比較してから、各モデルのパラメーターを調整することの十分な理由となります。

次のステップの準備

モデルの最終評価に進む前に、初期評価が十分であったかどうかを検討します。

タスク・リスト

- モデルの結果を理解できますか?
- モデルの結果は、純粋に論理的な観点から意味をなしていますか? さらに検討を必要とする明らかな矛盾がありますか?
- 第一印象として、結果が組織のビジネス上の問題に対処できるように思われますか?
- 分析ノードとリフト・グラフまたはゲイン・グラフを使用して、モデルの正確性を比較および評価しましたか?
- 複数のタイプのモデルを検討して、結果を比較しましたか?
- モデルの結果を展開できますか?

データ・モデリングの結果が正確かつ適切であると思われる場合は、最終的な展開を行う前により詳細な評価を実施します。

第 6 章 評価

評価の概要

この時点で、データ・マイニング・プロジェクトの大半が完了しました。モデリング・フェーズでは、先に定義した**データ・マイニングの成功基準**に従って、構築したモデルが技術的に適切かつ効果的であることも確認しました。

ただし、先に進む前に、プロジェクトの最初に作成した**ビジネスの成功基準**を使用して、作業の結果を評価する必要があります。これは、得られた結果を組織が確実に利用できるようにするための鍵となります。データ・マイニングにより、以下の 2 つのタイプの結果が作成されます。

- CRISP-DM の前のフェーズで選択された最終**モデル**。
- モデル自体およびデータ・マイニング・プロセスから導き出された結論または推論。これらは、**発見**と呼ばれます。

結果の評価

この段階では、プロジェクトの結果がビジネスの成功基準を満たしているかどうかの評価を定式化します。このステップでは、示されたビジネス目標を明確に理解する必要があります。そのため、プロジェクトの評価では、主要な意思決定者を必ず参加させるようにしてください。

タスク・リスト

最初に、データ・マイニングの結果がビジネスの成功基準を満たしているかどうかの評価を文書化する必要があります。レポートで以下の質問事項について検討します。

- 結果は、明快かつプレゼンテーションしやすい形式で示されていますか?
- 特に強調する必要のある新しい発見または独自の発見はありますか?
- ビジネス目標への適用可能性の順序でモデルおよび発見をランク付けできますか?
- 全般的に見て、これらの結果は、組織のビジネス目標をどの程度達成していますか?
- 結果からどのような問題が新たに生じましたか? これらの問題は、ビジネス用語でどのように表現できますか?

結果を評価した後、最終レポートに含める承認済みモデルのリストを編集します。このリストには、組織のデータ・マイニングの目標とビジネス目標の両方を達成するモデルを含める必要があります。

e-Commerce の例 -- 結果の評価

CRISP-DM を使用した Web マイニングのシナリオ

e-Commerce 業者が初めてデータ・マイニングを経験した場合、その全体的な結果をビジネスの観点から伝えることは非常に容易です。調査を通じて、よりよい製品のお勧めになる見込みがあるもの、および改善されたサイト設計が作成されました。サイト設計は、顧客のブラウズ順序に基づいて改善されます。顧客のブラウズ順序を調べると、顧客が望むサイトの機能が明らかになりますが、これを実現するにはいくつかのステップが必要です。製品のお勧めがよりよいという証拠を伝えることはさらに困難です。なぜなら、決定ルールが複雑になる場合があるからです。最終レポートを作成するために、アナリストは、ルール・セット内で、より簡単に説明できる一般的な傾向を識別することを試みます。

モデルのランク付け。 初期モデルのいくつかは、ビジネスとして理にかなうように思われたため、そのグループ内では、統計基準、解釈の容易さ、および多様性に基づいてランク付けが行われました。したがって、モデルでは、状況に応じて異なるお勧めが提供されました。

新しい問題。調査からわかった最も重要な問題は、どのようにすれば e-Commerce 業者は、その顧客に対する理解を深めることができるかということです。顧客データベースの情報は、お勧めのクラスターを作成する際に重要な役割を果たします。情報が欠落している顧客に対してお勧めを行うための特別なルールが用意されていますが、実際、そのようなお勧めは、登録顧客に対して行えるお勧めよりも一般的な内容となります。

レビュー・プロセス

通常、効果的な方法論には、完了したプロセスの利点と欠点を十分に検討する時間が含まれます。データ・マイニングでも、このことに変わりはありません。将来のデータ・マイニング・プロジェクトをより効果的なものとするために経験から学習することが CRISP-DM の一部となります。

タスク・リスト

最初に、各フェーズで行った活動や決定 (データの準備ステップやモデルの構築など) を要約する必要があります。次に、各フェーズごとに以下の質問事項を検討し、改善提案を作成します。

- このステージは、最終結果の価値を高めるために役立ちましたか?
- この特定のステージまたは操作を能率化または改善する方法はありますか?
- このフェーズにおける失敗または失策は何でしたか? 次回にこのような問題を回避するにはどうしたらよいでしょうか?
- 特定のモデルが役に立たないなどの行き詰まりはありましたか? 労力を無駄にしないためにも、そのような行き詰まりを予測する方法はありますか?
- このフェーズで何か注目すべき事柄はありましたか (よいことでも悪いことでも)? 後から考えて、そのような事柄の発生を予測する明確な手段はありましたか?
- 特定のフェーズで使用できた可能性がある代替の決定または戦略はありますか? あるならば、将来のデータ・マイニング・プロジェクトのためにそれらを記載してください。

e-Commerce の例 -- レポートのレビュー

CRISP-DM を使用した Web マイニングのシナリオ

最初のデータ・マイニング・プロジェクトのプロセスをレビューした結果、e-Commerce 業者は、プロセスのステップ間の相互関係をより深く理解するようになりました。 e-Commerce 業者は、最初は CRISP-DM プロセスの「後戻り」に気が進みませんでしたが、現在では、プロセスの循環性によりその威力が増すことを理解しています。プロセスのレビューを通じて、e-Commerce 業者は、以下のことを理解するようにもなりました。

- CRISP-DM プロセスの別のフェーズで何か異常が発生した場合、検討プロセスに戻ることが常に許可されています。
- データの準備、特に Web ログの準備には時間がかかるため、忍耐が必要です。
- 手近なビジネス上の問題に常に注目しておくことが重要です。なぜなら、データの分析準備が完了すると、大局にかかわらずにモデルの構築を開始することがよくあるからです。
- モデリング・フェーズが完了したら、結果の実装方法と、許可される詳細調査の判別方法を決定する上で、ビジネスの理解がより一層重要になります。

次のステップの決定

これまでに、結果を作成して、データ・マイニングの経験を評価しました。**次は何をするのか**と思われているでしょう。このフェーズでは、データ・マイニングのビジネス目標の観点から、この問いに対する答えを探していきます。基本的に、現時点では以下の 2 つの選択肢があります。

- **展開フェーズを続行する。** 次のフェーズは、モデルの結果をビジネス・プロセスに導入し、最終レポートを作成するのに役立ちます。データ・マイニング作業が失敗しても、CRISP-DM の展開フェーズを使用して、プロジェクトのスポンサーに配布する最終レポートを作成する必要があります。
- **前に戻ってモデルを洗練させるか、または置き換える。** 結果はほぼ最良ではあるが、完全ではないと思った場合は、もう一度モデリングを行うことを検討してください。このフェーズで学んだことを活用すれば、モデルを洗練させて、よりよい結果を得ることができます。

この時点での決定には、モデリング結果の正確性および関連性が関与します。結果がデータ・マイニングの目標とビジネス上の目標を達成している場合には、展開フェーズに進むことができます。どちらに決めた場合でも、必ず評価プロセスを詳細に文書化してください。

e-Commerce の例 -- 次のステップ

CRISP-DM を使用した Web マイニングのシナリオ

e-Commerce 業者は、プロジェクトの結果の正確性と関連性の両方にかなり確信を持ったので、引き続き展開フェーズに進みます。

同時に、プロジェクト・チームは、前に戻り、予測手法を取り入れていくつかのモデルを強化する準備も整っています。この時点で、プロジェクト・チームは、最終レポートの配布と、意思決定者からの認可を待機しています。

第 7 章 展開

展開の概要

展開とは、新しい洞察を使用して組織内を改善するプロセスのことです。これは、例えば、IBM SPSS Modeler モデルを実装し、チャーン・スコアを生成してデータウェアハウスに読み込むという、正式な統合を意味する場合があります。一方、展開は、データ・マイニングから得られた洞察を使用して、組織の変更を引き起こすことを意味する場合もあります。例えば、30 歳を超える顧客の行動変化を示す警告パターンがデータから発見されたとしましょう。この結果は、情報システムに正式に統合されないかもしれませんが、計画およびマーケティングの意思決定を行う際に明らかに役立ちます。

通常、CRISP-DM の展開フェーズには、以下の 2 つのタイプのアクティビティーが含まれます。

- 結果の展開の計画およびモニター
- 最終レポートの作成やプロジェクト・レビューの実施などの最終確認タスクの完了

組織の要件に応じて、これらのステップの一方または両方の実施が必要になる場合があります。

展開の計画

データ・マイニング作業で得られた結果をすぐにでも共有したいことと思いますが、ここでもう少し時間を割いて、結果の総合的な展開を円滑に行うための計画を作成しましょう。

タスク・リスト

- まず、モデルと発見の両方の結果を整理します。これは、データベース・システムにどのモデルを統合し、どの発見を他の人々に公開するかを決めるために役立ちます。
- 展開可能な各モデルに対して、手順を追った展開およびシステムへの統合計画を作成します。モデル出力のデータベース要件などの技術的な詳細も記載します。例えば、システムではモデリング出力を、タブ区切り形式で展開しなければならないこともあります。
- それぞれの確実な発見について、この情報を戦略担当に配布するための計画を作成します。
- 両方の結果に対して、言及する価値のある展開計画の代替案はありますか?
- 展開をどのようにモニターするかを検討します。例えば、IBM SPSS Modeler Solution Publisher を使用して展開されたモデルはどのように更新されますか? モデルが適用不能になった頃合いをどのように判断しますか？
- 展開上の問題を識別し、不測の事態に対する計画を作成します。例えば、意思決定者がモデリングの結果に関する詳細情報を必要として、より詳しい技術情報の提供を求める場合があります。

e-Commerce の例 -- 展開の計画

CRISP-DM を使用した Web マイニングのシナリオ

e-Commerce 業者のデータ・マイニングの結果を正常に展開するには、正確な情報を適切なユーザーに届けることが必要です。

意思決定者。 意思決定者には、サイトに対するお勧めおよび提案された変更を通知し、これらの変更がどのように役立つかについての簡単な説明を提供する必要があります。調査結果が受け入れられた場合は、変更を実装するユーザーにその内容を通知する必要があります。

Web 開発者。 Web サイトを保守するユーザーは、サイト・コンテンツの新しいお勧めおよび編成を導入する必要があります。将来の調査により発生する*可能性* がある変更をこれらのユーザーに通知し、直ちに下準備を行えるようにします。リアルタイム・シーケンス分析に基づくオンザフライ・サイト構築に対応できるようにチームに準備をさせると、後で役立つ場合があります。

データベースの専門家。 顧客データベース、購入データベース、および製品データベースを保守するユーザーには、データベース情報の使用方法、および将来のプロジェクトでデータベースに追加される可能性のある属性を常に通知する必要があります。

特に、プロジェクト・チームは、結果の展開および将来のプロジェクト計画を調整するために、これらの各グループと絶えず連絡を取り合う必要があります。

モニターおよび保守の計画

モデリングの結果を本格的に展開および統合しても、データ・マイニング作業が続くことがあります。例えば、Web サイトの買い物かごによる購入シーケンスを予測するモデルを展開した場合は、このモデルを定期的に評価することにより、その有効性を保証し、継続的な改善を図る必要があるでしょう。同様に、高い価値を持つ顧客の維持率を高めるために展開したモデルは、特定レベルの維持率に達した時点で微調整する必要があるでしょう。このモデルを変更して再利用することにより、価値ピラミッド中の価値はやや低いけれども収益性を見込める顧客を維持できるようになります。

タスク・リスト

以下の問題を記録して、最終レポートに必ず含めてください。

- それぞれのモデルや発見について、どの要素や影響 (市場価格や季節変動) を追跡する必要がありますか?
- 各モデルの妥当性と正確性はどのように測定およびモニターできますか?
- モデルが「古くなった」ことはどのように判断しますか? 正確性のしきい値や、データの期待される変化などに関する詳細な情報を指定します。
- モデルが古くなったらどのように対処しますか? モデルを新しいデータで再構築したり、微調整を行うことができますか? 変更は、新しいデータ・マイニング・プロジェクトが必要になるほど広範囲のものになりますか?
- このモデルが古くなった場合、似たようなビジネス上の問題に流用できますか? こうした場合、各データ・マイニング・プロジェクトのビジネス上の目標を評価するためによい文書が重要となります。

e-Commerce の例 -- モニターおよび保守

CRISP-DM を使用した Web マイニングのシナリオ

モニターにおける当面のタスクは、新しいサイト編成とお勧めの改善が実際にうまく機能しているかどうかを確認することです。例えば、ユーザーは、より直接的なルートで、探しているページに到達できますか? お勧め項目の抱き合わせ販売は増加しましたか? 数週間モニターを行えば、e-Commerce 業者は、調査が成功したかどうかを判断できます。

新規登録ユーザーの組み込みは、自動処理が可能です。顧客がサイトに登録したら、その情報に現在のルール・セットを適用することにより、提供すべきお勧めを判別できます。

お勧め判別用のルール・セットをいつ更新するかを判断することは、難しいタスクです。クラスターを作成するには、特定のクラスター・ソリューションの妥当性に関して人手による入力が必要です。そのため、ルール・セットの更新は、自動プロセスではありません。

将来のプロジェクトでより複雑なモデルが生成されるにつれて、モニターの必要性とその量は、ほぼ確実に増加します。可能なときは、モニターの大部分を自動化して、レビュー用の定期的なレポートを生成します。あるいは、その場で予測を行うモデルを作成して、企業の指針とする方法もあります。この場合は、最初のデータ・マイニング・プロジェクトよりも高度な知識がチームに必要となります。

最終レポートの作成

最終レポートを作成すると、以前の文書の未処理の部分が仕上がるだけではなく、結果を伝えることもできます。これは簡単なことのように思えるかもしれませんが、結果をさまざまな関係者に伝えることは重要です。これには、モデリング結果の実装を担当する技術管理者と、結果に基づいて意思決定を行うマーケティングおよび管理スポンサーの両方を含めてください。

タスク・リスト

最初にレポートの読者を検討します。読者は技術開発者ですか、それともマーケティング担当の管理者ですか? 読者のニーズが異なる場合は、読者ごとに別々のレポートを作成することが必要になる場合があります。いずれの場合も、レポートには、以下の事項の大半を含める必要があります。

- 元のビジネス上の問題の詳細な説明
- データ・マイニングを実施するために使用するプロセス
- プロジェクトのコスト
- 元のプロジェクト計画から逸脱した場合は、それに関するメモ
- データ・マイニングの結果の要約 (モデルおよび発見の両方)
- 提案される展開計画の概要
- 今後のデータ・マイニング作業の推奨事項 (検討およびモデリング中に発見された興味深い手掛かりを含む)

最終プレゼンテーションの準備

プロジェクト・レポートのほかに、プロジェクトの発見をスポンサーや関連部署のチームにプレゼンテーションすることが必要な場合があります。この場合、レポートとほぼ同じ情報を使用できますが、より幅広い観点からプレゼンテーションする必要があります。 IBM SPSS Modeler のグラフは、このタイプのプレゼンテーション用に簡単にエクスポートできます。

e-Commerce の例 -- 最終レポート

CRISP-DM を使用した Web マイニングのシナリオ

元のプロジェクト計画からの最大の逸脱は、今後のデータ・マイニング作業の興味深い手掛かりにもなります。元の計画では、1 回あたりのサイト訪問で顧客により多くの時間を費やさせ、より多くのページを参照させる方法を調べることが要求されました。

結局のところ、顧客を単にオンラインにしておけばよい顧客になるのではないことがわかりました。セッションあたりに費やされた時間の度数分布 (セッションが購入につながったかどうかで分割) から、ほとんどの場合、購入につながったセッションのセッション時間は、購入につながらなかったセッションの 2 つのクラスターのセッション時間の間に位置することがわかりました。

このことが明らかになったので、問題は、購入せずにサイトで長時間過ごす顧客が、ただブラウズしているのか、それとも単に探し物が見つからないのかを調査することです。次のステップでは、購入を促すために顧客の探し物を提供する方法を調べます。

最終プロジェクト・レビューの実施

これが CRISP-DM 方法論の最終ステップです。最終的な印象を系統立てて説明し、データ・マイニング・プロセスで学んだ知識を順序正しく揃えます。

タスク・リスト

データ・マイニング・プロセスに深く関与した人々に、簡単なインタビューを行う必要があります。これらのインタービューで検討すべき質問事項を以下に示します。

- プロジェクトに対する全体的な印象はどうでしたか?
- プロセスの間に何を学びましたか (データ・マイニング全般と使用可能なデータの両方について)?
- プロジェクトのどの部分がうまくいきましたか? どこで問題が発生しましたか? 混乱の回避に役立つ情報はありましたか?

データ・マイニングの結果を展開したら、その結果の影響を受けた顧客やビジネス・パートナーなどにもインタビューを行うことをお勧めします。ここでの目標は、プロジェクトは実施するだけの価値があったかどうか、および目指した利益がプロジェクトから得られたかどうかを判断することです。

これらのインタビュー結果は、プロジェクトの印象と共に最終レポートにまとめてください。この最終レポートでは、データ・ストアのマイニング経験から得られた知識に焦点を当てる必要があります。

e-Commerce の例 -- 最終レビュー

CRISP-DM を使用した Web マイニングのシナリオ

プロジェクト・メンバーのインタビュー。 調査に終始一貫して密接に関係しているプロジェクト・メンバーは、その大半が結果について熱心であり、将来のプロジェクトを楽しみにしているということを e-Commerce 業者は理解します。データベース・グループは、慎重ながら楽観的のようです。彼らは、調査の有用性を高く評価していますが、データベース・リソースへの負担が増加することを指摘します。調査の間、コンサルタントを利用できましたが、今後プロジェクトの範囲が拡大するにつれて、データベースの保守に専念する別の従業員が必要になります。

顧客のインタビュー。 顧客のフィードバックは、今までのところ概ね肯定的です。十分に考えられなかった問題の一つは、サイト設計の変更が既存顧客に与える影響です。数年後、登録顧客は、サイトの編成方法について、特定の期待を膨らませました。登録ユーザーからのフィードバックは、未登録顧客ほど肯定的ではなく、一部のユーザーは、変更に大きな抵抗感を持っています。 e-Commerce 業者は、この問題を念頭に置き、変更により既存顧客を失うリスクを補ってあまりあるほど、新規顧客を獲得できるかどうかを慎重に検討する必要があります。

特記事項

本書は IBM が世界各国で提供する製品およびサービスについて作成したものです。

本書に記載の製品、サービス、または機能が日本においては提供されていない場合があります。日本で利用可能な製品、サービス、および機能については、日本 IBM の営業担当員にお尋ねください。本書で IBM 製品、プログラム、またはサービスに言及していても、その IBM 製品、プログラム、またはサービスのみが使用可能であることを意味するものではありません。これらに代えて、IBM の知的所有権を侵害することのない、機能的に同等の製品、プログラム、またはサービスを使用することができます。ただし、IBM 以外の製品とプログラムの操作またはサービスの評価および検証は、お客様の責任で行っていただきます。

IBM は、本書に記載されている内容に関して特許権 (特許出願中のものを含む) を保有している場合があります。本書の提供は、お客様にこれらの特許権について実施権を許諾することを意味するものではありません。実施権についてのお問い合わせは、書面にて下記宛先にお送りください。

〒103-8510
東京都中央区日本橋箱崎町19番21号
日本アイ・ビー・エム株式会社
法務・知的財産
知的財産権ライセンス渉外

以下の保証は、国または地域の法律に沿わない場合は、適用されません。 IBM およびその直接または間接の子会社は、本書を特定物として現存するままの状態で提供し、商品性の保証、特定目的適合性の保証および法律上の瑕疵担保責任を含むすべての明示もしくは黙示の保証責任を負わないものとします。国または地域によっては、法律の強行規定により、保証責任の制限が禁じられる場合、強行規定の制限を受けるものとします。

この情報には、技術的に不適切な記述や誤植を含む場合があります。本書は定期的に見直され、必要な変更は本書の次版に組み込まれます。IBM は予告なしに、随時、この文書に記載されている製品またはプログラムに対して、改良または変更を行うことがあります。

本書において IBM 以外の Web サイトに言及している場合がありますが、便宜のため記載しただけであり、決してそれらの Web サイトを推奨するものではありません。それらの Web サイトにある資料は、この IBM 製品の資料の一部ではありません。それらの Web サイトは、お客様の責任でご使用ください。

IBM は、お客様が提供するいかなる情報も、お客様に対してなんら義務も負うことのない、自ら適切と信ずる方法で、使用もしくは配布することができるものとします。

本プログラムのライセンス保持者で、(i) 独自に作成したプログラムとその他のプログラム (本プログラムを含む) との間での情報交換、および (ii) 交換された情報の相互利用を可能にすることを目的として、本プログラムに関する情報を必要とする方は、下記に連絡してください。

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

本プログラムに関する上記の情報は、適切な使用条件の下で使用することができますが、有償の場合もあります。

本書で説明されているライセンス・プログラムまたはその他のライセンス資料は、IBM 所定のプログラム契約の契約条項、IBM プログラムのご使用条件、またはそれと同等の条項に基づいて、IBM より提供されます。

この文書に含まれるいかなるパフォーマンス・データも、管理環境下で決定されたものです。そのため、他の操作環境で得られた結果は、異なる可能性があります。一部の測定が、開発レベルのシステムで行われた可能性がありますが、その測定値が、一般に利用可能なシステムのものと同じである保証はありません。さらに、一部の測定値が、推定値である可能性があります。実際の結果は、異なる可能性があります。お客様は、お客様の特定の環境に適したデータを確かめる必要があります。

IBM 以外の製品に関する情報は、その製品の供給者、出版物、もしくはその他の公に利用可能なソースから入手したものです。 IBM は、それらの製品のテストは行っておりません。したがって、他社製品に関する実行性、互換性、またはその他の要求については確証できません。 IBM 以外の製品の性能に関する質問は、それらの製品の供給者にお願いします。

IBM の将来の方向または意向に関する記述については、予告なしに変更または撤回される場合があり、単に目標を示しているものです。

本書には、日常の業務処理で用いられるデータや報告書の例が含まれています。より具体性を与えるために、それらの例には、個人、企業、ブランド、あるいは製品などの名前が含まれている場合があります。これらの名称はすべて架空のものであり、名称や住所が類似する企業が実在しているとしても、それは偶然にすぎません。

この情報をソフトコピーでご覧になっている場合は、写真やカラーの図表は表示されない場合があります。

商標

IBM、IBM ロゴおよび ibm.com は、世界の多くの国で登録された International Business Machines Corporation の商標です。他の製品名およびサービス名等は、それぞれ IBM または各社の商標である場合があります。現時点での IBM の商標リストについては http://www.ibm.com/legal/copytrade.shtml をご覧ください。

インテル、Intel、Intel ロゴ、Intel Inside、Intel Inside ロゴ、Centrino、Intel Centrino ロゴ、Celeron、Xeon、Intel SpeedStep、Itanium、および Pentium は、Intel Corporation または子会社の米国およびその他の国における商標または登録商標です。

Linux は、Linus Torvalds の米国およびその他の国における登録商標です。

Microsoft、Windows、Windows NT および Windows ロゴは、Microsoft Corporation の米国およびその他の国における商標です。

UNIX は The Open Group の米国およびその他の国における登録商標です。

Java およびすべての Java 関連の商標およびロゴは Oracle やその関連会社の米国およびその他の国における商標または登録商標です。

索引

日本語, 数字, 英字, 特殊文字の順に配列されています。なお, 濁音と半濁音は清音と同等に扱われています。

[ア行]

[カ行]

[サ行]

[タ行]

[ナ行]

[ハ行]

［マ行］

［ヤ行］

［ラ行］

C

H

W

Printed in Japan

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