IBM SPSS Modeler CRISP-DM ガイド

注 ： この情報とサポートされている製品をご使用になる前に、「注意事項」(p.49)の一般情報をお読みください。

本版は IBM SPSS Modeler 15 ,および新版で指示されるまで後続するすべてのリリースおよび変更に対して適用されます。

Adobe 製品のスクリーンショットは Adobe Systems Incorporated の許可を得て転載しています。

Microsoft 製品のスクリーンショットは Microsoft 社の許可を得て転載しています。

はじめに

IBM® SPSS® Modeler は、IBM Corp. が開発した企業強化用のデータ マイニング ワークベンチです。SPSS Modeler を使用すると、企業はデータを詳しく調べることにより顧客および一般市民とのリレーションシップを強化することができます。企業は、SPSS Modeler を使って得られた情報に基づいて利益をもたらす顧客を獲得し、抱き合わせ販売の機会を見つけ、新規顧客を引き付け、不正を発見し、リスクを減少させ、政府機関へのサービスの提供を改善することができます。

SPSS Modeler の視覚的インターフェイスを使用すると、特定ビジネスの専門知識を適用し、より強力な予測モデルを実現し、解決までの時間を短縮します。SPSS Modeler では、予測、分類、セグメント化、および関連性検出アルゴリズムなど、さまざまなモデル作成手法を提供しています。モデルを作成した後は、IBM® SPSS® Modeler Solution Publisher により、企業全体の意思決定者やデータベースにモデルを配布することが可能になります。

IBM Business Analytics について

IBM Business Analytics ソフトウェアは、意思決定者がビジネス パフォーマンスを向上させるために信頼する完全で、一貫した正確な情報を提供します。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、および分析アプリケーションの包括的なポートフォリオを利用することによって、現在の実績を明確、迅速に理解し、将来の結果を予測することができます。豊富な業界のソリューション、実績ある実例、専門サービスと組み合わせ、さまざまな規模の組織が、高い生産性を実現、意思決定を自信を持って自動化し、より良い決定をもたらします。

このポートフォリオの一部として、IBM SPSS Predictive Analytics ソフトウェアを使用する組織は、将来のイベントを予測し、その洞察に基づいて積極的に行動し、より優れた業績を実現することができます。全世界の企業、政府、学術分野のお客様が IBM SPSS の技術を活用し、不正行為を減少させ、リスクを軽減させながら、顧客の獲得、保持、成長において、競争優位を高めることができます。IBM SPSS ソフトウェアを日々の業務に取り入れることによって、組織は業務目標を達成し、大きな競争的優位を獲得することができるよう、意思決定を方向付け、自動化することができるようになります。お問い合わせは、http://www.ibm.com/spss を参照してください。

テクニカル 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。IBM Corp. 製品の使用方法、または対応するハードウェア環境へのインストールについてサポートが必要な場合は、テクニカル サポートにご連絡ください。テクニカ

ル サポートの詳細は、IBM Corp. Web ページ http://www.ibm.com/support を参照してください。ご本人、組織、サポートの同意を確認できるものをご用意ください。

内容

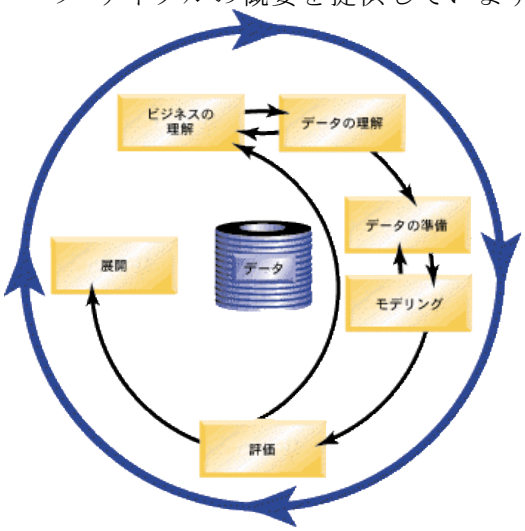
章 1

CRISP-DM の紹介

CRISP-DM ヘルプの概要

CRISP-DM (Cross-Industry Standard Process for Data Mining) は、すでに業界で実証されているデータ マイニング手法です。

- CRISP-DM には**手法**として、プロジェクトの一般的なフェーズの説明、各フェーズで行われる作業、およびこれらの作業間の関係の説明などが含まれています。
- また、**プロセス モデル**として、CRISP-DM はデータ マイニングのライフ サイクルの概要を提供しています。



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、重要で頻繁に発生するフェーズ間の依存関係を表しています。各フェーズの順序は固定されていません。実際に、大半のプロジェクトでは必要に応じてフェーズ間を前後に移動して作業を行います。

CRISP-DM モデルは柔軟で、簡単にカスタマイズすることができます。たとえば、マネー ロンダリングの検出を目的にしている場合、特定のモデル作成目標を持たずに大量のデータの取捨選択を行うことができます。この場合、モデルを作成する代わりに、金融データ中の疑わしいパターンを検出するための、データの探索と視覚化に専念します。CRISP-DM では、特定のニーズに適したデータ マイニング モデルを作成することができます。

このような状況下では、モデリング、評価、および展開フェーズは、データの理解および準備フェーズよりも関連性は低いものとなります。ただし、それでもこれらの後段階のフェーズで提起された問題を検討することは、長期計画や将来のデータ マイニングの成功のためにも大切です。

IBM SPSS Modeler の CRISP-DM

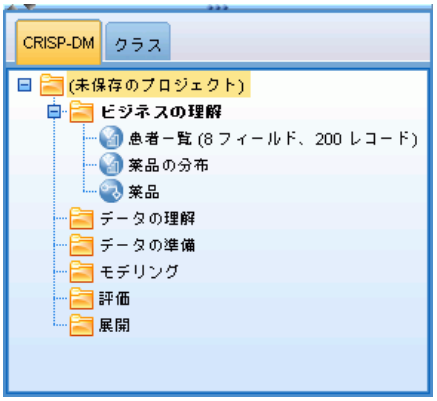
効果的なデータ マイニングをサポートするために、IBM® SPSS® Modeler には CRISP-DM 手法が 2 種類の方法で取り入れています。

- CRISP-DM プロジェクト ツールは、典型的なデータ マイニング プロジェクトのフェーズに関する、ストリーム、出力、および注釈を整理するために役立ちます。プロジェクト中の任意の時点で、ストリームや CRISP-DM フェーズで気づいたことを元にレポートを生成することができます。
- データ マイニング プロジェクトを実施する際のプロセスを案内する、CRISP-DM ヘルプが用意されています。ヘルプには、各ステップの作業リストや、実世界で CRISP-DM がどのように利用されるかの例などが記載されています。CRISP-DM ヘルプを参照するには、メイン ウィンドウの［ヘルプ］メニューから、[CRISP-DM ヘルプ] を選択してください。

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 マイニングへの構造的なアプローチ手段を提供し、プロジェクトの成功を確実なものとするために役立ちます。基本的にこのツールは、標準の IBM® SPSS® Modeler プロジェクト ツールを拡張するものです。実際に、CRISP-DM ビューと標準のクラス ビューを切り替えて、CRISP-DM のタイプやフェーズ別に編成されたストリームや出力を参照することができます。

図 1-1
CRISP-DM プロジェクト ツール



プロジェクト ツールの CRISP-DM ビューを使って、次の作業を行うことができます。

- データ マイニングのフェーズに応じたプロジェクトのストリームや出力の編成
- 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 ツールヒントの作成
- 特定のグラフやモデルから導き出された結論の記録
- プロジェクト チーム用の HTML レポートの生成またはアップデートの配布

CRISP-DM ヘルプ

IBM® SPSS® Modeler には、一般的な CRISP-DM プロセス モデルに関するオンライン ガイドが用意されています。このガイドはプロジェクトのフェーズごとに分かれており、次のような内容が記載されています。

- CRISP-DM の各フェーズの概要と作業リスト
- 様々な指標レポートの生成に関するヘルプ
- CRISP-DM を使ったデータ マイニング作業の進め方の実例
- CRISP-DM の他のリソースへのリンク

CRISP-DM ヘルプを参照するには、メイン ウィンドウの［ヘルプ］メニューから、［CRISP-DM ヘルプ］を選択してください。

他のリソース

IBM® SPSS® Modeler でサポートされている CRISP-DM のほかにも、データ マイニングの作業に関する知識や理解を深めるための様々な手段があります。

- CRISP-DM コンソーシアム Web サイト www.crisp-dm.org (http://www.crisp-dm.org) を参照する。
- CRISP-DM コンソーシアムが作成し、本リリースに付属する CRISP-DM マニュアルを参照する。
- 『Data Mining with Confidence』を参照する (Copyright 2002 by SPSS Inc., ISBN 1-56827-287-1)。

章
2

ビジネスの理解

ビジネスの理解の概要

まだ IBM® SPSS® Modeler で作業を行う前に、データ マイニングで何を見つけだそうとしているのかをあらかじめ調査しておく必要があります。調査には、関係する人々をできるだけ多く集め、打ち合わせた内容をきちんと文書に記録しておくことをお勧めします。CRISP-DM フェーズの最終段階では、ここで集めた情報を元にどのようにプロジェクト計画を作成するかを議論します。

このような調査は必要ないようにも思えますが、そうではありません。データ マイニング作業を実施する業務上の理由を理解しておくことにより、貴重なリソースを消費する前に全員が同じラインに並ぶことができます。

以下の各項目をクリックしてください。

ビジネス目標の決定

まずはじめに、データ マイニングのビジネス上の目標について、できるだけ多くの情報を集めその情報を検討します。この作業は思ったほど簡単ではありません。しかし、作業上の問題、目標、およびリソースなどを明確にすることによって、将来のリスクを最小限に抑えることができます。

CRISP-DM 手法は、この作業を達成するための系統立てられた方法を提供しています。

作業リスト

- 現在のビジネス状況に関する背景情報を収集します。
- 主要責任者が決定した特定のビジネス目標を記録します。
- 業務上の観点からデータ マイニングが成功したかどうかを判断するための基準を策定します。

e-Commerce の例 – ビジネス目標の調査

CRISP-DM を使った Web マイニングのシナリオ

Web 上で e-Commerce による販売に移行する企業が増加するにつれ、既存のコンピュータや電化製品などの e-Commerce 業者は新しいサイトとの競争にさらされています。Web ショップを利用する顧客数の増加速度を上回る Web ショップ数の急激な増加に伴い、各企業は顧客獲得にかける費用の増加にかかわらず収益をあげる手段を模索する必要がでてきました。解決策としては、既存の顧客関係を改善し、現在の企業顧客の価値を高めることが考えられます。

このためには、次のような目標の下に調査を行う必要があります。

- 魅力的なお勧め商品の提示による抱き合わせ販売の機会の増加。
- よりきめ細かいサービスによる顧客ローヤリティの増加。

また、以下の条件を満たした場合に、この試みが成功したと判断します。

- 抱き合わせ販売の 10 % の増加。
- 顧客がサイトを訪れた際の、滞在時間と参照ページ数の増加。
- 予定内の調査完了時間と経費支出。

ビジネス背景情報の収集

組織のビジネス状況を理解することは、次の事柄を把握しておくためにも役立ちます。

- 利用できるリソース（人材や材料）
- 問題点
- 目標

データ マイニング プロジェクトの結果に影響する問いに対して正しい答えを見つけるためには、現在の業務状況について調査する必要があります。

作業 1 – 組織構造の判断

- 組織図を作成し、事業部、部門、およびプロジェクト グループを確認します。このとき、各部門のマネージャ名や役割分担なども忘れないようにしてください。
- 組織内の主要責任者名の確認。
- 金融面の支援やドメインの専門家など、内部のスポンサーを確認します。
- 運営委員会があるかどうかを確認し、メンバー一覧を入手します。

- データ マイニング プロジェクトの影響を受ける業務部門を確認します。

作業 2 ― 問題領域の記述

- マーケティング、顧客サービス、ビジネス開発などの問題の領域を判断します。
- 一般的な用語を使い、問題を記述します。
- プロジェクトの前提条件を明確にします。プロジェクトの背景にある動機を調べます。また、すでに日常業務でデータ マイニングが使われているかなども確認します。
- ビジネス グループ内のデータ マイニング プロジェクトのステータスを確認します。目標が達成されたか、またデータ マイニングをビジネス グループの主要技術として公表する必要があるかどうかを確認します。
- 必要に応じて、データ マイニングに関する情報を組織内にプレゼンテーションしてください。

作業 3 ― 現在の解決法の記述

- 現在、業務上の問題に対処するために利用されている解決法を記述します。
- その解決法の利点と欠点を記述します。また、組織内における解決法の許容度も確認してください。

ビジネス目標の決定

ここで目標を具体的に定めます。調査や打ち合わせの結果として、明確な主目標を定め、プロジェクトのスポンサーや、結果によって影響を受ける業務部門の承認を得る必要がありすま。この目標は、「顧客離れの削減」などのような漠然としたものから、最終的には問題分析の手助けとなる特定の詳細なデータ マイニング目標へと修正されていきます。

作業リスト

後でプロジェクト プランに反映させるために、次の事項を忘れずに記録してください。目標は現実的なものでなければなりません。

- データ マイニングを使って解決したい問題を記述します。
- ビジネス上の疑問点をできるだけ正確に記入します。
- 他のビジネス要件を確認します（既存の顧客を失わずに抱き合わせ販売の機会を増やすなど）。
- 期待する利点や恩恵をビジネス用語で記述します（上得意顧客の顧客離れを 10 %減らすなど）。

ビジネス成功基準

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上の成功基準を定義しておくことが大切です。成功基準は、次の 2 種類のカテゴリに分類できます。

- **客観的：** この基準は、調査精度の増加や顧客離れの現象などの簡単なものにすることもできます。
- **主観的：** 「効果的な治療法群の発見」などの主観的な基準は、明確に定義するのが難しいものですが、誰が最終的な決定をくだすのかを決めておく必要があります。

作業リスト

- できる限り厳密に、このプロジェクトの成功基準を記述します。
- 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。できるならば、各責任者が何を求めているかを書き留めておいてください。

状況の評価

明確な目標を決めたら、次に現在の状況を評価します。このステップには、次のような問いが含まれます。

- どのようなデータを分析に利用できるか？
- プロジェクトを完了するために必要な人員が揃っているか？
- リスクが高い要素は何か？
- それぞれのリスクに対して代替プランを用意しているか？

現在の状況を評価するための詳細計画は、以下のリンク先で説明されています。

e-Commerce の例 － 状況の評価

CRISP-DM を使った Web マイニングのシナリオ

家電製品の e-Commerce 業者は、今回初めて Web マイニングを行います。この会社は、データ マイニングの専門家に支援を依頼しました。専門家が最初に行う作業の 1 つが、データ マイニングを行うために利用できる会社のリソースの調査です。

人員： 当然のごとく、社内にはサーバー ログや製品データベースおよび購入履歴データベースの管理者がいますが、分析を行うためのデータ ウェアハウジングやデータ クリーニングについては不慣れです。そのため、データベースの専門家にも支援を依頼する必要があります。この会社は、分析結果を継続的な Web マイニング プロセスの一環とすることを望んでいるため、今回の作業で決められる作業担当を、今後も常設的に設置するかどうかも検討する必要があります。

データ： この会社はすでに存在している企業のため、分析データの元となるさまざまな Web ログや購入履歴データなどが存在しています。この会社は、初期調査で分析するデータを、すでにサイトで「登録」している顧客だけに限定することを考えています。これが成功したら、対象を拡大していきます。

リスク： 今回の調査に伴う専門家への報酬や従業員の作業時間のほかには、さほど差し迫ったリスクというものはありません。しかし、時間は常に貴重です。そのため、今回の初期プロジェクトは第一四半期にスケジュールされました。

また、現在の時点でキャッシュ フローにさほど余裕はないことから、分析が予算内で完了することが必須条件となります。これらの目標を達成できない可能性があるようならば、プロジェクト範囲を狭めることを提案しなければなりません。

リソースの調査

利用できるリソースを正確に把握しておくことが必要不可欠です。ハードウェア、データ ソース、および人員上の問題を正確に把握しておくことにより、大幅に時間を節約し、潜在的な問題を減らすことができます。

作業 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るか？

作業 2 – データ ソースおよび利用できる情報の確認

- データ マイニングに利用できるデータ ソースは？利用できるデータ ソースのデータの種類やデータ形式を記録してください。
- データはどのように保存されているか？データ ウェアハウスや業務データベースに直接アクセスできるか？
- 地域情報などの外部データの購入を検討しているか？
- 必要なデータへアクセスする際にセキュリティ上の問題があるか？

作業 3 – 人員リソースの確認

- 業務内容やデータに関する専門家がいるか？
- 必要なデータベース管理者や他のサポート要員を用意したか？

これらの問題を確認したら、連絡先やリソースの一覧をフェーズ レポートに記入します。

要件、仮説、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、その後の努力がより報われたものになるでしょう。これらの問題点や検討事項をできる限り明確にしておけば、以降に発生する潜在的な問題を回避することができます。

作業 1 – 要件の決定

最初に策定した基本要件がビジネス上の目標ですが、次の事柄も検討する必要があります。

- データやプロジェクトの結果にセキュリティ上の制約や法的な制約がないか？
- 全員がプロジェクト スケジュール上の要件に適しているか？
- 結果の展開に関する要件はあるか（たとえば、Web 上に公開するとか、スコアをデータベースに記録するなど）？

作業 2 – 仮説の明確化

- プロジェクトに影響する可能性のある経済的な要素はあるか（たとえば、コンサルティング料金や競合製品など）？
- データ品質に関する仮説はあるか？
- 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？

作業 3 – 制約の検証

- データへのアクセスに必要なすべてのパスワードがあるか？
- データの使用における法的な制約をすべて確認したか？
- プロジェクトの予算として財務上の制約をすべて考慮しているか？

リスクと予想される事態

プロジェクト作業中に発生する可能性があるリスクを、あらかじめ検討しておくことをお勧めします。リスクは次の種類に分けられます。

- スケジュール（プロジェクトが予定よりも延びてしまったら？）
- 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ら？）

- 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？）
- 結果（初期の結果が予想していたほどでなかった場合は？）

さまざまなリスクを検討したら、それらのリスクを避けるための代替計画を策定してください。

作業リスト

- 可能性のあるリスクを書き出します。
- それぞれのリスクに対する代替計画を策定します。

用語集

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、わかりにくい技術用語や曖昧な専門用語を明確に定義することも検討する必要があります。たとえば、「顧客離れ」という用語が、業務上の特定で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 グラフの用法を明確化しておくことで、無用なトラブルや誤解を回避できることもあるでしょう。

作業リスト

- チーム メンバーが誤解しやすい、または理解しにくい単語や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入れてください。
- 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かを検討します。

コストと利益の分析

このステップは、「**収益は何ですか？**」の回答になります。最終評価の一環として、プロジェクトにかかる経費と成功時の潜在利益を比較することが重要です。

作業リスト

分析には、次の推定経費を含めます。

- データの収集と使用する外部データ
- 結果の展開
- オペレーティング コスト

次に、以下の利点を考慮します。

- 主目的の達成度

- データの探索で得られた他の洞察
- よりよいデータの理解から得られる利益

データ マイニングの目標の決定

ビジネス目標を明確に決めたら、それをデータ マイニングの観点から記述します。たとえば、ビジネス目標が「顧客離れを減らす」ならば、データ マイニング上の目標は次のようになります。

- 最近の購入履歴データから上得意客（より価値のある顧客）を識別する。
- 利用できる顧客データを使ってモデルを構築し、それぞれの顧客が他社に乗り換える可能性を予測する。
- 他社に乗り換える可能性とその顧客の価値に基づいて各顧客をランク付けする。

これらのデータ マイニングの目標は条件に合えば、ビジネスに適用して上得意客が他社に流れるのを減らすことができます。

このように、ビジネスと技術の両方の連携が、効果的なデータ マイニングには必要です。データ マイニングの目標の決定方法は、目的に応じたヒントを参照してください。

データ マイニングの目標

ビジネスおよびデータ アナリスト達とビジネス上の問題に関する技術的な解決法を作成するためには、さまざまな事柄を明確にしておくことを忘れないようにしてください。

作業リスト

- クラスタ化、予測、または分類などの、データ マイニングの問題の**種類** を記述します。
- 3 ヶ月の有効性を持つ予測など、特定の時間単位を使って技術上の目標を記録します。
- できるならば、既存顧客の 80 % の Churn Score の生成などのように、実際の数値目標を決定します。

e-Commerce の例 – データ マイニングの目標

CRISP-DM を使った Web マイニングのシナリオ

データ マイニングの専門家の支援の下、この業者は会社のビジネス目標をデータ マイニング用語に変換することができました。この四半期で完了する必要がある初期分析の目標は次の通りです。

- 今までの購入履歴情報を使って、「関連する」商品にリンクするモデルを生成する。ユーザーが商品の説明を参照した時に、関連商品グループ中の商品へのリンクを表示する（**マーケット バスケット分析**）。
- Web ログを使って、異なる顧客が何を探そうとしているのかを調べ、それらの商品が目立つようにサイトを設計しなおす。顧客の「タイプ」に応じて、表示するサイトのメイン ページも異なります（**プロファイリング**）。
- Web ログを使って、ユーザーがどこからきて、どのページをよく利用しているかなどから、ユーザーが次にどこへ移動しようとしているかを予測する（**シーケンス分析**）。

データ マイニングの成功基準

データ マイニング作業が順調に進んでいることを確認するために、技術用語における成功を定義する必要もあります。前に作成したデータ マイニングの目標を使って、成功の基準を決定します。IBM® SPSS® Modeler には、結果の精度や妥当性を分析するために役立つ、評価グラフ ノードや精度分析ノードのようなツールが用意されています。

作業リスト

- 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。特定の数値を定義します。
- できる限り主観的な基準を定義し、成功かどうかを判断する責任者を決定します。
- モデル結果の正常な展開が、データ マイニングの成功の一部となるかどうかを検討します。展開に関する計画を策定してください。

プロジェクト計画の策定

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、および決定したビジネス目標とデータ マイニングの目標が、このプロジェクト計画の基盤になります。

以下の手順にしたがって作業を行ってください。

プロジェクト計画の策定

プロジェクト計画は、すべてのデータ マイニング作業の基本文書となります。適切に作成すれば、プロジェクトの最終目標、リソース、リスク、およびスケジュールなどの、データ マイニングのあらゆるフェーズにおけるプロジェクト関連情報を全員に正しく伝えることができます。この計画と、その他収集した文書などは、社内イントラネットに公開することもできます。

作業リスト

計画を作成する際には、以下の事項が明確になっていることを確認してください。

- プロジェクトに関わる人員と、プロジェクト作業や推奨する計画について打ち合わせましたか？
- すべてのフェーズや作業を考慮して時間を見積もりましたか？
- 結果やビジネス 解決法の展開に必要な作業やリソースを考慮していますか？
- 計画中の意志決定を行う時点とレビュー要求を強調していますか？
- モデリングのように、一般的に反復作業が行われるようなフェーズをマークしましたか？

プロジェクト計画の例

この分析調査の計画概要は以下の通りです。

フェーズ	期間	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、技術上の問題
データの準備	5 週間	データ マイニングの専門家、一部データベース アナリストの協力	データの問題、技術上の問題
モデル作成	2 週間	データ マイニングの専門家、一部データベース アナリストの協力	技術上の問題、満足できるモデルを作成できない場合
評価	1 週間	すべてのアナリスト	経済上の変動、結果を導入できない場合
展開	1 週間	データ マイニングの専門家、一部データベース アナリストの協力	経済上の変動、結果を導入できない場合

評価ツールと評価技術

データ マイニングを行うためのツールとしてすでに IBM® SPSS® Modeler を選んでいるため、このステップでは、ビジネス上のニーズにあわせてどのデータ マイニング技術が最適かを調査することができます。SPSS Modeler は、データ マイニングの各フェーズに対応する幅広いツールを提供しています。さまざまな技術をいつ使用するかは、オンライン ヘルプのモデリングに関する説明を参照してください。

次の手順に進む用意ができましたか？

データを探索して IBM® SPSS® Modeler での作業を開始する前に、以下の事項を明確にしてください。

ビジネス上の観点

- このプロジェクトで期待しているビジネス上の結果は？
- 作業が成功したと判断する基準は？
- 目標に到達するために必要な予算やリソースがありますか？
- このプロジェクトに必要なすべてのデータにアクセスできますか？
- このプロジェクトに関するリスクや偶発事項について話し合いましたか？
- 費用と利益分析の結果、このプロジェクトに実施価値はありますか？

以上の事柄を明確にしたら、それらをデータ マイニングの最終目標に変換できますか？

データ マイニングの観点

- データ マイニングがどの程度、ビジネス目標を達成するための手助けになりますか？
- どのデータ マイニング技術が最良の結果を導き出せるかを判断できますか？
- どのようにして結果が正確または十分効果的だと判断しますか？(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？)
- モデル作成結果をどのように展開しますか？プロジェクト計画中に展開計画も入れていますか？
- 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？
- リスクと依存関係が計画中で明確にされていますか？

これらの質問の答えが「はい」の場合、データを実際に分析する準備は完了しています。

章
3

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、データ マイニングに利用できるデータを詳しく調査します。一般的にプロジェクトの中でもっとも時間がかかる、次段階のデータの準備フェーズで予期しない問題が発生するのを避けるためにも、このステップでの作業が重要になります。

データの理解作業には、データへのアクセスと、テーブルや画像を使った探索が含まれています。これらは、IBM® SPSS® Modeler で CRISP-DM プロジェクト ツールを使って編成することができます。この作業によって、データの品質を判断し、調査結果をプロジェクト文書に記入します。

以下の手順にしたがって作業を行ってください。

初期データの収集

CRISP-DM 中のこの段階では、データにアクセスしてそれを IBM® SPSS® Modeler に取り込む準備ができています。データはさまざまなソースから取得されます。次に例を示します。

- **既存データ:** トランザクション データ、調査データ、Web ログなどさまざまなデータが含まれます。既存データで十分かどうかを検討してください。
- **購入履歴データ:** 地域統計情報などの補助データを使っていますか？そうでない場合は、必要になるかどうかを検討してください。
- **その他のデータ:** 以上のデータ ソースで十分でない場合は、調査を行うか、または既存のデータを補足する他の情報を取得する必要があるかもしれません。

作業リスト

SPSS Modeler 中のデータを調べて、次の事柄を確認してください。調査結果などの情報は、きちんと記録するようにしてください。詳細は、p. 16 データ収集レポートの作成 を参照してください。

- データベース中のどの属性（列）が役に立つか？
- どの属性が調査に無関係で除外することができるか？
- 総括的な結論を導き出したり、正確な予測を行うため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？
- さまざまなデータ ソースからのデータを結合しますか？結合する場合、結合時に問題になりそうな領域がないですか？
- それぞれのデータ ソース中で欠損値をどのように処理するかを確認しましたか？

e-Commerce 業者の例 – 初期のデータ収集

CRISP-DM を使った Web マイニングのシナリオ

この例の業者は、次のような複数の重要なデータ ソースを使用しています。

Web ログ: この生のアクセス ログには、顧客が Web サイト内をどのように移動、参照したかなどの情報が含まれています。Web ログ中のイメージ ファイルへの参照や他の情報にならないエントリは、データの準備作業の一貫として除去する必要があります。

購入履歴データ: 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。購入履歴データベース中の注文は、対応する Web ログ中のセッションにマップする必要があります。

商品データベース: 商品属性は、「関連」商品を判断する際に役立ちます。商品情報は、対応する注文と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、サイトに登録した顧客から収集した情報が含まれています。このレコードは完全なものとは限りません。顧客が一部のデータを入力していない可能性もあります。顧客情報は、対応する購入履歴および Web ログ中のセッションにマップする必要があります。

現在の所、アナリストはデータの処理作業で多忙なため、外部データベースを購入したり、調査を行うために余分な予算をかける予定はありません。しかし、データ マイニングの結果からさらに詳細に調査、展開するような場合には、登録していない顧客に対する地域統計情報を購入することが有益なこともあるでしょう。また、地域統計情報を利用して、この業者の顧客ベースが他の Web 店舗の顧客とどのように違うかを確認することも役に立つことがあります。

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って、データ収集レポートを作成することができます。レポートを作成したら、それをプロジェクト用 Web サイトに公開したり、チームの各メンバーに配布することができます。この

レポートは、次のステップで用意するデータ詳細、データ探索、およびデータ品質検査レポートとまとめることもできます。これらのレポートは、データ準備フェーズ全体で大変役立ちます。

データの記述

データを記述するにはさまざまな方法がありますが、大部分の記述はどれくらいのデータを利用でき、データの状態はどうかという、データの量や質を重視します。データを記述する際に考慮する必要がある事柄を次に示します。

- **データ量**: たいていのモデリング技術では、データ サイズによって犠牲になるものがあります。大きいデータ セットはより正確なモデルを作成できますが、処理時間が長引いてしまいます。データの一部（サブセット）だけを利用できるかどうかを検討してください。最終レポートの作成時には、すべてのデータ セットのデータ サイズを忘れずに記入してください。また、データの記述時には、レコード数とフィールド数（属性）の両方を考慮することを忘れないでください。
- **値の種類**:データは、**数値**、**カテゴリ**（文字列）、または**ブール値**（真または偽）などのさまざまな種類で記録されています。値の種類に注意することにより、後で行うモデリング作業で発生する潜在的な問題を回避することができます。
- **コーディング スキーマ**: データベース中の値は、しばしば性別や商品タイプなどの特徴を表すことがあります。たとえば、あるデータ セットは男性および女性を表すために文字 M および F を使用しているのに、別のデータベースではこれを数値 1 および 2 で表されています。データレポート中のスキーマの矛盾に注意するようにしてください。

このような前提知識を元に、データ詳細レポートを作成し、それを他のメンバーと共有していきます。

e-Commerce 業者の例 – データの記述

CRISP-DM を使った Web マイニングのシナリオ

多数のレコードや属性が、Web マイニング アプリケーションで処理されます。このデータ マイニング プロジェクトで処理するデータを、Web サイトから登録した約 30,000 人の顧客に限定したとしても、それでもなお Web ログには何百万というレコードがあります。

日時、アクセスした Web ページ、および登録時の多値選択型の質問など、これらのデータ ソース中の大半の値の種類はシンボル値です。一部の変数は、Web ページに訪れた数および Web サイトで費やした時間などの、新たな数値型変数を作成するために用いられます。また、データ ソース中の数

少ない既存の数値変数には、各商品の注文数、購入までに要した時間、および商品データベースからの商品の重みや次元の指定などが含まれています。

データ ソースによって属性はかなり異なるため、さまざまなデータ ソースのコーディング スキーマ間には、ほとんど重複はありません。重複する変数は顧客 ID や商品コードのような、「キー」となる変数だけになります。これらの変数には、データ ソース間で同一のコーディング スキーマがなければなりません。そうでないと、データ ソースを結合することができません。結合するこれらのキー フィールドを記録するためには、それに応じたデータ準備作業を行う必要があります。

データ詳細レポートの作成

データ マイニング プロジェクトを効率的に進めるために、次のメトリックを使って正確なデータ詳細レポートを生成する値を検討していきます。

データ量

- データ フォーマットは？
- データの取得方法を確認します（例 ： ODBC)。
- データベースの大きさは（行数および列数）？

データ品質

- データにビジネス上の問いに関する特性が含まれていますか？
- 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量を算出しましたか？これが、ビジネス上の問いにどのような洞察を提供しますか？
- 関連属性を優先することができますか？できない場合、ビジネス アナリストがさらなる洞察を行えますか？

データの探索

この CRISP-DM フェーズでは、IBM® SPSS® Modeler で利用できるテーブル、グラフ、および他の視覚化ツールを使ってデータを探索します。このような分析は、ビジネスの理解フェーズで策定したデータ マイニングの最終目標を達成するために役立ちます。また分析内容は、前提条件を作成し、データ準備フェーズで行われるデータ変換作業を支援するためにも役立ちます。

e-Commerce 業者の例 – データの調査

CRISP-DM を使った Web マイニングのシナリオ

CRISP-DM ではこの時点で初期探索を行うことを指示していますが、Web ログにある生データの調査は不可能とは言わないまでも困難な作業です。一般的に、有益な探索を行えるデータを生成するには、データの準備フェーズの最初に Web ログ データを処理しておかなければなりません。このことは、プロセスは特定のデータ マイニングのニーズに合わせてカスタマイズできるし、またしなければならないことを意味しています。CRISP-DM は循環作業で、通常データ マイニングを行う際には各フェーズ間を前後に移動しながら作業が行われます。

探索を行う前に Web ログを処理する必要がありますが、この業者が持つ他のデータ ソースはより調査しやすいものとなっています。購入履歴データベースを探索に使用することによって、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の顧客に関する興味深い情報が得られます。顧客データベースの要約からは、登録時の質問によって得られた商品に対する反応の分布が得られます。

探索はデータ中のエラーを探すためにも役立ちます。大半のデータ ソースは自動的に生成されますが、商品データベース中の情報は手作業で入力されます。記載される商品次元の要約リストを作成することにより、「119インチ モニタ」（19インチが正しい）のような入力ミスを発見するのが容易になります。

データ探索レポートの作成

グラフを作成し、利用可能なデータの統計情報を作成したら、このデータから技術上およびビジネス上の目標に対する回答をどのようにして導き出せるかという仮説を作成します。

作業リスト

データ探索レポートには、ここで発見した内容を正しく記録してください。次の事項を確認してください。

- データに関するどのような仮説を作成しましたか？
- さらに分析を行う場合、どの属性が役立ちそうですか？
- 探索結果によりデータの新たな特性が表れましたか？
- 探索の結果、最初の仮説がどのように変わりましたか？

- 後で使用するデータのサブセットを識別できましたか？
- 以上の結果から、データ マイニングの最終目標をもう一度見直してください。この探索により、最終目標が変わりましたか？

データ品質の検証

データが完全であることは滅多にありません。実際に、大部分のデータには、分析を妨げるようなコーディング エラー、欠損値、または他の不整合などが含まれています。このような潜在的な問題を回避する 1 つの方法として、モデルを作成する前に、利用できるデータの完全な品質分析を行うことがあげられます。

IBM® SPSS® Modeler のレポート ツール（データ検査ノード、テーブル ノード、その他出力ノードなど）を使用して、次のような問題を検出できます。

- **欠損データ**は空白または応答なしと定義された値を含みます（$null$、?、または 999 など）。
- **データ エラー**は、一般的にデータ入力時に生じた入力ミスです。
- **測定エラー**は、正しく入力されたけれども不正な測定スキーマに基づいているデータを含みます。
- **コーディングの不整合**は、一般的に標準的ではない測定単位や値の不整合を含みます（性別に M と male の両方を使用するなど）。
- **不良メタデータ**は、フィールドの明確な意味と、フィールド名やフィールド定義に記述されている意味の不整合が含まれています。

品質上の問題や情報を忘れずに記録してください。詳細は、 p.21 データ品質レポートの作成 を参照してください。

e-Commerce 業者の例 – データ品質の検証

CRISP-DM を使った Web マイニングのシナリオ

データ品質の検証は、しばしばデータの記述および探索の過程で行われます。この例で見つかった問題には、次のようなものがありました。

欠損データ： 確認された欠損データとしては、登録顧客の一部に、未回答の項目がありました。特に足りない情報を追加しなければ、このような顧客は以降のモデリング段階で無視される可能性があります。

データ エラー： 大半のデータ ソースは自動的に生成されるため、さほど心配することはありません。商品データベース中のエラーは、探索フェーズで見つけることができます。

測定エラー 測定エラーが存在する可能性がもっとも大きいソースが、顧客へのアンケートです。質問内容が間違っていたり、説明が足りないと、業者が望む情報を得られない可能性があります。ここで繰り返しますが、探索の過程において、おかしい回答があった項目には、特に注意する必要があります。

データ品質レポートの作成

データ品質の探索および検証作業に基づいて、次の CRISP-DM フェーズに進むためのレポートを作成します。 詳細は、 p.20 データ品質の検証 を参照してください。

作業リスト

前に述べたように、データ品質に関しては、さまざまな種類の問題があります。次のステップに進む前に、次の観点から品質を検討し、対処方法を検討してください。データ品質レポート中のすべての回答を記録してください。

- 欠けている属性や空白フィールドを識別しましたか？その場合、その欠損値の背景には何か特別な意味がありますか？
- 後の結合や変換作業で問題になるようなスペリング上の不整合はありましたか？
- 偏差を調べて、それがノイズかまたは分析する価値がある現象かを確認しましたか？
- 値の妥当性を確認しましたか？明確な矛盾が見つかった場合は、それを記録してください（10 代なのに高収入など）。
- 仮説に影響しないデータの除外を検討しましたか？
- データはフラット ファイルに格納されていますか？その場合、各ファイル間で区切り文字は統一されていますか？各レコードのフィールド数は同じですか？

次の手順に進む用意ができましたか？

IBM® SPSS® Modeler でモデルを作成するデータの準備を開始する前に、次の事項を確認してください。

データをどの程度まで理解できていますか？

- すべてのデータ ソースが明確に識別、評価されましたか？気づいている問題や制約がありますか？
- 利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、仮説を作成するために役立ちますか？

- 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することができますか？
- 注目する各属性の基本統計量を算出しましたか？意味のある情報が表れましたか？
- 探索グラフィックを使ってキー属性を詳細に洞察しましたか？その洞察により、仮説を修正しましたか？
- このプロジェクトにおけるデータ品質上の問題は何ですか？これらの問題への対処方法は考えていますか？
- データの準備作業は明確になりましたか？たとえば、どのデータ ソースを結合して、どの属性を除外または選択するかが明確になりましたか？

これでビジネスおよびデータの両方を理解できました。次は、SPSS Modeler を使ってモデリング用のデータを準備していきます。

章
4

データの準備

データの準備の概要

データの準備フェーズは、データ マイニングにおける最も重要な作業で、しばしば時間がかかります。実際、一般的なデータの準備作業は、プロジェクトに費やされる時間や労力の 50～70% を占めると見積もられています。初期のビジネスの理解およびデータの理解フェーズに十分な労力を費やしていれば、このオーバーヘッドを最小限に抑えることができます。しかし、それでもデータの準備と事前処理にはかなりの時間を割く必要があります。

一般的にデータの準備フェーズでは、組織の業務内容や目標に応じて次の作業が行われます。

- データ セットやレコードの結合
- サンプルのデータ サブセットの選択
- レコードの集計
- 新しい属性の作成
- モデリングのためのデータのソート
- 空白や欠損値の削除または置換
- トレーニングへの分割とデータ セットのテスト

詳細は、以下のデータ準備作業から適切な項目を選択してください。

データの選択

前段階の CRISP-DM フェーズで行われた初期データの収集フェーズをベースに、データ マイニングの最終目標を達成するために必要なデータの選択を開始します。通常、データの選択には次の 2 種類の方法があります。

- **項目の選択 (行):** どのアカウント、商品、または顧客などを含めるかなどを決定します。
- **属性またはその特性の選択 (列):** 総トランザクション量や世帯あたりの収入などの、特性の使用に関する決定を行います。

e-Commerce 業者の例 – データの選択

CRISP-DM を使った Web マイニングのシナリオ

どのデータを選択するかという e-Commerce 業者の決定の大半は、初期のデータ マイニング フェーズで行われています。

項目の選択: 初期の分析調査は、サイトに登録した約 30,000 人の顧客に限定します。そのため、購入履歴ログや Web ログから、未登録の顧客を削除するようにフィルタを設定する必要があります。また、Web ログから、画像ファイルへの呼び出しや、他の情報にならないエントリを削除するフィルタも設定する必要があります。

属性の選択: 購入履歴データベースには、顧客に関する機密情報が含まれています。そのため、顧客名、住所、電話番号、およびクレジット カードの番号などの属性をフィルタリングする必要があります。

データの取り込みと除外

データのどの部分を取り込んでどの部分を除外するかを決めたら、それらの決定の根拠となった理由を忘れずに記録してください。

検討事項

- 選択した属性はデータ マイニングの最終目標を達成するために適切なものか？
- 特定のデータ セットや属性の品質により、結果の妥当性が損なわれないか？
- そのようなデータを修復できるか？
- 性別や人種などの特定のフィールドを使用することに関して制約はないか？

ここで行った決定内容が、データの理解フェーズで作成した前提と矛盾しないか？矛盾していたり異なっているような場合は、プロジェクト レポートに理由を明確に記録してください。

データのクリーニング

データのクリーニングでは、分析に利用するデータをより詳細に調査して、問題がないかどうかを確認します。IBM® SPSS® Modeler のレコードおよびフィールド設定ノードを使ったデータのクリーニングには、さまざまな方法があります。

データの問題	対処方法
欠損データ	行またはその特性を除外する。または、空白を予測フィール
データ エラー	ロジックを使って手作業でエラーを発見し修正する。または外する。
コードの不整合	単一のコーディング スキーマ（コード体系）を決定し、値を
メタデータの欠損または不良	問題のあるフィールドを探しだし、正しい内容に修正する。

データの理解フェーズで用意されたデータ品質レポートには、そのデータ特有の問題の種類が記載されています。これを使って SPSS Modeler でデータ操作を開始することができます。

e-Commerce 業者の例 – データのクリーニング

CRISP-DM を使った 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のクリーニングを行います。

欠損データ： オンラインの質問に回答していない顧客は、以降のモデルにおいて無視する必要がある場合があります。これらの顧客に対して未回答の部分を問い合わせることもできますが、これには時間や費用がかかるため容認できるものではありません。代わりに、質問に回答した顧客と回答しなかった顧客の購入内容の差異をモデル化することができます。これらの 2 種類の顧客群に同じような購入傾向がある場合、質問に対する回答の欠損は、さほど問題にはなりません。

データ エラー： 探索段階で見つかったエラーは、ここで修正することができます。しかし、たいていの場合は、顧客が回答をバックエンド データベースに送信する前に、Web サイトでデータ エントリが適切かどうか確認されます。

測定エラー： 質問に対して回答が曖昧だったり不十分な場合、それがデータの品質に大きく影響してしまいます。回答が欠けている場合のように、回答を問い合わせる時間や費用の面からも、これは難しい問題です。問題のある項目に対しては、選択プロセスに戻りこれらの項目をフィルタリングすることが最良の対処方法かもしれません。

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、データをどのように修正したかを確認、追跡するためにも必要不可欠です。今までに行った作業の詳細を記録しておくことにより、将来のデータ マイニング プロジェクトを行う際に非常に参考になることでしょう。

作業リスト

レポートを作成する際には、次の事項を明確にしておくことをお勧めします。

- データ中にどのようなノイズがあったか？
- そのノイズを除去するために、どのような手段を使ったか？どの手段が有効だったか？
- 修復できないケースや属性があったか？ノイズのために除外されたデータを記録しておくことも忘れないようにしてください。

新規データの作成

新しくデータを作成する必要があるような場合もよくあります。たとえば、各トランザクションに対して購入品の保証延長サービスを表すフラグを示す、新しい新規列を作成することができます。IBM® SPSS® Modeler のフラグ設定ノードを使用すれば、この新規フィールド、保証延長サービスを簡単に作成することができます。

新規データを作成するには、次の 2 種類の方法があります。

- 属性から取得（列や特性）
- レコードの生成（行）

SPSS Modeler では、レコードおよびフィールド設定ノードを使ってデータを作成するための、さまざまな手段が提供されています。

e-Commerce 業者の例 – データの作成

CRISP-DM を使った Web マイニングのシナリオ

Web ログを処理することにより、さまざまな新規属性を作成することができます。ログに記録されているイベントに対して、タイムスタンプの作成、訪問者やセッションの識別、およびアクセスされたページや操作の記録などの作業を行えます。これらの変数を使って、セッション内のイベント間の時間などさらなる属性を作成することもできます。

また、データの結合や再構成により新しい属性を作成することもできます。たとえば、行あたりのイベント Web ログが「ロール アップ」され、各行がセッションになると、総アクション数、総時間、およびそのセッション内での購入数などの属性を作成することができます。Web ログを顧客データベースとマージして、各行が顧客に対応するようにすれば、各顧客別にセッション数、総アクション数、総時間、および購入数などを作成することができます。

新しいデータを作成したら、探索作業を行って、データの作成が正しく行われているかどうかを確認します。

属性の取得

IBM® SPSS® Modeler では、次のフィールド設定ノードを使って新しい属性を作成することができます。

- **フィールド作成ノード**を使って、既存のフィールドから新規フィールドを作成する。
- **フラグ設定ノード**を使って、フラグ フィールドを作成する。

作業リスト

- 属性を取得する際には、モデリングのデータ要件を検討します。たとえば、モデリング アルゴリズムが数値などの特定のデータ型を必要とするかどうかを確認します。該当する場合は、必要に応じて適切なデータ型変換を行います。
- モデリングの前にデータを正規化する必要があるか？
- 欠けている属性を、集計、平均化、または算出などの方法で作成できるか？
- 事実背景を元に、既存のフィールドから取得できる重要な要素（Web サイトで費やした時間など）があるか？

データの統合

同じようなビジネス上の問いに対して、複数のデータ ソースがあるのは、さほど珍しいことではありません。たとえば、同じ顧客群に対して住宅ローン データと購入人口統計データの両方にアクセスすることもあります。これらのデータ セットに同じ固有の識別子（住民基本台帳番号など）がある場合、そのキー フィールドを使って IBM® SPSS® Modeler に結合することができます。

次の 2 種類の基本的な、データ統合方法があります。

- データの**結合**は、同じレコードがあるけれども属性が異なる 2 種類のデータ セットを結合します。データは同じキー識別子を使って結合されます（顧客 ID など）。結合の結果できるデータでは、列や特性が増加します。
- データの**追加**は、同じ属性があるけれどもレコードが異なる複数のデータ セットを結合します。データは、同じフィールドに基づいて統合されます（製品名や契約期間など）。

データの統合の詳細は、以下のリンクを参照してください。

e-Commerce 業者の例 – データの統合

CRISP-DM を使った Web マイニングのシナリオ

複数のデータ ソースがある場合、これらのデータを統合するにはさまざまな方法があります。

- **顧客および商品属性をイベント データに追加する：** 他のデータベースからの属性を使って Web ログ イベントのモデルを作成するには、各イベントに関連付けられている顧客 ID、商品番号、および商品購入番号を正しく識別し、対応する属性を処理された Web ログに結合する必要があります。結合されたファイルは、顧客や製品がイベントに関連付けられるたびに、顧客および製品情報を複製することに注意してください。
- **購入履歴ログおよび Web ログ情報を顧客データに追加する：** 顧客価値のモデルを作成するには、適切なデータベースから顧客の購入情報およびセッション情報を取得して合計し、それを顧客データベースと結合する必要があります。これには、データの構築作業で述べたように、新規属性の作成も含まれます。

データベースを統合したら、探索作業を行って、データの結合が正しく行われているかどうかを確認します。

統合作業

データの開発と理解に十分な時間を費やさないと、データの統合作業が複雑なものとなる可能性があります。データ マイニングの最終目標を達成するために重要と思われるデータの項目や属性を十分に調査してから、データの統合作業を開始してください。

作業リスト

- IBM® SPSS® Modeler のレコード結合ノードまたはレコード追加ノードを利用して、モデリングに役立つデータ セットを統合します。
- モデルの作成を開始する前に、結果となる出力を保存してください。

- 結合が完了したら、値を**集計**してデータを簡素化することができます。集計とは、複数のレコードやテーブルの情報を要約し、新しい値を算出することです。
- 新規レコードを生成する必要があることもあります（複数年の納税申告の平均控除など）。

データのフォーマット

モデリング前の最終作業として、ある技術を利用する場合に、特定の形式や並びを持つデータが必要になるかどうかを確認することをお勧めします。たとえば、シーケンス アルゴリズムを利用する場合、モデルを作成する前にデータをあらかじめソートする必要があることも珍しくはありません。モデルで自動的にソートが行われる場合でも、あらかじめソート ノードを使用することにより、処理時間を節約することができます。

作業リスト

データのフォーマットを行う際には、次の事項を明確にしてください。

- 使用するモデルは？
- 使用するモデルは特定のデータ形式や並び順を必要としていますか？

データを変更する必要がある場合、必要なデータ操作を行うために IBM® SPSS® Modeler の処理ツールが役立ちます。

モデルの作成準備ができましたか？

IBM® SPSS® Modeler でモデリングを開始する前に、次の事項を確認してください。

- すべてのデータを SPSS Modeler から利用できますか？
- データの初期探索と理解に基づいて、関連するデータのサブセットを選択できましたか？
- データの事前処理を適切に行い、修復不可能な項目を削除しましたか？判断事項、または決定事項はすべて最終レポートに記載してください。
- 複数のデータ セットを正しく統合できましたか？記載する必要がある結合上の問題がありましたか？
- 使用するモデル作成ツールの要件を調べましたか？
- モデルの作成前に対処できるフォーマット上の問題はありましたか？これには、データのフォーマットが必要な場合だけでなく、フォーマットによってモデリング時間を短縮できる場合も含まれます。

これらの事項がすべて大丈夫ならば、次にデータ マイニングでもっとも重要な作業であるモデリング作業に進みます。

章
5

モデル作成

モデル作成の概要

ここで、今までの労力が報われるかどうかが決まります。準備に時間を費やして IBM® SPSS® Modeler 分析ツールにデータを取り込み、そこから得られた結果により、ビジネスの理解フェーズで提起された問題を解決するための手がかりが得られました。

一般的にモデリングでは、繰り返し反復作業が行われます。データ マイニングの担当者は、デフォルトのパラメータを使ってさまざまなモデルを実行し、次にパラメータを細かく調整したり、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。単一のモデルや、1 回モデルを実行しただけで満足する回答が得られることは滅多にありません。これがデータ マイニング作業を興味深くしているもので、与えられた問題に対してさまざまな解決方法が存在し、SPSS Modeler はその解答を導き出すために役立つさまざまなツールが用意されています。

モデリングを開始するには、次の各ステップを参照してください。

モデリング手法の選択

どのモデルが組織ニーズに適しているかをすでにお考えかもしれませんが、ここではどのモデルを使用するのが適切かをより明確に決定していきます。適切なモデルの決定にあたっては、通常次のような事柄を検討していきます。

- **データ マイニングに利用できるデータ型：** たとえば、注目するフィールドはカテゴリ（シンボル値）ですか？
- **データ マイニングの最終目標：** 単にトランザクション データを分析調査して、興味深い購入パターンを発見したいだけですか？それとも、たとえば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：** モデルには、特定のデータ サイズやデータ型が必要ですか？簡単に結果を得られるモデルが必要ですか？

IBM® SPSS® Modeler で利用できるモデルの種類とその要件の詳細は、SPSS Modeler のマニュアルまたはオンライン ヘルプを参照してください。

e-Commerce 業者の例 – モデリング手法

この業者が採用するモデリング手法は、業者が設定したデータ マイニングの最終目標によって決まります。

お勧めの改善： 注文情報をクラスタ化して、どの製品が一緒に購入されるかを判断します。より優れた結果を得るために、顧客データやサイト訪問レコードなどを追加することもできます。このようなモデルを作成するには、TwoStep クラスタや Kohonen ネットワーク クラスタ手法が適しています。後ほど、C5.0 ルールセットを使ってクラスタをプロファイリングして、顧客の訪問時にどのような商品を勧めるのが最適かを判断することができます。

サイト ナビゲーションの改善： ここで、頻繁に使われるページを調べた所、顧客がそのページにたどり着くまでには何度もクリックを繰り返さなければならないことがわかりました。このことから、Web ログにシーケンス アルゴリズムを適用して顧客が Web サイトを移動するための「固有なパス」を生成し、次に頻繁に訪問されるけれども何も行われないページを持つセッションに注目します。後で、より詳細な分析を行う際に、クラスタリング技術を使って異なる訪問や訪問者の「種類」を識別し、その種類に応じてサイトの内容を編成したり、表示することもできます。

適切なモデリング手法の選択

IBM® SPSS® Modeler では、さまざまなモデリング手法を利用することができます。問題をさまざまな方向から検討するために、しばしば複数のモデルが利用されます。

作業リスト

使用するモデルを決める際には、選択に影響する次の事柄に注意してください。

- 利用するモデルでデータをテスト セットと学習セットに分割する必要がありますか？
- モデルに対して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？
- モデルに一定レベルのデータ品質が必要ですか？現在のデータはこの品質レベルを満たしていますか？
- データは特定のモデルに適切なタイプですか？そうでない場合、データ操作ノードを使って適切な形式に変換することができますか？

SPSS Modeler で利用できるモデルの種類とその要件の詳細は、SPSS Modeler のマニュアルまたはオンライン ヘルプを参照してください。

モデリングの仮説

モデリング ツールを選択、決定する際には、決定の過程を記録するようにしてください。データの仮説やモデルの要件に適合させるために行ったデータ操作なども、すべて記録してください。

たとえば、ロジスティック回帰およびニューラル ノードは両方とも、実行する前にデータ型を完全に**インスタンス化**（データ型を既知に）する必要があります。つまり、ストリームにデータ型ノードを追加して実行し、モデルの構築と実行の**??**にデータをインスタンス化しておく必要があります。同様に、C5.0 のような予測モデルで稀なイベントを予測するルールを見つける場合、データの再バランス化が有効になることもあります。この種類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、モデルによりバランス化されたサブセットを適用する方がしばしば良い結果を得られます。

これらの決定を行った場合には、忘れずに文書に記録してください。

テスト設計の生成

実際にモデルを作成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、次の 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
- 基準をテストするデータの定義

モデルの**適合度**は、さまざまな方法で測定できます。C5.0、GRI、および C&R Tree などの監視モデルでは、適合度の測定は特定のモデルの誤り率を推定することによって行われます。Kohonen クラスタ ネットのような非監視モデルでは、測定には解釈しやすさ、展開、または必要な処理時間などの基準が含まれることもあります。

モデリングは反復するプロセスだということを忘れないようにしてください。通常は複数のモデルの結果をテストしてから、使用して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の記述です。モデリングは反復的なプロセスのため、どのようなタイミングでパラメータの調整を中止して、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。

作業リスト

テスト設計を作成する際には、次の事柄を確認してください。

- 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アプローチ手法です。)
- 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 ネットなど）の適合度をどのように測定しますか？
- モデルの設定調整と実行を何回繰り返した後に、他の種類のモデルを試してみますか？

e-Commerce 業者の例 – テスト設計

CRISP-DM を使った Web マイニングのシナリオ

モデルの評価基準は、検討しているモデルとデータ マイニングの最終目標によって異なります。

お勧めの改善： 実際の顧客に対して改善されたお勧め商品が表示されるまでは、それらを純粋かつ客観的に評価する方法はありません。しかし、ビジネス上の観点から簡単かつ十分な意味を持つお勧め商品を生成するルールを作成することもできます。同様に、顧客やセッションごとに別のお勧め商品を生成するような複雑なルールを利用することもできます。

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を参照しているかという情報に基づいて、重要なページにアクセスしやすいようにするなど、サイト デザインの変更を客観的に評価することができます。しかし、お勧め商品の場合と同様に、顧客が設計し直したサイトにどのような印象を持つかをあらかじめ評価することは困難です。時間と資金に余裕があるならば、有用性テストを実施するのが望ましいこともあります。

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リングの準備も完了しているはずです。最終的な結論を出す前に、複数のモデルを試してみてください。たいていの場合、データ マイニング作業では複数のモデルを作成し、その結果を比較した上でモデルの展開や統合が行われます。

さまざまなモデルの進行状況を追跡するために、各モデルで使用した設定やデータを記録しておくことを忘れないようにしてください。このことは、結果を他のメンバーと協議し、必要に応じてステップを繰り返すため

に役立ちます。最終的にモデリング プロセスでは、データ マイニングの決定に使用する 3 種類の情報が得られます。

- 今までに試した最良の結果を得られる**パラメータ設定**。
- 実際に生成した**モデル**。
- **モデルの結果の詳細**、これには、モデルの実行および結果の調査時に発見されたパフォーマンスおよびデータ上の問題が含まれています。

e-Commerce 業者の例 – モデリング

CRISP-DM を使った Web マイニングのシナリオ

お勧めの改善： 購入履歴データベースから始まり、次に関連する顧客およびセッション情報を取り入れて、さまざまなレベルのデータ統合のためのクラスタが作成されました。各レベルの統合に対して、TwoStep や 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設定でクラスタが作成されました。これらのクラスタリングでは、異なるパラメータ設定を持ついくつかの C5.0 ルールセットが生成されました。

サイト ナビゲーションの改善： シーケンス モデル ノードを使って顧客パスが生成されます。アルゴリズムで、もっとも一般的な顧客パスに焦点をあてるために役立つ最小範囲の基準を指定できます。このパラメータに対して、さまざまな設定が試行されました。

パラメータの設定

大半のモデリング手法には、モデリング処理を制御するために調整できるさまざまなパラメータや設定が用意されています。たとえば、ディシジョン ツリーは、ツリーの深さ、分岐、および他のさまざまな設定を調整することができます。たいていの場合、最初はデフォルトのオプションを使ってモデルを作成し、次に以降のセッション中でパラメータを順次調整していきます。

もっとも正確な結果を得られるパラメータが決まったら、ストリームや生成したモデル ノードを保存してください。また、ノードの設定を最適化することは、自動化する時期の決定や、新しいデータによるモデルの再構築にも役立ちます。

モデルの実行

IBM® SPSS® Modeler では、モデルを実行するのは簡単な作業です。モデル ノードをストリームに挿入してパラメータを調整し、モデルを実行すれば結果を生成、表示することができます。結果は、作業領域の右側にある、

[モデル] タブに表示されます。生成されたモデルを右クリックして、結果を参照することができます。大部分のモデルでは、生成したモデルをストリームに挿入して、さらに詳細な評価や結果の展開を行うことができます。モデルを SPSS Modeler で保存して、後ほど再利用することもできます。

モデルの説明

モデルの結果を調査する場合、そのモデル関する有用な情報をきちんと書き留めるようにしてください。これらの情報は、ノードの注釈ダイアログやプロジェクト ツールを使って記録したり、それを保存することもできます。

作業リスト

各モデルに対して、次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ますか？
- モデルにより新しい洞察が行われたり、見慣れないパターンが表れましたか？
- モデルの実行時に問題がありましたか？処理時間は妥当ですか？
- モデルに欠損値が多いなどのデータ品質上の問題がありますか？
- 特に明記しておくべき計算の不整合はありますか？

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。ここで、モデルを詳細に調べて、目的通りの精度があり効果的かを判断していきます。ここで、「目的通り」という言葉は、「展開可能」、「興味深いパターンを発見できる」のような、さまざまな事柄を意味しています。さきほど作成したテスト設計を調べることは、このモデルを組織上の観点から評価するために役立ちます。

総合的なモデル評価

検討中の各モデルに対して、テスト計画時に策定した基準に基づいて秩序だった評価を行うことをお勧めします。ここで、生成したモデルをストリームに追加し、評価グラフ ノードや精度分析ノードを使って、結果の効率性を分析します。また、結果が論理的に意味を為すかどうか、または最終的なビジネス目標に対して結果が簡単すぎないかどうかも検討する必要があります（たとえば、ワイン > ワイン > ワインのような購入パターンを表すシーケンス）。

評価を行ったら、客観的基準（モデルの精度）および主観的基準（使いやすさや結果の解釈）の両面から、モデルを順番にランク付けしていきます。

作業リスト

- 評価グラフ ノード、精度分析ノード、および交差検証グラフなどの IBM® SPSS® Modeler データ マイニング ツールを使用して、モデルの結果を評価します。
- ビジネス上の問題の理解という観点から、結果のレビューを行います。特定の結果に関連する洞察を行ったデータ アナリストや他の専門家達にも助言を求めてください。
- 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。結果を Web 上に展開する必要がありますか？または、データ ウェアハウスに戻す必要がありますか？
- 成功基準に対する結果の効果を分析します。ビジネスの理解フェーズで策定した目標に到達していますか？

これらが問題に正しく対処でき、現在のモデルが最終目標に到達できたと判断できるなら、さらに詳細にモデルを評価して最終的な展開作業を行います。そうでない場合は、ここで学習した内容を元にパラメータの設定を調整し、モデルを再実行してください。

e-Commerce 業者の例 – モデルの評価

CRISP-DM を使った 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 1 つと、ある TwoStep クラスタリングから、それぞれ適切な結果を得ることができました。しかしまだ、どちらを選択するかを判断することは困難です。やがて、この業者は両方のモデルを使用することに決めました。両方の手法のお勧め商品に同意し、両者が異なる場合の状況についても綿密な調査を行いました。わずかな労力とビジネス知識を活用することにより、この業者は 2 つの手法間の差異を解決するルールを開発することができました。

また、セッション情報を含む結果が非常に良好であることもわかりました。結果には、お勧め商品とサイト内の移動を結びつけられる明確な情報があります。顧客が次に移動する可能性がある場所を定義したルールセットをリアル タイムに使用して、顧客のサイト移動に応じたコンテンツを直接表示することができます。

サイト ナビゲーションの改善： シーケンス モデルは特定の顧客パスを予測できる高レベルの確信度を持つ e-Commerce を提供します。これにより、適切な数の変更をサイト デザインに加えることを提案する結果が生成されます。

修正したパラメータの追跡

モデル評価時に学習した内容に基づいて、モデルを別の観点から調査してみましょう。ここでは、次の 2 種類の作業を行えます。

- 既存のモデルのパラメータを調整する。
- データ マイニングの問題に対処するための、別のモデルを選択する。

どちらの場合でも、モデリング作業に戻って、満足できる結果が得られるまで反復作業を行います。このステップを繰り返すことについては、あまり気にしないようにしてください。データ マイニングにおいては、評価を行った後モデルを再実行することを複数回繰り返し、ニーズに最適なモデルを見つけだすことが頻繁に行われています。一度に複数のモデルを構築し、それらの結果を比較しながらパラメータを調整していくのも良いでしょう。

次の手順に進む用意ができましたか？

モデルの最終評価に進む前に、反復評価作業が完全に完了したかどうかを確認してください。

作業リスト

- モデルの結果を理解することができますか？
- モデルの結果は、純粋に論理的な観点から意味を為していますか？さらに調査を行う必要があるような、明確な不整合がないですか？
- 第一印象として、組織のビジネス上の問題に対処できるような結果ですか？
- 精度分析ノードや、リフト グラフまたはレスポンス グラフなどを使って、モデルの精度を比較、評価しましたか？
- 複数の種類のモデルを調査して、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 モデルの結果が正確で関連性があるならば、最終的な展開を行う前により詳細な評価を行います。

章 6

評価

評価の概要

この時点で、データ マイニング プロジェクトの大部分の作業が完了しています。また、モデリング フェーズで、それ以前のフェーズで定義した**データ マイニングの成功基準**と比較して、作成したモデルが技術的に正しく効果的であることを確認しています。

しかし、作業を続行する前には、プロジェクトのはじめに作成した**ビジネス上の成功基準**を使って、結果を評価する必要があります。このことは、組織で結果を有効活用できるようにするためにも必要不可欠な作業です。データ マイニング作業により、次の 2 種類の結果が生成されます。

- 前の CRISP-DM フェーズで選ばれた、最終**モデル**。
- モデル自身やデータ マイニング プロセスから導き出された結論や推論。これらは**所見**として知られています。

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上の成功基準を満たしているかどうかを正式に評価します。この作業を行うには、あらかじめ作成されたビジネス目標を明確に理解していなければなりません。そのため、プロジェクトの評価には、主要担当責任者を参加させるようにしてください。

作業リスト

まず、データ マイニングの結果がビジネス上の成功基準を満たしているかどうかの評価を文書に記録する必要があります。レポートを作成するにあたっては、次の事項を確認してください。

- 結果は明確に記述され、簡単に公開できる形式ですか？
- 特に強調すべき新しいまたは固有の所見はありますか？
- モデルや所見を、ビジネス目標への適用可能度に応じてランク付けできますか？
- 全般的に、導き出された結果は、ビジネス上の目標に対して適切なものでしたか？
- 結果から提起された問題はありますか？これらの問いをビジネス用語でなんと呼びますか？

結果を評価したら、最終レポートに記載する、承認された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 マイニング上およびビジネス上の目標の両方を満たすモデルを入れる必要があります。

e-Commerce 業者の例 – 結果の評価

CRISP-DM を使った Web マイニングのシナリオ

この業者の初めてのデータ マイニングにおける結果は、業務上の観点からかなり簡単に伝えることができます。分析調査により、より状況に適したお勧め商品の表示と、サイト デザインの改善に関する情報を得ることができました。サイト デザインの改善は、顧客が目的の場所に到達するまでに何回も作業を行う必要があるページを示す、顧客の参照シーケンスに基づいています。お勧め商品に関する調査は、ディシジョン ルールが複雑なため、判断が困難なものとなります。最終レポートを作成するために、より簡単に説明できる、ルールセット中の一般的な傾向を探し出すことにします。

モデルのランク付け： 複数の初期モデルがビジネスに役立つようなので、統計基準、解釈しやすさ、および多様性に基づいて、それらのランク付けを行いました。このようにして、状況に応じて推奨モデルも異なることがわかりました。

新しい問題： この調査結果でもっとも重要な問題は、業者はどのようにしたらより詳細に顧客のことを把握できるかということにあります。顧客データベース中の情報は、お勧め商品のクラスタを作成するために重要な役割を果たします。情報が欠けている顧客に対してお勧め商品を表示する特殊なルールを採用することもできますが、本質的にこのようなお勧め商品は登録顧客に対して行うよりも、より一般的なものとなってしまいます。

プロセスの見直し

一般的に、効果的な手法には、完了したプロセスの利点と欠点を反映させるための時間が含まれています。データ マイニングでも、このことに変わりはありません。CRISP-DM では、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、今回の作業経験から学習していきます。

作業リスト

まず、データ準備作業やモデリング作業も含め、各フェーズで行った作業や決定内容をまとめる必要があります。次に各フェーズに対して、次の事項を確認し、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？
- 特定のステージや作業を効率化したり改善する手段はありますか？

- このフェーズにおける失敗や過ちは何でしたか？次回にこのような問題を回避するにはどうしたら良いでしょうか？
- 特定のモデルが役に立たないなどの問題はありましたか？労力を無駄にしないためにも、そのような問題を予測して回避する方法はありますか？
- このフェーズで特に注目すべき事柄はありましたか（良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手段はありましたか？
- それぞれのフェーズで利用できる代替手段、判断、戦略などはありましたか？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、それらの代替手段を記載してください。

e-Commerce 業者の例 – レビュー レポート

CRISP-DM を使った Web マイニングのシナリオ

初期データ 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、この業者はプロセス中のステップ間の相互関係に大変満足しました。当初は CRISP-DM プロセスの途中で「元に戻る」ことに気が進まなかった業者が、今ではプロセスの循環性のおかげでより優れた結果を導き出せることを理解しています。また、プロセスをレビューすることによって、この業者は次のような事柄も理解することができました。

- CRISP-DM プロセスのあるフェーズで何か普通でない事柄が発生した場合、探索プロセスに戻ることが常に保証されています。
- 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。
- ビジネス上の問題に常に注目しておくことが大切です。データの分析準備が完了したら、大きな図にとらわれずにモデルの作成を開始する方が簡単です。
- モデリング フェーズを完了したら、結果の導入および将来何を学習をするかを決めるために、ビジネスの理解がより重要になります。

次のステップの決定

現時点では、すでに結果を生成し、これまでのデータ マイニング作業の評価を完了しています。また、**次に何を行えば良いのか**と考えている方もいるでしょう。このフェーズでは、データ マイニングにおけるビジネス目標の観点からこの疑問に対する回答を探していきます。基本的に、現時点では次の 2 種類の選択肢があります。

- **展開フェーズに進む。**次のフェーズでは、モデルの結果を業務に導入し、最終レポートを作成します。データ マイニングの結果が失敗に終わっても、CRISP-DM の展開フェーズの作業を行って、最終レポートをプロジェクトのスポンサーに提出しなければなりません。
- **前に戻って、モデルを調整または変更する。**結果はおおむね良好だけれども完全ではないと考える場合は、もう一度モデリング作業を行うことも検討してください。このフェーズで学んだことを活用してモデルをさらにきめ細かく調整すれば、より良い結果を得ることができるでしょう。

この時点での決定事項が、モデル作成結果の正確性や適合性に影響を及ぼします。結果がデータ マイニング上のおよび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。どちらを選んだ場合でも、評価プロセスをきちんと文書に記録することを忘れないようにしてください。

e-Commerce 業者の例 – 次のステップ

CRISP-DM を使った Web マイニングのシナリオ

プロジェクト結果の精度や適合性にかなりの確信を持てたので、次に展開フェーズに進みます。

同時に、プロジェクト チームは、予測技術を取り入れてモデルの強化を行うことができます。この時点で、プロジェクト チームは、最終レポートの配布および責任者の指示を待っています。

章
7

展開

展開の概要

展開は、新しい洞察を使って組織の改善を行うプロセスです。このことは、IBM® SPSS® Modeler のモデルで生成され、後でデータ ウェアハウスに取り込まれる Churn Score の実装などの、正式な統合を意味しているということもできます。また、展開が、データ マイニングで得られた洞察を使って、組織の変更を実施することを意味することもあります。たとえば、データ中に 30 代以上の顧客に乗り換え傾向があるという、警告パターンが見つかったとします。これらの結果は正式に情報システムに統合されることはないかもしれませんが、マーケティング上の計画や意志決定を行うためには間違いなく有益な情報です。

通常、CRISP-DM の展開フェーズには、2 種類の作業が含まれます。

- 結果の展開作業のプランニングと監視
- 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残りの作業の完了

組織の要件に応じて、以上のどちらかまたは両方の作業を実施する必要があります。詳細は、以下の作業を参照してください。

展開のプランニング

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。

作業リスト

- まず、モデルと所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、どの所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説明も記載します。たとえば、システムではモデル出力を、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。
- それぞれの最終結論に対して、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？
- 展開作業をどのように監視するかを検討します。たとえば、IBM® SPSS® Modeler Solution Publisher を使って展開したモデルをどのように更新しますか？モデルが不適切になった場合は、どのように対処しますか？
- 展開上の問題を判断し、不測の事態に対する計画も作成してください。たとえば、意志決定責任者がモデルにより得られた結果に対してより詳細な情報を求め、技術詳細を提供する必要があることも考えられます。

e-Commerce 業者の例 – 展開のプランニング

CRISP-DM を使った Web マイニングのシナリオ

データ マイニングの結果を正しく展開するためには、正しい情報が適切な人々に渡されなければいけません。

責任者 (意志決定者)： 責任者には、サイトの推奨事項と変更の提案を行い、提案内容の利点や欠点を簡単に説明する必要があります。調査結果が承認された場合は、その変更を実際に導入する担当者にも通知する必要があります。

Web 開発者： Web サイトを管理する人々は、提案された推奨事項やサイトのコンテンツの編成を、実際の環境に反映させる必要があります。また、将来の分析調査に備えて、どのような変更が行われる可能性があるかを教え、それに対する準備も行わせてください。リアルタイム シーケンス分析に基づいて、サイト構築用のチームを準備しておけば後ほど役に立つことでしょう。

データベース担当者： 顧客データベース、購入履歴データベース、および商品データベースを保守、管理している人々には、データベースの情報がどのように使われているか、また将来のプロジェクトにおいてどのような属性がデータベースに追加される可能性があるかを教えてください。

特に、プロジェクト チームはこれらのグループと密接に連絡を取り合って、結果の展開作業と将来のプロジェクトのプランニングを行っていく必要があります。

監視と保守のプランニング

モデルから得られた結果の完全な展開と統合作業が完了しても、データ マイニング作業が続くことがあります。たとえば、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ためにモデルを導入した場合、このモデルを定期的に評価してその効果を確認し、継続して改善を図ること

ができます。同様に、高い価値を持つ顧客の保持率を増加するためのモデルを展開した場合、目標の保持レベルに達したらモデルを微調整する必要があるかもしれません。この場合、たとえば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も保持するために、モデルを調整、変更して再利用することができます。

作業リスト

次の問題に関する情報を記録してください。また、これらの情報は忘れずに最終レポートに記載してください。

- それぞれのモデルや所見に対して、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精度をどのように測定、監視しますか？
- モデルが「古くなった」ことはどのように判断しますか？このように判断するための、特定の精度閾値やデータの変更などの条件を確認してください。
- モデルが古くなった場合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ますか？それとも、パラメータを微調整しますか？または、全面的な変更になるため、新しいデータ マイニング プロジェクトを開始する必要がありますか？
- 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？このような場合に、今までの作業をきちんと文書化しておいたことが役立ちます。これらの文書は、各データ マイニング プロジェクトにおいて、ビジネス上の目的を評価するために必要不可欠です。

e-Commerce の例 – 監視と保守

CRISP-DM を使った Web マイニングのシナリオ

監視において差し迫った作業は、新しいサイト編成と推奨事項の改善が実際にうまくいっているかどうかを確認することです。つまり、ユーザーはより素早く目的のページを参照できているかどうか、またお勧め商品の抱き合わせ販売が順調に増加しているかどうかを確認する必要があります。数週間の監視を行えば、プロジェクトが成功したかどうかを判断することができるでしょう。

新しい登録ユーザーに対する自動処理を行うこともできます。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して、どのようなお勧め商品を表示するかを判断することができます。

お勧め商品を判断するルールセットをいつ更新するかを決める作業は、慎重に行わなければなりません。クラスタの作成には、適切さの観点から人手による入力が必要なため、ルールセットの更新は自動的に行える処理ではありません。

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伴い、監視の必要性と監視対象の数も増加していきます。できる限り、監視作業を自動化し、レビュー用の定期的なレポートを生成するようにしてください。代わりに、予測モデルを作成して、会社の指針とする方法もあります。この方法では、最初のデータ マイニング プロジェクト時に、よりきめ細かく洗練された作業が必要となります。

最終レポートの作成

最終レポートの作成は、これまでに作成した文書のまとめと総括であるだけでなく、結果を伝達するためにも用いられます。これは簡単なことのように思えますが、関係があるさまざまな人々に正しく結果を伝えることが非常に重要になります。これには、モデルの結果の実装を担当する技術管理者と、結果に基づいて意志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ることができます。

作業リスト

まず、レポートの対象読者を検討します。技術開発者を対象にしますか、それともマーケティング関係のマネージャですか？それぞれのニーズが異なる場合は、対象読者ごとに個別のレポートを作成する必要があります。どちらの場合でも、レポートには次の項目を記載しなければなりません。

- ビジネス上の問題の詳細な説明
- データ マイニングに用いられたプロセス
- プロジェクトの費用
- 元のプロジェクト計画から変更された事項
- データ マイニングの結果の要約（モデルと所見の両方）
- 提案する展開計画の概要
- 探索やモデリングの過程で見つかった手がかりなど、将来のデータ マイニング作業における推奨事項

最終プレゼンテーションの準備

プロジェクト レポートのほかにも、スポンサーや関連部門にプロジェクトの所見などをプレゼンテーションする必要があることもあります。この場合、レポートに記載した大部分の情報をそのまま流用できますが、

さらに幅広い観点からプレゼンテーションを行う必要があります。IBM® SPSS® Modeler のグラフをエクスポートして、このようなプレゼンテーションに利用することもできます。

e-Commerce 業者の例 – 最終レポート

CRISP-DM を使った Web マイニングのシナリオ

当初のプロジェクトからの大幅な方針変更は、将来のデータ マイニング作業において大変有益な情報になります。当初の計画では、顧客がサイトを訪れた際に、より時間をかけて各ページを参照させるようにするにはどうすれば良いかを見つけだすことを目的にしていました。

その後の調査により、単に顧客を満足させることがサイトの継続利用につながる訳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から、そのセッションが購入に結びついたかどうかを分割し、購入へとつながる大半のセッションが、購入されなかった 2 つセッションのクラスタ間に集中することがわかりました。

このことからは、サイトに長時間訪れながら商品を購入しない顧客は、サイトを参照しているだけなのか、それとも単に目的の商品を見つけだせないからかを判断することが問題になってきます。次のステップは、顧客の商品購入を促進するために、顧客が求める商品を提供する方法を探し出すことです。

最終プロジェクト レビューの実施

これが CRISP-DM 手法の最終ステップです。ここで、最終的な作業結果の編成と、データ マイニング作業時に学んだ留意事項などをまとめていきます。

作業リスト

データ マイニング プロセスに大きく関与した人々に、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は、次のようなことを問い合わせてください。

- プロジェクトに関する総合的な印象は？
- 作業の過程で何を学びましたか（データ マイニングに関する一般的な事柄や利用できるデータについて）？
- プロジェクトで順調に作業を行えた箇所は何ですか？どのような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 マイニングの結果の展開を完了したら、その結果の展開によって影響を受けた顧客や提携企業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、また結果を反映したことにより利点があったかどうかを判断することにあります。

これらのインタビュー結果は、自分のプロジェクトに対する印象とともに最終レポートにまとめることができます。この最終レポートでは、データ マイニング作業で学習した貴重な情報に重点を置いて記載するようにしてください。

e-Commerce 業者の例 – 最終レビュー

CRISP-DM を使った 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： プロジェクトの開始から終了まで密接に作業に関わったメンバーは、ほぼ結果に満足しており、将来のプロジェクトに期待を寄せています。データベース グループは慎重ながらも楽観的で、プロジェクトの結果の有用性に満足しながらも、一方でデータベース リソースの負担が増えたことを指摘しました。このプロジェクトではコンサルタントにも作業を依頼しましたが、今後プロジェクトの範囲が拡張するにつれて、データベース保守の専任者を配置することが必要になるでしょう。

顧客へのインタビュー： 現在までの所、顧客からのフィードバックの大半は良好なものです。見落としていた問題点としては、サイト デザインの変更が従来の顧客に対して与える影響について考慮していなかったことがあげられます。数年の間に、登録顧客はサイトの編成方法に関する期待や要望を寄せました。登録ユーザーからのフィードバックは、登録していない顧客からのフィードバックほど良好ではなく、一部の人々は変更を大変不便に感じていました。今後業者は、この問題を念頭において、この変更が既存の顧客を失うリスクを補ってあまりあるほどの新規顧客を開拓できるかどうかを注意深く検討する必要があります。

付録

A

注意事項

この情報は、世界各国で提供される製品およびサービス向けに作成されています。

IBM はこのドキュメントで説明する製品、サービス、機能は他の国では提供していない場合があります。現在お住まいの地域で利用可能な製品、サービス、および、情報については、お近くの IBM の担当者にお問い合わせください。IBM 製品、プログラム、またはサービスに対する参照は、IBM 製品、プログラム、またはサービスのみが使用することができることを説明したり意味するものではありません。IBM の知的所有権を侵害しない機能的に同等の製品、プログラム、またはサービスを代わりに使用することができます。ただし、IBM 以外の製品、プログラム、またはサービスの動作を評価および確認するのはユーザーの責任によるものです。

IBMは、本ドキュメントに記載されている内容に関し、特許または特許出願中の可能性があります。本ドキュメントの提供によって、これらの特許に関するいかなる権利も使用者に付与するものではありません。ライセンスのお問い合わせは、書面にて、下記住所に送ることができます。

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.

2 バイト文字セット (DBCS) 情報についてのライセンスに関するお問い合わせは、お住まいの国の IBM Intellectual Property Department に連絡するか、書面にて下記宛先にお送りください。

神奈川県大和市下鶴間1623番14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

以下の条項はは、イギリスまたはこのような条項が法律に反する他の国では適用されません。 International Business Machines は、明示的または黙示的に関わらず、第三者の権利の侵害しない、商品性または特定の目的に対する適合性の暗黙の保証を含むがこれに限定されない、いかなる保証なく、本出版物を「そのまま」提供します一部の州では、特定の取引の明示的または暗示的な保証の免責を許可していないため、この文が適用されない場合があります。

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。情報については変更が定期的に行われます。これらの変更は本書の新版に追加されます。IBM は、本書に記載されている製品およびプログラムについて、事前の告知なくいつでも改善および変更を行う場合があります。

IBM 以外の Web サイトに対するこの情報内のすべての参照は、便宜上提供されているものであり、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。これらの Web サイトの資料はこの IBM 製品の資料に含まれるものではなく、これらの Web サイトの使用はお客様の責任によるものとします。

IBM はお客様に対する一切の義務を負うことなく、自ら適切と考える方法で、情報を使用また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取得者が (i) 別途作成されたプログラムと他のプログラム (本プログラムを含む) との間の情報交換および (ii) 交換された情報の相互利用を目的とした本プログラムに関する情報の所有を希望する場合、下記住所にお問い合わせください。

IBM Software Group, Attention:Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

上記のような情報は、該当する条項および条件に従い、有料で利用できるものとします。

本ドキュメントに記載されている許可されたプログラムおよびそのプログラムに使用できるすべてのライセンス認証された資料は、IBM Customer Agreement、IBM International Program License Agreement、および当社とかわした同等の契約の条件に基づき、IBM によって提供されます。

ここに記載されているパフォーマンスデータは、すべて管理環境下で確認されたものです。そのため、他の操作環境で得られた結果は大きく異なる可能性があります。開発レベルのシステムで測定が行われている場合があり、これらの測定値は一般に利用可能なシステムと同じであることを保証するものではありません。また、測定値が推定値である可能性があり、実際の結果は異なる場合があります。本ドキュメントのユーザーは、特定の環境に適したデータを検証する必要があります。

IBM 以外の製品に関する情報は、それらの製品の供給業者、公開済みの発表、または公開で使用できるソースから取得しています。IBM は、それらの製品のテストは行っておらず、IBM 以外の製品に関連する性能、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業者に通知する必要があります。

IBM の将来の方向性または意向に関する記述については、予告なく変更または取り消すことがあり、目的や目標のみを示すものです。

この情報には、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。できる限り詳細に説明するため、例には、個人、企業、ブランド、製品などの名前が使用されています。これらの名称はすべて架空のものであり、実際の企業で使用される名称および住所とは一切関係ありません。

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

商標

IBM、IBM ロゴ、および ibm.com、SPSS は、世界の多くの国で登録された IBM Corporation の商標です。IBM の商標の現在のリストは、http://www.ibm.com/legal/copytrade.shtml を参照してください。

Intel、Intel のロゴ、Intel Inside、Intel Inside のロゴ、Intel Centrino、Intel Centrino のロゴ、Celeron、Intel Xeon、Intel SpeedStep、Itanium、および Pentium は、米国およびその他の国の Intel Corporation または関連会社の商標または登録商標です。

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

Microsoft、Windows、Windows NT、および Windows のロゴは、米国およびその他の国における Microsoft 社の商標です。

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

Java およびすべての Java ベースの商標およびロゴは、米国およびその他の国の Sun Microsystems, Inc. の商標です。

その他の製品名およびサービス名等は、IBM または他の会社の商標です。

索引