2016.10.27.THU

データサイエンスの舞台裏

成功の秘訣は基本にあった！

株式会社ef-prime
鈴木 了太

ef-prime, inc.

はじめに

- 本日の話題
  - データ分析のビジネス活用を成功させるには、どうすればよいか？
  - 分析の基本的な考え方と、成果に繋げる秘訣をご紹介

- 対象
  - データ分析のビジネス活用に興味がある方
  - データ分析に関わるすべての方
    - これから分析やデータ処理を担当
    - 社内部門や外部機関への委託、分析結果の運用など

自己紹介

- 鈴木 了太
  - 株式会社ef-prime（エフプライム）代表

- 会社概要
  - 2006年3月設立
  - データ分析コンサルティング
    - ビジネス課題を統計解析、機械学習、最適化で解決
    - 分析結果の納品からソフトウェアによる自動化、
      分析方法のご提案や社内研修など

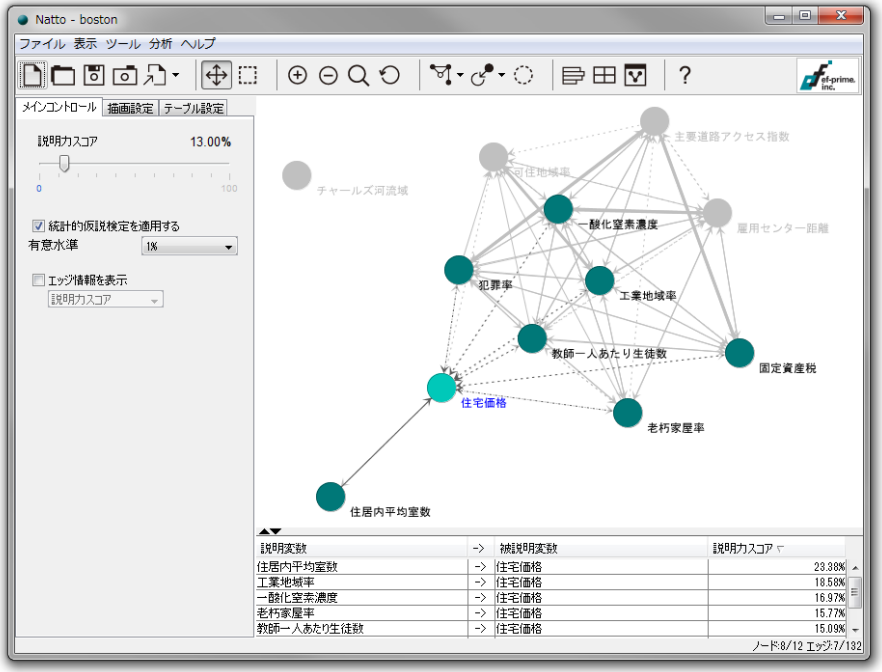
専門分野

- おもにビジネスデータ分析
  - 分析内容
    - 購買予測、キャンペーン最適化、広告効果測定、リコメンド、
      指標策定、需要予測、調査データ解析、異常検知、可視化など
  - クライアント様業種
    - 通販、教育、金融、医薬、消費財、流通、製造業、
      広告、通信、研究機関、調査会社など

ef-prime, inc.

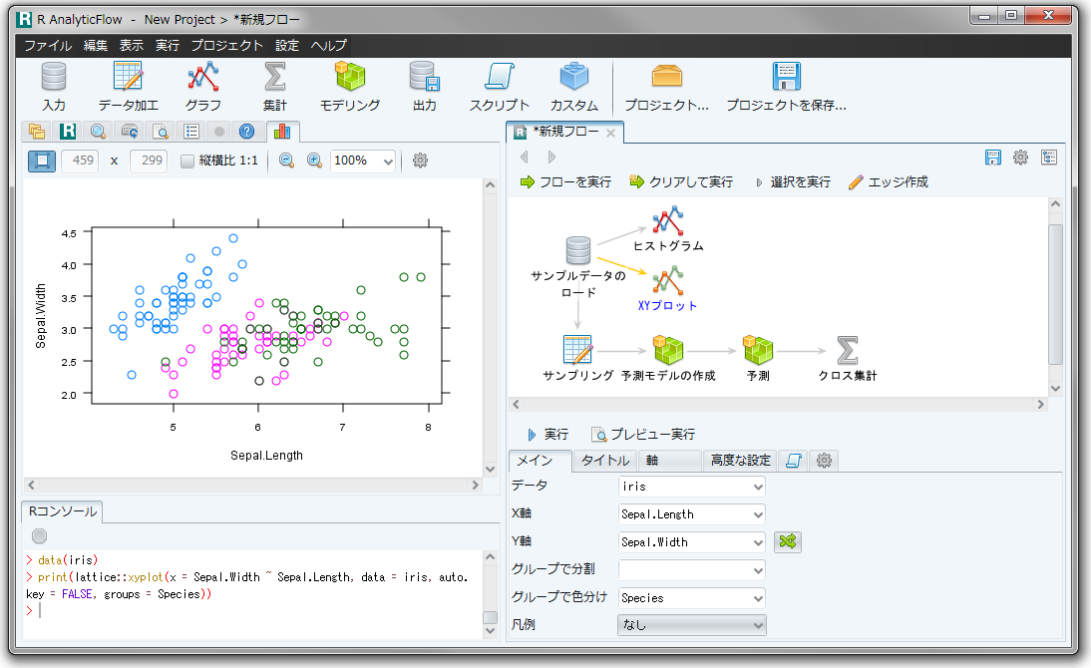
弊社開発ツール：Natto

データに含まれる関係性を視覚化し、直感的な操作で傾向を把握。
数値と文字が混ざった項目や欠損にも対応し、分析の初期段階で特に有用



弊社開発ツール：R AnalyticFlow

統計解析環境RのGUIで、ワークフロー形式で分析プロセスを記述。
マウス操作で分析を実施でき、プログラミング補助機能も充実



データ分析そもそも論

そもそもデータ分析とは何か？というところからビジネスへの応用まで、
事例を交えながらデータ分析の基本をご紹介します。

データについて

- データとは
  - 数値や文字などの形で記録された情報
    - 特にコンピュータで処理できるよう電子化されたもの

- データの種類
  - 整理された定型データ
    - 数値、カテゴリ、有無など。もっとも分析しやすい
  - 非定型データ
    - アンケートの自由回答文、画像、音声、動画など。
      定型データに変換したり、専用の分析ツールが必要

分析しやすいデータの例

- マーケティング分析の場合
  - 数値
    - 購入金額、利用回数、利用後の経過日数、
      ページ滞在時間、ウェブ広告のクリック率、クリック後の利用率、
      年齢、家族人数、年収…
  - カテゴリ
    - 性別、都道府県、会員区分、住居区分（持ち家、マンションなど）、
      職業区分（会社員、自営業など）…
  - 有無
    - 購入有無、クレジットカード利用有無、クーポン使用有無、
      問い合わせ有無、アンケート回答有無、家族会員有無…

事例：ダイレクトメールの発送最適化

- ダイレクトメール発送問題
  - 膨大な人数の見込み顧客リスト
  - 予算の範囲内で発送し、利益を最大化したい

- 蓄積されたデータ
  - デモグラフィックス（例：性別、年齢、住所）
  - コンタクト履歴（例：過去の問い合わせ有無）
  - アンケートへの回答（例：購入意向）

事例：ダイレクトメールの発送最適化

- 予測分析による解決策
  - ダイレクトメールを発送したときの購入確率を予測

事例：ダイレクトメールの発送最適化

- 予測分析による解決策
  - ダイレクトメールを発送したときの購入確率を予測
  - 購入確率スコアの高い順に発送し、効率よく顧客を獲得
  - 予測にもとづく期待売上を用いて利益を最大化

データ分析が実現すること

- データを何らかの役に立てる
  - 情報を整理する
  - 現象を理解する
  - 未知を予測する
  - 最適な行動を導き出す

情報を整理する

- 集計と視覚化
  - 表や集計値にまとめたり、グラフを描いて視覚化する
    - 例：購入金額を都道府県別に集計し、地図を塗り分け
  - 統計学では記述統計と呼ばれる分野
    - BIツールが得意とする分野でもある

- 分類
  - 似たもの同士をまとめる
    - 「クラスター分析」という分析手法が利用できる
      - 目的に応じて予測分析の方法論を援用するのが効果的

現象を理解する

- 構造や因果関係を明らかにする
  - 例：ある生活習慣が、将来の病気リスクを3倍に
    - 学術研究における分析に多い
    - 一般的な「分析」のイメージはこれでは？
  - 統計学では推測統計と呼ばれる分野
    - 完全にイコールではないが、関係は強い
  - 集計や視覚化によってわかることもある
    - 集計して眺めてみることで関係が見つかったり、分類ごとに集計したら違いが見つかるということも
      - データそのものを眺めることも重要

未知を予測する

- 予測とは
  - 既知の情報（X）から未知の情報（Y）を言い当てる
    - 例：過去の利用履歴や年収から、カード入会有無を予測
  - 未知の情報は将来のものとは限らない
    - 例：故障の原因を探る分析。原因はどこかにあるが、わからない

未知を予測する

- 現象の理解と予測の違い
  - パターンの発見という意味では同じ
    - 例：その生活習慣では、いずれ病気になるだろう
    - 実際、使用する分析手法も共通のものが多い
  - 理由はわからなくても、当たればよい
    - 例：猫が顔を洗っているから、雨が降るだろう
      ⇒　猫は（たぶん）雨の原因ではない！
      - 洗顔をやめさせても仕方ない
  - 利用する分析手法自体も「説明的」でなくてよい
    - それこそ「人工知能」に任せるやり方も
    - 挙動が読めない不確かさもあるので、ケースバイケース

最適な行動を導き出す

- コントロール可能な問題
  - 「XならばY」の問題のうち、Xに選択の余地がある
    - 例：
      ある顧客に電話したとき（X = 1）、予測では15%の確率で
      粗利1万円の商品が購入される。
      電話しなければ（X = 0）、購入確率は5%と予測される。

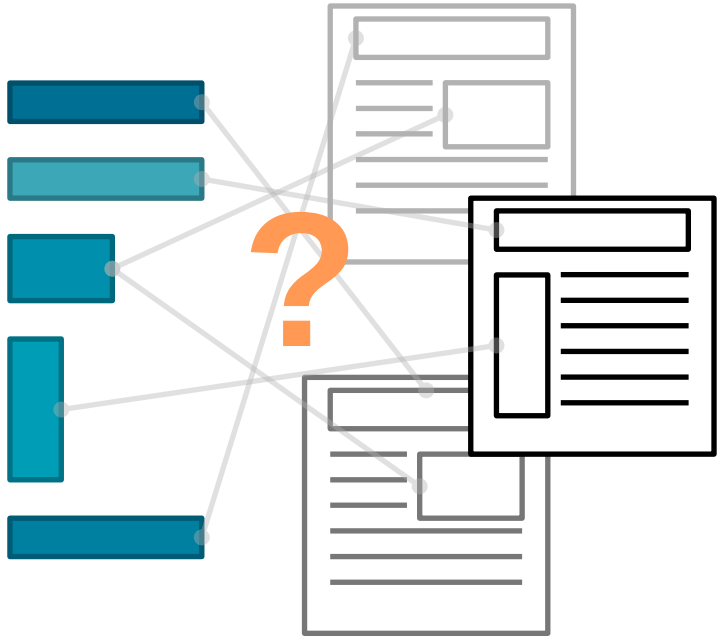
      期待利益は 10,000 ×（0.15 − 0.05）= 1,000円であるから、
      電話1回あたりの費用が1,000円未満なら電話をかける価値がある。

最適な行動を導き出す

- 特徴
  - 例のように、予測に基づいて問題を設定することが多い
  - 複雑な問題の場合、数理的最適化の手法が用いられる
    - 線形計画法、組み合わせ最適化など
  - Xのコントロール可能性に注意
    - 例：イベント参加者の購入率が高い ≠ イベントに来場させればよい
      - 「興味があるからイベントに来る」という因果関係である可能性
    - 別のX'との関連など、分析用データの選択にも注意が必要

事例：バナー広告の配信最適化

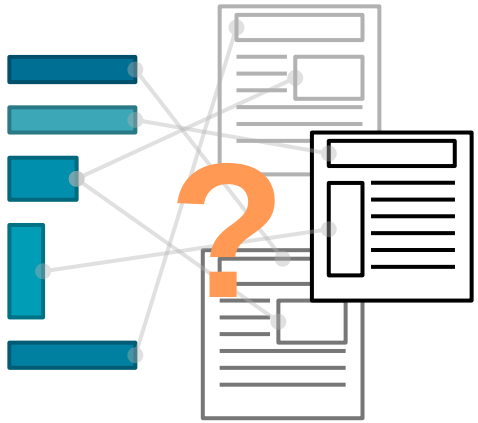
- このバナー広告、どこに表示する？



事例：バナー広告の配信最適化

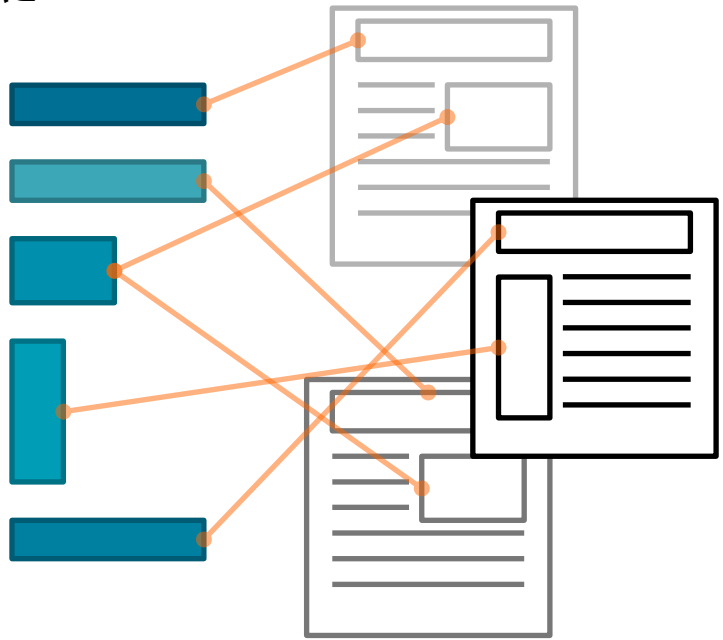
- 複数の媒体ウェブサイトでバナー広告を表示
  - バナーによって、サイズや成約時の利益額が異なる
  - 媒体によって費用の金額や基準が異なる（表示ごと、クリックごと）
  - どの媒体にどの広告を表示するかは任意に設定可能

- 成果データは毎日更新、最新のものが利用できる
  - インプレッション数（表示回数）、広告クリック回数、成約回数（成果ページ訪問など）などが取得可能



事例：バナー広告の配信最適化

- 予測＋最適化による解決策
  - 日時とバナーと媒体の組み合わせごとに利益額を予測、期待利益を算出
  - 利益最大化の観点から最適な配信計画を策定
  - 常に最新のデータを反映させ、配信計画を自動的に更新



すべての道は（だいたい）意思決定に通じる

- 最終的にはアクションに繋がる
  - 情報を整理して ⇒ 現象を理解すれば ⇒ 何をすべきかわかる
  - 未知を予測すれば ⇒ 最適な行動がわかる

すべての道は（だいたい）意思決定に通じる

- 最終的にはアクションに繋がる
  - 情報を整理して ⇒ 現象を理解すれば ⇒ 何をすべきかわかる
  - 未知を予測すれば ⇒ 最適な行動がわかる

- 目的から逆算する
  - 誰にアプローチする？ ⇐ 購入してくれる人 ⇐ 購入を予測
  - どこに広告を出す？ ⇐ 利益になる媒体 ⇐ 利益を予測
  - 顧客をどう分類する？ ⇐ 最適な販売チャネルごと
    ⇐ チャネル適合度を評価 ⇐ チャネルごとの予測売上

すべての道は（だいたい）予測分析に通じる

- 最終的にはアクションに繋がる
  - 情報を整理して ⇒ 現象を理解すれば ⇒ 何をすべきかわかる
  - 未知を予測すれば ⇒ 最適な行動がわかる

- 目的から逆算する
  - 誰にアプローチする？ ⇐ 購入してくれる人 ⇐ 購入を予測
  - どこに広告を出す？ ⇐ 利益になる媒体 ⇐ 利益を予測
  - 顧客をどう分類する？ ⇐ 最適な販売チャネルごと
    ⇐ チャネル適合度を評価 ⇐ チャネルごとの予測売上

すぐわかる予測分析

ここまで、予測分析の重要性について説明しました。
では予測分析とはどのようなもので、どうすればうまくいくのでしょうか？
基本から応用上の重要事項まで、一気にご紹介します。

予測分析

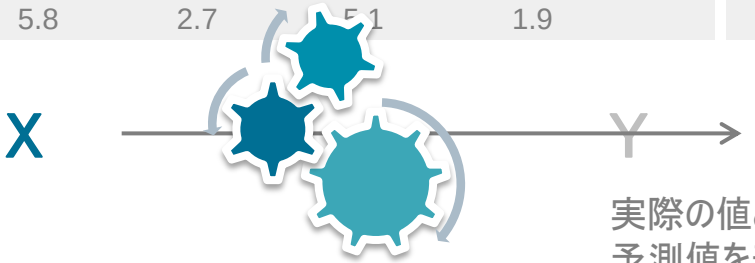
説明変数（X）の値から目的変数（Y）の値を予測します。
すなわち、Xの値だけを見て未知のYを言い当てることを考えます。

説明変数1	説明変数2	説明変数3	説明変数4	目的変数
5.1	3.5	1.4	0.2	?
4.9	3	1.4	0.2	
7	3.2	4.7	1.4	
6.4	3.2	4.5	1.5	
6.3	3.3	6	2.5	
5.8	2.7	5.1	1.9	

予測モデル

説明変数の値を入力すると、目的変数の予測値を出力する数式やロジック（アルゴリズム）を予測モデルと呼びます。

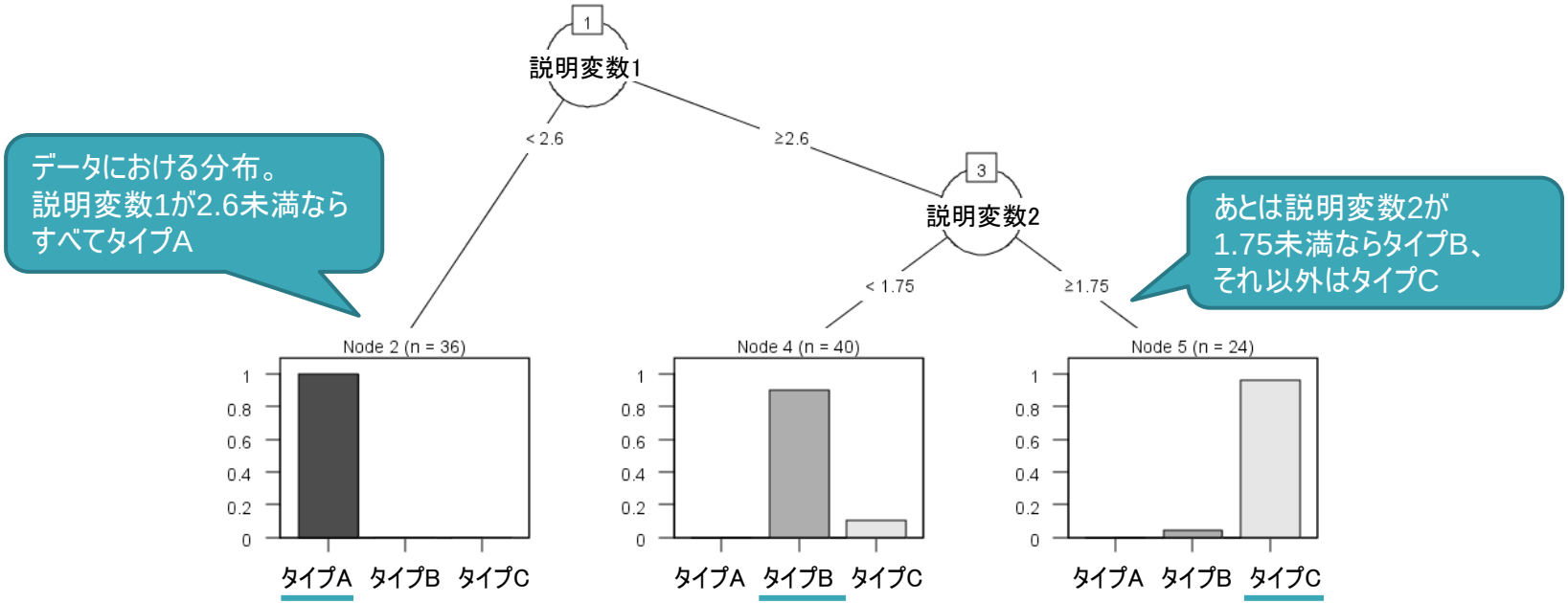
説明変数1	説明変数2	説明変数3	説明変数4	目的変数	予測値
5.1	3.5	1.4	0.2		
4.9	3	1.4	0.2		タイプA
7	3.2	4.7	1.4		
6.4	3.2	4.5	1.5		
6.3	3.3	6	2.5		
5.8	2.7	5.1	1.9		



実際の値と区別するため、Y に ^ をつけて予測値を表し、ワイハットと読みます。

予測モデルの例：決定木

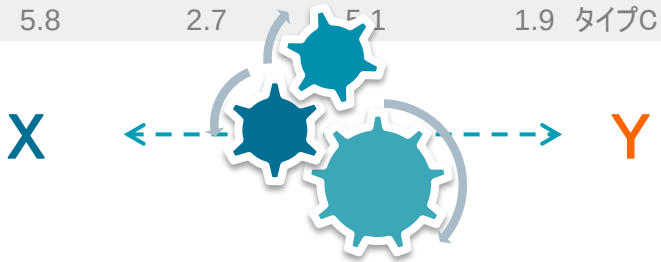
決定木モデルは説明変数を用いて自動的にルールを生成します。
ルールは木の形で表され、上から順に辿ることで予測値が得られます。



モデルの学習

決定木モデルにおけるルールの自動生成のように、
データから予測の仕組みを構築することをモデルの学習といいます。
このとき用いるデータを学習用データと呼びます。

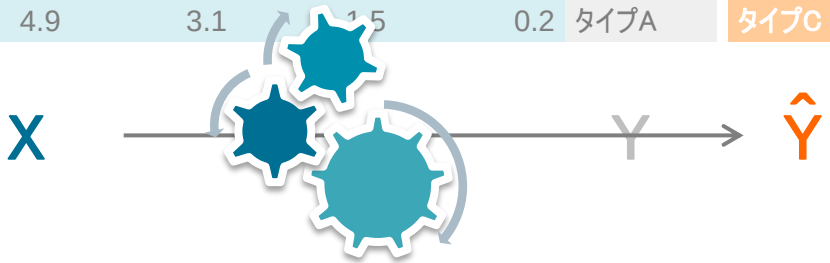
説明変数1	説明変数2	説明変数3	説明変数4	目的変数
5.1	3.5	1.4	0.2	タイプA
4.9	3	1.4	0.2	タイプA
7	3.2	4.7	1.4	タイプB
6.4	3.2	4.5	1.5	タイプB
6.3	3.3	6	2.5	タイプC
5.8	2.7	5.1	1.9	タイプC



モデルの評価

モデルに説明変数だけを与えて予測値を出力し、
実際の値と比較することで予測精度を評価します。
このとき用いるデータを検証用データまたはテストデータと呼びます。

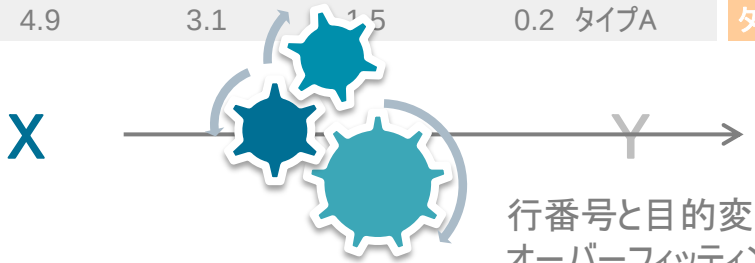
説明変数1	説明変数2	説明変数3	説明変数4	目的変数	予測値	
4.7	3.2	1.3	0.2	タイプA	タイプA	✔
6.3	2.9	5.6	1.8	タイプC	タイプA	✔
5.5	2.3	4	1.3	タイプB	タイプC	✘
7.1	3	5.9	2.1	タイプC	タイプB	✘
6.9	3.1	4.9	1.5	タイプB	タイプC	✔
4.9	3.1	1.5	0.2	タイプA	タイプC	✔



過学習（オーバーフィッティング）

モデルが学習用データに特化してしまい、他のデータに対する予測が外れる現象を過学習（オーバーフィッティング、過剰適合）といいます。
あらかじめデータを学習用と検証用に分けておくことで回避できます。

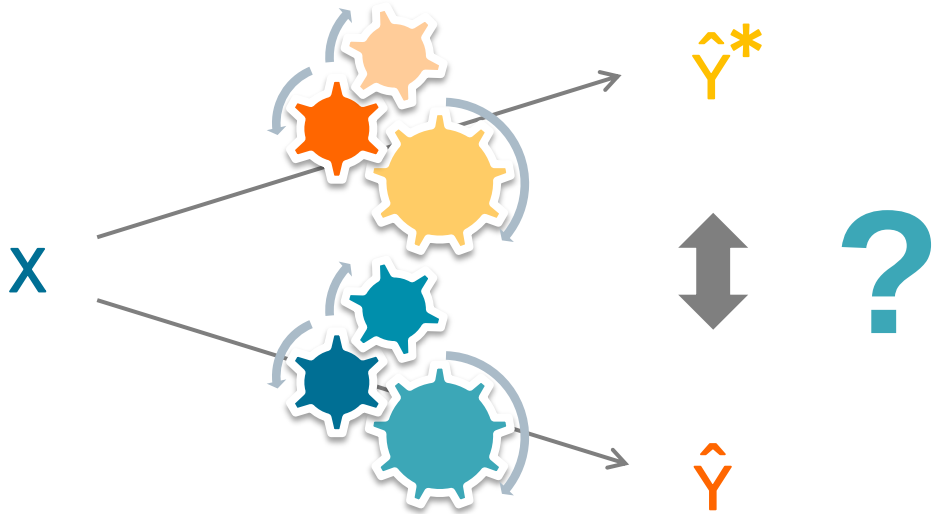
行番号	説明変数1	説明変数2	説明変数3	説明変数4	目的変数	予測値	
1	4.7	3.2	1.3	0.2	タイプA	タイプA	✔
2	6.3	2.9	5.6	1.8	タイプC	タイプA	✘
3	5.5	2.3	4	1.3	タイプB	タイプB	✔
4	7.1	3	5.9	2.1	タイプC	タイプB	✘
5	6.9	3.1	4.9	1.5	タイプB	タイプC	✘
6	4.9	3.1	1.5	0.2	タイプA	タイプC	✘



行番号と目的変数の対応を学習してしまったオーバーフィッティングの例。学習用データでは100%正解するが、他のデータには当てはまらない。

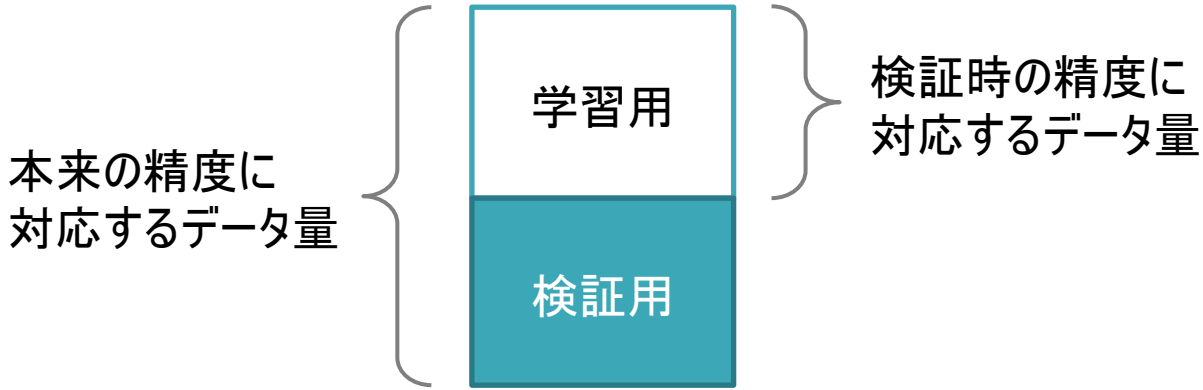
モデル選択

さまざまな予測手法があり、設定によっても予測値は異なります。
予測が目的の場合、基本的には精度の高いモデルを選択します。
検証用データで算出したモデルの予測精度を比較します。



予測精度の過小評価

一般的に予測モデルは使えるデータが多いほど精度が高くなるため、
データを学習用と検証用に分割すると本来の予測精度が得られません。



クロスバリデーション

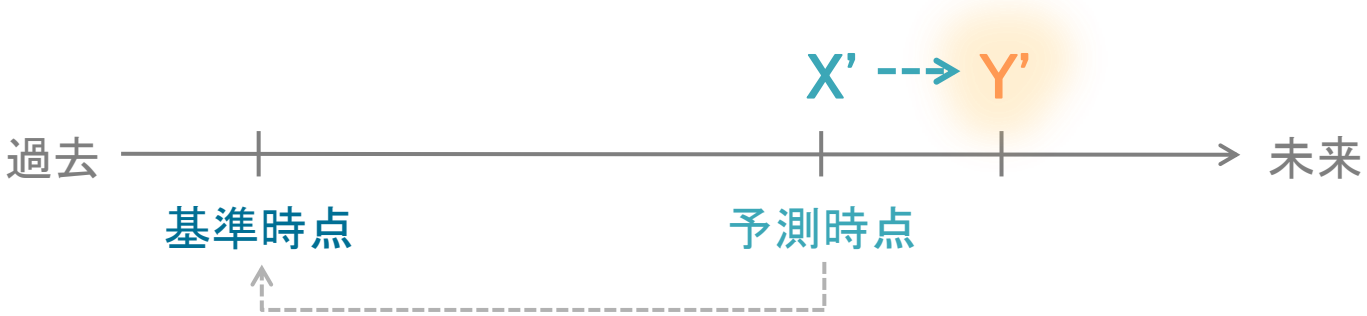
クロスバリデーション法では検証用として使うデータを少しずつずらし、
本来のデータ量に近い状態での検証を実現します。
分割数が多いほど計算時間がかかるため、5～10分割がよく使われます。

データの時点

予測対象が未来の値であるとき、分析用データの作成に注意が必要です。
ある時点（予測時点）で既知の情報 X’ から将来の Y’ を予測するとします。

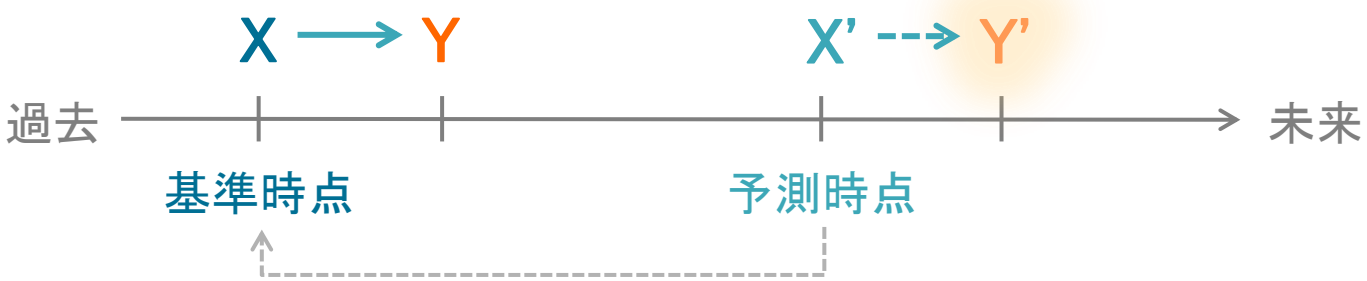
データの時点

これに対応する過去データを取得するため、予測時点から
過去に時間を遡って基準時点（例：1年前の同じ日付）を定めます。



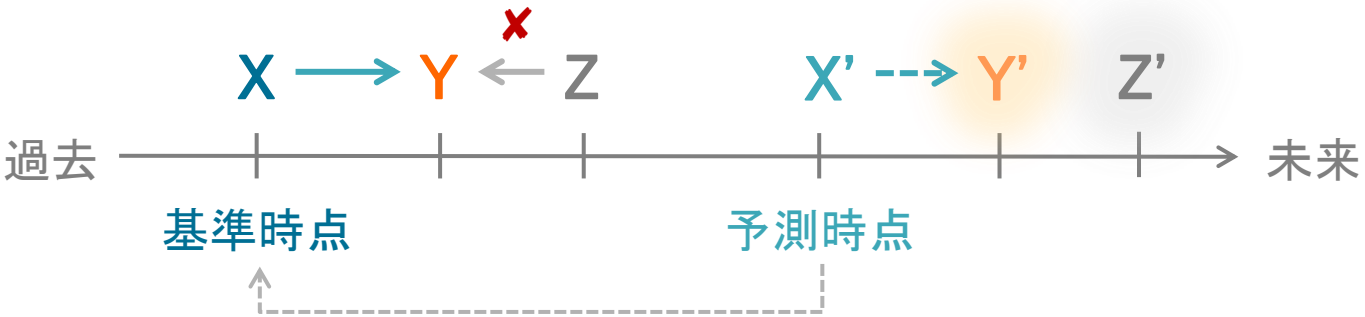
データの時点

当時の時点で既知の情報 X と、将来にあたる時点の Y からなるデータを作成し、X から Y を予測するモデルを作成します。
作成したモデルに予測時点の X' を与えて、将来の Y' を予測します。



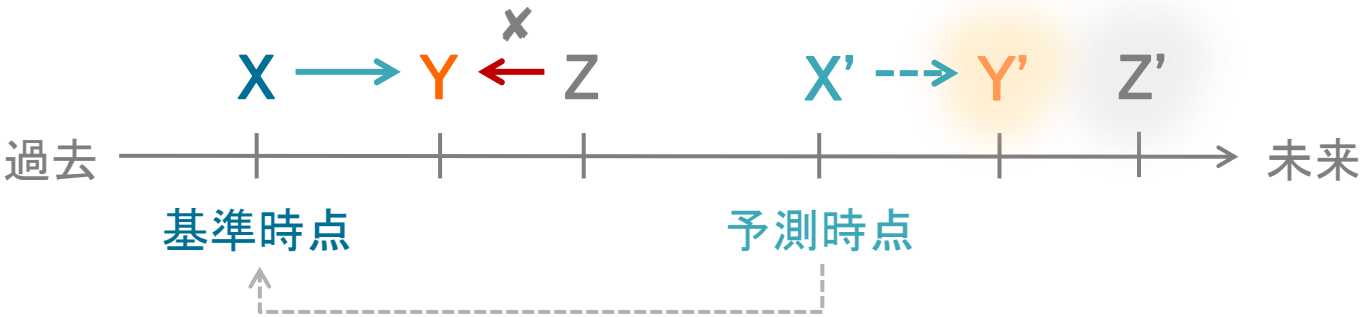
データの時点

分析用データの設計を誤ると、基準時点においては
未知であるはずの情報 Z が紛れ込んでしまう場合があります。
Z に対応する情報 Z' は予測時点で未知なので、使ってはいけません。



データの時点

ところが Z は Y より後の時点で得られる情報であることから、
Z を含むモデルは再現性のない非常に高い予測精度を達成してしまいます。
このような現象をLeakage（リーク、漏洩）と呼びます。



Leakageの例

データの設計時には、「使ってはいけないデータ」が含まれないように
特に注意が必要です。これは到底起こり得ないようで、
実は以下のように比較的発生しやすい現象です。

- マスタテーブルの値が変わってしまう
  - 無料会員登録の時点では存在しなかったクレジットカード情報。
    クレジットカード情報がある場合、全員が購入ありと予測されてしまう
  - 性別。見込みの時点では未登録が存在するが、正式入会時には
    必ず入力する場合、性別未登録は全員が入会なしと予測される

- トランザクションが残らない
  - 購入をキャンセルした場合にフラグが立つが、いつキャンセルしたか
    わからない場合。当時キャンセル済みだったかどうか判別できない

データ分析、成功の秘訣

ここまでデータ分析の基本から応用までを見てきました。
最後に、経験から導き出された成功の秘訣をご紹介します。

成功の秘訣

- 目的を定める
  - 目的が決まれば分析の方針が定まり、
    分析手法や必要なデータも見えてきます。
  - 分析方針に沿って慎重にデータを設計すれば、
    運用時にも高い予測精度が得られるでしょう。

泥沼への招待

- 目的を定めない
  - 目的がない状態で分析を始めれば、ともかく使えるデータを色々と放り込んで使い道のないモデルを作ってしまいがちです。
  - ぱっと見では使えそうなモデルができたとしても、無計画なデータには思いもよらぬ「使ってはいけないデータ」が息を潜めています。
    - 例：過去1年間に一度もサービスを利用していない人にクーポンを付与すると優良顧客になることがわかった
      ⇒ 実際は休眠中の優良顧客にクーポンを付与していた。クーポンをあげたから優良顧客になったわけではない

暫定的なものでも構いません。まずは目的を定めてみましょう。

成功の秘訣

- ビジネスと連携する
  - ビジネス（またはその他の応用分野）の知識を生かし、連携をとることは非常に重要です。
  - ビジネス上の知識があれば、何を予測すべきで、どんな情報が役立ちそうかのヒントが得られます。
    - 家族のすすめで買う人が多いと言われている
      ⇒　家族の利用有無データを取得してみる
  - 分析結果をビジネスに応用する方法が想定できれば、実施すべき分析も変わってくるかも知れません。
    - 高コストな対面営業やダイレクトメールは宛先を絞り込む分析が有効。発送数に対してコストが増加しにくい電子メールなら別のアプローチ

泥沼への招待

- ビジネスとの連携を疎かにする
  - まずは分析してみましょう。実に高い予測精度が得られました。
    なんと商品A購入者の80%が商品Zを購入しています！
    AとZを一緒に売ってみてはいかがでしょう？
    ⇒ プリンタとインクが一緒に売れただけで、そもそも併売している
  - 今度こそわかりました。なんと、電話営業をかけた先の99%が
    次回契約の更新をしています！もっと電話をかけましょう！
    ⇒ 契約を更新した人にオプション商品を訴求する電話だった。
    因果関係でいえばまったく逆

成功と泥沼の狭間

- 話を鵜呑みにし過ぎる
  - 一方、担当者ゆえに気になる細かすぎる情報や要求が本来の成果を妨げてしまうこともあります。
  - 更に、現場で囁かれるビジネス知識には古い情報や勘違い、都市伝説の類も交じっていると思ったほうが良いでしょう。
  - 正しい情報でも、分析上は役に立たないこともあります。
    - この商品は持ち家率の高い地域ほどよく売れるんですよ！
      ⇒ その商品自体の地域シェアを見た方が効率がよい

ビジネスへの応用を意識し、有用な知識は取り入れつつ、
真偽や有効性はデータに照らし合わせて確認することが重要です。

成功の秘訣？

- 精度を徹底的に追求する
  - 予測精度はデータ項目はもちろん、モデル学習時のアルゴリズムやパラメータによって変動します。
  - いま得られている結果に満足せず、精度を徹底的に追及する姿勢が重要でしょうか？
  - 当然、データは使えるものを全て使うべきでしょうか？（300万人の顧客1人たりとも取りこぼしてはならない、等）

成功の秘訣？

- 精度を徹底的に追求する
  - 予測精度はデータ項目はもちろん、モデル学習時の
    アルゴ パラメータによって変動します
  - いま得られ 満足せず
    精度を徹底的に しょうか？
  - 当然、データは使え きでしょうか？
    （300万人の も取り ない、等）

成功の秘訣

- 本質を見定める
  - 予測精度の追及に終わりはありませんが、
    分析に割り当てられるリソースには限界があります。
  - 将来起こることは過去のデータとは少しずつ異なるもの。
    シミュレーション上のギリギリの精度の違いは、
    現実の精度に必ずしも反映されません。
  - 一方、分析方針や評価軸の変更は大きな影響を持ち得ます。
    - 売上優先で販売していたところを利益優先に切り替えたら？
      キャンペーン中の売上ではなく、キャンペーン後の伸びを評価したら？

いま何が一番大切なのか。これを見定めることも重要な仕事です。

成功の秘訣：まとめ

- 目的を定める
  - 目的から逆算して分析方針を定め、正しくデータを設計する
- ビジネスと連携する
  - ビジネス上の知識を生かし、運用を意識した分析を行う
    - 一方で事前情報に惑わされ過ぎず、データに照らし合わせて判断する
- 本質を見定める
  - 分析そのものに拘泥せず、本来の成果を達成することを目指す

株式会社ef-prime　http://www.ef-prime.com
お問い合わせ　contact@ef-prime.com