第 2 章

機械学習を用いたデータサイエンスのプロセス

データサイエンスの分析プロセスには、物事の傾向を明らかにしたうえで、問題解決につながる優れたインサイトを見出すための、いくつかの決まった構成要素があります。

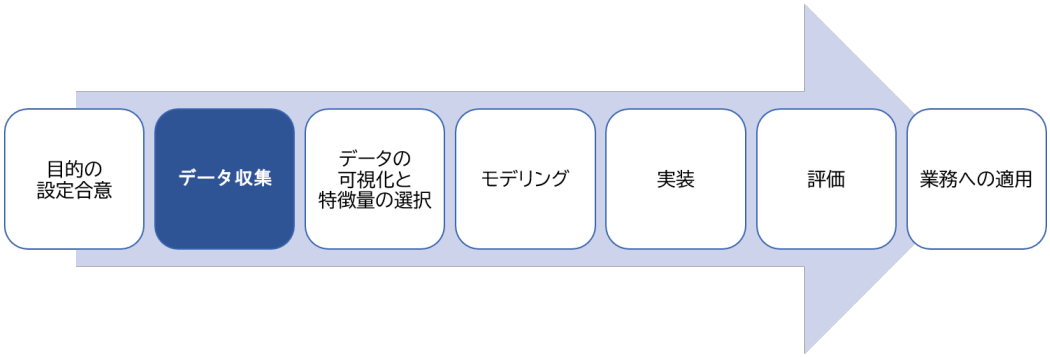
本章では、線形的というよりは反復的なデータサイエンスのプロセスにおいて、それぞれの工程と機械学習の役割や重要性を理解し、実行できるようになることを目指します。

この章のゴール

データサイエンスや AI のプロジェクトにおける機械学習の役割と重要性を理解し、プロセスに従って実行できるようになる。

データ収集

機械学習を用いた学習では、ある程度の量がまとまったデータが必要となります。データがなければ機械学習を行うことはできませんが、ひとことに「データ」といっても目的によって収集すべきデータはさまざまです。データ収集の工程では、分析目的に合致した必要なデータを見極め、それらをどのようにして集めるかを理解していることが重要です。



データの有用性とはなにか

私たちの身の回りは様々な情報が溢れており、それらの多くが、データとして各機関によって収集・蓄積されています。例えば、公共交通機関などを使う時に使用する IC カードは、改札などを通過する際にカードの ID や乗り継ぎなどの乗降履歴などのデータを取得しています。また、街中や店舗に設置された防犯カメラなどの映像もデータのひとつです。このように、現在はデータを取得する側のデバイスの発達も手伝って、人々の生活の中の様々な情報をデータ化し、取得することができるようになりました。

このように集められたデータは、以下の 2 つに分けられます。

一次データ･･･利用者が目的に沿って自ら収集したデータ

二次データ･･･利用目的以外のために、自らまたは他社が収集したデータ

一次データは分析目的に合わせて収集されているため、必要な情報が揃っており、分析に適合しているメリットがありますが、一方で新規にデータ収集を行うため、時間や費用を要するデメリットもあります。一方で二次データは、すでにまとめられた情報が多く収集が比較的容易です。ただし、信頼できるデータ元であるかどうか、必要な情報や加工しやすい形式になっているかなどの点で、分析に取り扱いづらい側面もあります。

また、一次データであっても、取得されたデータはメモ書きのような雑多な状態から、数値のみ集積されたもの、種類ごとにまとめられたものなど状態はさまざまであるため、取得後すぐに機械学習・データサイエンスに活用できるわけではありません。

機械学習・データサイエンスのプロセスでは、こうした精度にばらつきがある状態のデータを、活用できる状態に変換する作業も必要となります。

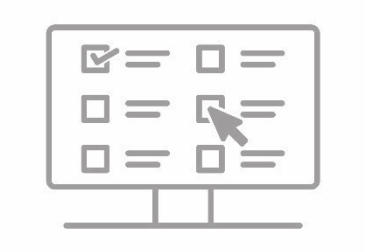
	Suicaの入場記録	廃盤商品の在庫	店舗独自の品番	etc…
A	10	4	60	5
B		35	40	25
C	42	2		8
D	5			30
E	60	65	4	8
F	22	1	10	5

データの取得方法

世の中にあふれる様々なデータを収集するには、いくつか方法があります。例えば、アンケートなどで直接回答を求めてデータを集めることもあれば、公的機関や企業などによって公開されているデータ(オープンデータという)、SNS などの WEB サービスに掲載されているデータを API というプログラムを使って取得する方法などです。

民間企業などでは、自社サービスの顧客の行動などの履歴をログと言われる記録情報として日々データを取得しています。

- アンケート収集



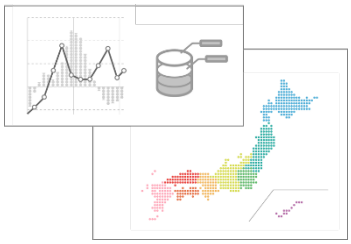
目的に合わせて、適切な対象者に調査したいことを直接問いかける方法です。分析用に新たに取得する場合は、設問を自由に決めることができるため、より分析内容に合致したデータを取得することができます。一方で、設問設計の精度や分析に有用なサンプル数の確保などは、事前にしっかり考慮する必要があります。

- 行動履歴データ

位置情報などによるオフラインでの移動履歴や店舗の来店記録、インターネット・端末上に保存された、オンラインでの商品の購買や WEB サイトの閲覧履歴、検索履歴などがこれに該当します。

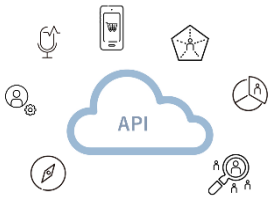
膨大な量となる一方で、得られる情報量は多く、顧客の嗜好性や行動属性を探るためにマーケティング分析で良く活用されています。

- オープンデータ



世の中には、多くのデータがオープンデータとして一般公開されています。代表的なものは、政府が公開している人口統計データや気象庁が公開している降水・積雪などの期間別気象データです。これらのデータは、二次利用が可能な状態に整備されているほか、膨大な量を確保できていることが多く、機械学習を使った分析対象データとしては有効なことが多いです。

- API を使った WEB 情報の取得



API とは、アプリケーション・プログラミング・インターフェースの略称です。ソフトウェアを公開しておくことで、外部からのアクセスを可能とします。オープン型の SNS などで公開されていることが多く、テキストデータ、画像データ、音声データなど様々なデータが外部から取得可能です。

有効データの見極め方

取得されたデータを分析し、問題解決や新たな課題発見につなげるためには、取得されたデータを「信憑性」「量」「偏り」などの項目で評価することが必要です。

➢ データの信憑性

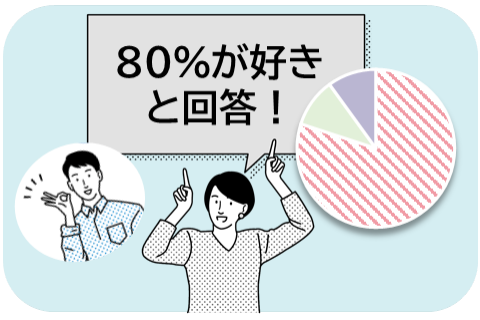
示されたデータが、信じるに値する科学的根拠となり得るかどうかを判定します。信頼できる機関によって取得されたものかどうか、集計・分析済のものであればその方法は正しいかどうかなどを確認する必要があります。

➢ データの量

収集されたデータを使って、全体の傾向を把握したり、示唆を得る上で、データ量が十分かどうかを判断する必要があります。データは多ければ多いほど良いというわけではありませんが、あまりにデータが少なすぎると、母集団に対する推定に誤差が生じ、解釈やその後の意思決定が事実にそぐわない危険性があります。

➢ データの偏り（バイアス）

データを収集する際は、収集の対象となる**母集団**について、標本の大きさや標本の抽出方法を選択する必要があります。これらを考慮せずにデータを収集し分析を行った場合、不正確な分析結果を招いてしまう原因となります。取得方法や対象者が恣意的に選定されていないかなどの確認が必要です。



偏りが出ない方法のひとつとして、母集団全体を調査する**全数調査**があり、総務省が 5 年に一度行う日本国内の全国民を対象とした国勢調査が代表例です。全数調査が難しい場合には一部の抽出を行いますが、抽出方法には、母集団の様々なカテゴリから適切な割合を観ながら調査対象となる標本を抽出する有意抽出と、特定条件を持った対象者をランダムに抽出する無作為抽出とがあります。

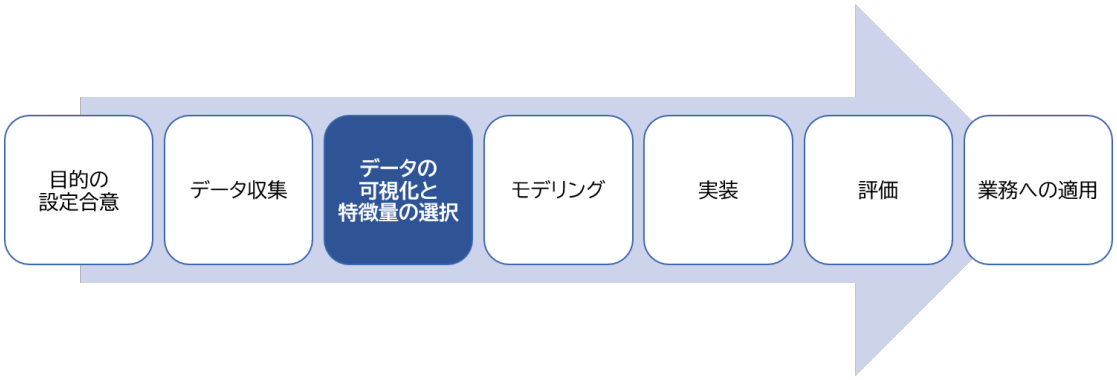
➢ **欠損データ**

何らかの理由により取得ができなかったデータが存在する場合を指します。データは常に全ての項目が揃っているとは限りません。アンケートなどでも未回答などが発生すると、データが取得できない項目が発生します。精度の高い分析を行うためにはデータが揃っていることが望ましいですが、機械学習のアルゴリズムでは、こうした欠損値に対して補完をする手法も存在します。

データの可視化と特徴量の選択

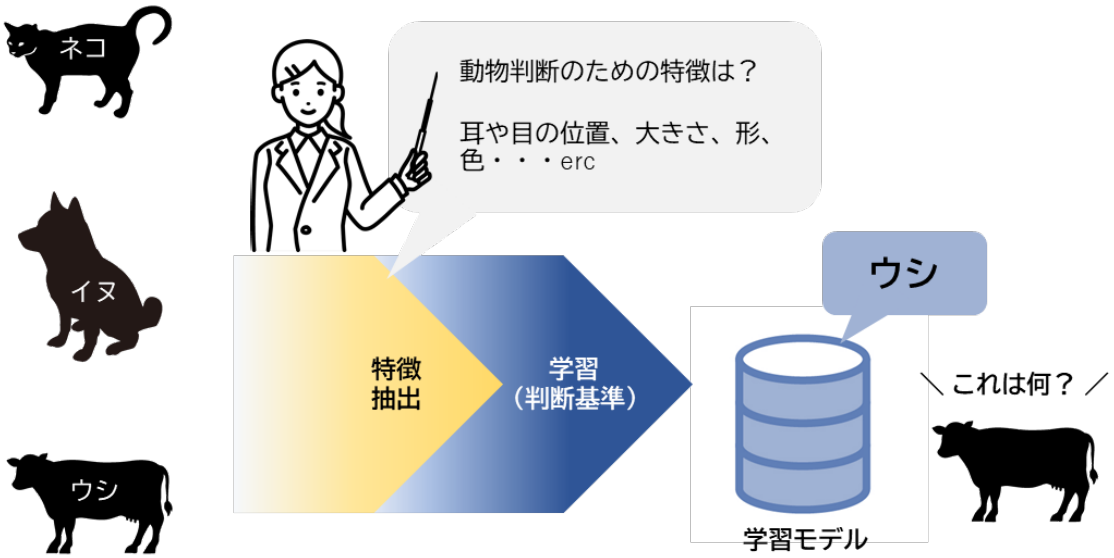
集めた膨大な量のデータについて機械学習を実行する上で、収集されたデータをそのまま使用することはできません。事前に、そのデータの傾向や特徴を把握するための情報加工が必要となります。

まずは、どのような項目を含むデータなのかなどの基本状態を確認し、その後、機械学習に使用すべき特徴量の選定を行います。



特徴量とは

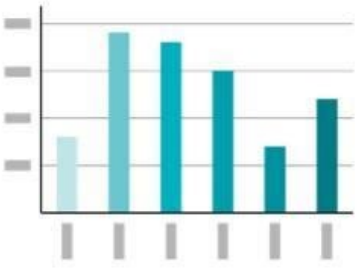
特徴量とは、分析のためにそろえたデータセットの中から選定された、機械学習に読み込ませるために必要なデータのことを指します。特徴量は、コンピュータが読み込み可能な状態に数値化されている必要があります。

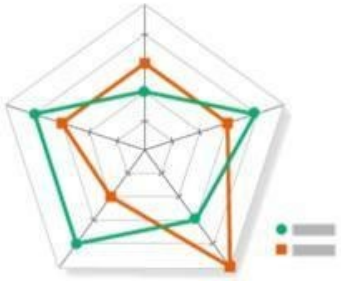


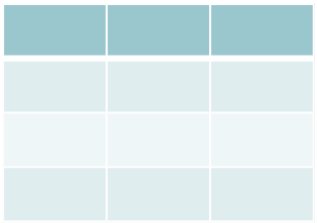
データの可視化

特徴量の選定を行うためには、用意したデータセットについてグラフや表などを使用して、数値データを一目見て状態を把握できるように可視化する必要があります。数字の羅列だけでは、その関係性や他の数字と掛け合わせて分析を行うには多くの時間がかかります。

以下のようなグラフや図、チャートなどを用いてデータの特徴を把握しやすくすることを**データの可視化**といいます。







➢ 可視化の手法

グラフを使用したデータの可視化

データ可視化において最もよく使われる手法はグラフ化です。

グラフにはいくつかの種類がありますが、自分がどのようなことを伝えたいか、目的に応じて適切なグラフを選ぶことで、説明力が高まります。
以下に、いくつかのグラフの種類とそれぞれの用途について説明します。

・棒グラフ
棒の長さで、量の大小を比較するときに使用します。

・積み上げグラフ
項目ごとの集計値の全体に対する割合と、全体の合計値を比較するときに使用します。

・折れ線グラフ
量が増えているか減っているか、変化をみるときに使用します。

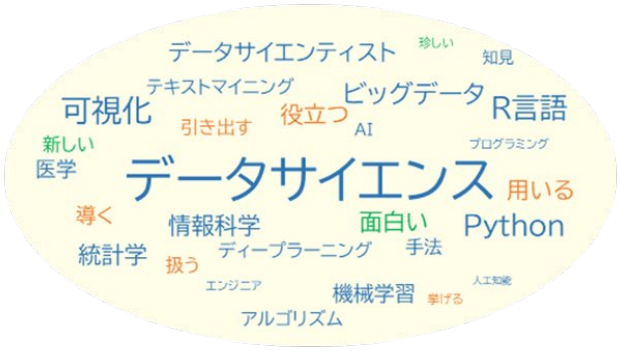
・レーダーチャート
複数の指標について、まとめて 1 つで確認したいときに使用します。

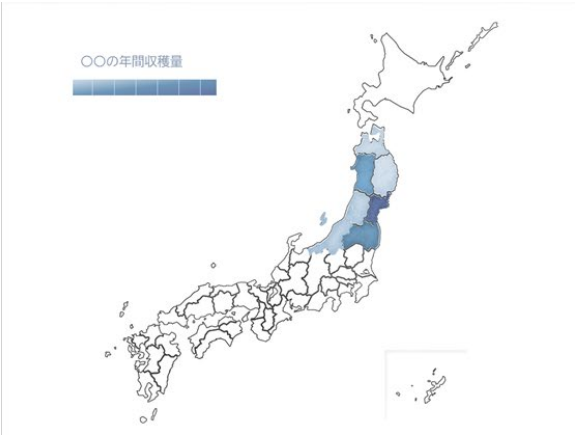
・円グラフ
複数の項目について、全体の中の構成比を見るときに使用します。

・散布図
二項目の量や大きさについて、座標上にプロットされたデータ。二つの変数の関係性や、データの密度を把握したい場合に使用します。

・色や面積で表すデータの可視化

表示サイズの大小で割合が一目でわかるようにしたり、色の濃淡によって差を表現する方法です。グラフなどでもよく利用されますが、テキスト表現やイラストを使った見せ方も可能です。





モデリング

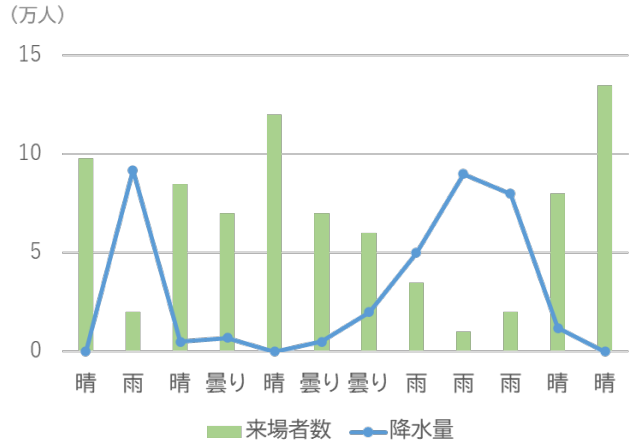
モデリングとは

モデリングとは、収集したデータをコンピュータが分かる形の形式でコンピュータに入力し、機械学習のプログラムを実行して結果を出力するプロセスを指します。「モデルを作る」「モデリングを実行する」などとも呼ばれています。

教師あり機械学習のモデリング

教師あり機械学習におけるモデリングの成果は、機械が発見した法則やパターンから識別や予測を行う数式・アルゴリズムです。これに新たに未知のデータを投入することにより、その性能を検証することができます。「予測モデル」「識別モデル」など、機械学習の目的に応じて呼びわけられることもあります。

※モデル式の例:イベントの来場者数と天気データの関係から得られた予測モデル



教師なし機械学習におけるモデリングの成果は、実行結果そのものです。データを似ている者同士に分割した分割数や、分かれ度合いなどを分析者が解釈します。

ミニコラム:「間違えた理由」を知っているのは、人間だけ

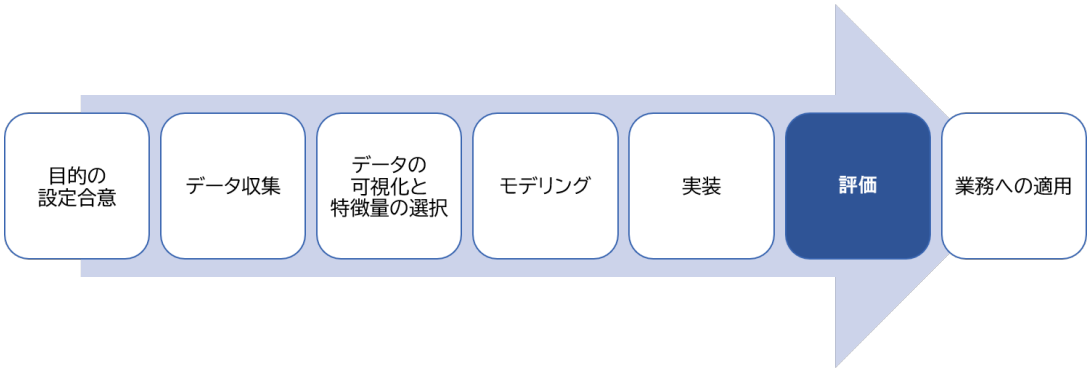
様々な物事を高い精度で予測することができる機械学習のモデルですが、モデルの精度がどれほど高くても、ミスや事故を完全に防ぐことはできません。何かアクシデントが起きたときには、原因究明とモデルの見直しが必要となります。その際に、モデルがどのようにして結果を判断したかは後から究明できるようにしておく必要があります。

予測モデルそのものが、誤った原因を究明することはできず、そのモデルを選定した背景や説明性は、人間だけが把握することができます。

モデルを選定する際は、導き出された結果の値だけでなく、「人が説明を受けて納得ができるものか」という視点も重要になります。

評価

モデルができ上がったら、そのモデルが業務に適用できそうなものかどうか確認するために評価を行います。具体的には収集したデータを2つに分割、一方のデータ群でモデリングを行い、もう一方のデータ群でモデルの評価を行います。データ量が不十分な場合は第三章で後述するホールドアウト法やk分割法を用いてモデリングと評価を行います。



評価とは簡単に言えば「作ったモデルが実際の業務へ適用可能なものであるか判断すること」です。いかなる手法においてもモデルの評価をすると、その結果の数値が出力されます。結果は自動的に出力されますが、その数値を見て適用できるかどうかを判断するのは人間です。どの手法においても予測と実測値の差が少なかったり、判別した結果が合っている方が適合しているモデルということになりますが、手法や取り組んでいる課題によって適用できるかどうかの閾値は異なってきます。

	本当は陽性	本当は陰性
予測が陽性	【A】 レントゲン画像を見て予測した結果が陽性で実際の検査結果も陽性	【B】 レントゲン画像を見て予測した結果が陽性だが実際の検査結果は陰性
予測が陰性	【C】 レントゲン画像を見て予測した結果が陰性だが実際の検査結果は陽性	【D】 レントゲン画像を見て予測した結果が陰性で実際の検査結果も陰性

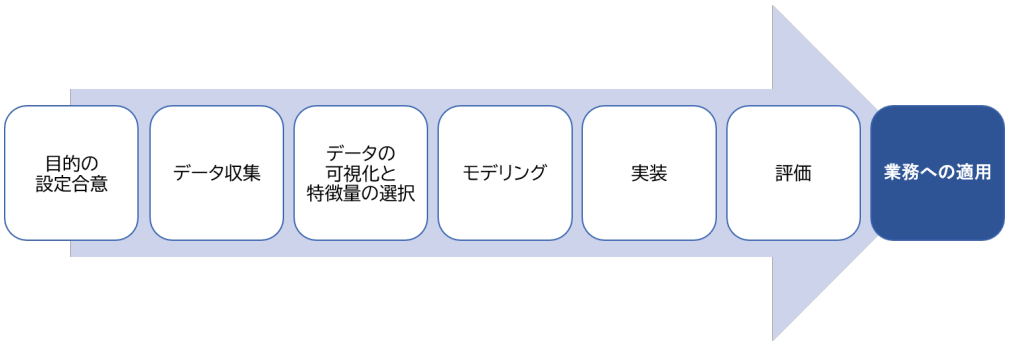
予測と実際の結果が合っている　予測と実際の結果が異なっている

例えば予測を行う機械学習の場合、その結果は上記の図のような四象限に分かれます。【A】と【D】のように予測と結果が一致している場合は問題ないのですが、【B】や【C】のように、予測と実際の結果が異なるところに含まれるデータが存在した場合は慎重に考える必要があります。レントゲン画像を見てある病気にかかっているかどうかを判別するようなモデルを評価する場合、【B】のように予測では陽性だが実際の検査結果が陰性の場合は罹患していなかったということで安心できますが、【C】のように陰性と予測されたが実は陽性だった場合は病気の見落としが発生して大変なことになってしまいます。

取り組む課題によっては多少【C】にデータが含まれても問題になりにくい場合もありますが、人の命が関わるような課題では慎重になるべき場面もあるなど、手法や課題によって閾値やモデルを適用するかどうかの判断は変わってきます。

業務への適用

前項の工程を経て機械学習のモデルを作成したら、次はいよいよ実際の業務への適用を行います。具体的には、既存のシステムなどに組み込んで、新しく取得されるデータに対して、モデルを適用して予測や判別を継続的に行うことができる環境を作ります。



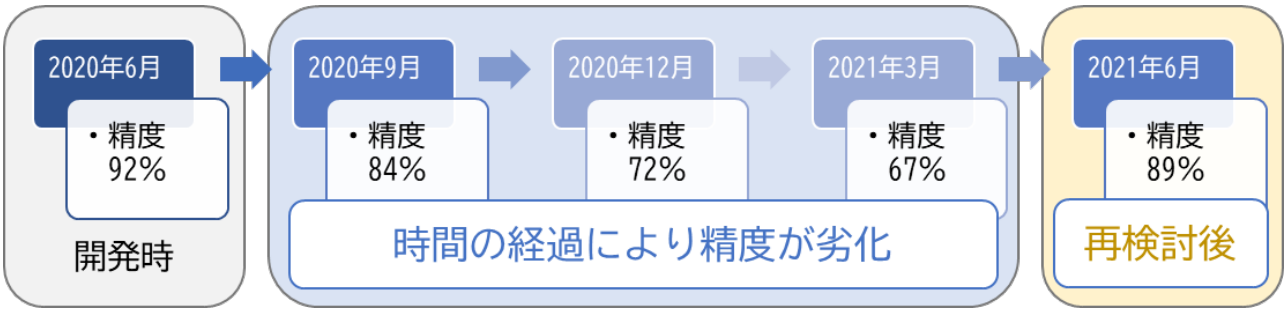
効果検証

データ上でのモデルの当てはまり具合を確認するとともに大事なことは、もともと設定していた課題が解決されているか、その改善度合いを確認することです。コンピュータ外で行われることですが、そもそも業務改善のために行っていることですのでこの確認は重要です。効果が感じられるのであればこのモデルは必要ですし、そうでなければモデルの再検討や違うアプローチを考える必要があります。

モデルの再検討

システムに実装したあと、適用したモデルの当てはまりがよいかどうか、もしくは判別結果が実際の結果と合っているかは常に確認することが重要です。当てはまりがよくない場合は、再度モデルを作り直すことやアプローチ手法の変更を検討することも考えなくてはなりません。実態に合わない判別結果や、予測の精度が低いままだと、モデルを作った意味がなくなってしまいます。

また、実装したてのときにはモデル精度が良かった場合でも時間の経過により適合しなくなってしまう場合があります。これを**モデルの陳腐化**と呼び、時間の経過とともにデータを取り巻く環境の変化や季節、時代が移り行くことなどが原因と考えられます。



モデルの精度が悪くなってきたらその状況を鑑みて要因が何であるのか仮説を立て検証を行い、再度「現在」の状況に合ったモデルを再度構築することを検討する必要があります。

モデリングを行い業務へ適用することが1つのゴールではありますが、意味のあるものとして使い続けるためには業務へ適用した後もモデルが状況に合ったものであるか確認し続けることが必要であり、そうすることによって意味のある運用が実現可能となります。

コラム： 膳所高校野球班が実現したデータサイエンス

著者
青山学院大学 経営学部 准教授　 保科　架風

第90回記念選抜高校野球大会に出場した滋賀県立膳所高等学校の野球班（※ 以下 ”膳所高野球班”，膳所高校では部活動ではなく班活動という名称を利用）は、企業でも難しい「データ分析による問題解決」を意思決定に活かしました。膳所高野球班では以前から監督の指導の下、合理的な判断をするチームでした。

例えば投手のコントロールが悪ければ「ストライクゾーン真ん中低めだけを狙う」という意思決定を行い、また、野手の守備範囲が狭く難しい打球が取れなければ「打球が飛んでくる場所に野手を配置し、可能な限り打球を取りやすい状況を作る」というものです。

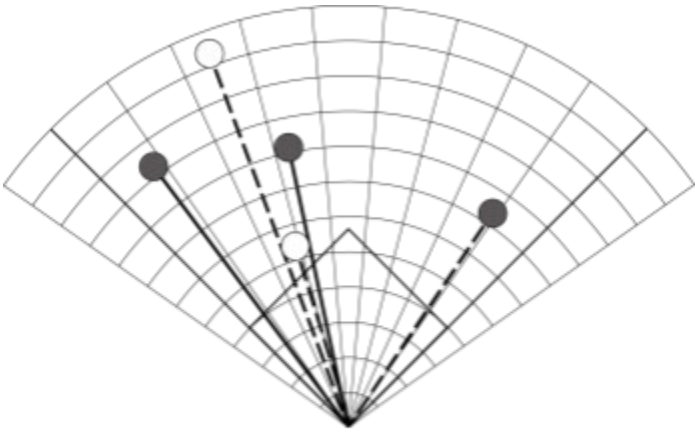
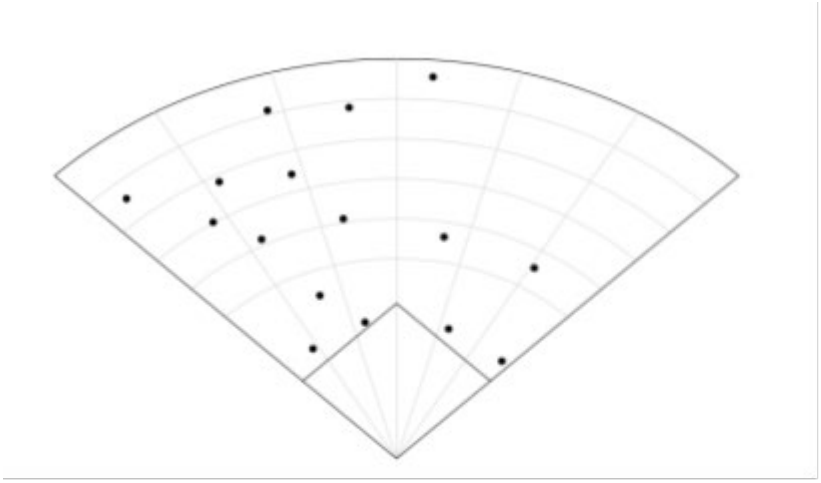


図1. 打球傾向レポートのイメージ
打球の落ちた位置のマーカーや線の種類によってどのような打球だったのか，その打球はヒットになったかアウトになったのかなどを表現している．

しかし、この野手の守備位置の決定のためには、そもそも打球がどこに飛んでくるのかを予測しなければいけません。そこで膳所高野球班のデータ分析チーム”データ班”の生徒が対戦相手の試合や試合映像を見て、全ての打者の打球情報（打球の飛んだ位置、打球の種類、打ったボールの種類など）をフィールドが描かれた紙に手書きで記録し、そこからさらに「打球傾向レポート」を Excel で作成して守備位置の決定に活用していました。

この打球傾向レポートにはデータ班の生徒が手作業で作成したグラフなども含まれており（図 1）、ミーティング時にそのレポートを選手の人数分印刷しなければならないなども含め、データ班の生徒の負担はとても大きくなっていました。

このような状況で私はこのデータ班によるレポート作成作業の省力化をお手伝いさせて頂きました。具体的には統計解析ソフト R を用いて、紙に記録していた打球情報を PC 画面で入力できるようにし（図 2）、そこで入力されたデータや打撃成績のデータを結合させて自動でレポートを作成する Shiny アプリ（※ R の計算を Web ブラウザ上から実行できるアプリケーション）の制作を手伝いました。これにより、「手書き記録・手入力・手作業でグラフ化」からデータ班の生徒たちは開放され、ミーティング時のレポート印刷もファイルの共有と簡素化できるようになりました。



x	y
-17.2855692276877	27.4633634660568
-6.52669941084021	34.0997689524545
-15.8561008100798	40.7774285135707
10.7435201896025	32.4276037905275
21.7530012597488	24.3729748910262
9.80921720902617	55.4143742237275
-37.9235195069015	59.1189901334571
-55.8365456150503	64.9944871306007
-26.7043124030278	87.0828354921514
-21.6966301392256	71.0973399173046
7.52873051737631	95.4454948162181
-9.81114538307696	87.8006563922549
28.4566630020689	47.6760265637354
-11.0157450930421	59.9679073154442
-27.9089121129929	54.8294831199394
-36.6545634327962	69.209736809625

図2. R を利用した打球の位置情報の入力イメージ
左側のように画面のフィールドの図上に打球の位置をマウスでクリックすると、その座標情報が右側のように記録される。

ただ、膳所高野球班のデータサイエンスはここでは終わりませんでした．最大のポイントは「選手がレポートから守備位置を判断する」ということです。一般に、データ分析をうまく活用できない組織では、現場とデータ分析の部署それぞれのお互いに対する理解不足が存在することがあります。それによりせっかくデータ分析をしても問題解決に活かすことはできず、また実施したデータ分析もどこか現場から見ると実態と乖離している印象を持たれてしまいます。これに対し、データ班の生徒は可能な限りそれまでミーティングで使用していた打球傾向レポートと同じレイアウトのレポートを出力するように Shiny アプリを改良しました。これにより、選手たちはそれまでのレポートと同様にそこから意思決定をすることができるようになり、現場でデータ分析結果を活用し、価値を生むことができたのです。

また、膳所高野球班のデータ班が行った「グラフで打球傾向を表現する」という分析はデータ分析の観点からでも一定の妥当性が存在します。例えばこの守備位置の決定という課題に対し、フィールド上のどこにどのぐらいの確率で打球が飛んでくるのかをカーネル密度推定やニューラルネットワークなどで分析するという手段もあります。しかし、打者ごとにそのフィールド上の打球確率を推定するには多くの計算コストが高くなり、また高い推定精度を出すには打者ごとに莫大な量のデータが必要であり、それを高校野球で行うことは現実的ではありません。一方、データ班が行ったグラフ化という分析は、手法こそ簡単なものではありますが、ニューラルネットワークのようにデータへの過適合を心配する必要はありませんし、コードを組んでしまえば作業・計算コストも極めて小さくできます。

これらのデータ班を含めた膳所高野球班のデータサイエンスにより、第90回記念選抜高校野球大会では極端な守備位置を取る膳所高野球班と、相手チームの強烈な打球がその守備の正面に繰り返し飛んだことが話題となりました。残念ながら試合に勝つことはできませんでしたが、「効果的な守備位置を決める」という課題に対して膳所高野球班が実施したデータ分析に基づく意思決定には一定の正しさがあったことが確かめられたのです。

このように、膳所高野球班のデータサイエンスは泥臭くデータを集め、グラフを作成し、選手が理解しやすいレポートを作成するという、決して高度な取り組みではありませんでしたが、データ分析による問題解決は十分に実現できるものでした。また、私も少しお手伝いさせて頂きましたが、実際には「こんなやり方があるよ」と R を使った省力化を紹介し、少しだけやり方を教えただけで、あとはデータ班の生徒がほとんど独力でアウトプットを設計し、コードを組み、実際に運用までしてしまいました。データサイエンスと聞くと分析手法やプログラミングに着目してしまいがちですが、データ班の生徒が行ったようなこの現場で活かすための工夫も大切ですし、それは高校生でもできることです。是非、問題解決につながるデータサイエンスを実践してみてください。

本コラムの著者

【プロフィール】

保科　架風(ほしな　いぶき)

2017 年　中央大学大学院理工学研究科博士課程後期課程修了(博士(理学))。2017 年より滋賀大学データサイエンス教育研究センター助教として企業との共同研究やデータサイエンス教育に従事し、2019 年より青山学院大学経営学部准教授に着任。専門は統計的モデリング。