“DataPrep”

機械学習と
データ・プレパレーション

予測モデルの精度を上げる10のヒント

予測モデルの精度を上げる10のヒント

すぐれた機械学習プラットフォームの登場によって、 AI・機械学習へのチャレンジはこれまでになく身近になり、あらゆる事業部門が最新のテクノロジーを簡単に利用しながら、未来を予測できるデータの使い方ができるようになってきました。

ところが、機械学習の取り組みが進むほど「学習させるデータによって予測精度が大きく上下する」「予測分析テーマを繰り返し検証したいのに、学習データの準備に時間がかかり検証が限られてしまう」といった課題が顕在化し始めています。

本資料では、このような課題の解決策として注目される「データ・プレパレーション」をテーマに、Paxata（パクサタ）でのデータ準備の手法をわかりやすくご紹介します。

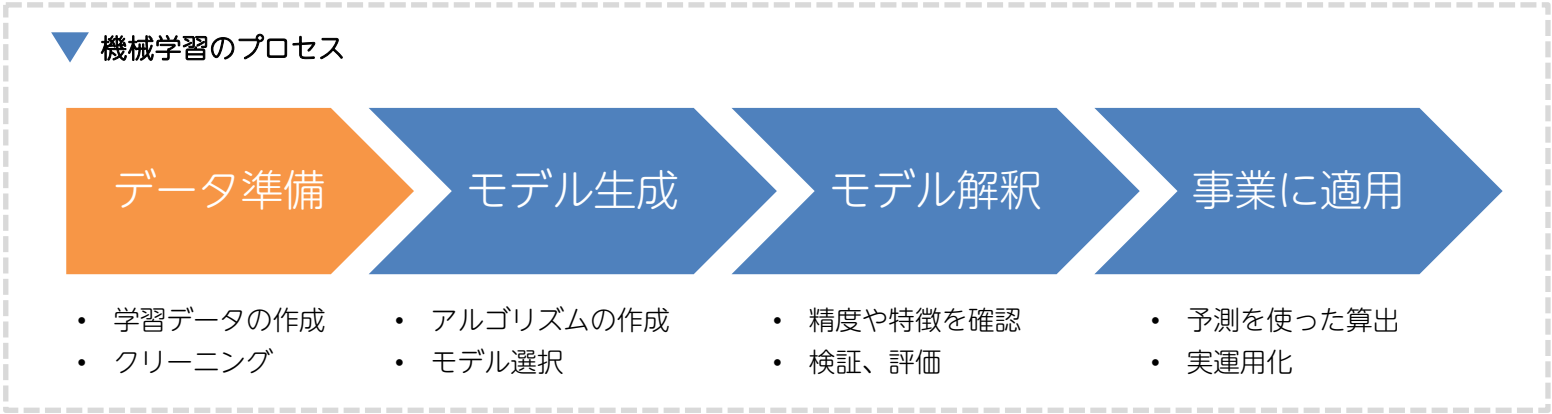
超サポ 愉快カンパニー アシスト

DataPrep 1

機械学習のスタートは「データ準備」

AIの民主化により、これまではデータサイエンティストに限定されていた機械学習が、ビジネス部門のユーザーにも広く利用されるようになりました。難しい数学や統計学をまったく意識することなく、最先端のアルゴリズムに基づいた予測を誰もが手に入れられるようになっています。

機械学習というと予測モデルの生成に主眼が置かれがちですが、機械学習のプロセス全体では、モデルの生成はその一部であることがわかります。ここでは、機械学習のスタート地点である「データ準備」に焦点を当てて、機械学習を成功に導くポイントを探っていきます。



DataPrep 2

機械学習におけるデータ準備とは

“機械学習では、何度も何度もデータを作り直さなければならないので、”
いいモデルを作るためにはデータ準備は非常に重要なステップです。

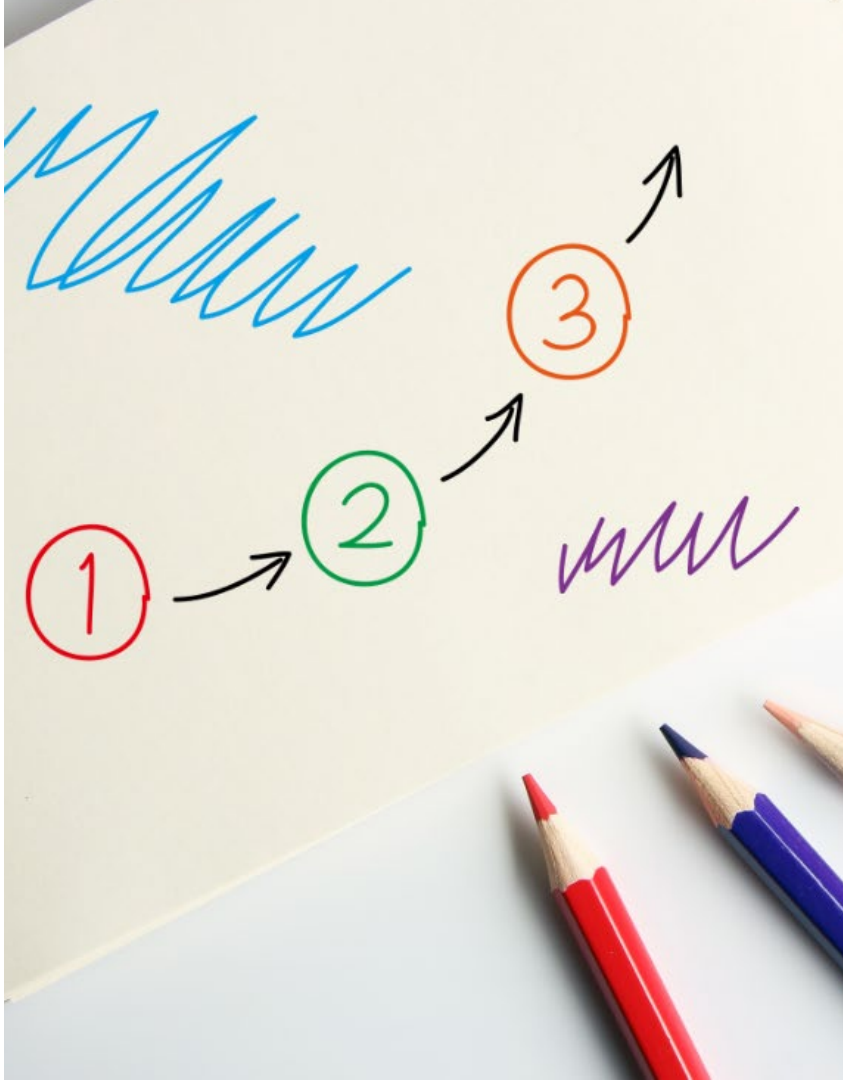
データ準備の手間や時間を減らせれば、モデルの生成もより簡単になり、
モデルの精度が高ければ、モデルを検証する時間も短くなります。

— DataRobot チーフデータサイエンティスト シバタアキラ氏 ※

機械学習では、元となるデータのクリーニングや欠損、変換などにすべて対応して、ようやくアルゴリズムにデータを入れられるようになります。

学習データをもとに予測モデルが構築され、その予測が事業に導入されて、意思決定や行動を変えていくインパクトを考えると、機械学習におけるデータ準備の重要性を理解できます。

※2017年10月20日 アシスト主催セミナーでの講演より

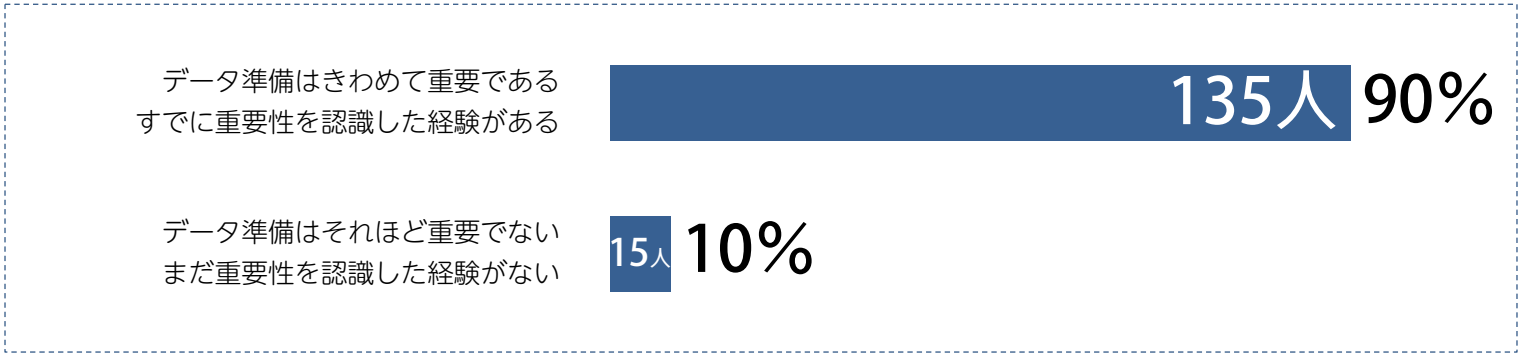


DataPrep 3

90%がデータ準備の重要性を実感！
機械学習の成功のカギは「データ」にある

機械学習の現場では、ユーザーの90%が「事前のデータ準備が重要である」と考えています。
機械学習の実践が長く、より多くの予測分析テーマを検証しているユーザーほど、その重要性を実感しています。

▼「機械学習やアナリティクスでは、事前のデータ準備はどのように位置づけられますか？」



※アシストが2017年に実施したアンケートによる回答結果

DataPrep 4

つまづきやすいデータの準備

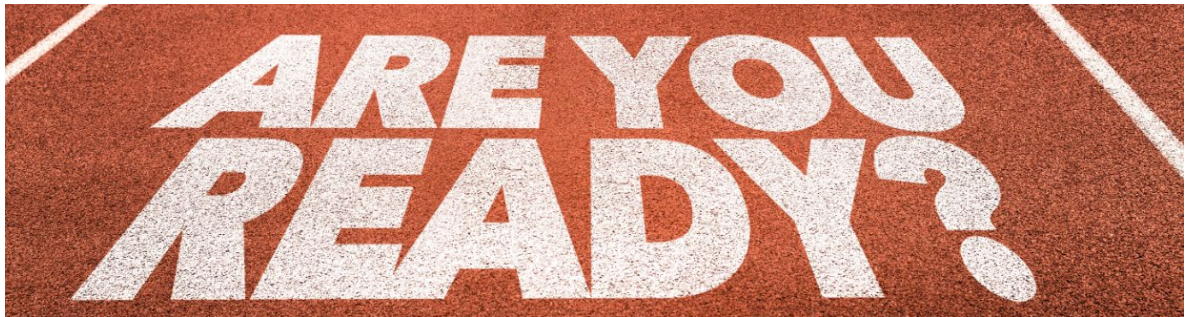
機械学習のためのデータは、重要性が認識されながらも、その準備は簡単ではありません。
すぐれた機械学習プラットフォームによって、モデルの生成プロセスは自動化が進む一方、元となるデータは私たちが手動で準備しなければならないからです。モデルの生成スピードが上がるほど、データ準備にかかる時間が機械学習のボトルネックになってきています。

機械学習にデータを入れると、
わずか数分でモデルを生成できるのに、
そのデータ準備に数10時間かかる

ExcelやPythonでの
データの加工はもう限界

センサーの時系列データは
コーディングでしか
整形できない

高度で複雑なデータ加工は
専門スキルを持った人しかできない



DataPrep 5

なぜデータの準備は難しいのか？

機械学習では、データに含まれる値からパターンを検出し、予測モデルを構築していきます。どの値がインパクトを与えているのか、要因や特徴量はすべてデータから導き出されるため、足りないデータがあれば追加し、不要なデータは削除する作業を繰り返していきます。

誤差の無いデータを準備するには補正やクリーニングを行い、よりリッチなデータへと強化するには複数のデータソースを組み合わせ、名寄せや計算を行いながらデータセットの準備を進めます。

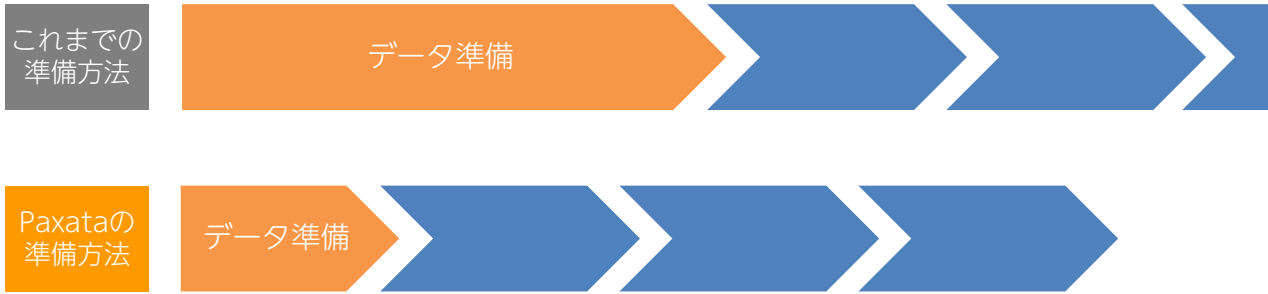
ビジネス部門のユーザーは事業データの理解を持ち合わせていますから、生成された予測モデルのアルゴリズムを読み解き、次に必要なデータが何かがわかります。しかしこの作業を自分で行うには、依然として高度なITスキルが求められるのです。

DataPrep 6

機械学習のスタートボタンをもっと早く押すには

機械学習のプラットフォームを導入して、これまでにないほど早く多くの予測ができるようになったなら、データ準備における時間のロスは大きな機会損失を招いてしまうかもしれません。

機械学習のスタートボタンをもっと早く押すには、これまでのデータ準備の方法を疑い、新しい方法に置き換えることで、従来は当たり前のようにかかっていた時間と手間を圧倒的に短縮することができます。



DataPrep 7 これまでのデータ準備の課題をクローズアップ

機械学習のデータ準備には、これまで主に3つの方法が採用されてきました。
注目すべきは、AIの民主化によって、ビジネス部門のユーザーが自分たちで機械学習に取り組めるようになってきたにもかかわらず、
データ準備は従来のままExcelで行われるか、データサイエンティストやIT部門、外注先に依頼して必要なデータを準備してもらう選択肢しかなく、機械学習のプロセスにおいてデータ準備の民主化やセルフサービス化が進んでいないことにあります。

Excel

ビジネス部門ユーザー向け

- 扱えるデータ件数に制限がある
- 大量データは分割するか、サンプリングしたデータを用意する
- マクロや関数の記述が必要
- 複雑な加工になると、リテラシーのある人に作業が依存し、属人化しやすい
- 個人のローカルPCでの作業になる
- データ加工の過程がブラックボックス化する
- データマネジメントやガバナンス、セキュリティへの対応が必要

R Python SQL

データサイエンティスト
IT部門向け

- 言語の習得に時間がかかる
- プログラムによって加工内容が定義されるため、可読性が低く属人化しやすい
- 類似の加工定義を別のデータに適用しづらい
- 繰り返しや再利用に不向き

EAI ETL

SIerなどの外注先
IT部門向け

- 高度なデータ加工ができるが、高度なITスキルが求められる
- データを必要とする人と、データを加工する人が別になる
- 加工の依頼と実施にあたり、説明と理解が必要になる
- 要件を決めてからでないと依頼できず、要件定義からデータ提供までに時間がかかる
- 外注する場合はコストが都度発生する

DataPrep 8

Paxataでデータ準備を一新する

Paxataは、機械学習の発展とともに拡大するデータをめぐる乖離を解決できます。データの扱いに長けた事業を知らないIT部門と、データの扱いには詳しくないけれどデータの意味と事業に精通しているビジネス部門のギャップにアプローチできるデータ・プレパレーション製品です。

機械学習に求められるデータ準備は、厳密に定義されたデータウェアハウスに向かって検索や抽出を繰り返しながらデータを作り込んでいく従来型の手法ではなく、まだ整理されていないローデータを見ながら試行錯誤を繰り返せる新しい手法です。

Paxataならこれまでの方法を一新し、桁違いのスピードで、桁違いの仮説や分析テーマを導き出せるデータを即座に準備できます。

DataPrep 9

思いどおりにデータを準備できる
セルフサービス型のデータ・プレパレーション

Paxata

データを必要とするすべてのユーザー向け

- Excelのようなスプレッドシートで、データの読み込みから加工、出力までのデータ準備を行える
- 簡単なクリック操作で、あらゆる形式のデータを扱える
- 加工を選択するだけで瞬時に加工後のデータイメージをプレビューでき、手戻りなく進められる
- 大量データも全件を対象にでき、サンプリングによる精度低下を解消できる
- データ加工の手順を再利用できるため、組織でデータ準備の課題を解決できる
- サーバー環境でデータを準備するため、ガバナンスとマネジメントに役立てられる
- IT部門やデータサイエンティストに依頼しなくても、ビジネス部門のユーザーが自分でデータを準備できる
- データ準備をスピードアップできるため、より多くの予測分析テーマを検証できる

DataPrep 10

もっとチャレンジできる機械学習へ

機械学習はまだまだ新しい分野ですが、もうすでにビジネスの現場で使うことができます。
機械学習の価値は、ビジネス部門のユーザーが現場にある課題を見つけて仮説を立て、
機械学習のテクノロジーを使って検証し、課題を解いてアクションにつなげていくことに
あるといわれています。

予測モデルの生成が自動化された今、機械学習の成功のカギは、ビジネスの現場にいる
ユーザーが事業の知見を活かしながらデータを活用できるかどうかにかかっています。
データ・プレパレーションを新たな手段に、さらなるチャレンジを進めてみませんか？

Paxata

Paxata（パクサタ）は、エンタープライズのデータ・プレパレーションを実現するまったく新しいツールです。

セルフサービスの発想をいち早くデータ・プレパレーションの領域に導入したPaxataは、ビジネスデータに知見を持つ現場のユーザーが自分のほしいデータを自分で準備できる環境を提供します。ユーザーは視覚化されたデータを見ながら、クリックだけのシンプルな操作で、すばやく簡単にノンコー ディングでデータを作成できるようになります。

すべてをサーバで管理するため、セキュリティとガバナンスの効いたプラットフォームと して、企業のデータ・プレパレーションへの取り組みをご支援します。

---

https://www.ashisuto.co.jp/paxata/

超サポ 愉快カンパニー アシスト