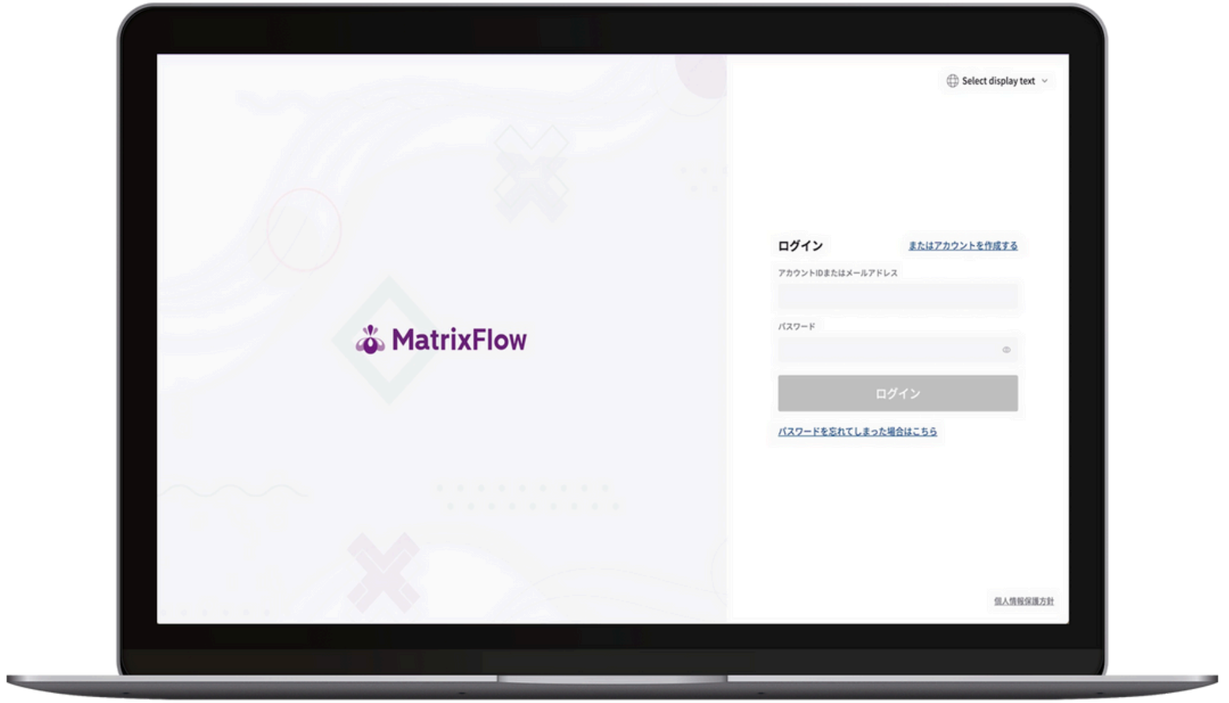
MatrixFlow

生成AIを活用した自動前処理機能
＋
生成AIによる行削除機能
＋
データ結合機能



会社概要

【社名】株式会社MatrixFlow

【代表者】田本　芳文

【設立】2018年10月

【資本金】1.71億円（資本準備金含）

【住所】

東京都台東区上野3丁目16番2号

天翔上野末広町ビル206号室

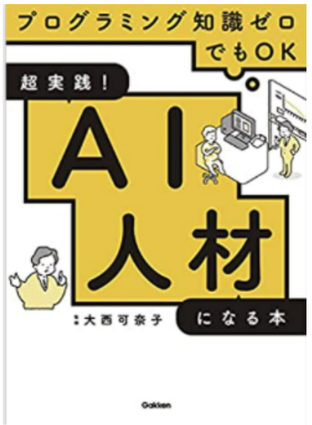
【事業内容】

『MatrixFlow』のSaaSサービス提供

AIの受託開発・研究・コンサル

実績/許認可

・株式会社dipのAIアクセラレータープログラム採択
・Plug　and　Playのアクセラレータプログラム採択
・2020年10月　1億円の資金調達を実施
・セキュリティマネジメントシステム（ISMS）認証取得
・プライバシーマーク（Pマーク）認証取得
・総務省後援ASPIC　IoT・AI・クラウドアワード2021
　　AI部門ニュービジネスモデル賞受賞
・2022年4月　資金調達を実施



掲載メディア

テクノロジーで世界をつくる

第一弾

AIの民主化を推進

一部の企業や技術者がAI技術を独占するのではなく誰でもAIが活用できる社会を実現

MatrixFlow

ビジネスのための AI活用プラットフォーム

プログラミングなしで、
短期間でAIを構築・活用し、
幅広い課題を解決します

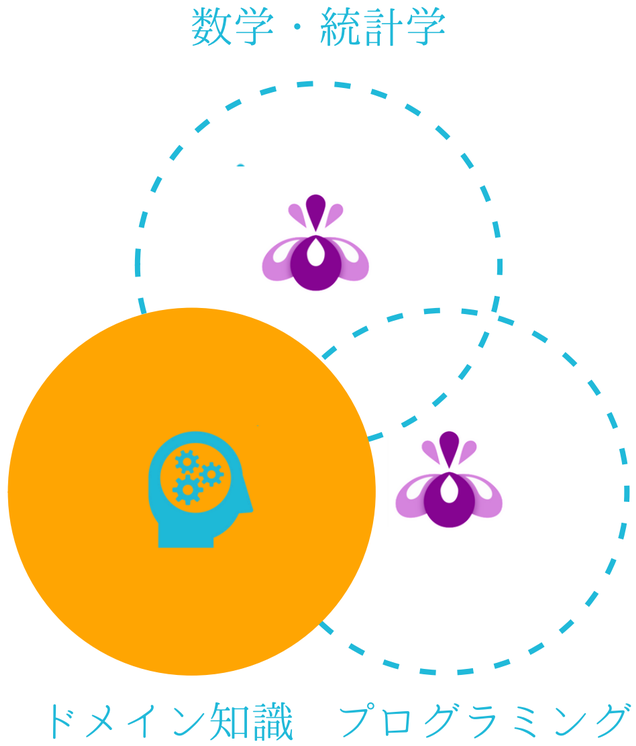
数学・統計学，プログラミングの知識は一切不要

プログラミングすることなく（No-Code），
データサイエンスなどの高度なスキルがなくてもAIを構築・活用できるSaaSサービスです。

MatrixFlowがない場合

AI化には多くのスキルが必要

MatrixFlowがある場合



ビジネスの知識だけあればよい

本日の説明会要旨

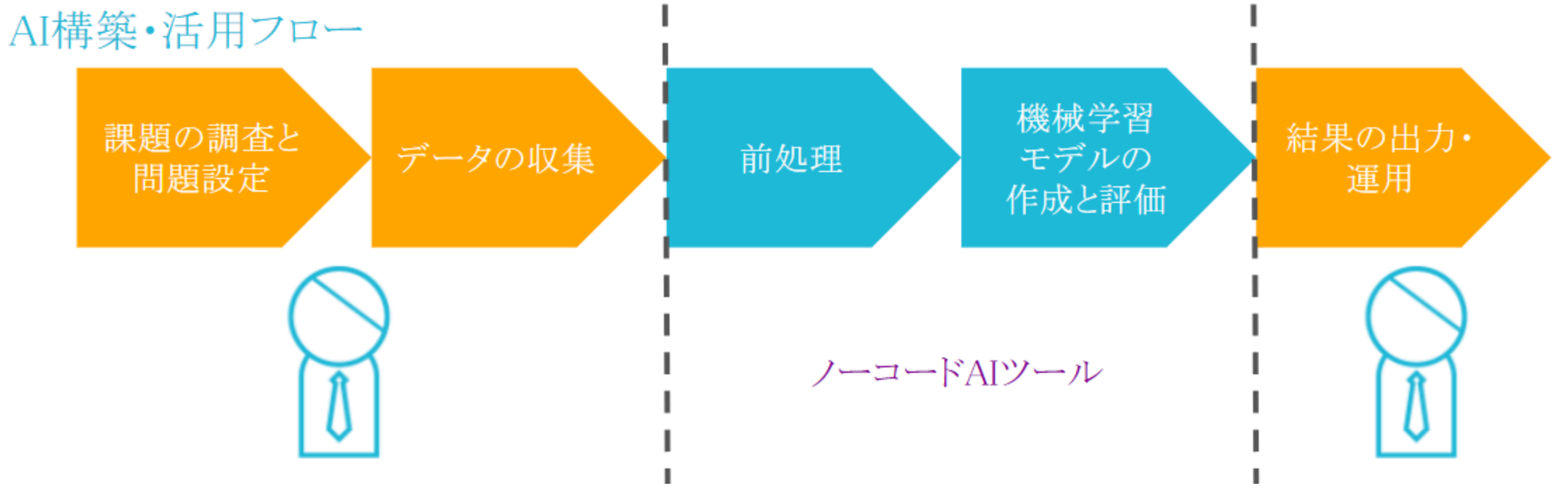
はじめに

これまで、AIモデルのアルゴリズム選択やハイパーパラメータのチューニングは自動化が進んでいましたが、データの前処理は自動化が難しい領域とされてきました。

データ前処理は、データの構造や意味を深く理解した上で行う必要があり、これがデータサイエンス業務全体の8割を占めるとさえ言われています。

しかし、大規模言語モデル（LLM）の登場により、データの構造や意味を理解し、人間のような判断をもとに処理を行うことが可能になりました。

この技術的進化は、前処理を自動化する新たな道を切り開き、データ分析の生産性を飛躍的に向上させる可能性を示しています。



MatrixFlowの前処理新機能

- 生成AIによる自動前処理機能
  - 読み込んだデータを解析し、次ページ以降でご説明する「MatrixFlowがもつ前処理機能」を最適に自動的に実行してくれる機能です。
- 生成AIによる行削除機能
  - 前処理「行の削除」が追加されました。この機能は従来の空白の行を消す機能ではなく、生成AIに希望する「列」と「行」を伝えることで、該当箇所の行を消してくれる機能です。
- データ結合機能
  - データ結合機能は、データ管理にある２つのデータを統合する機能です。対象のデータ種類はテーブルデータ（CSV）です。

生成AIによる自動前処理機能

生成AIによる自動前処理機能

- データクレンジングの自動化
  - 各種エンコーディング、名寄せ、対数変換、単位の統一、損値や外れ値の検出・処理、重複データの削除などのクレンジング処理を自動的に実行し、正確で信頼性の高いデータを生成します。
- 自然言語処理による列名の意味解析
  - カラム名やデータの文脈を理解し、自動的に適切な前処理方法を選択しています。
- 直感的なUIで操作を簡単に
  - データサイエンスの専門知識がないユーザーでも、ワンクリックで前処理を完了できます。
- 前処理の透明性を担保
  - 前処理の実行内容は明示されますので透明性を担保できます。

MatrixFlowがもつ前処理機能

数値に変換する

AIモデルが学習に使えるように、文字列を数値に置き換える

・One-Hotエンコーディング
・ダミーコーディング
・ラベルエンコーディング

欠損値を変換する（欠損値がある列のみ表示されます）

AIモデルが学習に使えるように、欠落している値を補完する

・欠損値を含む行を削除する
・欠損値を（数値で）埋める

列を削除する

外れ値を除去する

測定ミス、記録ミス、データの異常、データの偏りなどに起因した外れ値でAIモデルの学習にゆがみを生じさせるおそれがある値を除去する

データ変換

・正規化：表記ゆれの名寄せ、重複データの削除、データの単位やスケールを共通の基準に整える
・標準化：平均が0、標準偏差が1の正規分布にスケーリングする
・対数変換：Yeo-Jonson変換（負値にも対応した対数変換）を実行する
・離散化：グループにまとめて離散化する

前処理の具体例とその理由

数値に変換する

- One-Hotエンコーディング

  具体例:

  - 「動物」列が　["犬"，　"猫"，　"鳥"]　の場合、以下のように変換

動物
犬
猫
鳥

犬	猫	鳥
1	0	0
0	1	0
0	0	1

  理由:

  - 順序性がないカテゴリ変数を適切に処理し、モデルが順序性を誤解しないようにするため
  - カテゴリ変数をバイナリ特徴量に変換し、各カテゴリを独立した次元で表現する
  - 効果的に機械学習アルゴリズム（特に線形モデルや距離計算に基づくモデル）で利用可能

前処理の具体例とその理由

数値に変換する

- ダミーコーディング

  具体例:

  - One-Hotエンコーディングと同様ですが、基準カテゴリを1つ削除（例:　「鳥」を基準）する

動物
犬
猫
鳥

犬	猫
1	0
0	1
0	0

  理由:

  - 順序性がないカテゴリ変数を適切に処理し、モデルが順序性を誤解しないようにするため
  - One-Hotエンコーディングに似ているが、1つのカテゴリを基準として省略する（kカテゴリに対しk-1次元に削減）
  - 基準カテゴリをモデルの基準に設定することで、多重共線性を回避しやすくなる

前処理の具体例とその理由

数値に変換する

- ラベルエンコーディング

  具体例:

  - カテゴリごとに数値を割り当てる

評価
低
中
高

評価
0
1
2

  理由:

  - 順序があるカテゴリ（例:　評価スコア「低」「中」「高」）に適している
  - 各カテゴリに一意の数値を割り当て、カテゴリ変数を数値変数に変換する
  - 軽量で、カテゴリ数が多くてもメモリ消費が少ない

前処理の具体例とその理由

欠損値を変換する

- 欠損値を含む行を削除する

  具体例:

  - 「年齢」がNaNの行を削除する

名前	年齢	性別
山田	25	男性
佐藤	NaN	女性
高橋	30	男性

名前	年齢	性別
山田	25	男性
高橋	30	男性

  理由:

  - 欠損値の多い行がデータ全体に与える影響を排除する
  - 小規模データセットではデータ不足のリスクがあるが、大規模データでは欠損値の影響を小さくできる

前処理の具体例とその理由

欠損値を変換する

- 欠損値を（数値で）埋める

  具体例:

  - 「年齢」が欠損している場合、平均値（27.5）で埋める

名前	年齢	性別
山田	25	男性
佐藤	NaN	女性
高橋	30	男性

名前	年齢	性別
山田	25	男性
佐藤	27.5	女性
高橋	30	男性

  理由:

  - データ量を維持し、モデル学習に利用できるデータを最大化する
  - 平均値、中央値、最頻値、または他のデータから推定された値で埋めることで、統計的な一貫性を保つ
  - 欠損がデータの一部の特徴を反映する場合（例:　欠損が特定の条件下で起こる）には慎重に処理が必要

前処理の具体例とその理由

外れ値を除去する

具体例:

- 年齢が範囲外（例:　正常範囲を0～120歳とする）の値を削除

名前	年齢	性別
山田	25	男性
佐藤	27	女性
高橋	300	男性

名前	年齢	性別
山田	25	男性
佐藤	27	女性

理由:

- 外れ値はモデルの学習を歪め、性能低下や誤解を引き起こす可能性がある
- 外れ値を削除または適切に処理することで、モデルの汎化性能（モデルが訓練データだけでなく、未知のデータに対してもどの程度適応し、正確な予測ができるかを示す指標　）を向上させる
- ただし、外れ値がデータの重要な特徴を示す場合があるため、慎重に判断する

前処理の具体例とその理由

列を削除する

具体例:

- 「あきらかに無意味な列」「相関が高過ぎる列」や「重要度が低い列」を削除

名前	年齢	性別	身長	体重	BMI
山田	25	男性	171	64	21.89
佐藤	27	女性	155	52	21.64
高橋	30	男性	175	105	34.29

名前	年齢	性別	身長	BMI
山田	25	男性	171	21.89
佐藤	27	女性	155	21.64
高橋	30	男性	175	34.29

理由:

- 高い欠損率を持つ列や、予測タスクに寄与しない列を削除することでデータのノイズを減らし、モデルの精度を向上させる
- 相関が高すぎる特徴量を削除することで、多重共線性のリスク（重回帰モデルにおいて、説明変数の中に、相関係数が高い組み合わせがあることをいう（例:　体重とBMI））を軽減する
- 特徴量が多すぎる場合に、次元削減によって計算効率を向上させる

前処理の具体例とその理由

データ変換

- 正規化

  具体例:

  - 表記ゆれやスケールを整える：表記ゆれ修正：「東京都」「東京」を「東京」に統一
  - 重複データ削除：同じレコードを1つだけ残す
  - スケール統一：単位が異なる場合（例:　距離がkmとm）を統一する

名前	出身地	社員番号	通勤時間
山田	東京都	E2211	1h
佐藤	東京	E2031	80min
高橋	愛知	A1701	30min
山田	東京都	E2211	1h

名前	出身地	社員番号	通勤時間
山田	東京	E2211	60min
佐藤	東京	E2031	80min
高橋	愛知	A1701	30min

  理由:

  - 異なる単位やスケールを持つデータを統一基準に合わせることで、機械学習モデルの計算を安定化させる
  - 距離計算に基づくモデル（例:　k近傍法、SVM）で、スケール差による影響を抑える
  - 重複データの削除や表記ゆれの修正は、データの一貫性と品質を向上させる

前処理の具体例とその理由

データ変換

- 標準化

  具体例:

  - 「収入」列：[300,　500,　700]　を標準化。　平均:　500、標準偏差（標準的な平均値との差）:　200

名前	収入
山田	300
佐藤	500
高橋	700

名前	収入
山田	-1.0
佐藤	0.0
高橋	1.0

  理由:

  - データを平均0、標準偏差1にスケールすることで、統計的性質を標準化する
  - 勾配降下法を用いるモデル（例:　ロジスティック回帰、ニューラルネットワーク）の収束を安定化・高速化する

前処理の具体例とその理由

データ変換

- 対数変換

  具体例:

  - 「売上高」列：[10,　100,　1000]　を対数変換する

店舗名	売上高
東京店	10
名古屋店	100
大阪店	1000

店舗名	売上高
東京店	1.0
名古屋店	2.0
大阪店	3.0

  理由:

  - データの分布を正規分布に近づけることで、モデルの性能を向上させる
  - 大きな値の影響を抑え、データを扱いやすくする
  - Yeo-Johnson変換では負の値にも対応しており、幅広いデータセットに適用可能

前処理の具体例とその理由

データ変換

- 離散化

  具体例:

  - 年齢データを以下のグループに分類

名前	年齢
山田	25
佐藤	27
高橋	30
鈴木	19
田中	60

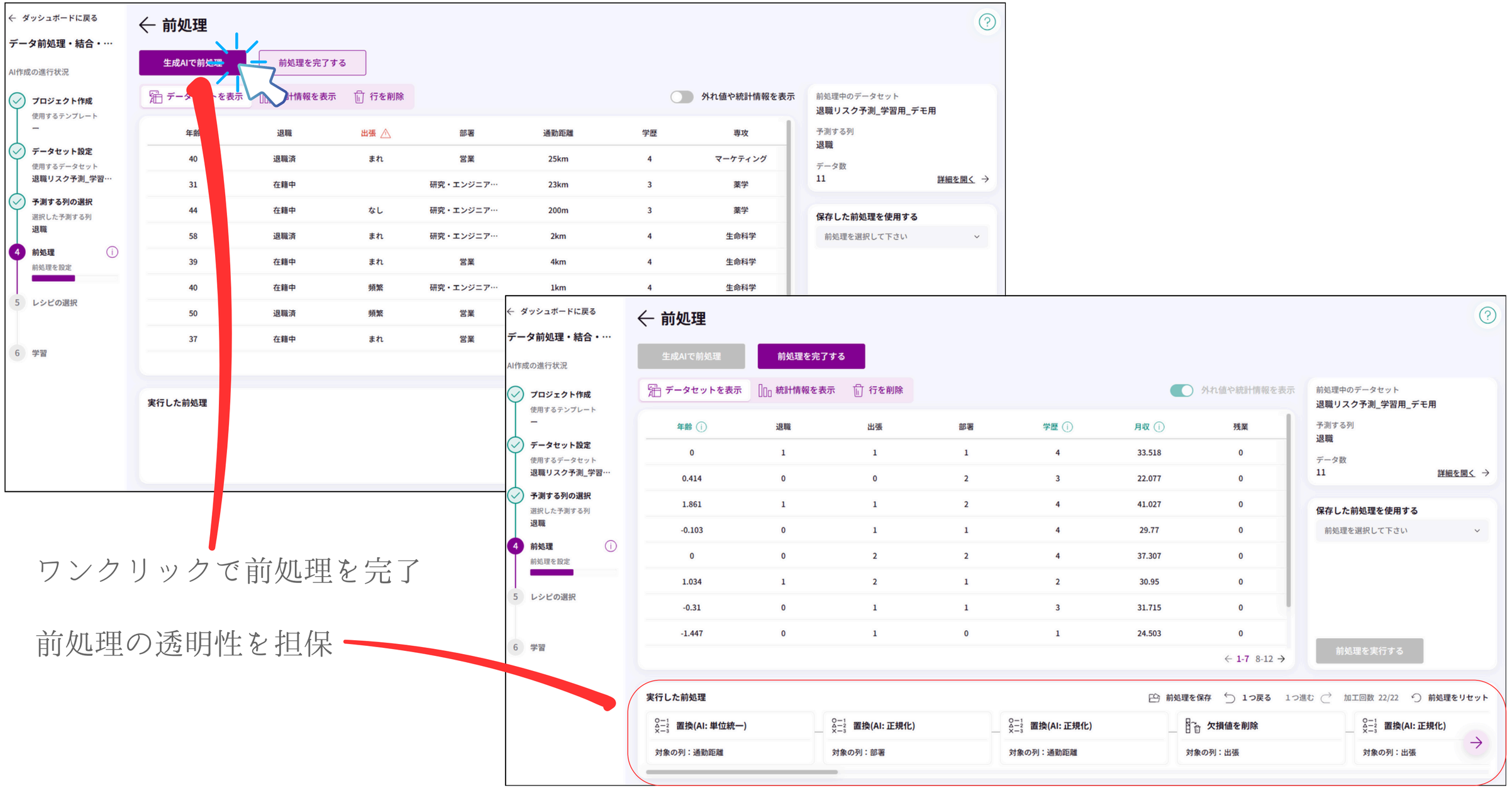
10-20:　若年層
21-40:　中年層
41-60:　壮年層

名前	年齢
山田	中年層
佐藤	中年層
高橋	中年層
鈴木	若年層
田中	壮年層

  理由:

  - 数値データをカテゴリに分けることで、特定の区間に対する解釈性を高める
  - 複雑な数値変数を簡素化し、モデルの学習を容易にする
  - 例:　年齢を「10代」「20代」などのグループに分割する

MatrixFlowでの「生成AI自動前処理」の実行はカンタン

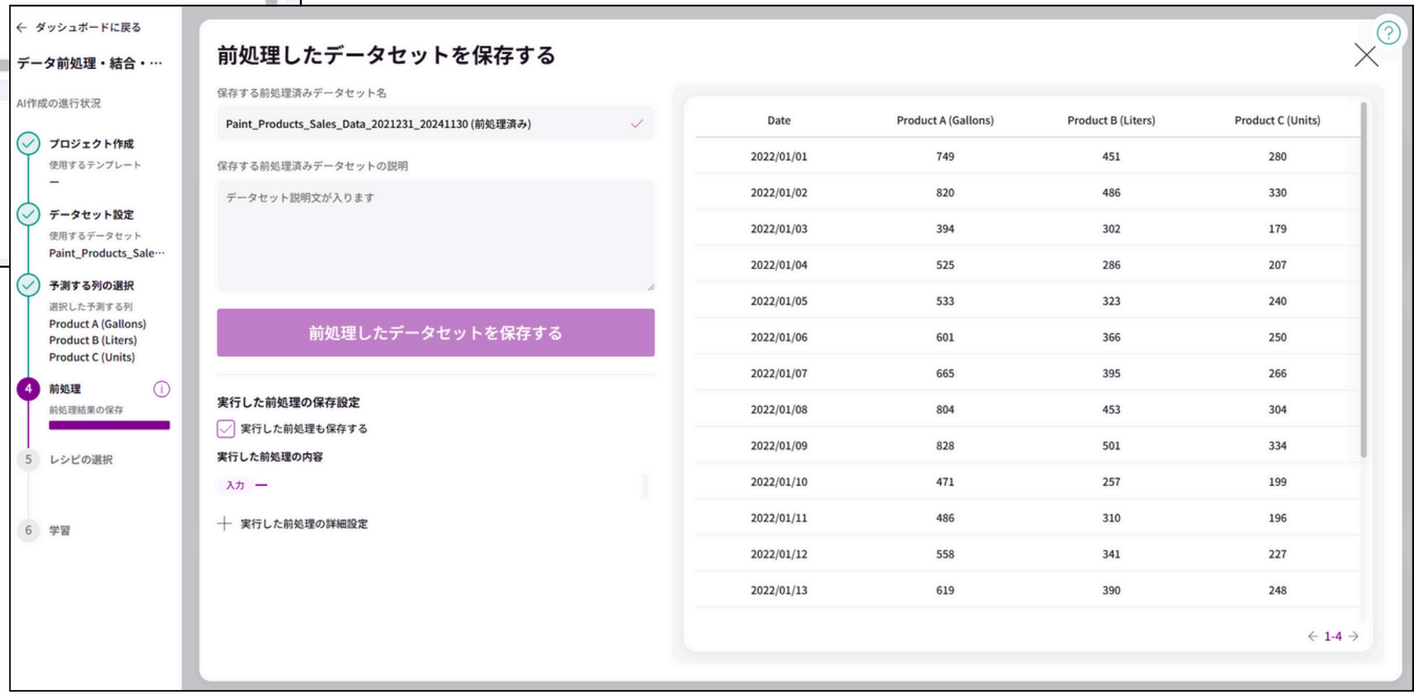


ワンクリックで前処理を完了

前処理の透明性を担保

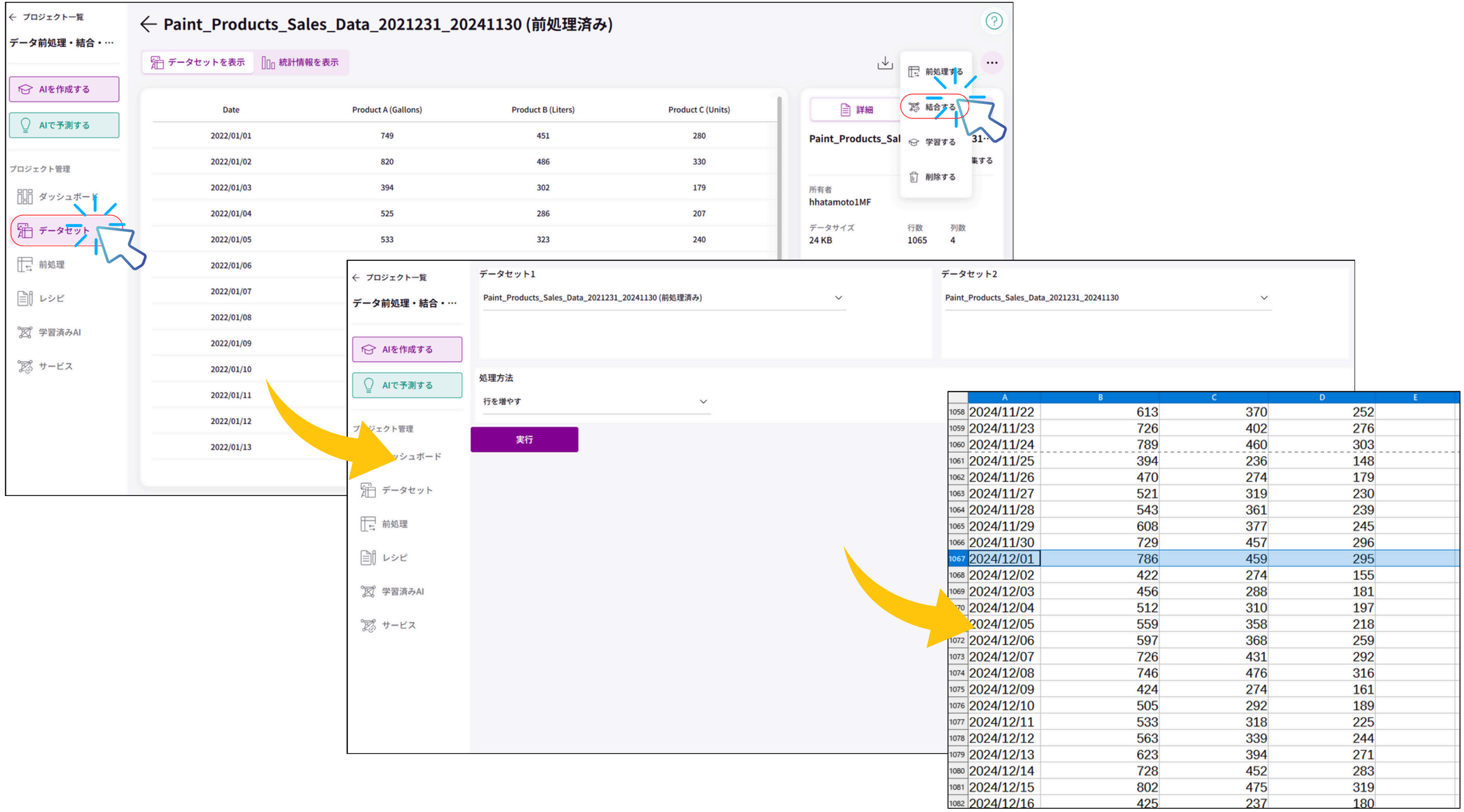
生成AIによる行削除機能

テキストから指定して行削除が可能



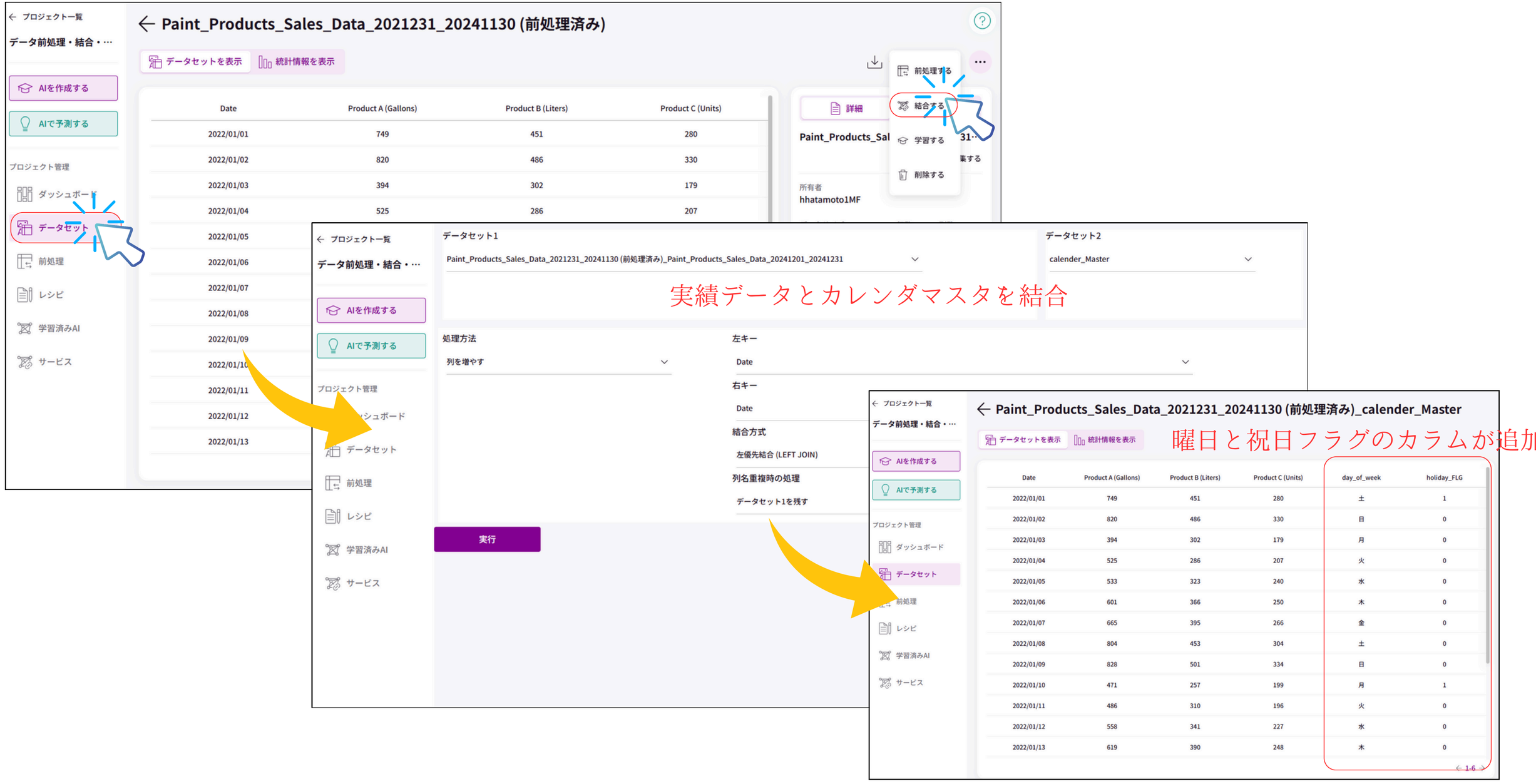
データ結合機能

データ管理にある２つのデータを統合も簡単



	A	B	C	D	E
1058	2024/11/22	613	370	252	
1059	2024/11/23	726	402	276	
1060	2024/11/24	789	460	303	
1061	2024/11/25	394	236	148	
1062	2024/11/26	470	274	179	
1063	2024/11/27	521	319	230	
1064	2024/11/28	543	361	239	
1065	2024/11/29	608	377	245	
1066	2024/11/30	729	457	296	
1067	2024/12/01	786	459	295	
1068	2024/12/02	422	274	155	
1069	2024/12/03	456	288	181	
1070	2024/12/04	512	310	197	
1071	2024/12/05	559	358	218	
1072	2024/12/06	597	368	259	
1073	2024/12/07	726	431	292	
1074	2024/12/08	746	476	316	
1075	2024/12/09	424	274	161	
1076	2024/12/10	505	292	189	
1077	2024/12/11	533	318	225	
1078	2024/12/12	563	339	244	
1079	2024/12/13	623	394	271	
1080	2024/12/14	728	452	283	
1081	2024/12/15	802	475	319	
1082	2024/12/16	425	237	180	

データ管理にある２つのデータを統合も簡単



データ収集・整理からAI構築までサポート

AI活用の壁として、「データの収集・整理」があります。
MatrixFlowではデータの収集・前処理から、AIアルゴリズム構築まで「現場で使える＆高い精度のAI」を実現。

また、プロのデータサイエンティストが、現場のAI活用をサポートいたします。

MatrixFlow　AI構築サポート

課題設定/計画立案サポート

▼

データ収集計画/データ収集サポート

▼

データクレンジング/前処理支援サポート

▼

AIモデル作成/推論サポート

▼

AIモデル評価/実運用サポート

2週間～4週間に1回Web面談でのサポート
を提供しております。
ベーシックプラン月額料金に含まれています。