Fujitsu Kozuchi Auto Data Wrangling 利用手順書

2023年12月19日

富士通株式会社

Fujitsu Auto Data Wrangling

生成AIを用いた表データの前処理自動化

複数の生成AIを活用したデータ整形とデータ強化により、表データの前処理を自動化！AI適用におけるデータ準備の工数削減と、AIの精度向上を実現します。

課題

- 現場の表データに機械学習を適用するには、データ整形などの前処理に工数がかかっている
- 通常の機械学習技術では多種多様なテキスト項目をそのまま扱えず、精度向上に限界がある

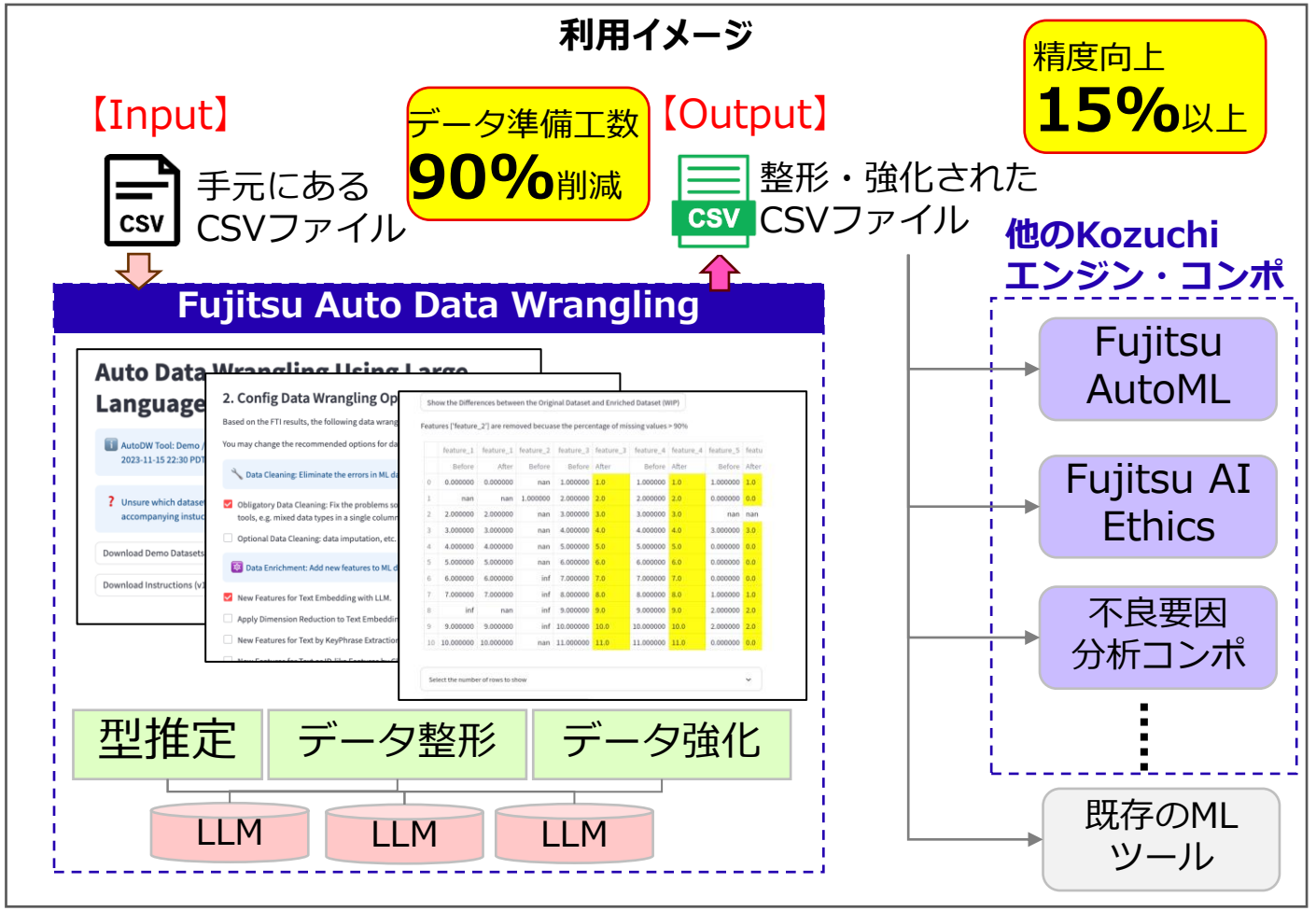
ソリューション

- LLMを用いた**型推定**にもとづく**自動データ整形**
- LLMを用いてテキスト項目を分析し、新たなデータ項目を追加する**自動データ強化**
- 処理内容に応じて複数のLLMを使い分けることによる、**自動化とスケーラビリティの両立**

強み

- データ準備の工数削減（手動でのData Wranglingと比べて90%の削減）
- 他社のData Wranglingツールでは対処できないテキスト項目に対する自動データ強化の実現
- 本技術の適用により機械学習の精度が15%以上向上（Fujitsu AutoMLを用いた場合）

利用イメージ



できること・できないこと

- できること：機械学習適用前のデータ前処理
  - Data Cleaning（データ整形）
    - Fujitsu AutoML等の機械学習ツールに適用する際に、エラーが起きないよう、入力データの形式を整える
  - Data Enrichment（データ強化）
    - 入力データの特徴列を分析し、予測精度の向上に寄与しそうな特徴列を新たに作成・追加する

- できないこと：機械学習そのもの
  - データの学習・分類・予測等の機械学習処理については、Fujitsu AutoML等の既存の機械学習ツールを用いてください

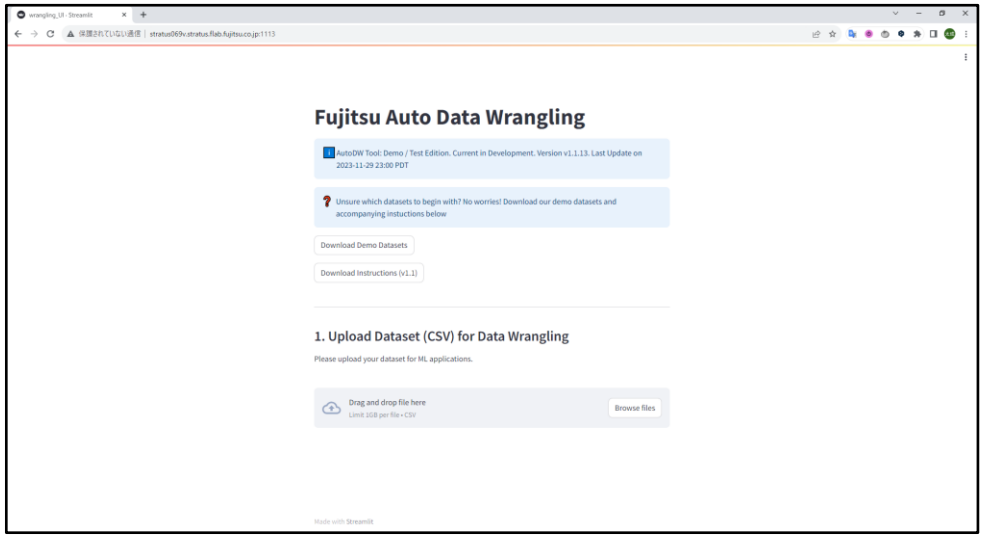
2通りの使い方

- **インタラクティブデモ**
  - 当方が用意したサンプルデータを用いて、Fujitsu Auto Data Wranglingの動作を体験いただけます
- **PoC**
  - ユーザー自身のデータを用いてFujitsu Auto Data Wranglingを利用することができます

環境へのアクセス

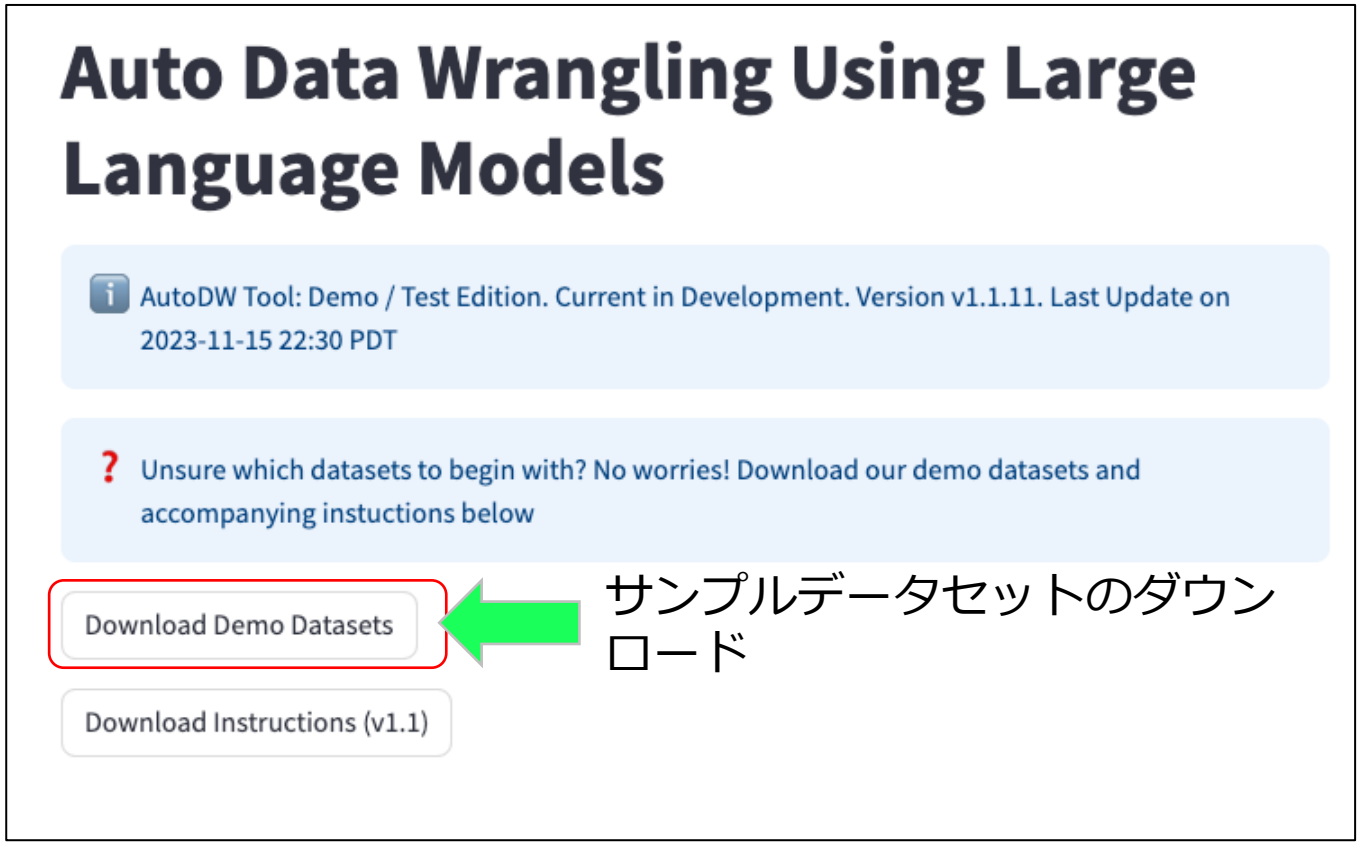
- 準備作業
  - 環境アクセス手順書をaccess manualからダウンロード
  - **PoC**
    - 「環境アクセス手順書」の「Azure VPN Gatewayへの接続方法」セクション
  - **インタラクティブデモ**
    - 「Remote DesktopでBastion接続」セクション

- 環境へのアクセス
  - Webブラウザで http://10.0.0.139:8550/ へアクセス
  - 右図の開始画面が表示されたら動作確認完了



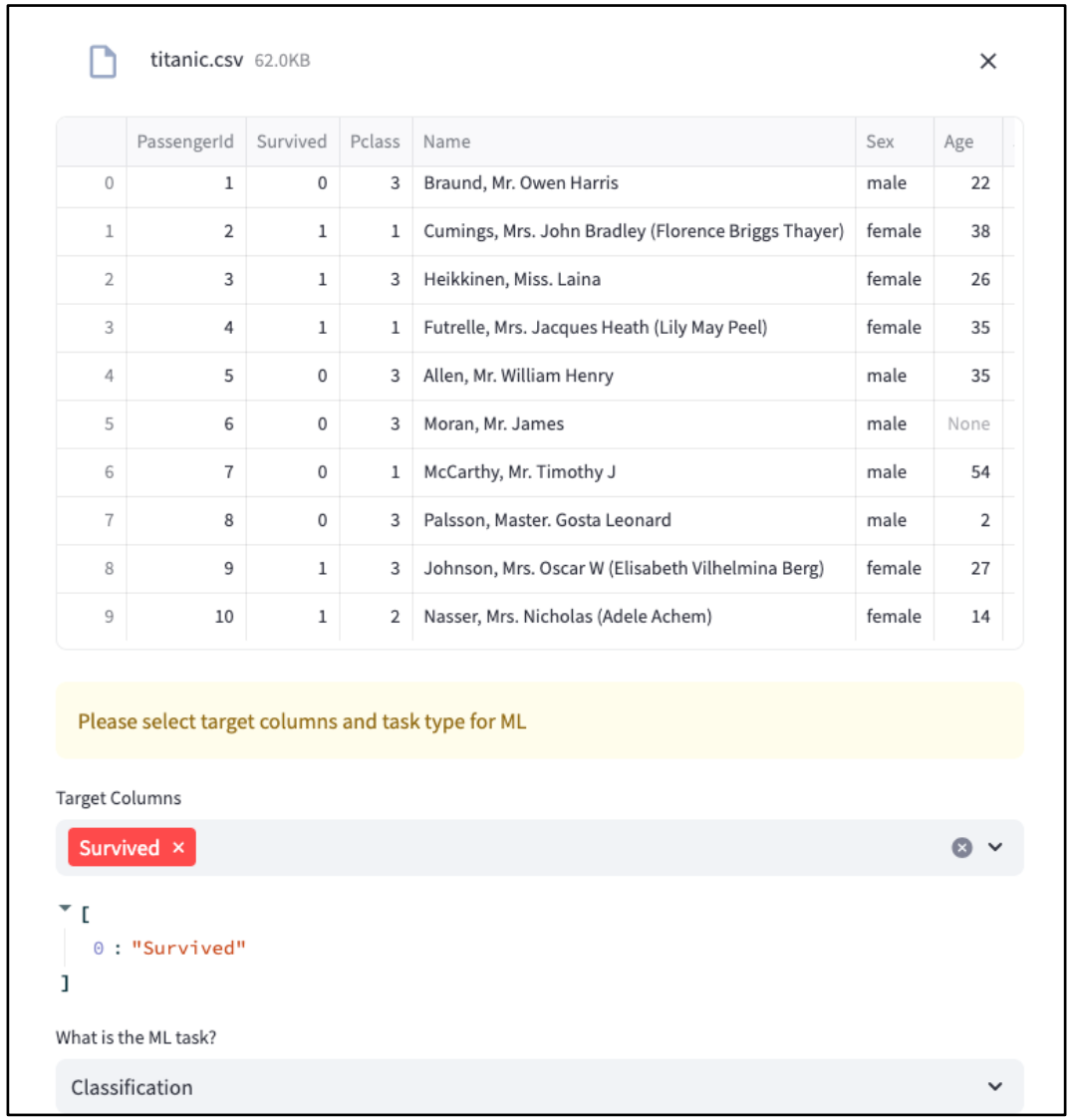
2. Fujitsu Auto Data Wrangling ウェブアプリ使用方法

2.1 サンプルデータセットでの始め方（任意）



- どのデータセットから始めていいか分からない場合、サンプルデータセットをダウンロードしてFujitsu Auto Data Wranglingの効果を確かめることが出来ます
- 以下では、NYC_Airbnb(changed)データセットを元に説明します。

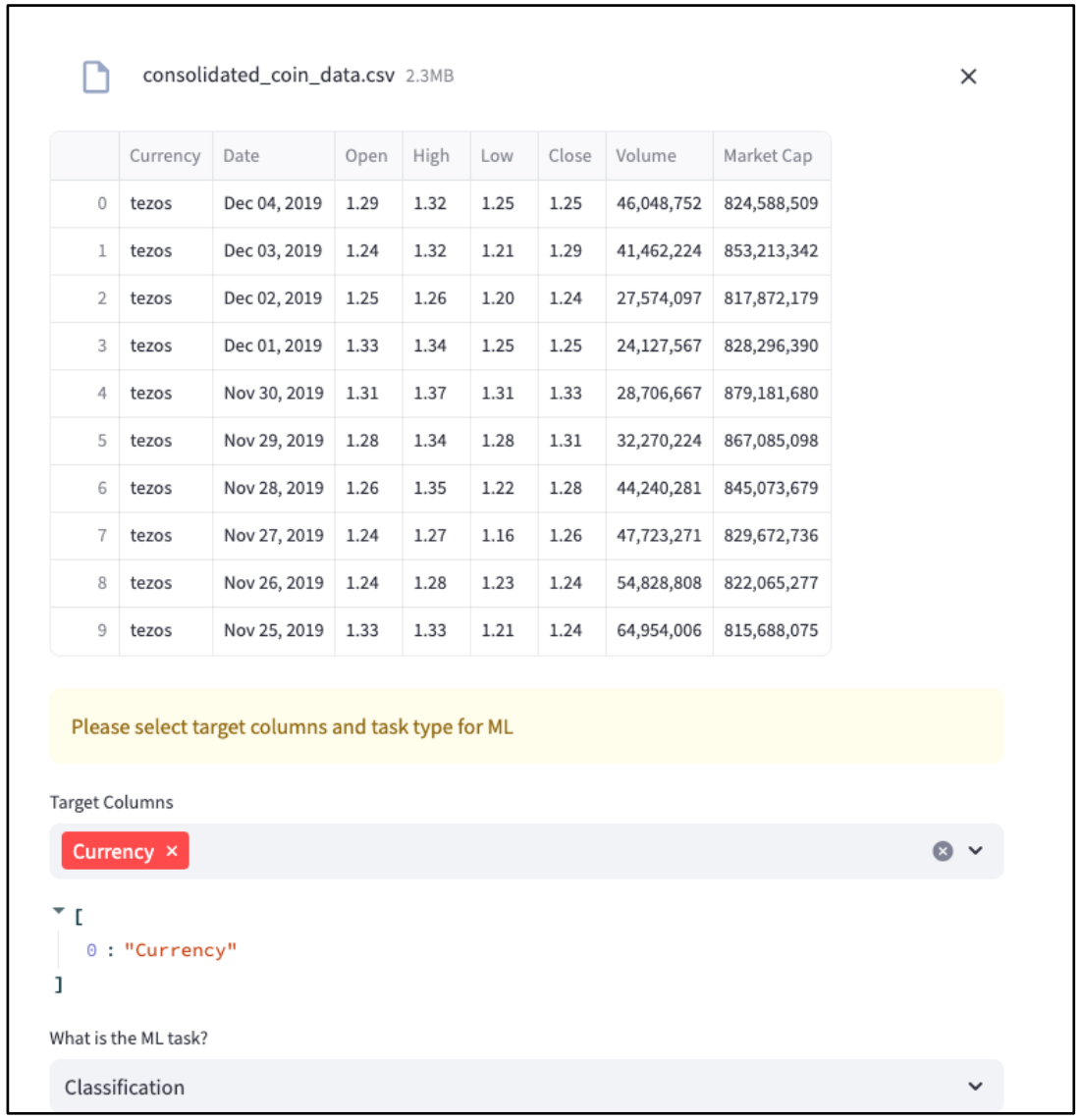
サンプルデータセット #1



	PassengerId	Survived	Pclass	Name	Sex	Age
0	1	0	3	Braund, Mr. Owen Harris	male	22
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38
2	3	1	3	Heikkinen, Miss. Laina	female	26
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35
4	5	0	3	Allen, Mr. William Henry	male	35
5	6	0	3	Moran, Mr. James	male	None
6	7	0	1	McCarthy, Mr. Timothy J	male	54
7	8	0	3	Palsson, Master. Gosta Leonard	male	2
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14

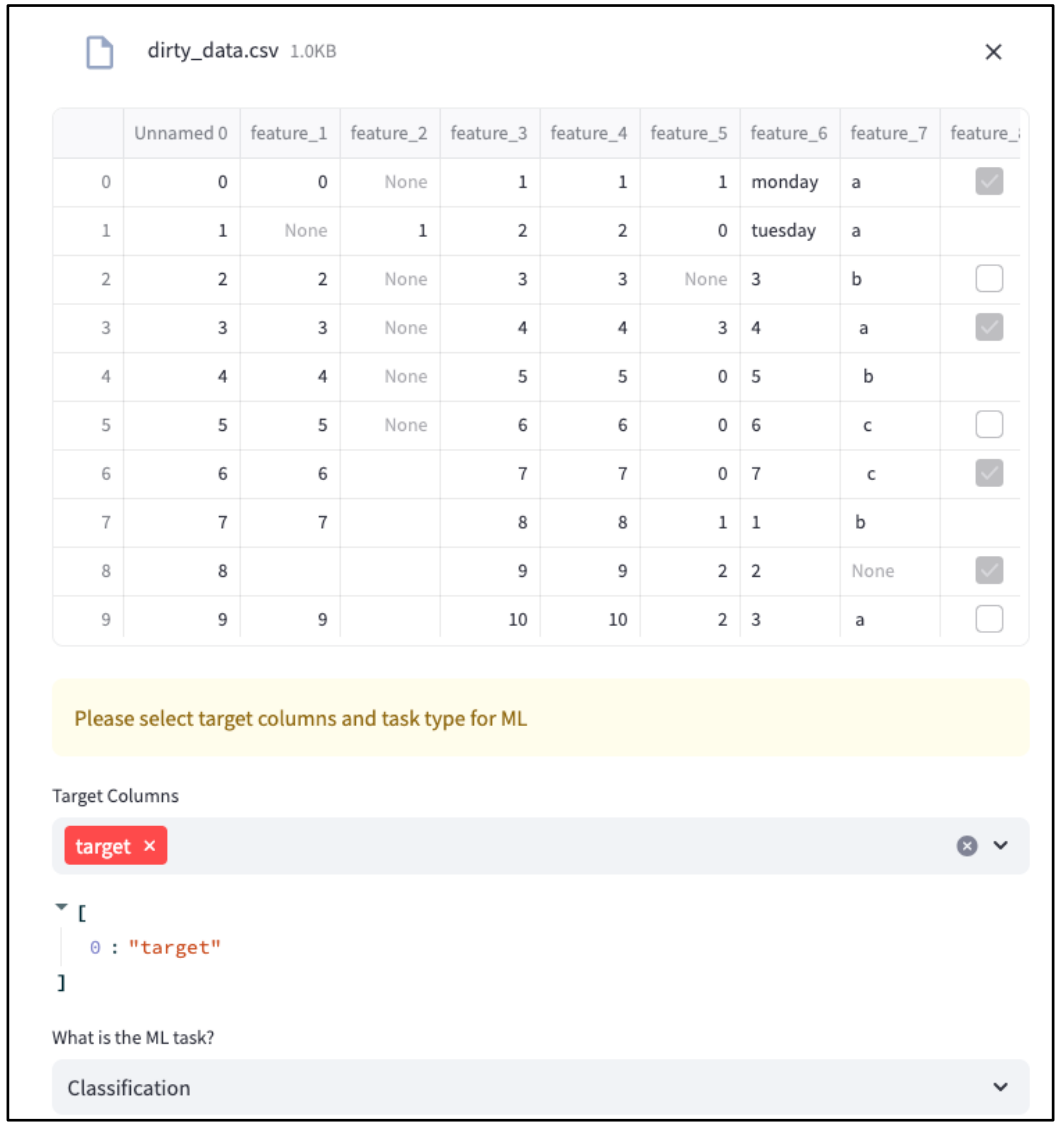
- titanic.csv
  - デモ用データセット
  - Target column: Survived
  - ML task: Classification

サンプルデータセット #2



- consolidated_coin_data.csv
  - Feature type inferenceの効果を示す代表的データセット
    - 数値型に見えますがコンマが埋め込まれているため文字列になります (もしくは文字列に埋め込まれた数字)
  - Target column: Currency
  - ML task: Classification

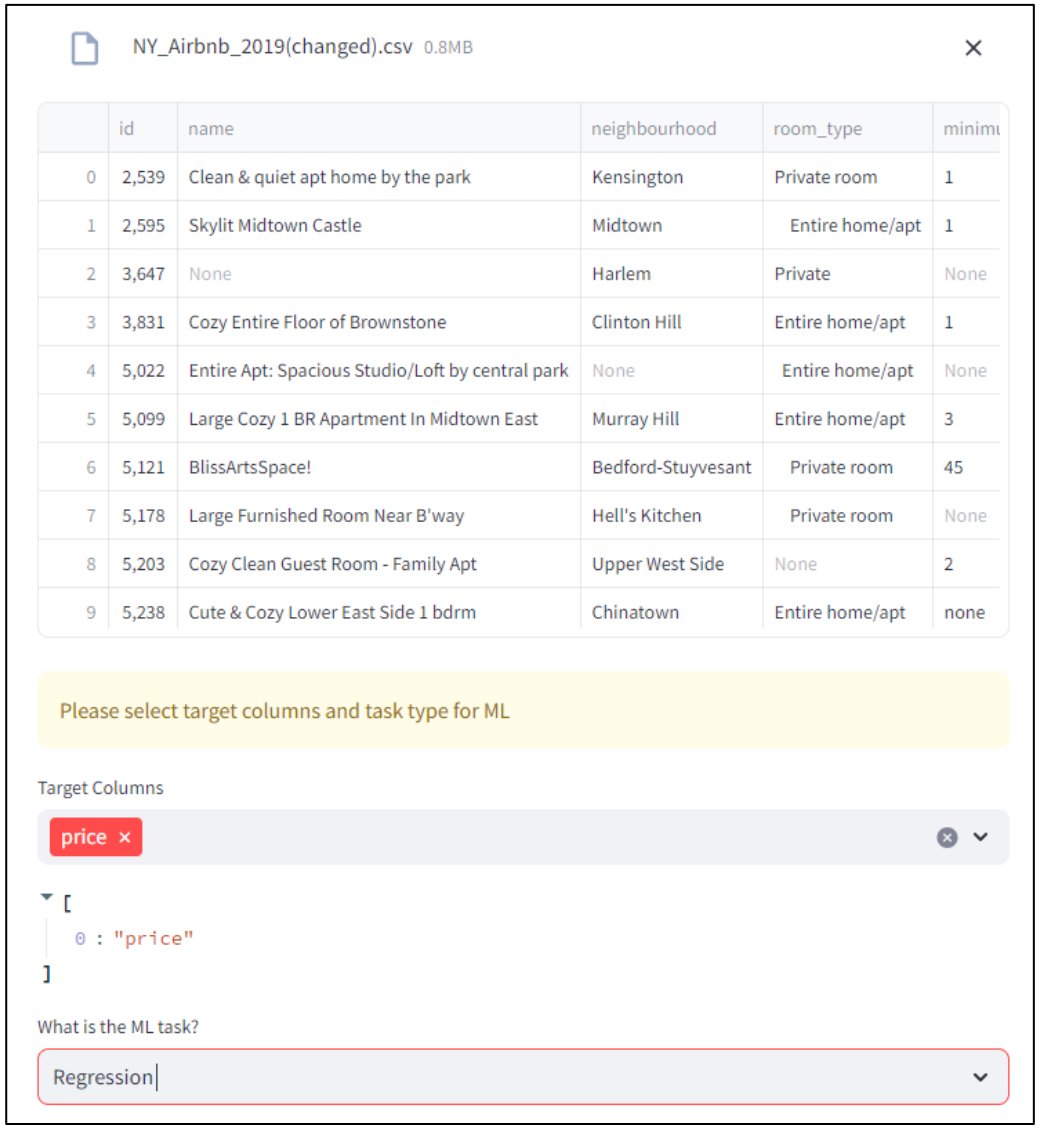
サンプルデータセット #3



	Unnamed 0	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_
0	0	0	None	1	1	1	monday	a	☑
1	1	None	1	2	2	0	tuesday	a	
2	2	2	None	3	3	None	3	b	☐
3	3	3	None	4	4	3	4	a	☑
4	4	4	None	5	5	0	5	b	
5	5	5	None	6	6	0	6	c	☐
6	6	6		7	7	0	7	c	☑
7	7	7		8	8	1	1	b	
8	8			9	9	2	2	None	☑
9	9	9		10	10	2	3	a	☐

- dirty_data.csv
  - Data cleaningの効果を示す代表的データセット
    - 様々な種類のエラーが埋め込まれた小規模データセット (例: 単一カラム内の混合データ型、”?”で表された欠損値等)
  - Target column: target
  - ML task: Classification

サンプルデータセット #4



	id	name	neighbourhood	room_type	minimu
0	2,539	Clean & quiet apt home by the park	Kensington	Private room	1
1	2,595	Skylit Midtown Castle	Midtown	Entire home/apt	1
2	3,647	None	Harlem	Private	None
3	3,831	Cozy Entire Floor of Brownstone	Clinton Hill	Entire home/apt	1
4	5,022	Entire Apt: Spacious Studio/Loft by central park	None	Entire home/apt	None
5	5,099	Large Cozy 1 BR Apartment In Midtown East	Murray Hill	Entire home/apt	3
6	5,121	BlissArtsSpace!	Bedford-Stuyvesant	Private room	45
7	5,178	Large Furnished Room Near B'way	Hell's Kitchen	Private room	None
8	5,203	Cozy Clean Guest Room - Family Apt	Upper West Side	None	2
9	5,238	Cute & Cozy Lower East Side 1 bdrm	Chinatown	Entire home/apt	none

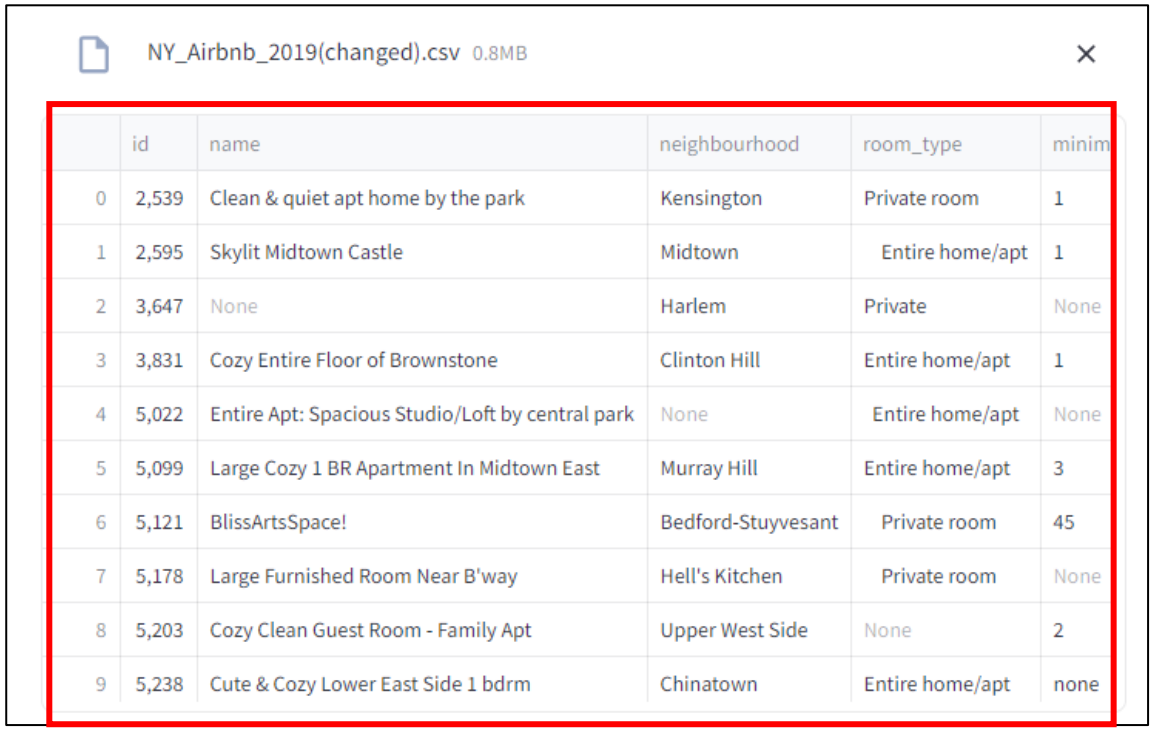
- NYC_Airbnb_2019(changed).csv
  - Data Cleaning, FTI, DataEnrichmentの効果を示すデータセット
    - テキスト型・日付型・数値型の特徴列が含まれております。
    - dirty_data.csvと同様のエラーを編集により埋め込み
    - AB_NYC_2019.csv© http://insideairbnb.com/ クリエイティブ・コモンズ・ライセンス（表示4.0 国際）を改変して作成
  - Target column: Price
  - ML task:Regression

以降のページはこのサンプルデータを用いて説明
デモ動画ではこのファイルを使用

2.2 データセットと問題設定の指定（1）

- データセット(CSVファイル)をアップロードするには二つの方法があります
  - CSVファイルをドラッグ＆ドロップ
  - “Browser files”ボタンをクリックしてCSVファイルを選択

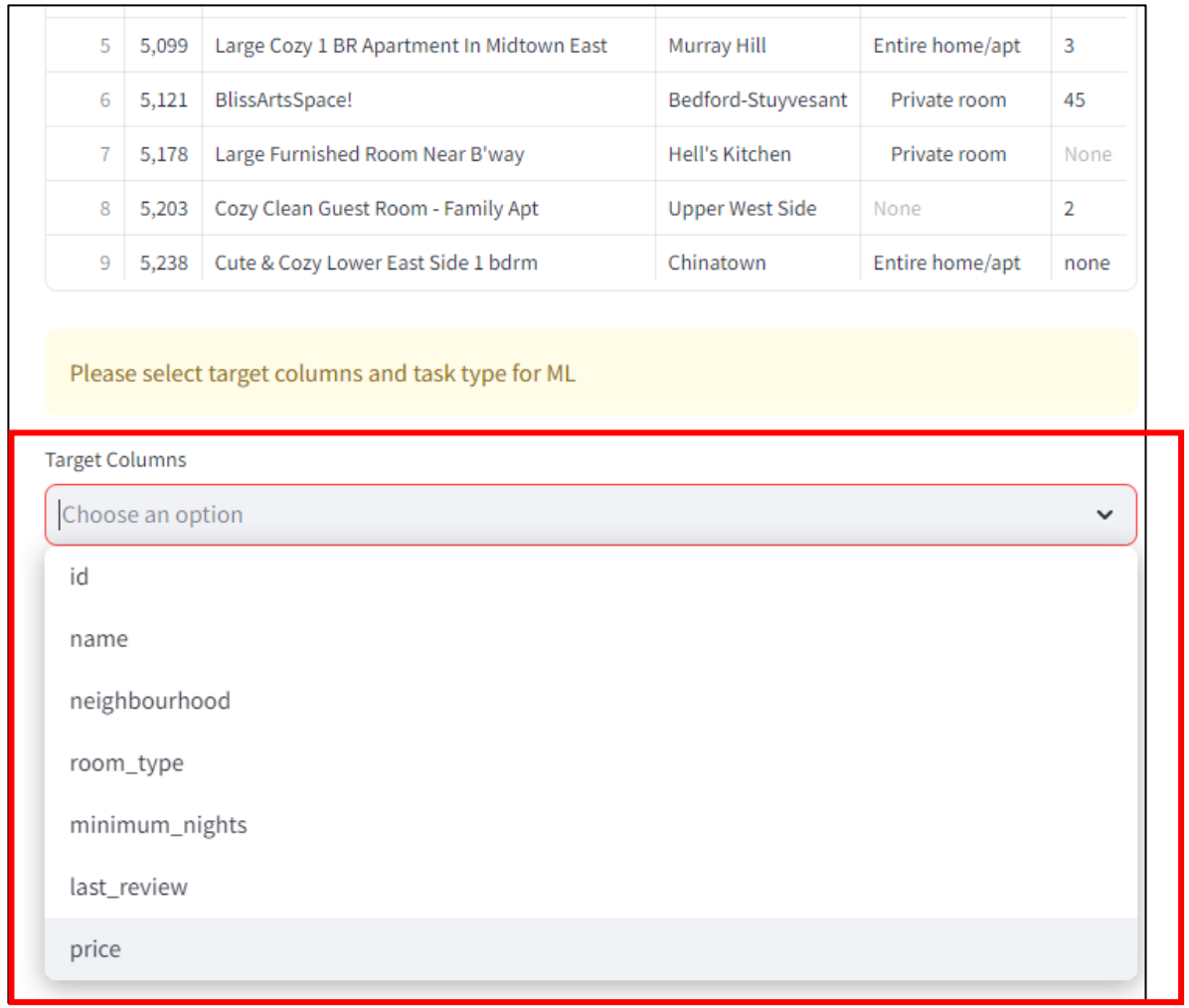
2.2 データセットと問題設定の指定（2）



	id	name	neighbourhood	room_type	minim
0	2,539	Clean & quiet apt home by the park	Kensington	Private room	1
1	2,595	Skylit Midtown Castle	Midtown	Entire home/apt	1
2	3,647	None	Harlem	Private	None
3	3,831	Cozy Entire Floor of Brownstone	Clinton Hill	Entire home/apt	1
4	5,022	Entire Apt: Spacious Studio/Loft by central park	None	Entire home/apt	None
5	5,099	Large Cozy 1 BR Apartment In Midtown East	Murray Hill	Entire home/apt	3
6	5,121	BlissArtsSpace!	Bedford-Stuyvesant	Private room	45
7	5,178	Large Furnished Room Near B'way	Hell's Kitchen	Private room	None
8	5,203	Cozy Clean Guest Room - Family Apt	Upper West Side	None	2
9	5,238	Cute & Cozy Lower East Side 1 bdrm	Chinatown	Entire home/apt	none

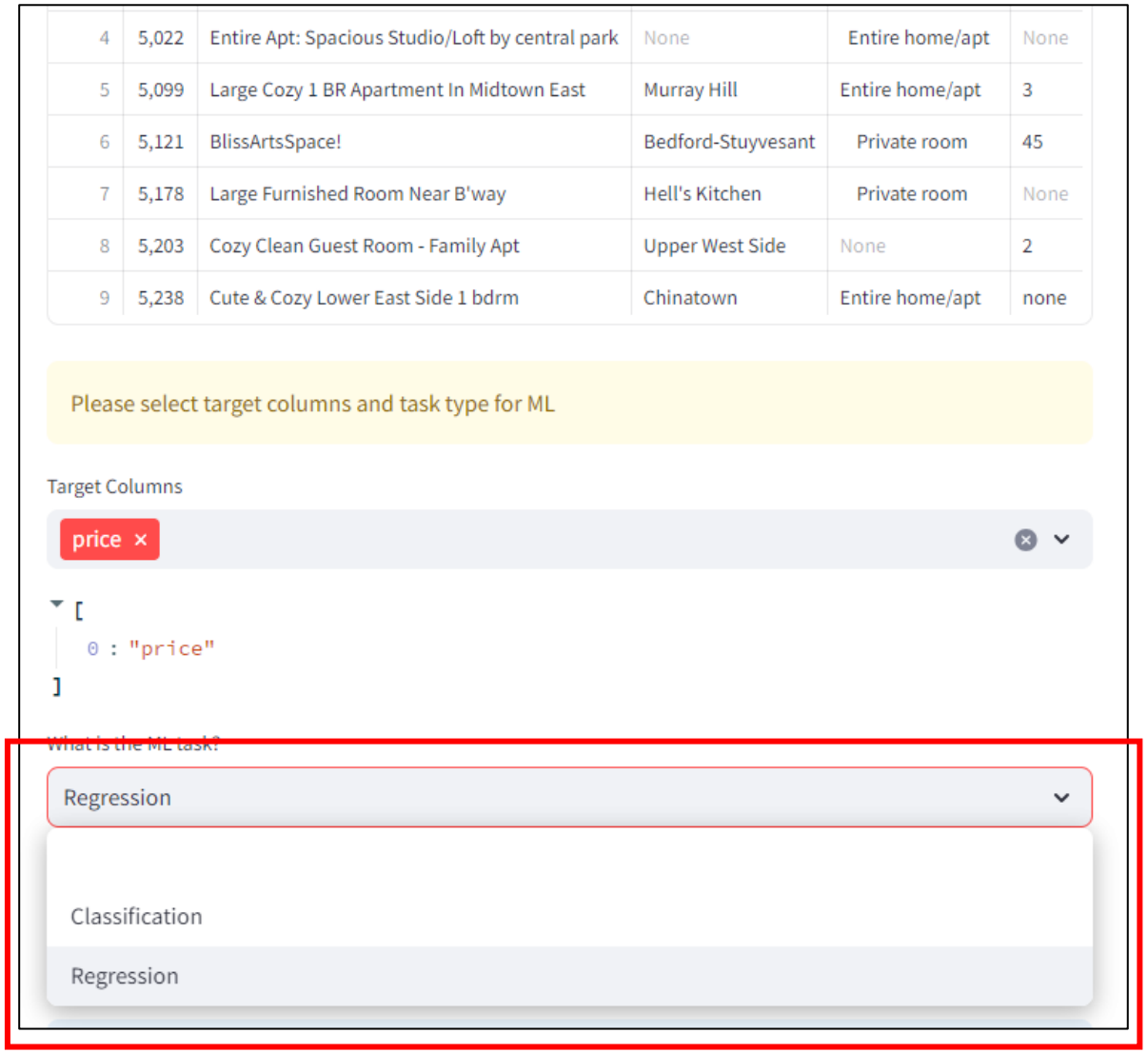
- アップロードしたデータセットのプレビューが表示されます

2.2 データセットと問題設定の指定（3）



- “Target Columns”において目的変数(予測したいカラム)を選択してください

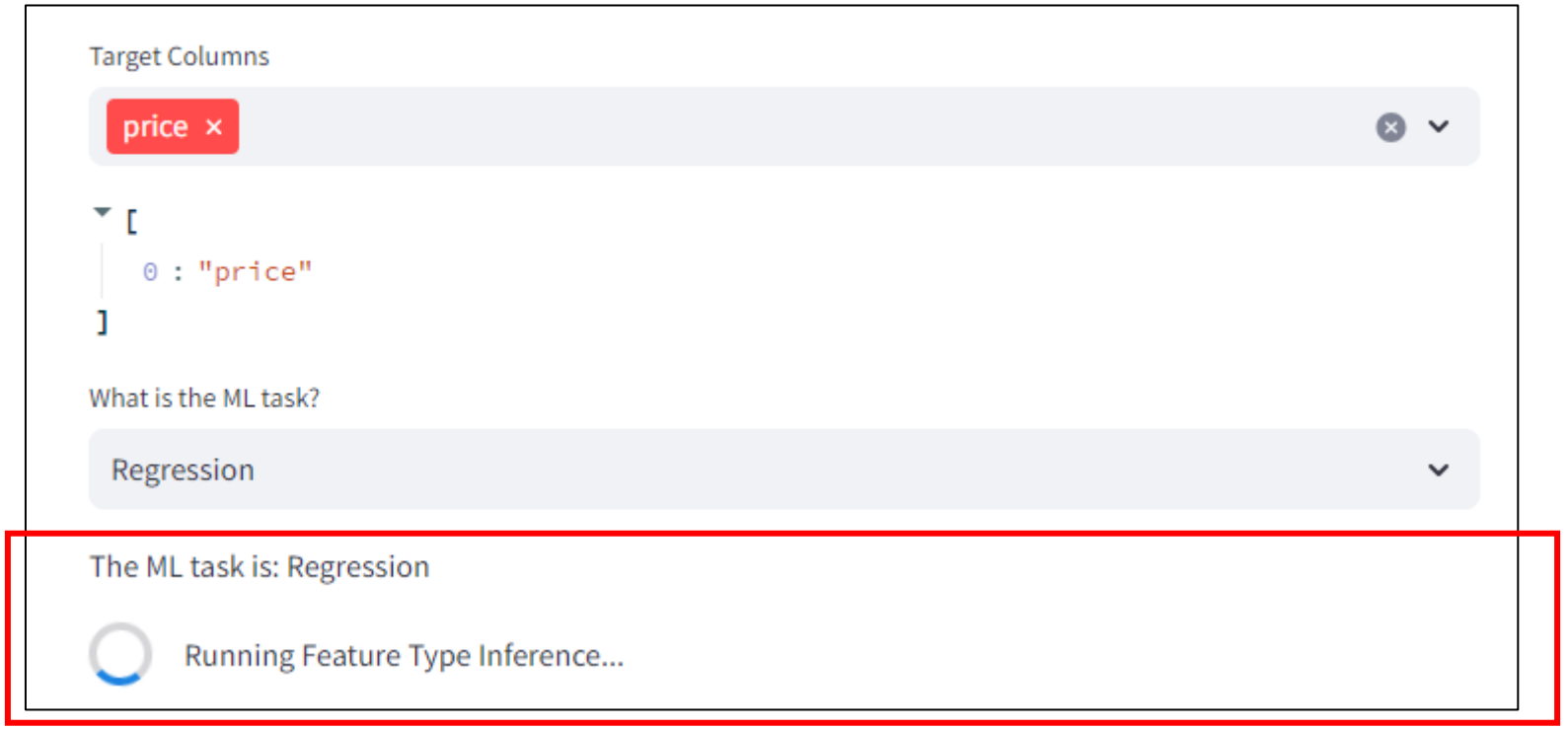
2.2 データセットと問題設定の指定（4）



4	5,022	Entire Apt: Spacious Studio/Loft by central park	None	Entire home/apt	None
5	5,099	Large Cozy 1 BR Apartment In Midtown East	Murray Hill	Entire home/apt	3
6	5,121	BlissArtsSpace!	Bedford-Stuyvesant	Private room	45
7	5,178	Large Furnished Room Near B'way	Hell's Kitchen	Private room	None
8	5,203	Cozy Clean Guest Room - Family Apt	Upper West Side	None	2
9	5,238	Cute & Cozy Lower East Side 1 bdrm	Chinatown	Entire home/apt	none

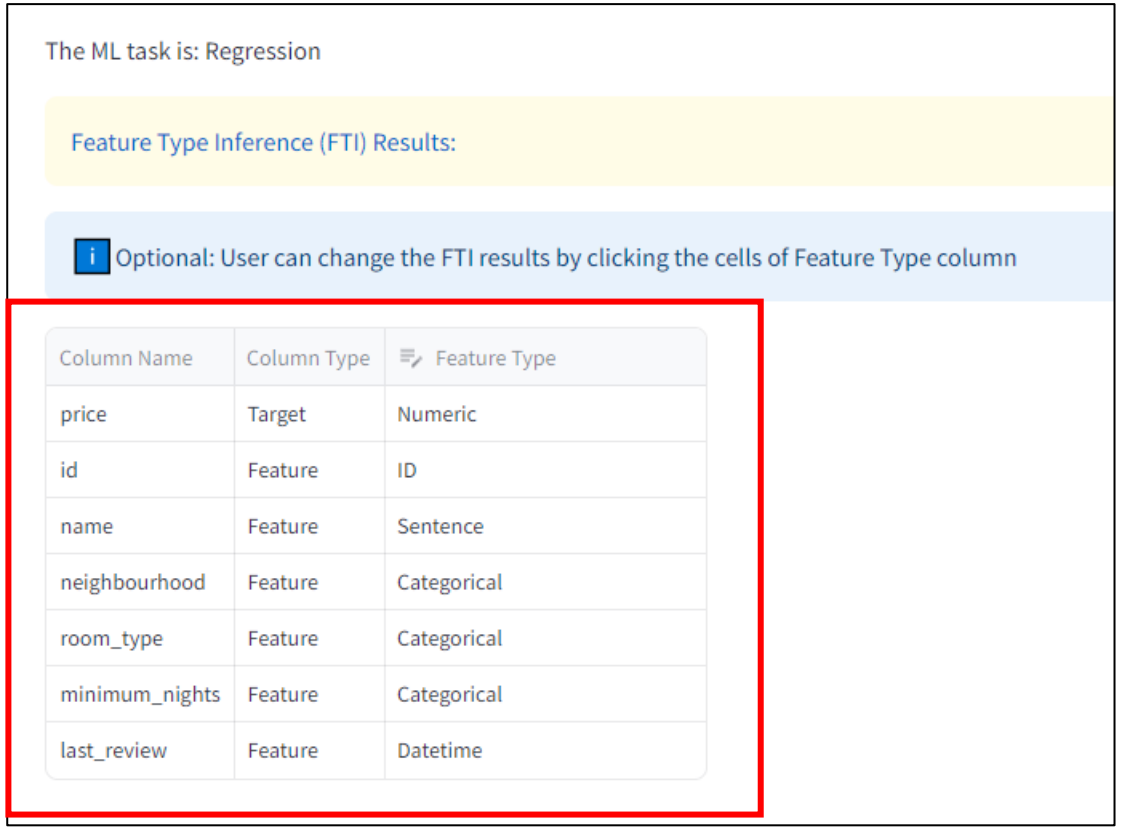
- 機械学習タスクを選択してください。以下のように二つの選択肢があります。
  - Classification: 分類。離散的なカテゴリーを予測するタスク
  - Regression: 回帰。連続値を予測するタスク

2.3 Feature Type Inference (1)



- Target ColumnsとML Taskを選択すると、Feature type inference (FTI)が自動的に実行され、各カラムのFeature typeが推論されます

2.3 Feature Type Inference (2)



Column Name	Column Type	Feature Type
price	Target	Numeric
id	Feature	ID
name	Feature	Sentence
neighbourhood	Feature	Categorical
room_type	Feature	Categorical
minimum_nights	Feature	Categorical
last_review	Feature	Datetime

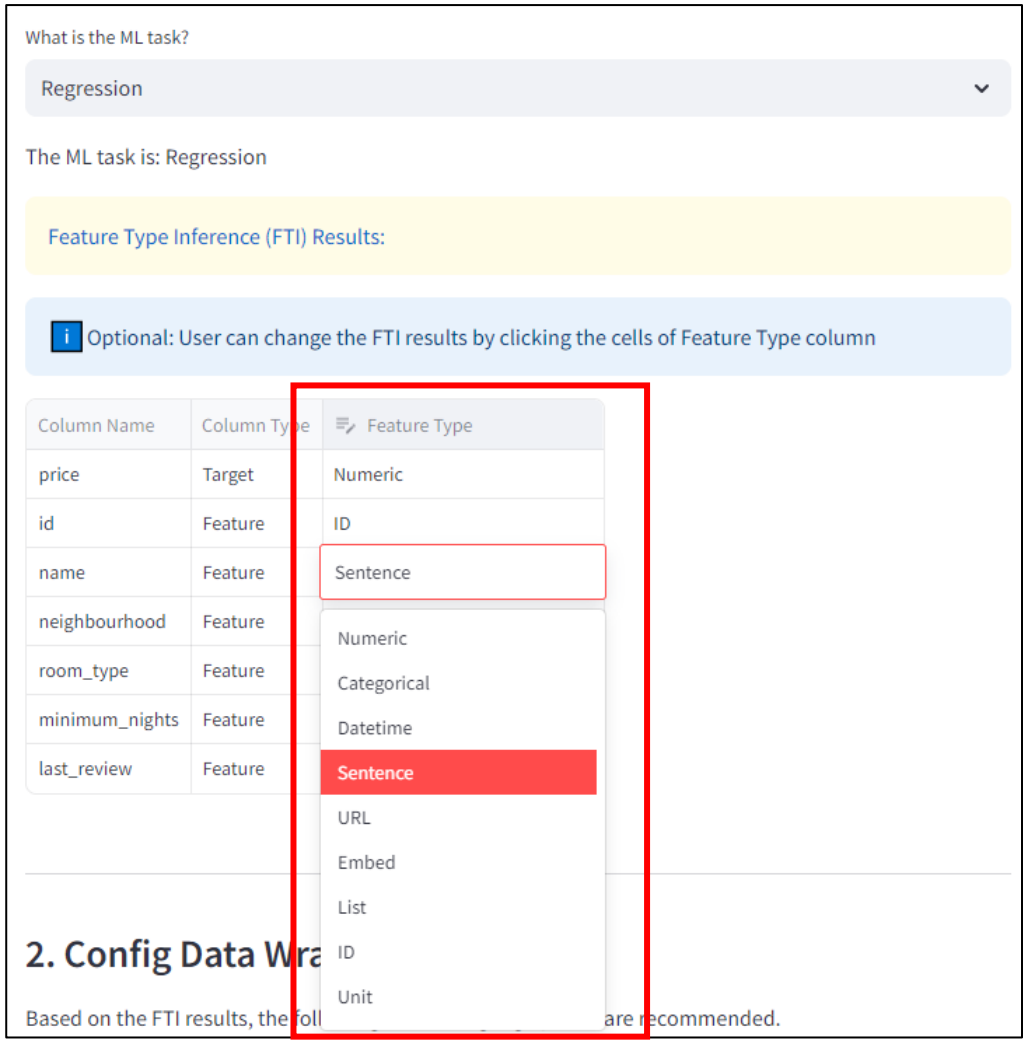
- FTIが完了すると3つのカラムが表示されます
  - Column Name: データセットの各カラムの名前を表示します
  - Column Type: 各カラムがTarget (目的変数)かFeature (特徴量)かを表示します
  - Feature Type: 各カラムのFeature type (次ページで説明)を表示します

2.3 Feature Type Inference (3)

FUJITSU

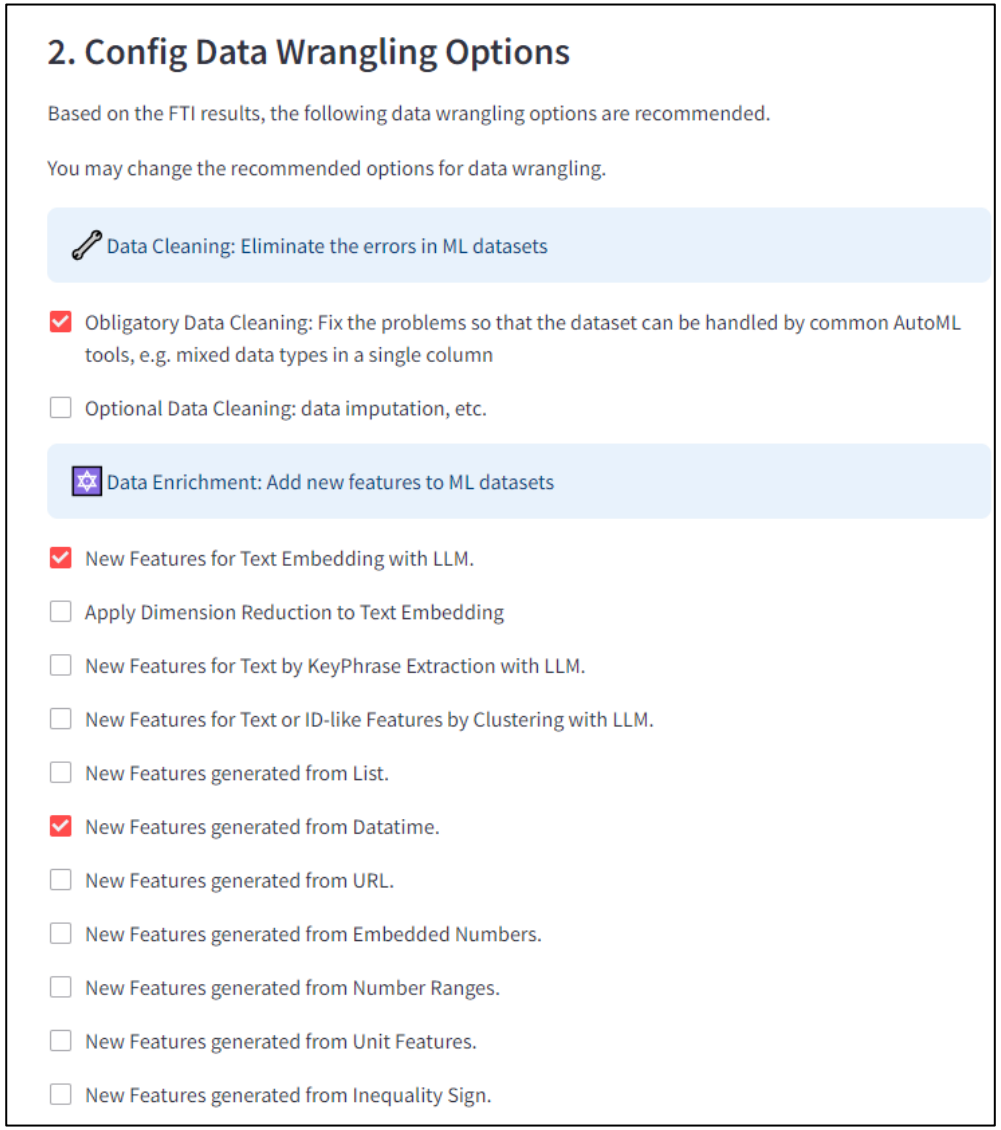
- Feature Types
  - **Numeric**: 数値。例: 1, 2, 3, …
  - **Categorical**: カテゴリー。例: male, female
  - **Datetime**: 日付。例: 11-23-2022, 15:20PM, …
  - **Sentence**: 単語の集合体。例: “auto data wrangling tools are useful”
  - **URL**: URL。例: https://www.fujitsu.com/global/about/research/
  - **Embed**: 数値が含まれた文字列(文字部分は共通)。例: $1,000
  - **List**: リスト。例: [A, B, C], [1, 2, 3]
  - **ID**: ID。例: インデックス
  - **Unit**: 単位と数値が含まれる文字列。例: 100m, 60kg
  - **Sign**: >や<=等の大小を表す記号と数値が含まれる文字列。例: >50, <=100
  - **Range**: 二つの値の範囲を示す文字列。例: 60-100

FTI結果の編集（任意）



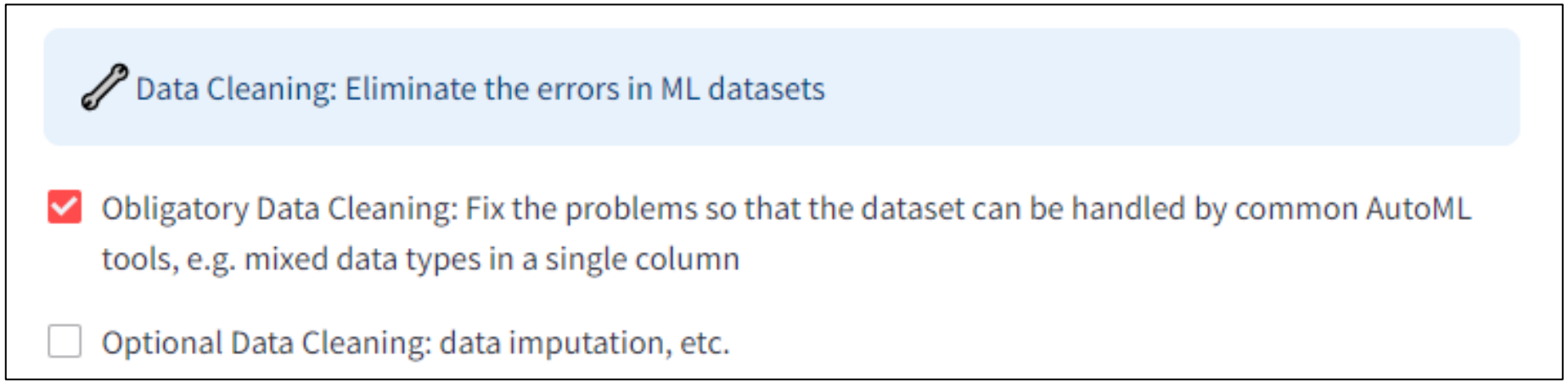
- “Feature Type”カラムは編集可能です
- FTI結果が正確ではない場合、対応セルをクリックしてFeature Typeを変更可能です
- 以後のデータラングリングはユーザーが選択したFeature typeを元に実行されます

2.4 データラングリング設定（1）



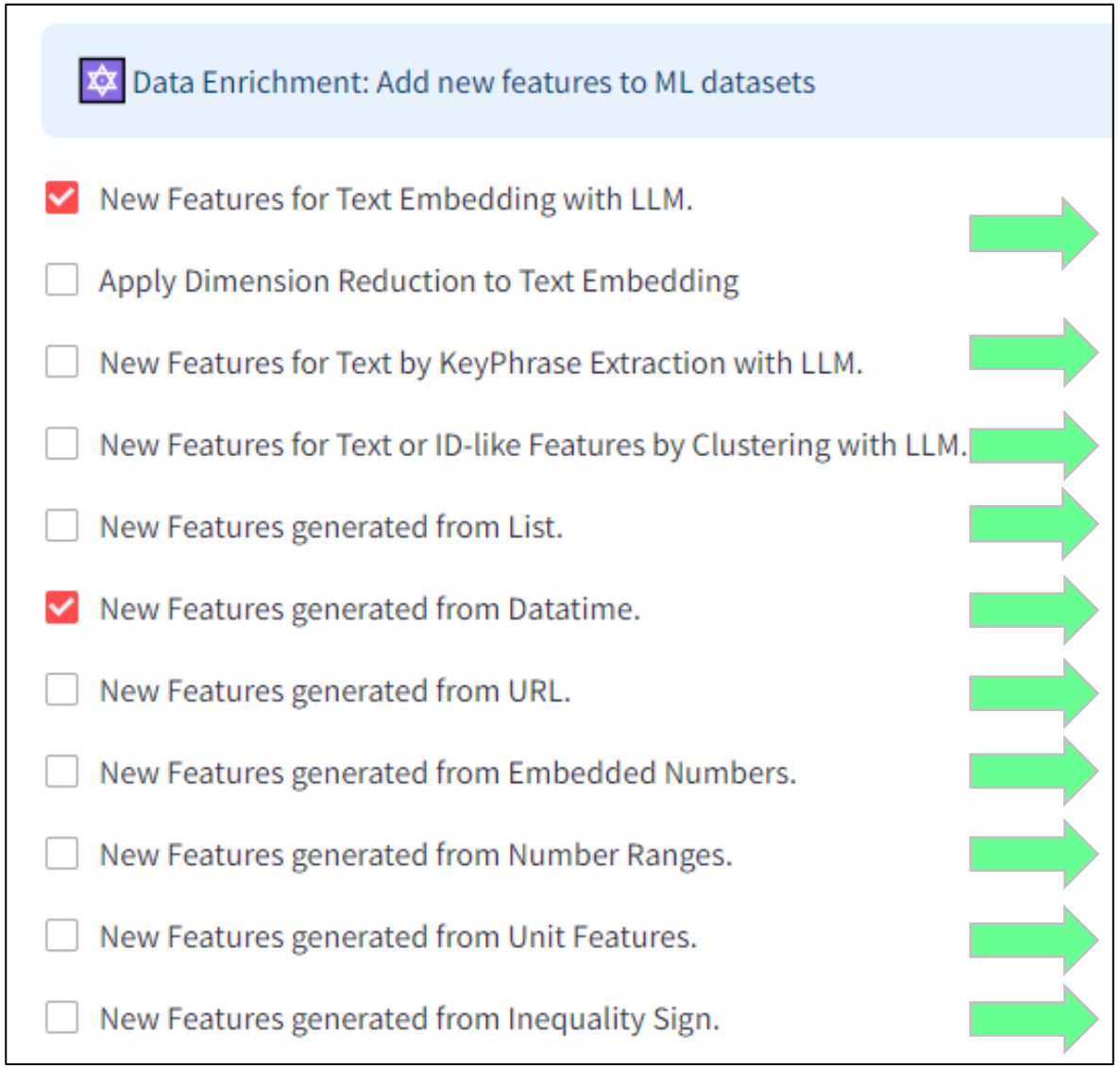
- FTI結果を元に、データラングリング設定が自動的に提案されます
- ユーザーは提案された設定を修正できます
- データラングリングの設定として主要なものが二つあります
  - Data Cleaning: エラー除去
  - Data enrichment: 新しい特徴量の追加

2.4 データラングリング設定 (2)



- Data Cleaningは二項目あります
  - **Obligatory Data Cleaning**: 主要AutoMLツールへの互換を可能にするため、データセットの様々な問題に対応します。データセットの適切なデコード、ヘッダーの修正、無関係な特徴量の除去、目的カラム内の欠損値を含むセルの対処、Inf値の対処、混合データ型を含むカラムやテキスト型カラムの処理等の種々の致命的エラーへの対応を行います
  - **Optional Data Cleaning**: AutoMLツールへの入力を万全にするため、Obligatory Data Cleaningに加えて処理を行います。欠損値処理や目的カラムのエンコード等が行われます。多くの主要AutoMLツールはこれらの処理が可能なため、選択は任意です

2.4 データラングリング設定 (3)



- FTIを元にしたData enrichment

“Sentence”に対して大規模言語モデル(LLM)を用いたテキスト埋め込み特徴量が追加されます (詳細は次ページ)

“Sentence”に対してLLMによるキーフレーズ特徴量が追加されます

“Sentence”に対してLLMによるクラスタ特徴量が追加されます

“List”に対して新たな特徴量が追加されます

“Datetime”に対して新たな特徴量が追加されます

“URL”に対して新たな特徴量が追加されます

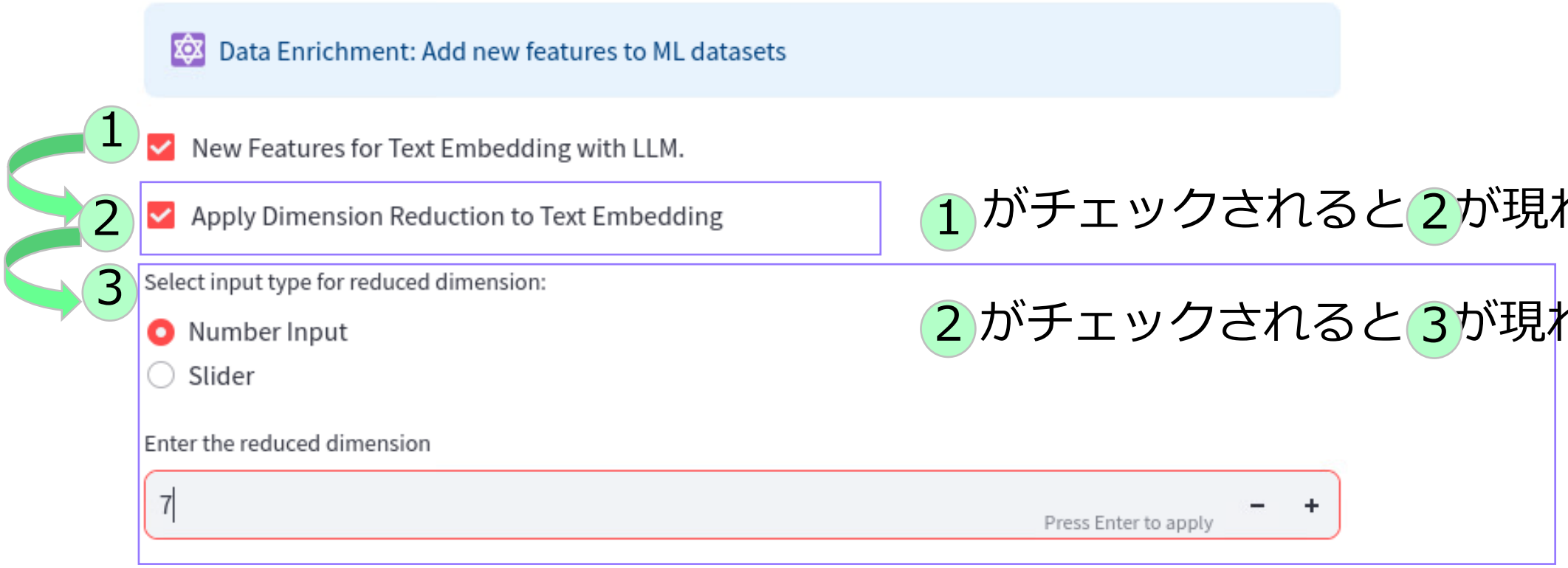
“Embed”に対して新たな特徴量が追加されます

“Range”に対して新たな特徴量が追加されます

“Unit”に対して新たな特徴量が追加されます

“Sign”に対して新たな特徴量が追加されます

2.4 データラングリング設定 (4)



- テキスト埋め込み設定
  - 2 を選択すると、次元削減がテキスト埋め込み特徴量に対して適用されます
  - 3 のように、手入力かスライダーを用いて次元数(1から768の間)を設定できます

2.4 データラングリング設定（5）

FUJITSU

Actions to be Performed & Explanations

Obligatory Data Cleaning will be conducted. The Data cleaning module will check for possible dataset errors and fix them, including but not limited to: decode datasets as necessary, clean headers, remove irrelevant features, drop NaN cells in target, process and replace infinite values, handle columns with mixed data types, encode the target column for machine learning compatibility, text column cleaning, etc.

LLM Embedding will be conducted

Embedding dimension is:768

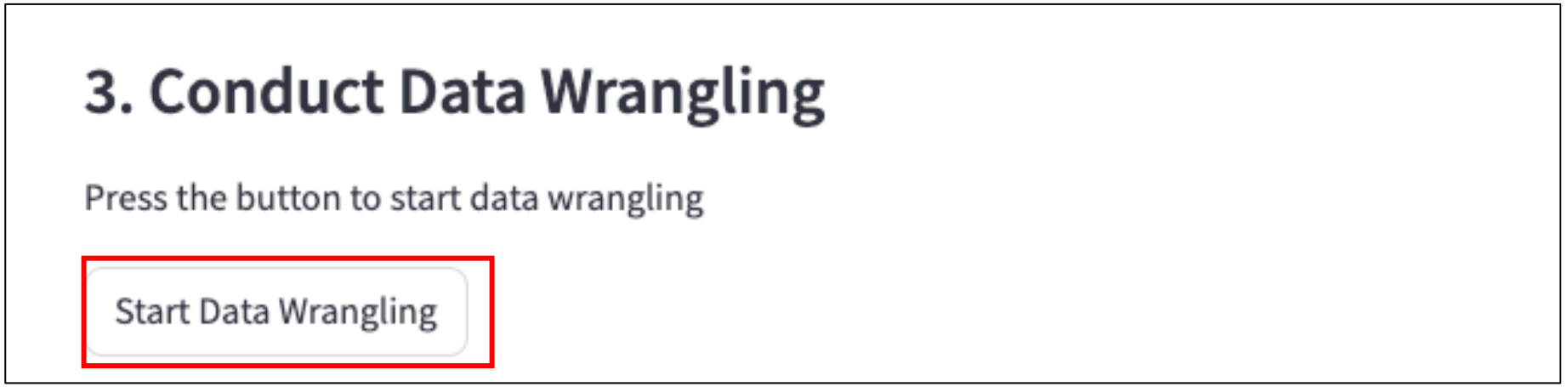
Because columns ['last_review', 'reviews_per_month'] are detected as Datatime features, new features will be generated from these Datatime features, for example, MM/DD/YYYY => MM, DD, YYYY

Columns ['price'] are string instead of numerical values. To better use the features, the string will be converted into numerical values, or the numerical values embedded in the string will be extracted.

Because columns ['price'] are detected as Embedded Number features, new features will be generated from these Embedded Number features, for example, $1,000 => 1000

- どのようなデータラングリングが適用されるかの説明がUI上に表示されます

2.5 データラングリング実行と結果（1）



- データラングリング設定後、”Start Data Wrangling”ボタンを押すと設定を元にData CleaningとData Enrichmentが開始されます

2.5 データラングリング実行と結果 (2)

FUJITSU

3. Conduct Data Wrangling

Press the button to start data wrangling

Start Data Wrangling

	neighbourhood	room_type	minimum_nights	price	name_dimension_0	name_dimensio
0	Kensington	Private room	1	149.5	-0.0475	0.0
1	Midtown	Entire home/apt	1	225	-0.0469	0.0
2	Harlem	Private	None	150	-0.0125	0.0
3	Clinton Hill	Entire home/apt	1	89.5	-0.0686	0.0
4	None	Entire home/apt	None	80	-0.046	0.0
5	Murray Hill	Entire home/apt	3	200	-0.0486	-0.0
6	Bedford-Stuyvesant	Private room	45	62.5	0.0158	0.
7	Hell's Kitchen	Private room	None	79.5	-0.0356	-0.0
8	Upper West Side	None	2	79	-0.0449	-0.0
9	Chinatown	Entire home/apt	None	150	-0.07	-0.0

- データラングリング終了後、処理されたデータセットが表示されます

2.5 データラングリング実行と結果 (3)

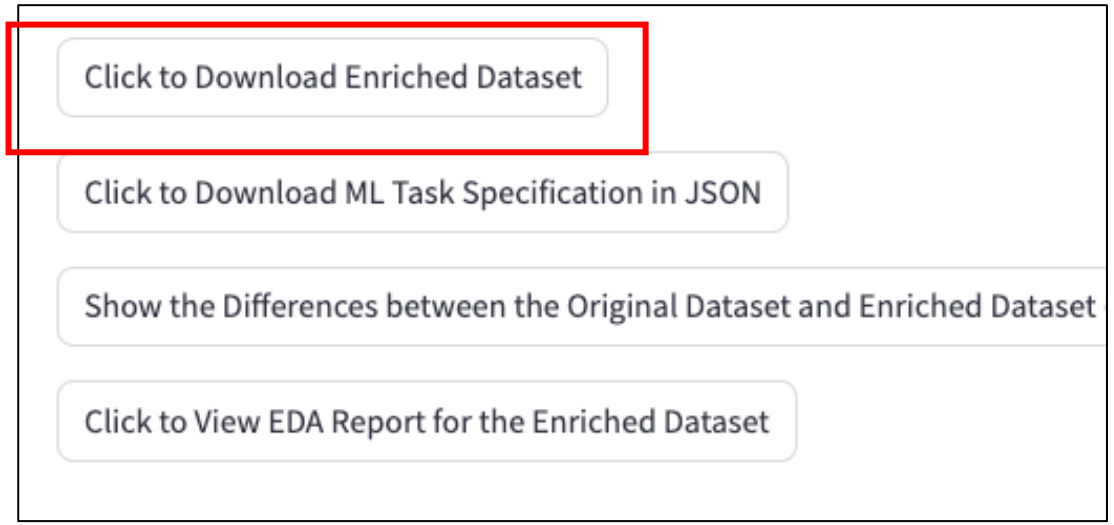
FUJITSU

Descriptive statistics

	57	last_review_Year	last_review_Month	last_review_Day	last_review_WeekDay	last_review_Hour
count	56	8,660	8,660	8,660	8,660	8,660
unique	ne	None	None	None	None	None
top	ne	None	None	None	None	None
freq	ne	None	None	None	None	None
mean	37	2,017.7411	6.3028	15.9072	3.1374	0
std	35	1.6403	2.6837	9.8054	2.2032	0
min	78	2,008	1	1	0	0
25%	16	2,016	5	6	1	0
50%	39	2,019	6	17	3	0
75%	53	2,019	8	24	5	0

- 記述統計情報も併せて表示されます

2.5 データラングリング実行と結果 (4)

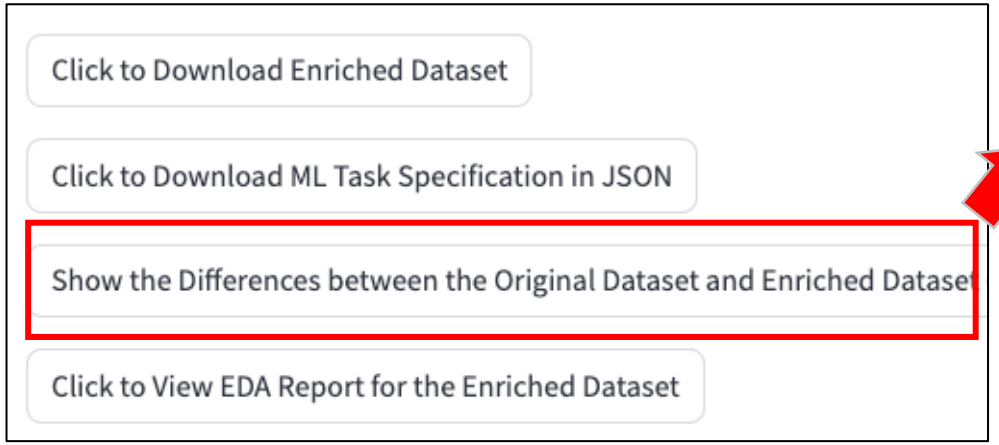


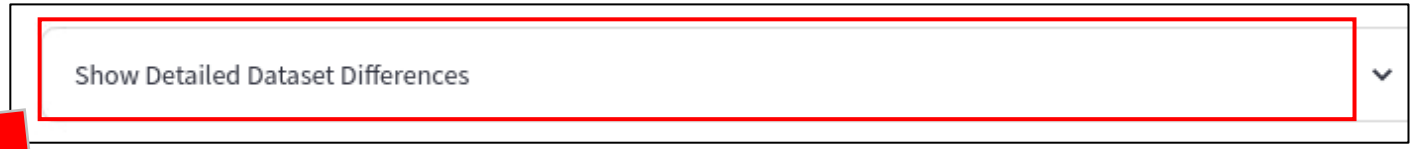
- “Click to Download Enriched Dataset”ボタンを押すと処理されたデータセットがローカルマシンにダウンロードされます。ダウンロードされたデータセットはCSV形式で、AutoML等の更なるアプリケーションに利用可能です

2.6 データラングリング実行と結果 (5)

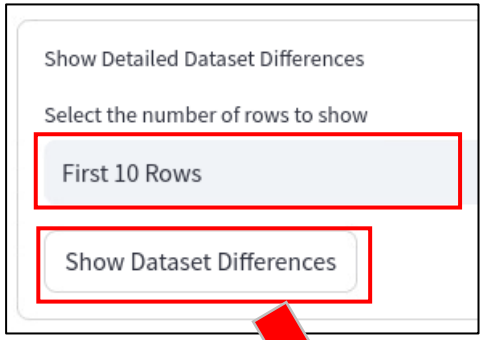
- “Click to Download ML Task Specification”ボタンを押すと機械学習タスク設定がローカルマシンにダウンロードされます。機械学習タスク設定はJSON形式で、AutoML等の更なるアプリケーションに利用可能です

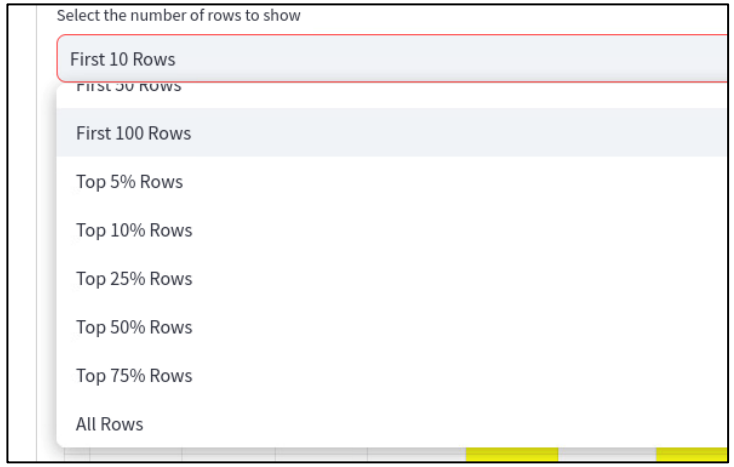
2.6 データラングリング実行と結果（6）

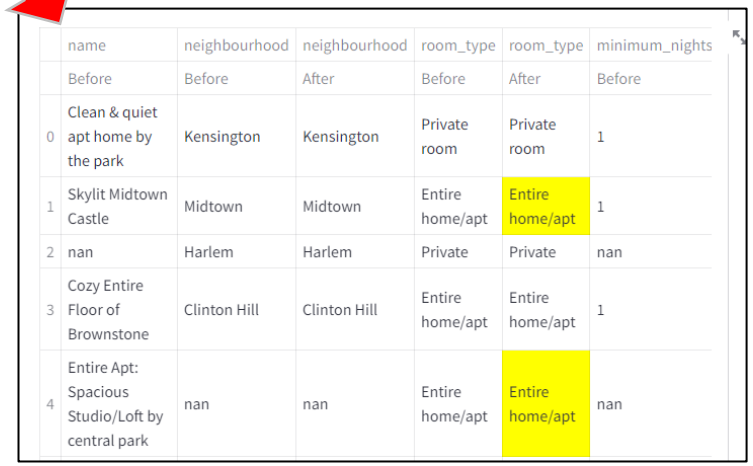




展開



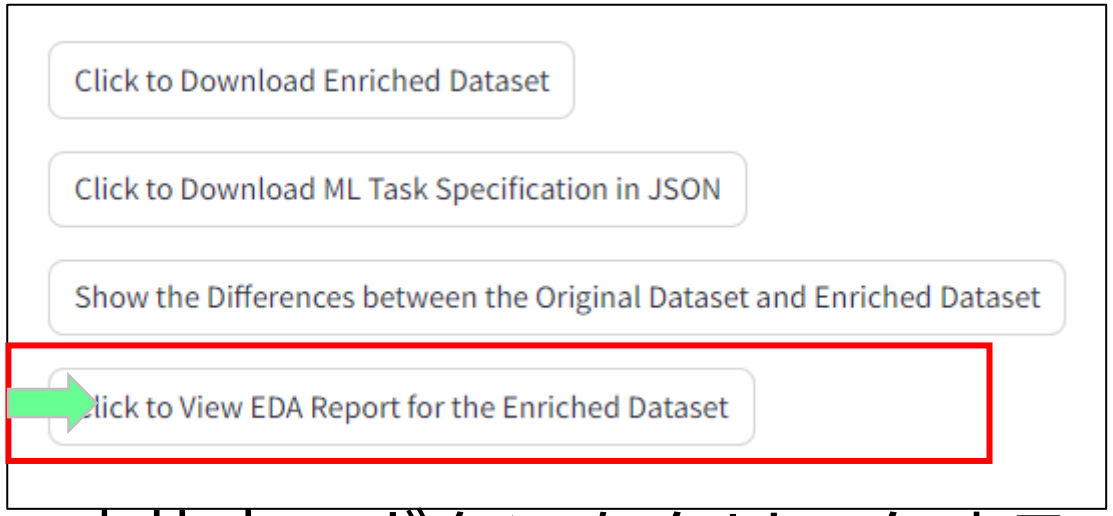




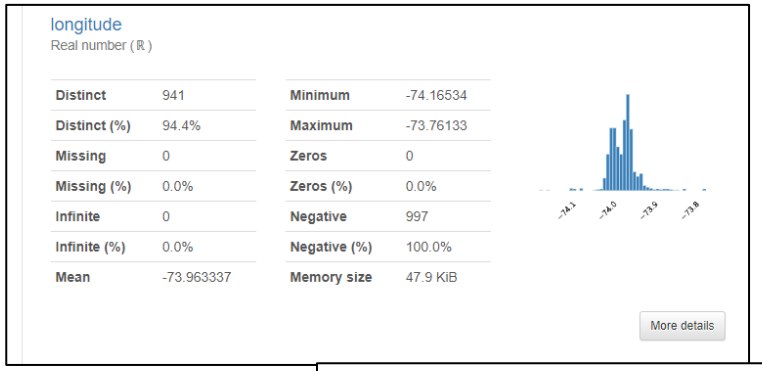
	name	neighbourhood	neighbourhood	room_type	room_type	minimum_nights
	Before	Before	After	Before	After	Before
0	Clean & quiet apt home by the park	Kensington	Kensington	Private room	Private room	1
1	Skylit Midtown Castle	Midtown	Midtown	Entire home/apt	Entire home/apt	1
2	nan	Harlem	Harlem	Private	Private	nan
3	Cozy Entire Floor of Brownstone	Clinton Hill	Clinton Hill	Entire home/apt	Entire home/apt	1
4	Entire Apt: Spacious Studio/Loft by central park	nan	nan	Entire home/apt	Entire home/apt	nan

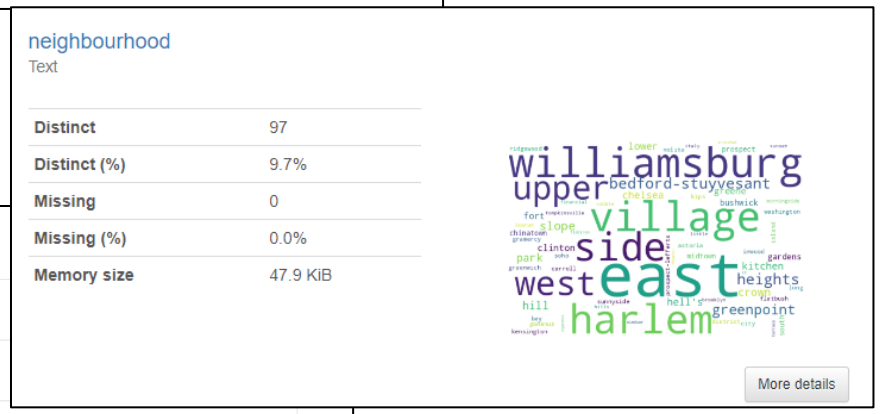
- 赤枠のボタンをクリック
  - 処理前と処理後データセットの違い(黄色で強調)が表示されます
    - デフォルトでは10列表示されますが、”Select the number of rows to show”ボタンで表示列数を変更できます。
  - カラムが除去される場合、理由が表示されます

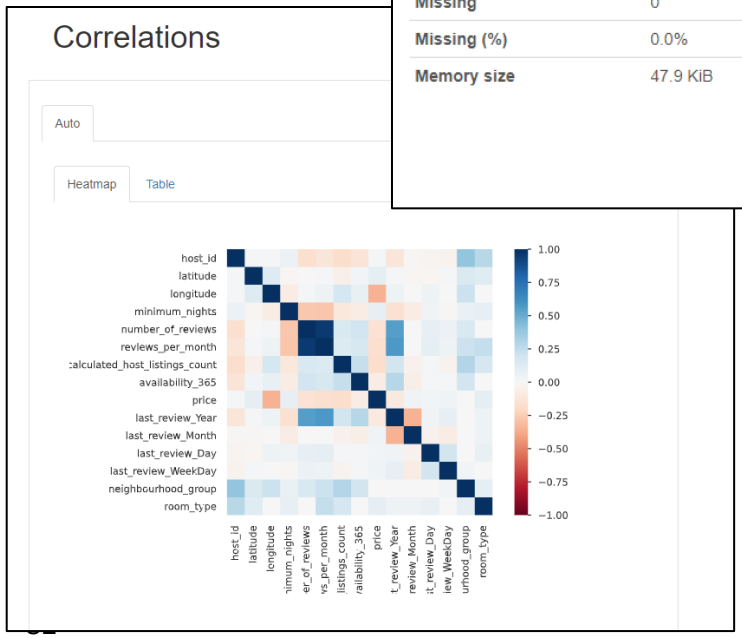
2.6 データラングリング実行と結果 (7)



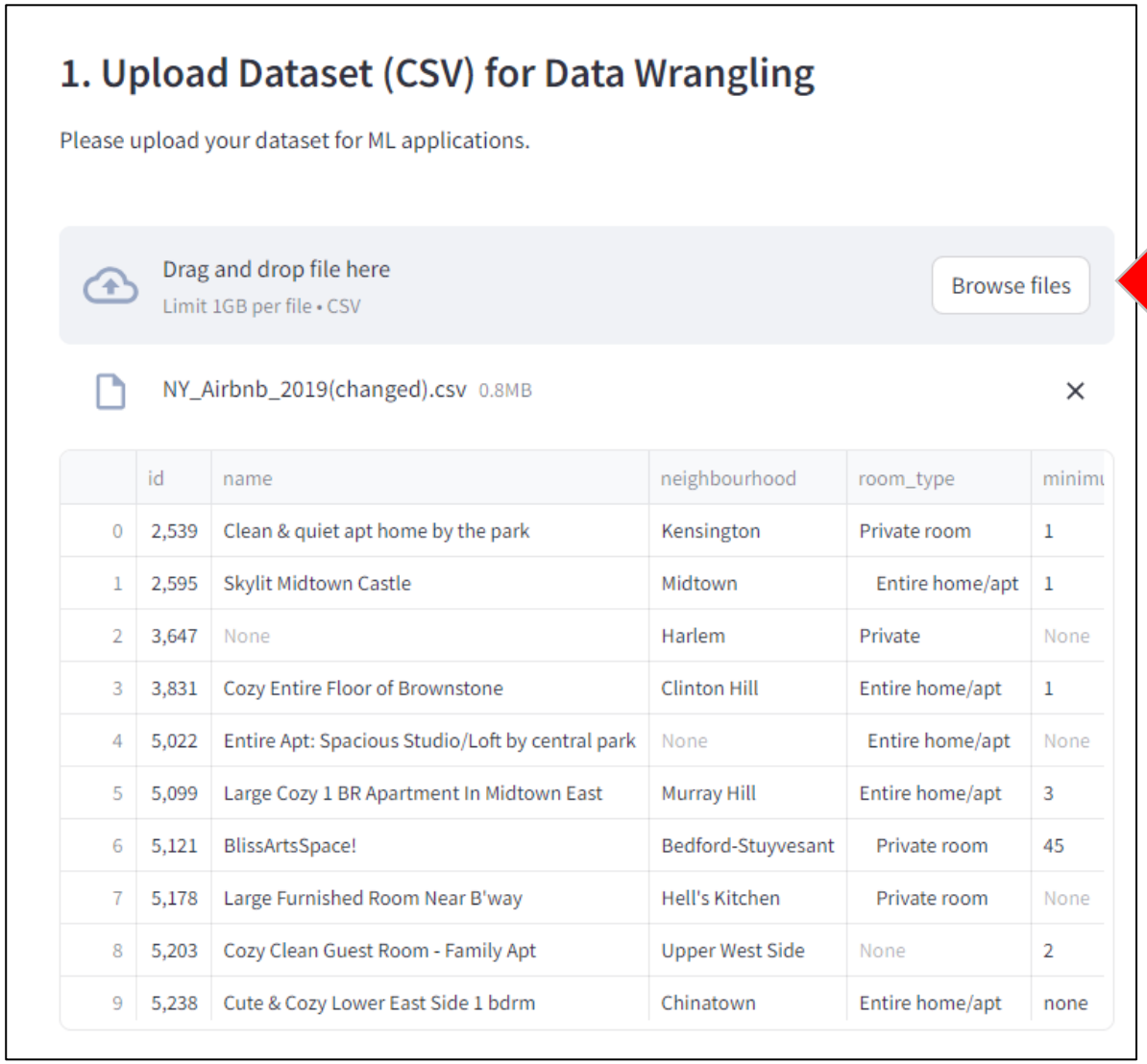
- 赤枠内のボタンをクリックすると処理されたデータセットに対してEDA結果が生成されます
  - 注: 処理後データセットに非常に多くのカラムがある場合 (例: テキスト埋め込み特徴量を追加)、EDA結果の生成に時間がかかります







新しいデータセットを試す場合



	id	name	neighbourhood	room_type	minimu
0	2,539	Clean & quiet apt home by the park	Kensington	Private room	1
1	2,595	Skylit Midtown Castle	Midtown	Entire home/apt	1
2	3,647	None	Harlem	Private	None
3	3,831	Cozy Entire Floor of Brownstone	Clinton Hill	Entire home/apt	1
4	5,022	Entire Apt: Spacious Studio/Loft by central park	None	Entire home/apt	None
5	5,099	Large Cozy 1 BR Apartment In Midtown East	Murray Hill	Entire home/apt	3
6	5,121	BlissArtsSpace!	Bedford-Stuyvesant	Private room	45
7	5,178	Large Furnished Room Near B'way	Hell's Kitchen	Private room	None
8	5,203	Cozy Clean Guest Room - Family Apt	Upper West Side	None	2
9	5,238	Cute & Cozy Lower East Side 1 bdrm	Chinatown	Entire home/apt	none

- 対象データセットに対してデータラングリングが終了し新しいデータセットを試す場合、こちらをクリックしてデータラングリングを始めてください

お問合わせ（お客様向け）

- 弊社の担当者にお問い合わせください

Thank you