

データを一から分析してみよう (Part 1)

Kaggle¹で公開されているアイオワ州住宅価格のデータセットを使って、EDA（探索的データ解析）からモデル作成までのデータ分析のプロセスをシリーズ連載していきます。

第1回目(Part1)では、データの読み込み、データ型の確認と修正まで行います。作業中のエラーやトラブルの対処方法もご紹介いたします。プログラムで行うデータ型の確認と修正については、使用する csv ファイルとサンプルプロセスをダウンロードいただけますので、ぜひお手元の RapidMiner で再現ください。



アイオワ州の住宅イメージ画像

使用データ

Ames Housing dataset

アメリカ合衆国アイオワ州エイムズ都市の住宅に関するデータセット

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview>

販売価格、外装素材、地下室の高さ、ガレージのサイズなど、79 の属性（変数）

各属性（変数）の特徴は data_description.txt を参照

分析の目的

各住宅の販売価格を予測すること

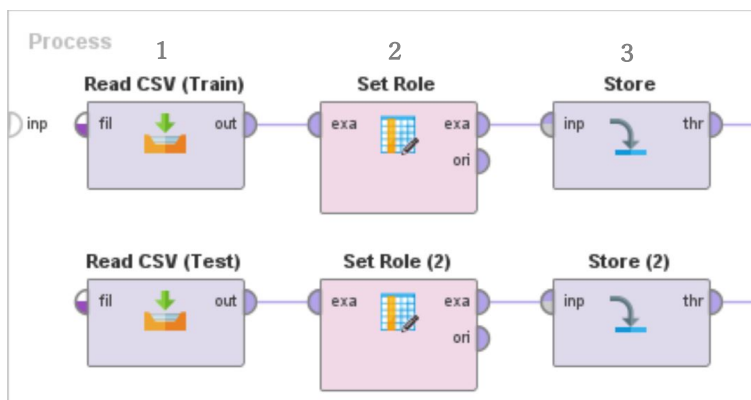
¹ Kaggle：世界最大の機械学習・データ分析コンペティションのためのプラットフォーム

Part1 で行うプロセス

- データの読み込み
- データ型の確認・修正

データの読み込み

訓練データとテストデータに属性の役割を定義し、リポジトリに格納



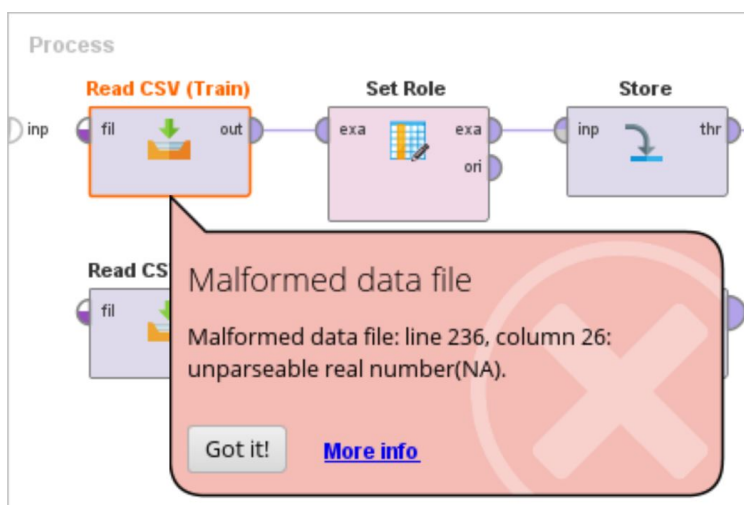
Step 1 CSV ファイルの読み込み (Kaggle からダウンロードした train.csv と test.csv)

Step 2 属性の役割を定義 → 訓練データは id と label (Id : id、SalePrice : label)
テストデータは ID のみ (Id : id)

Step 3 リポジトリに格納 → 格納したオブジェクトは他のプロセスで使用



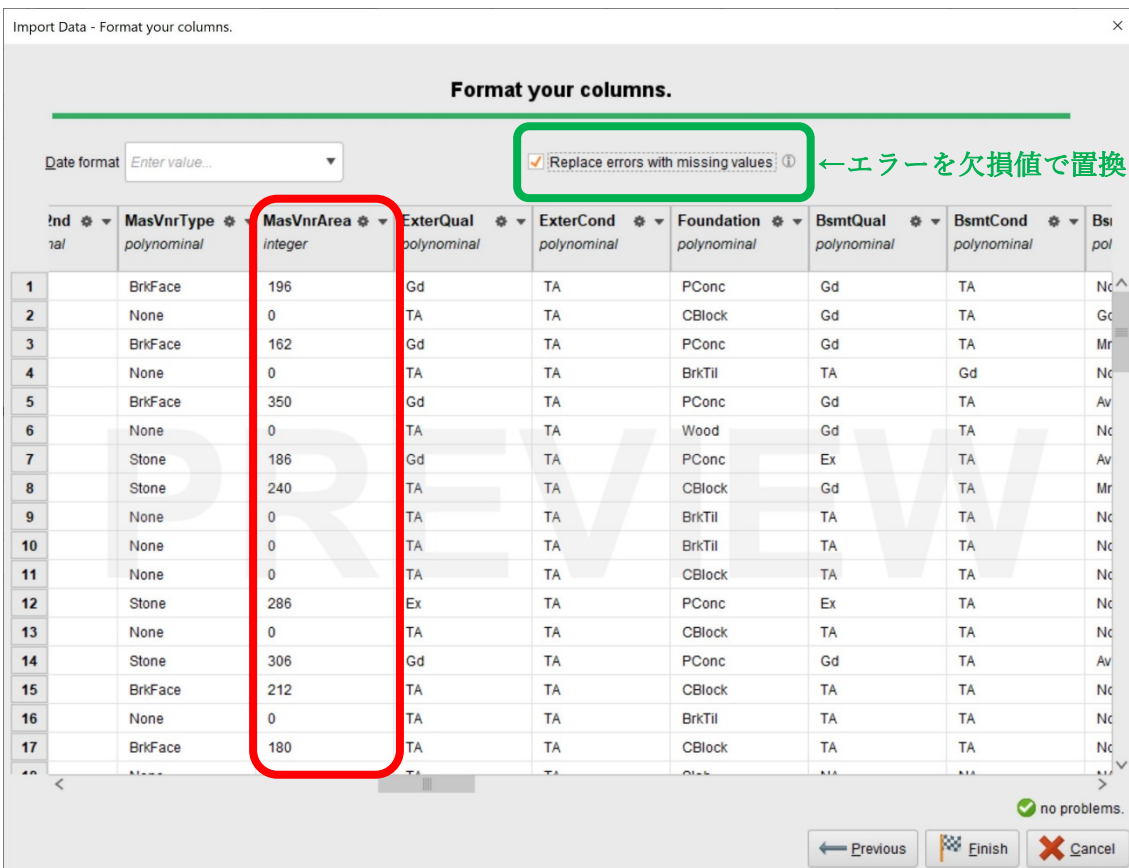
▶ 実行すると、データの読み込みでエラーが発生しました。何が起こったのでしょうか？



エラーの原因は、列 26 (MasVnrArea)のデータ型²は integer(整数)と設定されており、データの中に整数と解析できないテキストが存在していたためです。データ型が Integer の場合、データは全て整数である必要があります。

train と test の CSV ファイル読み込み時に発生したエラーに対し以下の方法で対処します。

※エラーの対処法



Import Data - Format your columns.

Format your columns.

Date format:

☒ Replace errors with missing values ⓘ ←エラーを欠損値で置換

Ind	MasVnrType	MasVnrArea	ExterQual	ExterCond	Foundation	BsmtQual	BsmtCond	Bsi
al	polynomial	integer	polynomial	polynomial	polynomial	polynomial	polynomial	pol
1	BrkFace	196	Gd	TA	PConc	Gd	TA	Nc
2	None	0	TA	TA	CBlock	Gd	TA	Gc
3	BrkFace	162	Gd	TA	PConc	Gd	TA	Mr
4	None	0	TA	TA	BrkTil	TA	Gd	Nc
5	BrkFace	350	Gd	TA	PConc	Gd	TA	Av
6	None	0	TA	TA	Wood	Gd	TA	Nc
7	Stone	186	Gd	TA	PConc	Ex	TA	Av
8	Stone	240	TA	TA	CBlock	Gd	TA	Mr
9	None	0	TA	TA	BrkTil	TA	TA	Nc
10	None	0	TA	TA	BrkTil	TA	TA	Nc
11	None	0	TA	TA	CBlock	TA	TA	Nc
12	Stone	286	Ex	TA	PConc	Ex	TA	Nc
13	None	0	TA	TA	CBlock	TA	TA	Nc
14	Stone	306	Gd	TA	PConc	Gd	TA	Av
15	BrkFace	212	TA	TA	CBlock	TA	TA	Nc
16	None	0	TA	TA	BrkTil	TA	TA	Nc
17	BrkFace	180	TA	TA	CBlock	TA	TA	Nc

no problems.

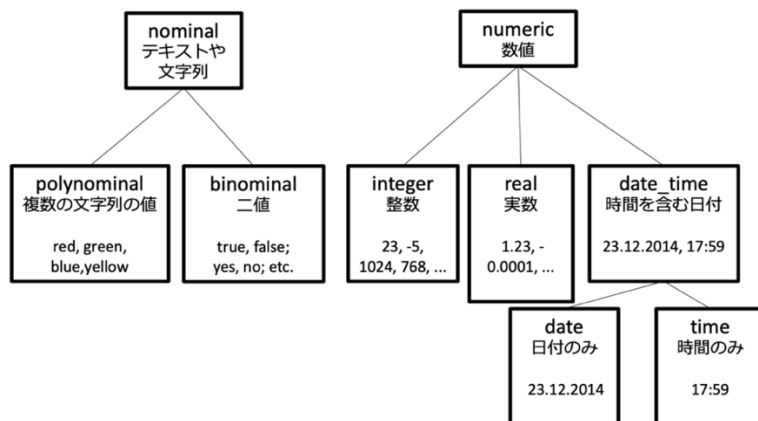
Previous Finish Cancel

データを全て整数にするため、整数以外を削除します。データの読み込みフェーズ中に、画面緑色の枠 [Replace errors with missing values](=エラーを欠損値で置換)をクリックすることで修正ができます。

² データ型：P.4 データ型とは 参照

データ型とは

テキストや文字列、数値など、列（属性）が持つデータの種類のことです。データ型にはさまざまな種類があり、数値型には色々なデータ型があることを知っておくことが重要です。



データ型の確認・修正


データの読み込み後は、RapidMiner がデータ型を正しく認識できているかを確認することが重要です。もし確認しなければ、分析結果が異なってしまう場合があります。例えば、郵便番号は住所を表す文字列ですが、データを読み込むと整数として認識されます。整数型のまま回帰式にすると無意味な関係式が成り立ってしまい、分析結果に影響しない変数であると判断されてしまいます。データを正しく取り扱うために、データ型を意識し分析対象であるデータの特徴を掴むようにしましょう。

データ型の確認方法には、手動またはプログラムで行う 2 つの方法があります。プログラムで行う方法は応用になります。いずれの方法も Kaggle の data_description.txt からデータの特徴と照らし合わせ、データ型に誤りがないかを確認します。

■手動で確認・修正

データ型は、データ読み込みフェーズ(P.3)または読み込み後に基本統計量から確認できます。

データ型を修正する方法には、以下の 2 つがあります。

- ・データ読み込みフェーズで属性名の右横  からデータ型を変更(P.3 画面参照)
- ・基本統計量でデータ型を確認し、プロセス画面で対象オペレータを配置しデータ型を変更

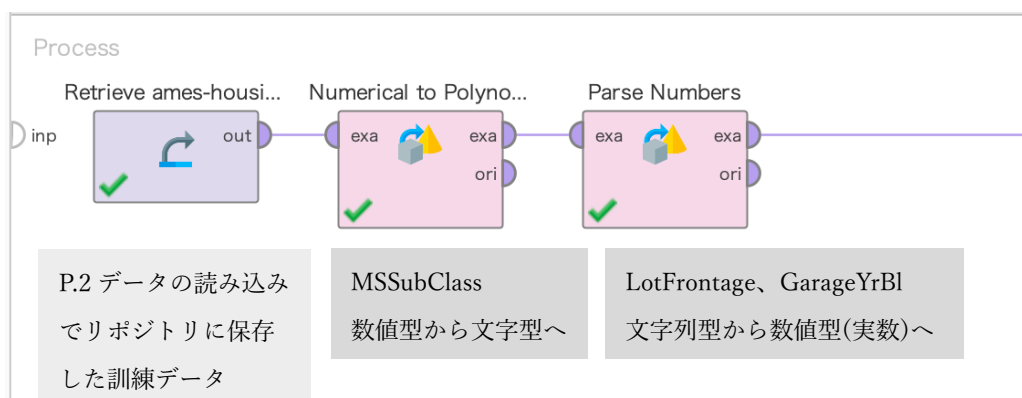
結果画面の基本統計量でデータ型を確認すると、3つの属性のデータ型に誤りがあることが分かりました。

▼ MSSubClass	Integer
▼ LotFrontage	Nominal
▼ GarageYrBlt	Nominal

各属性には以下の特徴があることから、正しいデータ型へ変更する必要があります。

- ・ MSSubClass : Identifies the type of dwelling involved in the sale. (Nominal 型)
- ・ LotFrontage : Linear feet of street connected to property (Numeric 型)
- ・ GarageYrBlt : Year garage was built (Numeric 型)

選択した属性のデータ型を修正



<各パラメータの設定>

Parameters

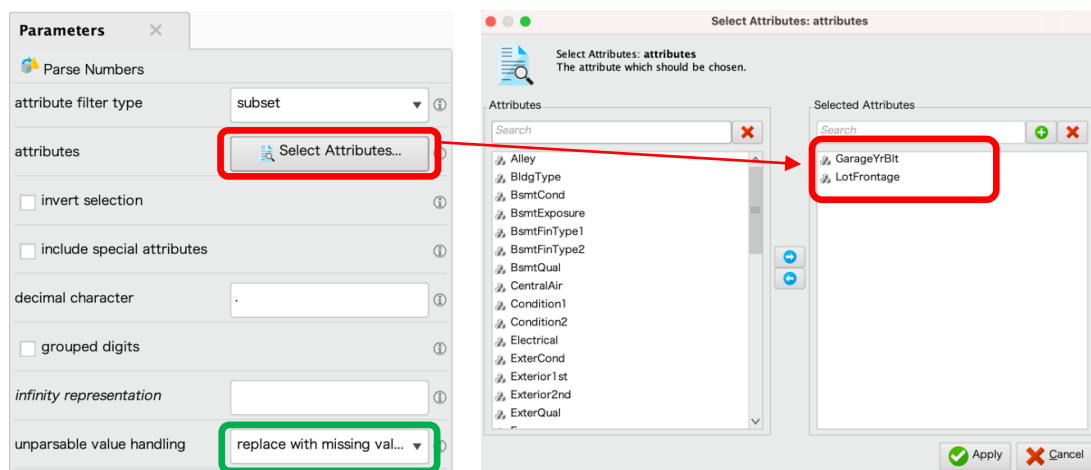
Numerical to Polynomial

attribute filter type single

attribute MSSubClass

☐ invert selection

☐ include special attributes



数値と解釈できない値を欠損値として処理するため[replace with missing values]を選択

■プログラムで確認・修正

今回のようにデータに属性が多い場合は、プログラムで実行することで効率的にデータ型の確認と修正を行うことができます。作成したプロセスは今後予測モデルに新規データを適応させる際にも再利用することができます。

手順としては、まずデータの特徴を確認しながら各属性にデータ型を振り分け“正しいデータ型リスト”を作成します。そして、正しいデータ型リストと RapidMiner に認識されたデータ型を照合させ、エラーがあれば修正するプロセスを作成しプログラムを実行します。正しいデータ型リストとプロセスはダウンロードいただけますので、RapidMiner を動かしながらプロセスをご確認いただけたらと思います。

【事前準備】

正しいデータ型リストの作成方法

data_description.txt でデータの特徴を確認しながら、csv ファイルを作成します。属性(name)とデータ型(correct_type)を以下のように手入力します。

	A	B	C
1	name	correct_type	
2	MSSubClass	polynomial	
3	MSZoning	polynomial	
4	LotFrontage	integer	
5	LotArea	integer	

※正しいデータ型リスト(csv ファイル)はダウンロードいただけます

オペレータのインストール方法

オペレータ一覧にないオペレータは、RapidMiner メニューバーにある

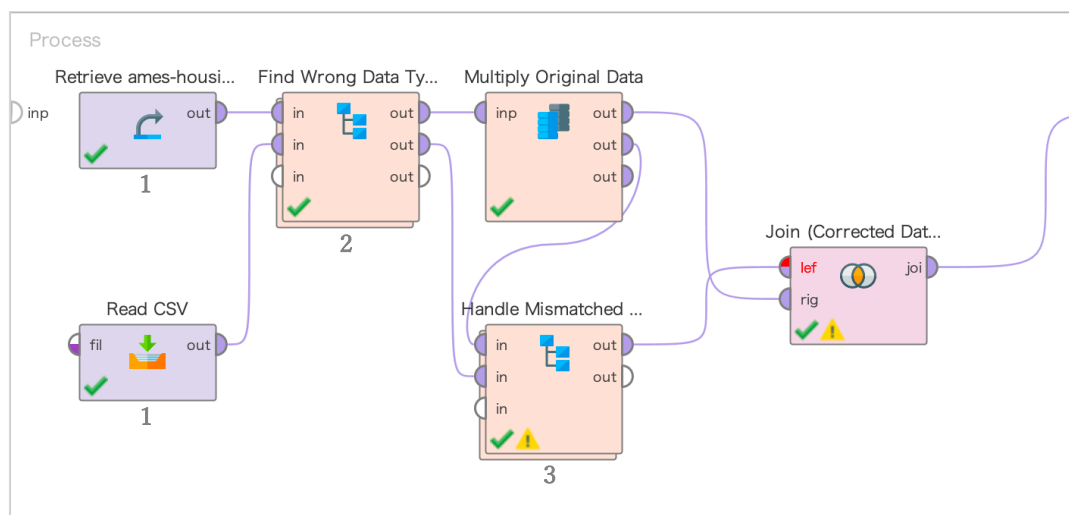
Extensions(拡張機能)>Marketplace から検索しインストールいただけます。

(拡張機能のインストール方法は[こちら](#))

※Extract Statistics オペレータは、Operator Toolbox をインストールするとご使用いただけます。Operator Toolbox は Text Processing と依存関係があるため、「Text Processing」の拡張機能もインストールする必要があります。

(Operator Toolbox が利用できない時の対処法は[こちら](#))

【プロセスの作成】



以下の手順でプログラムを組みます。

Step 1 訓練データと、正しいデータ型リスト(CSV ファイル)の読み込み

Step 2 間違ったデータ型を見つけるサブプロセス

Step 3 選択した属性の数値を解析し、データ型を修正するサブプロセス

※プロセスはダウンロードいただけます。各オペレータの動作は、ブレイクポイント(オペレータを右クリックしてブレイクポイント(後)または F7)を置いて確認してみてください。

プロセスのインポート方法：RapidMiner メニューバーの File>Import Process>ファイル指定

終わりに

RapidMiner がどのようにデータ型を認識し、どのような問題が発生するのかを学んでいただけたと思います。データ型がすべて正しく修正できたので、EDA や前処理を行うことができます。次回は、欠損値の扱いに焦点を当てて、EDA と前処理を続けていきたいと思っています。