2024年7月
Fracta コーディングテスト

課題概要

本課題は、フラクタの実務と実データに沿って作成されたものです。通常クライアントが提供する生データに近いものを処理し、データに異常等がないことを確認することがタスクとなります。

顧客から管路データ(GM2022_assets.csv)と漏水データ(WO_EXPORT.csv)をいただいているとします。このデータを読み込み、下記の通り2つの出力ファイルを作成し、データ処理のために使用したJupyter Notebookを提出してください。また、課題を進めている中で自分の理屈や意思決定の説明をNotebook上に明記してください。

管路ファイル　(Pipe.csv)

Column	Data type	Format	Description
native_pipe_id	str		顧客が使用しているパイプ管理ID。重複してもOK。
asset_id	int		フラクタシステム内に使用されるユニークな管理ID。重複してはならない。
material	str		{‘DIP’, ‘CAS’, ‘PVC’, ‘AC’, ‘SP’}　の中の一つ。
install_year	str	YYYY-MM-DD	布設年度
diameter	float		口径（mm）
abandoned	bool		撤去ステータス

漏水ファイル　(Break.csv)

Column	Data type	Format	Description
native_pipe_id	str		顧客が使用しているパイプ管理ID。重複してもOK。
asset_id	int		システム内に使用されるユニークなパイプ管理ID。漏水事故を管路に紐づける際に使われる。
break_date	str	YYYY-MM-DD	漏水発生日

注意点:

- 有り得ない数値は NULL に変更したままでよい。
- 上記の必然カラム以外の情報があれば、そのままデータに残してもよい。最低上記のカラムがデータにあればOK。（必然データ意外の情報を保存しておくと、また後で使えることもある）。
- カラム名は上記のものと完全に合致しないとアルゴリズムに通すことができない。
- システムで使用されている管種IDの意味は下記の通り:
    - DIP:ダクタイル鋳鉄管
    - CAS:ネズミ鋳鉄管
    - AC :石綿管
    - PVC:塩化ビニル管
    - SP:鋼管
- 漏水事故は自然に発生した漏水に絞る必要がある。（モデリングの観点からは、入力変数とターゲット変数の相関関係があるのが前提）。

今回の分析では、下記の情報を既に顧客から得たとする:

- 使用している管種コードは下記の通り:
    - CIP: ネズミ鋳鉄
    - PVC: 塩ビ管
    - DUC: ダクタイル (＋継ぎ手種類, 耐震化状況によりGかK)
    - ACON: アスベスト管
    - STW: 鋼管（ポリスリーブ有り）
    - STP: 鋼管（エポキシ樹脂有り）
- 布設年は和暦で記録されているため、西暦に変更する必要がある。