Complex pipelines in Sklearn

sages

kodo/amacz

Ros Apostol

O mnie

Ros Apostol

Data Scientist w NorthGravity

Certyfikowany specjalista ML

Trener w Sages

https://www.linkedin.com/in/apostolros/

sages

Agenda

1. Co to jest pipeline?
2. Omówienie zbioru danych
3. Określenie typów transformacji na zbiorze danych
4. Implementacja poszczególnych transformatorów
5. Budowa pipeline
    - Pipeline
    - Custom transformers
    - Feature Union
    - ColumnTransfomers

Pipeline

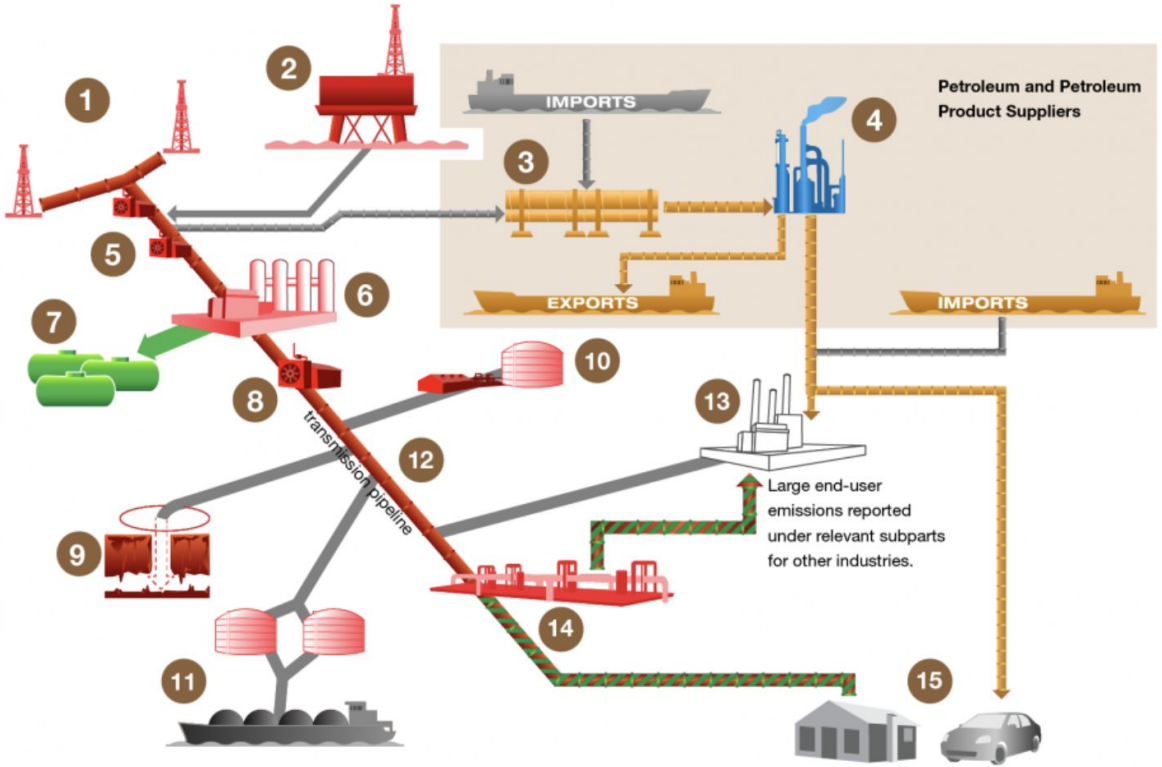
Pipeline to nie takie proste

Wejście:

Mieszanka z ziemi

Wyjście:

Nasz samochód jeździ



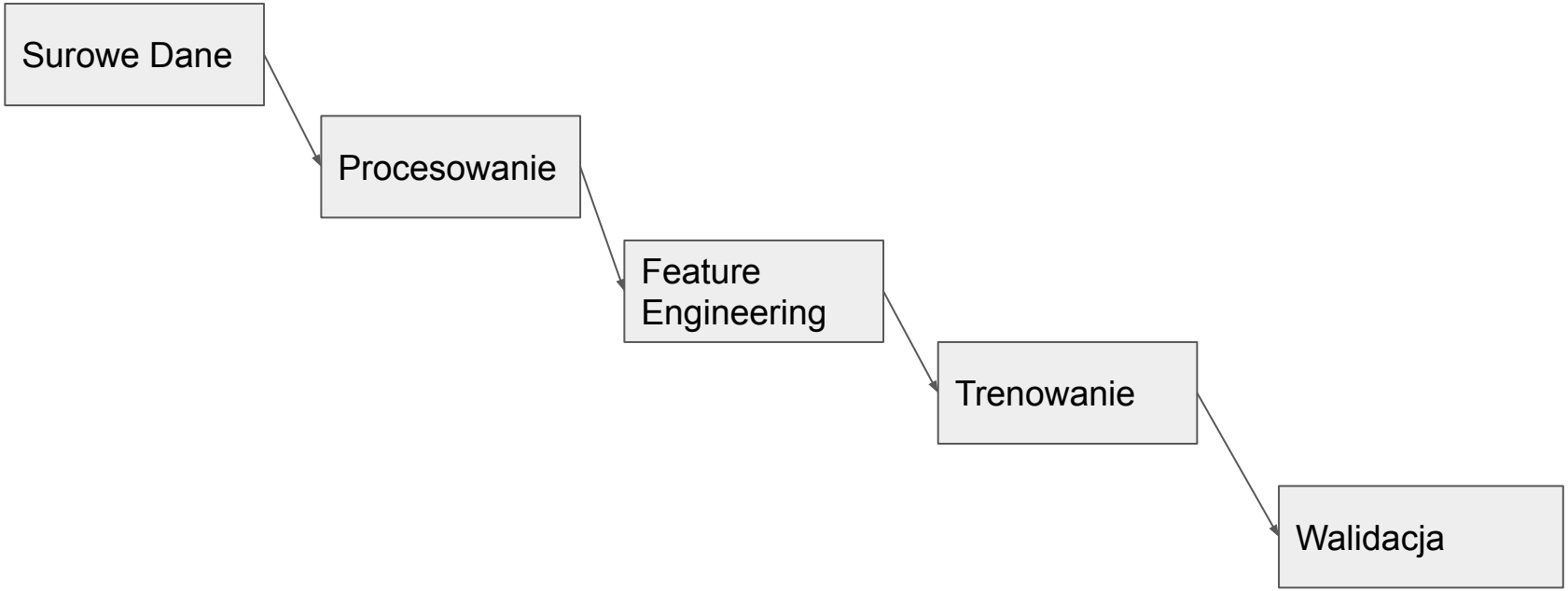
Życie Data Scientista

Życie Data Scientista średnio na 60-70% składa się z przygotowania i procesowania danych.

Im więcej czasu poświęcamy na czyszczenie i przygotowanie danych, tym mniej zostaje na rzeczy naprawdę ciekawe.

Jak zrobić życie Data Scientista łatwiejszym?

Jak to się przekłada na dane?



Do czego pipeline’y służą?

1) Automatyzacja procesu przetwarzania danych
2) Zapobieganie wycieku danych

Do czego to prowadzi?

1) Lepsze utrzymanie procesów przetwarzania danych
2) Bardziej niezawodne modele

Zbiór danych Titanic

1. Jest kultowy
2. Przedstawia dane, które możemy łatwo zrozumieć
3. Jest różnorodny, jeśli chodzi o dane:
    - Kolumny numeryczne
    - Kolumny kategoryczne
    - Wartości brakujące dla obu typów kolumn
    - Wartości odstające itd.

Teoria i przegląd dostępnych modułów

Biblioteka - Sklearn

- Pipeline
- ColumnTransformer
- FeatureUnion
- Custom Transformer

Alternatywy: Feature Engine

Struktura projektu

1. Pobieranie danych
2. Analiza eksploracyjna - zrozumienie danych
3. Budowa pipeline’u
    a. Wartości brakujące
    b. Wartości odstające
    c. Skalowanie danych
    d. Redukcja danych
    e. Feature Engineering
    f. Enkodowanie zmiennych kategorycznych
    g. Trenowanie modelu
    h. Walidacja modelu

Cel projektu

Zbudować pipeline:

- Na wejście: podajemy surowe dane.
- Na wyjściu: mamy predykcję i dokładność modelu.

Cała magia tworzy się w środku.

Pipeline Sklearn

1) Pipeline - ustrukturyzowany ciąg transformacji na pewnym zbiorze danych.
2) Kroki - poszczególne transformacje datasetu. Każdy tranformer musi mieć dwie metody: fit() i transform()
3) Ilość kroków w pipelinie - dowolna.
4) Ostatni krok może być modelem - implementuje dwie metody: fit i predict.
5) Obiekt pipelinu można stosować na innych danych o podobnej strukturze (pipeline trenowany na train set, a wykonywany ponownie na test set).

Custom Transformer

1) Pozwala zaimplementować dowolną transformację danych
2) Może być krokiem w pipeline.
3) Musi implementować dwie metody: fit() i transform()
4) Może dziedziczyć BaseEstimator, TransformerMixin dla kompatybilności z obiektem Pipeline.

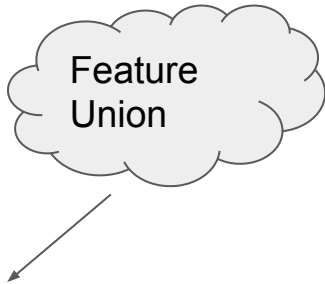
Feature Union

Pozwala połączyć razem dwa datasety w stylu ‘horizontal stack’

Warunek - ilość wierzy w obu datasetach musi się zgadzać.

Zastosowanie:

1) Procesujemy kolumny numeryczne => dataset_1
2) W inny sposób procesujemy kolumny kategoryczne => dataset_2
3) Łączymy dwa datasety razem => dataset = dataset_1 + dataset_2
4) Trenujemy model i robimy predykcje



Column Transformer

Pozwala stosować określone typy transformacji na określonych kolumnach.

Przykład:

1) Transformacje specyficzne dla kolumn numerycznych zastosować na kolumnach numerycznych
2) Transformacje specyficzne dla kolumn kategorycznych zastosować na kolumnach numerycznych
3) Wytrenować model

Columns Transformer vs. Feature Union

Feature Union - łączenie różnych reprezentacji tego samego datasetu.

Przykład:

Połączyć ilość macierzy i macierz ważności słów tego samego tekstu w jedną macierz, a następnie trenować model klasyfikacji dokumentów.

Podsumowanie

1) Zbudowaliśmy pipeline, gdzie na wejście podajemy surowe dane, a na wejściu otrzymujemy predykcje.
2) Pipeline’y pomagają zautomatyzować procesowanie danych i zapobiec wycieku danych.
3) Umiemy implementować customowe transformery i wkładać je jako kolejne kroki pipeline’u.
4) Feature Union łaczy dwa datasety o takiej samej ilości wierszy.
5) ColumnTransformer pozwala stosować różne transformacje dla różnych kolumn.

Dziękuję!
Pytania?