Wydział Nauk Ekonomicznych i Zarządzania
Studia podyplomowe - Data Science w Biznesie
Raport: Projekt Transformatory
Maria Kowalska
Grudzień 2020

Spis treści

1 Cel biznesowy

Przykładowe przedsiębiorstwo E dysponując danymi historycznymi dotyczących awarii transformatorów oraz wykorzystując naukę o danych chce:

1. wskazanie najważniejszych czynników zwiększających prawdopodobieństwo zaistnienia awarii transformatora,
2. ustalenie progu dotarcia (odsetka populacji transformatorów), przy którym wykorzystanie predykcji awarii jest opłacalne (lepsze niż objęcie konserwacją zapobiegawczą całej populacji),
3. określenie optymalnej ze względu na koszty polityki konserwacji dla 20 kolejnych urządzeń w stanie Georgia

2 Eksploracyjna analiza danych

2.1 Opis danych

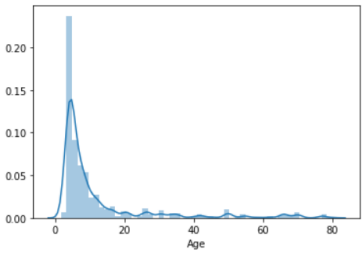
Dane do wykonania projeku 'awarie_transf_hist.xlsx' nie zawierają ani pustych wartosci czy też powtarzających się wierszy.
Jednak 2 cechy miały błedne wartosci: AssetZipCode oraz AssetCity.
Folder './krok1' zawiera:

1. naiwną analize cech z opisem "./krok1/analiza_zmiennych.xlsx"
2. notebook: './krok1/Eksploracyjna analiza danych.ipynb'

2.2 Dane ilosciowe, opis, wizualizacja, modyfikacja

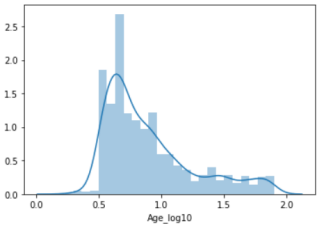
Do danych ilosciowych zaliczam: ['AssetZip','Lat', 'Long', 'AvgRepairCost', 'Age']

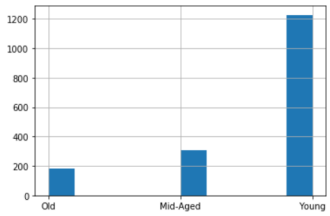
1. Age - cecha o rozkładzie:



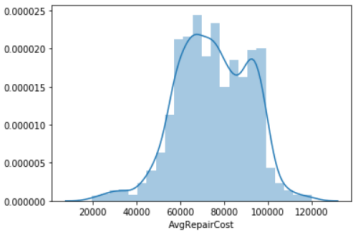
Transformatory mamy w przedziale wiekowym do 80 lat. Większosc

transformatorow jest poniżej sredniej wieku. Cechę poddałam 2 obróbkom, logarytmowaniu i podzia na kubełki, i zapisałam 2 dwoch różnych zbiorach danych odpowiednio.
Rozkład cechy po zamianie logarytmicznej oraz podziale na przedziały wiekowe- [Old, Mid, Young]:

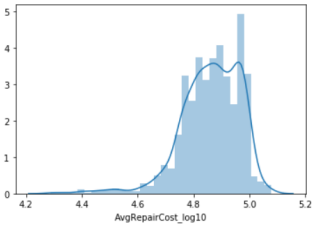


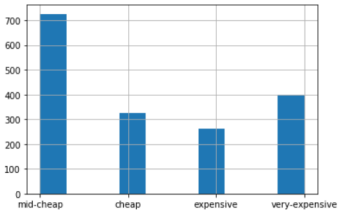


2. AvgRepairCost - cecha o rozkładzie:

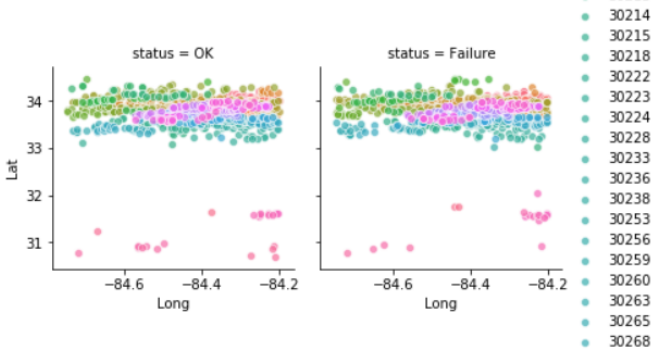


Podobnie jak cecha Age, poddana została logarytmizacji oraz podziałowi na kubełki i zapisana w osobnych zbiorach odpowiednio:





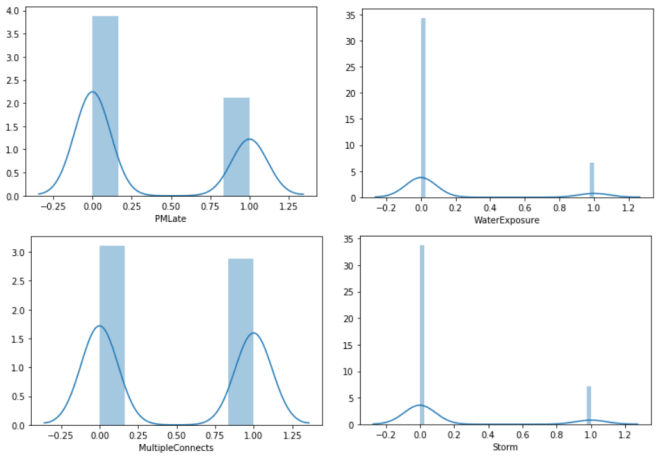
3. Dla cechy: Lat, Long i AssetZip zauważalna jest zależnosć:



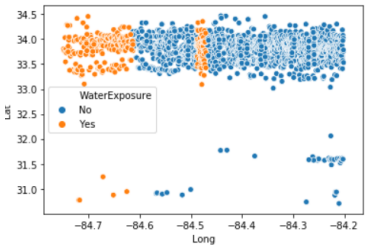
2.3 Dane jakosciowe,opis, wizualizacja, modyfikacje

Zbiór cech możemy podzielić na:

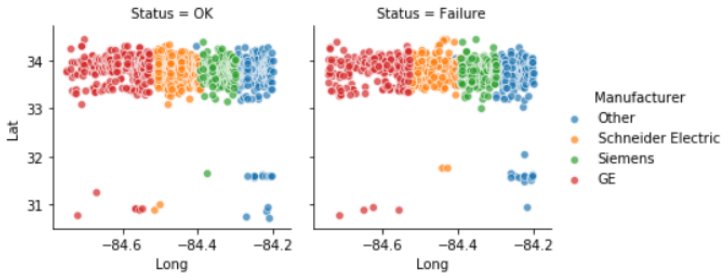
1. Binarne : 'PMLate','WaterExposure','MultipleConnects','Storm'



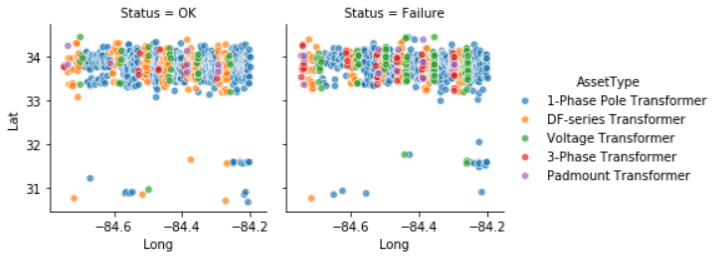
Cechy: Storm i WaterExposure nierówno rozłożone.
Dodatkowo widoczna jest zależnosć położenia (LongxLat) a cechy WaterExposure



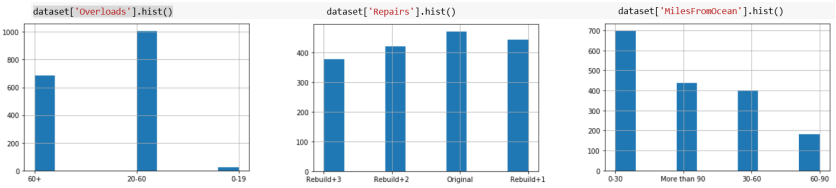
2. Nominalne:'Manufacturer','AssetType'
Mamy 4 kategorie producentów transformatorów, których umieszczenie zależy od położenia:



5 typów transformatorów, a ich położenie:



3. Porządkowe: 'Overloads', 'MilesFromOcean','Repairs'
   Widoczny nierównomierny rozkład dla cech 'Overloads' oraz 'MilesFromOcean'



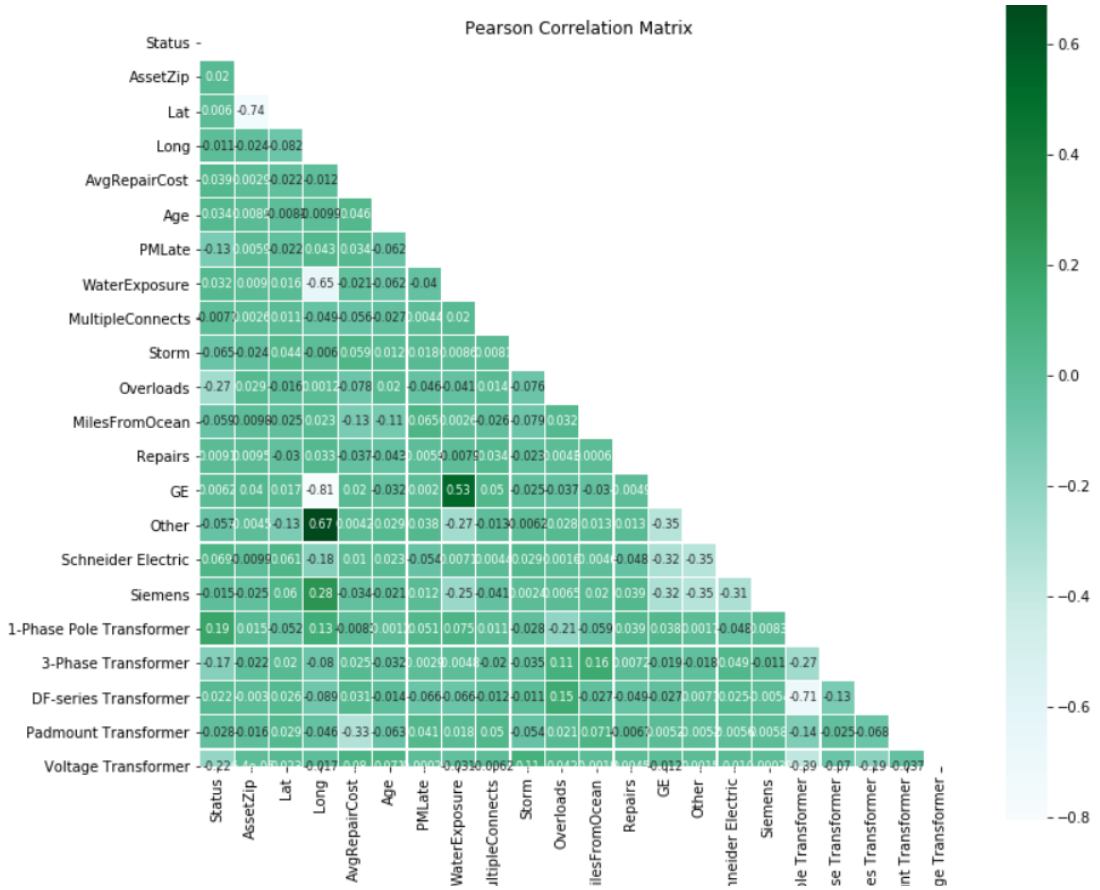
2.4 Cechy usunięte

1. AssetId - unikalna wartosc dla każdego wiersza w zbiorze
2. AssetLocation - ulica, za duzy chaos informacyjny, trudno sprawdzić czy adresy są poprawne
3. AssetCity - błędne nazwy miast w danych
4. AssetState - w każdym wierszu ta sama wartosc: GA

2.5 Korelacje

Z macierzy korelacji można wyczytać:

1. Cechy: 'PMLate' i 'Storm' są najmocniej skorelowane ze zmienną celu 'Status'
2. Najsłabiej skorelowana ze zmienną celu 'Status' jest cecha 'Lat'.
3. Cechy najbardziej skorelowane ze sobą: 'Long' z 'WaterExposure', 'Other' oraz z 'GE'; 'AssetZip' z 'Lat'



3 Podział na zbiór treningowy i testowy, trenowanie drzewa

Modelem klasyfikującym jest drzewo z pakietu sklearn. Podział na zbiór treningowy i testowy z próbą zachowania rozkładu cech nominalnych nierówno rozłożonych w zbiorze.

W folderze krok2 znajdują się:

1. Zbiory danych przygotowane w 'krok1': 'zb1.xlsx', 'zb2.xlsx', 'zb3.xlsx'
2. Notatnik: 'dobór cech i trenowanie drzewka.ipynb' z analizą treningu drzewa na podstawie doboru cech i parametrów.
3. Raport z wynikami: 'raport.xlsx'

3.1 Trenowanie klasyfikatora

Klasyfikator był trenowany na podstawie różnych parametrów:

1. Dobór cech na podstawie korelacji
2. Manualny dobór cech
3. Rekurencyjny dobór cech
4. Dostrajanie parametrów, pojedynczo: max_depth, min_samples_split, min_samples_leaf, max_features

Wszystkie wyniki zapisałam w raport.xlsx.

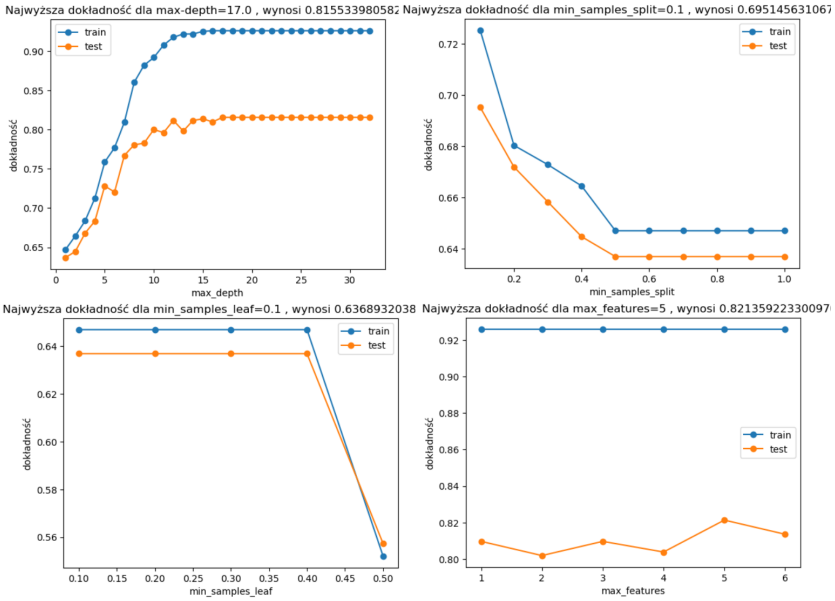
5 najlepszych wyników wg. miary dokładnosc w tabelce:

	Cechy	zbiór	parametry	dokładność
2	['Age', 'Schneider Electric', 'Siemens', '1-Phase Pole Transformer', '3-Phase Transformer', 'Overloads', 'MilesFromOcean']	zb0	manualnu dobór cech, dobór cech rekurencyjny	0.815533981
3	['PMLate', 'WaterExposure', 'Storm', 'AssetZip', 'AvgRepairCost', 'Age', 'Schneider Electric', 'Siemens', '1-Phase Pole Transformer', '3-Phase Transformer', 'Overloads', 'MilesFromOcean']	zb0	manualnu dobór cech, max-depth= 12.0	0.811650485
4	['PMLate', 'WaterExposure', 'Storm', 'AssetZip', 'AvgRepairCost_log10', 'Age_log10', 'Schneider Electric', 'Siemens', '1-Phase Pole Transformer', '3-Phase Transformer', 'Overloads', 'MilesFromOcean']	zb1	manualnu dobór cech, max-depth= 12.0	0.80776699
5	['PMLate', 'WaterExposure', 'Storm', 'AssetZip', 'AvgRepairCost', 'Age', 'Schneider Electric', 'Siemens', '1-Phase Pole Transformer', '3-Phase Transformer', 'Overloads', 'MilesFromOcean']	zb0	manualnu dobór cech, max-depth= 11.0	0.803883495
6	['PMLate', 'WaterExposure', 'Storm', 'AssetZip', 'AvgRepairCost_log10', 'Age_log10', 'Schneider Electric', 'Siemens', '1-Phase Pole Transformer', '3-Phase Transformer', 'Overloads', 'MilesFromOcean']	zb1	manualnu dobór cech, max-depth= 11.0	0.8

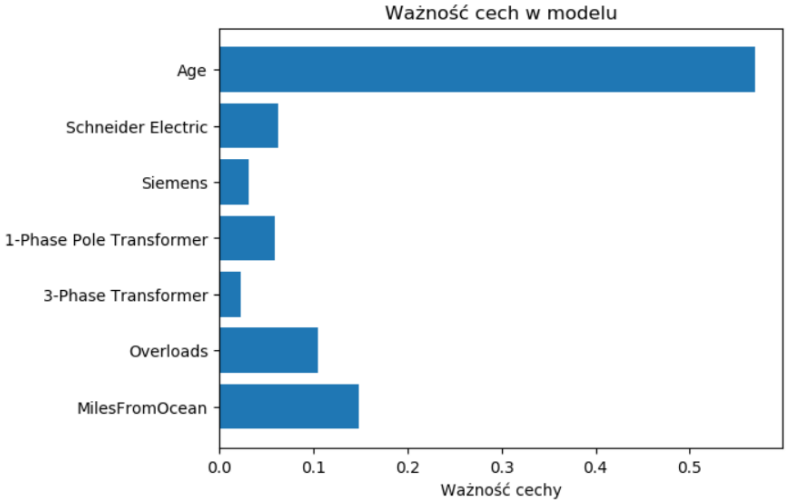
4 Wpływ cech uczących oraz parametrów na dokładnosć

W kroku3 dla 3 najlepszych wyników próbuję jeszcze różnych parametrów, plus obrazuję wagę cech dla klasyfikatora. Nazwy notatników: 1,2,3.

1. Przykład 1 z tabeli wyników, na zbiorze zb0 = zb1.xlsx, wyniki zmiany parametrów :



dokładność się poprawia przy ograniczeniu parametwó do 5. Najważniejsze cechy to: Age, MilesFromOcean, Overloads

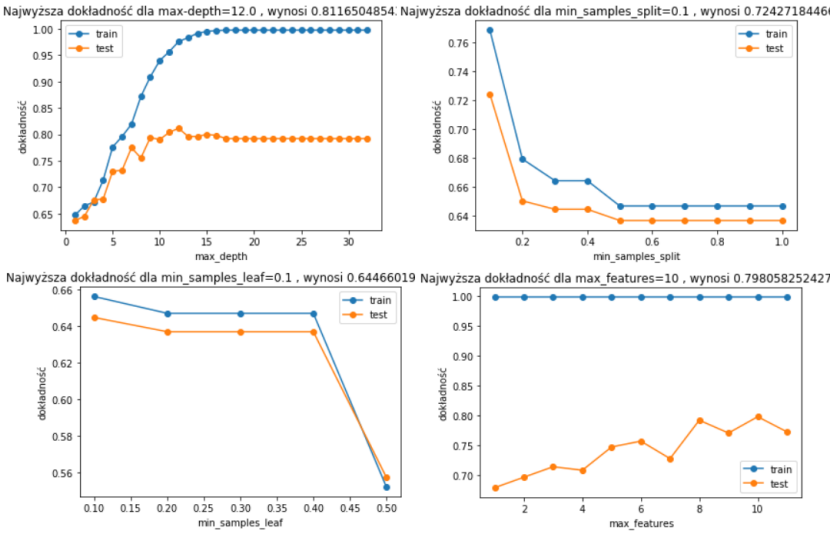


Drzewko z parametrem max-depth=2 i parametrem max_features=5,

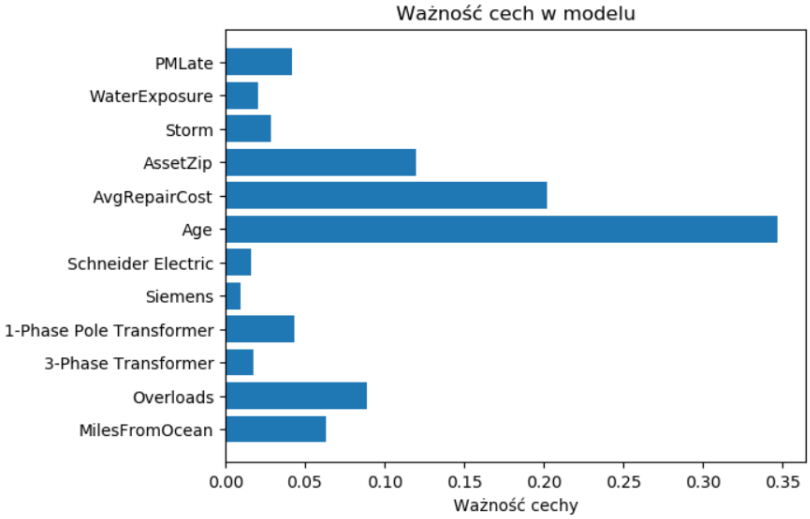
prezentuje się następująco:

```
|--- Overloads <= 0.22
|   |--- Age <= -0.40
|   |   |--- 3-Phase Transformer <= 2.18
|   |   |   |--- truncated branch of depth 15
|   |   |--- 3-Phase Transformer >  2.18
|   |   |   |--- class: 0
|   |--- Age >  -0.40
|   |   |--- MilesFromOcean <= 1.06
|   |   |   |--- truncated branch of depth 13
|   |   |--- MilesFromOcean >  1.06
|   |   |   |--- truncated branch of depth 11
|--- Overloads >  0.22
|   |--- Age <= 1.41
|   |   |--- 1-Phase Pole Transformer <= -0.20
|   |   |   |--- truncated branch of depth 13
|   |   |--- 1-Phase Pole Transformer >  -0.20
|   |   |   |--- truncated branch of depth 12
|   |--- Age >  1.41
|   |   |--- Age <= 3.80
|   |   |   |--- truncated branch of depth 4
|   |   |--- Age >  3.80
|   |   |   |--- truncated branch of depth 6
```

2. Przykład 2 z tabeli wyników, na zbiorze zb0 = zb1.xlsx, wyniki zmiany parametrów nie przyniosły poprawy:



Najważniejsze cechy to: Age, AvgRepairCost, AssetZip, Overloads

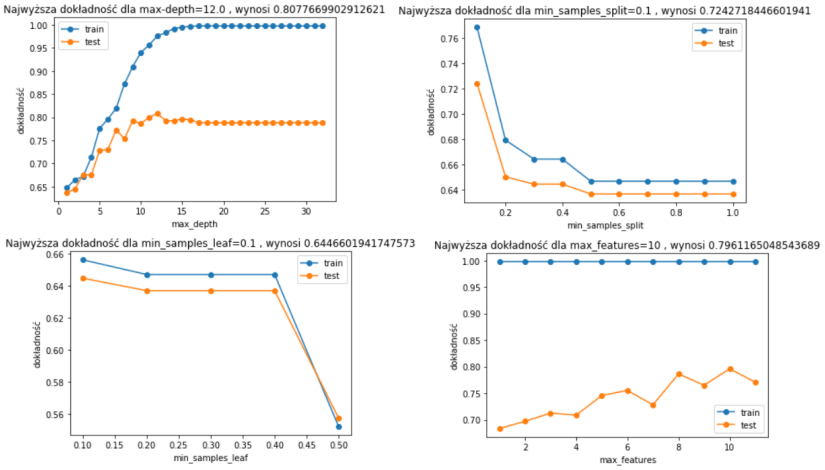


Drzewko z parametrem max-depth=2 , prezentuje się następująco:

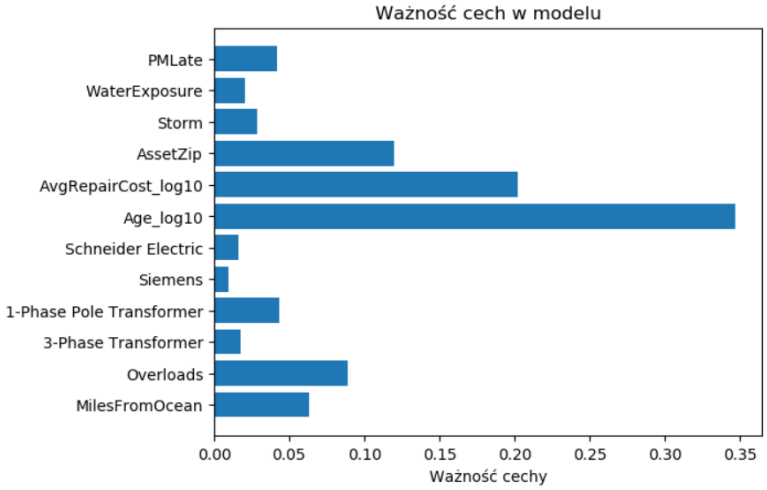
```
|--- Overloads <= 0.22
|   |--- PMLate <= 0.31
|   |   |--- Age <= -0.40
|   |   |   |--- truncated branch of depth 10
|   |   |--- Age >  -0.40
|   |   |   |--- truncated branch of depth 10
|   |--- PMLate >  0.31
|   |   |--- AvgRepairCost <= -2.76
|   |   |   |--- truncated branch of depth 3
|   |   |--- AvgRepairCost >  -2.76
|   |   |   |--- truncated branch of depth 10
|--- Overloads >  0.22
|   |--- Age <= 1.41
|   |   |--- 1-Phase Pole Transformer <= -0.20
|   |   |   |--- truncated branch of depth 10
|   |   |--- 1-Phase Pole Transformer >  -0.20
|   |   |   |--- truncated branch of depth 10
|   |--- Age >  1.41
|   |   |--- Age <= 3.80
|   |   |   |--- truncated branch of depth 2
|   |   |--- Age >  3.80
|   |   |   |--- truncated branch of depth 5
```

3. Przykład 3 z tabeli wyników, na zbiorze zb1 = zb2.xlsx, wyniki zmiany

parametrów nie przyniosły poprawy:



Najważniejsze cechy, które należy brać pod uwagę to Age, AVGRepairCost, AssetZip oaz Overloads i MilesFromOcean:



Drzewko z parametrem max-depth=2, prezentuje się następująco:

```
|--- Overloads <= 0.22
|   |--- PMLate <= 0.31
|   |   |--- Age_log10 <= -0.31
|   |   |   |--- truncated branch of depth 10
|   |   |--- Age_log10 >  -0.31
|   |   |   |--- truncated branch of depth 10
|   |--- PMLate >  0.31
|   |   |--- AvgRepairCost_log10 <= -3.70
|   |   |   |--- truncated branch of depth 3
|   |   |--- AvgRepairCost_log10 >  -3.70
|   |   |   |--- truncated branch of depth 10
|--- Overloads >  0.22
|   |--- Age_log10 <= 1.77
|   |   |--- 1-Phase Pole Transformer <= -0.20
|   |   |   |--- truncated branch of depth 10
|   |   |--- 1-Phase Pole Transformer >  -0.20
|   |   |   |--- truncated branch of depth 10
|   |--- Age_log10 >  1.77
|   |   |--- Age_log10 <= 2.67
|   |   |   |--- truncated branch of depth 2
|   |   |--- Age_log10 >  2.67
|   |   |   |--- truncated branch of depth 5
```

Cechy wspólne, które wydają się być najważniejsze:Age, MilesFromOcean, Overloads.

5 Ustalenie progu dotarcia; polityka konserwacji dla kolejnych 20 urządzeń

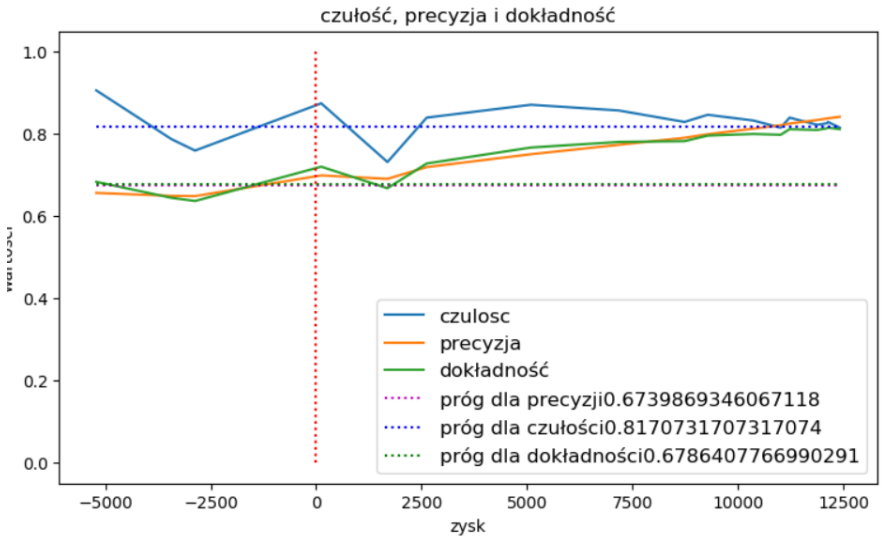
W folderze krok4, notatniku pt. zestawienie, analiza progu dotarcia oraz polityki konserwacji dla 20 nowych urządzeń.

Funkcja kosztu = tp*(100) + tn*(-30) + fp*(-100) + fn*(0), gdzie
tp - liczba uszkodzonych transformatorów, które zostały zaklasyfikowane jako uczkodzone. Mamy dla pojedynczego poprawnie sklasyfikowanego transformatora zysk +100 tys. dolarów.
tn - transformatory przewidziane jako niepoprawnie jako uszkodzone. Koszt sprawdzenia -30tys. dolarów.

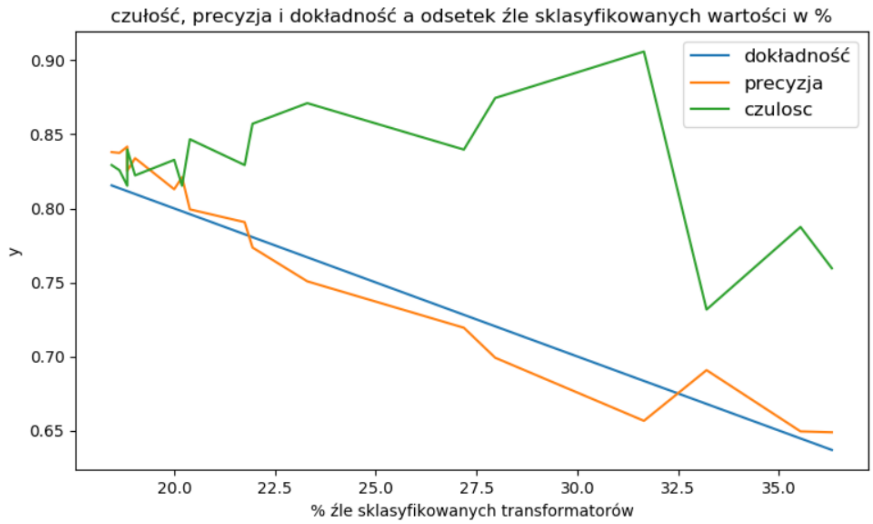
fp - uszkodzone transformatory, które sklasyfikowane zostały jako działające. Koszt -100tys. dolarów.
fn - transformatory działające, sklasyfikowane jako działające. Nie ma kosztów.

Dla tak zdefiniowanej funkcji kosztu na na zbiorze testowym nie mamy strat od momentu, gdy:

	próg	precyzja	czułość	dokładność	tp	tn	fp	fn	zysk_lub_strata
3	0.0	0.672926	0.905263	0.703963	348.0	90.0	418.0	860.0	-9700.0
1	0.0	0.667279	0.764211	0.658508	404.0	224.0	362.0	726.0	-2520.0
0	0.0	0.664403	0.721053	0.643939	420.0	265.0	346.0	685.0	-550.0
5	0.0	0.732699	0.891579	0.759907	457.0	103.0	309.0	847.0	11710.0
2	0.0	0.714747	0.698947	0.678904	501.0	286.0	265.0	664.0	15020.0
4	0.0	0.756917	0.806316	0.749417	520.0	184.0	246.0	766.0	21880.0
6	0.0	0.797820	0.847368	0.796620	562.0	145.0	204.0	805.0	31450.0

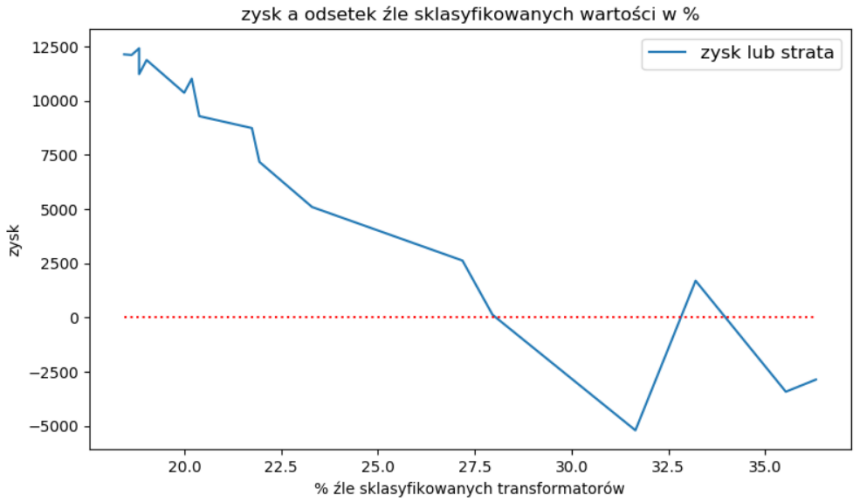


Procent błędnie sklasyfikowanych transformatorów, a zmiany parametrów klasyfikatora:



Dla najlepszego klasyfikatora, estymujemy błędnie ponizej 20 procent populacji. Czułosc i precyzja sa nie mniejsze niż dokładnosc na poziomie 0,81, natomiast nie wieksze niż 0,85.

Jeżeli procent błednie zaklasyfikowanych transformatorów nie będzie większy niż 27%, to klasyfikacjia nie przyniesie strat.



Wykorzystując najlepszy klasyfikator, predykcja dla danych: awarie_transf_new.xlsx wygląda następująco:

```
nowe['Predykcja'].value_counts()
```

```
OK      16
Fail     4
Name: Predykcja, dtype: int64
```

Wyniki zostały zapisane w pliku: predykcja.xlsx