titanic-cardinality-rj

April 12, 2025

0.1 Titanic - cardinality

0.1.1 Roksana Jandura grupa 4 nr. 416314

0.1.2 Wczytanie danych

[24]:
```
import pandas as pd
import numpy as np
```

[25]:
```
df = pd.read_csv("Zbiór danych Titanic.csv",na_values='?')
```

[26]:
```
df.head()
```

[26]:
```
   'pclass'  'survived'                                             'name'  \
0         1           1                    Allen, Miss. Elisabeth Walton
1         1           1                   Allison, Master. Hudson Trevor
2         1           0                     Allison, Miss. Helen Loraine
3         1           0             Allison, Mr. Hudson Joshua Creighton
4         1           0  Allison, Mrs. Hudson J C (Bessie Waldo Daniels)

    'sex'    'age'  'sibsp'  'parch' 'ticket'    'fare'  'cabin' 'embarked'  \
0  female  29.0000        0        0    24160  211.3375       B5          S
1    male   0.9167        1        2   113781  151.5500  C22 C26          S
2  female   2.0000        1        2   113781  151.5500  C22 C26          S
3    male  30.0000        1        2   113781  151.5500  C22 C26          S
4  female  25.0000        1        2   113781  151.5500  C22 C26          S

  'boat'  'body'                        'home.dest'
0      2     NaN                       St Louis, MO
1     11     NaN  Montreal, PQ / Chesterville, ON
2    NaN     NaN  Montreal, PQ / Chesterville, ON
3    NaN   135.0  Montreal, PQ / Chesterville, ON
4    NaN     NaN  Montreal, PQ / Chesterville, ON
```

0.1.3 Sprawdzenie liczebność poszczególnych etykiet dla danych

[27]:
```
for col in df.columns:
    print('Liczba etykiet zmiennej {}: {}'.format(col, len(df[col].unique())))
#Musimy pamiętać, że unique() traktuje wartość NaN jako oddzielną kategorię
```

```
Liczba etykiet zmiennej 'pclass': 3
Liczba etykiet zmiennej 'survived': 2
Liczba etykiet zmiennej 'name': 1307
Liczba etykiet zmiennej 'sex': 2
Liczba etykiet zmiennej 'age': 99
Liczba etykiet zmiennej 'sibsp': 7
Liczba etykiet zmiennej 'parch': 8
Liczba etykiet zmiennej 'ticket': 929
Liczba etykiet zmiennej 'fare': 282
Liczba etykiet zmiennej 'cabin': 187
Liczba etykiet zmiennej 'embarked': 4
Liczba etykiet zmiennej 'boat': 28
Liczba etykiet zmiennej 'body': 122
Liczba etykiet zmiennej 'home.dest': 370
```

Do dalszej analizy wybieram następujące zmienne jakościowe: pclass, survived, sex, ticket, cabin, embarked, boat, home.dest.

Odrzucam:

*name, ponieważ jest unikalna dla każdego (zawiera 1307 unikalnych wartości na 1309 możliwych), brak wartości analitycznej

*age- ma wartość ciągłą

*sibsp, parch - zmienna ilościowa, liczba rodzeństwa/małżonków, rodziców/dzieci

*fare - jest to kwota biletu więc zmienna ilościowa

*body - były przypisywane tylko osobom zmarłym odnalezionym, unikalne dla każdego odnalezionego ciała

[29]:
```
qualitative_cols = ["'pclass'","'survived'", "'sex'", "'ticket'", "'cabin'",
                    "'embarked'", "'boat'", "'home.dest'"] #zmienne jakościowe

for col in qualitative_cols:
    print("Liczba etykiet zmiennej '{}': {}".format(col, len(df[col].unique())))
```

```
Liczba etykiet zmiennej ''pclass'': 3
Liczba etykiet zmiennej ''survived'': 2
Liczba etykiet zmiennej ''sex'': 2
Liczba etykiet zmiennej ''ticket'': 929
Liczba etykiet zmiennej ''cabin'': 187
Liczba etykiet zmiennej ''embarked'': 4
Liczba etykiet zmiennej ''boat'': 28
Liczba etykiet zmiennej ''home.dest'': 370
```

0.1.4 Liczba wszystkich pasażerów

[30]:
```
print('Liczba wszystkich pasażerów: {}'.format(len(df)))
```

```
Liczba wszystkich pasażerów: 1309
```

0.1.5 Podział zmiennych ze względu na dużą i małą moc zbioru (kardynalność)

Niska kardynalność (mała moc zbioru):

'pclass': 3 klasy pasażerskie (1, 2, 3)

'survived' : 2 unikalne wartości 1,0 - informacja, czy dany pasażer przeżył, czy nie

'sex': 2 wartości („male", „female")

'embarked': 4 porty zaokrętowania

Średnia kardynalność (umiarkowana moc zbioru):

'boat': 28 etykiet (różne szalupy ratunkowe)

Wysoka kardynalność (duża moc zbioru) — może powodować problemy przy kodowaniu:

'ticket': 929 różnych biletów

'cabin': 187 różnych kabin

'home.dest': 370 miejsc docelowych

Zmienne o wysokiej kardynalności (ticket, cabin, home.dest) wymagają dodatkowego przetworzenia lub redukcji liczby unikalnych kategorii, aby mogły zostać skutecznie wykorzystane w modelach uczenia maszynowego.

0.1.6 Ilość unikalnych etykiet zmiennej mówiącej o kabinie danego pasażera

[31]:
```
unique_cabins = df["'cabin'"].unique()
#print(type(unique_cabins))
print('Liczba unikalnych kabin:', len(unique_cabins))
```

```
Liczba unikalnych kabin: 187
```

0.1.7 Zredukowanie liczby cech dla zmiennej opisującej kabiny poprzez zastąpienie obecnych etykiet w formacie LL11 do etykiet zawierających tylko pierwszą literę

[32]:
```
df['CabinReduced'] = df["'cabin'"].astype(str).str[0]
print(df[["'cabin'", 'CabinReduced']].head(20))
```

```
    'cabin' CabinReduced
0        B5            B
1   C22 C26            C
2   C22 C26            C
3   C22 C26            C
4   C22 C26            C
```

```
5       E12           E
6        D7           D
7       A36           A
8      C101           C
9       NaN           n
10  C62 C64           C
11  C62 C64           C
12      B35           B
13      NaN           n
14      A23           A
15      NaN           n
16  B58 B60           B
17  B58 B60           B
18      D15           D
19       C6           C
```

0.1.8 O ile procent zredukowano kardynalność zbioru zmiennej opisującej kabiny?

[33]:
```
reduced_count = len(df['CabinReduced'].unique())
print('Liczba unikalnych etykiet w kolumnie "cabin": {}'.
 ↪format(len(unique_cabins)))
print('Liczba unikalnych etykiet w kolumnie "CabinReduced": {}'.
 ↪format(reduced_count))

reduction_percentage = ((len(unique_cabins) - reduced_count) /␣
 ↪len(unique_cabins)) * 100
print('Kardynalność została zredukowana o: {:.2f}%'.
 ↪format(reduction_percentage))
```

```
Liczba unikalnych etykiet w kolumnie "cabin": 187
Liczba unikalnych etykiet w kolumnie "CabinReduced": 9
Kardynalność została zredukowana o: 95.19%
```

0.1.9 Uzasadnij dlaczego dokonujesz redukcji akurat tej zmiennej. Jak to wpływa na przyszłe analizy. Czy powoduje jakieś negatywne skutki?

Wybrałam tą kolumnę, ponieważ zawiera ona 186 unikatowych wartości. Tak wysoka liczba unikalnych wartości utrudnia analizę i może powodować poważne problemy w dalszym przetwarzaniu danych Dodatkowo, wiele z tych wartości może być unikalnych tylko z pozoru – np. „C85” i „C86” mogą prawdopodobnie oznaczać kabiny znajdujące się na tym samym pokładzie i różniące się tylko numerem lub wręcz sąsiadujące, identyczne kabiny.

Dlatego dokonałam redukcji zmiennej cabin poprzez wydzielenie jedynie pierwszej litery i utworzenie nowej zmiennej CabinReduced. W efekcie liczba unikalnych etykiet spadła z 186 do 9, co stanowi redukcję o 95,19%. Taka transformacja pozwala zachować istotną informację o lokalizacji kabiny, przy jednoczesnym znacznym ograniczeniu liczby kategorii, co ułatwia dalsze analizy i modelowanie.

Redukcja ta przynosi wiele korzyści: zmniejsza złożoność zbioru danych, ułatwia analizę i kodowanie

zmiennej, a także ogranicza ryzyko przeuczenia modeli predykcyjnych.

Potencjalnym minusem tej redukcji jest utrata szczegółowej informacji o konkretnych numerach kabin - utrata infomracji, którzy pasażerowie mieli przydzieloną tą samą kabinę.