PD2

March 22, 2021

[1]:
```
import pandas as pd
import numpy as np
import sklearn
import category_encoders as ce
from sklearn.impute import KNNImputer
import math
import matplotlib.pyplot as plt

import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
allegro = pd.read_csv("allegro-api-transactions.csv")
```

[3]:
```
allegro.head()
```

[3]:
```
   lp                 date     item_id  \
0   0  2016-04-03 21:21:08  4753602474
1   1  2016-04-03 15:35:26  4773181874
2   2  2016-04-03 14:14:31  4781627074
3   3  2016-04-03 19:55:44  4783971474
4   4  2016-04-03 18:05:54  4787908274

                                          categories  pay_option_on_delivery  \
0  ['Komputery', 'Dyski i napędy', 'Nośniki', 'No…                       1
1  ['Odzież, Obuwie, Dodatki', 'Bielizna damska',…                       1
2  ['Dom i Ogród', 'Budownictwo i Akcesoria', 'Śc…                       1
3  ['Książki i Komiksy', 'Poradniki i albumy', 'Z…                       1
4  ['Odzież, Obuwie, Dodatki', 'Ślub i wesele', '…                       1

   pay_option_transfer         seller   price  it_is_allegro_standard  \
0                    1    radzioch666   59.99                       1
1                    1  InwestycjeNET    4.90                       1
2                    1    otostyl_com  109.90                       1
3                    1        Matfel1   18.50                       0
4                    1      PPHU_RICO   19.90                       1

   it_quantity  it_is_brand_zone  it_seller_rating         it_location  \
```

```
0          997                 0             50177              Warszawa
1         9288                 0             12428              Warszawa
2          895                 0              7389                Leszno
3          971                 0             15006  Wola Krzysztoporska
4          950                 0             32975             BIAŁYSTOK

              main_category
0                 Komputery
1  Odzież, Obuwie, Dodatki
2               Dom i Ogród
3         Książki i Komiksy
4  Odzież, Obuwie, Dodatki
```

1 Target encoding na allegro_data

[5]:
```
allegro_target_encoding = allegro.copy(deep = True)
means = allegro.groupby("it_location")['price'].mean()
allegro_target_encoding['it_location'] = allegro_target_encoding['it_location'].
 ↪map(means)
allegro_target_encoding.head()
```

[5]:
```
   lp                 date     item_id  \
0   0  2016-04-03 21:21:08  4753602474
1   1  2016-04-03 15:35:26  4773181874
2   2  2016-04-03 14:14:31  4781627074
3   3  2016-04-03 19:55:44  4783971474
4   4  2016-04-03 18:05:54  4787908274

                                          categories  pay_option_on_delivery  \
0  ['Komputery', 'Dyski i napędy', 'Nośniki', 'No…                       1
1  ['Odzież, Obuwie, Dodatki', 'Bielizna damska',…                       1
2  ['Dom i Ogród', 'Budownictwo i Akcesoria', 'Śc…                       1
3  ['Książki i Komiksy', 'Poradniki i albumy', 'Z…                       1
4  ['Odzież, Obuwie, Dodatki', 'Ślub i wesele', '…                       1

   pay_option_transfer         seller   price  it_is_allegro_standard  \
0                    1    radzioch666   59.99                       1
1                    1  InwestycjeNET    4.90                       1
2                    1    otostyl_com  109.90                       1
3                    1        Matfel1   18.50                       0
4                    1      PPHU_RICO   19.90                       1

   it_quantity  it_is_brand_zone  it_seller_rating  it_location  \
0          997                 0             50177    85.423398
1         9288                 0             12428    85.423398
2          895                 0              7389    61.990914
```

```
3          971                 0             15006    35.433365
4          950                 0             32975   117.191956

              main_category
0                 Komputery
1  Odzież, Obuwie, Dodatki
2               Dom i Ogród
3         Książki i Komiksy
4  Odzież, Obuwie, Dodatki
```

W przypadku onehot encodingu, ze względu na to że mamy bardzo dużo wartości w kolumnie 'it_location', pojawi się bardzo dużo nowych kolumn. Może okazać się, że będziemy mieli za dużą ilość kolumn żeby wytrenować nasz model. One hot encoding będzie dobrym rozwiązaniem jeśli tabela kategoryczna, która przekształcamy ma mało różnych wartości.

2 One hot encoding

[6]:
```
allegro_one_hot = allegro.copy(deep = True)
main_categories = pd.get_dummies(allegro['main_category'])
allegro_one_hot = pd.concat([allegro_one_hot, main_categories], axis=1)
allegro_one_hot.drop(labels = ['main_category'], axis = 1,inplace = True)
allegro_one_hot
```

[6]:
```
            lp                 date     item_id  \
0            0  2016-04-03 21:21:08  4753602474
1            1  2016-04-03 15:35:26  4773181874
2            2  2016-04-03 14:14:31  4781627074
3            3  2016-04-03 19:55:44  4783971474
4            4  2016-04-03 18:05:54  4787908274
…          …                    …          …
420015  420015  2016-04-03 20:27:13  6099625607
420016  420016  2016-04-03 22:35:02  6099634607
420017  420017  2016-04-03 22:38:57  6099780407
420018  420018  2016-04-03 22:44:17  6099801007
420019  420019  2016-04-03 23:08:23  6099873207

                                                categories  \
0       ['Komputery', 'Dyski i napędy', 'Nośniki', 'No…
1       ['Odzież, Obuwie, Dodatki', 'Bielizna damska',…
2       ['Dom i Ogród', 'Budownictwo i Akcesoria', 'Śc…
3       ['Książki i Komiksy', 'Poradniki i albumy', 'Z…
4       ['Odzież, Obuwie, Dodatki', 'Ślub i wesele', '…
…                                                     …
420015  ['RTV i AGD', 'Sprzęt audio dla domu', 'Odtwar…
420016     ['Uroda', 'Makijaż', 'Oczy', 'Tusze do rzęs']
420017  ['Odzież, Obuwie, Dodatki', 'Przebrania, kosti…
```

```
420018  ['Dla Dzieci', 'Rowery i pojazdy', 'Rowery bie…
420019  ['Motoryzacja', 'Części samochodowe', 'Koła, f…

        pay_option_on_delivery  pay_option_transfer           seller   price  \
0                            1                     1      radzioch666   59.99
1                            1                     1    InwestycjeNET    4.90
2                            1                     1      otostyl_com  109.90
3                            1                     1          Matfel1   18.50
4                            1                     1        PPHU_RICO   19.90
…                          …                     …              …     …
420015                       0                     0        iwona7012  180.00
420016                       1                     1  Dolce_Cosmetics   14.99
420017                       1                     1        pewex4all    5.99
420018                       1                     0         kostasia  200.00
420019                       0                     0        Malami172  500.00

        it_is_allegro_standard  it_quantity  …  Nieruchomości  \
0                            1          997  …              0
1                            1         9288  …              0
2                            1          895  …              0
3                            0          971  …              0
4                            1          950  …              0
…                          …          …  …              …
420015                       0            0  …              0
420016                       1            2  …              0
420017                       1          470  …              0
420018                       0            0  …              0
420019                       0            0  …              0

        Odzież, Obuwie, Dodatki Przemysł  RTV i AGD  Rękodzieło  \
0                             0        0          0           0
1                             1        0          0           0
2                             0        0          0           0
3                             0        0          0           0
4                             1        0          0           0
…                           …      …        …         …
420015                        0        0          1           0
420016                        0        0          0           0
420017                        1        0          0           0
420018                        0        0          0           0
420019                        0        0          0           0

        Sport i Turystyka  Sprzęt estradowy, studyjny i DJ-ski  \
0                       0                                    0
1                       0                                    0
2                       0                                    0
3                       0                                    0
```

```
4                        0                                        0
…                      …                                        …
420015                   0                                        0
420016                   0                                        0
420017                   0                                        0
420018                   0                                        0
420019                   0                                        0

        Telefony i Akcesoria  Uroda  Zdrowie
0                          0      0        0
1                          0      0        0
2                          0      0        0
3                          0      0        0
4                          0      0        0
…                        …    …       …
420015                     0      0        0
420016                     0      1        0
420017                     0      0        0
420018                     0      0        0
420019                     0      0        0

[420020 rows x 40 columns]
```

Metoda one hot encoding tworzy nam z kolumny o wyrazach kategorycznych macierz (n x m), gdzie n - liczba wierszy, m - liczba kategorii w kolumnie. Każdy wiersz ma jedynkę w tej kolumnie z której miał wartośc i zera w innych kolumnach. Dobra metoda gdy mamy mało kategorii.

[8]:
```
allegro_cat = allegro.copy(deep = True)
CatBoostEncoder = ce.CatBoostEncoder()
CatBoostEncoder = CatBoostEncoder.fit(allegro['main_category'], y =␣
 ↪allegro['price'])
allegro_cat['main_category'] = CatBoostEncoder.
 ↪transform(allegro_cat["main_category"])
allegro_cat
```

[8]:
```
            lp                 date     item_id  \
0            0  2016-04-03 21:21:08  4753602474
1            1  2016-04-03 15:35:26  4773181874
2            2  2016-04-03 14:14:31  4781627074
3            3  2016-04-03 19:55:44  4783971474
4            4  2016-04-03 18:05:54  4787908274
…          …                  …           …
420015  420015  2016-04-03 20:27:13  6099625607
420016  420016  2016-04-03 22:35:02  6099634607
420017  420017  2016-04-03 22:38:57  6099780407
420018  420018  2016-04-03 22:44:17  6099801007
420019  420019  2016-04-03 23:08:23  6099873207
```

```
                                                categories  \
0       ['Komputery', 'Dyski i napędy', 'Nośniki', 'No…
1       ['Odzież, Obuwie, Dodatki', 'Bielizna damska',…
2       ['Dom i Ogród', 'Budownictwo i Akcesoria', 'Śc…
3       ['Książki i Komiksy', 'Poradniki i albumy', 'Z…
4       ['Odzież, Obuwie, Dodatki', 'Ślub i wesele', '…
…                                                     …
420015  ['RTV i AGD', 'Sprzęt audio dla domu', 'Odtwar…
420016      ['Uroda', 'Makijaż', 'Oczy', 'Tusze do rzęs']
420017  ['Odzież, Obuwie, Dodatki', 'Przebrania, kosti…
420018  ['Dla Dzieci', 'Rowery i pojazdy', 'Rowery bie…
420019  ['Motoryzacja', 'Części samochodowe', 'Koła, f…

        pay_option_on_delivery  pay_option_transfer           seller   price  \
0                            1                    1      radzioch666   59.99
1                            1                    1    InwestycjeNET    4.90
2                            1                    1      otostyl_com  109.90
3                            1                    1          Matfel1   18.50
4                            1                    1        PPHU_RICO   19.90
…                          …                    …                …     …
420015                       0                    0        iwona7012  180.00
420016                       1                    1  Dolce_Cosmetics   14.99
420017                       1                    1        pewex4all    5.99
420018                       1                    0         kostasia  200.00
420019                       0                    0        Malami172  500.00

        it_is_allegro_standard  it_quantity  it_is_brand_zone  \
0                            1          997                 0
1                            1         9288                 0
2                            1          895                 0
3                            0          971                 0
4                            1          950                 0
…                          …            …                 …
420015                       0            0                 0
420016                       1            2                 0
420017                       1          470                 0
420018                       0            0                 0
420019                       0            0                 0

        it_seller_rating          it_location  main_category
0                  50177            Warszawa     121.806959
1                  12428            Warszawa      75.858066
2                   7389              Leszno      72.434802
3                  15006  Wola Krzysztoporska      25.031971
4                  32975           BIAŁYSTOK      75.858066
…                      …                   …            …
```

```
420015              176              Kraśnik     107.532762
420016            34851          Dzierżoniów      28.130309
420017              983              Supraśl      75.858066
420018              163               Poznań      71.206519
420019              265                Pszów     134.425547

[420020 rows x 14 columns]
```

Metoda działa tak samo jak target encoding z tą różnicą, że pomija wartość targetu aktualnego wiersza, aby zminimalizować efekt outlierów. Metoda działa w locie.

[9]:
```
allegro_count = allegro.copy(deep = True)
CountEncoder = ce.CountEncoder()
CountEncoder = CountEncoder.fit(allegro_count['main_category'])
allegro_count['main_category'] = CountEncoder.
 ↪transform(allegro_count["main_category"])
allegro_count
```

[9]:
```
            lp                 date     item_id  \
0            0  2016-04-03 21:21:08  4753602474
1            1  2016-04-03 15:35:26  4773181874
2            2  2016-04-03 14:14:31  4781627074
3            3  2016-04-03 19:55:44  4783971474
4            4  2016-04-03 18:05:54  4787908274
…          …                  …          …
420015  420015  2016-04-03 20:27:13  6099625607
420016  420016  2016-04-03 22:35:02  6099634607
420017  420017  2016-04-03 22:38:57  6099780407
420018  420018  2016-04-03 22:44:17  6099801007
420019  420019  2016-04-03 23:08:23  6099873207

                                               categories  \
0       ['Komputery', 'Dyski i napędy', 'Nośniki', 'No…
1       ['Odzież, Obuwie, Dodatki', 'Bielizna damska',…
2       ['Dom i Ogród', 'Budownictwo i Akcesoria', 'Śc…
3       ['Książki i Komiksy', 'Poradniki i albumy', 'Z…
4       ['Odzież, Obuwie, Dodatki', 'Ślub i wesele', '…
…                                                     …
420015  ['RTV i AGD', 'Sprzęt audio dla domu', 'Odtwar…
420016     ['Uroda', 'Makijaż', 'Oczy', 'Tusze do rzęs']
420017  ['Odzież, Obuwie, Dodatki', 'Przebrania, kosti…
420018  ['Dla Dzieci', 'Rowery i pojazdy', 'Rowery bie…
420019  ['Motoryzacja', 'Części samochodowe', 'Koła, f…

        pay_option_on_delivery  pay_option_transfer           seller   price  \
0                            1                    1      radzioch666   59.99
1                            1                    1    InwestycjeNET    4.90
```

```
2                              1                    1       otostyl_com  109.90
3                              1                    1           Matfel1   18.50
4                              1                    1         PPHU_RICO   19.90
…                            …                    …               …      …
420015                         0                    0         iwona7012  180.00
420016                         1                    1   Dolce_Cosmetics   14.99
420017                         1                    1         pewex4all    5.99
420018                         1                    0          kostasia  200.00
420019                         0                    0         Malami172  500.00

        it_is_allegro_standard  it_quantity  it_is_brand_zone  \
0                            1          997                 0
1                            1         9288                 0
2                            1          895                 0
3                            0          971                 0
4                            1          950                 0
…                          …            …                 …
420015                       0            0                 0
420016                       1            2                 0
420017                       1          470                 0
420018                       0            0                 0
420019                       0            0                 0

        it_seller_rating          it_location  main_category
0                  50177              Warszawa          14491
1                  12428              Warszawa          54257
2                   7389                Leszno          91042
3                  15006  Wola Krzysztoporska          11572
4                  32975             BIAŁYSTOK          54257
…                  …                       …             …
420015               176               Kraśnik          20341
420016             34851            Dzierżoniów          28096
420017               983               Supraśl          54257
420018               163                Poznań          42107
420019               265                 Pszów          45941

[420020 rows x 14 columns]
```

Metoda zamienia każdą wartość na ilość wystąpień danej kategorii w kolumnie.

[10]:
```
allegro_reduced = allegro[['price','it_seller_rating', 'it_quantity']]
allegro_reduced = allegro_reduced[0:40000]
n = len(allegro_reduced['price'])
allegro_reduced
```

[10]:
```
        price  it_seller_rating  it_quantity
0       59.99             50177          997
```

```
1        4.90             12428          9288
2      109.90              7389           895
3       18.50             15006           971
4       19.90             32975           950
…         …                 …             …
39995   94.90             40273           231
39996    9.99             10214           999
39997  209.00              6056             9
39998    6.90             18052             4
39999    6.19             30348           446

[40000 rows x 3 columns]
```

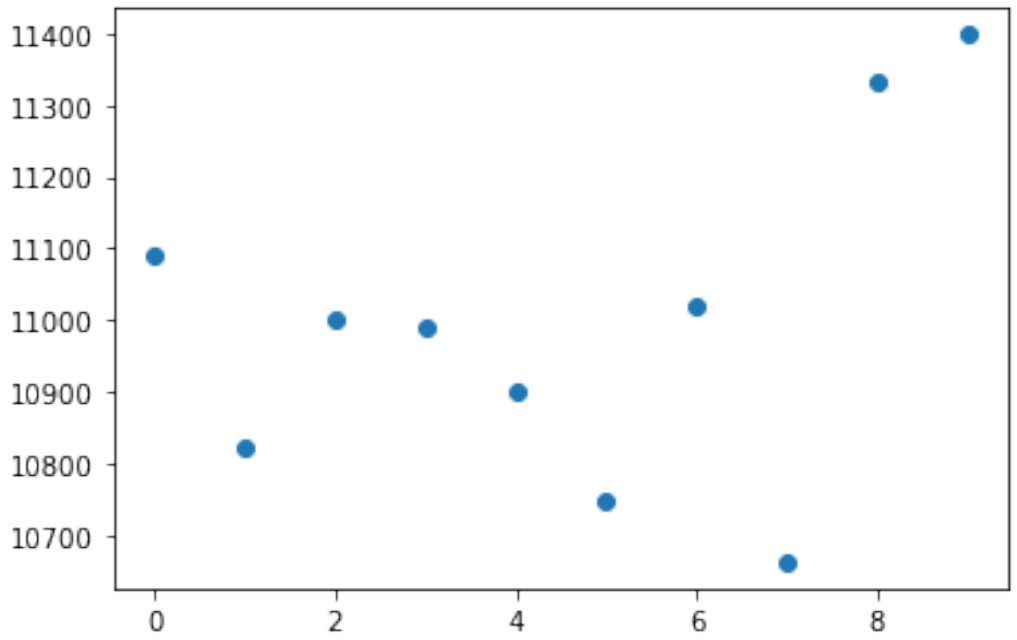
Zmniejszylem liczbe rekordow bo bardzo wolno mi dzialal KKNIImputer :(

[85]:
```
results = [0 for i in range(10)]
for i in range(10):
    X = allegro_reduced.copy(deep=True)
    # zmieniam ziarno za kazdym razem zeby miec rozne wiersze w roznich␣
 ↪eksperymentach
    np.random.seed((i+1) * 10)
    chosen_idx = np.random.choice(n , replace = False, size = 4000)
    X.it_seller_rating[chosen_idx] = np.nan
    imputer = KNNImputer(n_neighbors=5, weights="uniform")
    X = imputer.fit_transform(X)
    results[i] = math.sqrt(np.sum(pow(X[:,1] -␣
 ↪allegro_reduced["it_seller_rating"], 2)) / n)
```

[98]:
```
results
```

[98]:
```
array([11091.30265326, 10824.09304193, 10999.58255539, 10990.81347631,
       10899.04809758, 10749.58908422, 11020.95200936, 10662.82232402,
       11333.87726358, 11398.77064044])
```

[111]:
```
x = [i for i in range(10)]
plt.scatter(x,results)
plt.show()
```



[100]:
```
standard_deviation = math.sqrt(np.sum(pow(results - results.mean(),2)))
standard_deviation
```

[100]: 704.8374063029624

To odchylenie wyszlo bardzo duże i wgl jakoś kiepsko to wygląda

[76]:
```
```

[76]:
```
array([[5.99900e+01, 4.51266e+04, 9.97000e+02],
       [4.90000e+00, 1.24280e+04, 9.28800e+03],
       [1.09900e+02, 7.38900e+03, 8.95000e+02],
       …,
       [2.09000e+02, 6.05600e+03, 9.00000e+00],
       [6.90000e+00, 1.80520e+04, 4.00000e+00],
       [6.19000e+00, 3.03480e+04, 4.46000e+02]])
```

[112]:
```
second_results = [0 for i in range(10)]
for i in range(10):
    X = allegro_reduced.copy(deep=True)
    # zmieniam ziarno za kazdym razem zeby miec rozne wiersze w roznich␣
 ↪eksperymentach
    np.random.seed((i+1) * 10)
    chosen_idx = np.random.choice(n , replace = False, size = 4000)
    X.it_seller_rating[chosen_idx] = np.nan
    X.it_quantity[chosen_idx] = np.nan
```

```
    imputer = KNNImputer(n_neighbors=5, weights="uniform")
    X = imputer.fit_transform(X)
    second_results[i] = math.sqrt(np.sum(pow(X[:,1] -␣
↪allegro_reduced["it_seller_rating"], 2)) / n)
```

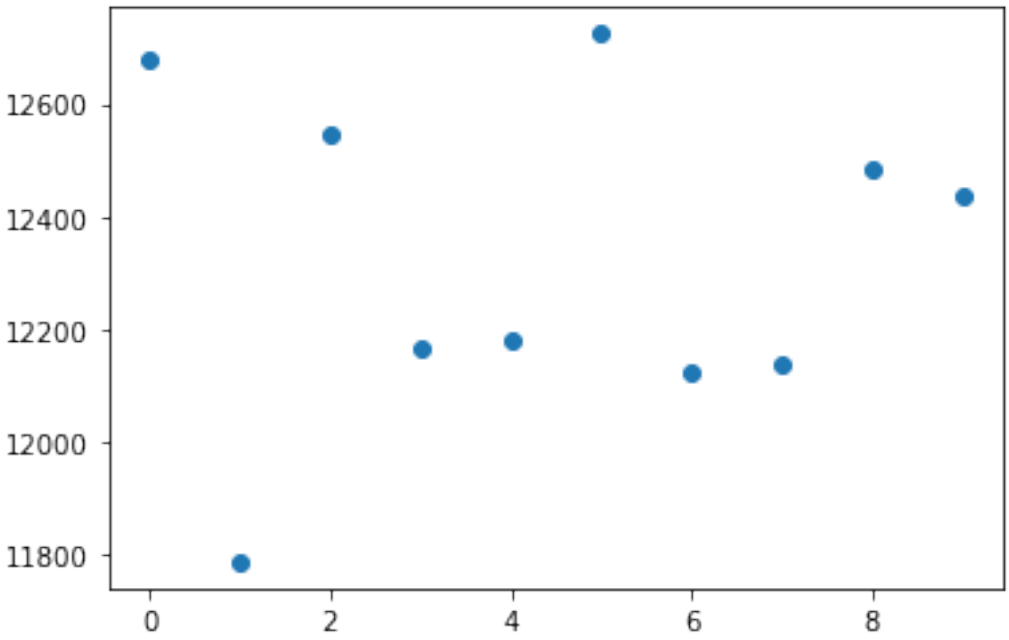
[113]:
```
second_results
```

[113]:
```
[12677.534103081522,
 11786.433902690416,
 12544.189375938566,
 12168.1075889494,
 12182.160840345321,
 12724.979093205025,
 12121.543752533256,
 12135.839755103681,
 12486.038002537434,
 12435.73023295114]
```

[114]:
```
second_standard_deviation = math.sqrt(np.sum(pow(second_results - results.
↪mean(),2)))
second_standard_deviation
```

[114]:
```
4295.308773371413
```

[115]:
```
x = [i for i in range(10)]
plt.scatter(x,second_results)
plt.show()
```



Po usunięciu wartości z obu kolumn różnice w wynikach pomiędzy prawdziwą kolumną it_seller_rating, a to stworzoną po algorytmie k-nearest-neighbours są bardzo duże, również odchylenie standardowe jest znacznie większe. Wydaje mi się, że w obydwu przypadkach błąd jest stosunkowo duży.