Визуализация данных “Digital Reputation Challenge”

Курс “Прикладные задачи анализа данных 2019”

Цыпин Артем, 517 группа

Данные

Датасет с платформы boosters.pro. Содержит три таблицы с анонимизированными данными.

```
X1.head()
```

	id	1	2	3	4	5	6	7	8	9	...	16	17	18	19	20	21	22	23	24	25
0	3	1	-1.0	-1.0	107.0	255.0	537.0	10.0	41.0	0.0	...	0	0	0	0	0	0	1	0	1	0
1	5	0	0.0	0.0	20.0	0.0	188.0	1.0	25.0	2.0	...	0	0	0	0	0	0	0	0	0	0
2	6	1	0.0	0.0	158.0	155.0	3092.0	3.0	218.0	29.0	...	0	0	0	0	0	0	0	1	0	0
3	8	1	0.0	0.0	102.0	343.0	341.0	0.0	24.0	2.0	...	0	0	0	0	0	0	1	0	0	0
4	10	1	0.0	0.0	1.0	1.0	33.0	0.0	41.0	1.0	...	0	0	0	0	0	0	1	0	1	0

```
X2.head()
```

	id	A
0	3	5
1	3	70340
2	3	72868
3	3	73471
4	3	74998

Данные

```
X3.head()
```

	id	1	2	3	4	5	6	7	8	9	...	443	444	445	446	447	448	449	450	451	452
0	3	0.0	0.0	0.00	0.0	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.00	0.0	0.0	0.022222	0.0	0.0	0.0	0.000000
1	5	0.0	0.0	0.00	0.0	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.00	0.0	0.0	0.029703	0.0	0.0	0.0	0.000000
2	6	0.0	0.0	0.00	0.0	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.00	0.0	0.0	0.222222	0.0	0.0	0.0	0.111111
3	8	0.0	0.0	0.02	0.0	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.02	0.0	0.0	0.060000	0.0	0.0	0.0	0.000000
4	10	0.0	0.0	0.00	0.0	0.055556	0.055556	0.0	0.0	0.0	...	0.0	0.0	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000

Таблица X1

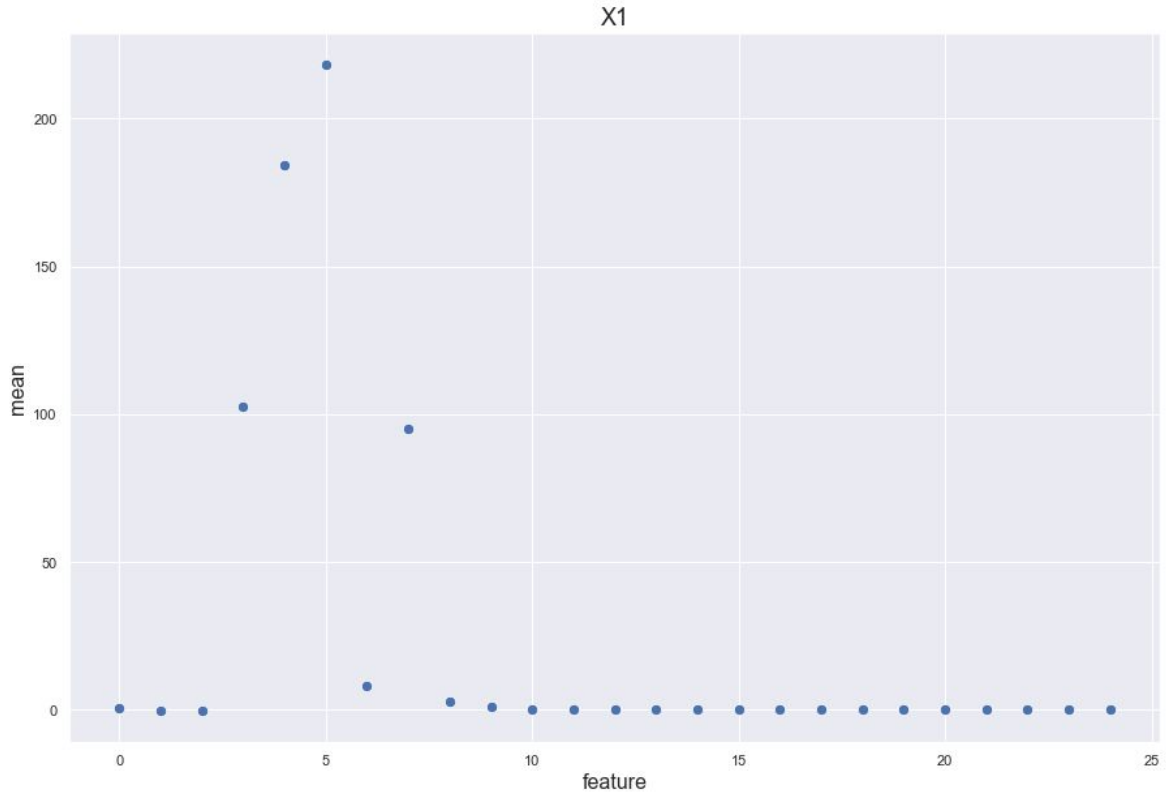


Таблица X1

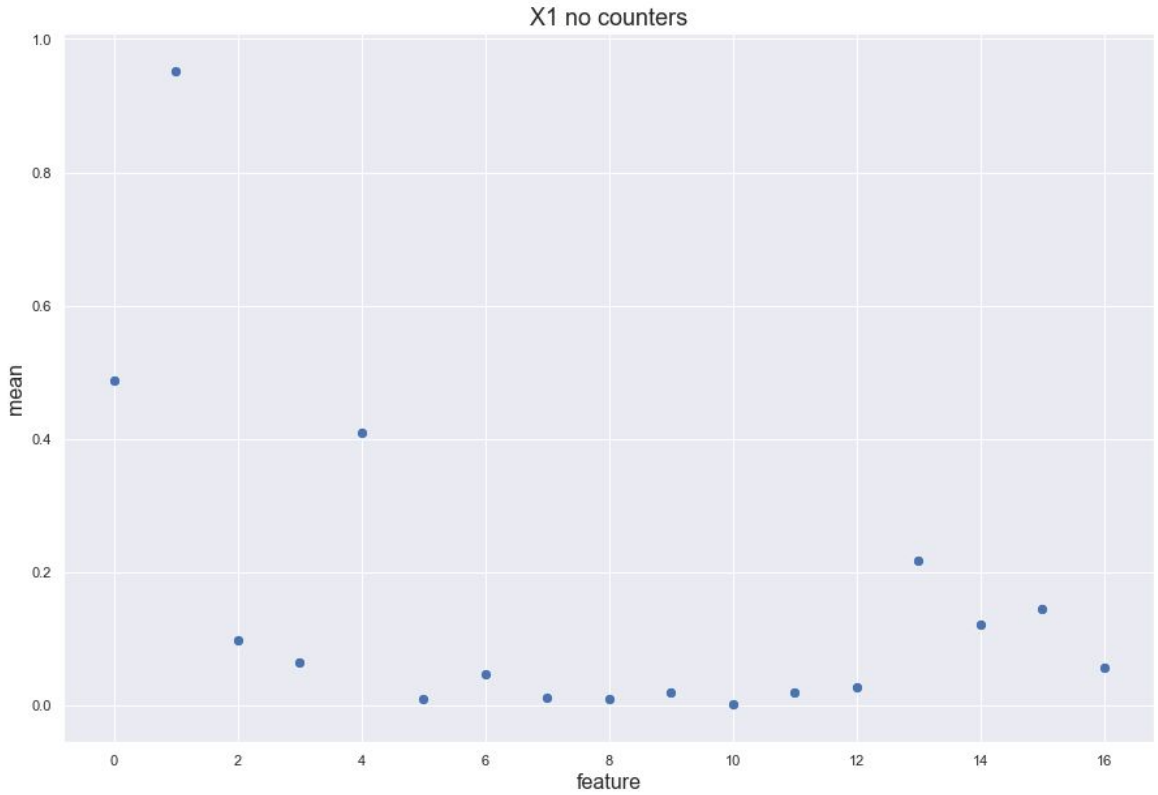


Таблица X1

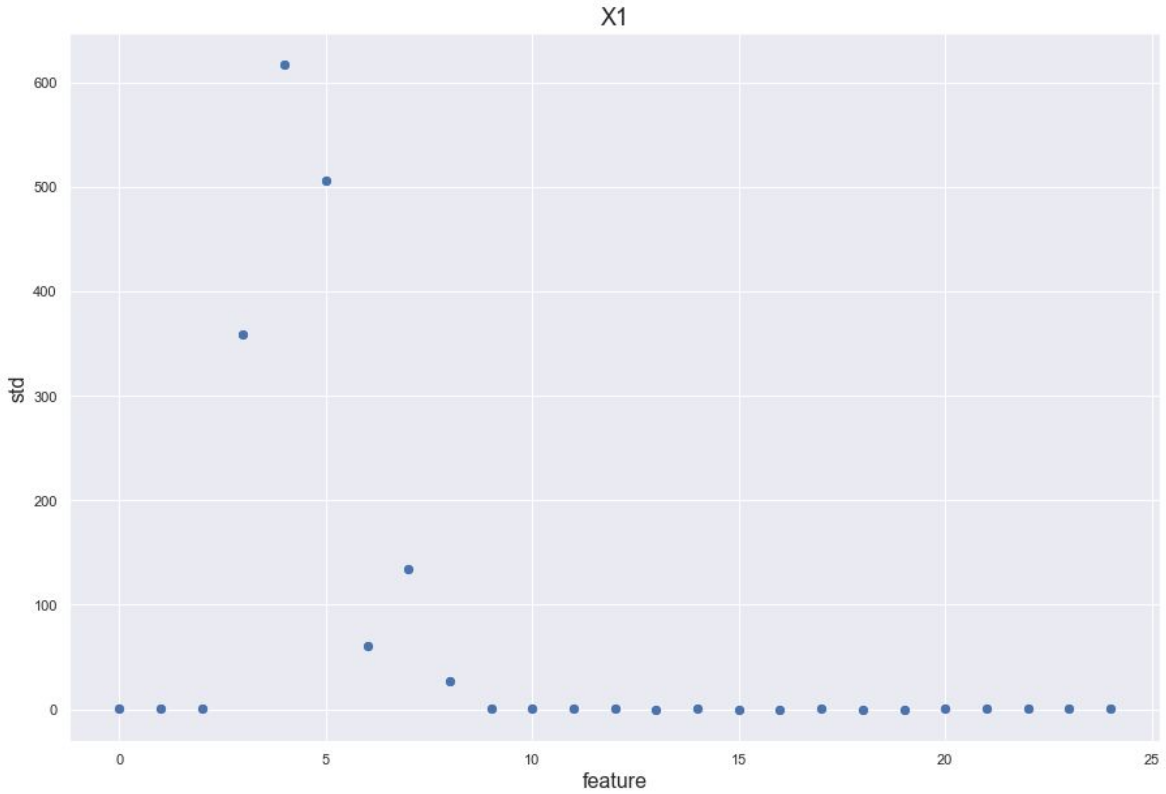


Таблица X1

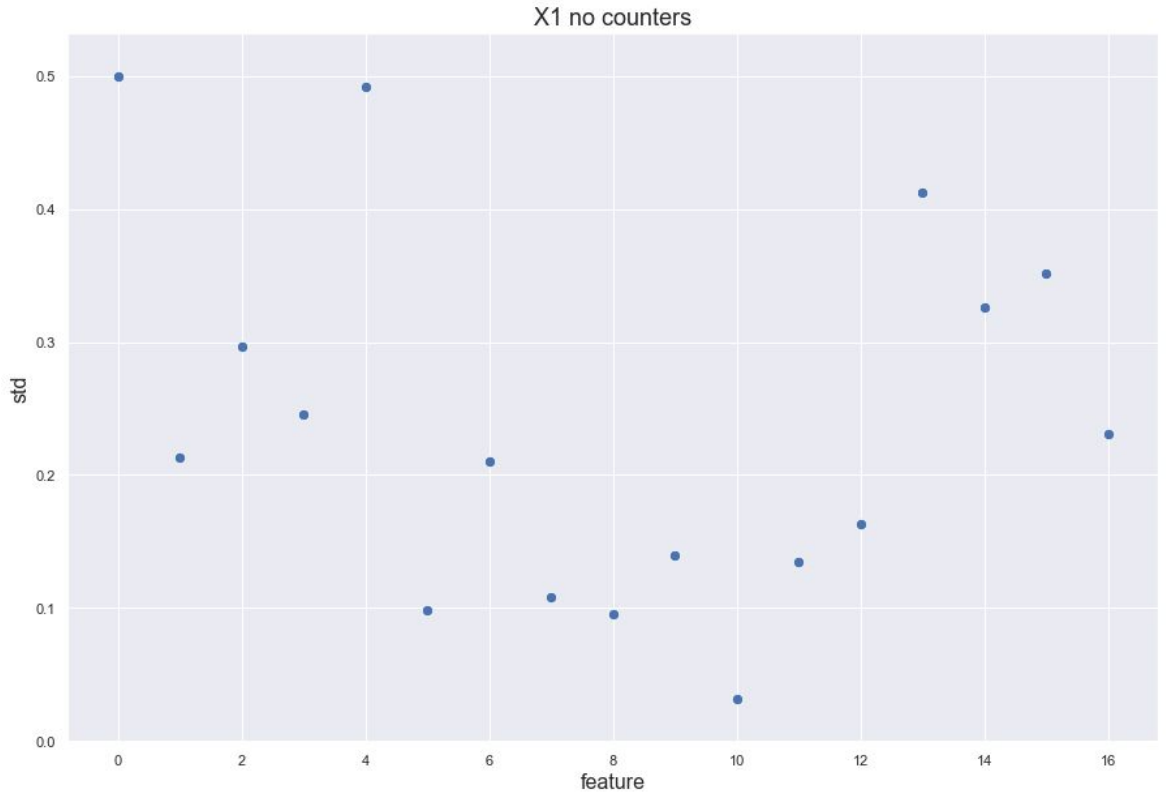


Таблица X1

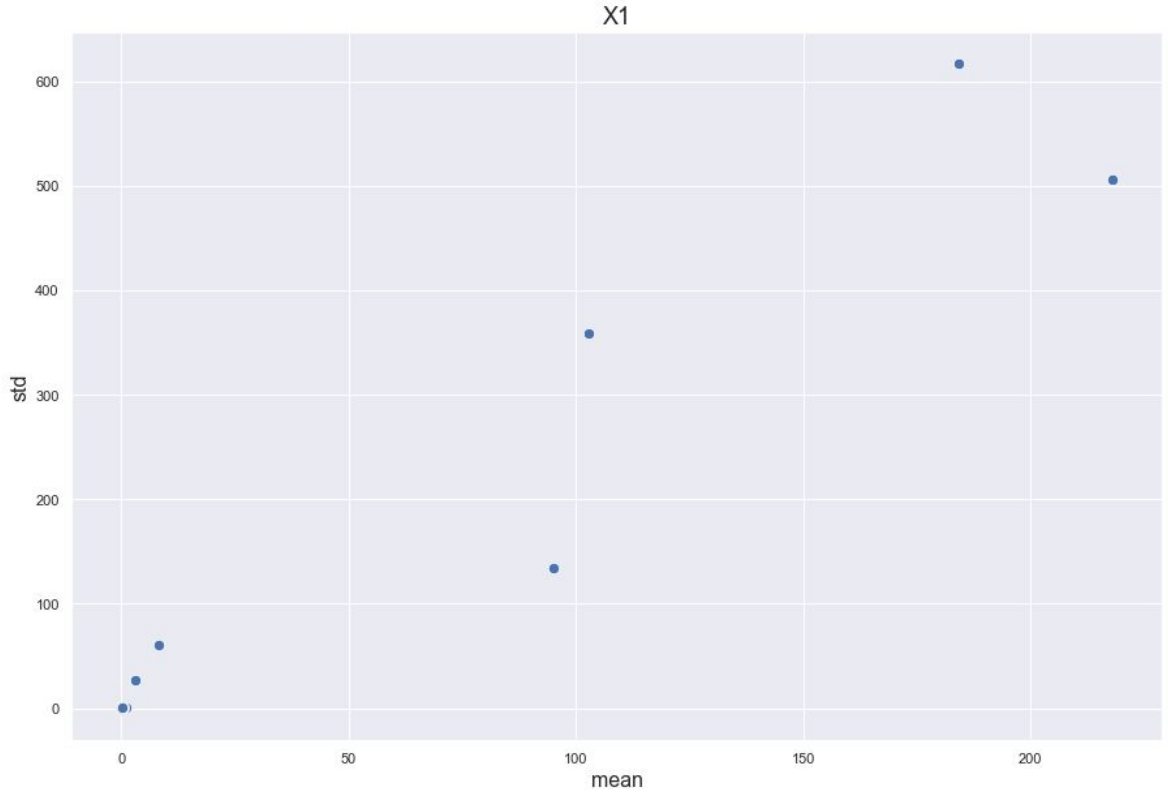


Таблица X1

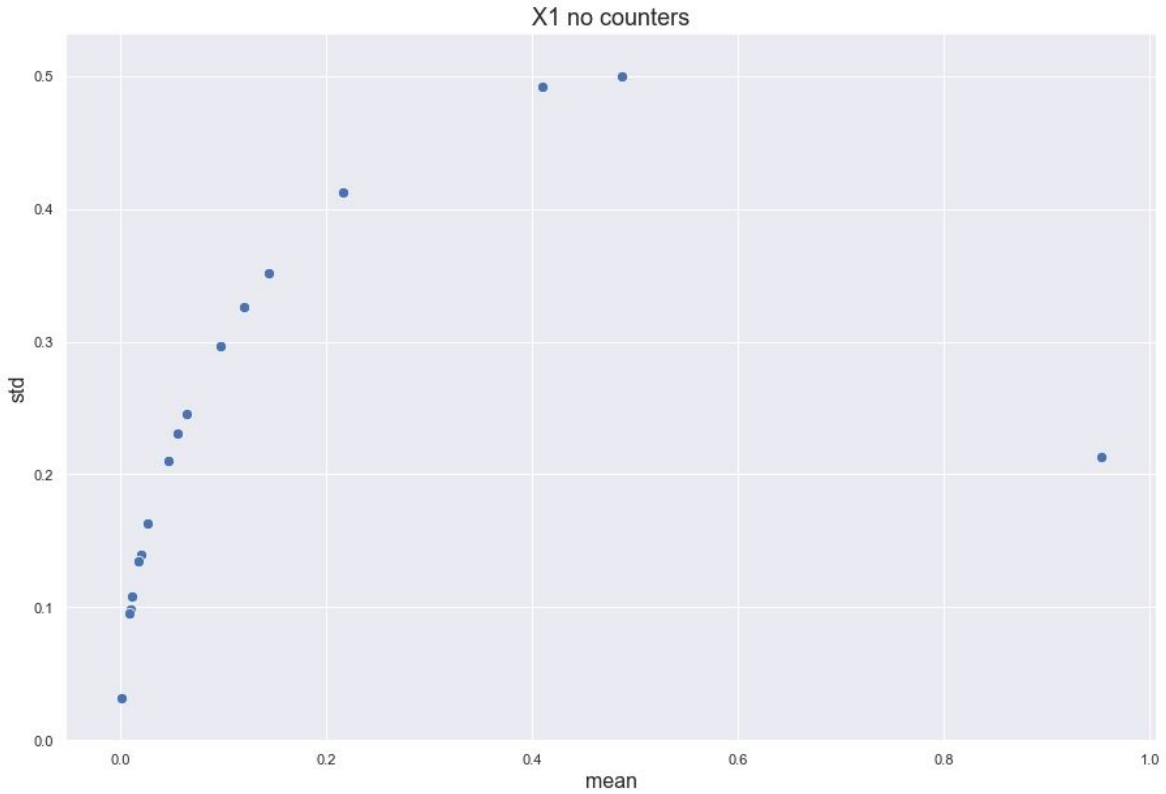


Таблица X1

Анализ значений не бинарных переменных таблицы X1.

```
'2'     5  -2     2   -0.038    0.412 [-2 -1  0  1  2] [  97  109 3662  113   19]
'3'     5  -2     2   -0.020    0.306 [-2 -1  0  1  2] [  52   64 3803   73    8]
'4'   531   0  9967  102.645  359.097 [0 1 2 3 4 5 6 7] [549 277 198 152 130 125  97  86]
'5'   721   0 10000  184.296  616.335 [ 0  1  2  3  4  5  9 10] [2061   92   59   38   33   30   20   19]
'6'   677   0 10000  218.048  505.995 [ 0 73 68 69 48 35 41 47] [35 28 28 27 25 25 25 25]
'7'   115   0  1989    8.085   60.907 [0 1 2 3 4 5 6 7] [1656  691  423  253  185  115   98   77]
'8'   452   0  2409   95.189  133.882 [ 2  3 11  1  9  5  0  8] [62 59 57 56 51 50 49 48]
'9'    52   0  1396    2.919   27.434 [0 1 2 3 4 5 6 7] [1876  836  351  238  137  108   87   63]
```

Таблица X3

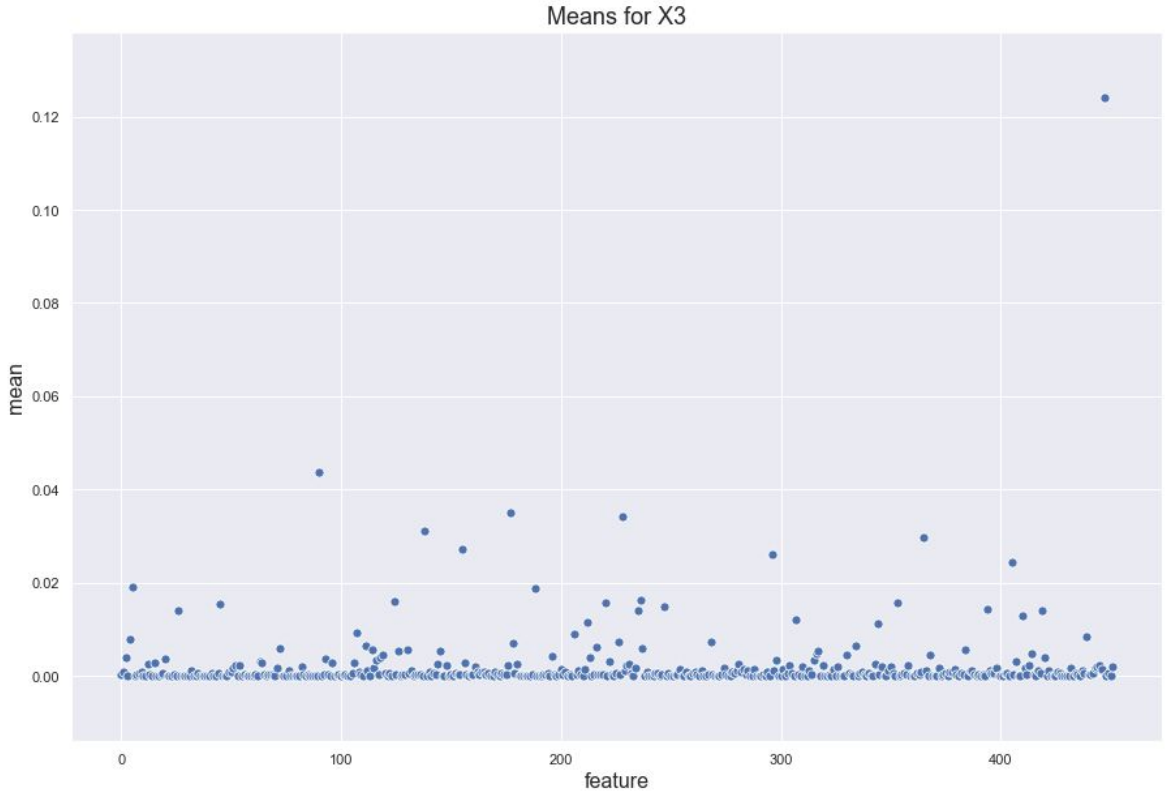

Means for X3
mean
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
100
200
300
400
feature

Таблица X3

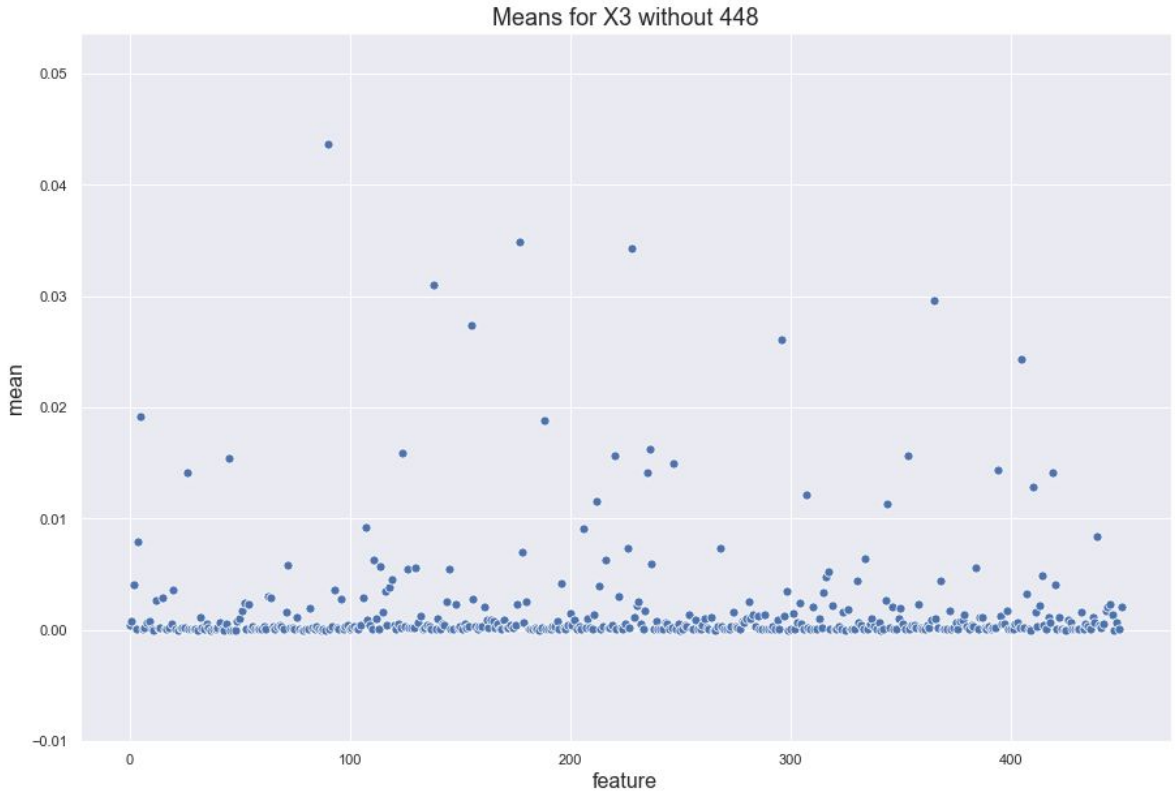

Means for X3 without 448
mean
feature
0.05
0.04
0.03
0.02
0.01
0.00
−0.01
0
100
200
300
400

Таблица X3

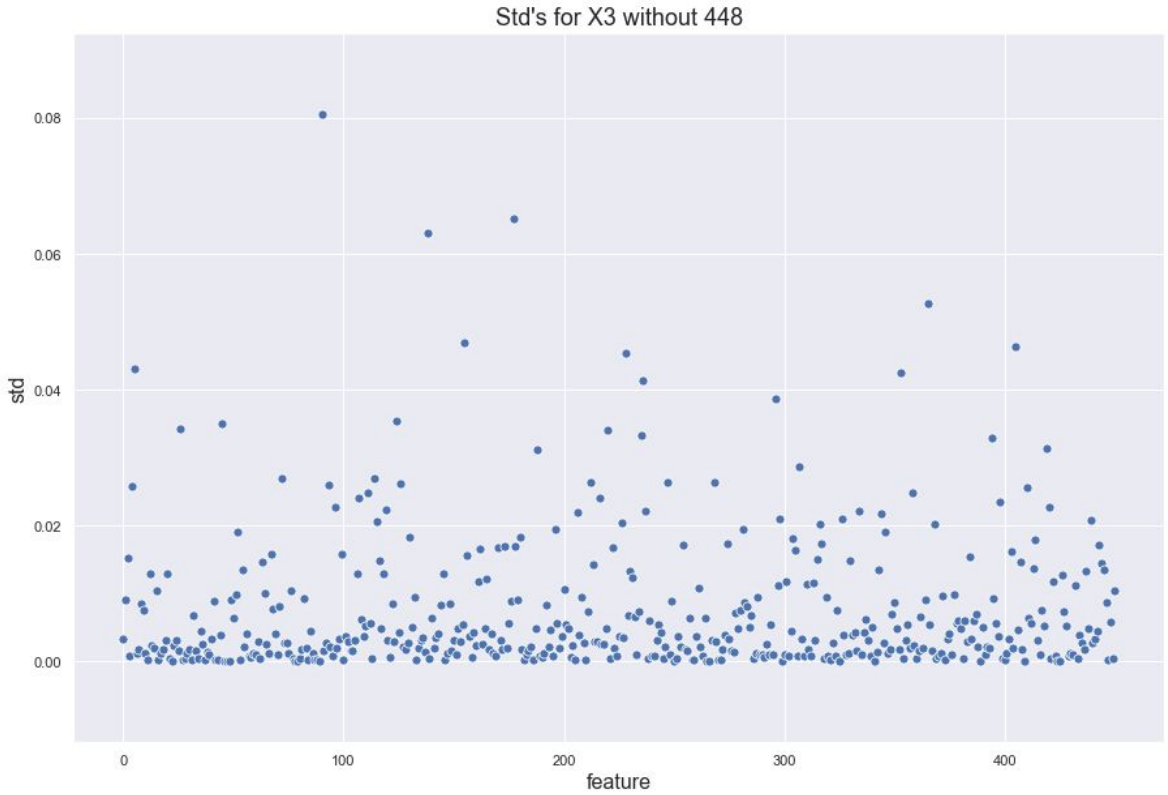

Std's for X3 without 448
std
0.08
0.06
0.04
0.02
0.00
0
100
200
300
400
feature

Таблица X3

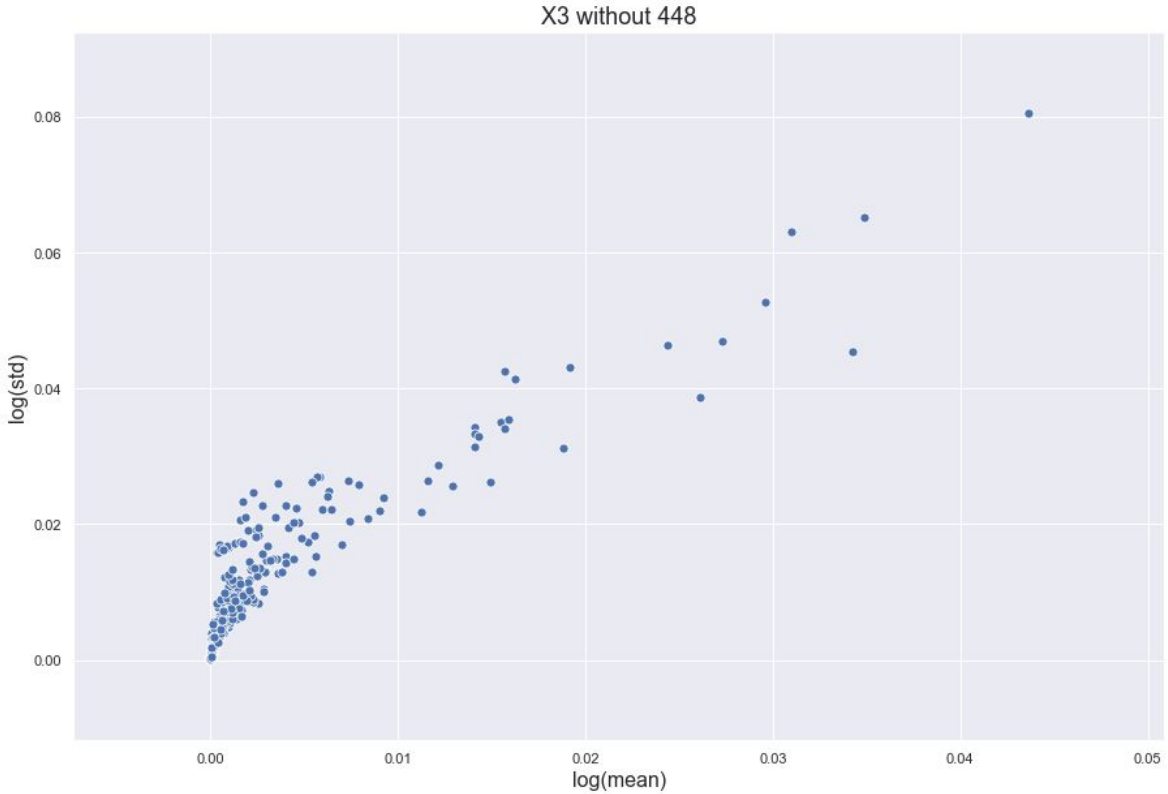
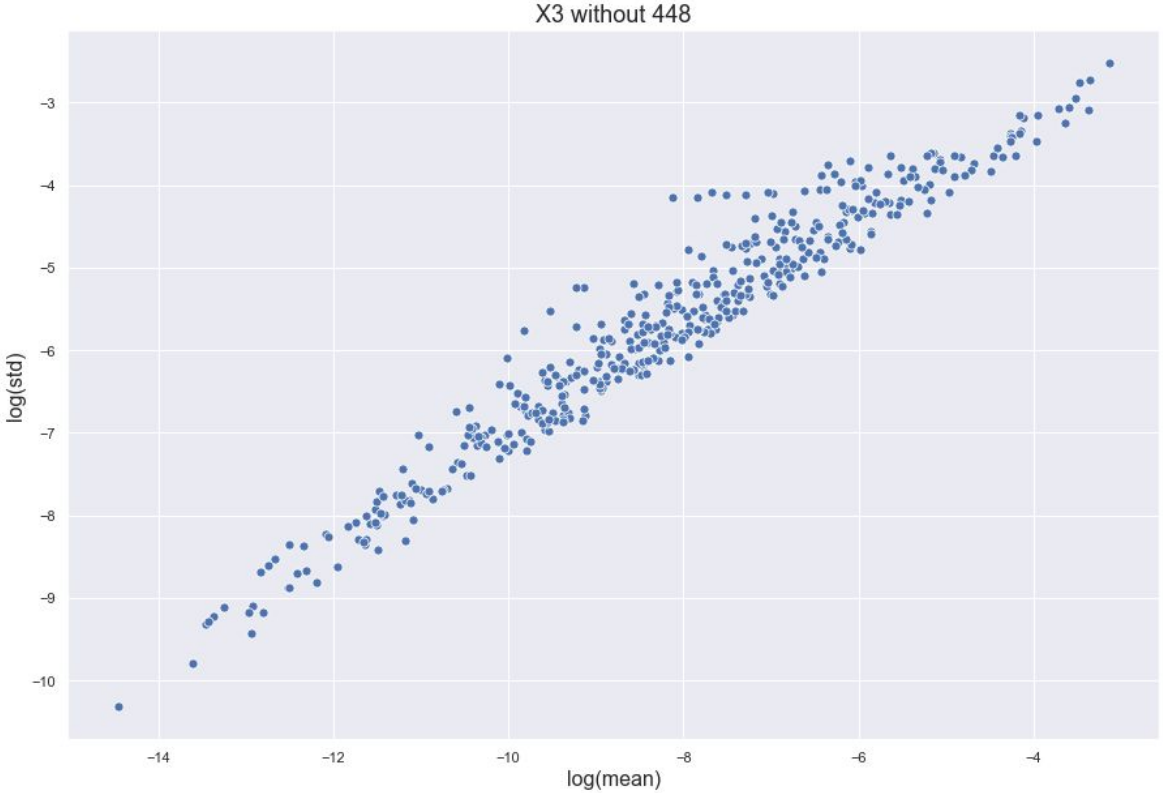

X3 without 448
log(std)
log(mean)
0.00
0.02
0.04
0.06
0.08
0.00
0.01
0.02
0.03
0.04
0.05

Таблица X3



$$\mathbb{D}\xi \simeq 2.24 * \mathbb{E}\xi^{1.3}$$

Таблица X3

В обучающей выборке есть нулевые признаки! При этом в контрольной выборке у таких признаков есть и другие значения.

```
410 train:  1 [0.] [4000]
    test:  2 [0. 0.0227] [4057    1]
23  train:  1 [0.] [4000]
    test:  4 [0. 0.0013 0.0017 0.0151] [4055    1    1    1]
266 train:  1 [0.] [4000]
    test:  2 [0. 0.0010] [4057    1]
426 train:  1 [0.] [4000]
    test:  2 [0. 0.0068] [4057    1]
12  train:  2 [0. 0.0106] [3999    1]
    test:  7 [0. 0.0020 0.0024 0.0056 0.0091 0.0172 0.02] [4051    2    1    1    1    1    1]
```

Таблица X3

Бывает и наоборот:

```
67  train:  5 [0. 0.0037 0.0038 0.0204 0.0714] [3996    1    1    1    1]
    test:  5 [0. 0.0010 0.0024 0.004 0.0196] [4054    1    1    1    1]
81  train:  5 [0. 0.0020 0.008 0.0126 0.0169] [3996    1    1    1    1]
    test:  1 [0.] [4058]
```

Таблица X3

```
np.unique(X3['172'], return_counts=True)
```

```
(array([0.        , 0.00104058, 0.00112613, 0.00135318, 0.00153374,
        0.00173913, 0.00179211, 0.00185874, 0.00189036, 0.00204499,
        0.00218341, 0.00225225, 0.00230415, 0.00238095, 0.00255754,
        0.00271003, 0.00274725, 0.0028169 , 0.00288184, 0.00292398,
        0.00294985, 0.00295858, 0.00307692, 0.00362319, 0.00367647,
        0.00381679, 0.00383142, 0.00389105, 0.00413223, 0.00425532,
        0.004329  , 0.00440529, 0.00460829, 0.00497512, 0.00505051,
        0.00534759, 0.00540541, 0.00558659, 0.00589971, 0.00595238,
        0.00628931, 0.00645161, 0.00671141, 0.00680272, 0.00684932,
        0.00694444, 0.00724638, 0.00740741, 0.00769231, 0.00819672,
        0.00847458, 0.00854701, 0.00900901, 0.00909091, 0.00952381,
        0.00990099, 0.0106383 , 0.01075269, 0.01086957, 0.01149425,
        0.01204819, 0.01234568, 0.01449275, 0.015625  , 0.01666667,
        0.01694915, 0.01785714, 0.01960784, 0.02      , 0.02325581,
        0.02631579, 0.02702703, 0.02941176, 0.03333333, 0.04      ,
        0.05263158, 0.16666667]),
 array([3916,    1,    1,    1,    1,    1,    1,    2,    1,    1,    1,
           1,    1,    1,    1,    1,    1,    1,    1,    1,    1,    1,
           1,    1,    1,    1,    1,    1,    1,    1,    1,    1,    1,
           1,    1,    1,    2,    1,    1,    1,    3,    1,    2,    1,
           1,    1,    1,    1,    1,    1,    1,    1,    1,    1,    1,
           1,    1,    1,    1,    1,    1,    1,    2,    1,    1,    1,
           1,    3,    1,    1,    1,    1,    1,    1,    1,    1,    1]))
```

Большинство признаков почти всегда равны нулю. Даже признаки с большим числом уникальных значений принимают эти значения по одному разу.

Таблица X2

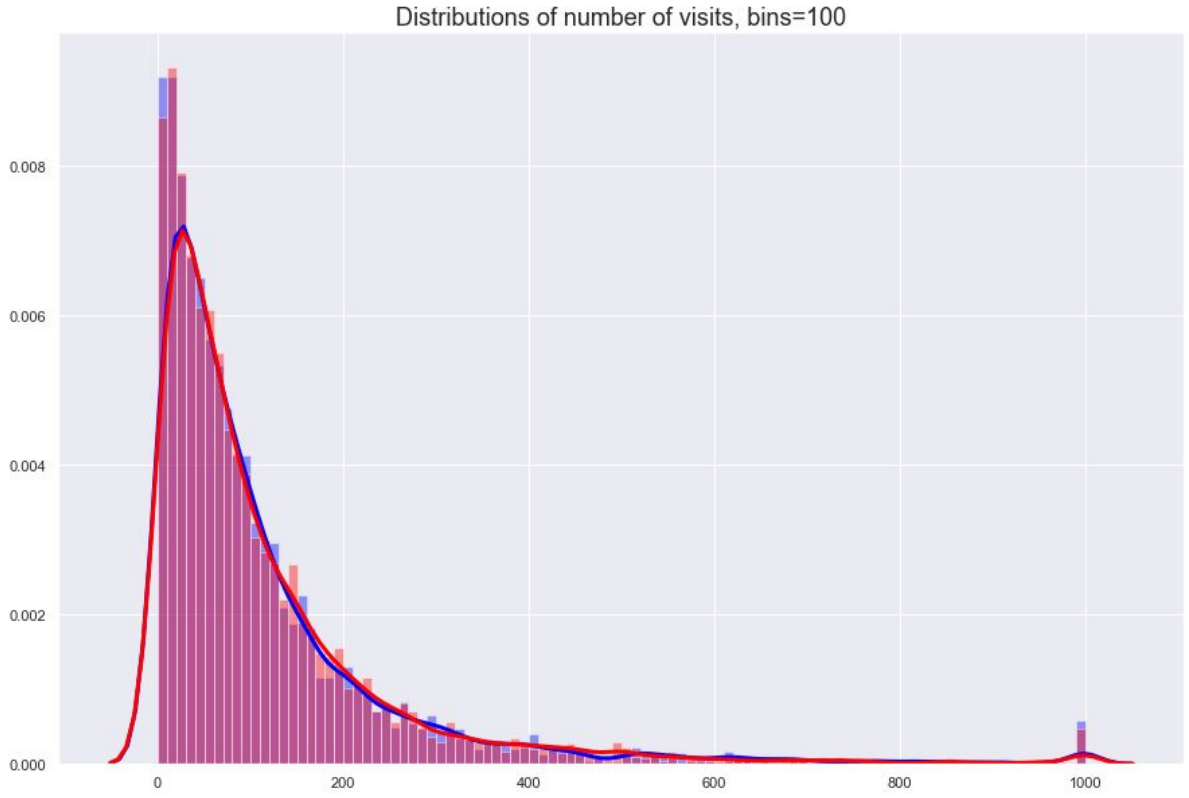

Distributions of number of visits, bins=100
0.008
0.006
0.004
0.002
0.000
0
200
400
600
800
1000

Таблица X2

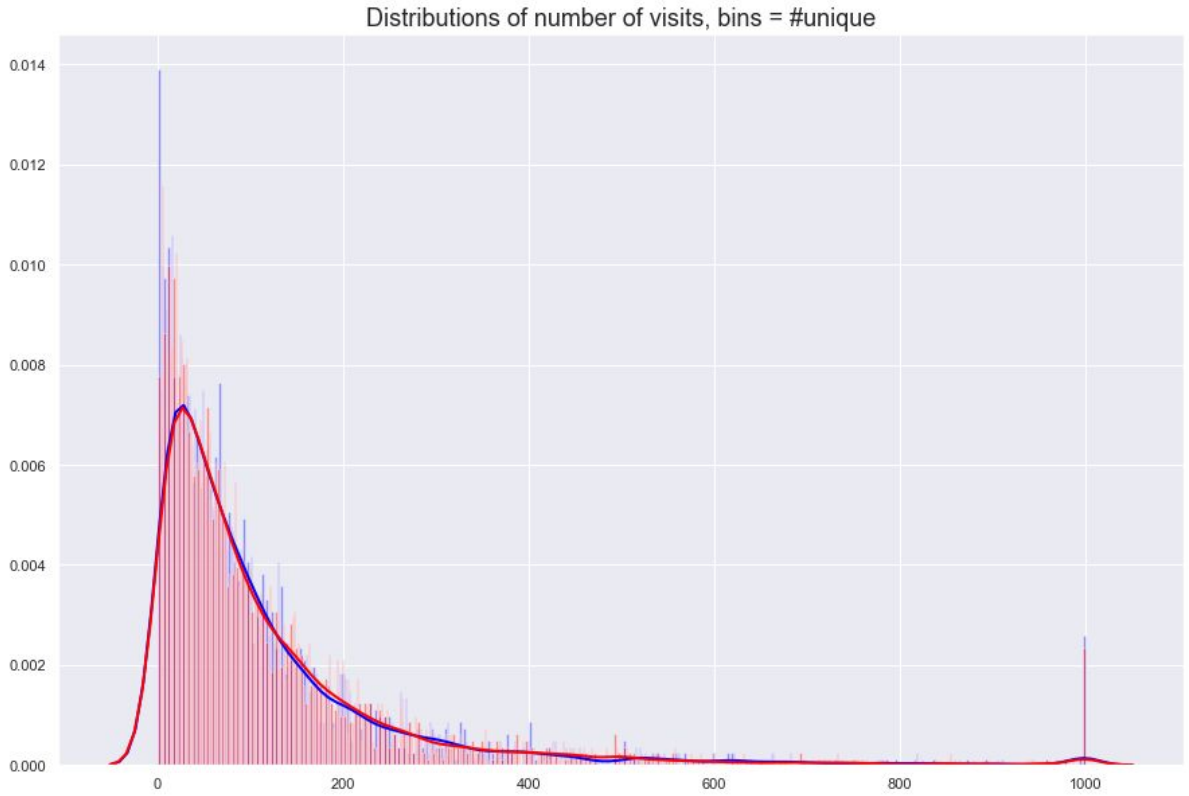

Distributions of number of visits, bins = #unique
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
200
400
600
800
1000

Таблица X2


Distributions of log(number of visits), bins=100
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8

Таблица X2

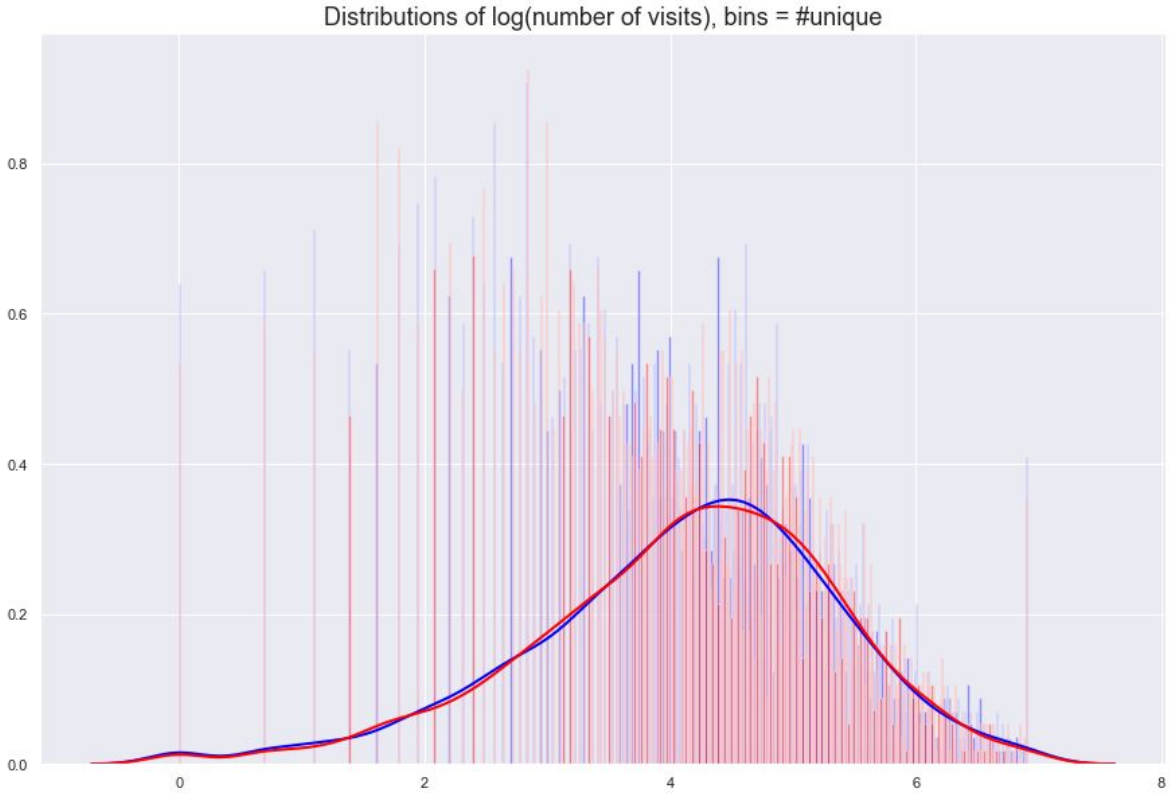

Distributions of log(number of visits), bins = #unique
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8

Таблица X2

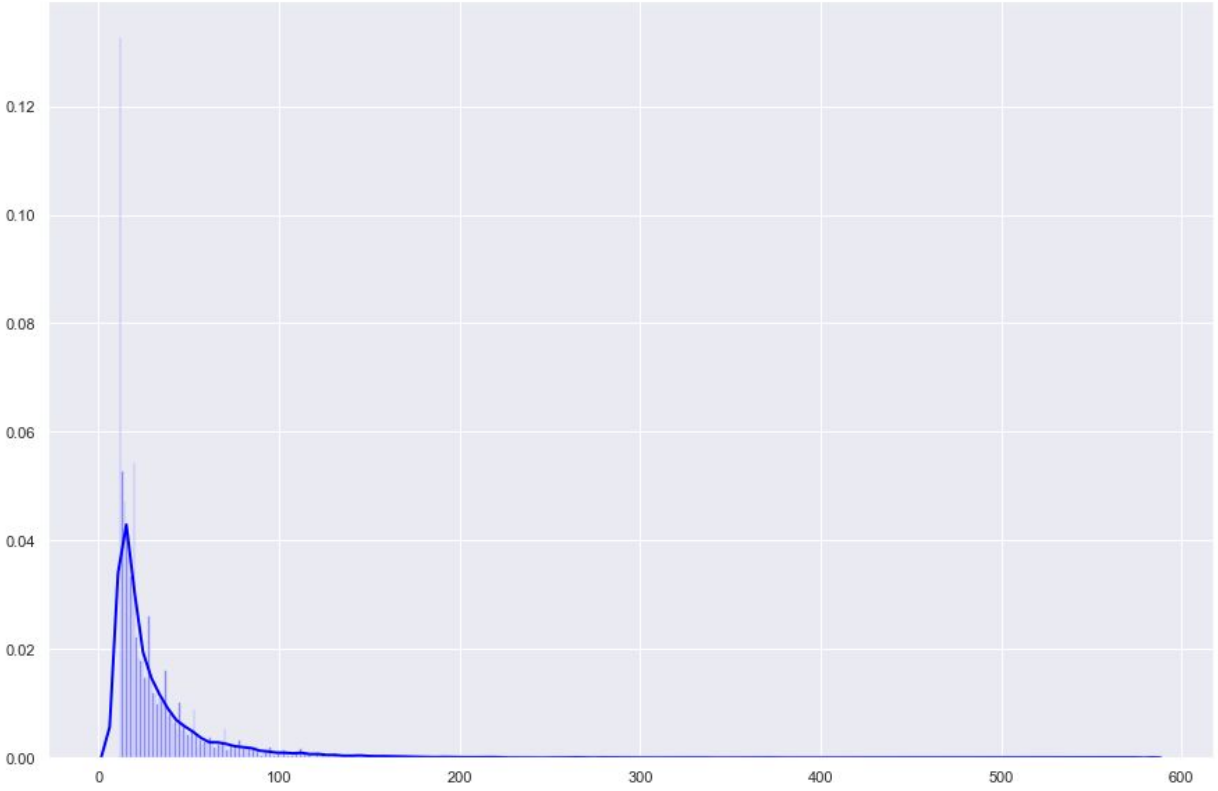


Таблица X2

```
Top 10 most popular urls in train:  [52464 20263 30159 68094 15593 44123 14038 67589 21542 49258]
Times visited:  [579 555 503 435 414 413 392 392 371 368]

Top 10 most popular urls in test:  [20263 52464 30159 44123 68094 15593 14038 49258 21542 47634]
Times visited:  [584 576 526 459 442 420 404 398 397 379]
```

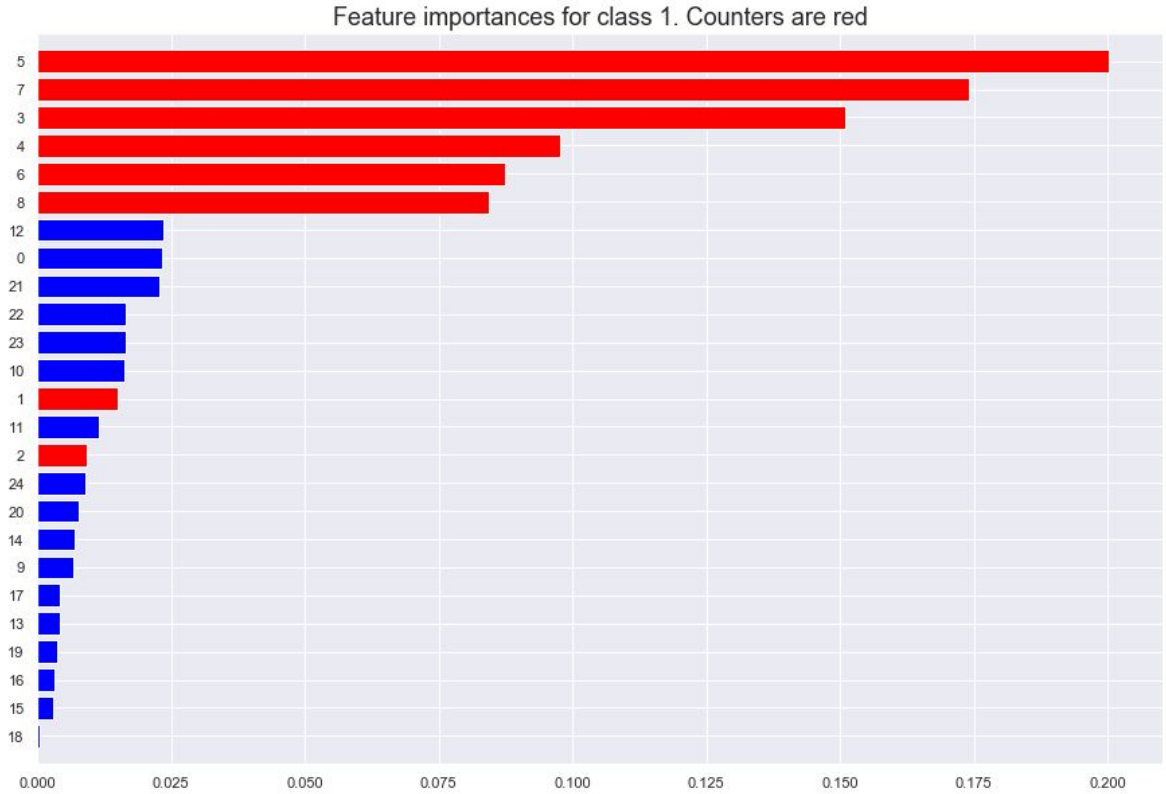
```
Number of urls with more than 10 views in train:  7409
Number of urls with more than 10 views in test:  7606
```

Посещения сайтов совпадают на обучающей и контрольной выборках.

Целевая переменная

```
Percentage of ones for each target:
id    406031.125
1         30.375
2         34.225
3         32.825
4         31.850
5         35.675
dtype: float64
```

Важность признаков в X1

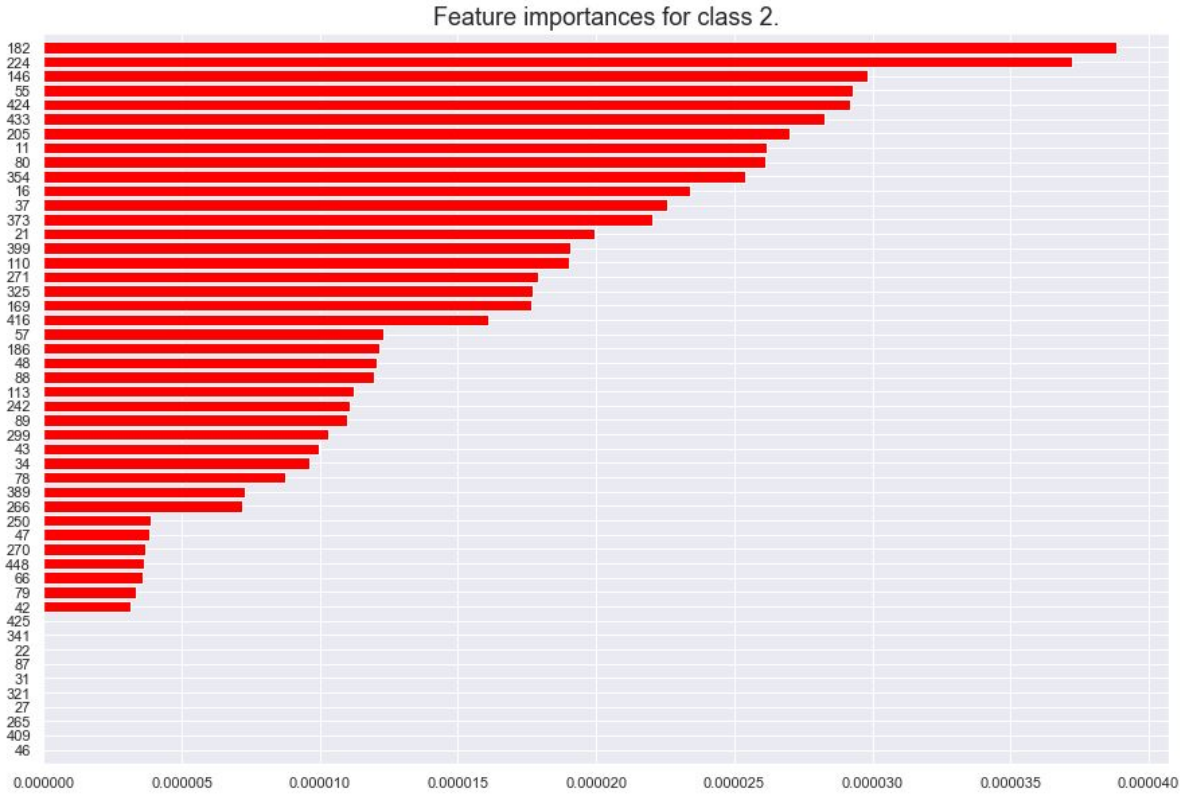


Mean AUC-ROC = 0.5614

По классам есть небольшой разброс в значении AUC-ROC.

```
Score for class 0 = 0.569
Score for class 1 = 0.544
Score for class 2 = 0.583
Score for class 3 = 0.561
Score for class 4 = 0.551
```

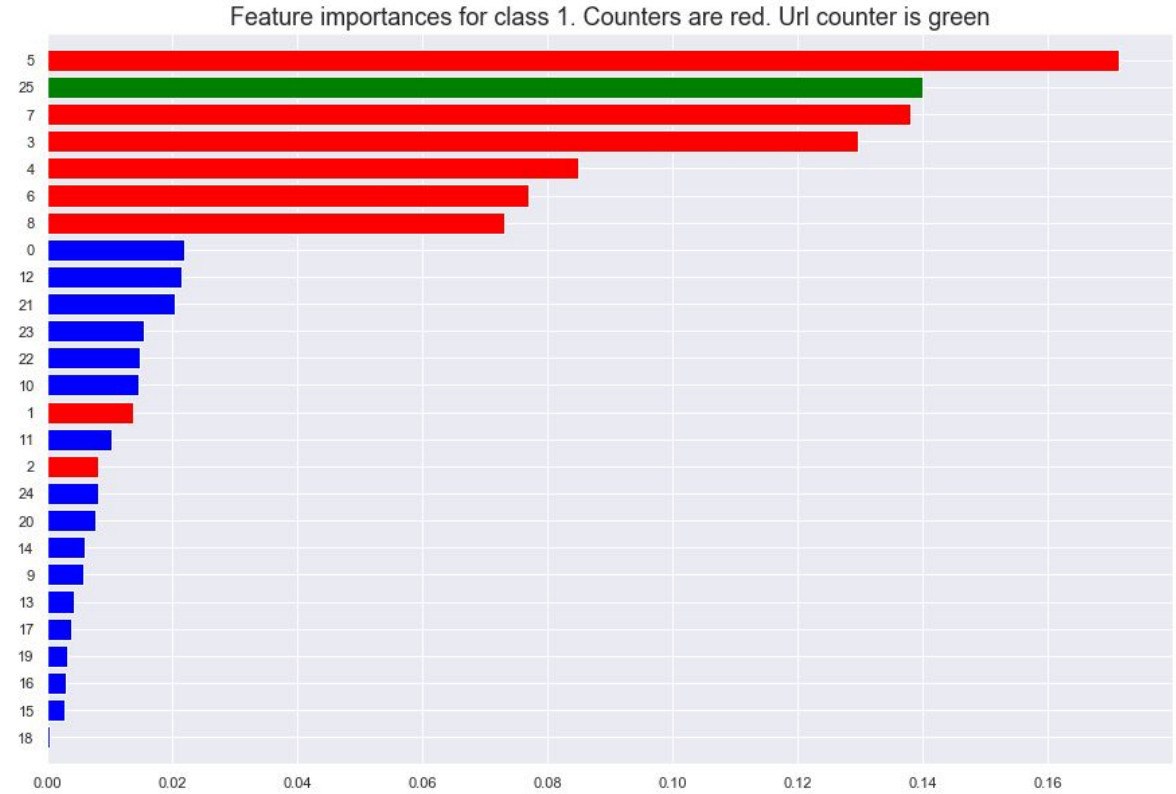
Важность признаков в X3



Mean AUC-ROC = 0.5049

```
Score for class 0 = 0.497
Score for class 1 = 0.518
Score for class 2 = 0.507
Score for class 3 = 0.494
Score for class 4 = 0.508
```

Важность признаков в X1 + счётчик посещений



Mean AUC-ROC = 0.5621

```
Score for class 0 = 0.576
Score for class 1 = 0.550
Score for class 2 = 0.580
Score for class 3 = 0.562
Score for class 4 = 0.542
```