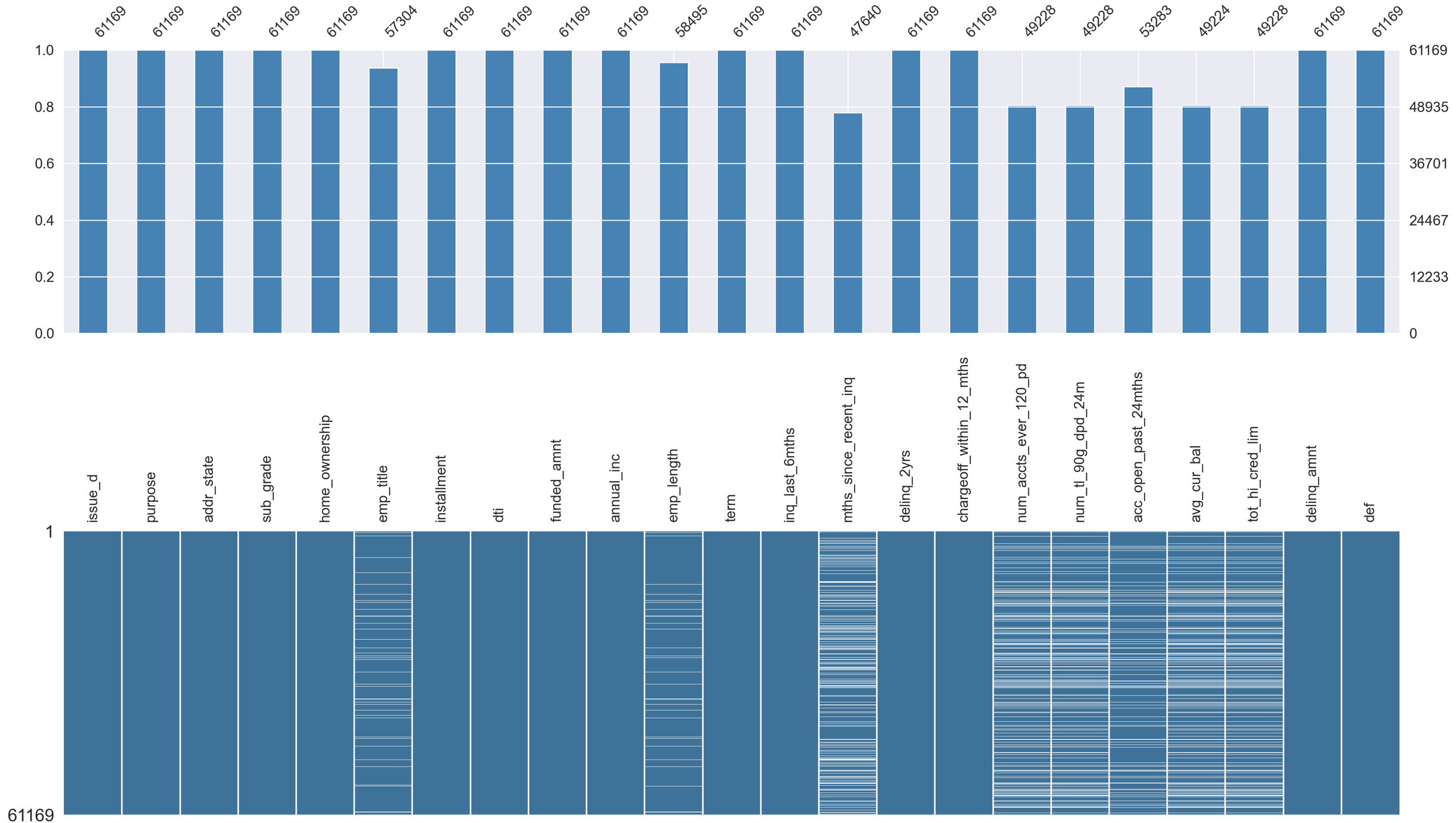
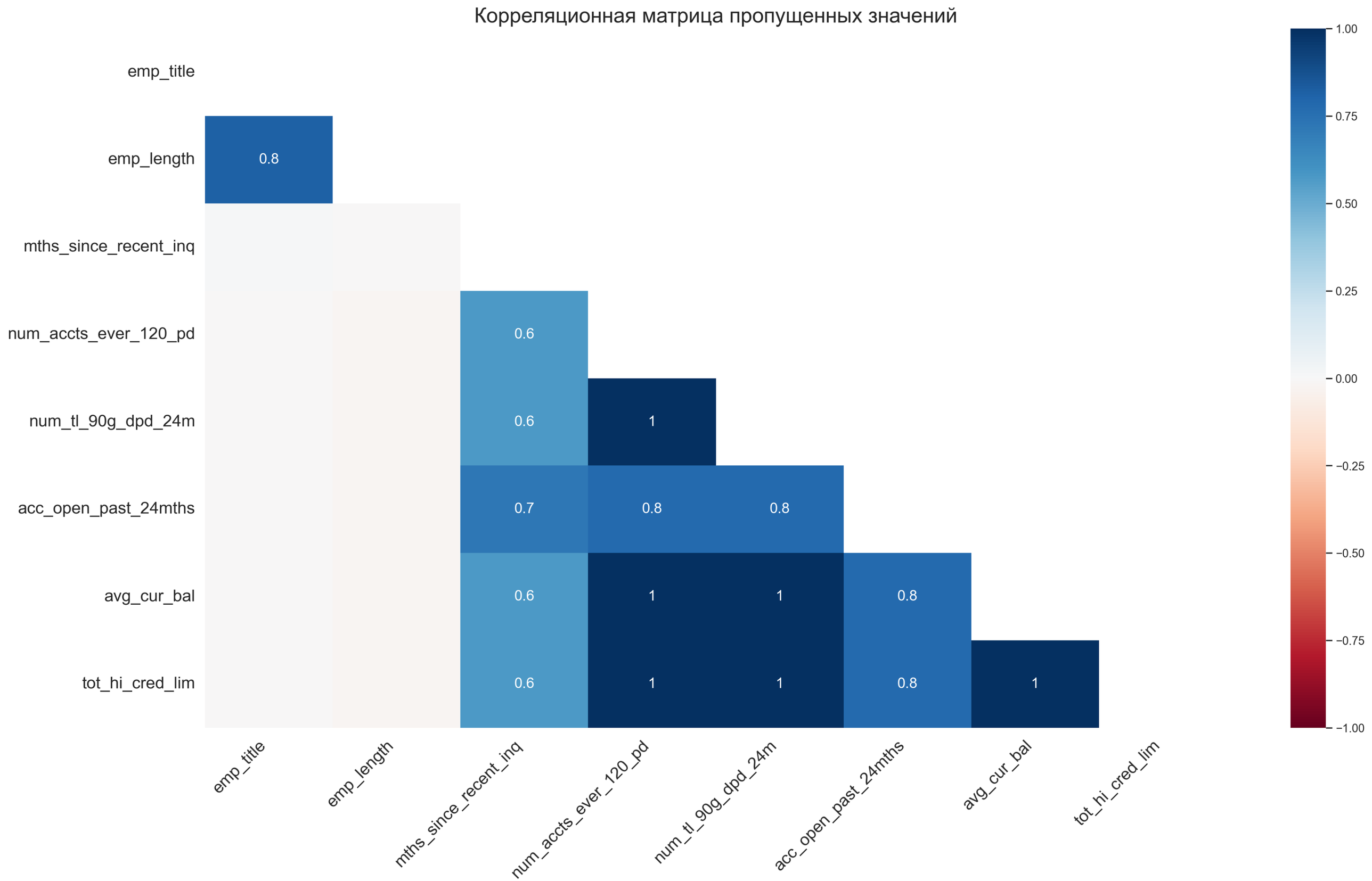
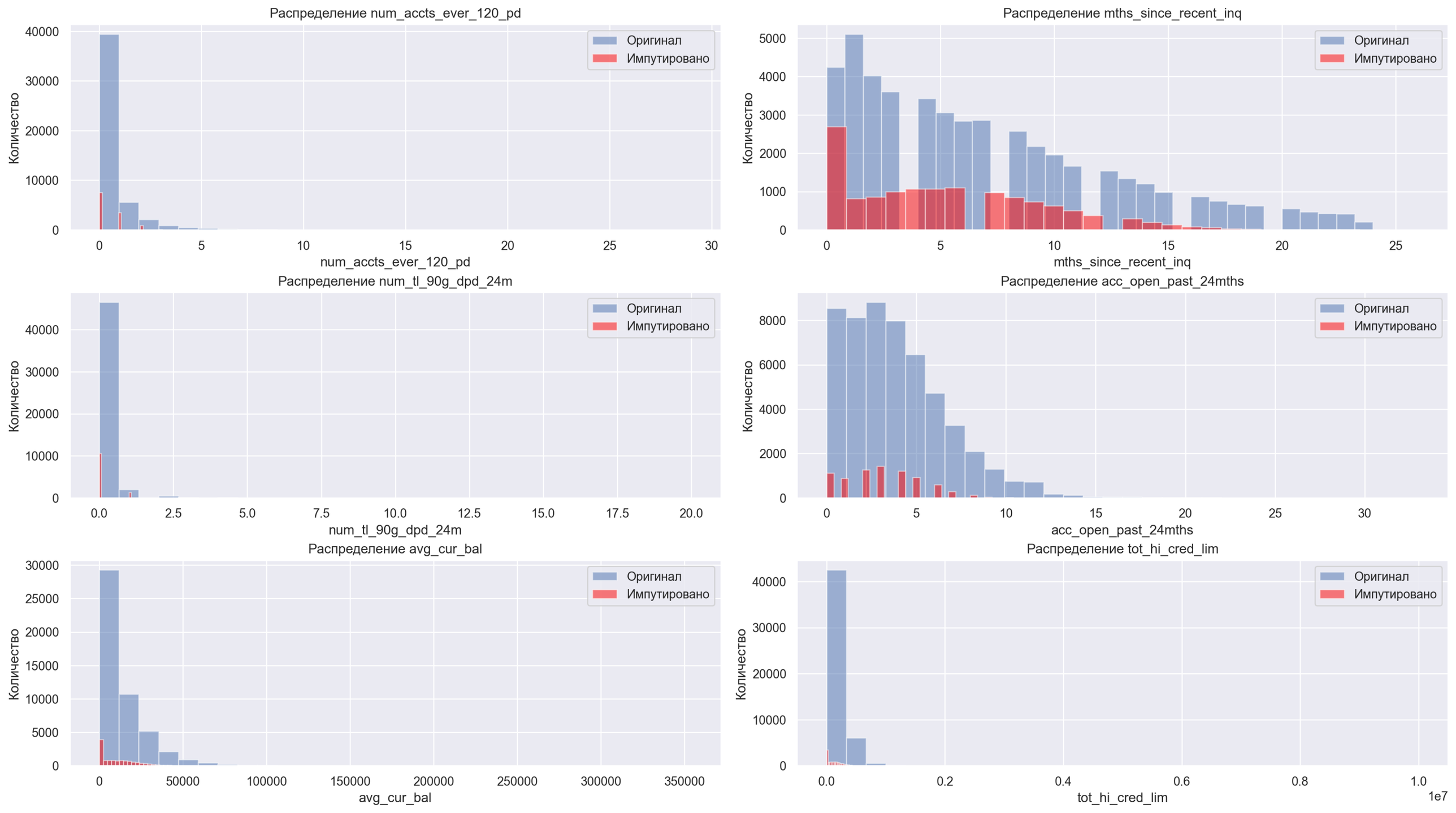
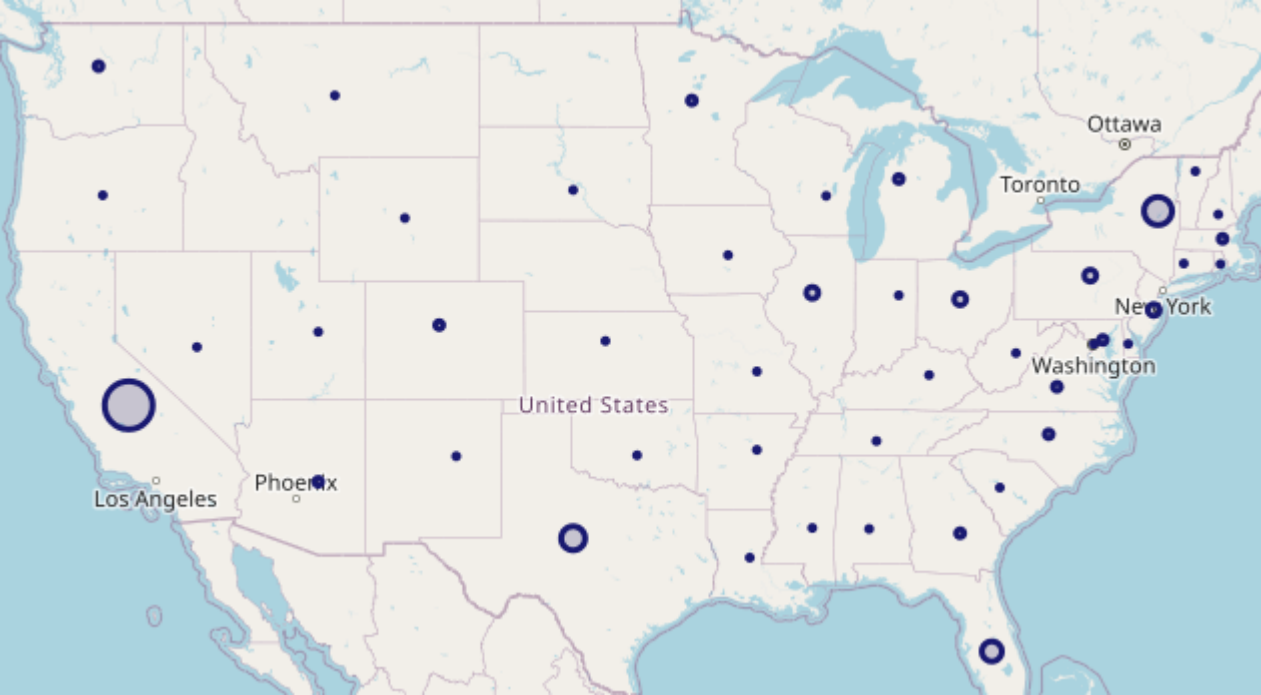
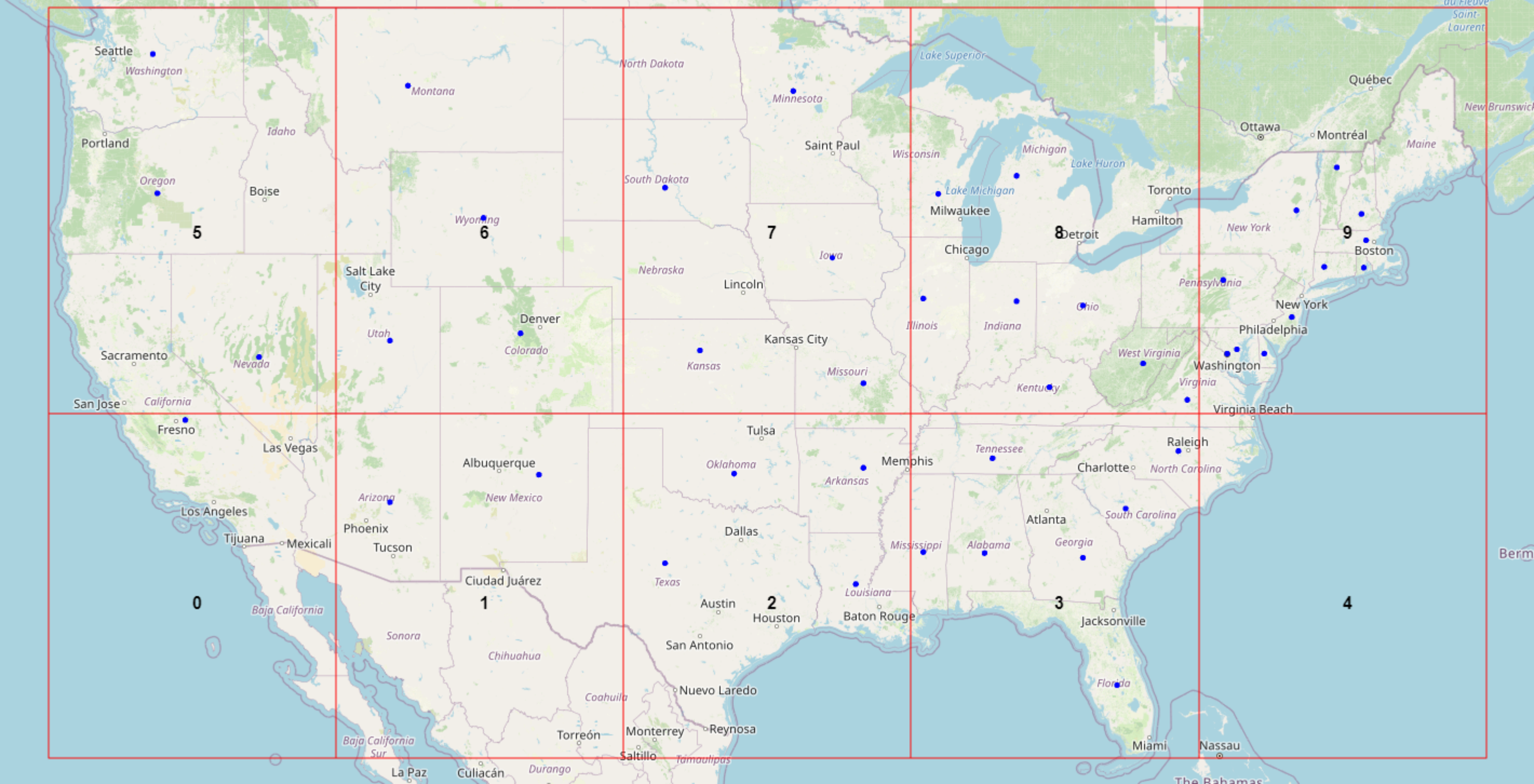
Гистограмма и матрица пропущенных значений



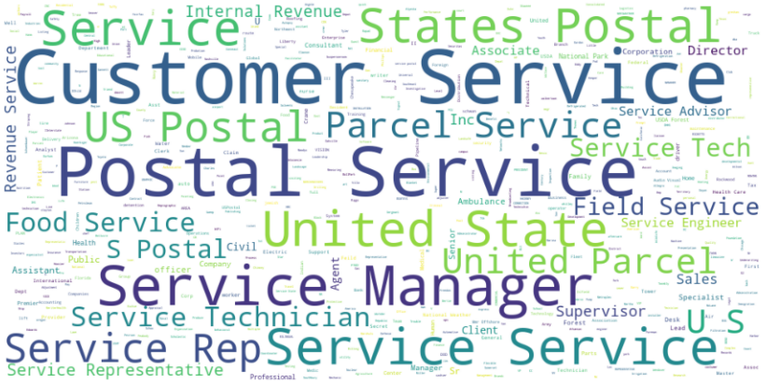

Корреляционная матрица пропущенных значений
emp_title
emp_length
mths_since_recent_inq
num_accts_ever_120_pd
num_tl_90g_dpd_24m
acc_open_past_24mths
avg_cur_bal
tot_hi_cred_lim
0.8
0.6
0.6
1
0.7
0.8
0.8
0.6
1
1
0.8
0.6
1
1
0.8
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Распределение num_accts_ever_120_pd
Оригинал
Импутировано
Количество
num_accts_ever_120_pd
Распределение mths_since_recent_inq
Оригинал
Импутировано
Количество
mths_since_recent_inq
Распределение num_tl_90g_dpd_24m
Оригинал
Импутировано
Количество
num_tl_90g_dpd_24m
Распределение acc_open_past_24mths
Оригинал
Импутировано
Количество
acc_open_past_24mths
Распределение avg_cur_bal
Оригинал
Импутировано
Количество
avg_cur_bal
Распределение tot_hi_cred_lim
Оригинал
Импутировано
Количество
tot_hi_cred_lim
1e7

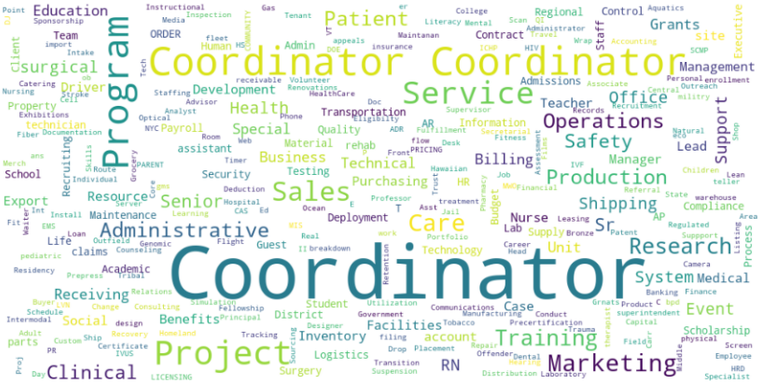

Ottawa
Toronto
New York
Washington
United States
Phoenix
Los Angeles


Seattle
Washington
Portland
Oregon
Idaho
Boise
Montana
North Dakota
Minnesota
Lake Superior
Saint Paul
Wisconsin
Michigan
Lake Huron
Lake Michigan
Milwaukee
Chicago
Toronto
Hamilton
Ottawa
Montréal
Québec
New Brunswick
Maine
New York
Boston
Pennsylvania
New York
Philadelphia
Washington
Virginia
Virginia Beach
Raleigh
North Carolina
Charlotte
South Carolina
Atlanta
Georgia
Jacksonville
Florida
Miami
Nassau
The Bahamas
Berm
Wyoming
South Dakota
Nebraska
Lincoln
Iowa
Salt Lake City
Utah
Denver
Colorado
Kansas
Kansas City
Missouri
Illinois
Indiana
Ohio
West Virginia
Kentucky
Sacramento
Nevada
San Jose
California
Fresno
Las Vegas
Albuquerque
New Mexico
Arizona
Phoenix
Tucson
Los Angeles
Tijuana
Mexicali
Baja California
Sonora
Chihuahua
Ciudad Juárez
Tulsa
Oklahoma
Arkansas
Memphis
Tennessee
Dallas
Texas
Austin
Houston
Louisiana
Baton Rouge
Mississippi
Alabama
San Antonio
Nuevo Laredo
Coahuila
Torreón
Monterrey
Reynosa
Saltillo
Tamaulipas
Baja California Sur
La Paz
Culiacán
Durango
0
1
2
3
4
5
6
7
8
9

Кластер 1

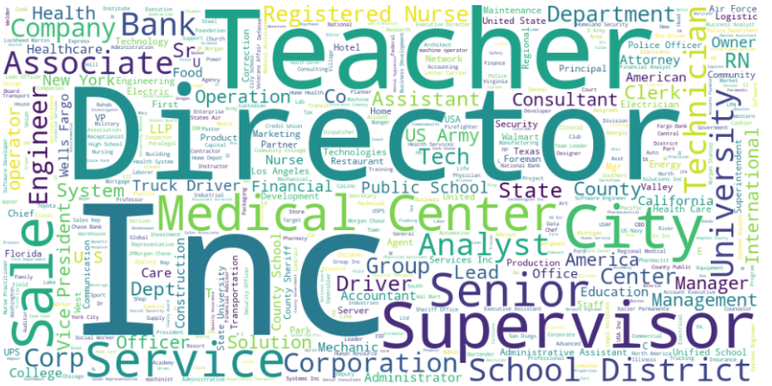


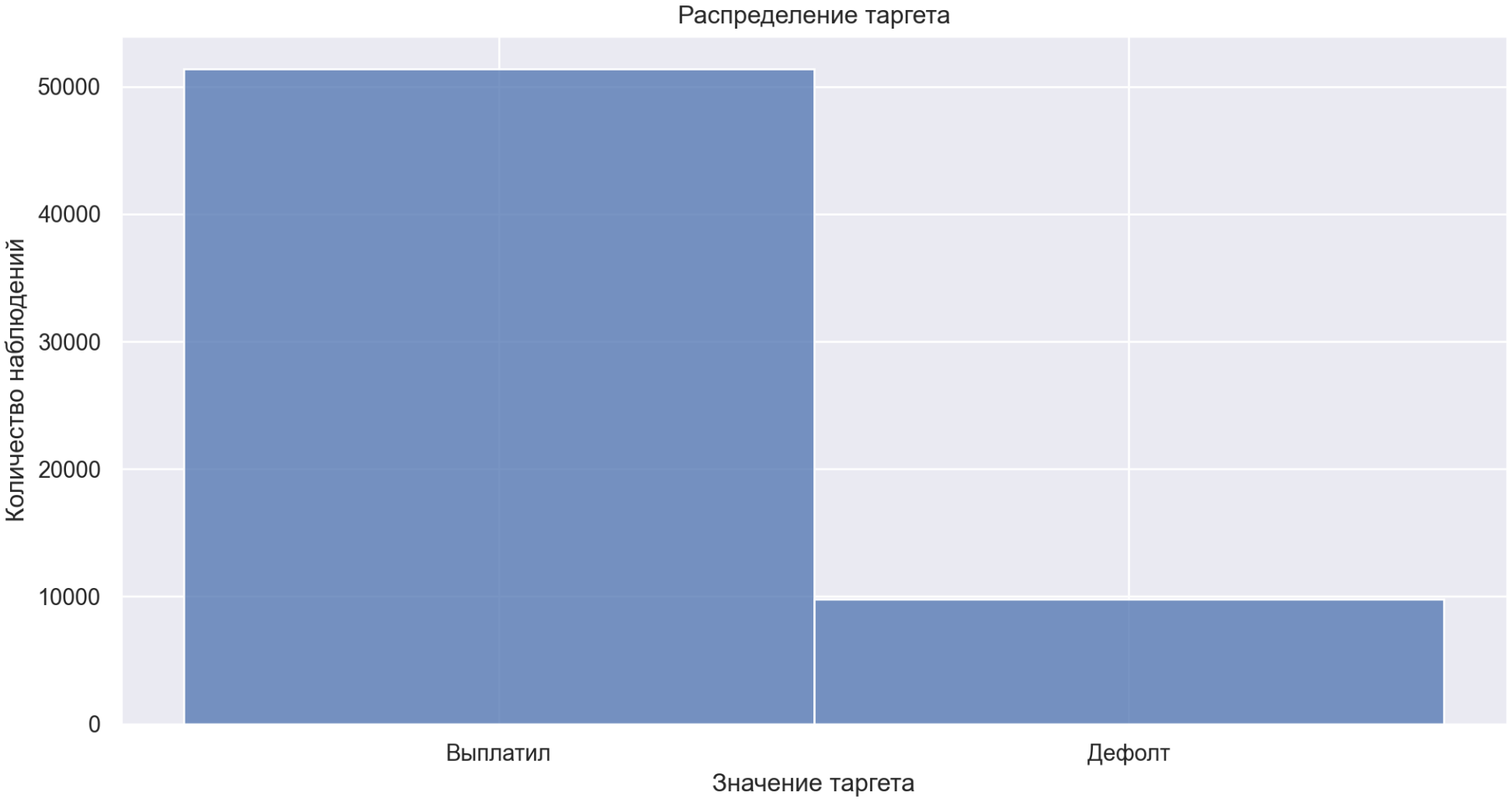
Кластер 2



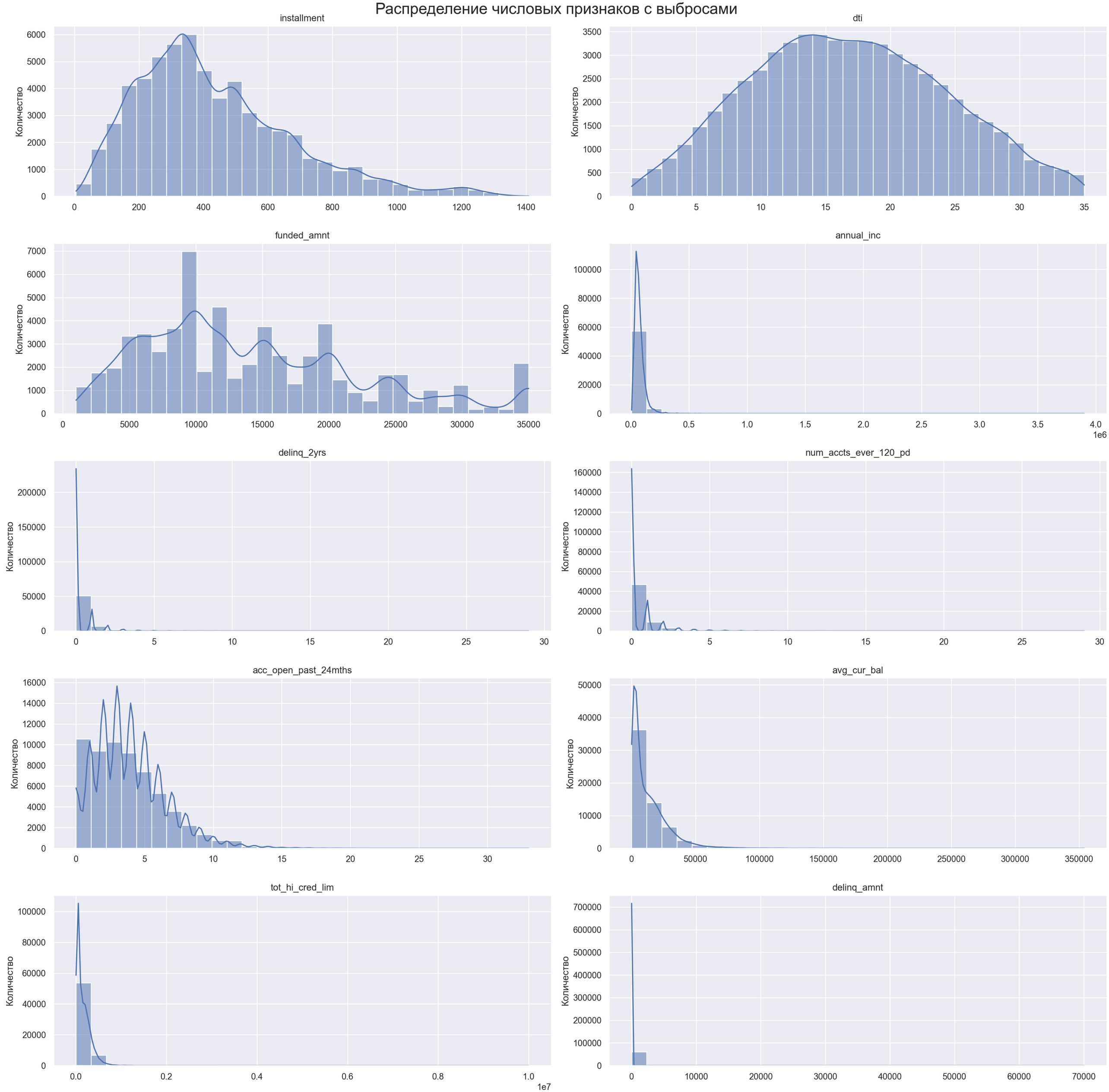
Кластер 3

Кластер 4

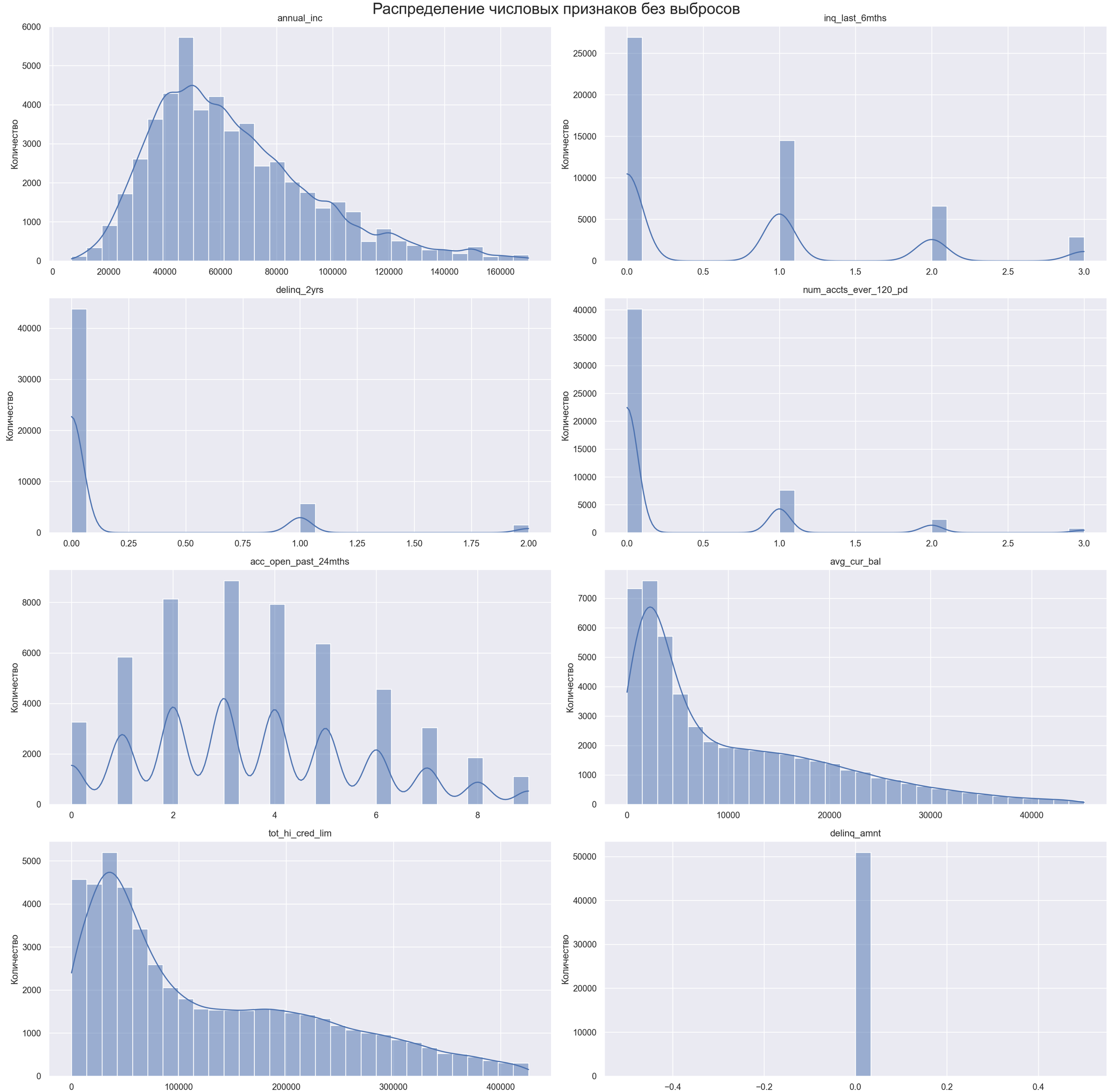

Teacher
Director
Inc
Medical Center
City
Supervisor
Senior
Service
Corporation
School District
Sale
Engineer
Analyst
Consultant
Assistant
Technician
University
Bank
Company
Associate
Registered Nurse
Department
Tech
Truck Driver
Public School
State County
Operation
Manager
Group
Lead
Driver
Accountant
Office
Solution
Mechanic
Administrator
Corp
Dept
Construction
Transportation
Security
Clerk
Attorney
Owner
RN
Engineering
Financial
Nurse
Walmart
Texas
Los Angeles
Foreman
Restaurant
Product
Marketing
Partner
Health Care
Medical
Technologies
Development
Business
Accounting
Management
Director
Supervisor
Senior
Manager
Department
Warehouse
Marketing
Engineering
Operation
Accounting
Area
Manager
Operations
Office
Branch
Store
Sales
Project
Manager
Assistant
General
Office
Operation
Account
Case
Plant
Production
Finance
Business Development
Regional
Store
Program
Property
Quality
Purchasing
Contract
Procurement
Facilities
Administrative
Logistic
Clinical
Senior Project
Division
Credit
Planning
Construction
Corporate
Supply
Warehouse
Assistant Store
Regional Sales
Marketing
Engineering
Manager
Operation Manager
Accounting
Area
Operations
Manager
Office
Manager
Branch
Manager
Senior
Manager
Operation
Manager
Store
Manager
Sales
Manager
Project
Manager
General
Manager
Office
Manager
Manager
Project
Manager
Manager
Assistant
General
Manager
Manager
Manager


Распределение таргета
Количество наблюдений
50000
40000
30000
20000
10000
0
Выплатил
Дефолт
Значение таргета

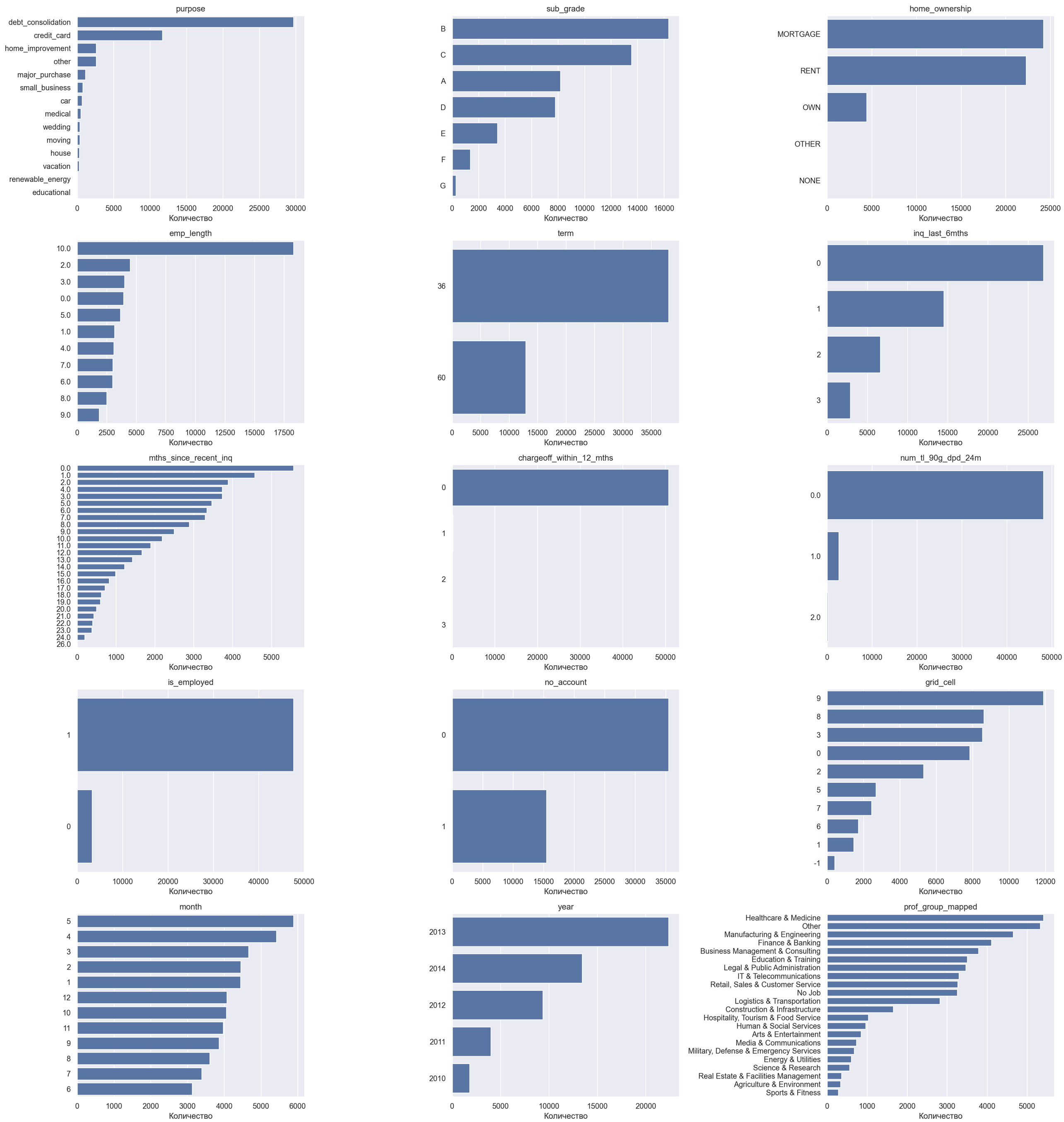
Распределение числовых признаков с выбросами



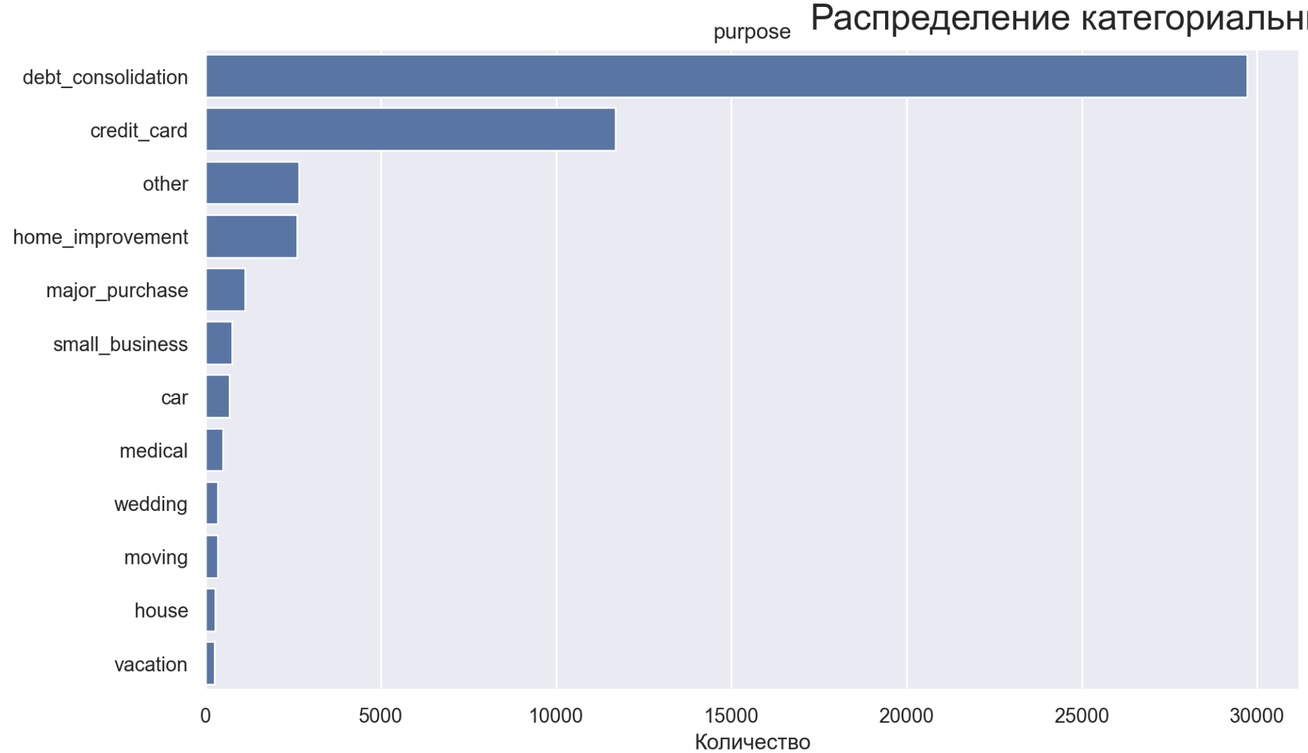
Распределение числовых признаков без выбросов

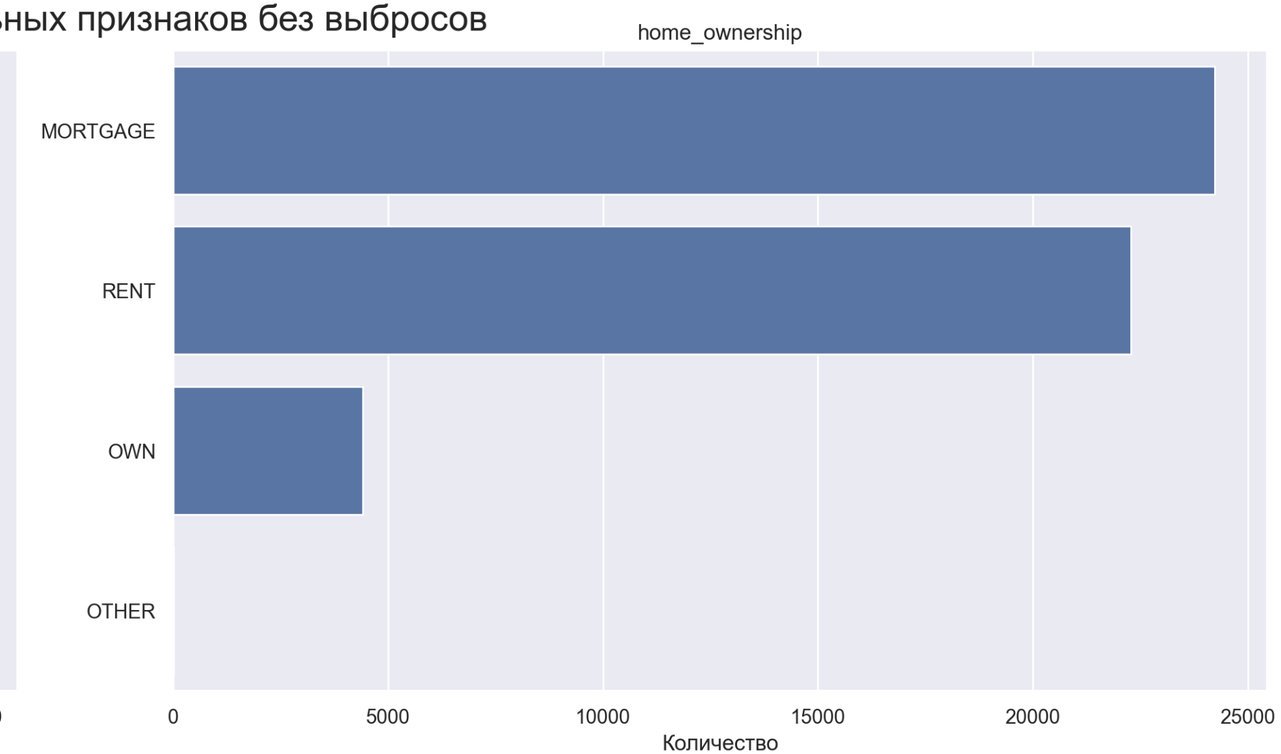


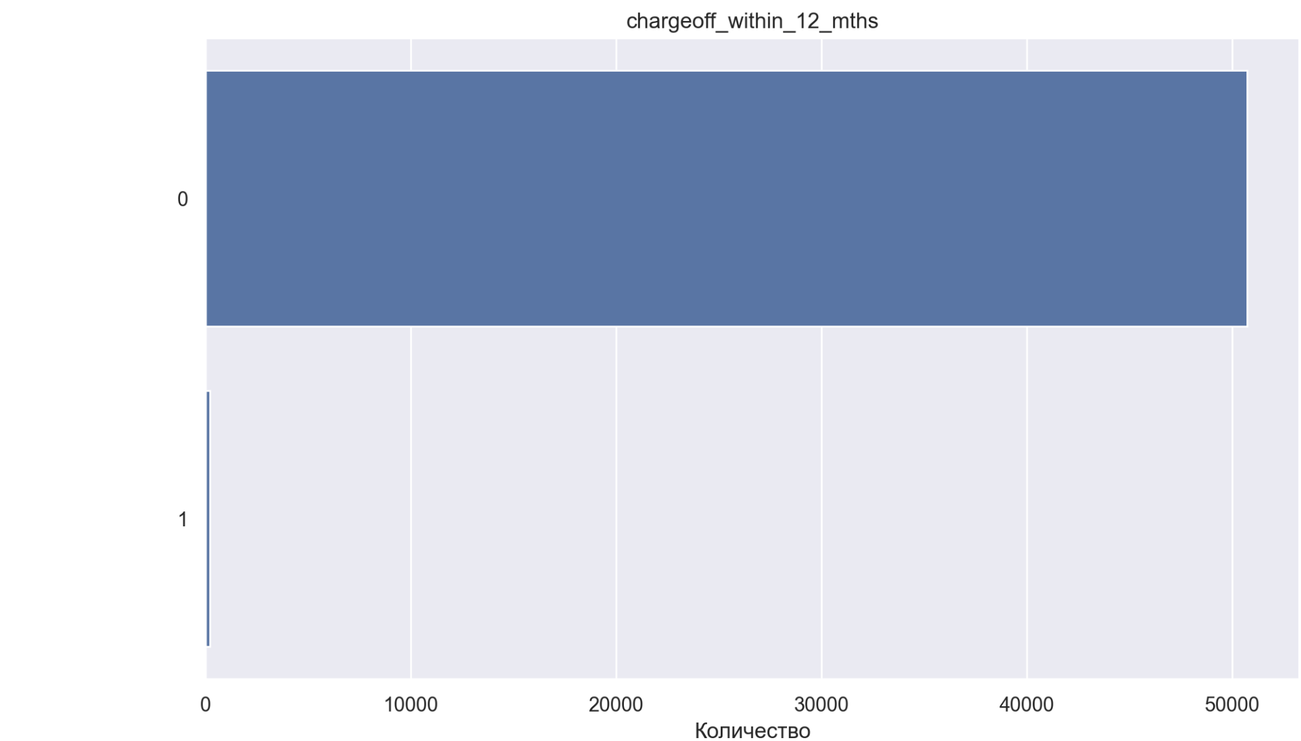
Распределение категориальных признаков с выбросами

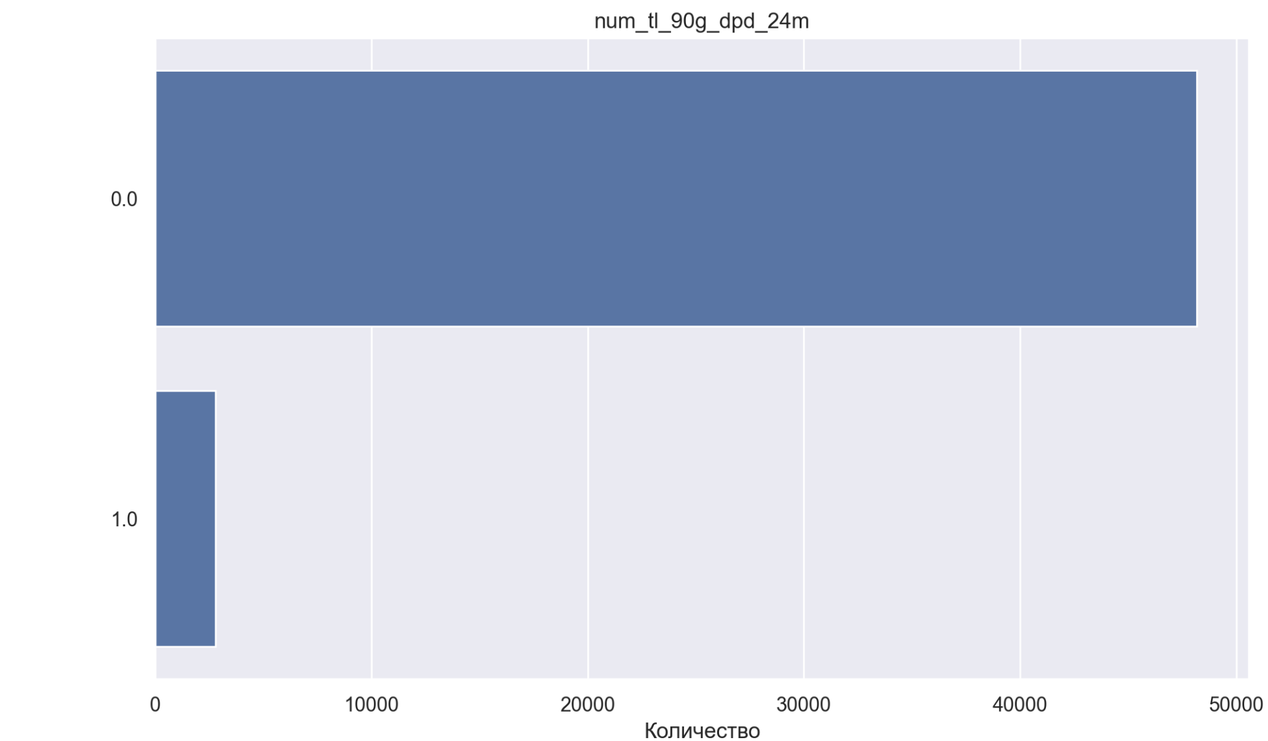


Распределение категориальных признаков без выбросов

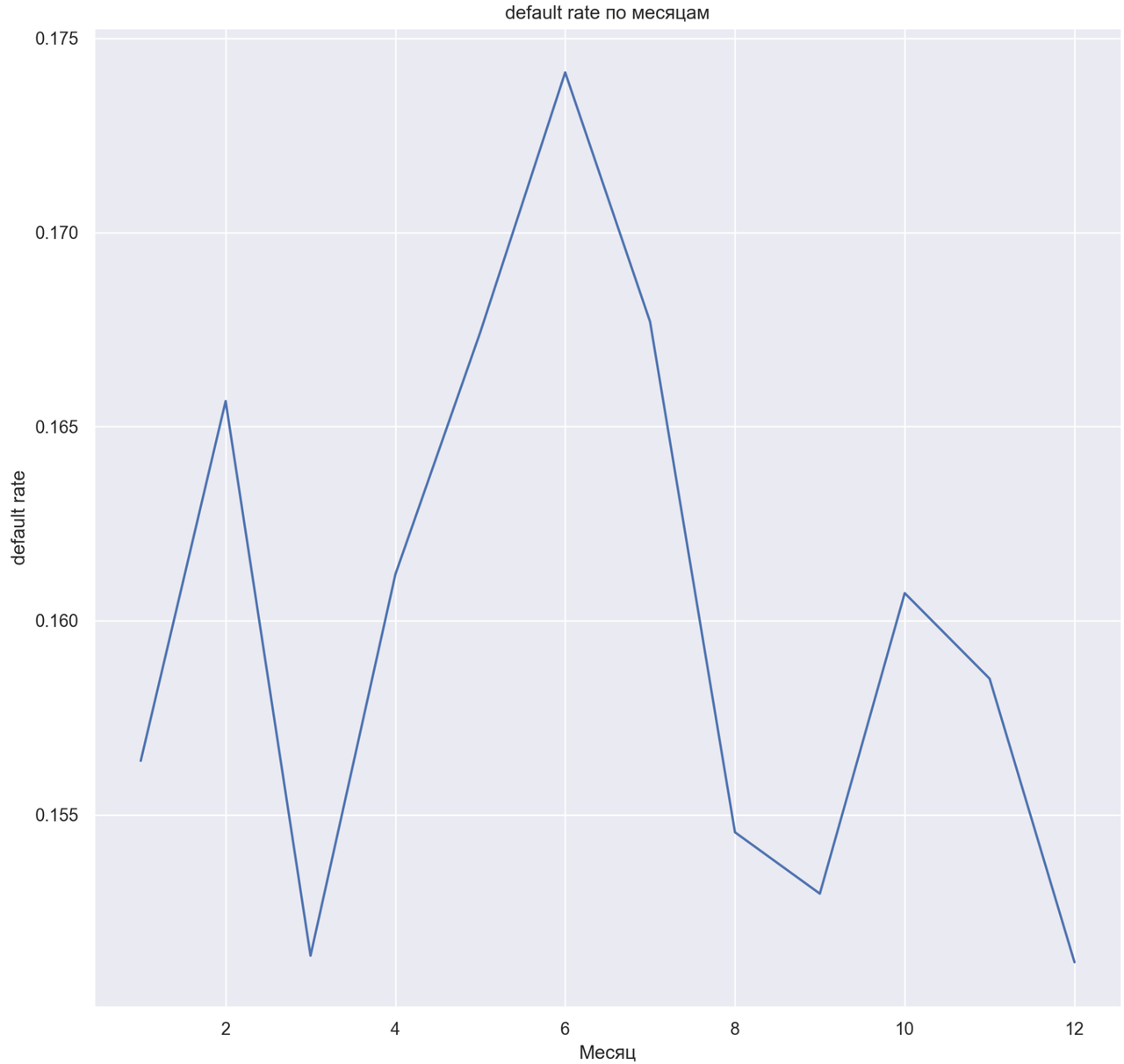


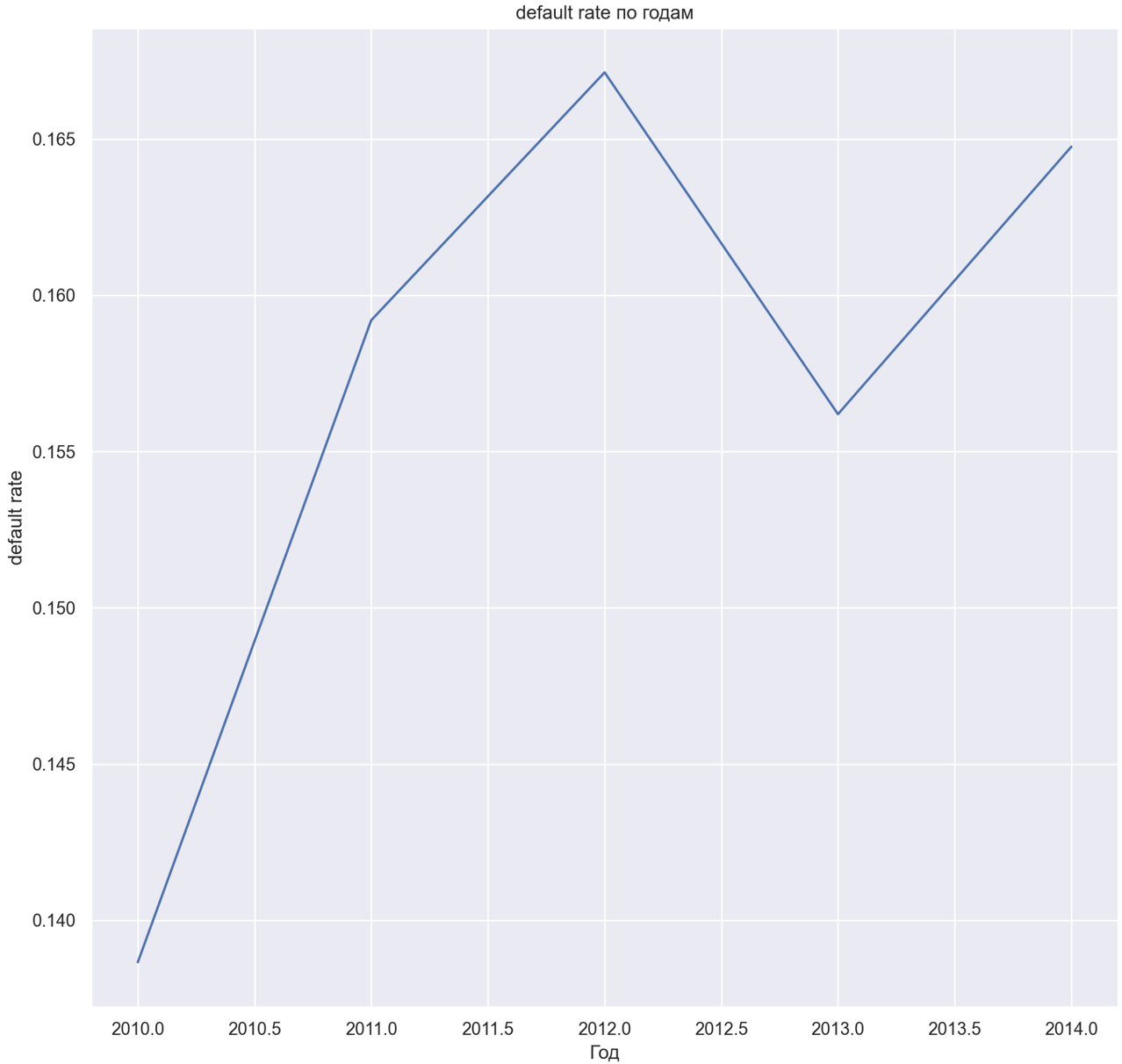




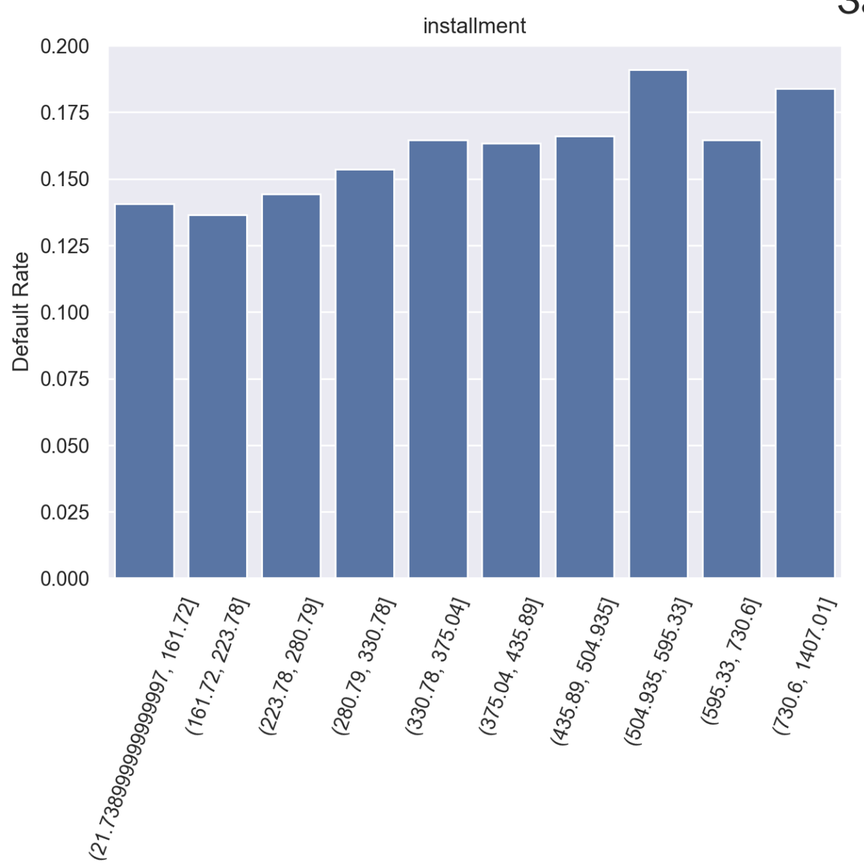


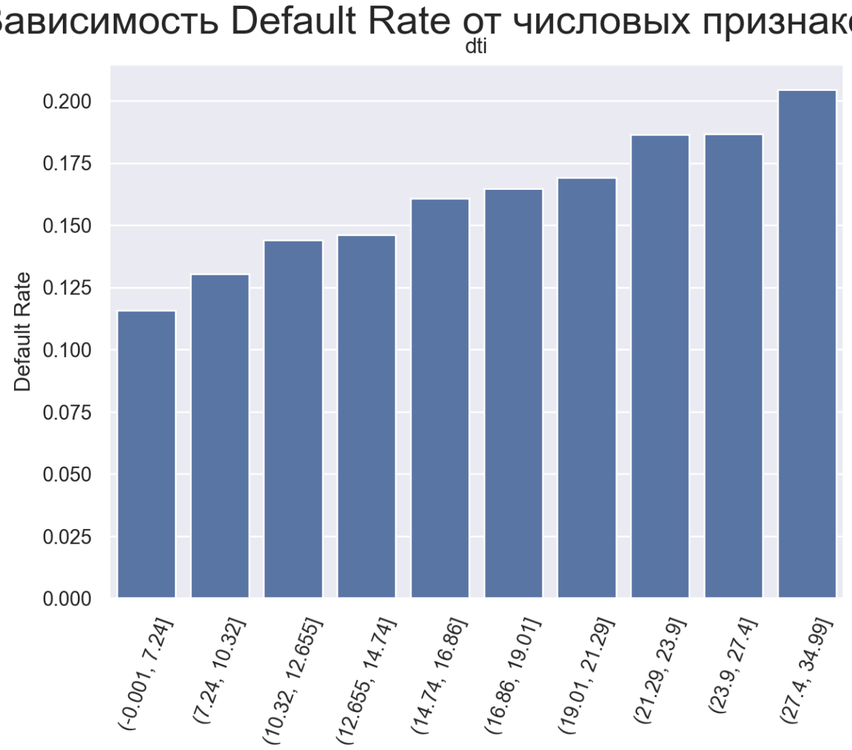
default rate во времени

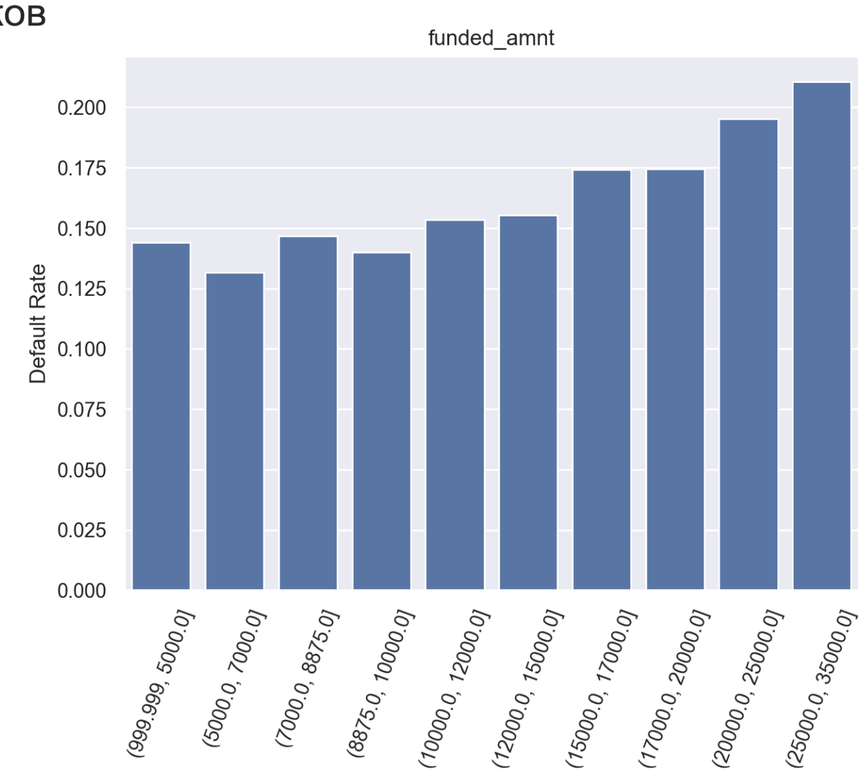


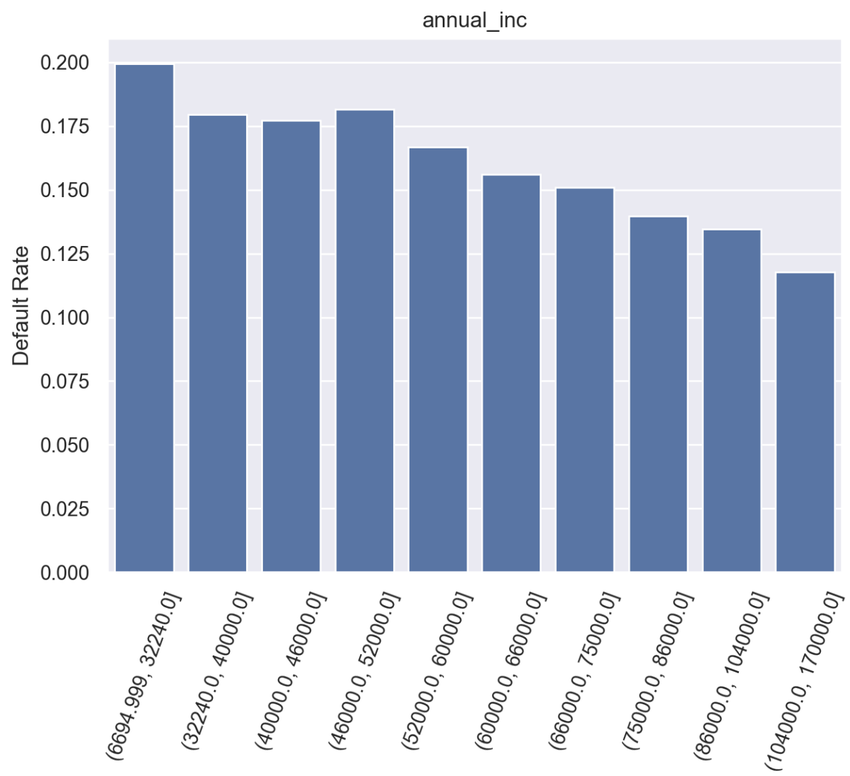


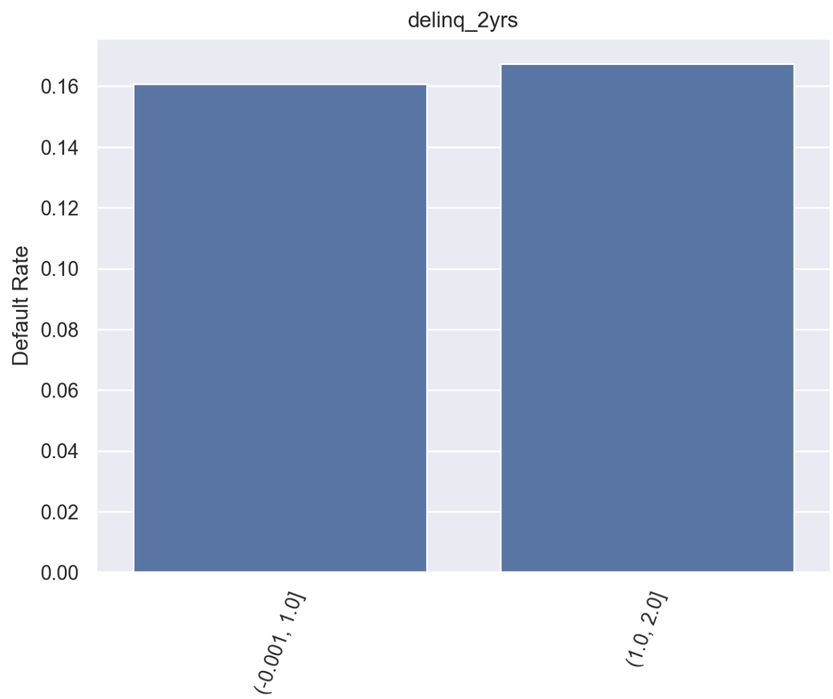
Зависимость Default Rate от числовых признаков

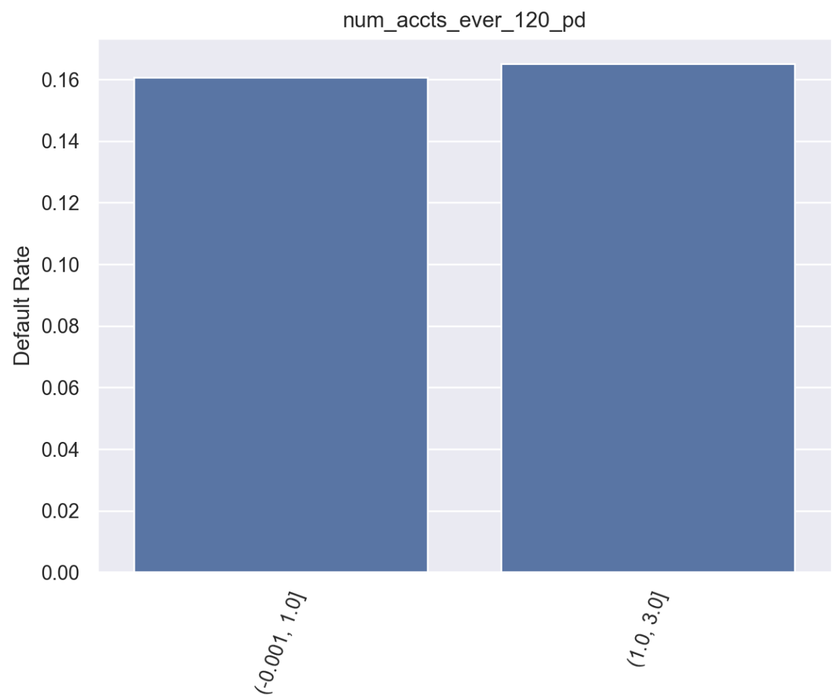


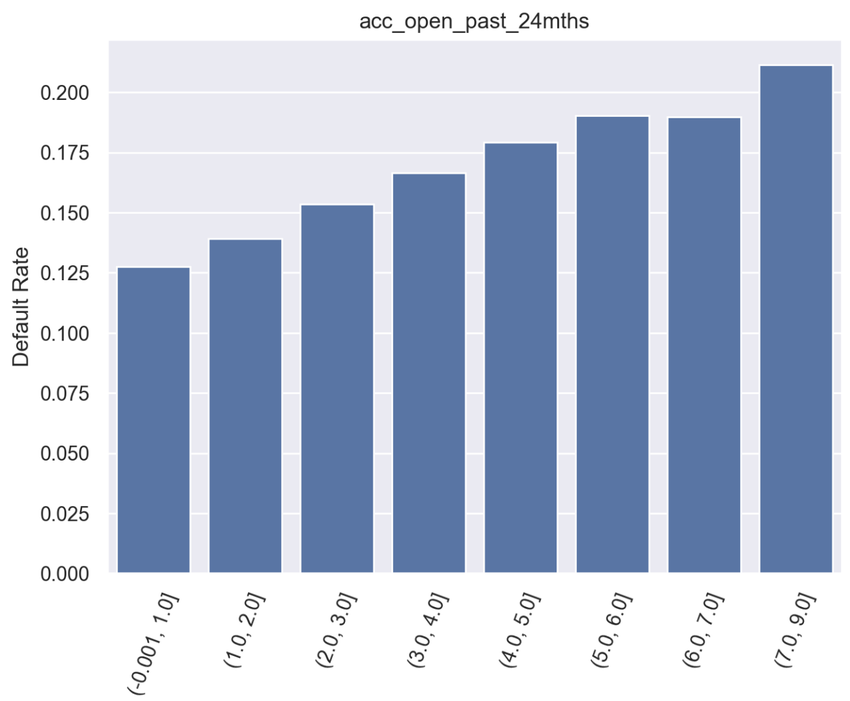


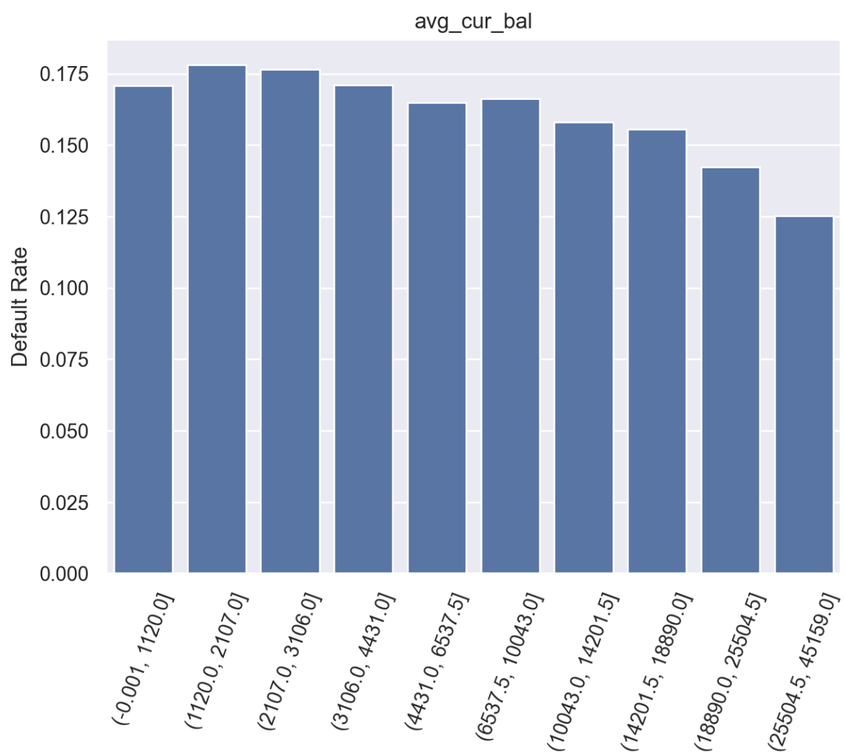


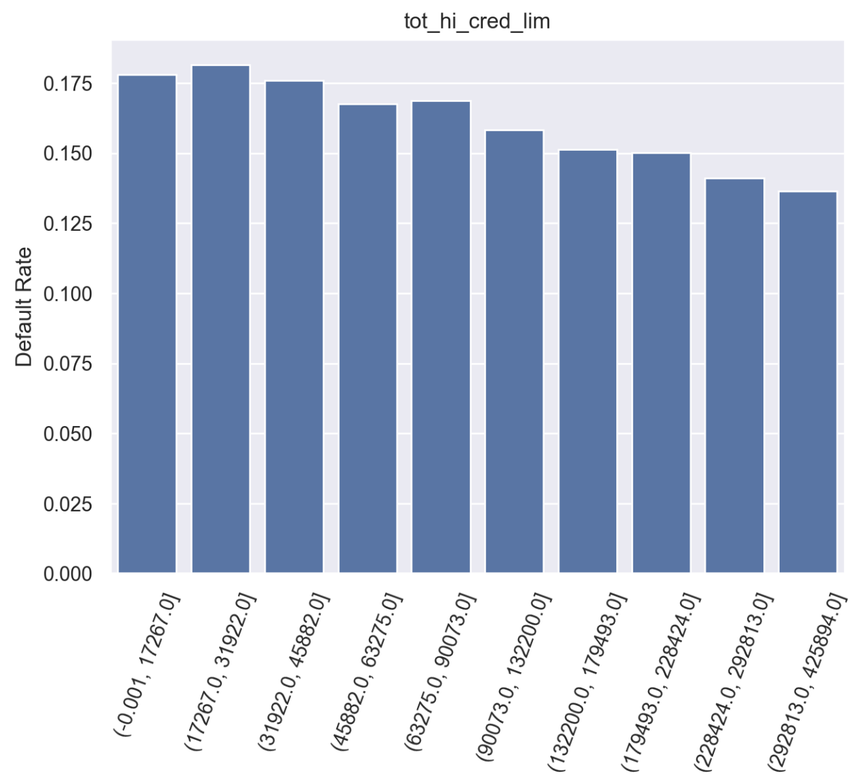




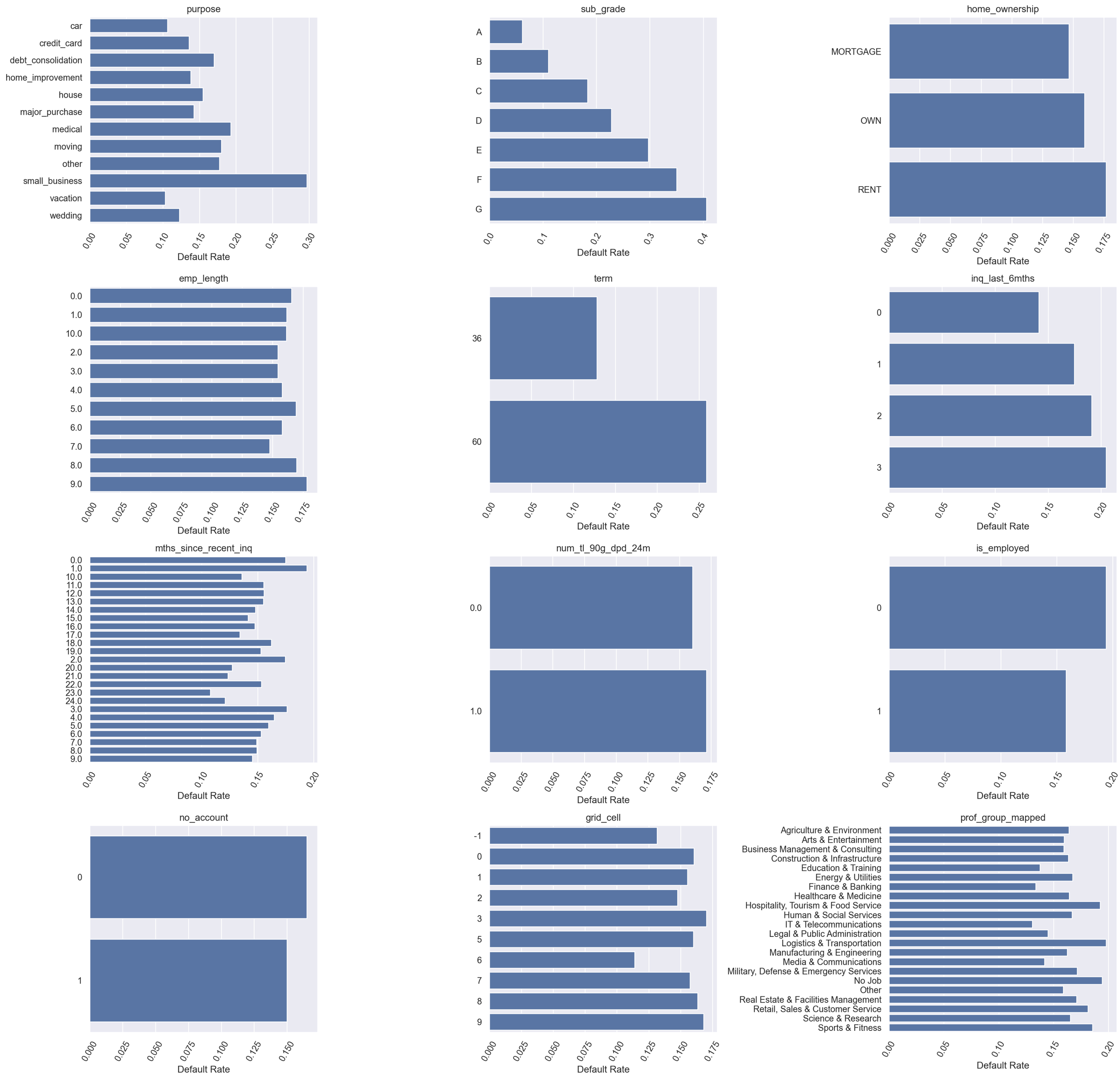


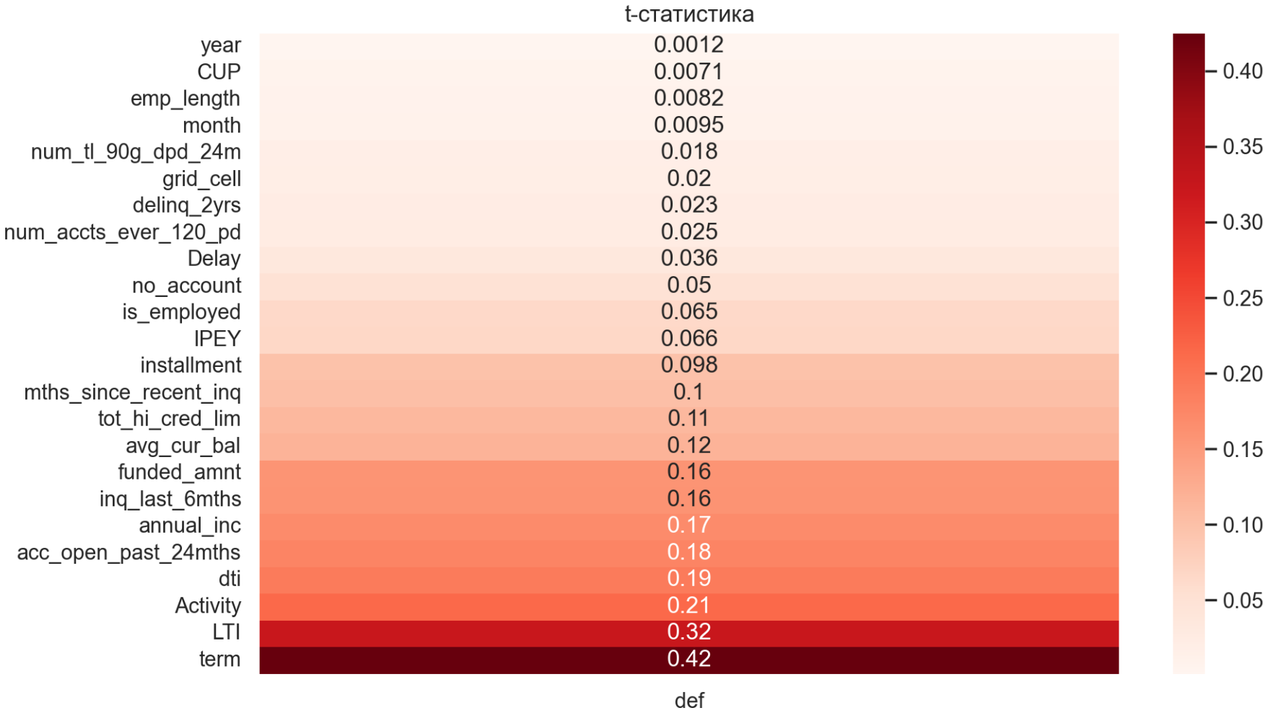


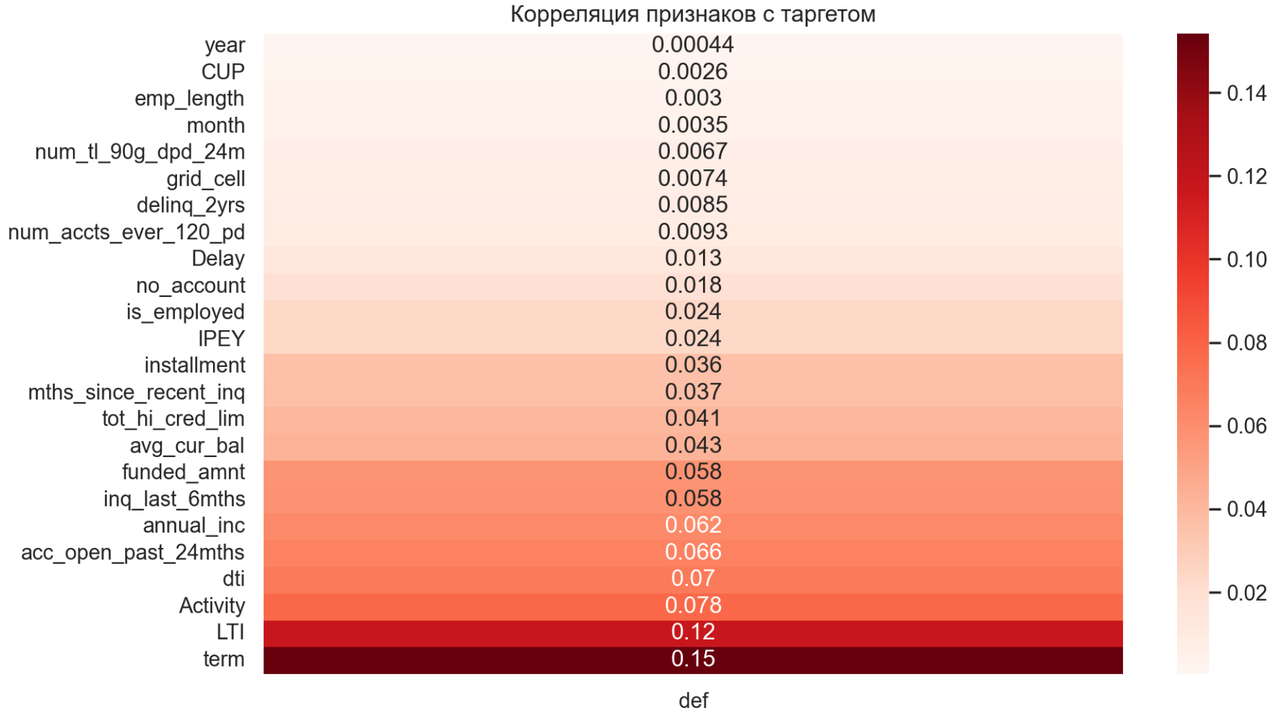




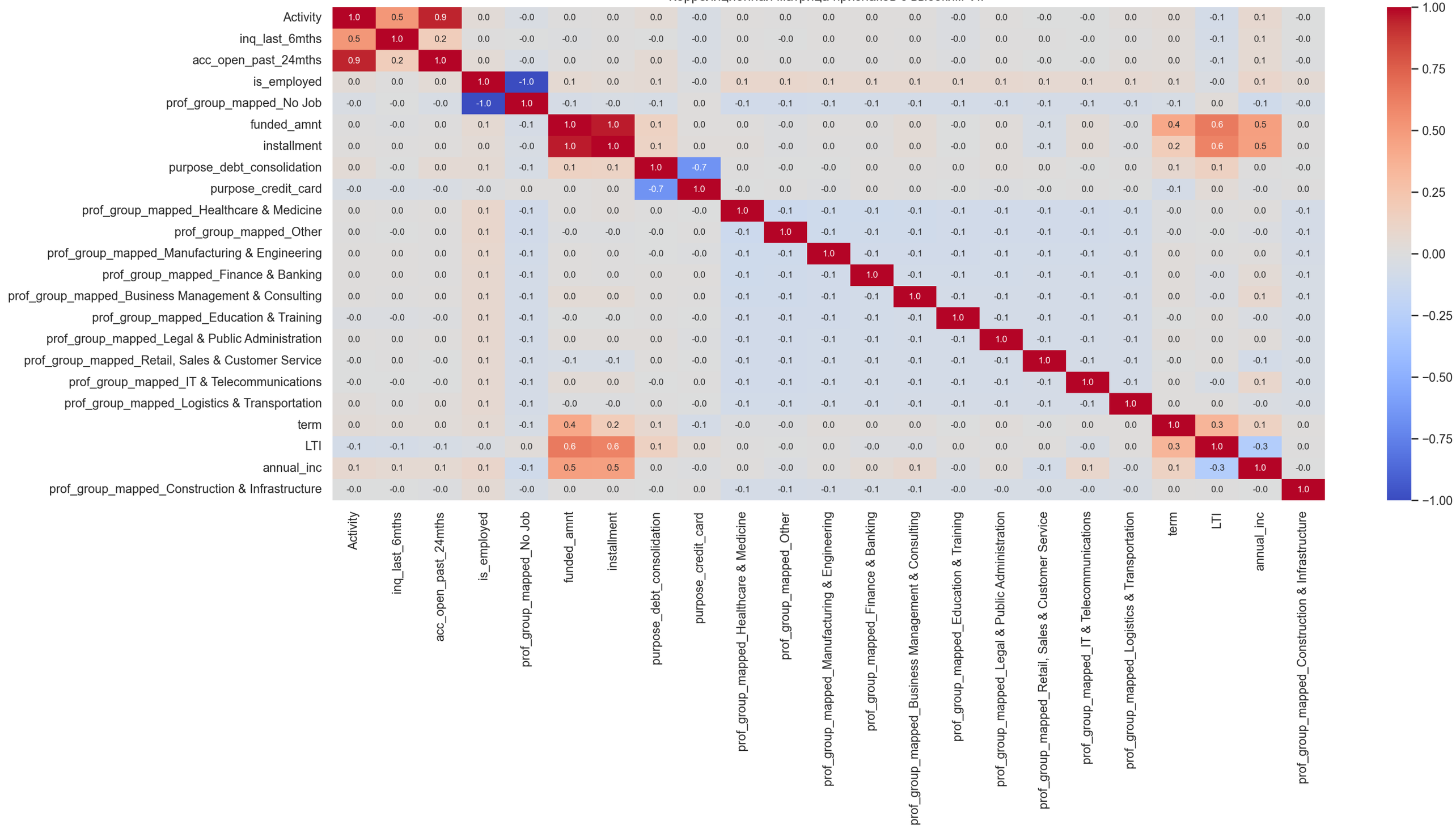
Зависимость Default Rate от категориальных признаков

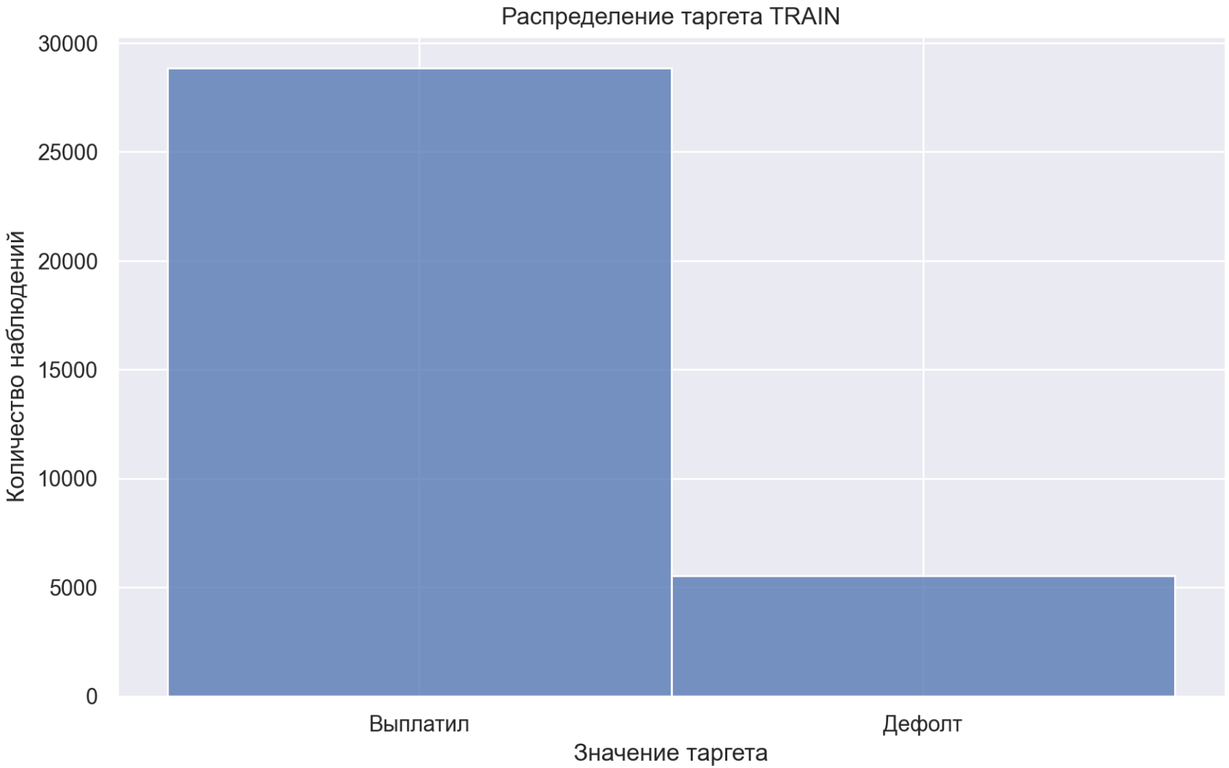


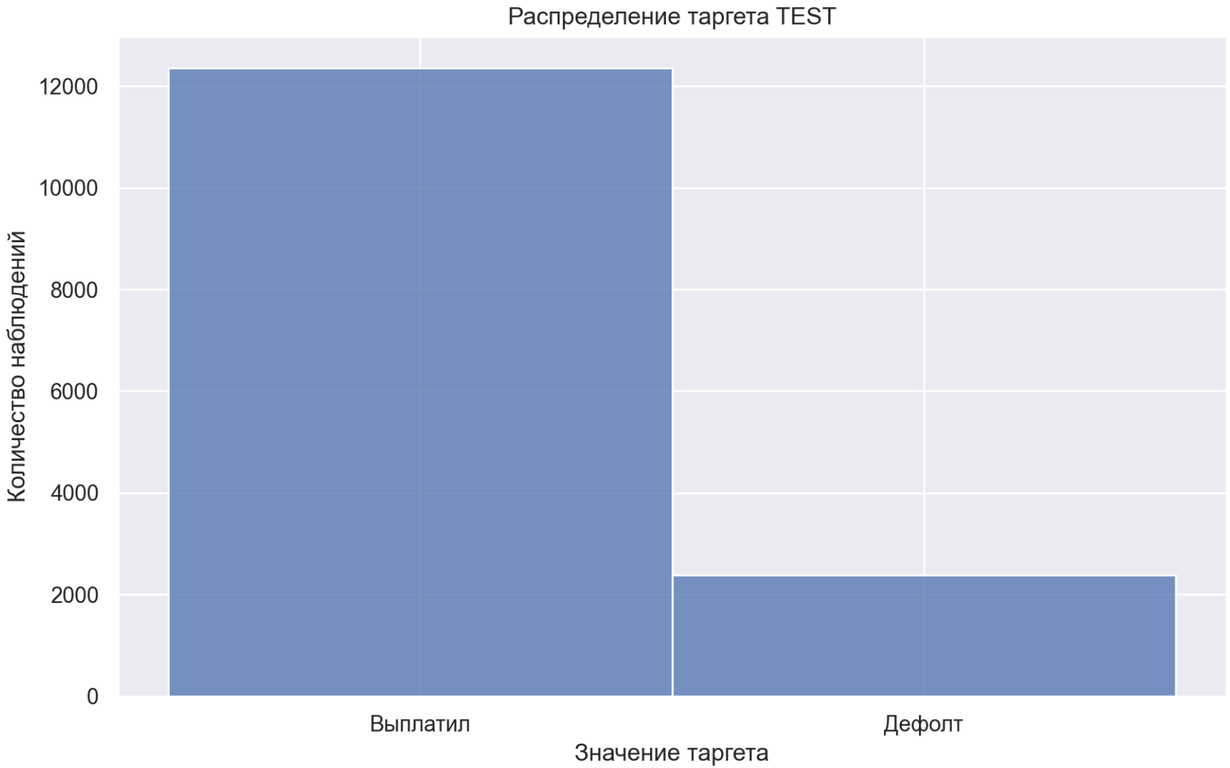

t-статистика
year 0.0012
CUP 0.0071
emp_length 0.0082
month 0.0095
num_tl_90g_dpd_24m 0.018
grid_cell 0.02
delinq_2yrs 0.023
num_accts_ever_120_pd 0.025
Delay 0.036
no_account 0.05
is_employed 0.065
IPEY 0.066
installment 0.098
mths_since_recent_inq 0.1
tot_hi_cred_lim 0.11
avg_cur_bal 0.12
funded_amnt 0.16
inq_last_6mths 0.16
annual_inc 0.17
acc_open_past_24mths 0.18
dti 0.19
Activity 0.21
LTI 0.32
term 0.42
def
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05

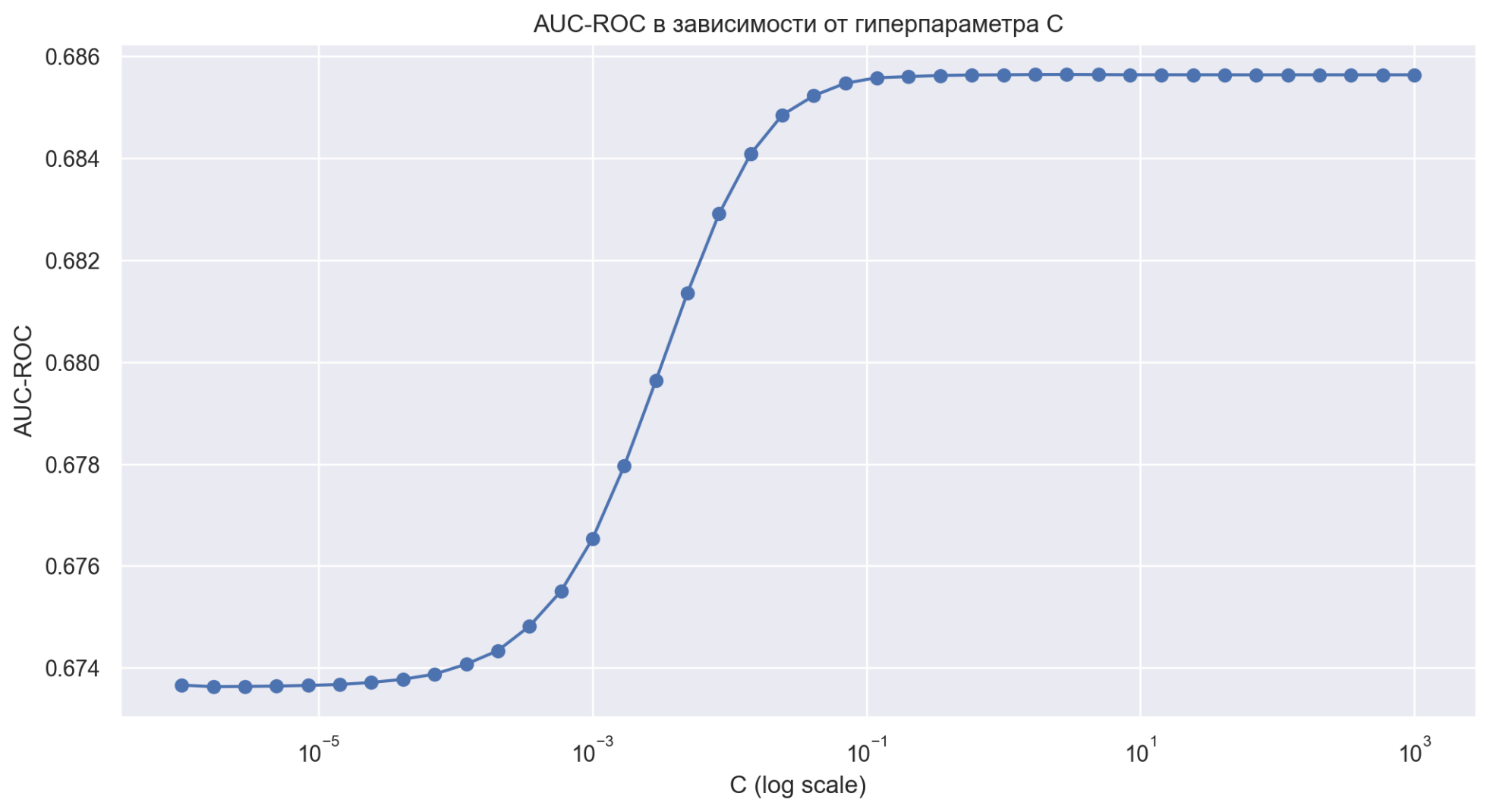

Корреляция признаков с таргетом
year 0.00044
CUP 0.0026
emp_length 0.003
month 0.0035
num_tl_90g_dpd_24m 0.0067
grid_cell 0.0074
delinq_2yrs 0.0085
num_accts_ever_120_pd 0.0093
Delay 0.013
no_account 0.018
is_employed 0.024
IPEY 0.024
installment 0.036
mths_since_recent_inq 0.037
tot_hi_cred_lim 0.041
avg_cur_bal 0.043
funded_amnt 0.058
inq_last_6mths 0.058
annual_inc 0.062
acc_open_past_24mths 0.066
dti 0.07
Activity 0.078
LTI 0.12
term 0.15
def
0.14
0.12
0.10
0.08
0.06
0.04
0.02

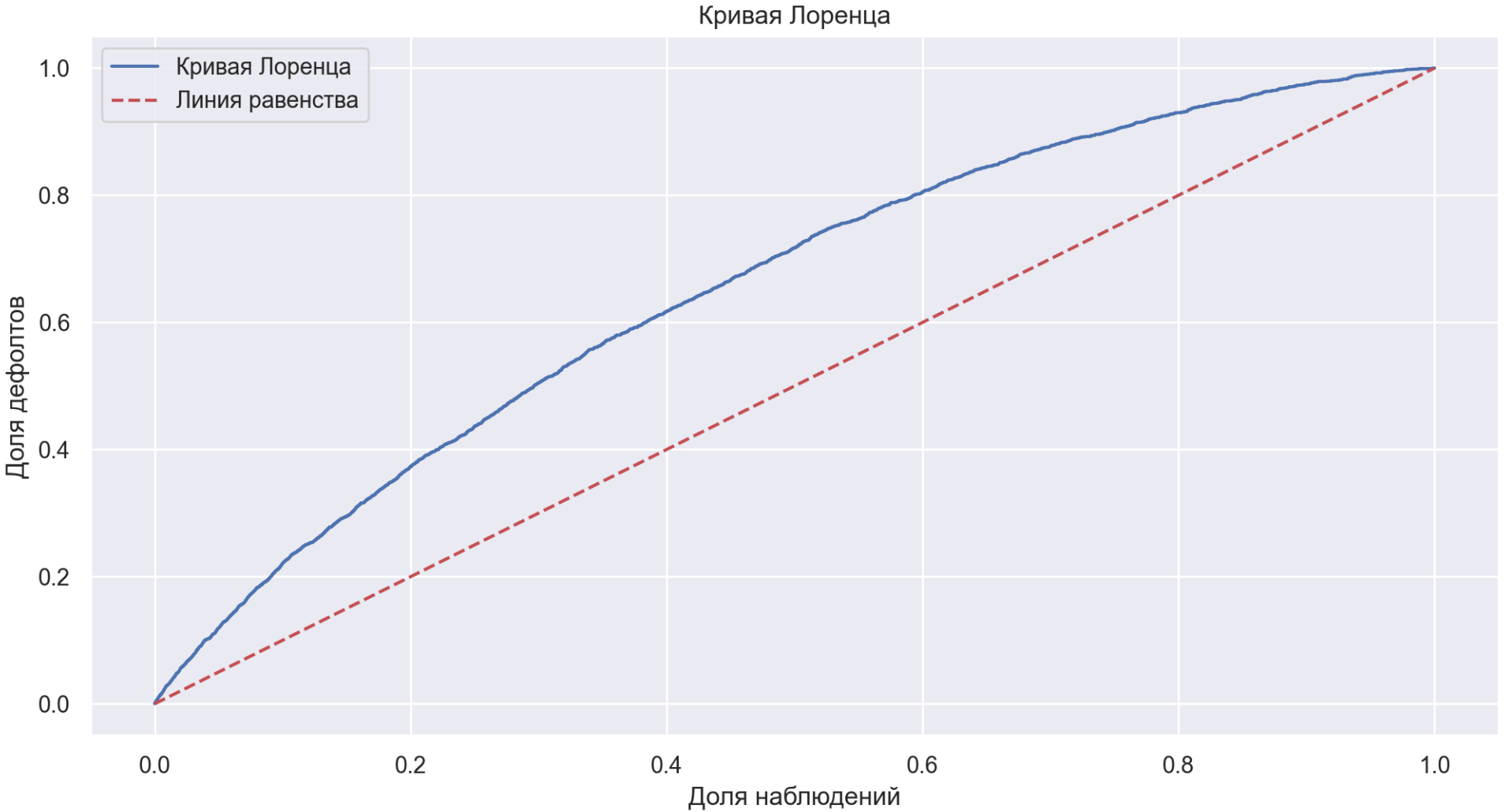
Корреляционная матрица признаков с высоким VIF

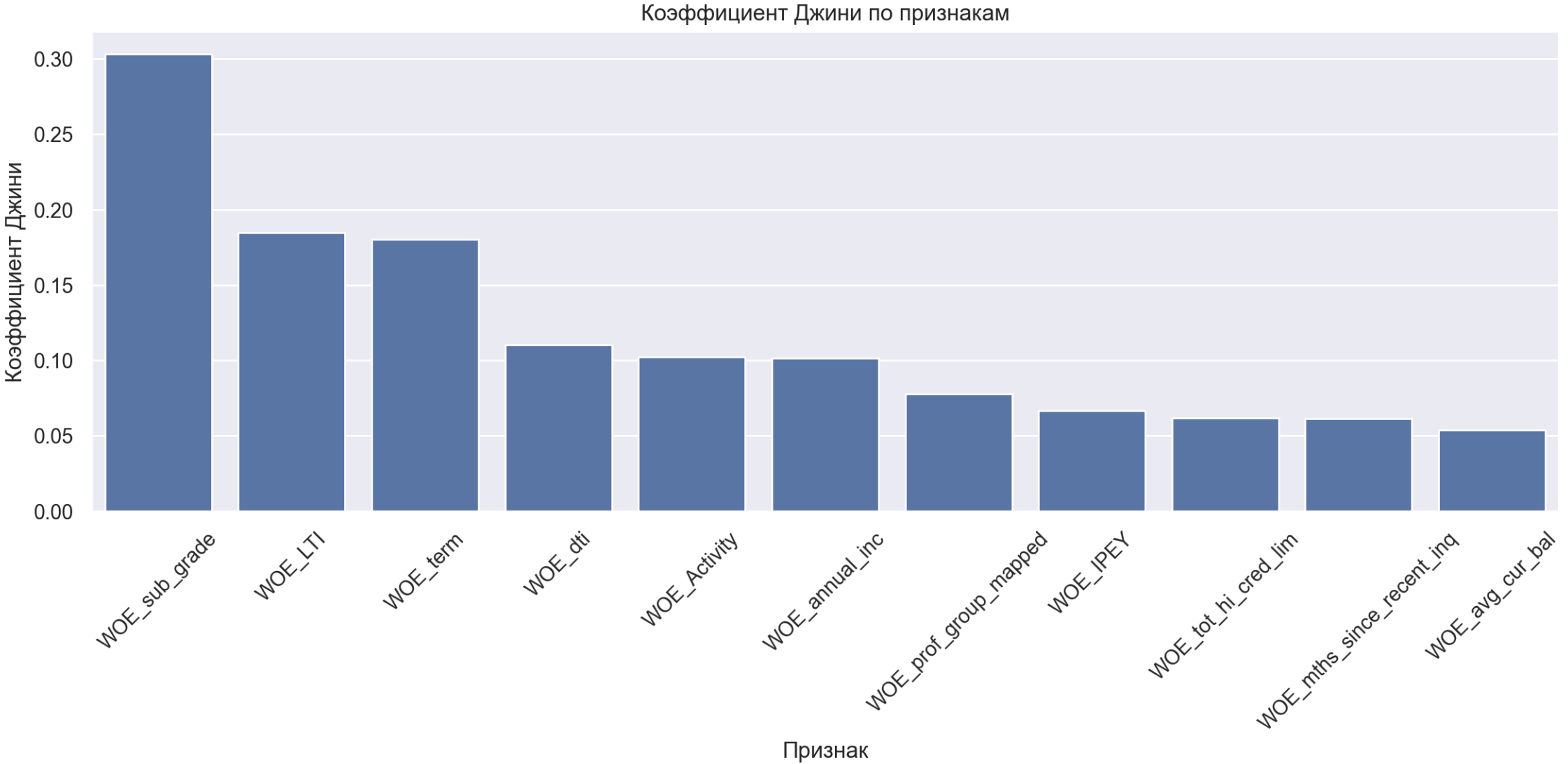


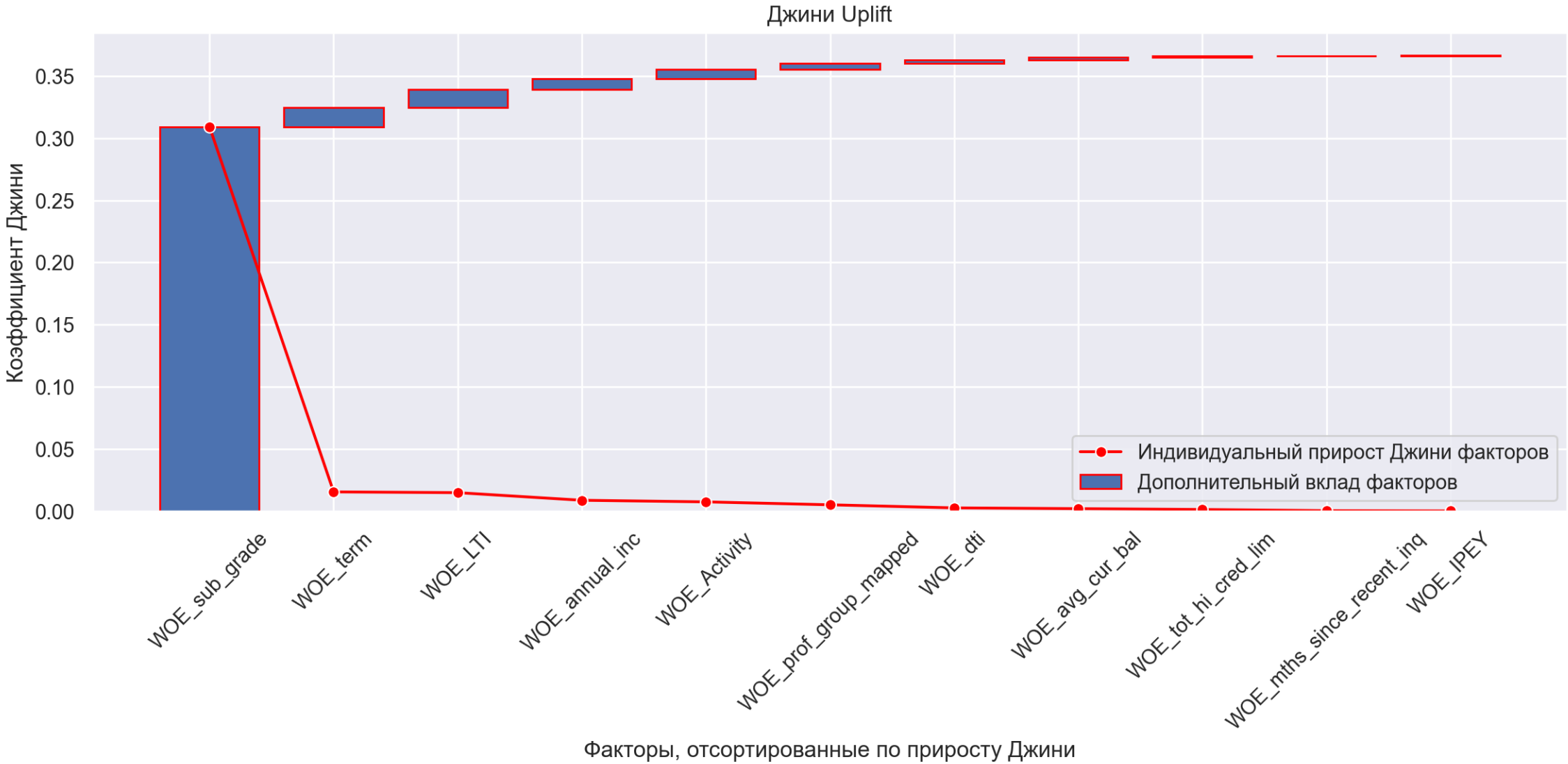

Распределение таргета TRAIN
30000
25000
20000
15000
10000
5000
0
Количество наблюдений
Выплатил
Дефолт
Значение таргета

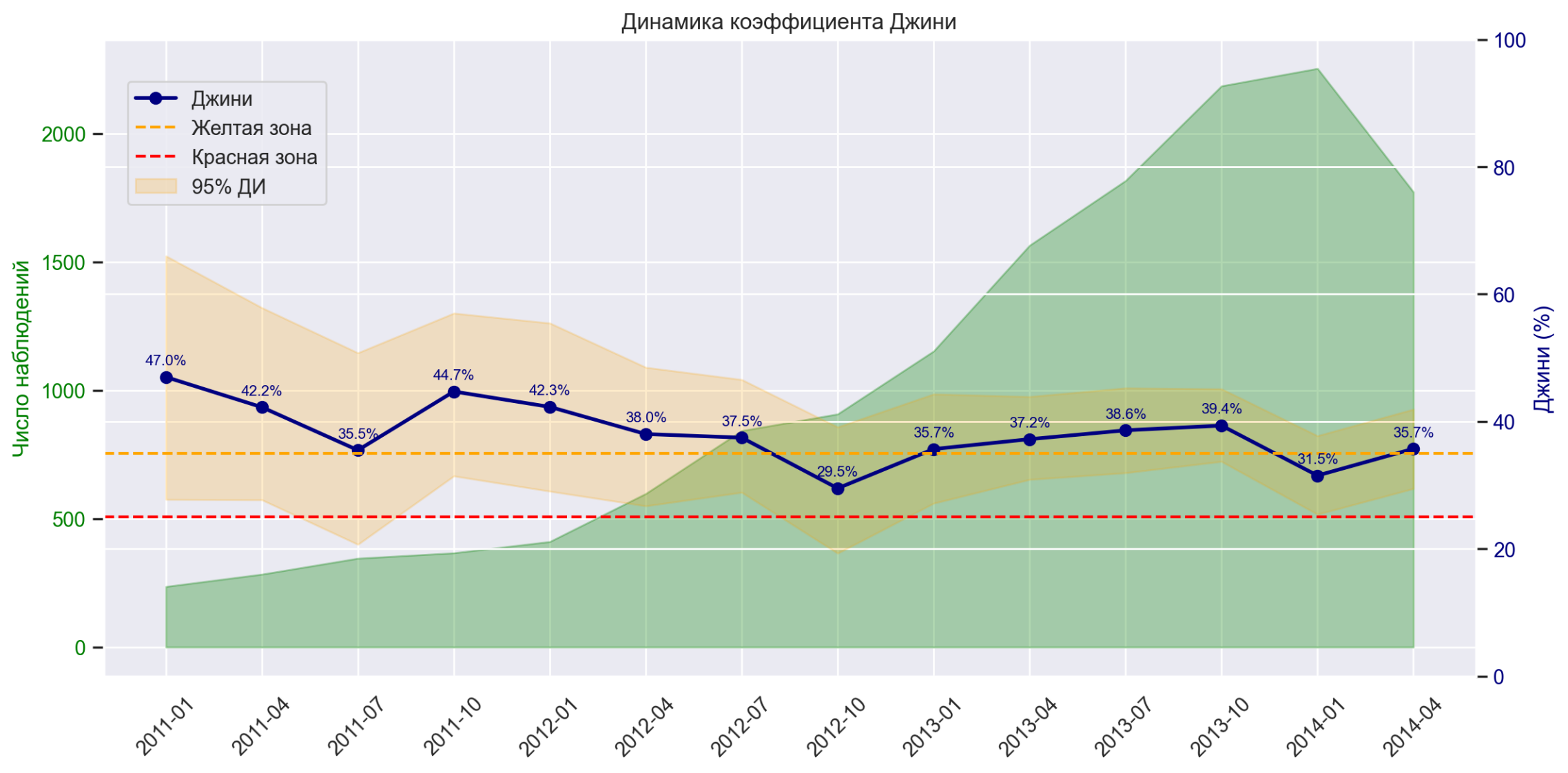

Распределение таргета TEST
12000
10000
8000
6000
4000
2000
0
Количество наблюдений
Выплатил
Дефолт
Значение таргета

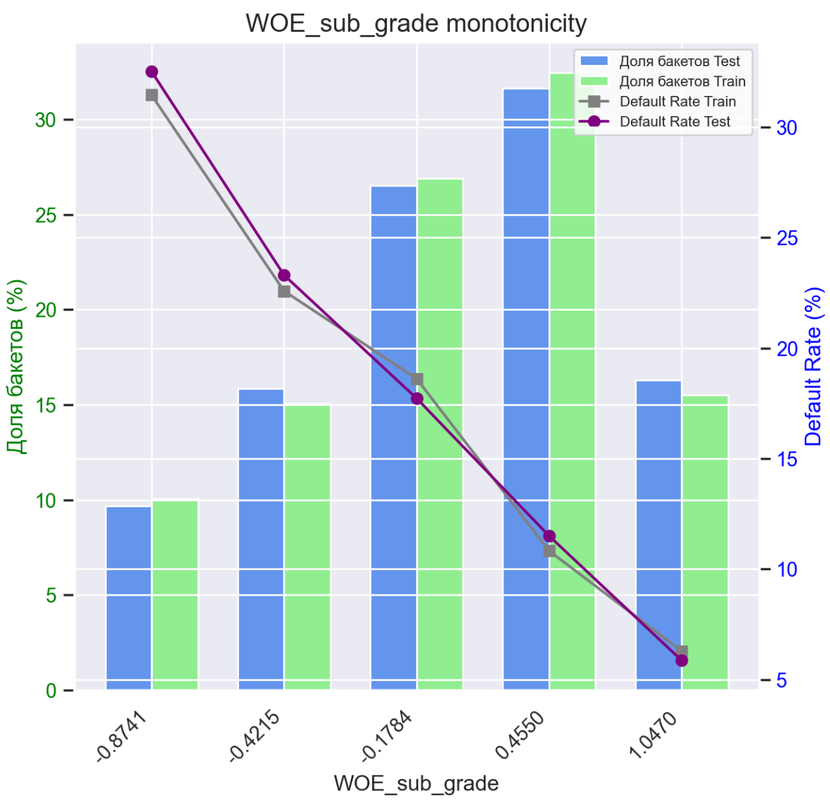

AUC-ROC в зависимости от гиперпараметра C
0.686
0.684
0.682
0.680
0.678
0.676
0.674
AUC-ROC
10^{-5}
10^{-3}
10^{-1}
10^{1}
10^{3}
C (log scale)

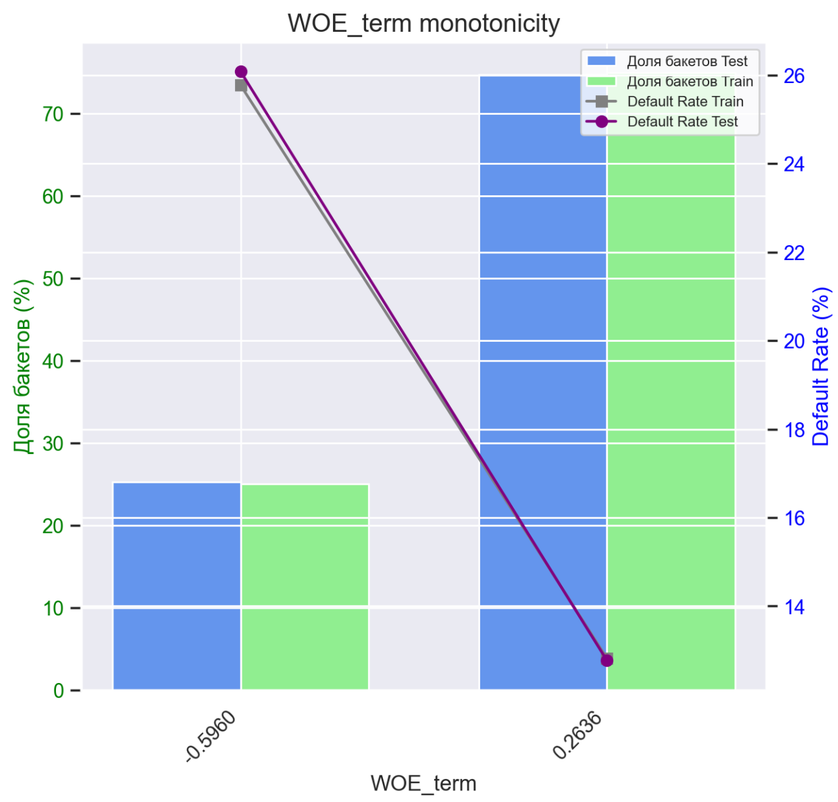

Кривая Лоренца
Кривая Лоренца
Линия равенства
Доля дефолтов
Доля наблюдений
0.0
0.2
0.4
0.6
0.8
1.0

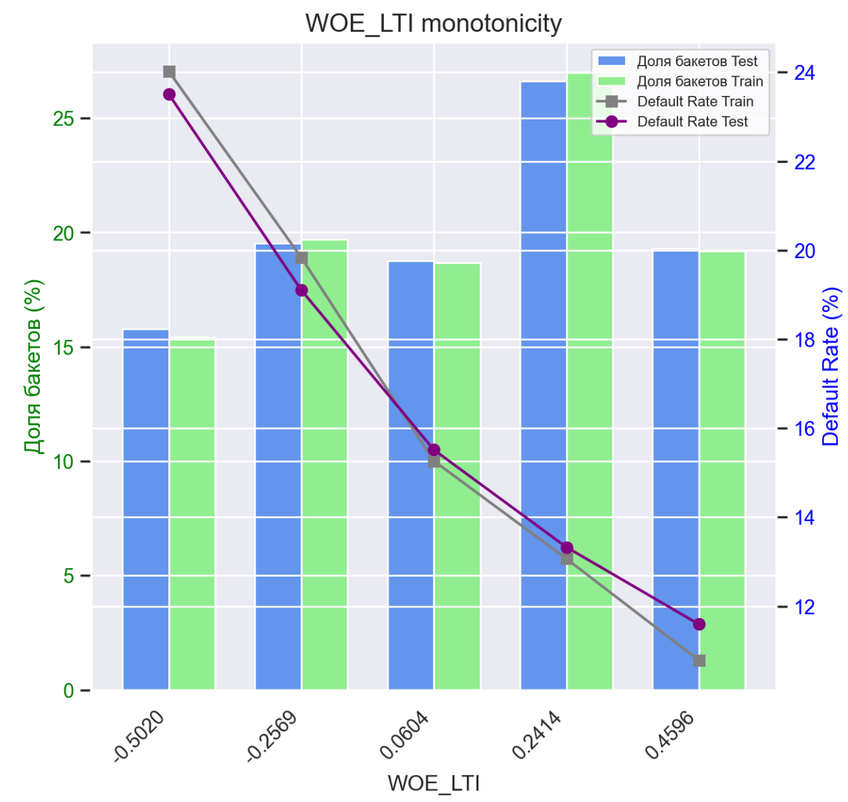

Коэффициент Джини по признакам
Коэффициент Джини
0.30
0.25
0.20
0.15
0.10
0.05
0.00
WOE_sub_grade
WOE_LTI
WOE_term
WOE_dti
WOE_Activity
WOE_annual_inc
WOE_prof_group_mapped
WOE_IPEY
WOE_tot_hi_cred_lim
WOE_mths_since_recent_inq
WOE_avg_cur_bal
Признак

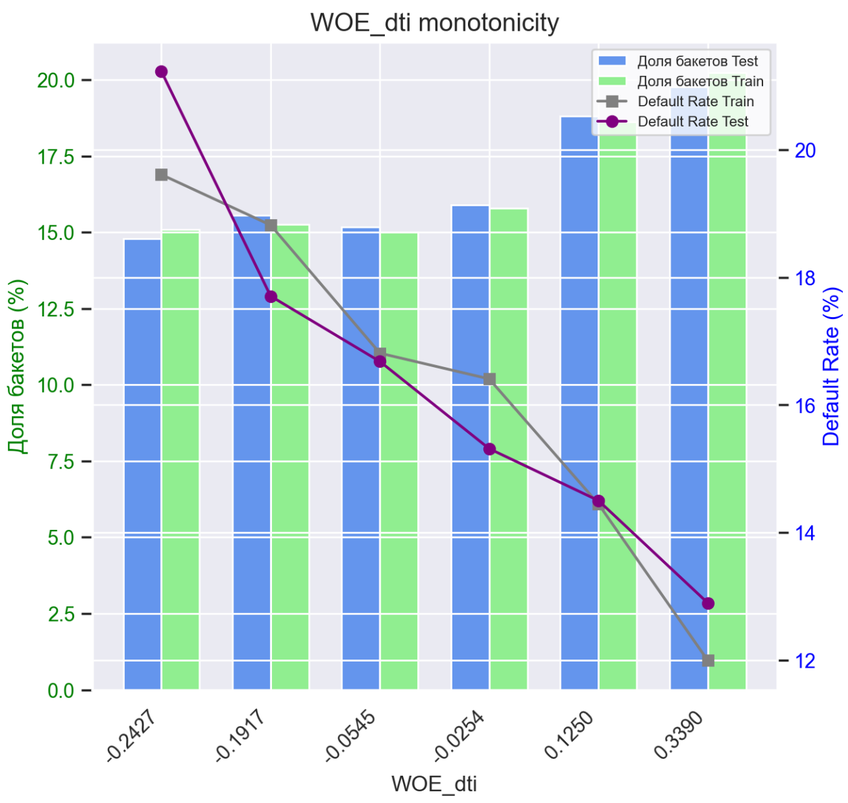

Джини Uplift
Коэффициент Джини
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Индивидуальный прирост Джини факторов
Дополнительный вклад факторов
WOE_sub_grade
WOE_term
WOE_LTI
WOE_annual_inc
WOE_Activity
WOE_prof_group_mapped
WOE_dti
WOE_avg_cur_bal
WOE_tot_hi_cred_lim
WOE_mths_since_recent_inq
WOE_IPEY
Факторы, отсортированные по приросту Джини

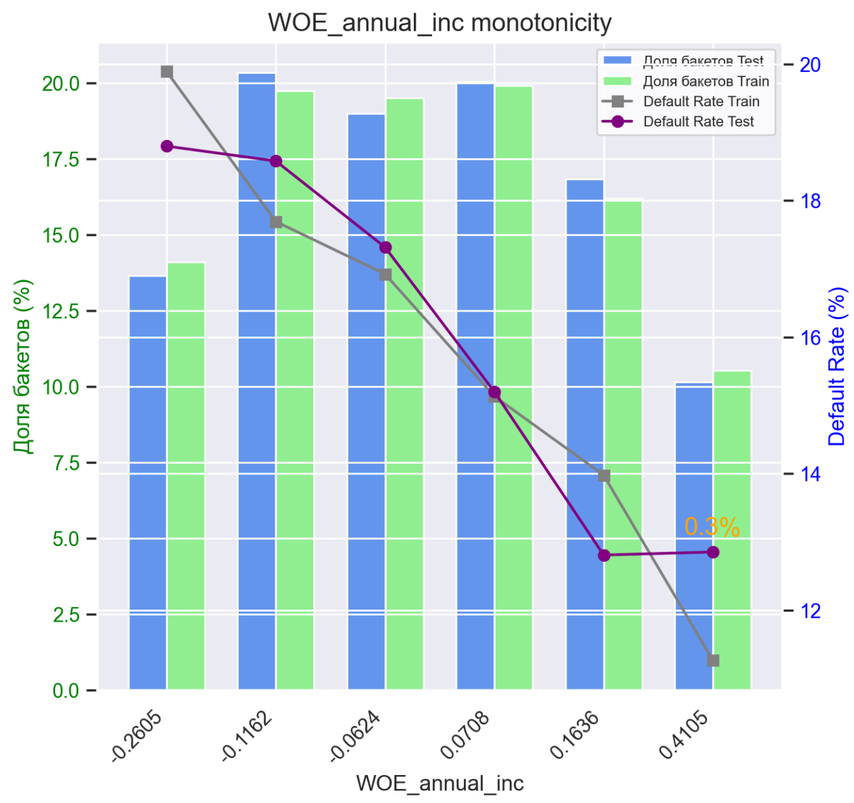

Динамика коэффициента Джини
Джини
Желтая зона
Красная зона
95% ДИ
Число наблюдений
Джини (%)
0
500
1000
1500
2000
0
20
40
60
80
100
47.0%
42.2%
35.5%
44.7%
42.3%
38.0%
37.5%
29.5%
35.7%
37.2%
38.6%
39.4%
31.5%
35.7%
2011-01
2011-04
2011-07
2011-10
2012-01
2012-04
2012-07
2012-10
2013-01
2013-04
2013-07
2013-10
2014-01
2014-04

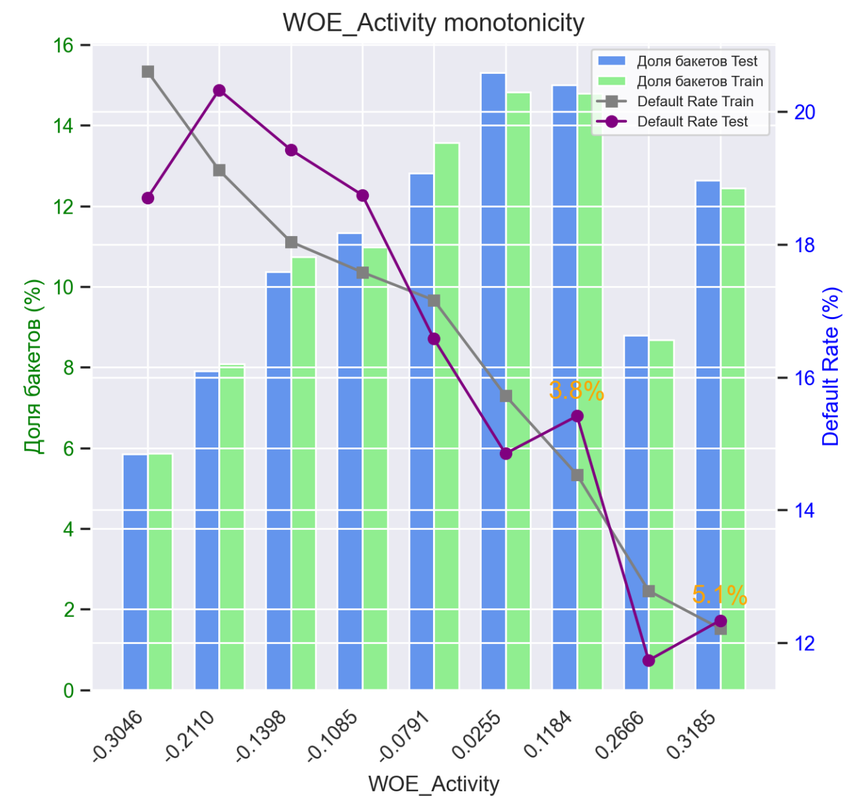

WOE_sub_grade monotonicity
Доля бакетов Test
Доля бакетов Train
Default Rate Train
Default Rate Test
Доля бакетов (%)
Default Rate (%)
-0.8741
-0.4215
-0.1784
0.4550
1.0470
WOE_sub_grade

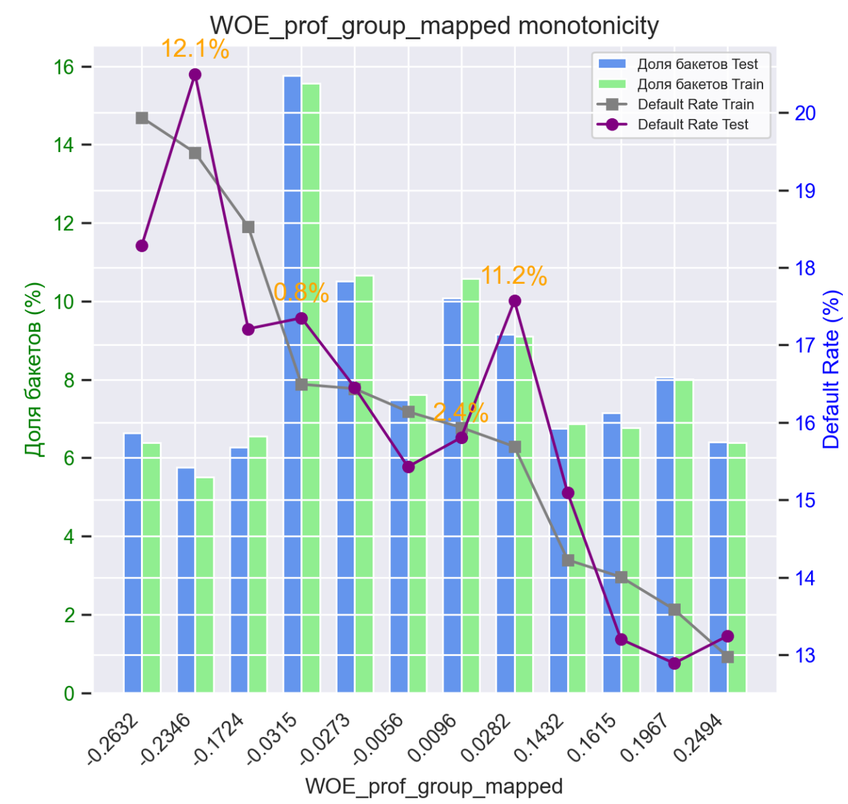

WOE_term monotonicity
Доля бакетов Test
Доля бакетов Train
Default Rate Train
Default Rate Test
Доля бакетов (%)
Default Rate (%)
-0.5960
0.2636
WOE_term

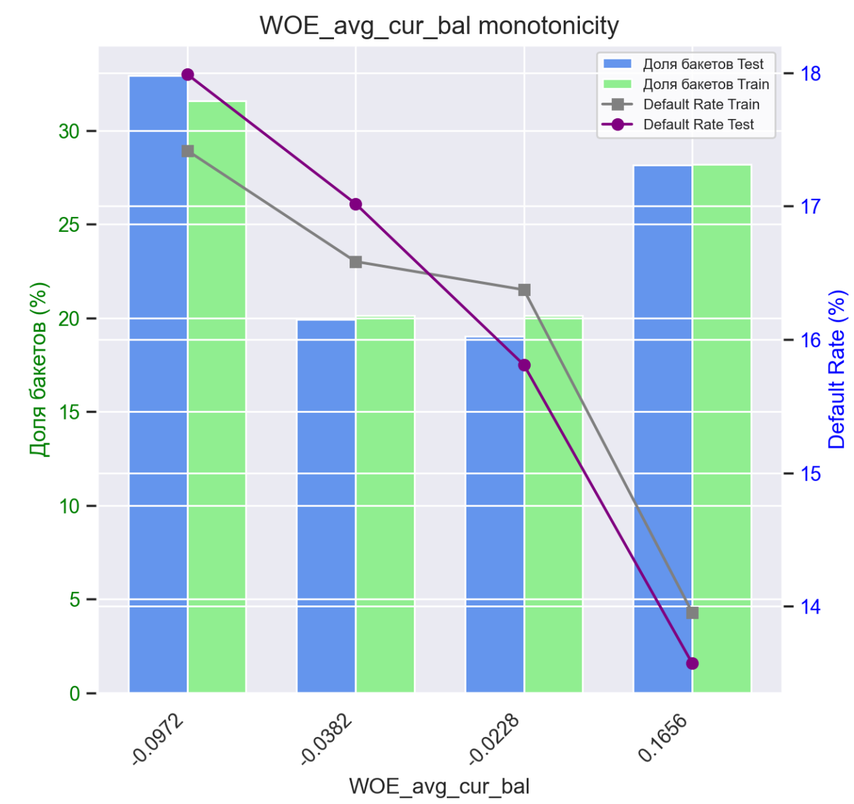

WOE_LTI monotonicity
Доля бакетов Test
Доля бакетов Train
Default Rate Train
Default Rate Test
Доля бакетов (%)
Default Rate (%)
-0.5020
-0.2569
0.0604
0.2414
0.4596
WOE_LTI

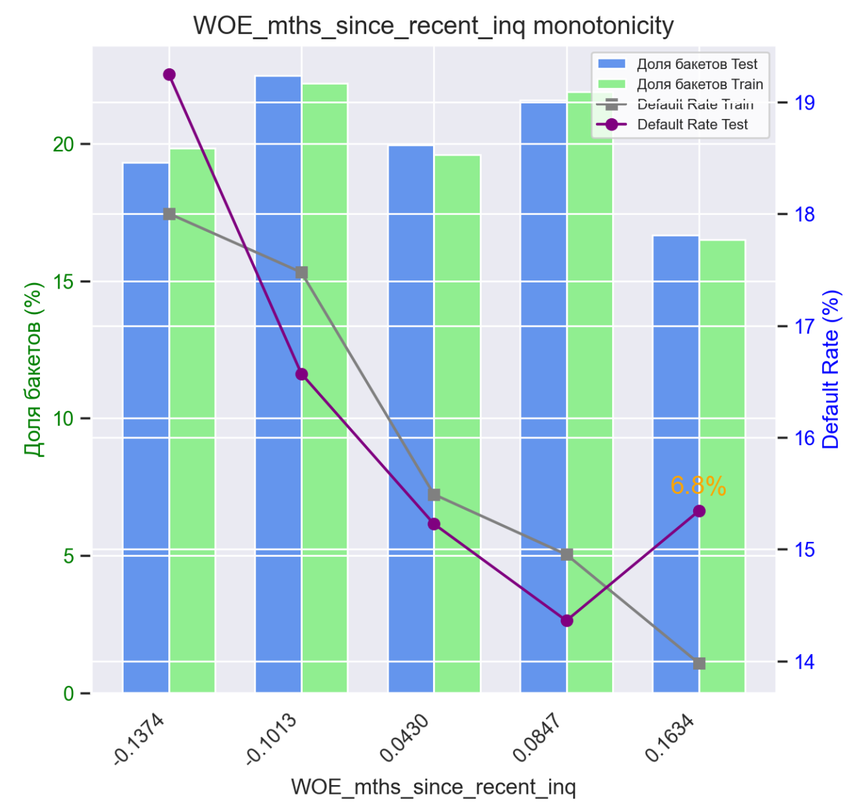

WOE_dti monotonicity
Доля бакетов Test
Доля бакетов Train
Default Rate Train
Default Rate Test
Доля бакетов (%)
Default Rate (%)
-0.2427
-0.1917
-0.0545
-0.0254
0.1250
0.3390
WOE_dti

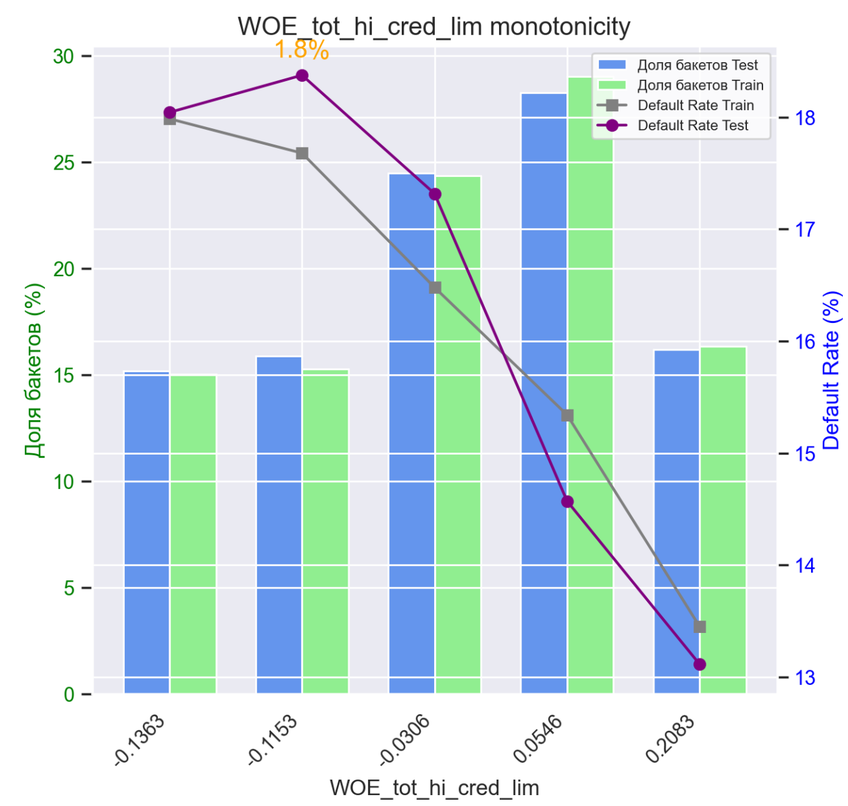

WOE_annual_inc monotonicity
Доля бакетов Test
Доля бакетов Train
Default Rate Train
Default Rate Test
0.3%
Доля бакетов (%)
Default Rate (%)
-0.2605
-0.1162
-0.0624
0.0708
0.1636
0.4105
WOE_annual_inc


WOE_Activity monotonicity
Доля бакетов Test
Доля бакетов Train
Default Rate Train
Default Rate Test
3.8%
5.1%
Доля бакетов (%)
Default Rate (%)
-0.3046
-0.2110
-0.1398
-0.1085
-0.0791
0.0255
0.1184
0.2666
0.3185
WOE_Activity

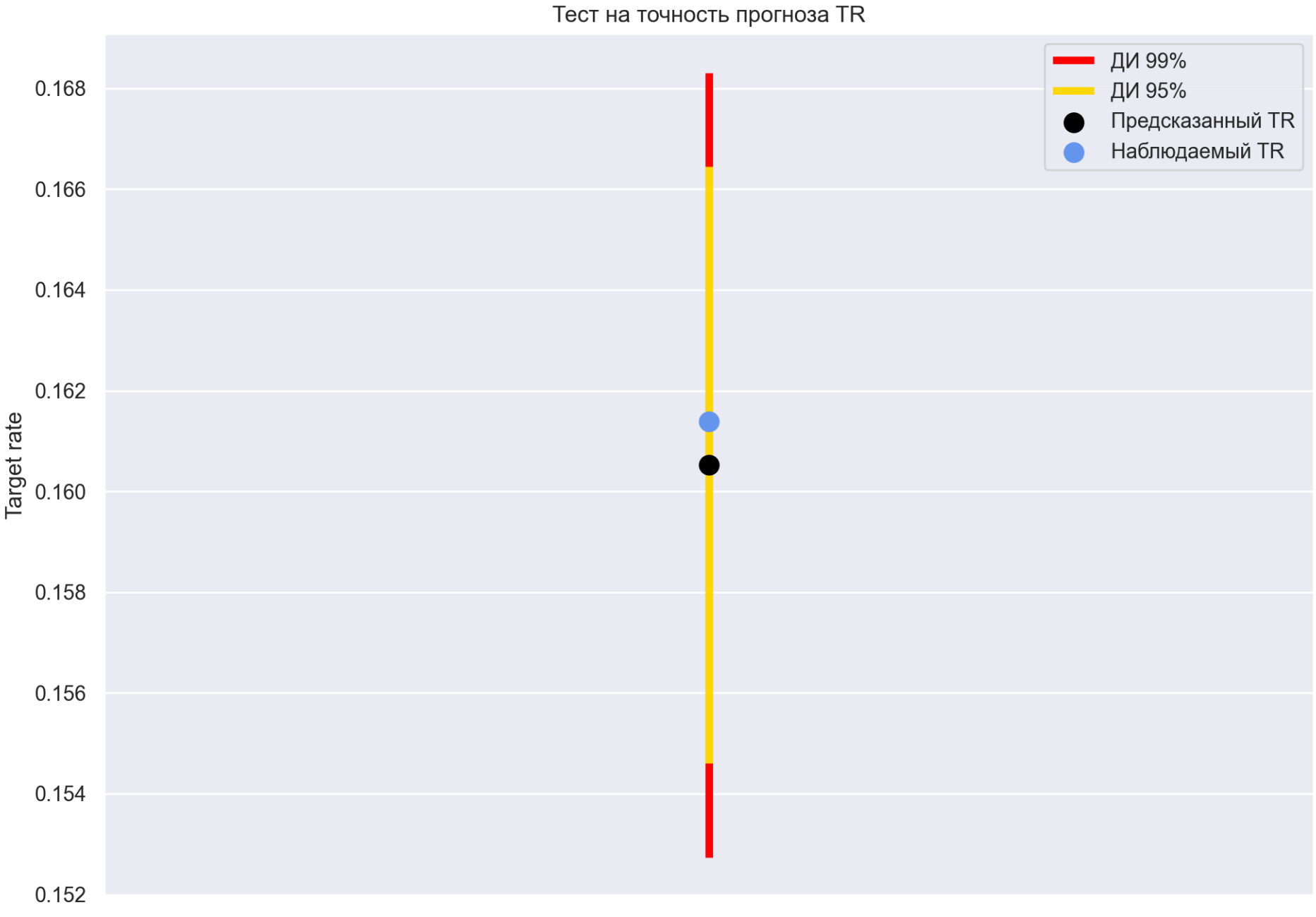

WOE_prof_group_mapped monotonicity
Доля бакетов Test
Доля бакетов Train
Default Rate Train
Default Rate Test
12.1%
0.8%
2.4%
11.2%
Доля бакетов (%)
Default Rate (%)
-0.2632
-0.2346
-0.1724
-0.0315
-0.0273
-0.0056
0.0096
0.0282
0.1432
0.1615
0.1967
0.2494
WOE_prof_group_mapped

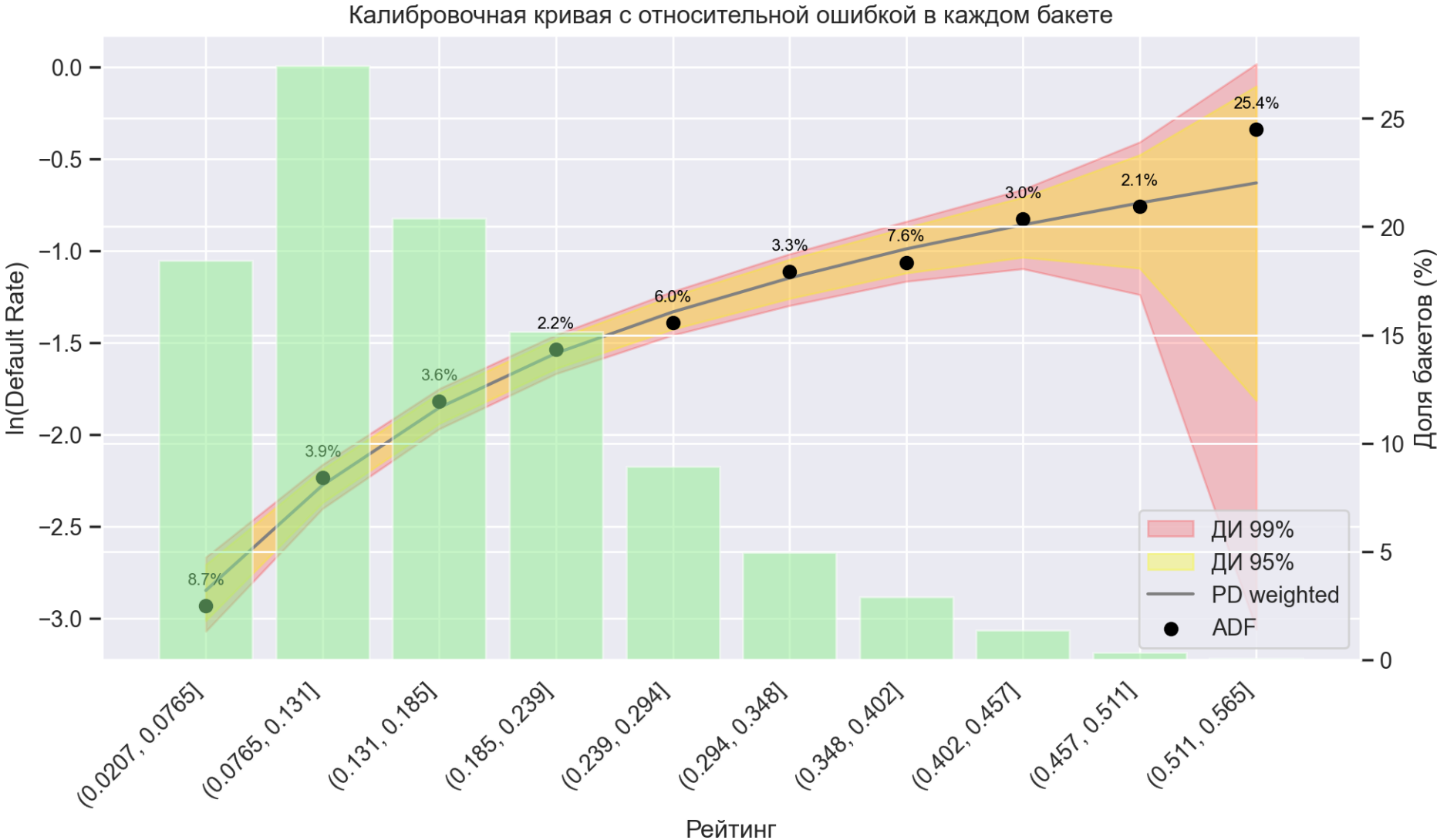

WOE_avg_cur_bal monotonicity
Доля бакетов Test
Доля бакетов Train
Default Rate Train
Default Rate Test
Доля бакетов (%)
Default Rate (%)
-0.0972
-0.0382
-0.0228
0.1656
WOE_avg_cur_bal

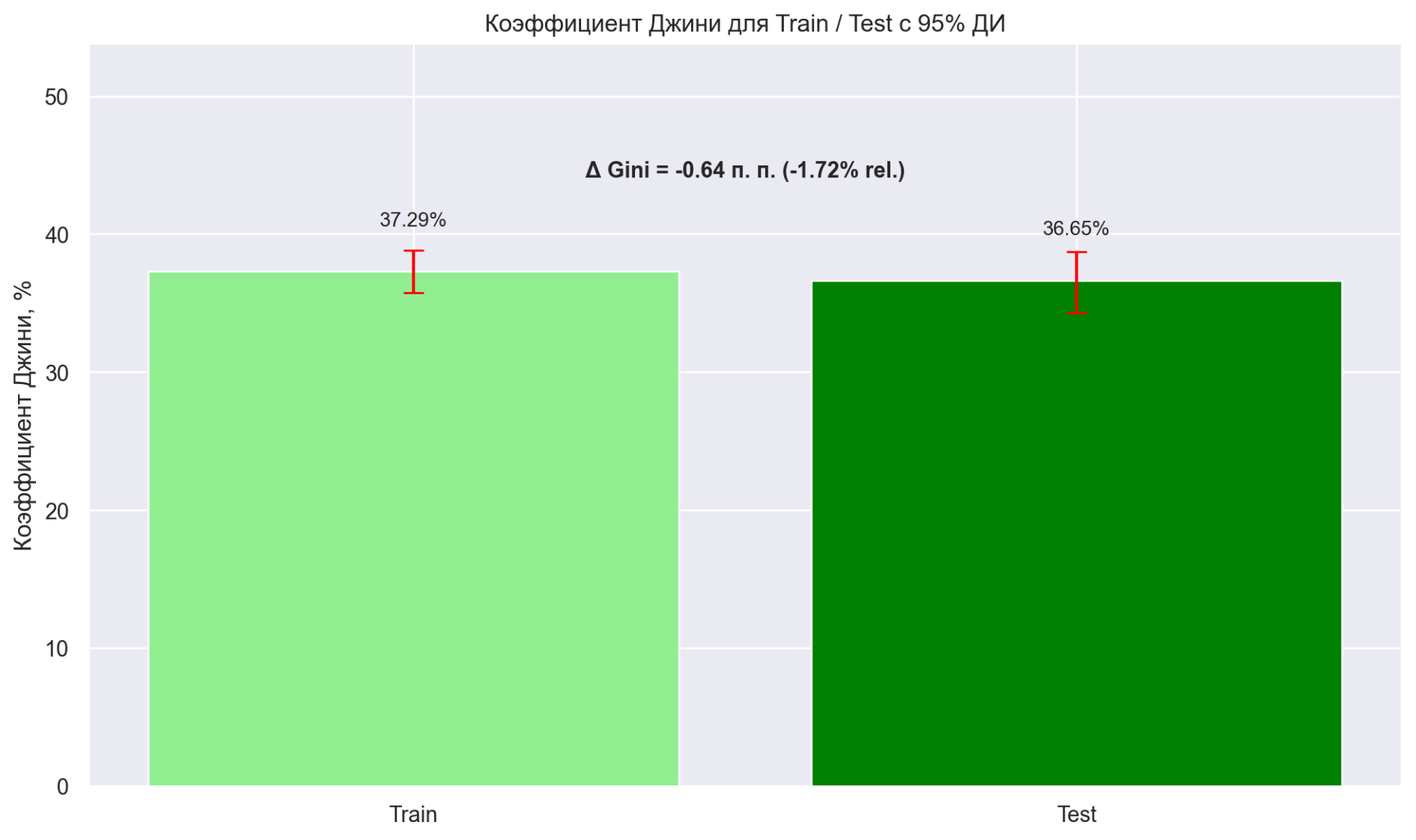

WOE_mths_since_recent_inq monotonicity
Доля бакетов Test
Доля бакетов Train
Default Rate Train
Default Rate Test
6.8%
Доля бакетов (%)
Default Rate (%)
-0.1374
-0.1013
0.0430
0.0847
0.1634
WOE_mths_since_recent_inq

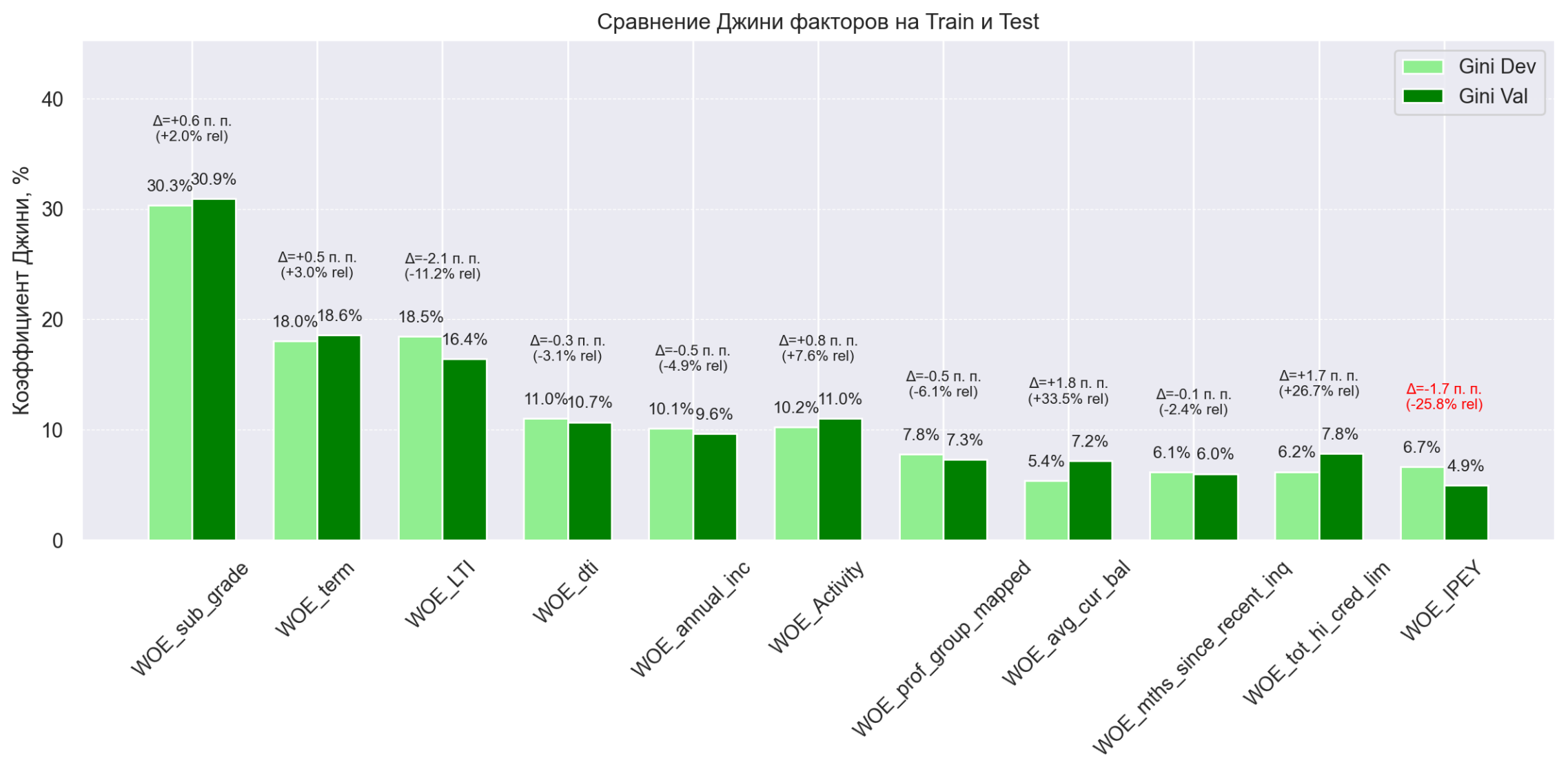

WOE_tot_hi_cred_lim monotonicity
Доля бакетов Test
Доля бакетов Train
Default Rate Train
Default Rate Test
1.8%
Доля бакетов (%)
Default Rate (%)
-0.1363
-0.1153
-0.0306
0.0546
0.2083
WOE_tot_hi_cred_lim

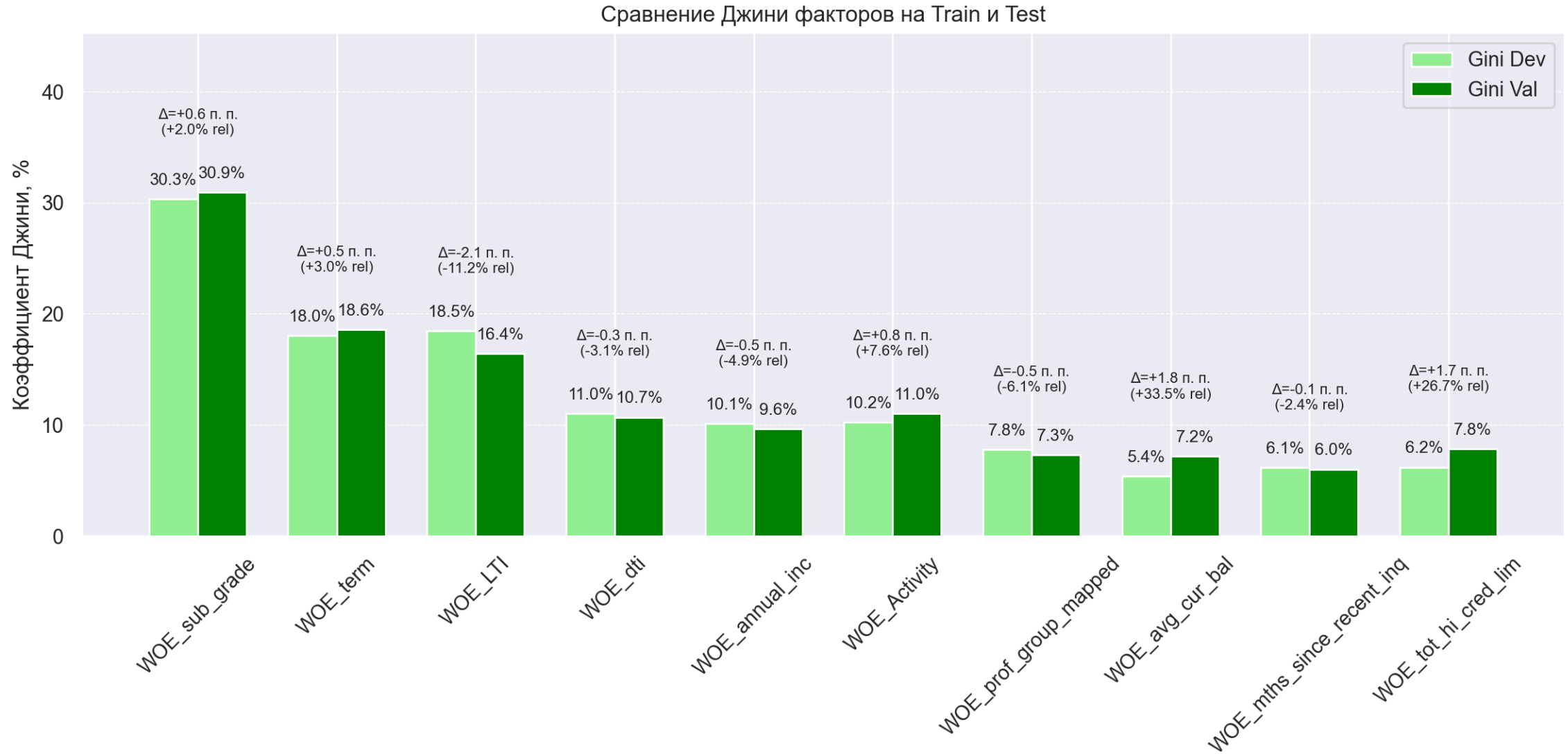
WOE_IPEY monotonicity
Доля бакетов Test
Доля бакетов Train
Default Rate Train
Default Rate Test
Доля бакетов (%)
Default Rate (%)
-0.1841
-0.0066
0.1181
WOE_IPEY

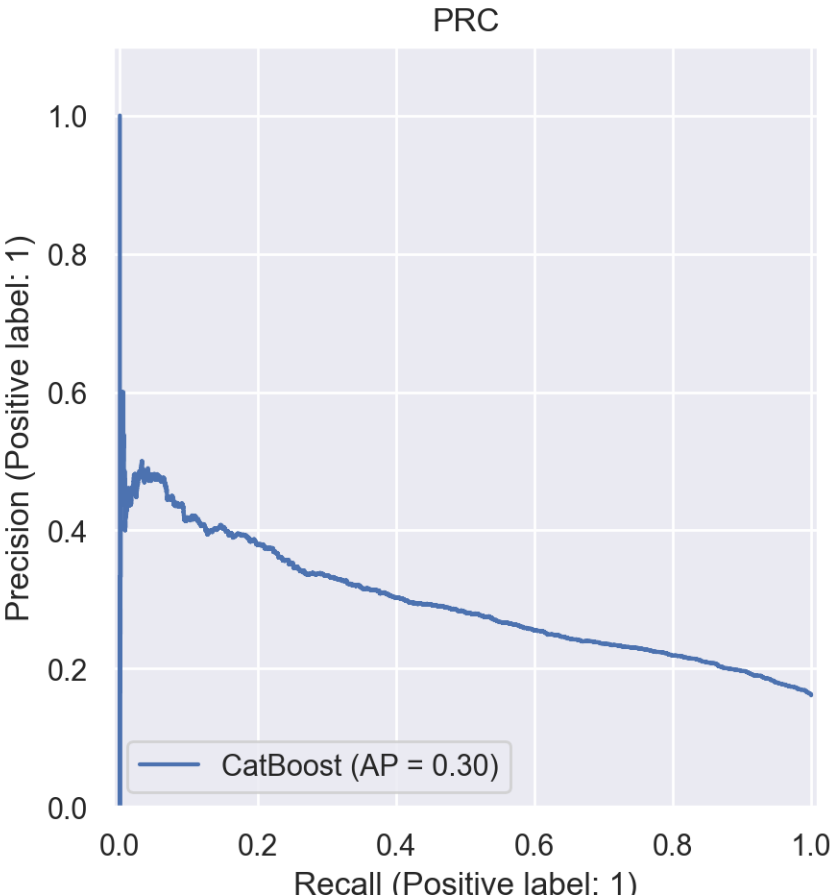

Тест на точность прогноза TR
ДИ 99%
ДИ 95%
Предсказанный TR
Наблюдаемый TR
Target rate
0.168
0.166
0.164
0.162
0.160
0.158
0.156
0.154
0.152

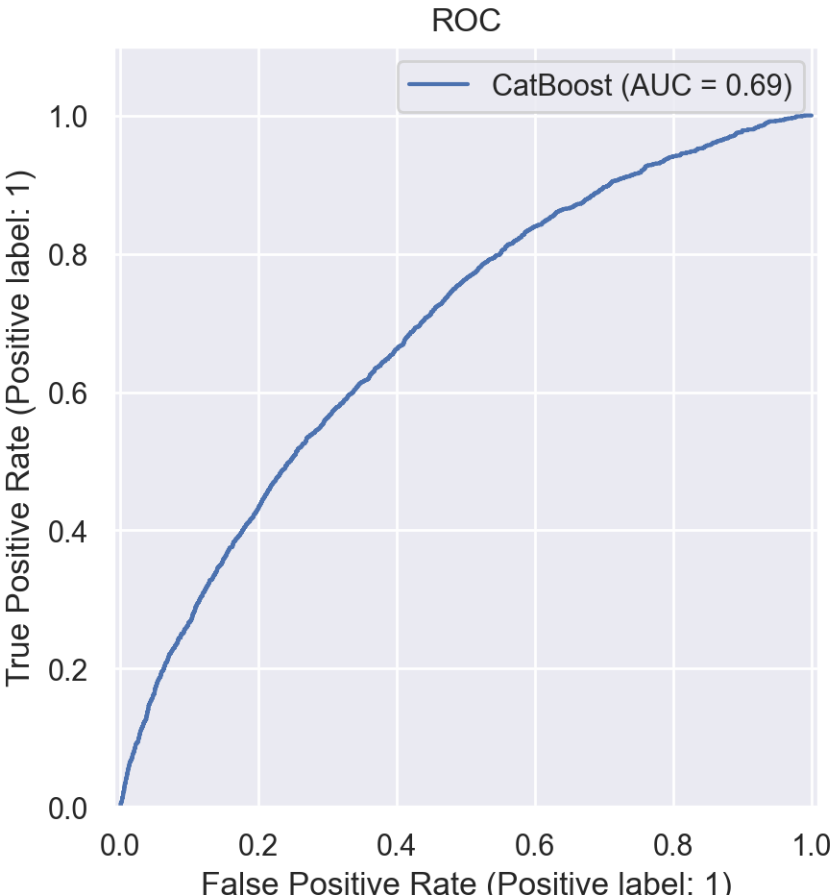

Калибровочная кривая с относительной ошибкой в каждом бакете
ln(Default Rate)
Доля бакетов (%)
Рейтинг
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
0
5
10
15
20
25
8.7%
3.9%
3.6%
2.2%
6.0%
3.3%
7.6%
3.0%
2.1%
25.4%
ДИ 99%
ДИ 95%
PD weighted
ADF
(0.0207, 0.0765]
(0.0765, 0.131]
(0.131, 0.185]
(0.185, 0.239]
(0.239, 0.294]
(0.294, 0.348]
(0.348, 0.402]
(0.402, 0.457]
(0.457, 0.511]
(0.511, 0.565]

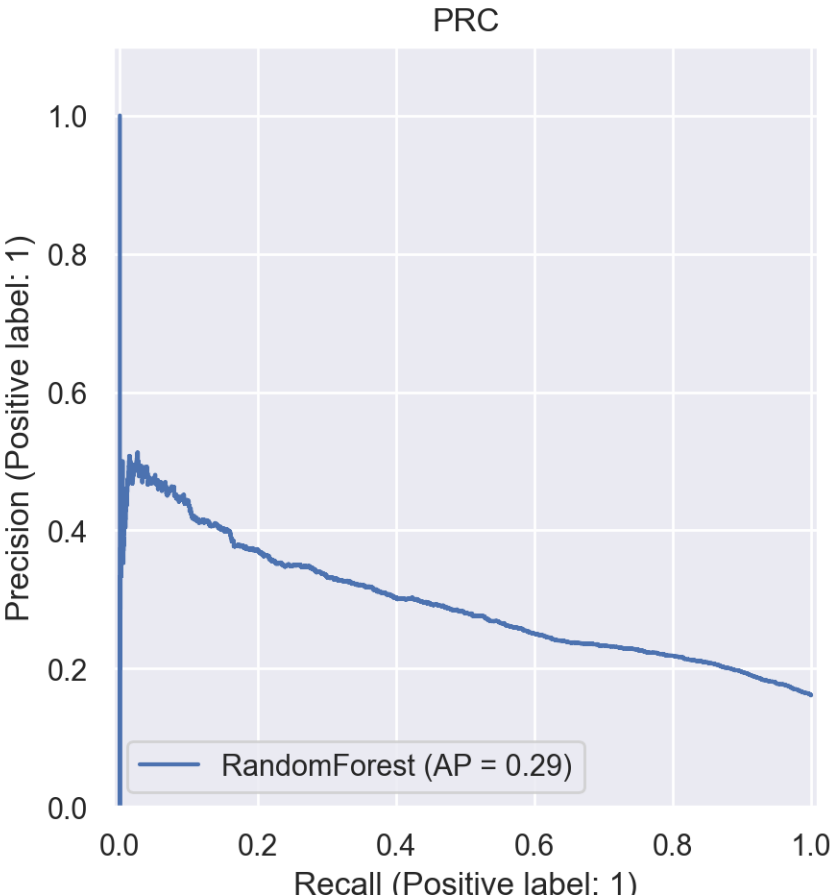

Коэффициент Джини для Train / Test с 95% ДИ
Δ Gini = -0.64 п. п. (-1.72% rel.)
37.29%
36.65%
Коэффициент Джини, %
0
10
20
30
40
50
Train
Test

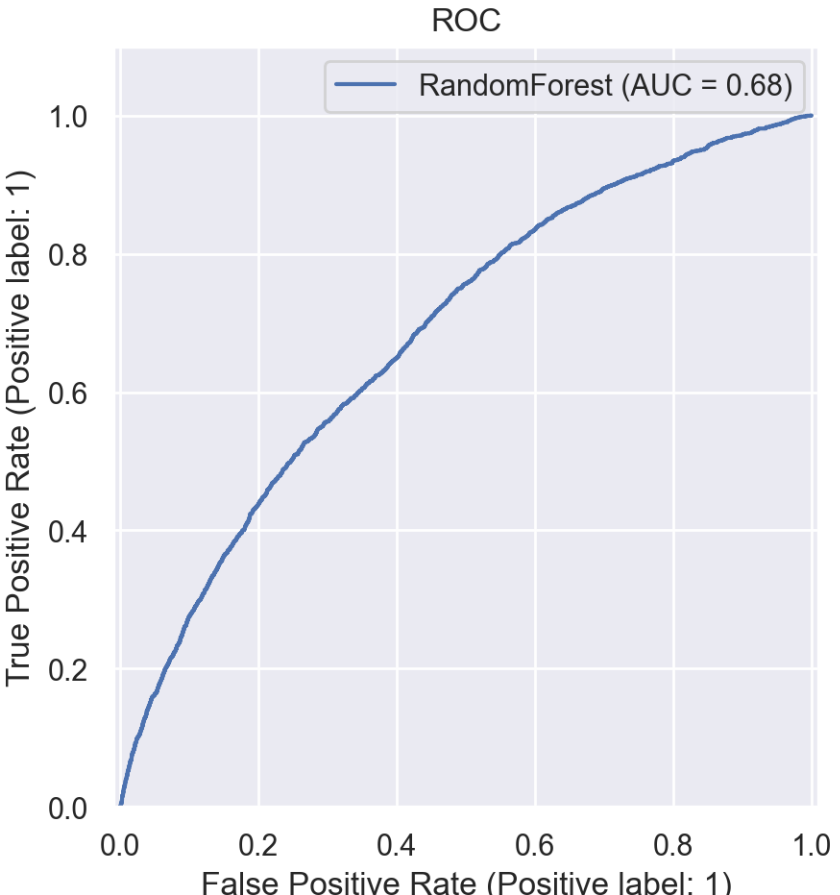

Сравнение Джини факторов на Train и Test
Коэффициент Джини, %
Gini Dev
Gini Val
0
10
20
30
40
Δ=+0.6 п. п.
(+2.0% rel)
30.3%
30.9%
Δ=+0.5 п. п.
(+3.0% rel)
18.0%
18.6%
Δ=-2.1 п. п.
(-11.2% rel)
18.5%
16.4%
Δ=-0.3 п. п.
(-3.1% rel)
11.0%
10.7%
Δ=-0.5 п. п.
(-4.9% rel)
10.1%
9.6%
Δ=+0.8 п. п.
(+7.6% rel)
10.2%
11.0%
Δ=-0.5 п. п.
(-6.1% rel)
7.8%
7.3%
Δ=+1.8 п. п.
(+33.5% rel)
5.4%
7.2%
Δ=-0.1 п. п.
(-2.4% rel)
6.1%
6.0%
Δ=+1.7 п. п.
(+26.7% rel)
6.2%
7.8%
Δ=-1.7 п. п.
(-25.8% rel)
6.7%
4.9%
WOE_sub_grade
WOE_term
WOE_LTI
WOE_dti
WOE_annual_inc
WOE_Activity
WOE_prof_group_mapped
WOE_avg_cur_bal
WOE_mths_since_recent_inq
WOE_tot_hi_cred_lim
WOE_IPEY

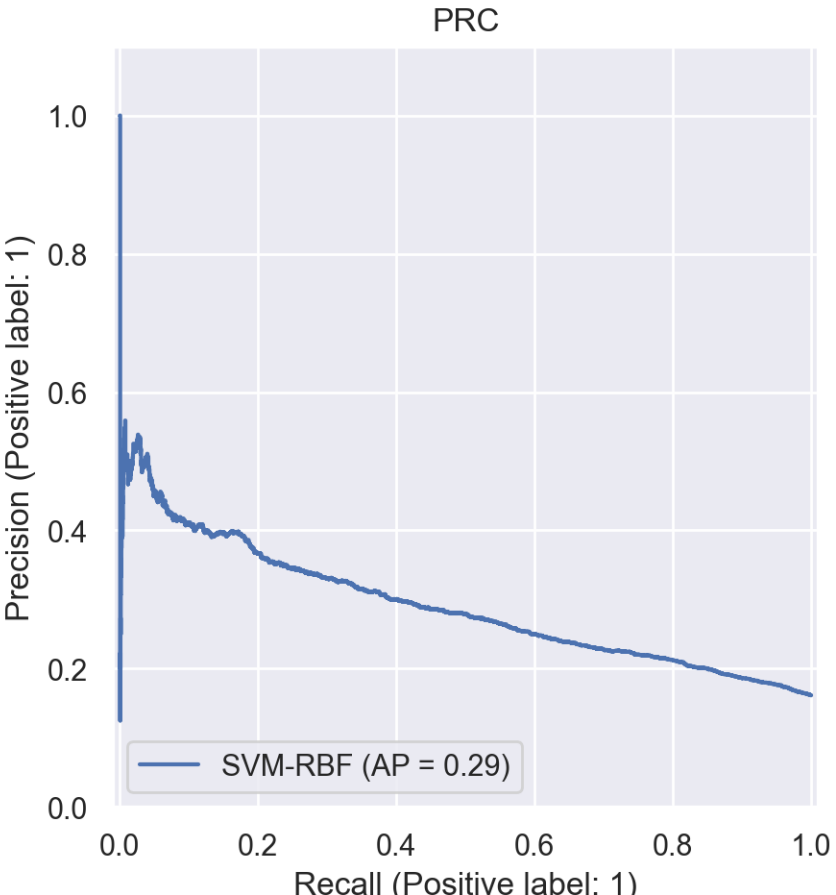

Сравнение Джини факторов на Train и Test
Коэффициент Джини, %
Gini Dev
Gini Val
0
10
20
30
40
Δ=+0.6 п. п.
(+2.0% rel)
30.3%
30.9%
Δ=+0.5 п. п.
(+3.0% rel)
18.0%
18.6%
Δ=-2.1 п. п.
(-11.2% rel)
18.5%
16.4%
Δ=-0.3 п. п.
(-3.1% rel)
11.0%
10.7%
Δ=-0.5 п. п.
(-4.9% rel)
10.1%
9.6%
Δ=+0.8 п. п.
(+7.6% rel)
10.2%
11.0%
Δ=-0.5 п. п.
(-6.1% rel)
7.8%
7.3%
Δ=+1.8 п. п.
(+33.5% rel)
5.4%
7.2%
Δ=-0.1 п. п.
(-2.4% rel)
6.1%
6.0%
Δ=+1.7 п. п.
(+26.7% rel)
6.2%
7.8%
WOE_sub_grade
WOE_term
WOE_LTI
WOE_dti
WOE_annual_inc
WOE_Activity
WOE_prof_group_mapped
WOE_avg_cur_bal
WOE_mths_since_recent_inq
WOE_tot_hi_cred_lim

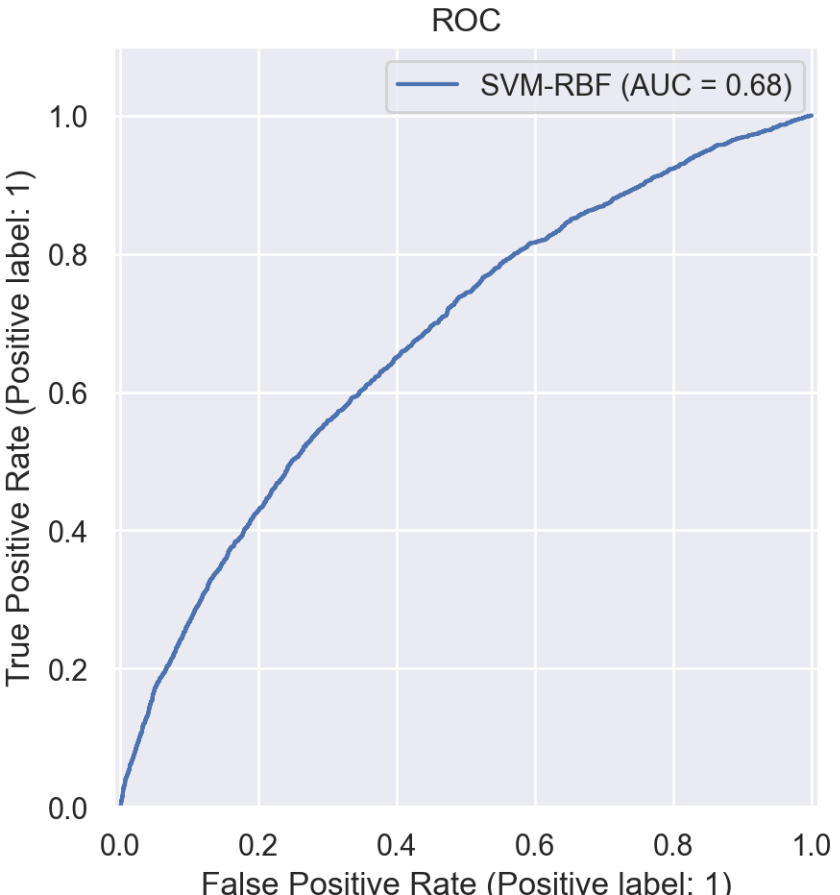

PRC
Precision (Positive label: 1)
Recall (Positive label: 1)
CatBoost (AP = 0.30)
0.0
0.2
0.4
0.6
0.8
1.0

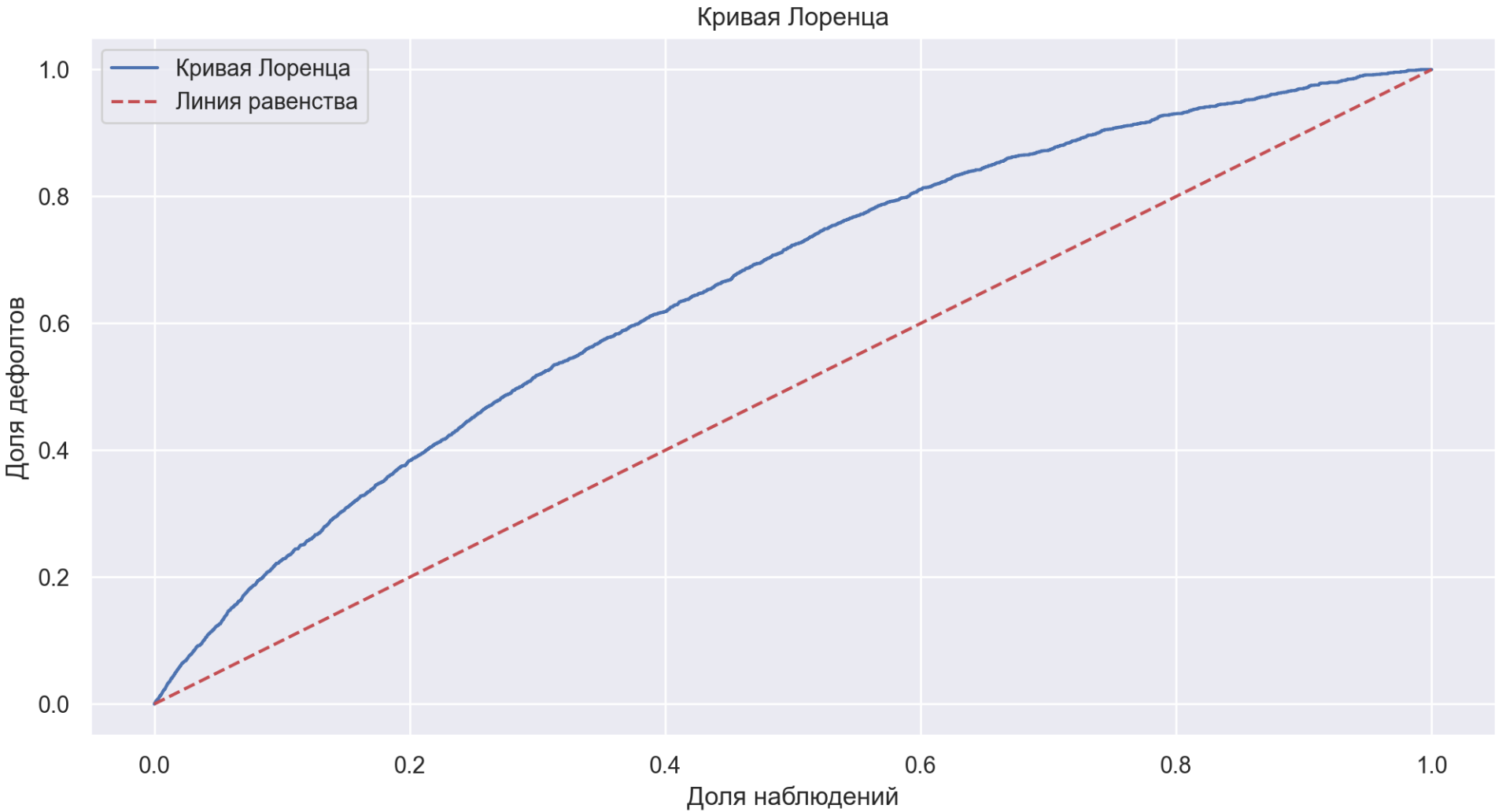

ROC
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
CatBoost (AUC = 0.69)
0.0
0.2
0.4
0.6
0.8
1.0

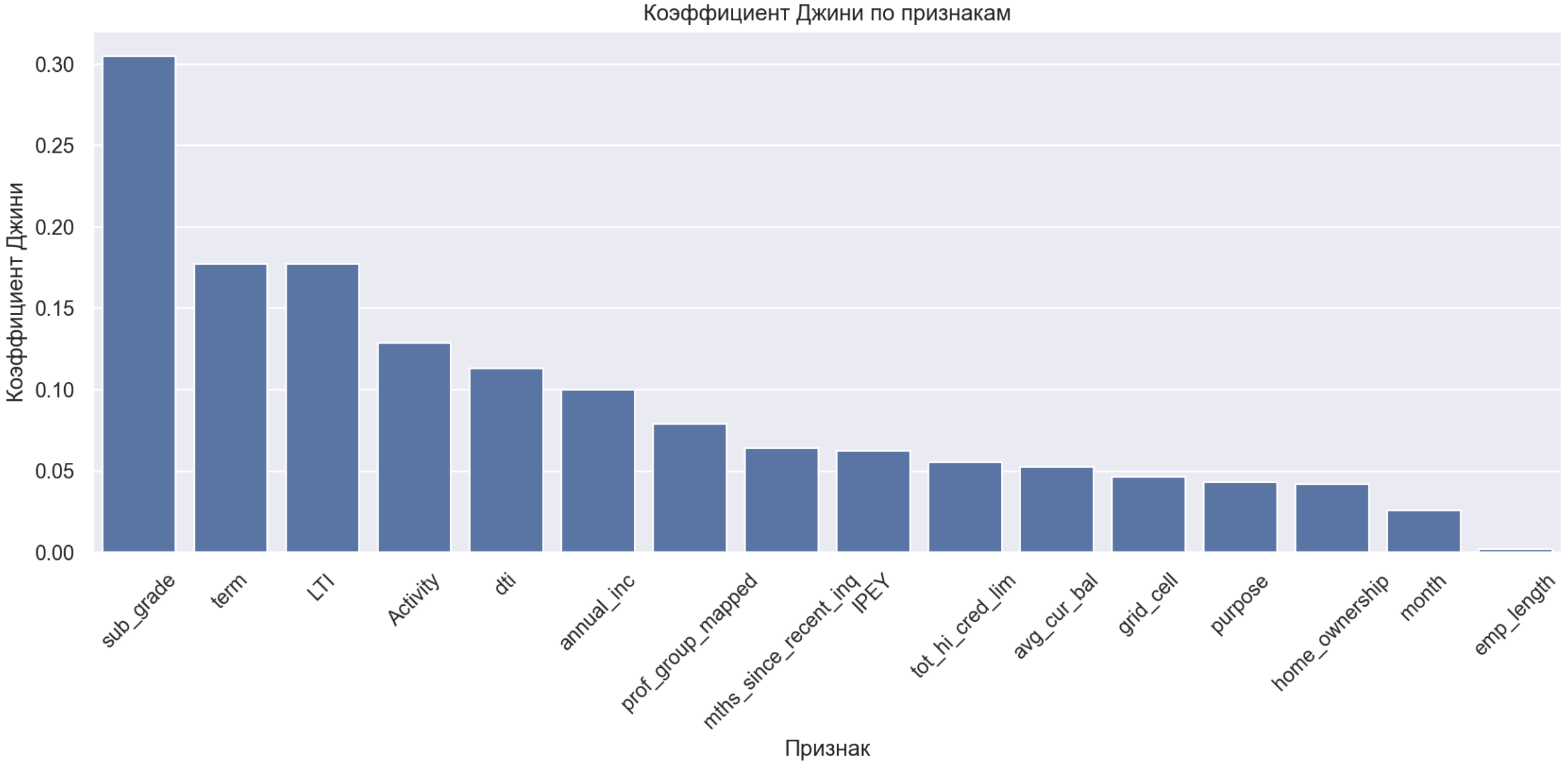

PRC
Precision (Positive label: 1)
Recall (Positive label: 1)
RandomForest (AP = 0.29)
0.0
0.2
0.4
0.6
0.8
1.0

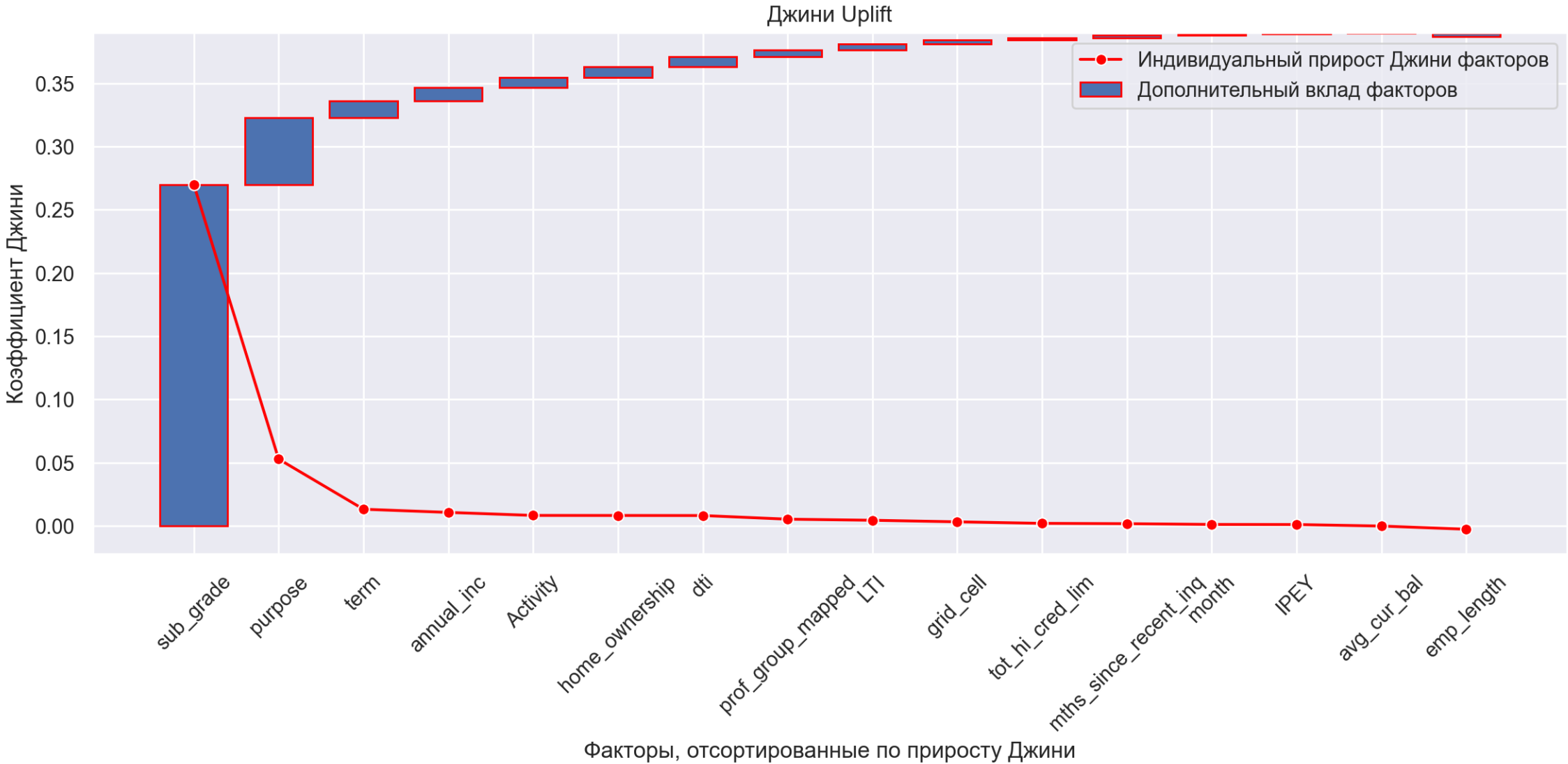

ROC
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
RandomForest (AUC = 0.68)
0.0
0.2
0.4
0.6
0.8
1.0

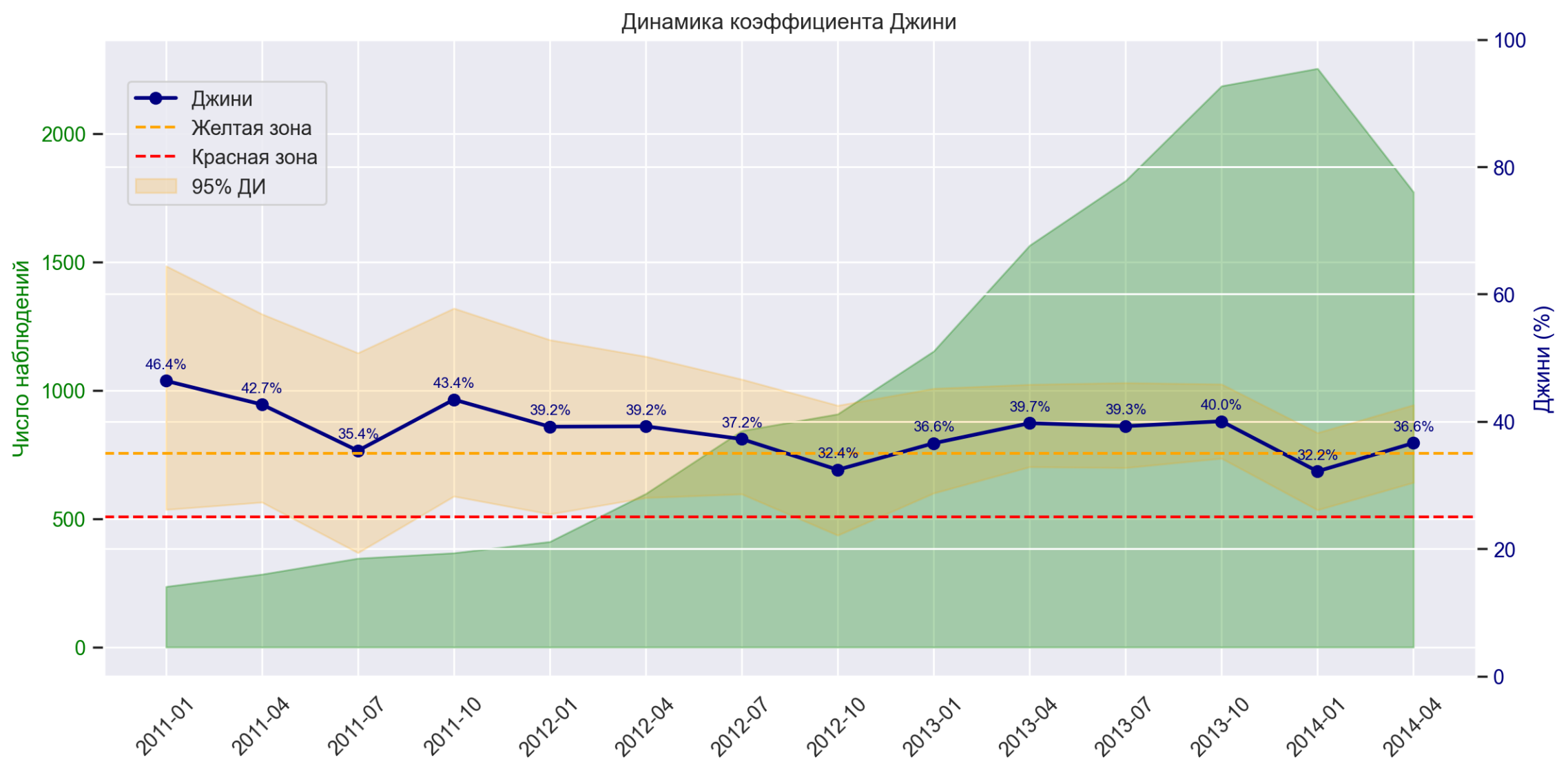

PRC
Precision (Positive label: 1)
Recall (Positive label: 1)
SVM-RBF (AP = 0.29)
0.0
0.2
0.4
0.6
0.8
1.0

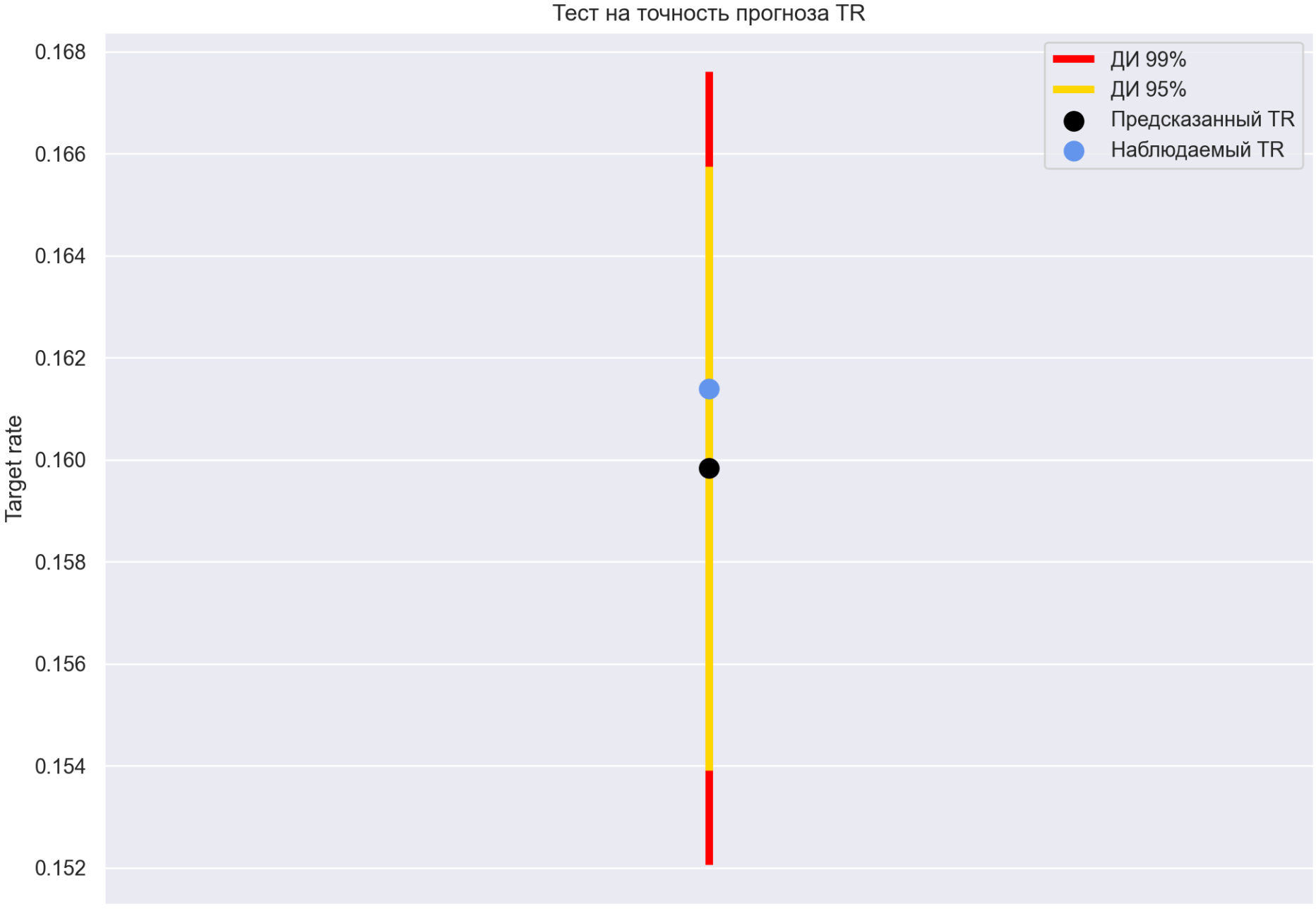

ROC
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
SVM-RBF (AUC = 0.68)
0.0
0.2
0.4
0.6
0.8
1.0

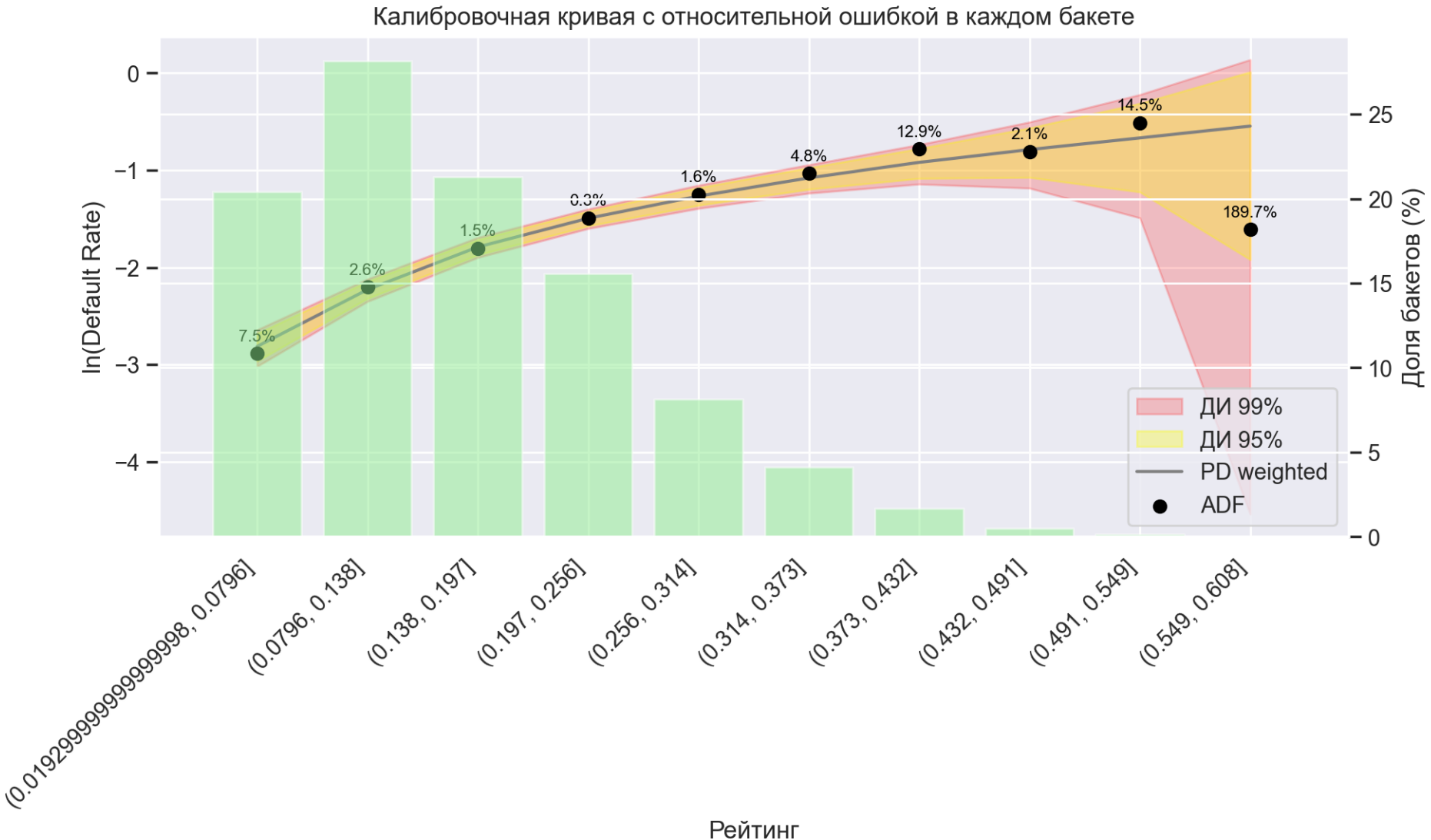

Кривая Лоренца
Кривая Лоренца
Линия равенства
Доля дефолтов
Доля наблюдений
0.0
0.2
0.4
0.6
0.8
1.0

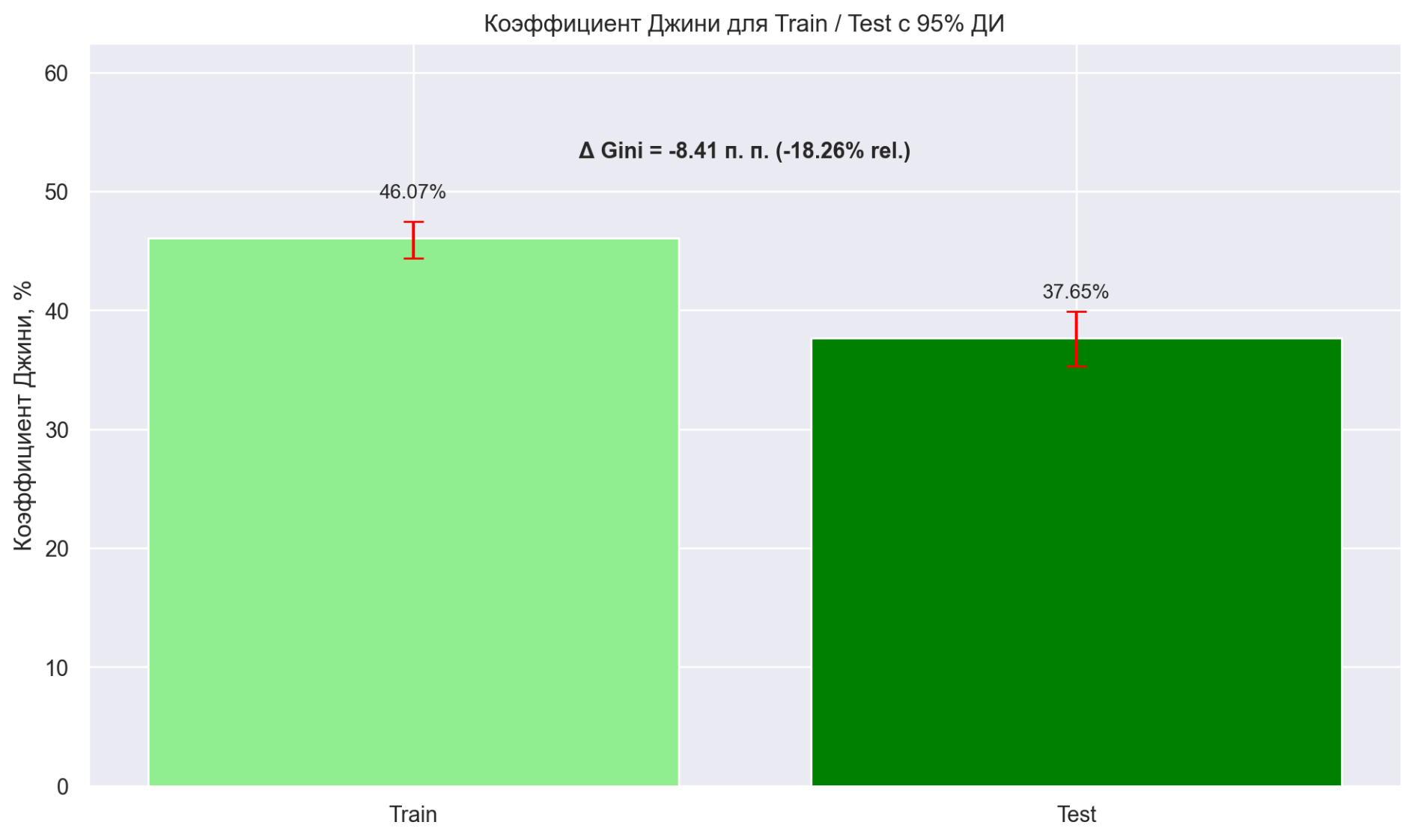

Коэффициент Джини по признакам
Коэффициент Джини
0.30
0.25
0.20
0.15
0.10
0.05
0.00
sub_grade
term
LTI
Activity
dti
annual_inc
prof_group_mapped
mths_since_recent_inq
IPEY
tot_hi_cred_lim
avg_cur_bal
grid_cell
purpose
home_ownership
month
emp_length
Признак

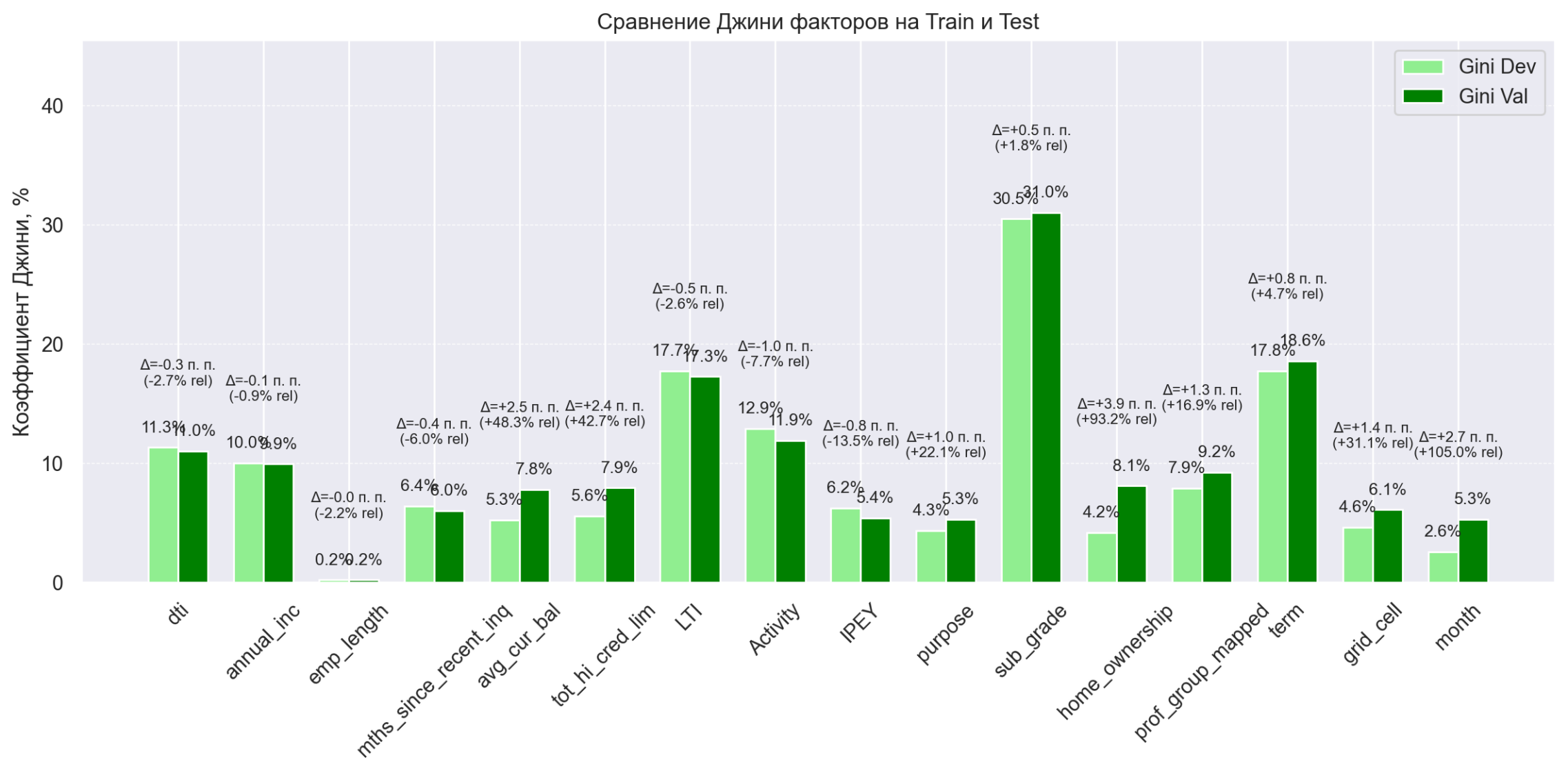

Джини Uplift
Индивидуальный прирост Джини факторов
Дополнительный вклад факторов
Коэффициент Джини
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
sub_grade
purpose
term
annual_inc
Activity
home_ownership
dti
prof_group_mapped
LTI
grid_cell
tot_hi_cred_lim
mths_since_recent_inq
month
IPEY
avg_cur_bal
emp_length
Факторы, отсортированные по приросту Джини

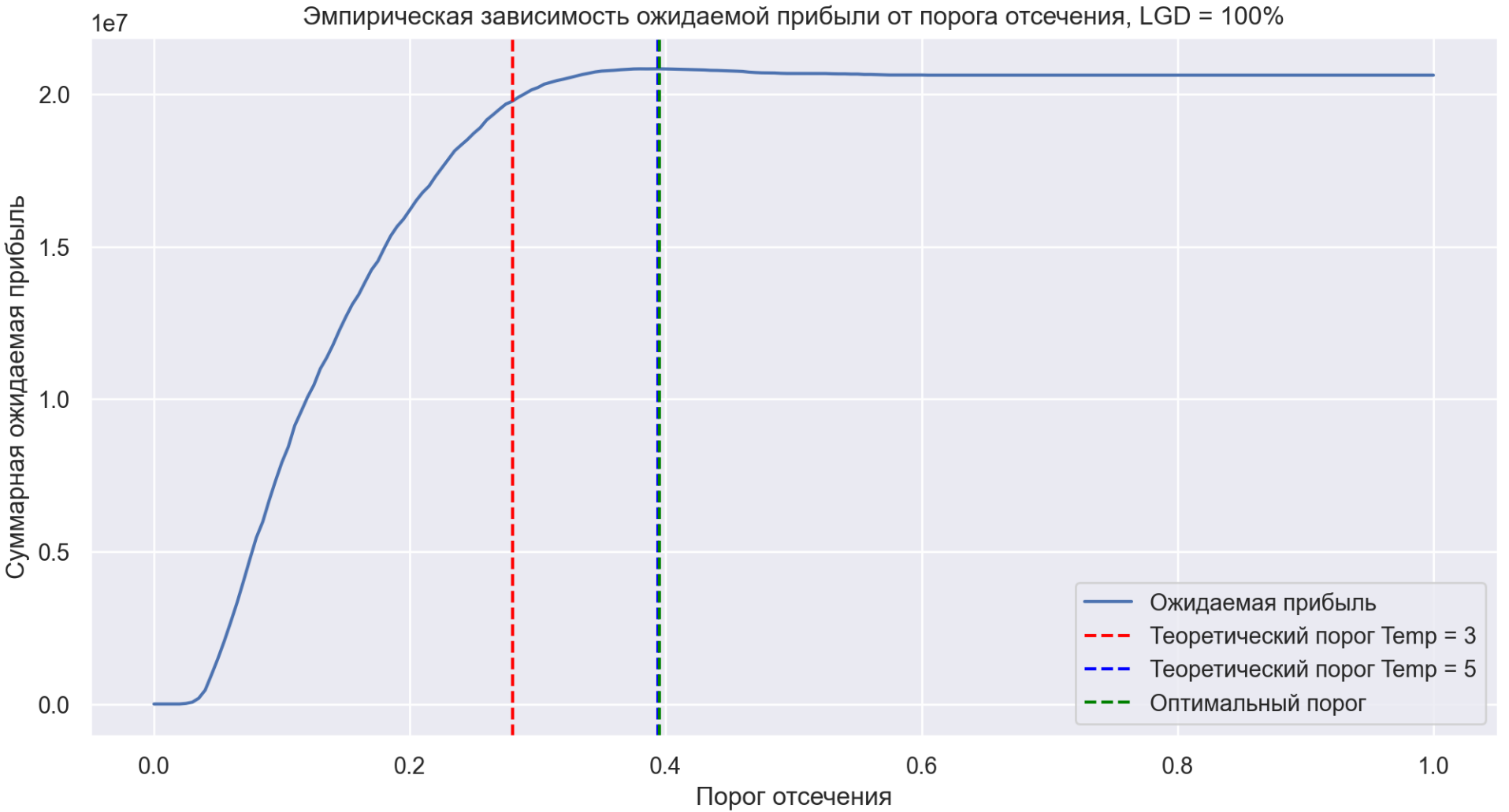

Динамика коэффициента Джини
Джини
Желтая зона
Красная зона
95% ДИ
Число наблюдений
Джини (%)
0
500
1000
1500
2000
0
20
40
60
80
100
46.4%
42.7%
35.4%
43.4%
39.2%
39.2%
37.2%
32.4%
36.6%
39.7%
39.3%
40.0%
32.2%
36.6%
2011-01
2011-04
2011-07
2011-10
2012-01
2012-04
2012-07
2012-10
2013-01
2013-04
2013-07
2013-10
2014-01
2014-04

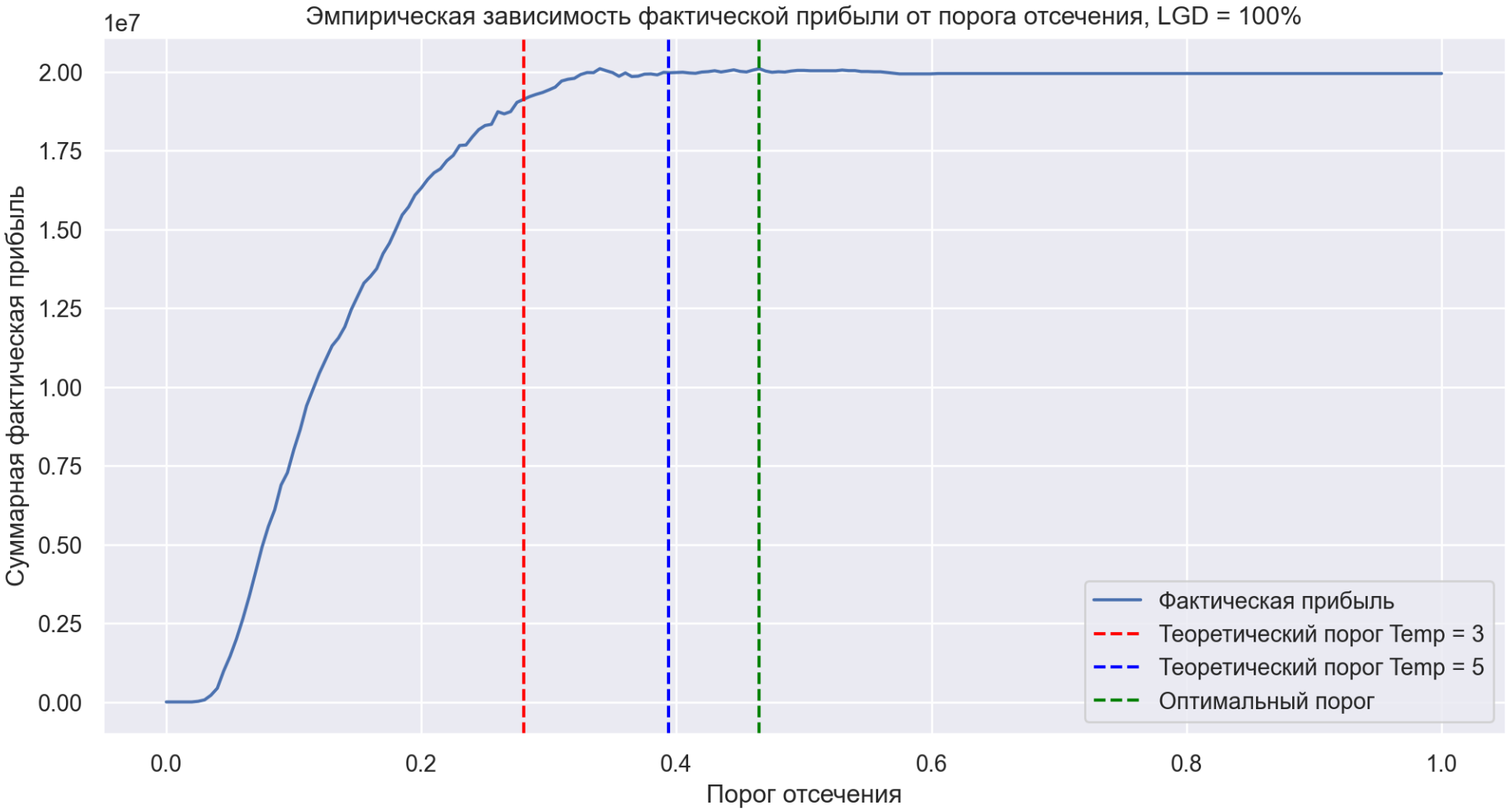

Тест на точность прогноза TR
Target rate
0.168
0.166
0.164
0.162
0.160
0.158
0.156
0.154
0.152
ДИ 99%
ДИ 95%
Предсказанный TR
Наблюдаемый TR

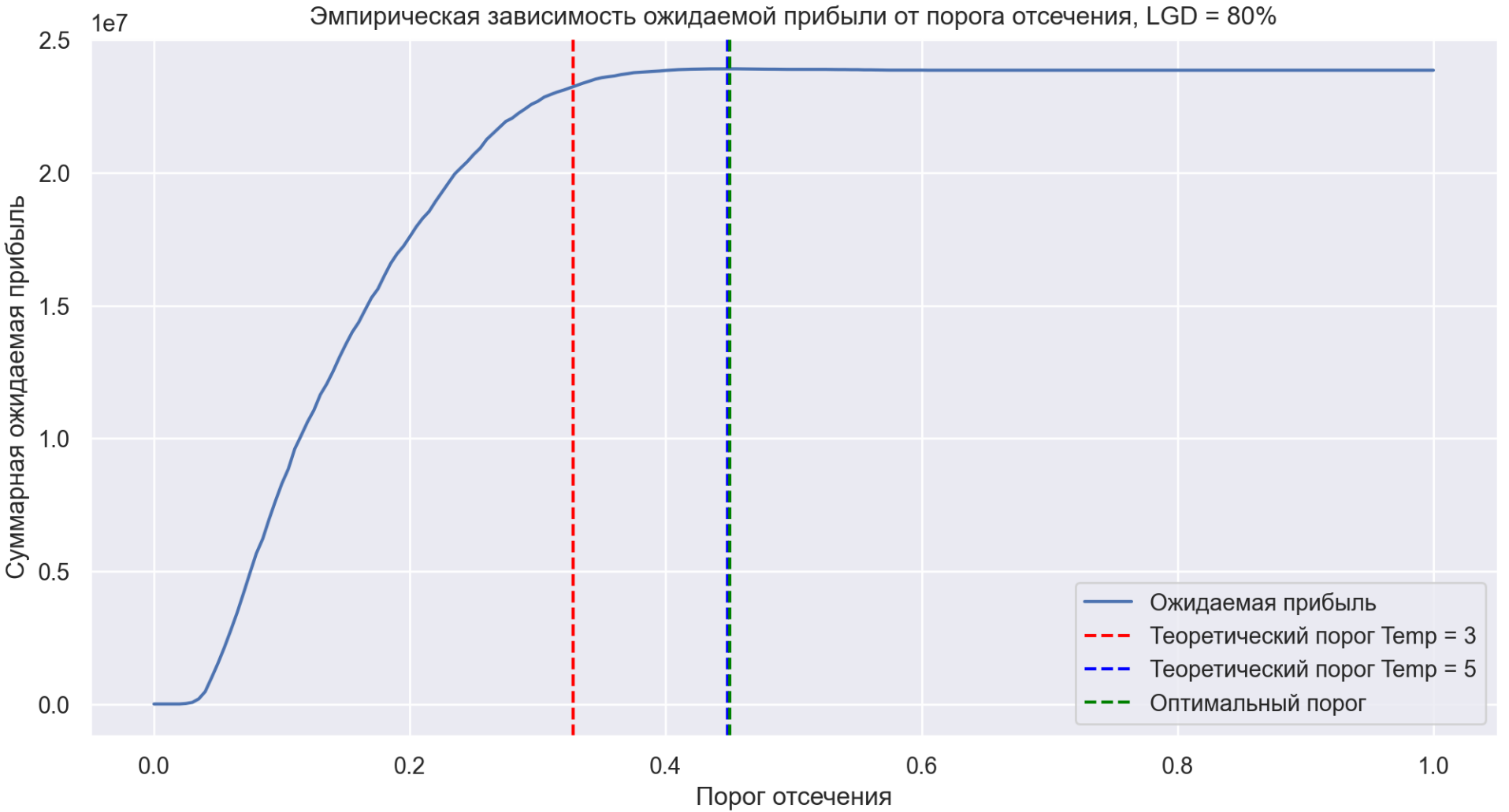

Калибровочная кривая с относительной ошибкой в каждом бакете
ln(Default Rate)
Доля бакетов (%)
Рейтинг
0
−1
−2
−3
−4
25
20
15
10
5
0
7.5%
2.6%
1.5%
0.3%
1.6%
4.8%
12.9%
2.1%
14.5%
189.7%
ДИ 99%
ДИ 95%
PD weighted
ADF
(0.019299999999998, 0.0796]
(0.0796, 0.138]
(0.138, 0.197]
(0.197, 0.256]
(0.256, 0.314]
(0.314, 0.373]
(0.373, 0.432]
(0.432, 0.491]
(0.491, 0.549]
(0.549, 0.608]

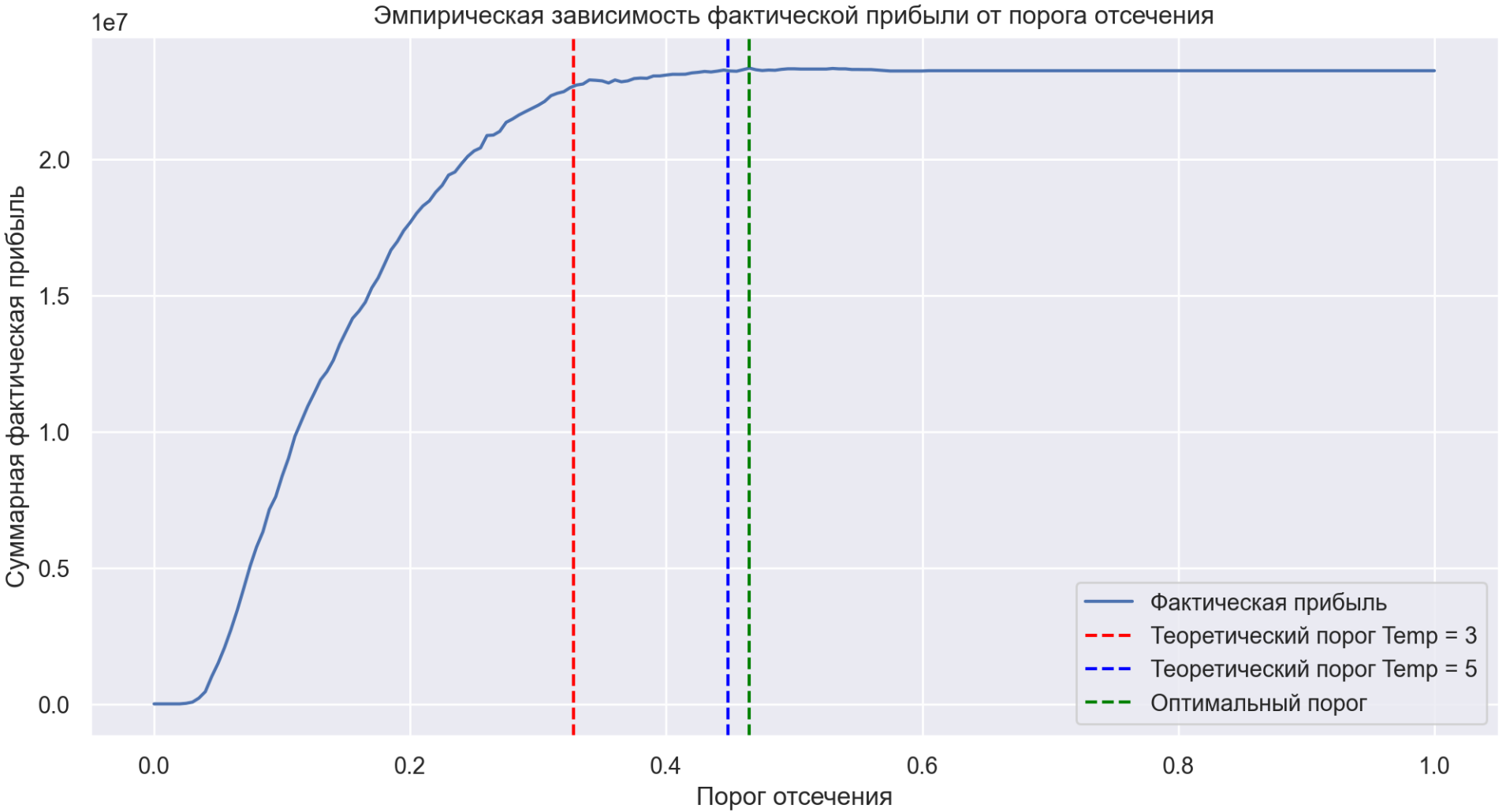

Коэффициент Джини для Train / Test с 95% ДИ
Δ Gini = -8.41 п. п. (-18.26% rel.)
46.07%
37.65%
Коэффициент Джини, %
0
10
20
30
40
50
60
Train
Test


Сравнение Джини факторов на Train и Test
Gini Dev
Gini Val
Коэффициент Джини, %
0
10
20
30
40
Δ=-0.3 п. п. (-2.7% rel)
11.3%
11.0%
Δ=-0.1 п. п. (-0.9% rel)
10.0%
9.9%
Δ=-0.0 п. п. (-2.2% rel)
0.2%
0.2%
Δ=-0.4 п. п. (-6.0% rel)
6.4%
6.0%
Δ=+2.5 п. п. (+48.3% rel)
5.3%
7.8%
Δ=+2.4 п. п. (+42.7% rel)
5.6%
7.9%
Δ=-0.5 п. п. (-2.6% rel)
17.7%
17.3%
Δ=-1.0 п. п. (-7.7% rel)
12.9%
11.9%
Δ=-0.8 п. п. (-13.5% rel)
6.2%
5.4%
Δ=+1.0 п. п. (+22.1% rel)
4.3%
5.3%
Δ=+0.5 п. п. (+1.8% rel)
30.5%
31.0%
Δ=+3.9 п. п. (+93.2% rel)
4.2%
8.1%
Δ=+1.3 п. п. (+16.9% rel)
7.9%
9.2%
Δ=+0.8 п. п. (+4.7% rel)
17.8%
18.6%
Δ=+1.4 п. п. (+31.1% rel)
4.6%
6.1%
Δ=+2.7 п. п. (+105.0% rel)
2.6%
5.3%
dti
annual_inc
emp_length
mths_since_recent_inq
avg_cur_bal
tot_hi_cred_lim
LTI
Activity
IPEY
purpose
sub_grade
home_ownership
prof_group_mapped
term
grid_cell
month


Эмпирическая зависимость ожидаемой прибыли от порога отсечения, LGD = 100%
1e7
2.0
1.5
1.0
0.5
0.0
Суммарная ожидаемая прибыль
0.0
0.2
0.4
0.6
0.8
1.0
Порог отсечения
Ожидаемая прибыль
Теоретический порог Temp = 3
Теоретический порог Temp = 5
Оптимальный порог


Эмпирическая зависимость фактической прибыли от порога отсечения, LGD = 100%
1e7
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Суммарная фактическая прибыль
0.0
0.2
0.4
0.6
0.8
1.0
Порог отсечения
Фактическая прибыль
Теоретический порог Temp = 3
Теоретический порог Temp = 5
Оптимальный порог


Эмпирическая зависимость ожидаемой прибыли от порога отсечения, LGD = 80%
1e7
2.5
2.0
1.5
1.0
0.5
0.0
Суммарная ожидаемая прибыль
0.0
0.2
0.4
0.6
0.8
1.0
Порог отсечения
Ожидаемая прибыль
Теоретический порог Temp = 3
Теоретический порог Temp = 5
Оптимальный порог


Эмпирическая зависимость фактической прибыли от порога отсечения
1e7
2.0
1.5
1.0
0.5
0.0
Суммарная фактическая прибыль
0.0
0.2
0.4
0.6
0.8
1.0
Порог отсечения
Фактическая прибыль
Теоретический порог Temp = 3
Теоретический порог Temp = 5
Оптимальный порог