

MegaQuant



Задача:

Разработать модель предсказания дефолта

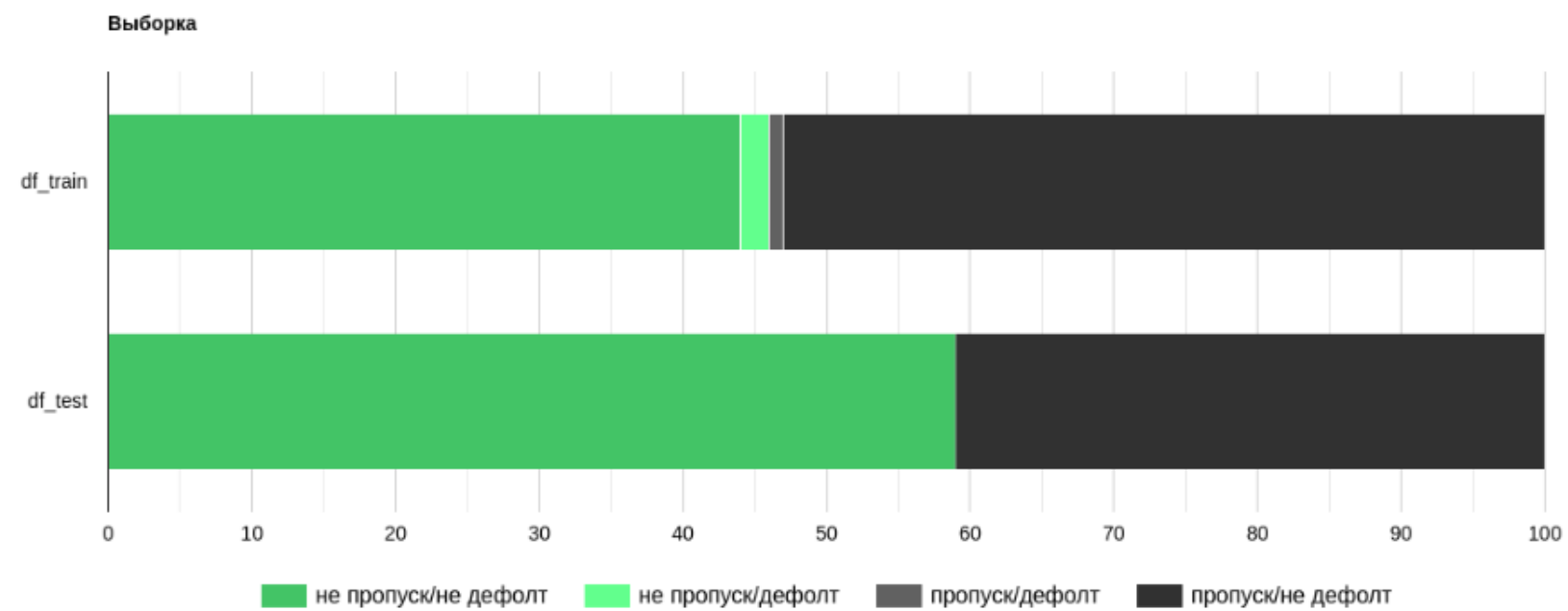
- Даны финансовые данные 32 395 компаний (выручка, активы, пассивы и т.д)
- Результат будет оцениваться на скрытой выборке из 200 компаний по метрике качества `accuracy_score`

Ход работы

- 1 Анализ данных
- 2 Анализ предметной области
- 3 Отбор признаков
- 4 Выбор модели
- 5 Результаты работы и интерпретация
- 6 Преимущества модели

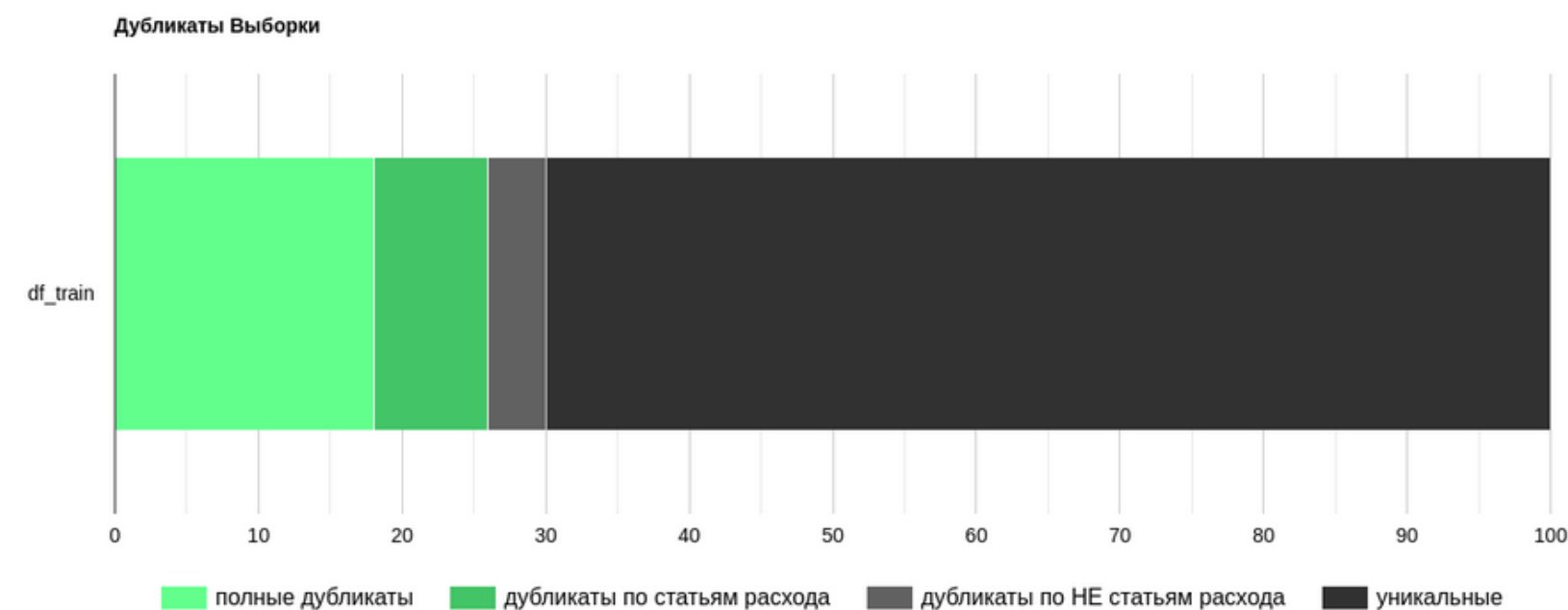
Анализ данных

Пропуски и дубликаты



~50% пропусков

Компании с пропусками просто не имеют/не предоставли своих бухгалтерских отчетов.



~18/26/30% дубликатов

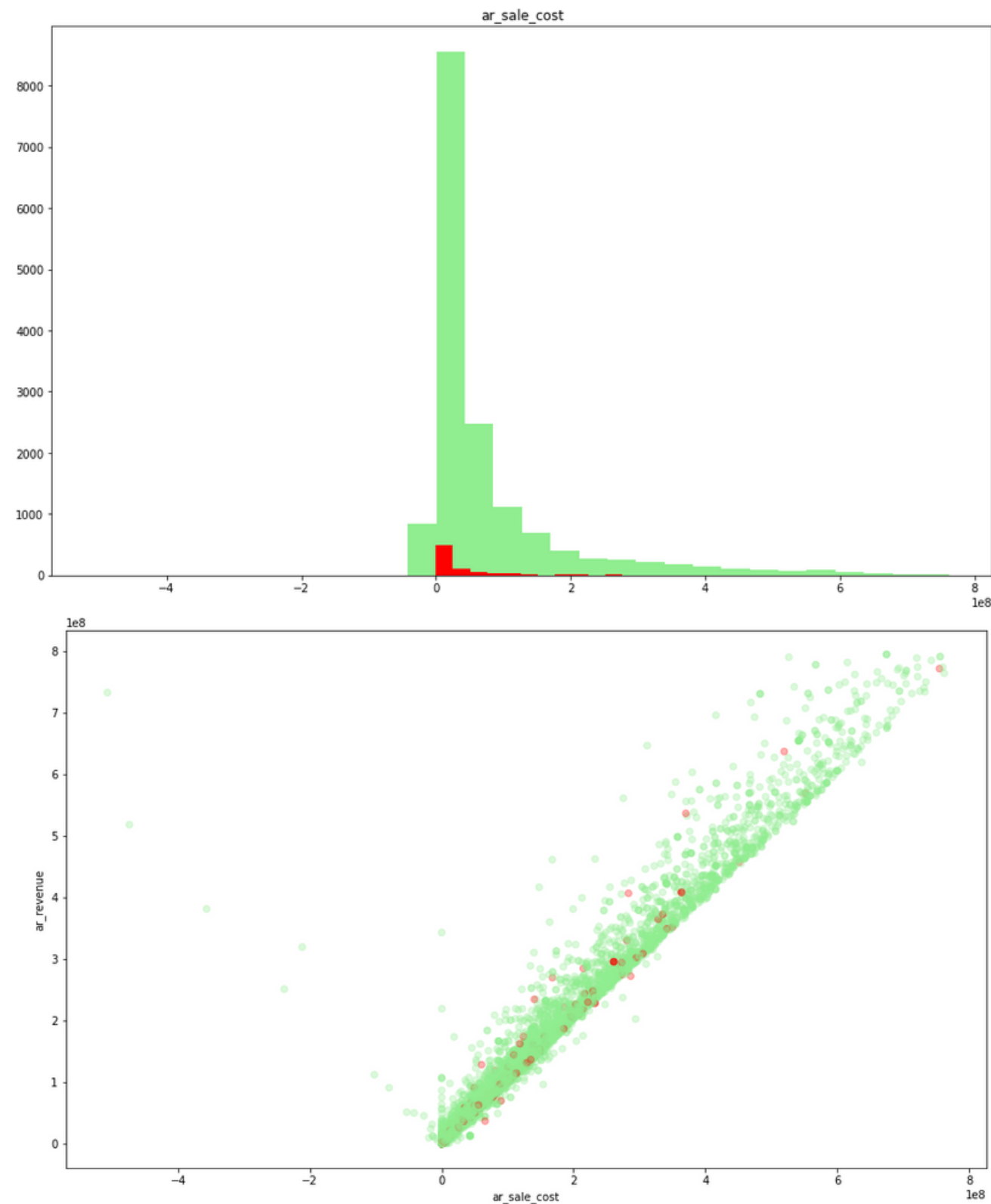
18% полных

26% по состоянию баланса

30% по предприятию

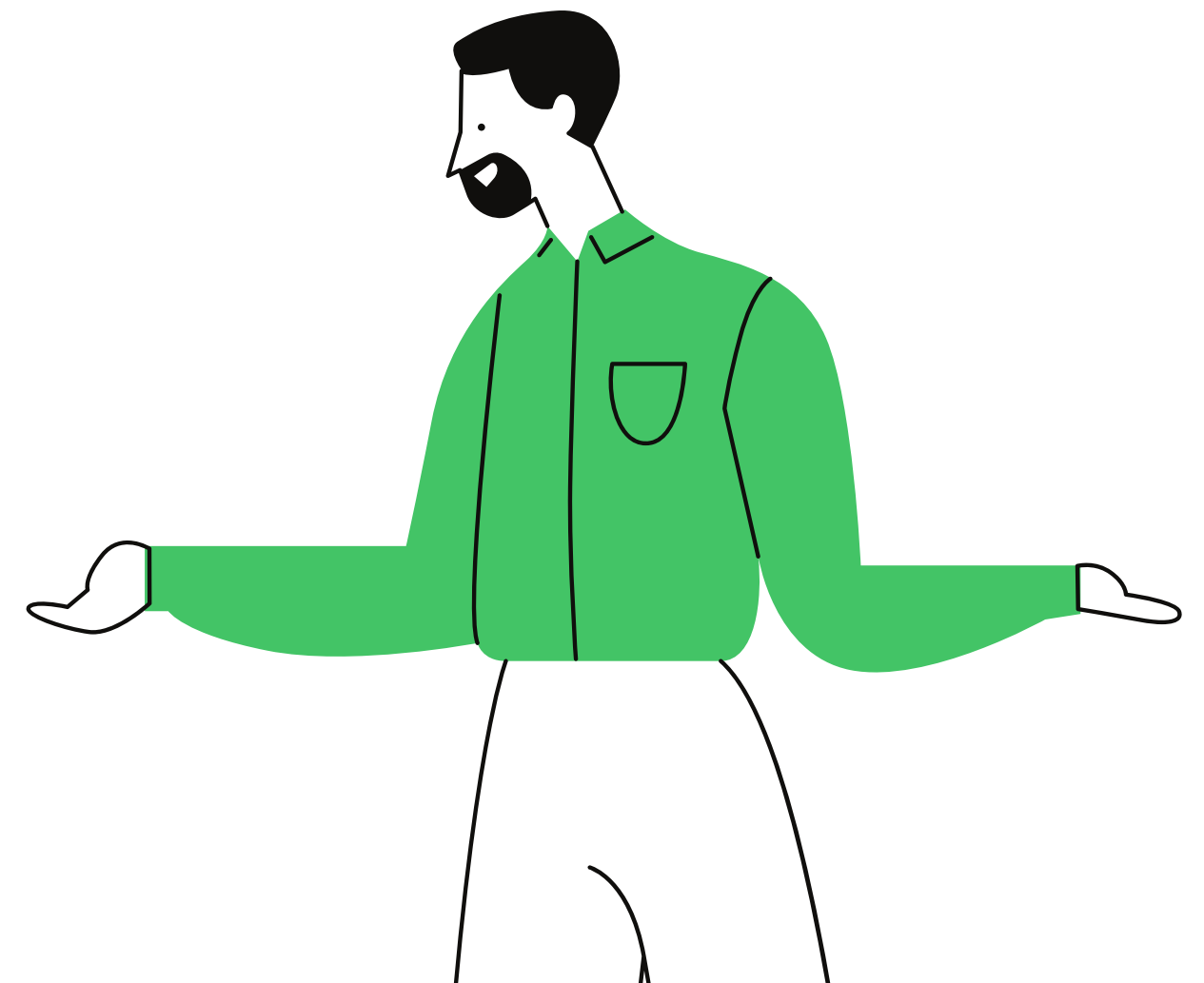
Анализ данных

Пропуски и дубликаты



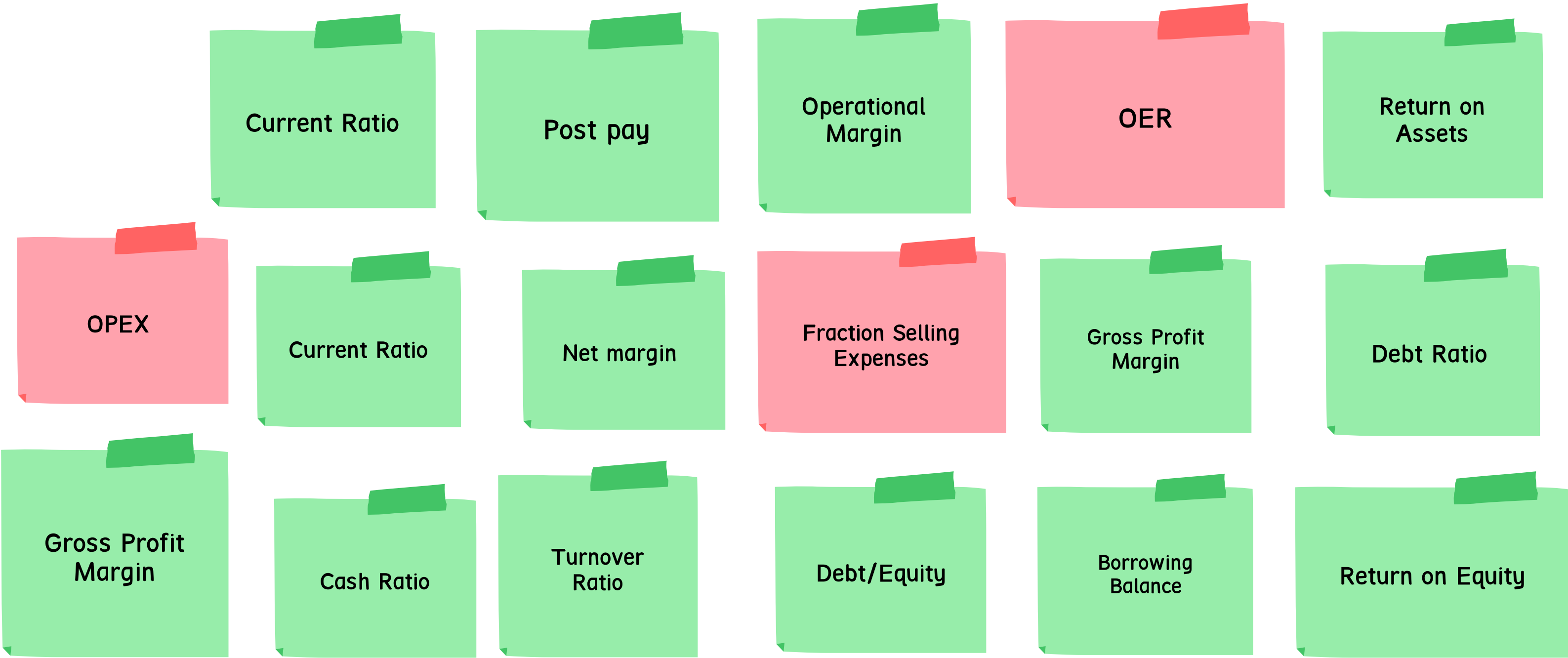
~8% не по МСФО

Редкие отрицательные значения в
некотрых статьях



Анализ предметной области

Создание новых признаков на основе данных



WoE-binning

числовых и категориальных переменных

- 1 Выявление сложных нелинейных взаимосвязей
- 2 Преобразование основано на логарифмическом распределении
- 3 Нет необходимости в dummy variables
- 4 Снижение риска переобучения модели
- 5 Не учитывание выбросов
- 6 Преобразовываете независимой переменной для установления монотонной связи с зависимой переменной

WoE-трансформация

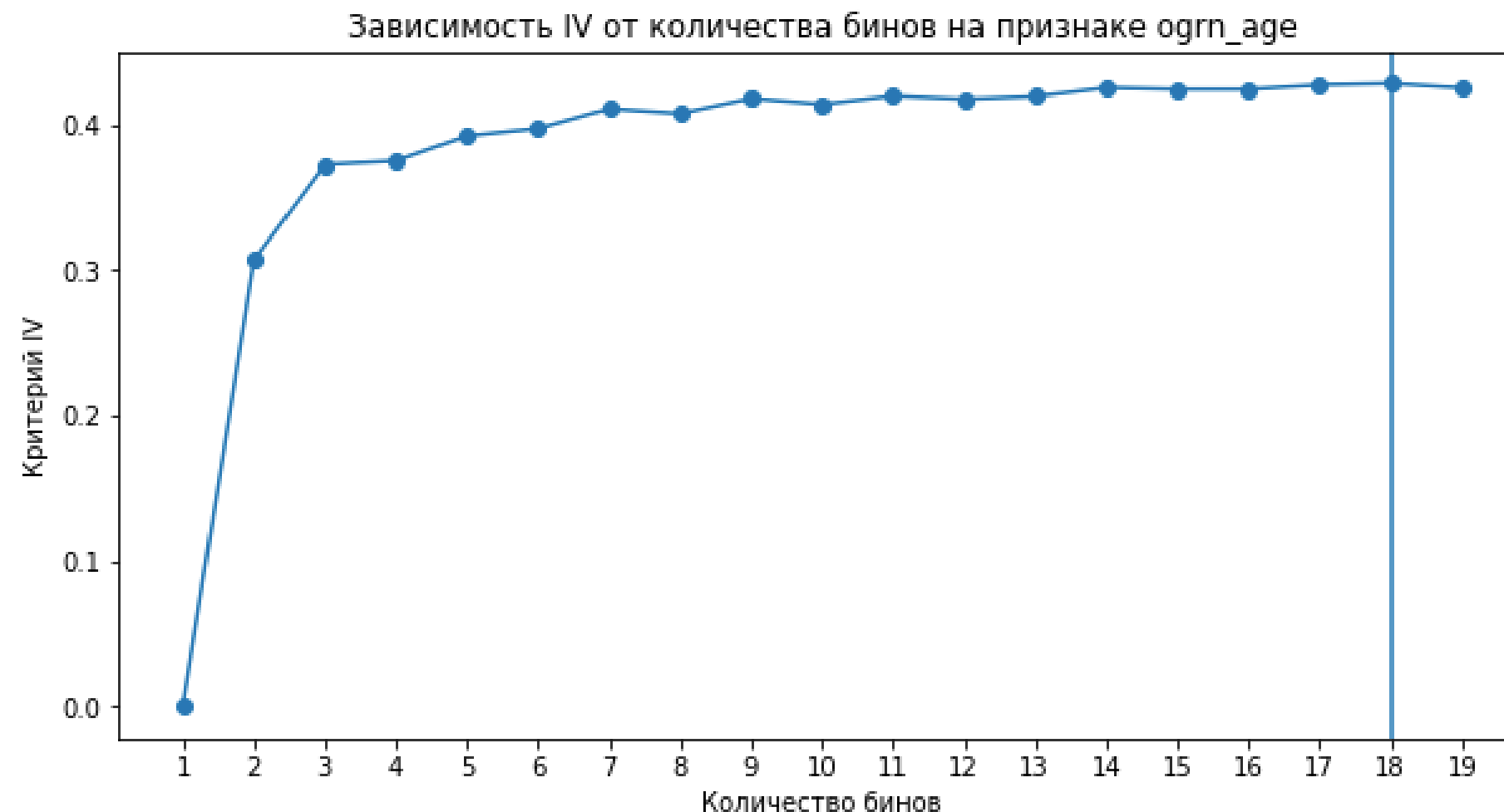
1

Определяем **оптимальное количество бинов** в каждом признаке **на основе критерия информативности IV** (Information Value).

Лучшее разбиение – то, которое показало наибольший показатель IV. IV рассчитывается по следующей формуле:

$$IV = \sum (\text{Event\%} - \text{Non Event\%}) * (\text{WOE})$$

$$\text{WOE} = \ln \left(\frac{\text{Event\%}}{\text{Non Event\%}} \right)$$



WoE-трансформация

2

Поделив каждый признак на бины, заменим значения бина на его WoE (Weight of evidence).


3

Применим то же разделение на бины к каждому признаку тестовой выборки и заменим значение бинов на WoE

	value	count_all	default	non_default	Distr_non_default	Distr_default	WoE_cut ogrn_age	IV
0	(131.471, 140.235]	5477	164	5313	0.219991	0.096527	0.823760	0.101704
1	(26.294, 35.059]	1902	231	1671	0.069190	0.135962	-0.675526	0.045107
2	(96.412, 105.176]	1530	68	1462	0.060536	0.040024	0.413767	0.008487
3	(17.529, 26.294]	1853	227	1626	0.067326	0.133608	-0.685358	0.045427
4	(43.824, 52.588]	1538	151	1387	0.057430	0.088876	-0.436667	0.013731
5	(105.176, 113.941]	849	27	822	0.034036	0.015892	0.761618	0.013819
6	(70.118, 78.882]	1105	73	1032	0.042731	0.042966	-0.005491	0.000001
7	(35.059, 43.824]	1416	135	1281	0.053041	0.079459	-0.404164	0.010677
8	(52.588, 61.353]	1377	93	1284	0.053166	0.054738	-0.029150	0.000046
9	(140.235, 149.0]	1993	39	1954	0.080908	0.022955	1.259786	0.073008
10	(113.941, 122.706]	760	36	724	0.029978	0.021189	0.346987	0.003050
11	(8.765, 17.529]	1101	180	921	0.038135	0.105945	-1.021783	0.069287
12	(78.882, 87.647]	1171	65	1106	0.045795	0.038258	0.179832	0.001355
13	(87.647, 96.412]	1499	63	1436	0.059459	0.037081	0.472196	0.010567
14	(122.706, 131.471]	798	29	769	0.031841	0.017069	0.623509	0.009211
15	(61.353, 70.118]	1217	71	1146	0.047451	0.041789	0.127067	0.000719
16	(-0.149, 8.765]	264	47	217	0.008985	0.027663	-1.124536	0.021004

Модель

Создание модели предсказания дефолта



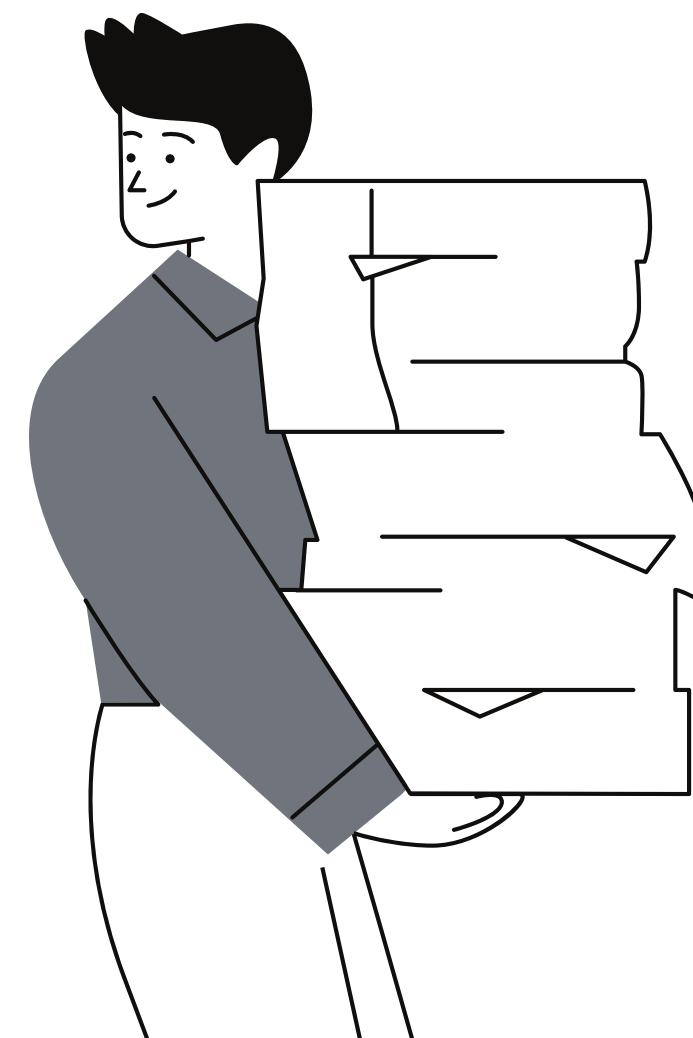
Logloss Regression +
StandardScaler +
SMOTE +
GridSearchCV +
подбор порога

Результат

0.65 accuracy

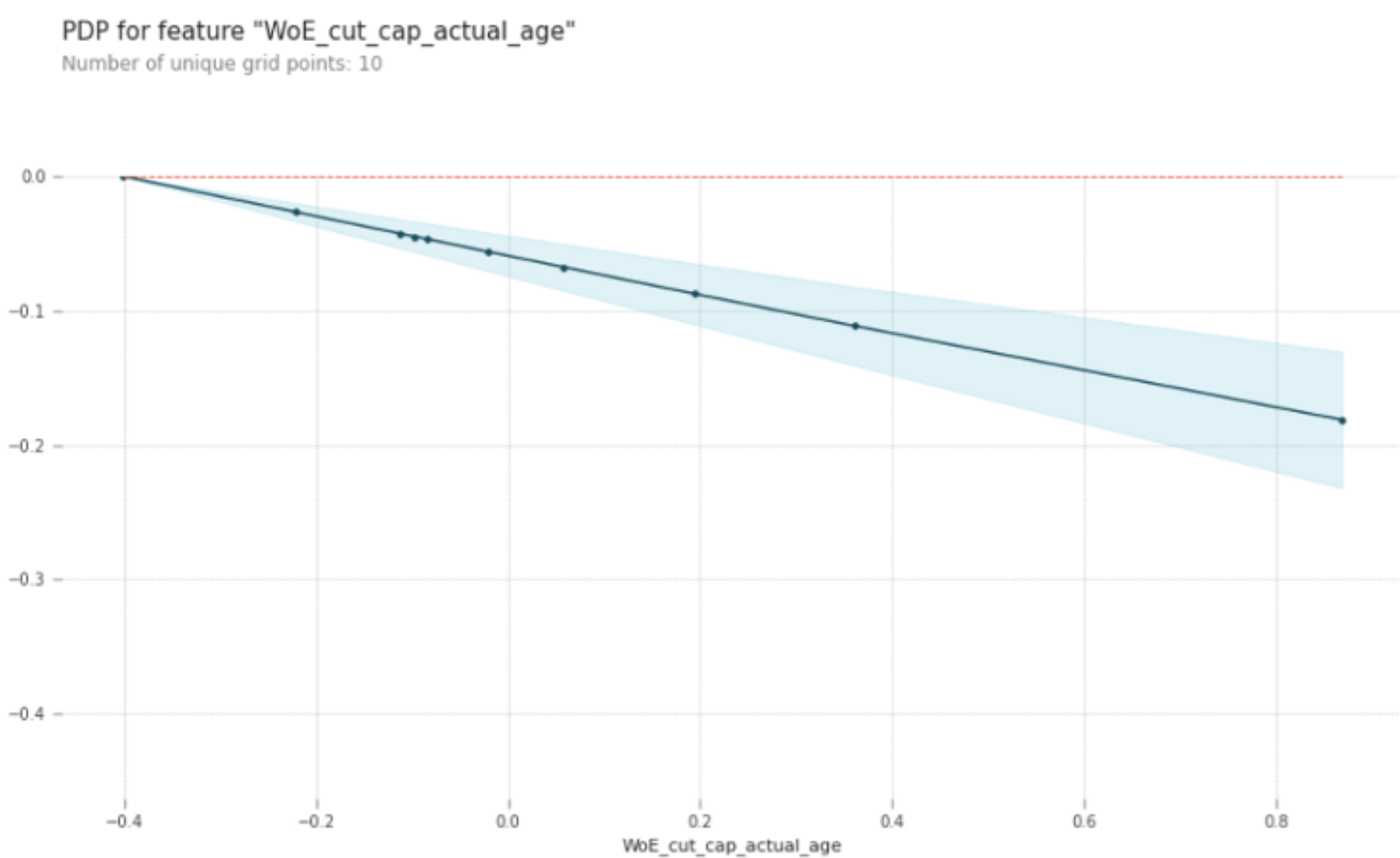
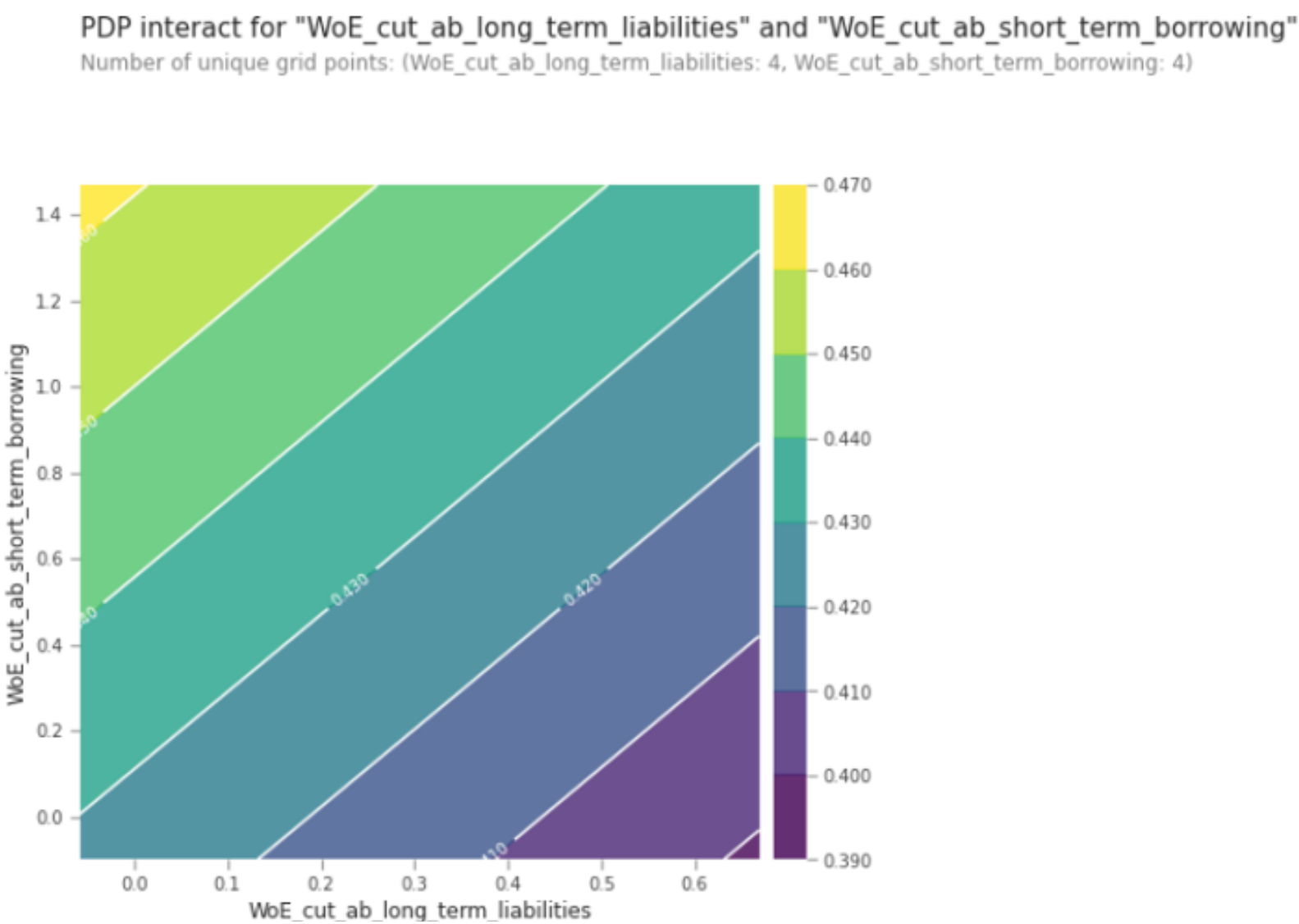
0.68 ROC AUC

Интерпретация модели



Интерпретация модели

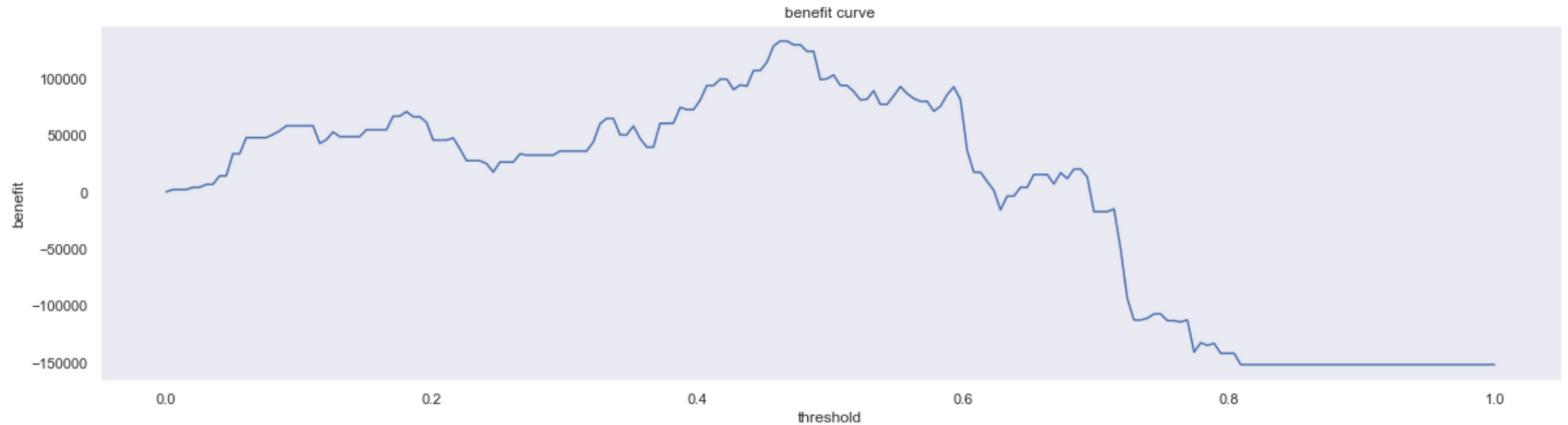
PDP plots



Интерпретация модели

benefit-кривая

```
gini 0.36  
max benefit 132945.39000000004  
best_threshold 0.4623115577889447
```



$\text{Benefit} = \text{Количество правильно классифицированных нулей} \times (r \times Debt) - \text{Количество неправильно классифицированных нулей} \times (LGD \times Debt)$

Интерпретация модели

Скоринговая карта

	value	WoE_cut_ogrn_age	scoring_points
0	(131.471, 140.235]	0.823760	49.098911
1	(26.294, 35.059]	-0.675526	2.106103
2	(96.412, 105.176]	0.413767	36.248324
3	(17.529, 26.294]	-0.685358	1.797949
4	(43.824, 52.588]	-0.436667	9.592765
5	(105.176, 113.941]	0.761618	47.151166
6	(70.118, 78.882]	-0.005491	23.107312
7	(35.059, 43.824]	-0.404164	10.611519
8	(52.588, 61.353]	-0.029150	22.365772
9	(140.235, 149.0]	1.259786	62.765498
10	(113.941, 122.706]	0.346987	34.155192
11	(8.765, 17.529]	-1.021783	-8.746779
12	(78.882, 87.647]	0.179832	28.915990
13	(87.647, 96.412]	0.472196	38.079694
14	(122.706, 131.471]	0.623509	42.822371
15	(61.353, 70.118]	0.127067	27.262155
16	(-0.149, 8.765]	-1.124536	-11.967428

Скоры по значениям срока
с момента присваивания ОГРН

	value	WoE_cut_cap_actual_age	scoring_points
0	(2.882, 4.765]	-0.113424	22.092369
1	(6.647, 8.529]	-0.097720	22.256722
2	(25.471, 27.353]	0.057207	23.878142
3	(14.176, 16.059]	-0.020600	23.063837
4	(0.968, 2.882]	-0.401428	19.078195
5	(21.706, 23.588]	0.186330	25.229517
6	(10.412, 12.294]	-0.100493	22.227696
7	(4.765, 6.647]	-0.220459	20.972169
8	(8.529, 10.412]	-0.043516	22.824004
9	(27.353, 29.235]	0.354692	26.991542
10	(19.824, 21.706]	0.194413	25.314110
11	(16.059, 17.941]	0.055845	23.863896
12	(23.588, 25.471]	0.360789	27.055360
13	(31.118, 33.0]	0.868608	32.370062
14	(17.941, 19.824]	-0.083612	22.404366
15	(29.235, 31.118]	0.490267	28.410442
16	(12.294, 14.176]	-0.083625	22.404239

Скоры по значениям срока
с момента установки капитала

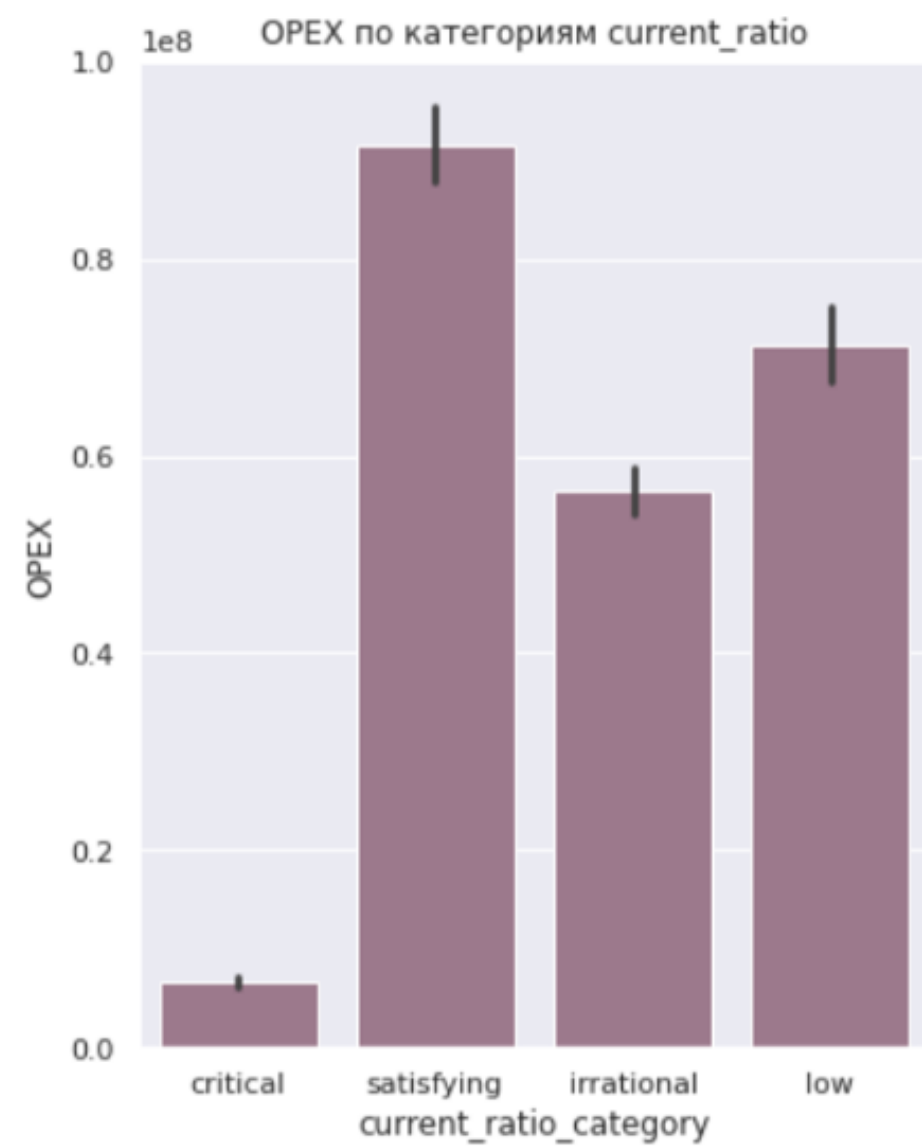
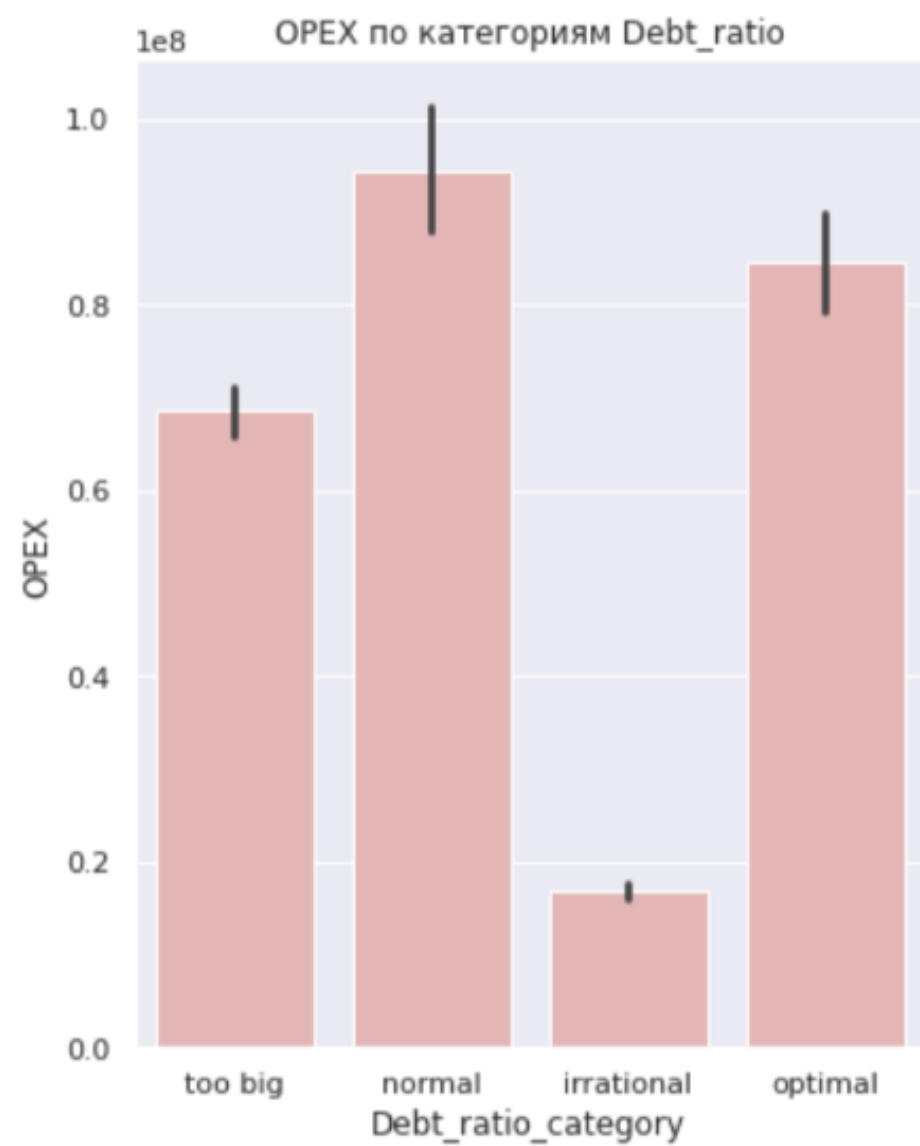
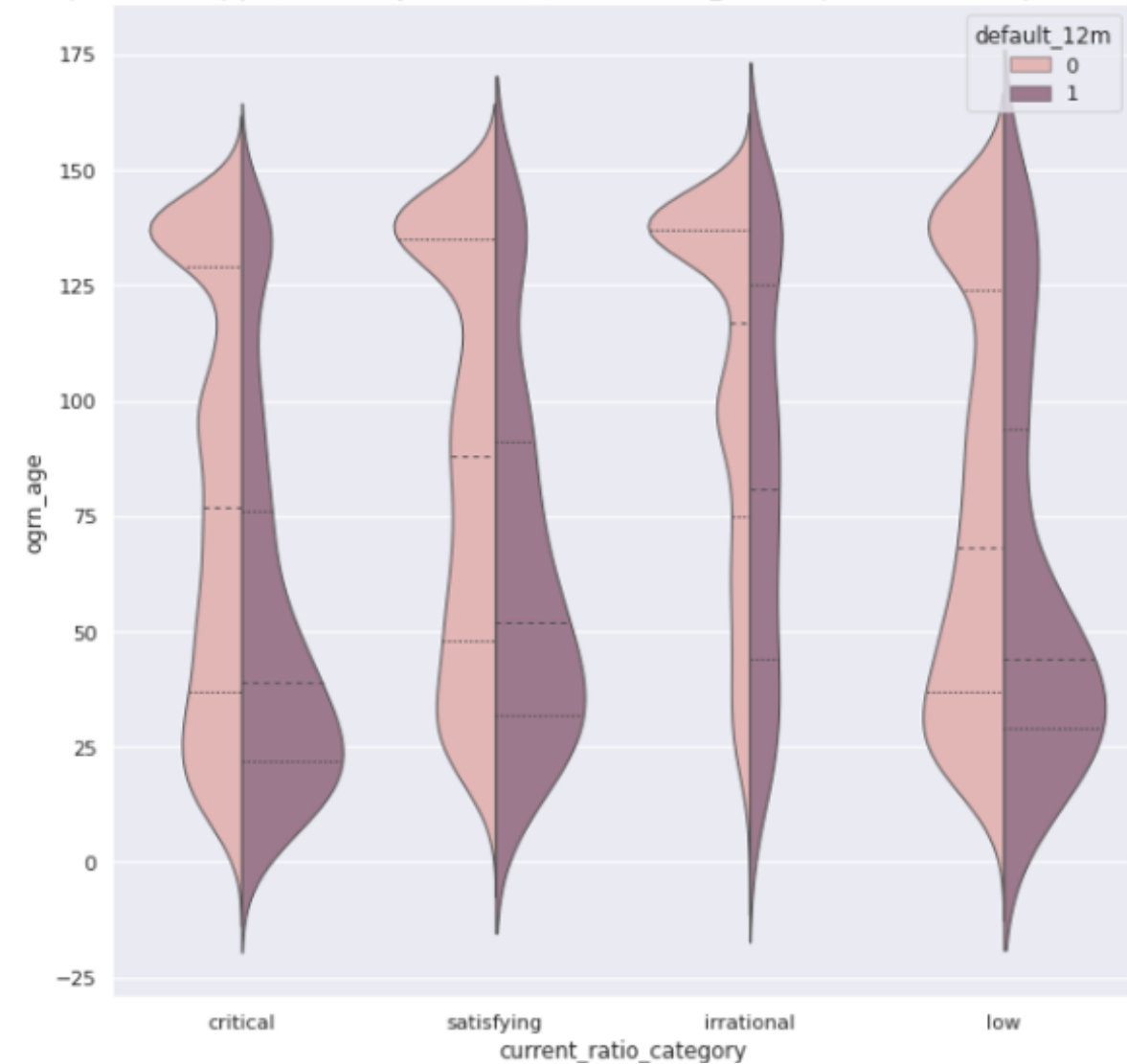
	value	WoE_cut_frac_comer_exp	scoring_points
0	(-0.001, 0.0588]	-0.033621	22.978285
1	(0.0588, 0.118]	1.150354	33.583331
2	(0.824, 0.882]	-2.654286	-0.495411
3	(0.176, 0.235]	0.901062	31.350383
4	(0.118, 0.176]	0.944547	31.739886
5	(0.235, 0.294]	0.564590	28.336550
6	(0.294, 0.353]	1.316006	35.067098
7	(0.471, 0.529]	0.523768	27.970901
8	(0.412, 0.471]	-0.015229	23.143028
9	(0.353, 0.412]	0.086554	24.054711
10	(0.588, 0.647]	0.000000	23.279433
11	(0.882, 0.941]	0.000000	23.279433
12	(0.529, 0.588]	-0.256391	20.982902
13	(0.941, 1.0]	-0.064019	22.706007
14	(0.706, 0.765]	-1.044848	13.920570
15	(0.647, 0.706]	0.000000	23.279433
16	(0.765, 0.824]	0.000000	23.279433

Скоры по значениям
доли коммерческих расходов в
общих расходах

Интерпретация модели

Бизнес-инсайты по данным и рекомендации

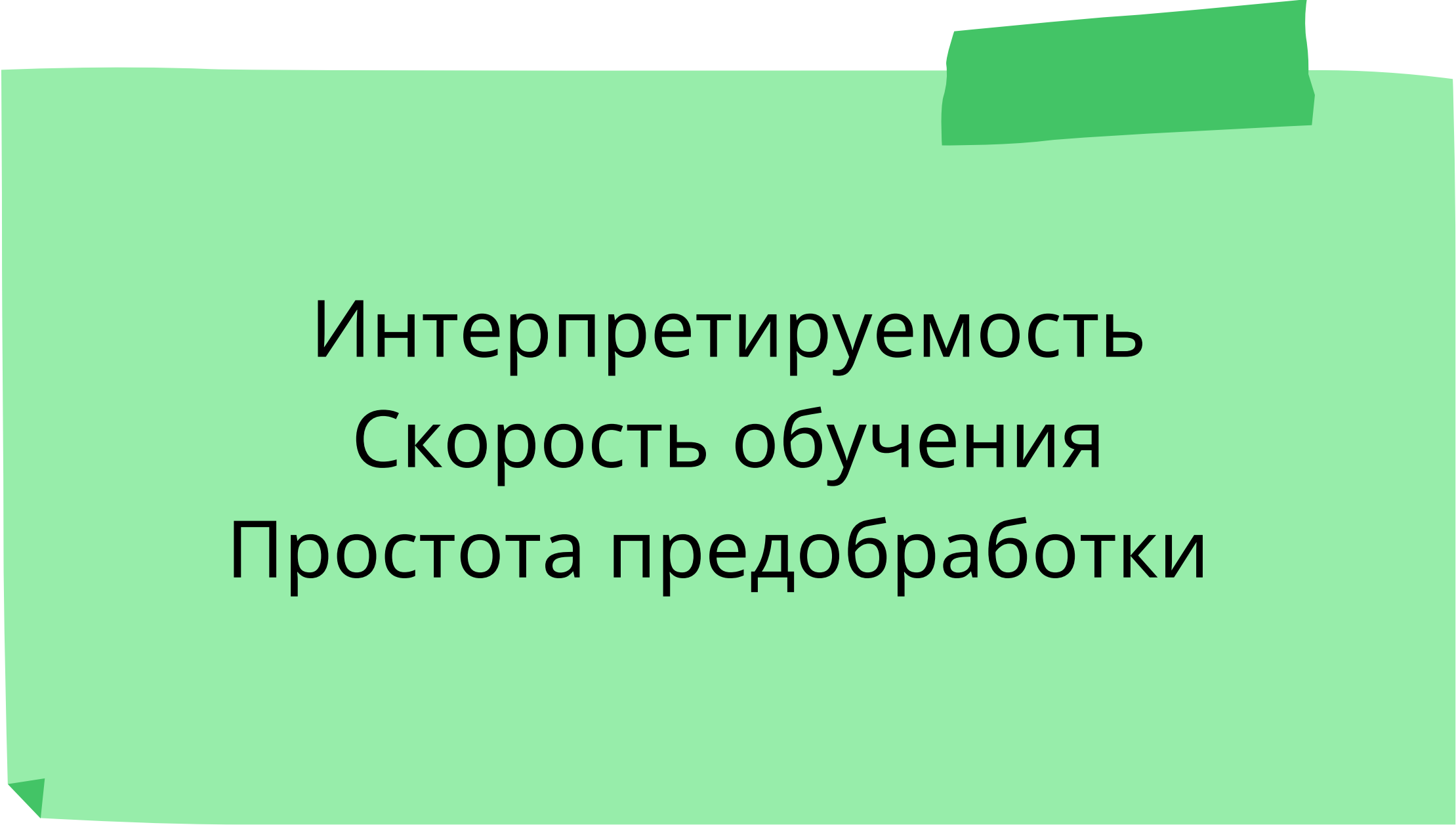
Зависимость дефолта от коэффициента текущей ликвидности (current_ratio) и срока с момента присваивания ОГРН (ogrn_age)



Значение Current Ratio	Норматив
CR < 1.5	Критическая платежеспособность
1.5 < CR <=2	Низкая платежеспособность
2 < CR <=3	Удовлетворительная платежеспособность
3 < CR	Высокая платежеспособность (Нерациональная структура капитала)

Значение Debt Ratio	Норматив
DR <= 0.4	Иррациональная зависимость (упущенные возможности, слишком осторожный подход)
0.4 < DR <=0.6	Оптимальная зависимость
0.6 < DR <=0.75	Нормальная зависимость
0.75 < DR	Слишком высокая зависимость (от кредиторов)

Преимущества модели



Интерпретируемость
Скорость обучения
Простота предобработки

Рекомендации

$$\text{ДПП} = c_1 \cdot \text{Доходность} + c_2 \cdot \text{Порог принятия} + c_3 \cdot \text{Порог согласия}$$

$$\text{Доходность} = \sum_{i=1}^n \left(\frac{\text{Сумма всех платежей по } i\text{—ому кредиту (в руб.)}}{\text{Выданная сумма } i\text{—го кредита (в руб.)}} - 1 \right)$$

$$\text{Порог принятия} = \frac{\text{Количество одобренных заявок на кредит (в шт.)}}{\text{Количество поступивших заявок на кредит (в шт.)}} \cdot \frac{1}{n} \sum_{i=1}^n \text{Сумма выданного кредита}_i$$

$$\text{Порог согласия} = \frac{\text{Количество выданных кредитов (в шт.)}}{\text{Количество одобренных банком кредитов (в шт.)}}$$

c_1, c_2, c_3 – веса, которые зависят от ключевой стратегии конкретного банка, применяющего нашу модель

Команда

MegaQuant



Антон Сметанин



Алина Мусина



Камран Абдулхаев



Петяева Елизавета



Екатерина Филимошина



Ангелина Юдина