Факультет экономических наук

Экономика и анализ данных

Москва 2025

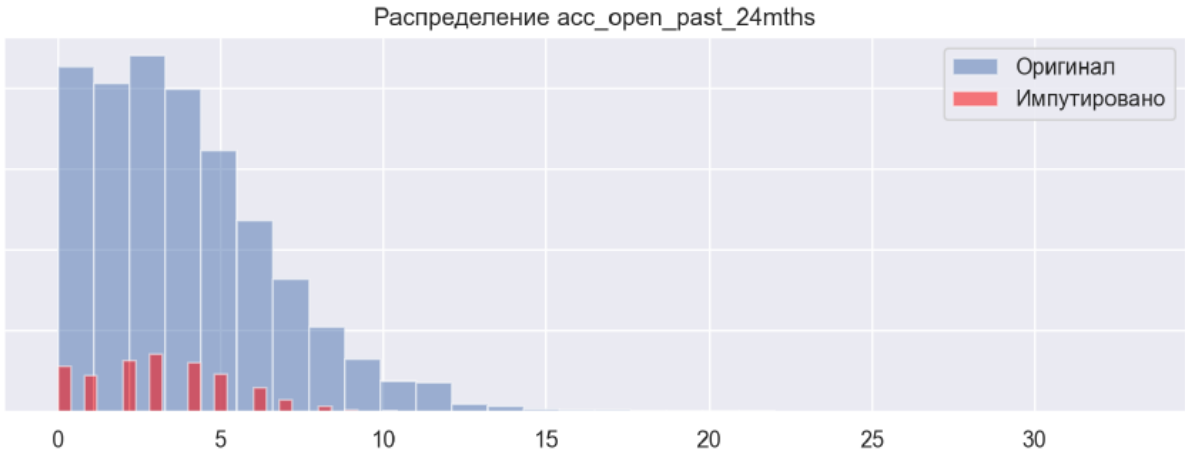
Разработка модели кредитного скоринга физических лиц.

Андреев Иван Васильевич БЭАД223

Научный руководитель:
Васильева Наталья Васильевна

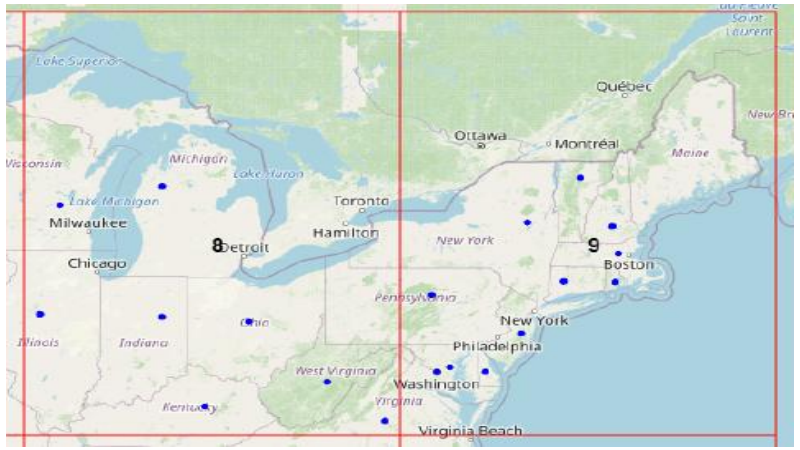
Пропущенные значения.

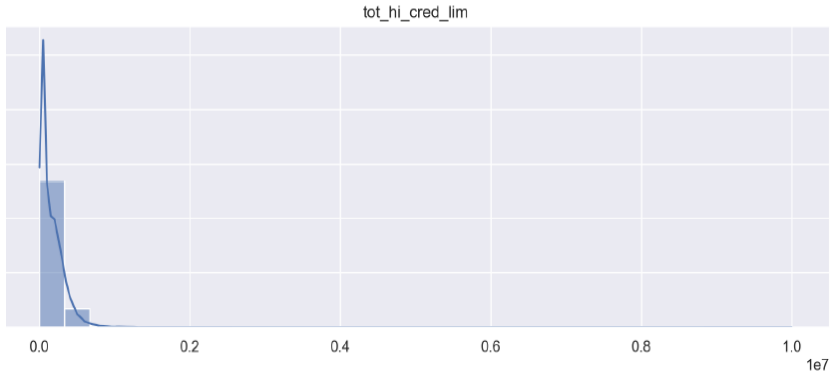
1. Принадлежность пропусков категории:
   - MCAR;
   - MAR;
   - MNAR
2. Создание дамми-переменных.
3. EM-алгоритм.

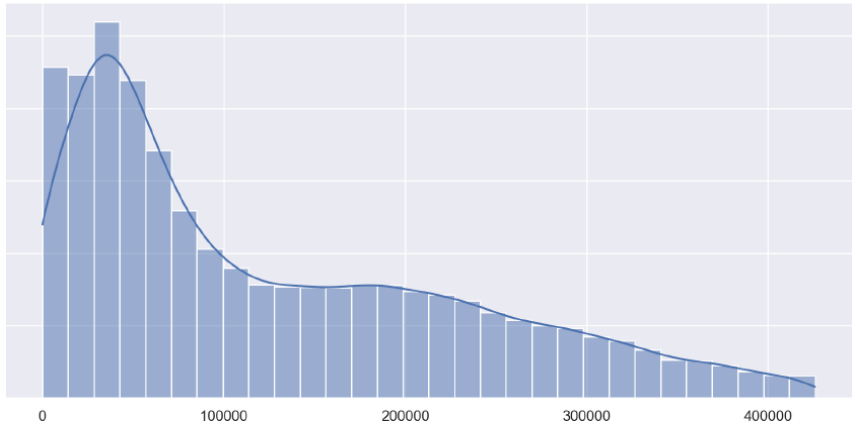


Распределения.

1. Уникальные значения:
   - Геокодирование;
   - Кластеризация;
   - GigaChat.
2. Целевая переменная.
3. Числовые признаки.
4. Категориальные признаки.
5. Исключение выбросов.



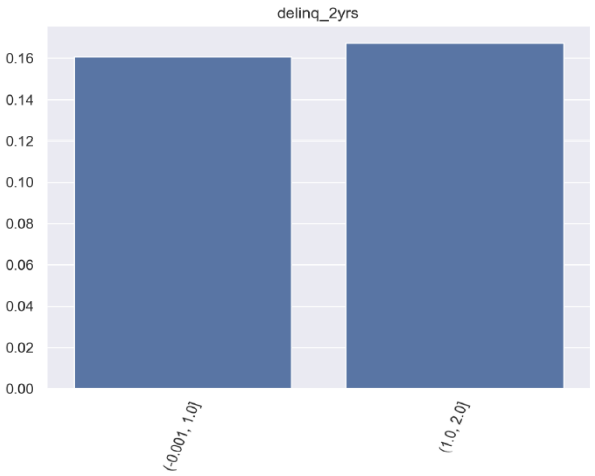


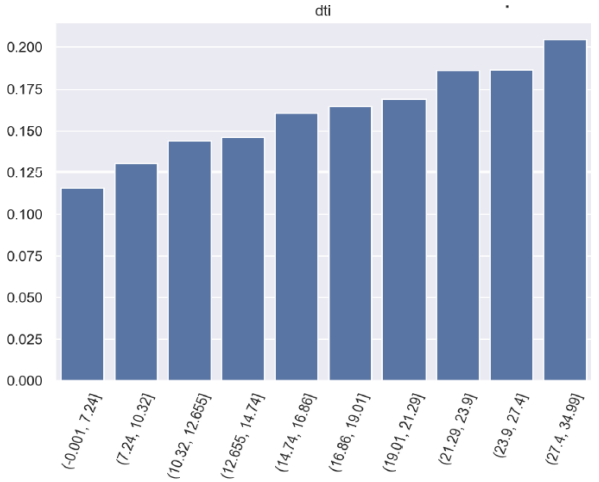


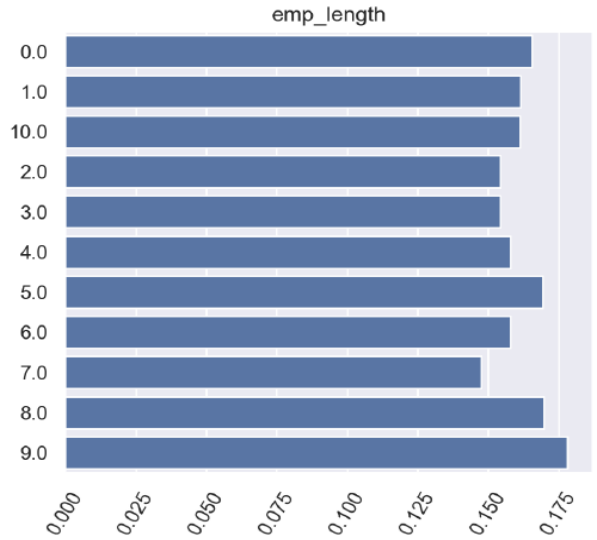
Зависимости.

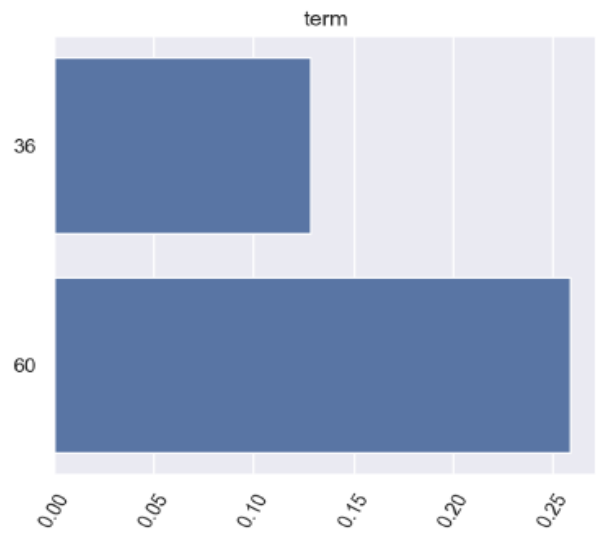
1. Default Rate во времени.
2. Числовые признаки.
3. Категориальные признаки.

$$\mathrm{DF}_{\mathrm{bin}_k} = \frac{1}{n}\sum_{i,j=1}^{n} 1\left[(\mathrm{y}_i = 1) \bigvee (x_{ij} \in \mathrm{bin}_k)\right]$$





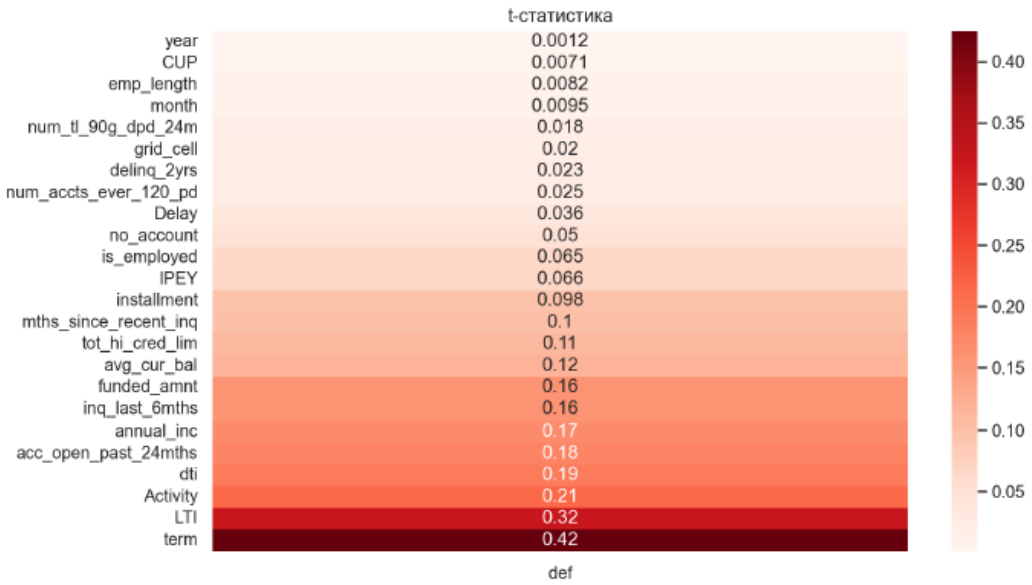


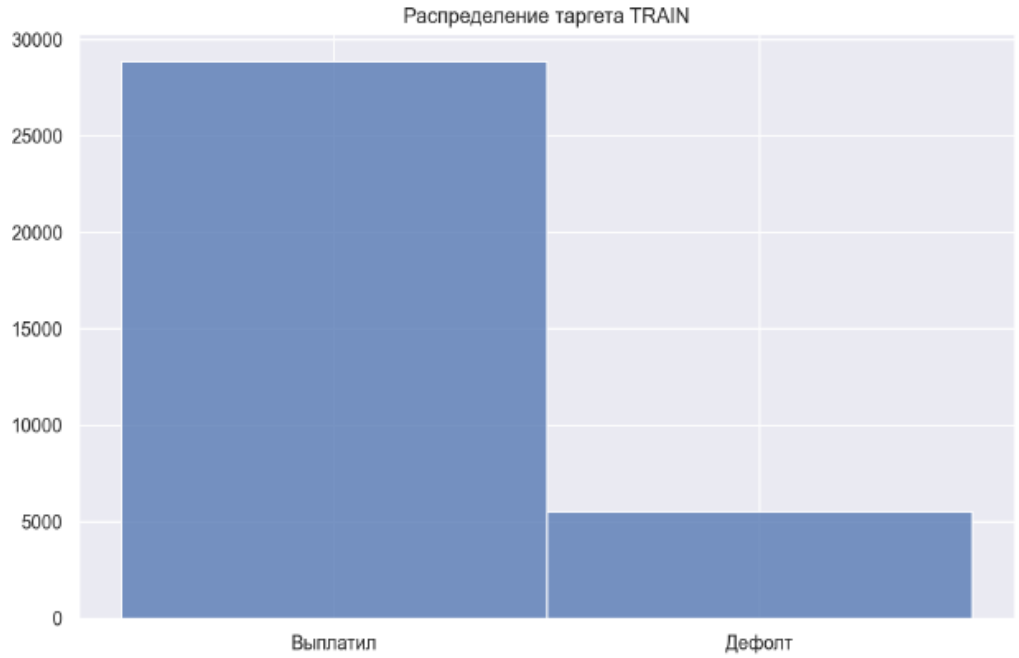


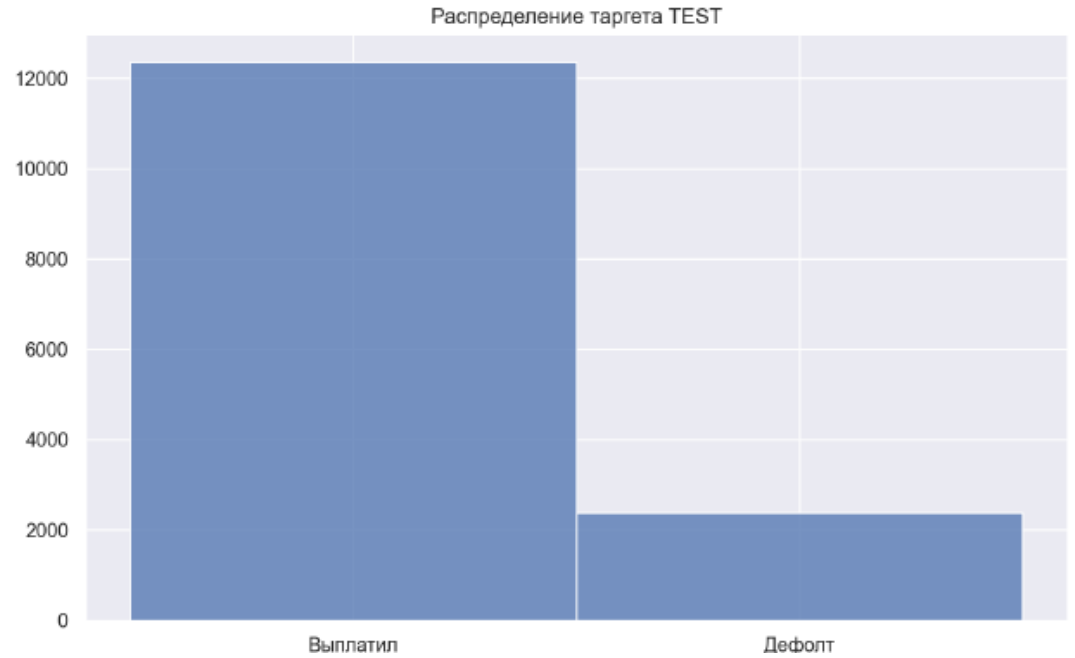
Feature Engineering.

1. Новые переменные:
   - $\text{LTI} = \frac{\text{funded\_amnt}}{\text{anual\_inc}}$
   - $\text{CUP} = \frac{\text{avg\_cur\_bal}}{\text{tot\_hi\_cred\_lim}}$
   - $\text{IPEY} = \frac{\text{anual\_inc}}{\text{emp\_length}}$
   - $\text{Activity} = \text{acc} + \text{open\_past\_24mths} + \text{inq\_last\_6mths}$
   - $\text{Delay} = 1[(\text{num\_accts\_ever\_120\_pd} > 0) \vee (\text{num\_tl\_90g\_dpg\_24m} > 0) \vee (\text{delinq\_2yrs} > 0)]$
2. Описательные статистики.
3. VIF.
4. Корр. матрица.

Признак	$VIF = \frac{1}{1 - R_j^2}$
Activity	inf
inq_last_6mths	inf
is_employed	11 128 327
prof_group_mapped_No Job	769 132







Train / Test split.

1. Случайное разделение.
2. Проверка Default Rate.

Отбор признаков.

1. WOE-преобразование:

$$\mathrm{WOE}_i = \log\left(\frac{\mathrm{GoodRate}_i}{\mathrm{BadRate}_i}\right)$$

2. Information Value:

$$IV = \sum_{i=1}^{n} (\mathrm{BadRate}_i - GoodRate_i) \times \mathrm{WOE}_i$$

Признак	***IV***
sub_grade	0.3152
term	0.1551
dti	0.0396

Значение *IV*	**Интерпретация**
$IV < 0.01$	Нет предсказательной силы
$0.01 \le IV < 0.1$	Слабая предсказательная сила
$0.1 \le IV < 0.3$	Средняя предсказательная сила
$0.3 \le IV < 0.5$	Высокая предсказательная сила
$IV \ge 0.5$	Слишком высокая сила (возможна утечка)

Логистическая регрессия.

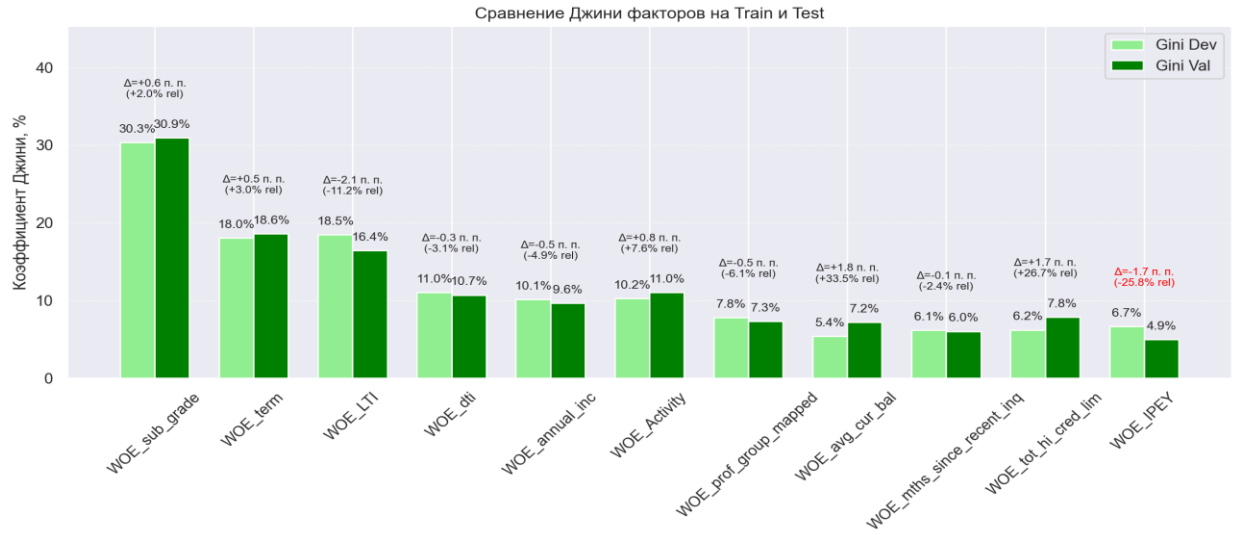
1. Модель:

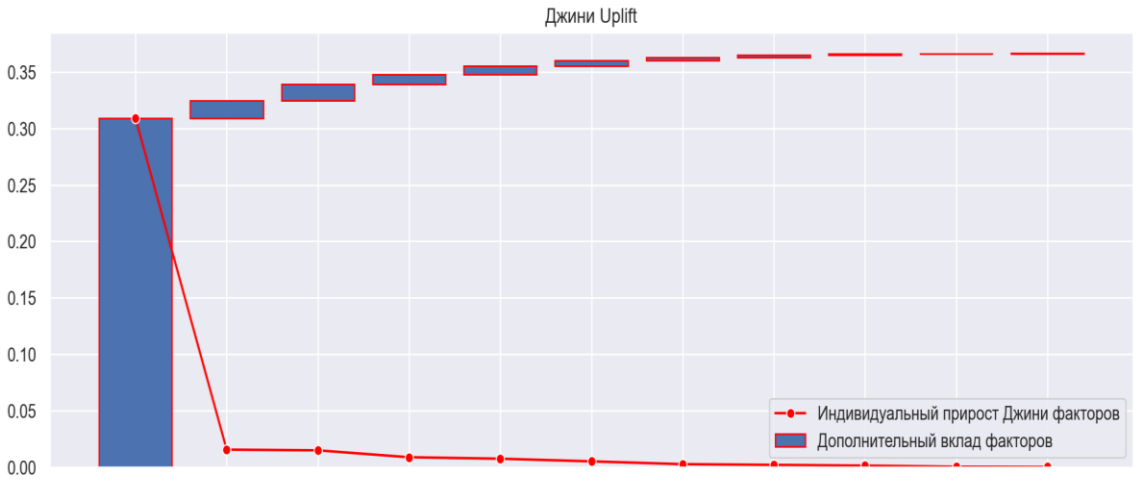
$$P(y = 1 \mid x) = \sigma(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

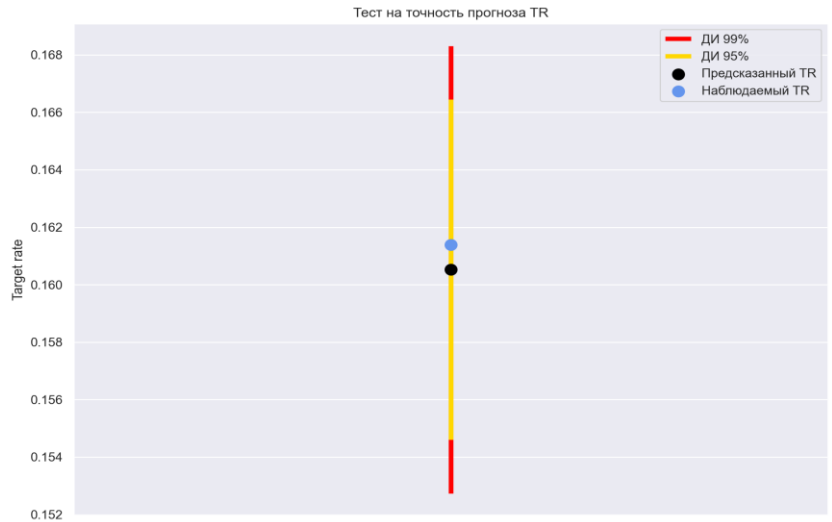
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

2. Подбор гиперпараметров и обучение.
3. Валидация.

Gini = 0.3664



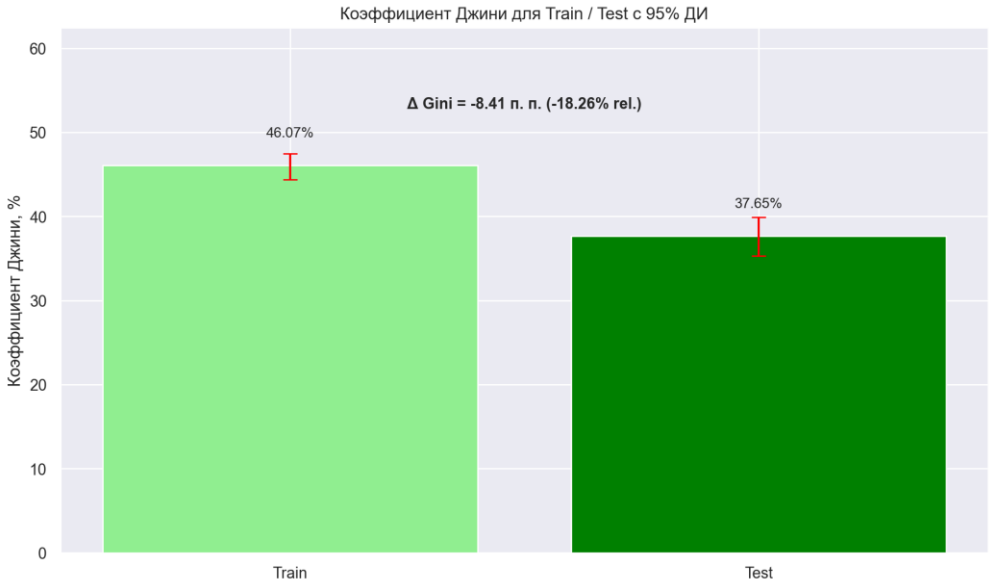




Неинтерпретируемая модель.

1. Модели:
   - CatBoost;
   - RandomForest;
   - SVM с RBF-ядром;
2. MeanTargetEncoder + StandartScaler.
3. Подбор гиперпараметров.
4. Выбор лучшей модели.
5. Валидация.

Модель	Gini
CatBoost	**0.3765**
RandomForest	0.3698
SVM-RBF	0.3521



Сравнение моделей.

Качество

Бустинг дает значимый прирост коэффициента Джини в сравнении с логистической регрессией.

Важность признаков

Бустинг с MeanTargetEncoder’ом извлек больше информации из категориальных переменных в сравнении с LR на WOE-преобразованиях.

Динамика Джини

Бустинг дает более стабильный во времени Джини модели в сравнении с логистической регрессией.

Вероятности

Бустинг хуже приближает вероятности в сравнении с логистической регрессией.

Переобучение

Бустинг переобучается сильнее в сравнении с логистической регрессией.

Прибыль.

1. Теоретический порог отсечения:

$$rT_i - p_i(rT_i + 1) \geq 0 \rightarrow p_i \leq \frac{rT_i}{rT_i + 1}$$

2. Оптимальный порог ожидаемой прибыли:

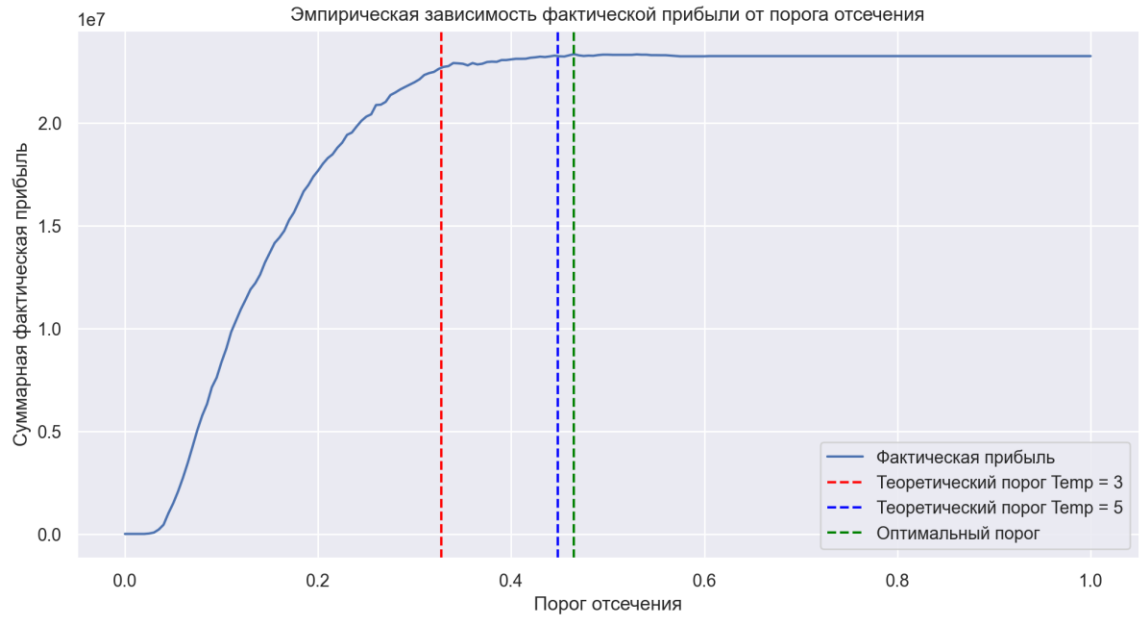
$$t_e^* = argmax_{t_j}\left(\sum_{i:p_i<t_j} F_i(0.13T_i - p_i(0.13T_i + 1))\right)$$

3. Оптимальный порог фактической прибыли:

$$t_r^* = argmax_{t_j}\left(\sum_{i:p_i<t_j} [(1 - y_i)F_iT_ir - y_iF_i]\right)$$

4. LGD = 80%.

LGD	t_e^*	$\pi_e(t_e^*)$	t_r^*	$\pi_r(t_r^*)$
100%	0.395	45 377 168.42$	0.465	44 277 073$
80%	0.45	52 033 888.14$	0.465	51 215 868$



Скоринговая карта.

Признак	Бин / Категория	WOE	Скор
Intercept			48
sub_grade	A	1.05	22
sub_grade	G	-0.87	-18
term	36	0.26	4
term	60	-0.6	-9
LTI	< 0.1225	0.46	6
LTI	> 0.3484	-0.5	-6

Скоринговая карта.

Признак	Бин / Категория	WOE	Скор
dti	< 10.345	0.34	3
dti	> 25.455	-0.24	-3
annual_inc	< 35 528	-0.26	-6
annual_inc	> 102 673.5	0.41	9
prof_group_mapped	IT & Telecommunications	0.25	7
prof_group_mapped	No Job	-0.26	-7
avg_cur_bal	< 3 339.5	-0.1	-1
avg_cur_bal	> 14 983.5	0.17	1

Скоринговая карта.

Признак	Бин / Категория	WOE	Скор
Activity	0	-0.11	-2
Activity	2	0.32	7
Activity	8	-0.3	-7
mths_since_recent_inq	< 4.5	-0.1; -0.14	-1
mths_since_recent_inq	> 7.5	0.08; 0.16	1
tot_hi_cred_lim	< 25 096.5	-0.14	-2
tot_hi_cred_lim	> 248 908.5	0.21	2


ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •