

ПОСТРОЕНИЕ МОДЕЛИ КРЕДИТНОГО СКОРИНГА

Студентки: Семавина Юлия,
Шарф Мира и Оронова Софья

ЦЕЛЬ НАШЕЙ РАБОТЫ

Научиться предсказывать,
вернет ли клиент кредит, на
основе данных Home Credit
Group

АКТУАЛЬНОСТЬ



ПЛАН ПРЕЗЕНТАЦИИ

01

Обзор используемых данных

03

Модель кредитного scoringа

02

Разведочный анализ данных + WOE-IV

04

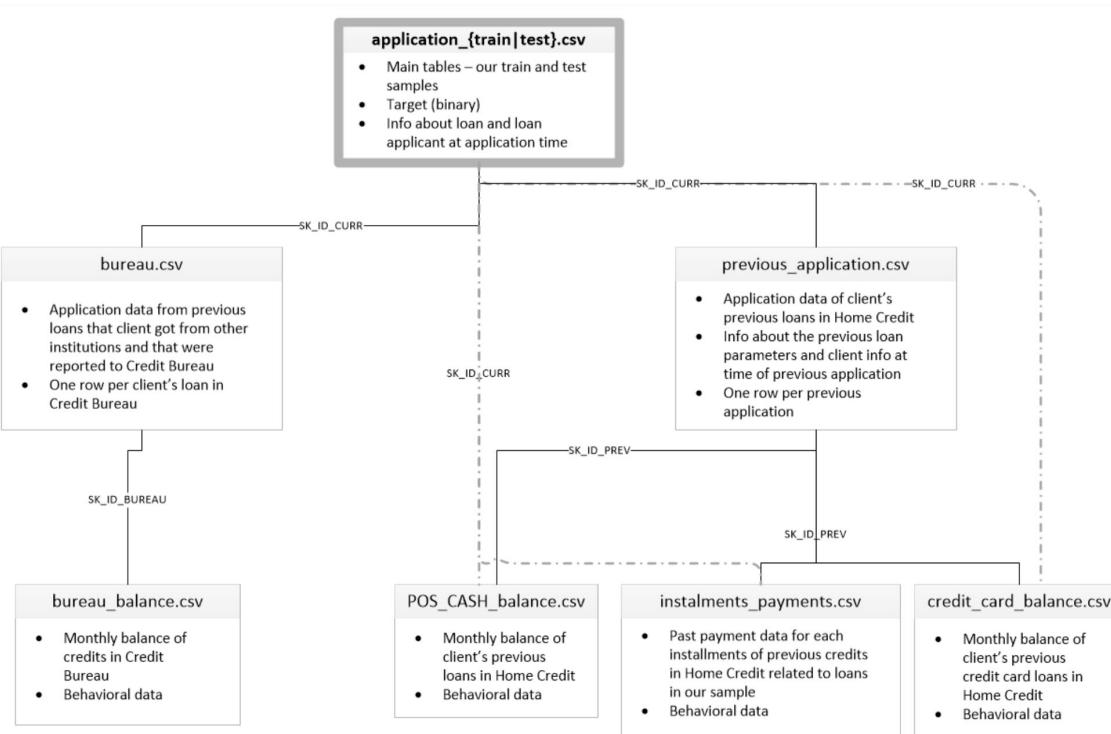
Результаты



ДАННЫЕ

01

ОПИСАНИЕ ДАННЫХ



ОБЪЕКТ ВЫБОРКИ

Заявка на кредит и информация о заявителе на момент заявки в Home Credit Group, с уникальным идентификатором **SK_ID_CURR**

ЦЕЛЕВАЯ ПЕРЕМЕННАЯ

TARGET:

- 1 - клиент с дефолтами
- 0 - остальные случаи

ОПИСАНИЕ ТАБЛИЦ

bureau

- Информация о предыдущих кредитах клиентов
- Размер 162.14 MB
- 17 столбцов, 1'716'428 строк (в среднем по 5.6 на клиента)
- Информация о валюте, задолженностях, статусе, ...

application

- Основная таблица - одна строка соответствует одной заявке
- Размер: 158.44 MB
- 122 столбца, 307'511 строк
- Общая информация о пользователе, его заявке, месте проживания, предоставленных документах
- Информация о его оценках (=“скорах”) из внешних источников

previous_application

- Информация о предыдущих заявках в Home Credit
- Размер 386.21 MB
- 37 столбцов, 1'670'214 строк (в среднем по 5.4 на клиента)
- Информация о процентной ставке, первоначальном взносе, размере кредита, ...

ОПИСАНИЕ ТАБЛИЦ

POS_CASH_balance

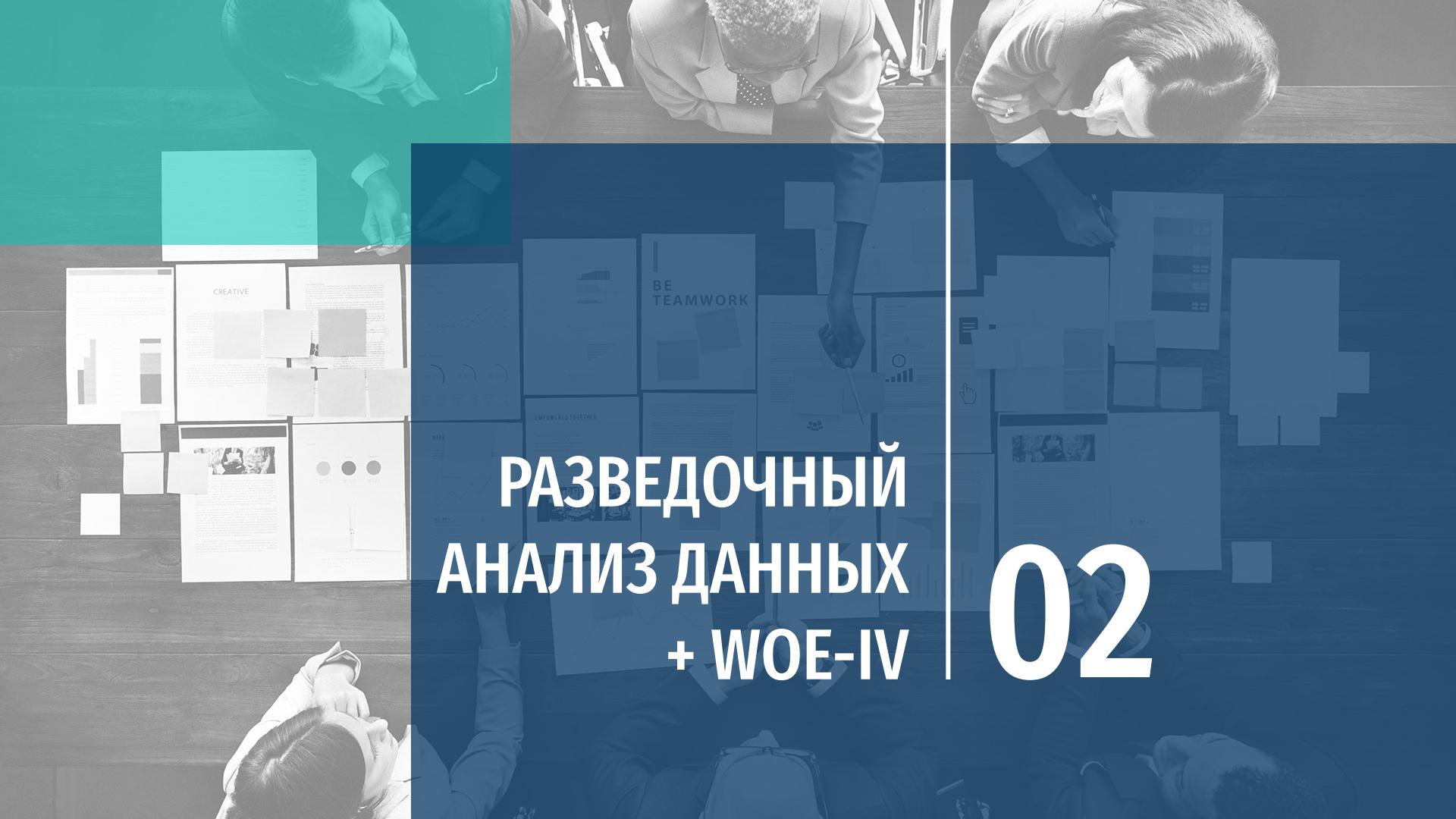
- Балансы по предыдущим заемам типа "кредит наличными" и "рассрочка"
- Размер: 374.51 MB
- 8 столбцов, 10'001'358 строк (в среднем 32.5 на клиента)

installments_payments

- Информация о предыдущих платежах клиентов в Home Credit Group
- Размер 689.62 MB
- 8 столбцов, 13'605'401 строк (в среднем по 44 на клиента)

credit_card_balance

- Ежемесячная информация о балансах кредитных картах клиентов и о поведении клиентов
- Размер 404.91 MB
- 23 столбцов, 3'840'312 строк (в среднем по 12.5 на клиента)
- Информация о балансе, сумме снятий, ...



РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ + WOE-IV

02

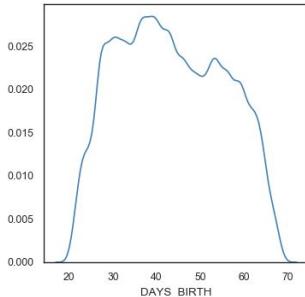
ОБЩАЯ ИНФОРМАЦИЯ О ЗАЯВКАХ

34% ♂ **66%** ♀

Распределение заявок по полу

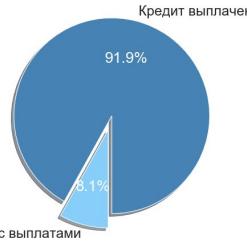
20 - 69

Возраст



\$ 599'026,00

Средняя сумма кредита



Распределение целевой переменной

ПРОПУСКИ/ОТСУТСТВИЕ ИСТОРИИ

55% 

Столбцы с пропусками (application)

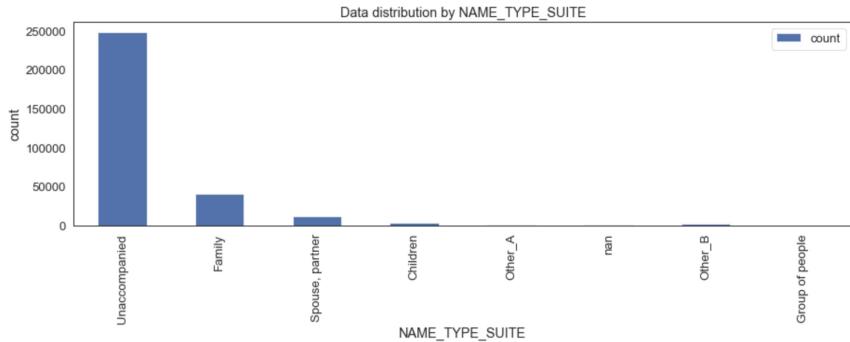
\$ 33,6%

Столбцы с ≥50% пропусков (application)

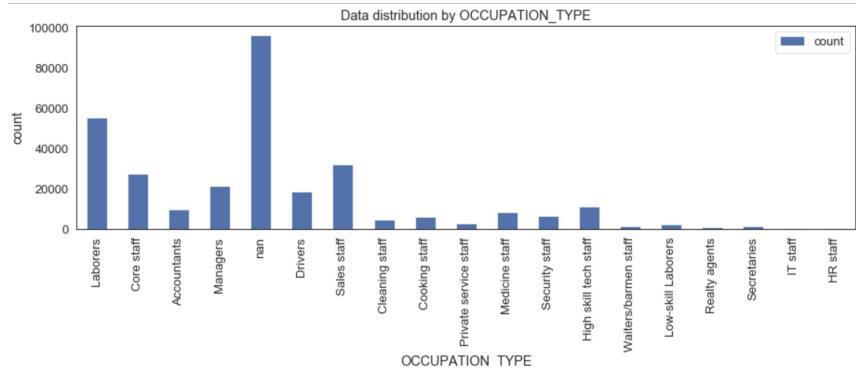
Таблица	Число отсутствующих заявок	Процент отсутствующих заявок от общего числа заявок
credit_card_balance	220606	71.0
bureau	44020	14.0
pos_cash	18067	5.0
installments_payment	15868	5.0
previous_application	16454	5.0

Отсутствие заявок в справочниках

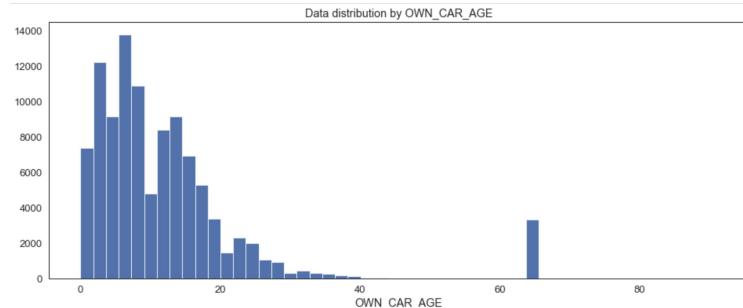
ОСОБЕННОСТИ ДАННЫХ



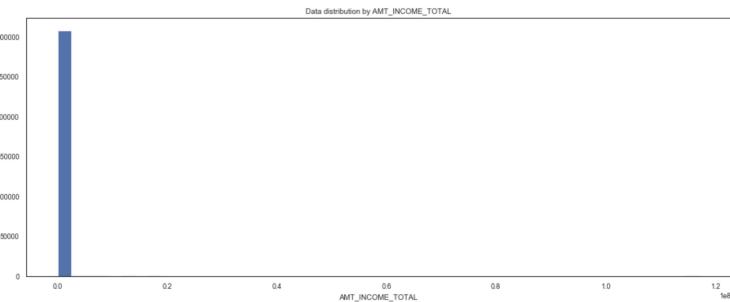
Категориальные признаки с недостаточным числом объектов в классах



Несбалансированные классы

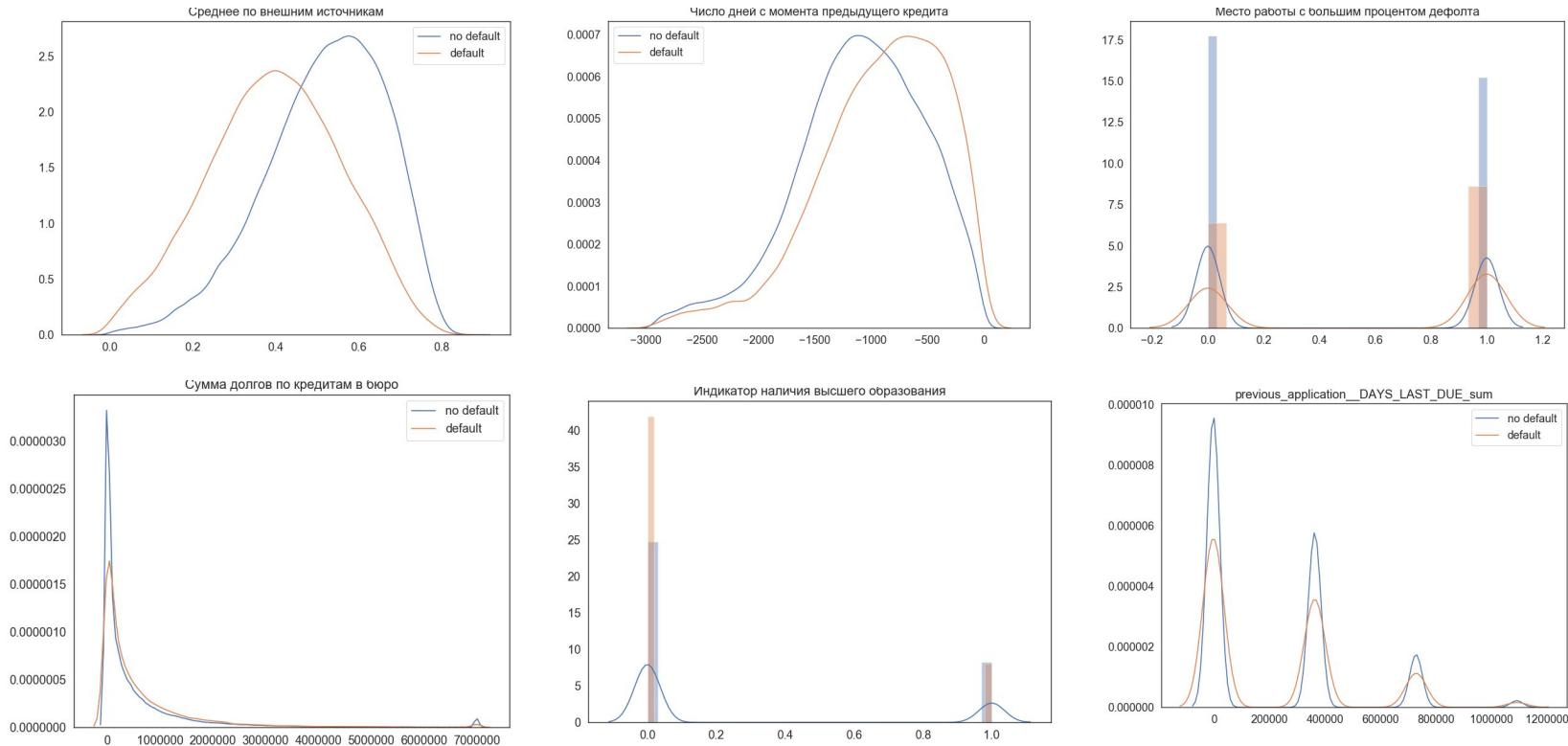


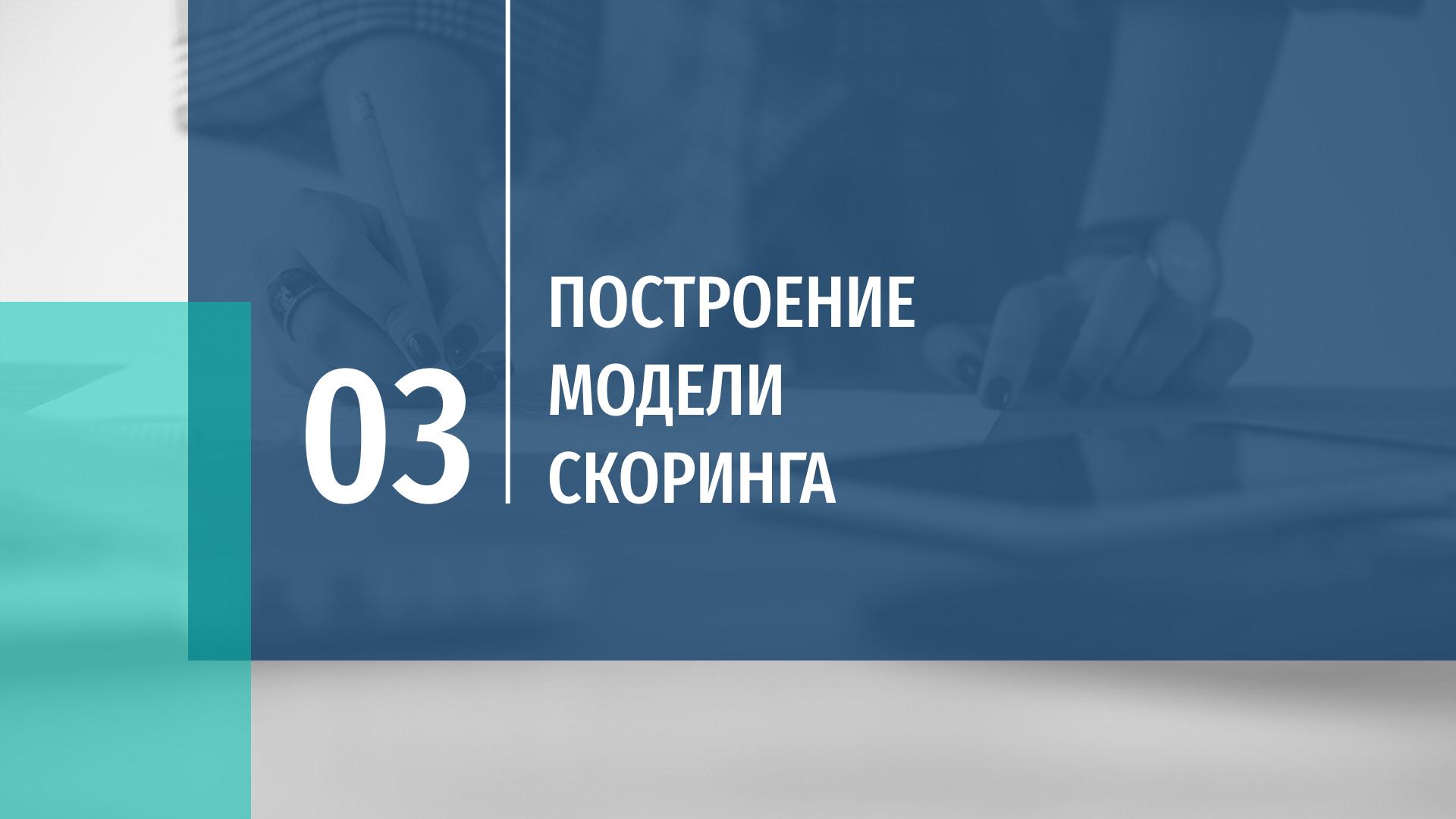
Странные распределения



Выбросы

WOE АНАЛИЗ: САМЫЕ ПОЛЕЗНЫЕ ПРИЗНАКИ ПО IV





03

ПОСТРОЕНИЕ МОДЕЛИ СКОРИНГА

BASELINE: ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ + WOE

	tables	train_auc_roc	test_auc_roc	n_features
application+bureau+credit_card_balance+pos_cash		0.733942	0.733056	155
application+bureau+credit_card_balance+pos_cash_installments_payment+previous_application		0.730945	0.731665	252
application+bureau		0.731111	0.726421	127
application+bureau+pos_cash		0.731429	0.726180	142
application+bureau+credit_card_balance+pos_cash_installments_payment		0.731106	0.725705	186
application+bureau+installments_payment		0.724307	0.722969	158
application+bureau+credit_card_balance		0.722718	0.721913	140
application		0.734995	0.718749	81
application+pos_cash		0.728257	0.718506	96
application+credit_card_balance		0.710881	0.717056	94
application+installments_payment		0.715072	0.709649	112
application+previous_application		0.710867	0.709636	147
credit_card_balance		0.459850	0.625869	13
bureau		0.677262	0.542534	46
previous_application		0.626521	0.459197	66
installments_payment		0.459850	0.459197	31
pos_cash		0.000000	0.000000	15

СЛОЖНОСТИ С ОБУЧАЮЩЕЙ ВЫБОРКОЙ

- Сильно несбалансированная целевая переменная
- Сильно несбалансированные классы признаков
- Отсутствие линейных зависимостей
- Большое число пропусков в данных
- Большое число признаков
- Слабые корреляции отдельных признаков с целевой переменной

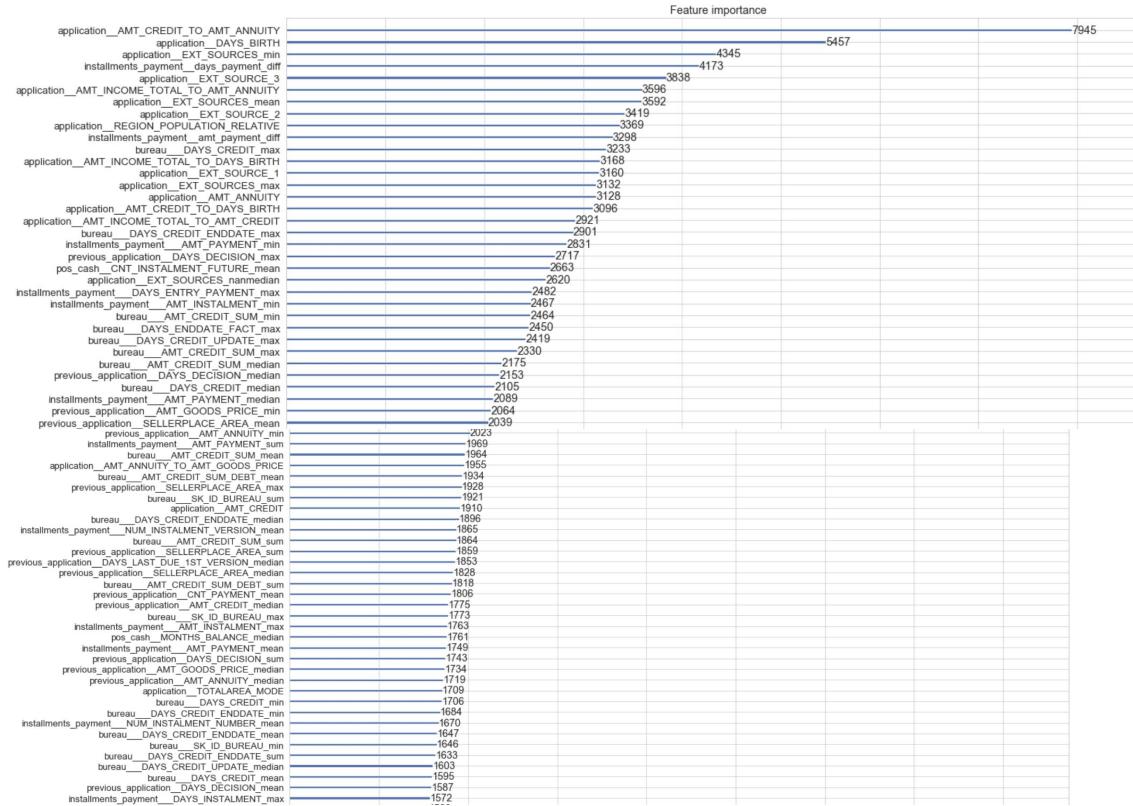
ГРАДИЕНТНЫЙ БУСТИНГ НА ДЕРЕВЬЯХ

- Устойчивость к шуму и к выбросам
- Способен находить сложные зависимости
- Умеет работать с пропусками в данных
- Особенно хорошо работает с большими выборками
- Очень сильный алгоритм
- Достаточно прост в использовании

ГРАДИЕНТНЫЙ БУСТИНГ + KFOLDS

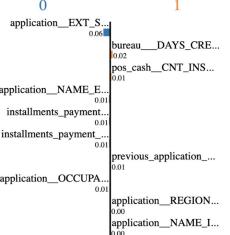
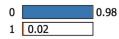
	tables	train_auc_roc	test_auc_roc	n_features
application+bureau+credit_card_balance+pos_cash_installments_payment+previous_application	0.781473	0.782800	252	
application+bureau+credit_card_balance+pos_cash_installments_payment	0.780183	0.781648	186	
application+bureau+installments_payment	0.775998	0.778639	158	
application+bureau+credit_card_balance+pos_cash	0.774154	0.777690	155	
application+bureau+pos_cash	0.771315	0.775580	142	
application+bureau+credit_card_balance	0.767542	0.772913	140	
application+installments_payment	0.771149	0.772334	112	
application+previous_application	0.768845	0.771938	147	
application+bureau	0.764195	0.770307	127	
application+pos_cash	0.765973	0.768128	96	
application+credit_card_balance	0.760119	0.766077	94	
application	0.756867	0.762634	81	
bureau	0.646444	0.653117	46	
previous_application	0.635226	0.641317	66	
installments_payment	0.635428	0.633367	31	
pos_cash	0.585440	0.586427	15	
credit_card_balance	0.547007	0.548121	13	

FEATURE IMPORTANCE

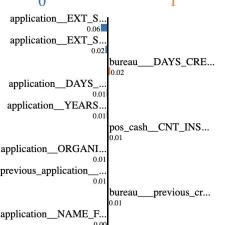
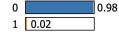


LIME

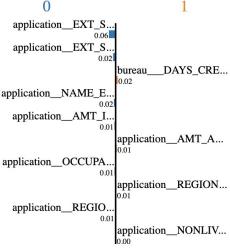
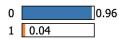
Prediction probabilities



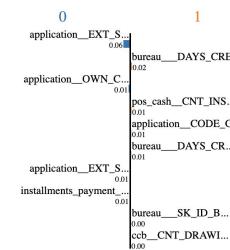
Prediction probabilities



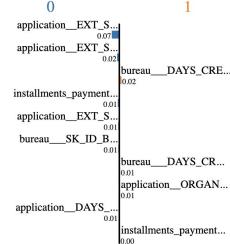
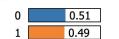
Prediction probabilities



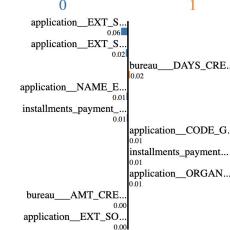
Prediction probabilities



Prediction probabilities



Prediction probabilities





РЕЗУЛЬТАТЫ

04

ТАБЛИЦЫ



application



bureau



previous application



installments_payments



POS_CASH_balance



credit_card_balance

ПОСТРОЕНИЕ МОДЕЛИ



РЕЗУЛЬТАТ В СОРЕВНОВАНИИ KAGGLE “Home Credit Default Risk”



0.78588

СПАСИБО

Вопросы?

WOE АНАЛИЗ


$$S_{True}(f, v) = \frac{\sum_{x \in data} [x_f=v \text{ and } target=1]}{\sum_{x \in data} [target=1]}$$


$$S_{False}(f, v) = \frac{\sum_{x \in data} [x_f=v \text{ and } target=0]}{\sum_{x \in data} [target=0]}$$


$$woe((f, v)) = \ln \left(\frac{S_{True}(f, v)}{S_{False}(f, v)} \right)$$


$$IV((f, v)) = (S_{True}(f, v) - S_{False}(f, v)) \cdot woe((f, v))$$


$$IV(f) = \sum_v IV((f, v))$$

WOE АНАЛИЗ

Information Value	Variable Predictiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive Power
0.1 to 0.3	Medium predictive Power
0.3 to 0.5	Strong predictive Power
>0.5	Suspicious Predictive Power

КОЭФИЦИЕНТ КОРРЕЛЯЦИИ ПИРСОНА



$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

.00-.19	“very weak”
.20-.39	“weak”
.40-.59	“moderate”
.60-.79	“strong”
.80-1.0	“very strong”