

Решение Андеррайтера

Быстрое одобрение

Цель – найти диапазоны

- Найти диапазоны значений рассмотренных признаков, в которых заявки могут быть одобрены с полезной для нас вероятностью.
- Например:
 - Возраст – от 23 до 40
 - Скоринговый бал ОКБ – от 300 до 800

Входные данные

Заявки в диапазоне дат: 2020-01-01 по 2021-10-10

Витрина MART_NORMA_AUTO: 256070 строк, 150 столбцов.

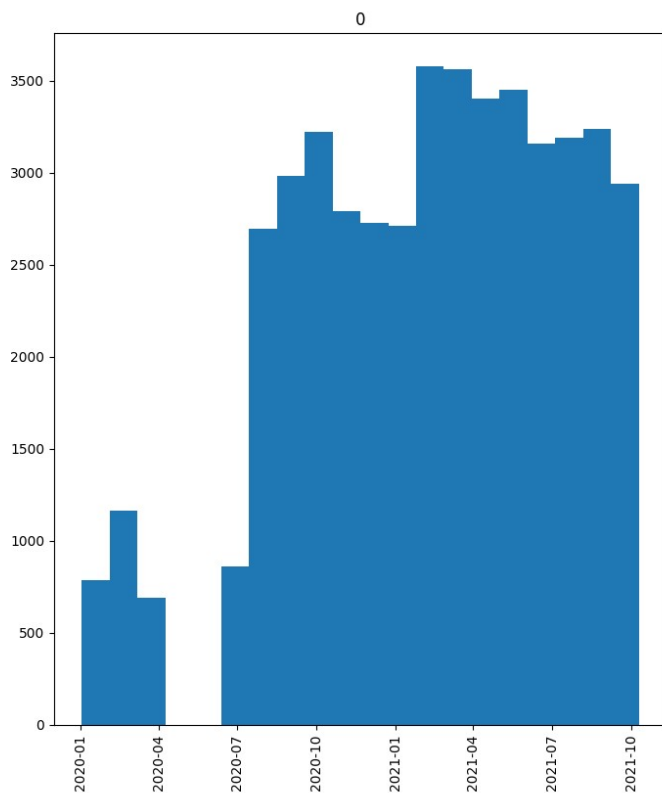
- Одобрено: 23498
- Отклонено: 59710
- Всего: 83208

После обработки

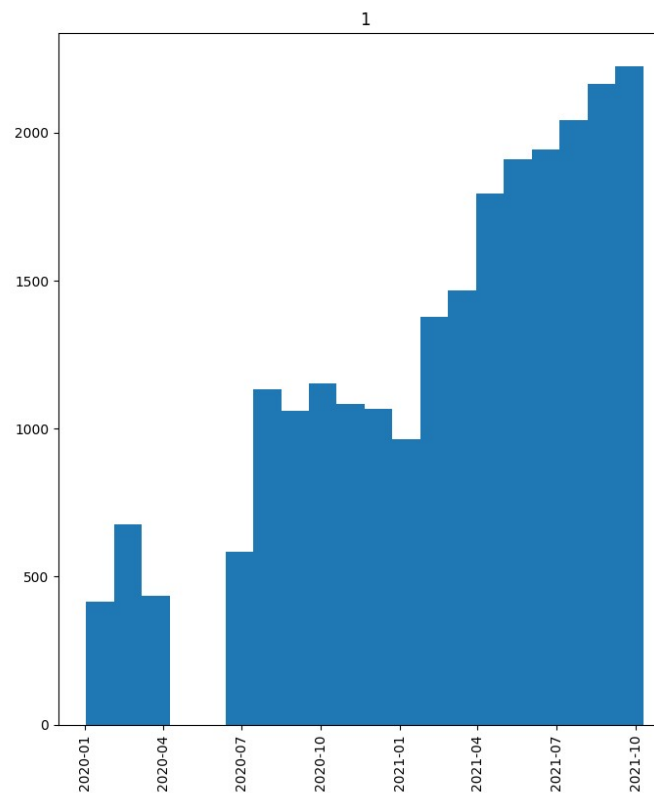
- Всего 81751, столбцов 32 (на основании 13 выбранных признаков)
- Одобрено 23091
- Отклонено 58660

Входные данные — по времени

Отклонено



Одобрено



Входные данные

Статистика:

Одобрено без отмененных пользователем - 22956

Отклонено без отмененных пользователем - 47128

Отменено пользователем и андер.-ом - 12582

Отменено пользователем, одобрено андером - 542

Отменено только пользователем - 3161

Где,

Одобрено без отмененных пользователем: Первое решение андерайтера == 'Одобрено'

Отклонено без отмененных пользователем: Первое решение андерайтера == 'Отмена'

Отменено только пользователем: Наличие кодов отказа — 01 или 091

Принимаем:

Одобренные — «Одобрено без отмененных пользователем» + «Отменено пользователем, одобрено андером»

Отклоненные — «Отклонено без отмененных пользователем»

Пояснения:

Везде считаем, «Заявка дошла до андерайтера» = Верно

Если заявка одобрена Андерайтером и отклонена пользователем, то решение Андерайтера было, иначе, решения не было.

01 - "Отказ клиента"

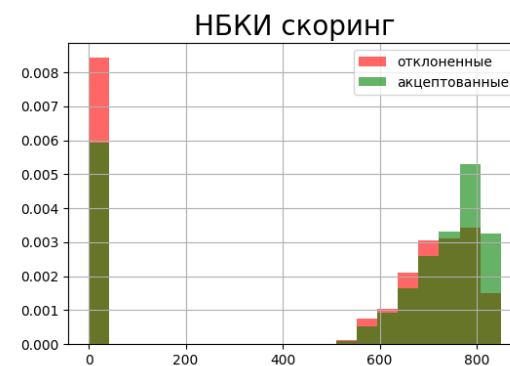
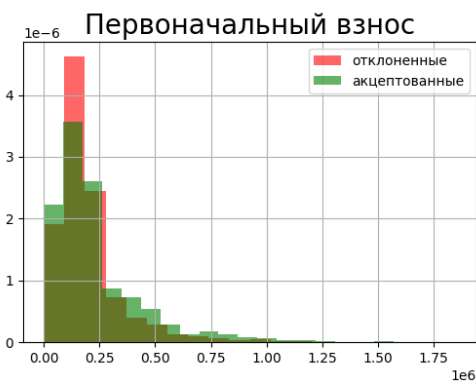
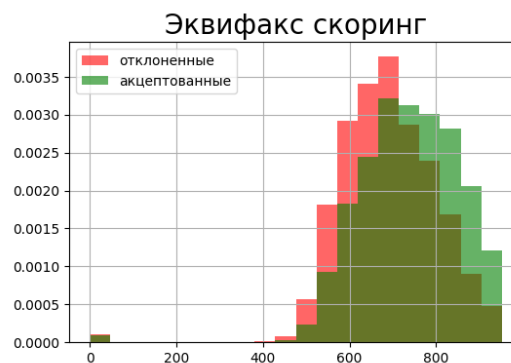
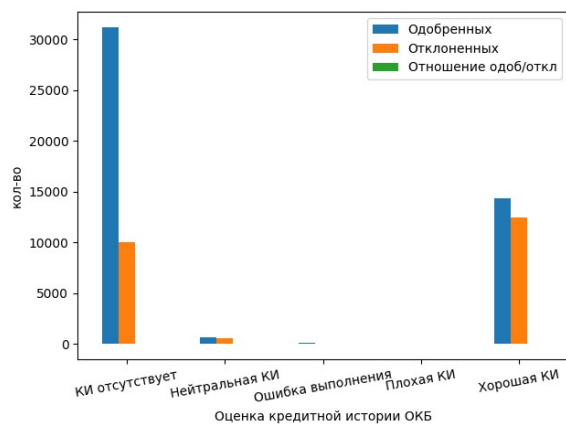
091 - "Отказ клиента от кредита, заявка не актуальна"

Просроченные заявки — код 06, выставляются автоматически, заявки дошедшие до андерайтера проходят его полноценно.

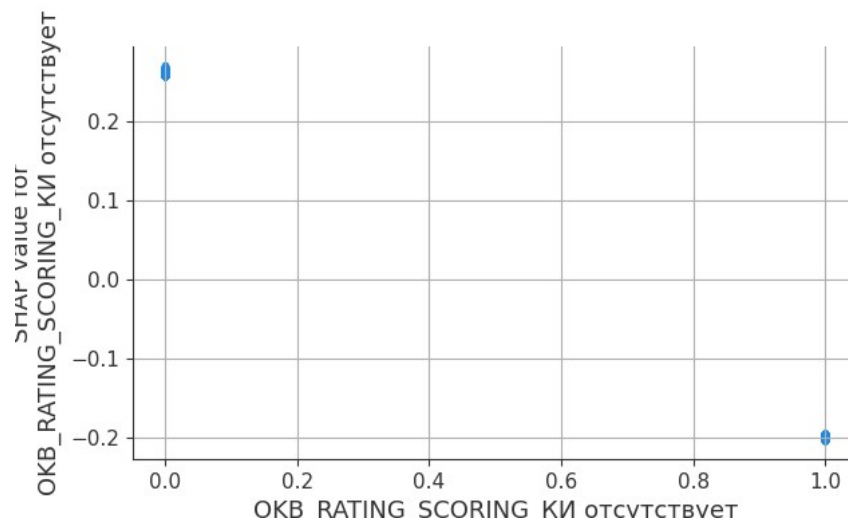
Входные данные

- Возраст клиента на дату выдачи кредитного договора
- Семейное положение
- Кол-во иждивенцев
- Стаж в организации (мес.)
- Скоринговый балл клиента (анкетный скоринг)
- Скоринговый балл ОКБ
- Скоринг Бюро Эквифакс 4Score
- Скоринговый балл НБКИ
- Первоначальный взнос
- Оценка кредитной истории ОКБ
- Оценка кредитной истории НБКИ
- Оценка кредитной истории Эквифакс
- Скоринг МБКИ

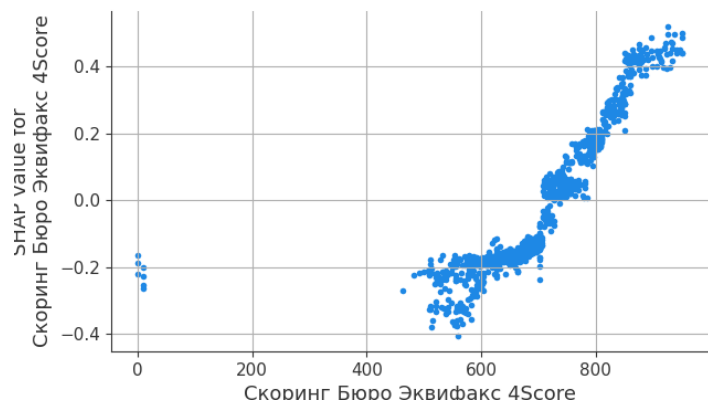
Несколько важных признаков



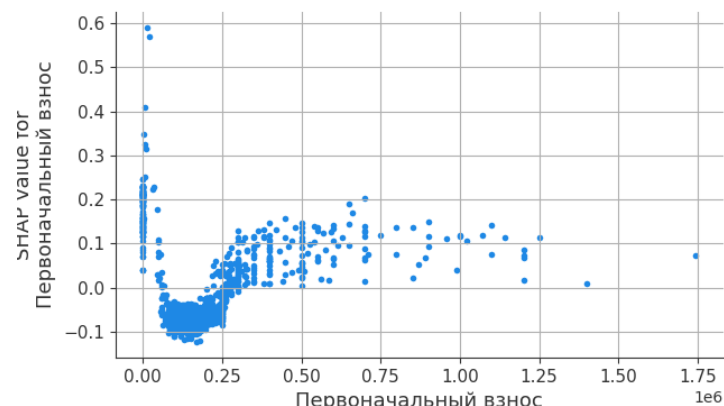
Вес признаков в решении



Если «ОКБ рейтинг» принимает значение КИ отстутствует — то он отрицательно влияет на заявка, если другое значение, то положительно (на всех заяв

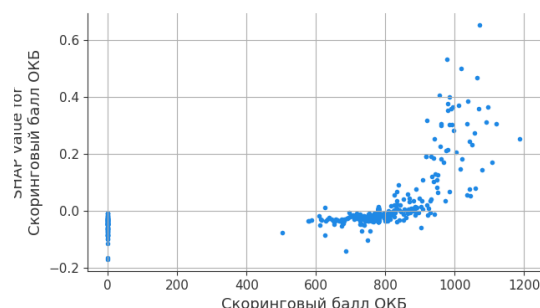


Монотонная зависимость с переломом в 700.

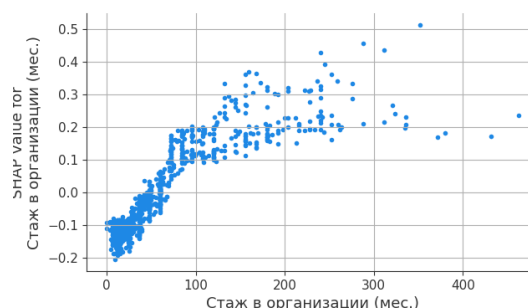


Отрицательно влияет между 50000 и 250000

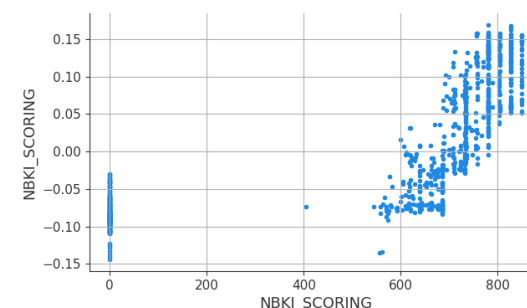
Вес признаков в решении



Положитель
он влияет
после 900.



Монотонно,
с переломом
в 50.



После 800,
точно
положитель
но.

Особые случаи

Особыми случаями являются на дату 2021-10-10:

- МБКИ:
- «Требуется проверка информации о действительности паспорта» - отклонено 19, одобрено 2
- «Недостаточно данных для выполнения проверки» — отклонено 1403, одобрено 375
- «Данные по клиенту не найдены» — отклонено 710, одобрено 94
- «Лица, находящиеся в розыске.» — отклонено 22, одобрено 0
- «Наличие информации о постановке клиента на специальный учет (наркозависимые, алкоголики, психически больные и т.д.).» — отклонено 22, одобрено 0

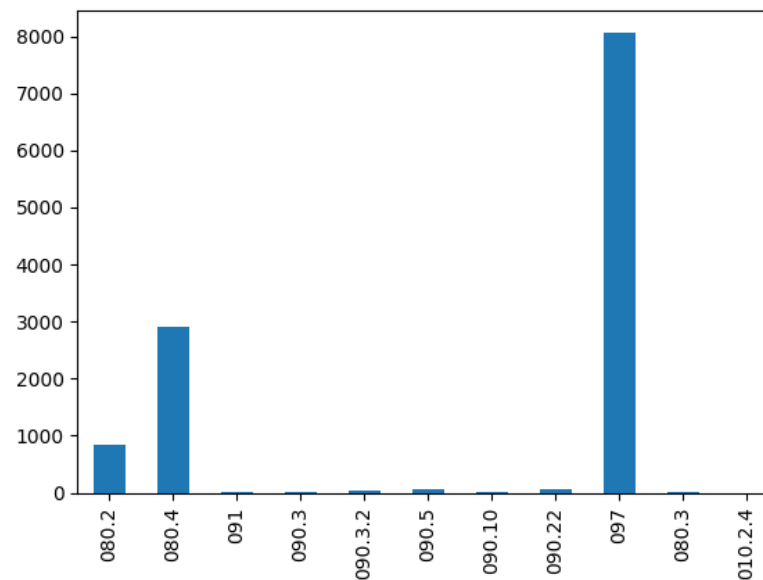
Таких случаев 2579 или 4 % на всех данных и 1754 в обучающей выборке. Исключим их из обучающей выборки, так как это добавляет точность.

Коды отказа не характеризующие клиента

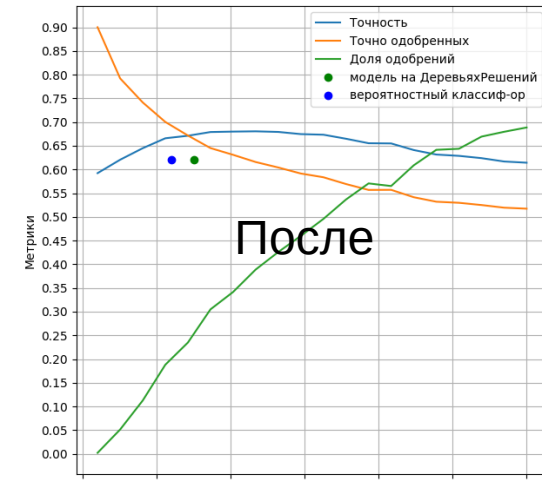
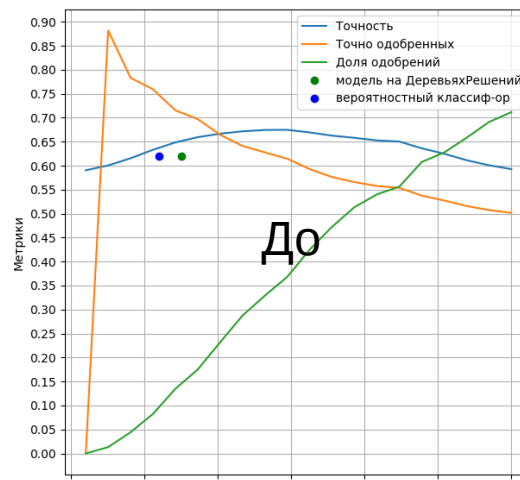
- 080.2 Не набрано достаточное количество баллов верификации и не хватает обязательных компонентов телефонной верификации
- 080.4 Не хватает обязательных компонентов верификации
- 091 Отказ клиента от кредита, заявка не актуальна
- 090.3 Адрес регистрации не соответствует паспорту продукта
- 090.3.2 Регистрация или фактическое проживание заемщика в местах проблемных или удаленных территорий
- 090.5 Нет водительского удостоверения клиента/ родственника первой очереди (муж, жена, мать, отец, сын, дочь)
- 090.10 Адрес постоянной регистрации клиента – общежитие
- 090.22 Регистрация клиента в населенных пунктах, в которых отсутствуют названия улиц, номера домов
- 097 Доходов клиента недостаточно для формирования положительного решения
- 080.3 Какой-либо из указанных телефонов не существует, не обслуживается или временно заблокирован
- 010.2.4 Код подразделения, оформившего паспорт клиента или поручителя/супруга(-и) клиента, расположен в субъекте РФ, который не кредитруется Банком

Исключим их из обучающей выборки, тех которые отклонены андеррайтером.

На диаграмме видно, что на треть увеличилось количество одобренных выгода после выбора точности может составить около 6%



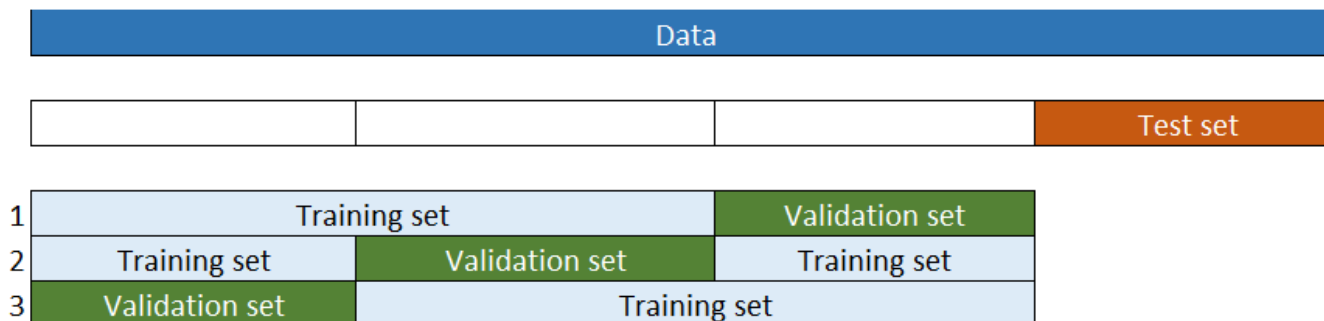
Точность своих моделей старая



Кросс-валидация(обучение) и Тестирование

Обучение с 2020-01-01 по 2021-07-24 (кросс валидация)

- Одобренных: 16933 (0.44)
 - Отклоненных: 38107
-
- Проверка с 2021-07-25 по 2021-10-10 на ~20%
 - Одобренных: 5591 (0.68)
 - Отклоненных: 8170



Содержание

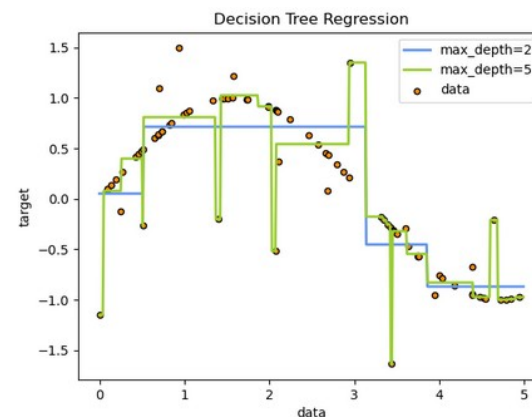
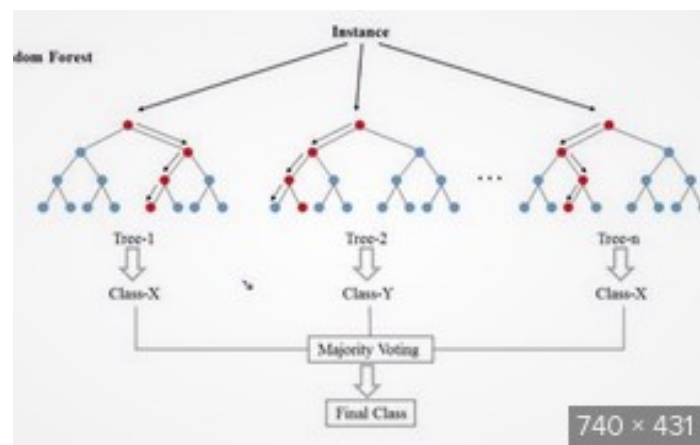
1. Максимальная эффективность — 12 слайд
2. Выбор точности модели — 13 слайд
2. Диапазоны Древа Решений — прогресс - слайд 15,18,22 слайд
- 3. Вероятностный классификатор — слайд 54,55 слайд
4. Важность признаков — 41 слайд

Грубый перебор диапазонов — 80%

Кластеризация — по одобренным и отклоненным — 70%

Классификатор — R Forest

- Это несколько деревьев решений
- Сложно интерпретировать
- Высокая эффективность
- Альтернатива XGBoost для более объективной оценки важности признаков.



Классификатор — XGBoost

- Как и Random Forest - это ансамбль из деревьев.
- Сложно интерпретировать
- Самая высокая эффективность.
- Позволяет оценить максимально достижимые метрики других моделей.

Выбранные метрики - Accuracy

- (правильные решения) / (все заявки)
- Показывает общую точность классификатора
- Вероятность того, что заявка будет отклонена или принята верно
- В 71% случаев мы правильно определяем статус заявки (одобрена\отклонена)
- $(tp+tn)/(tp+tn+fp+fn)$

Выбранные метрики - Precision



- (корректно одобренные) / (одобренные моделью)
- Слева — отличный Precision, но низкий Recall
- Справа — низкий Precision, но высокий Recall
- В 54% случаев мы правильно присваиваем положительный статус заявке
- $tp/(tp+fp)$

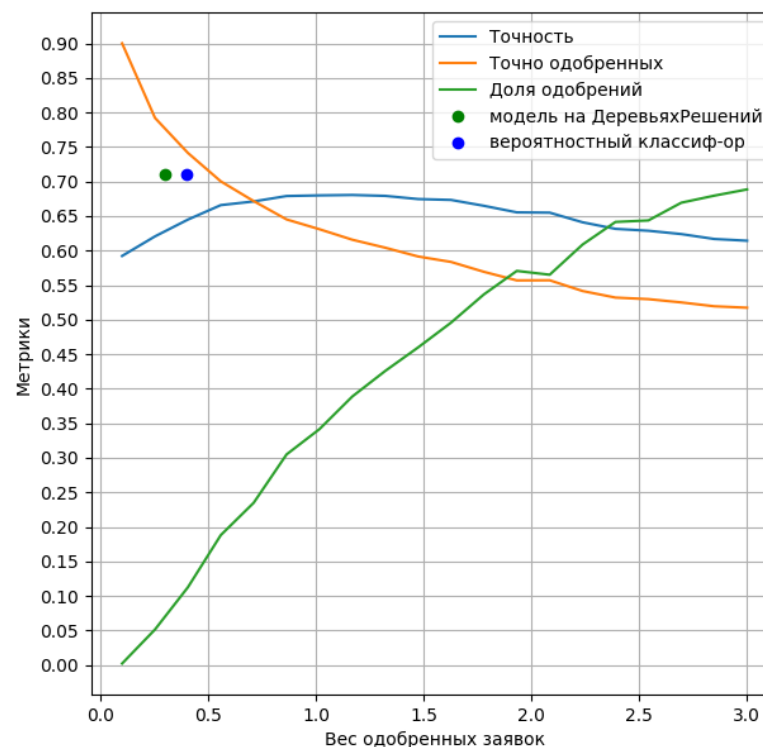
Выбранные метрики - Recall



- (корректные одобрённые) / (хорошие заявки на входе)
- Слева — отличный Recall, но плохой Precision
- Справа — плохой Recall, но отличный Precision
- $tp/(tp+fn)$

Выбор точности

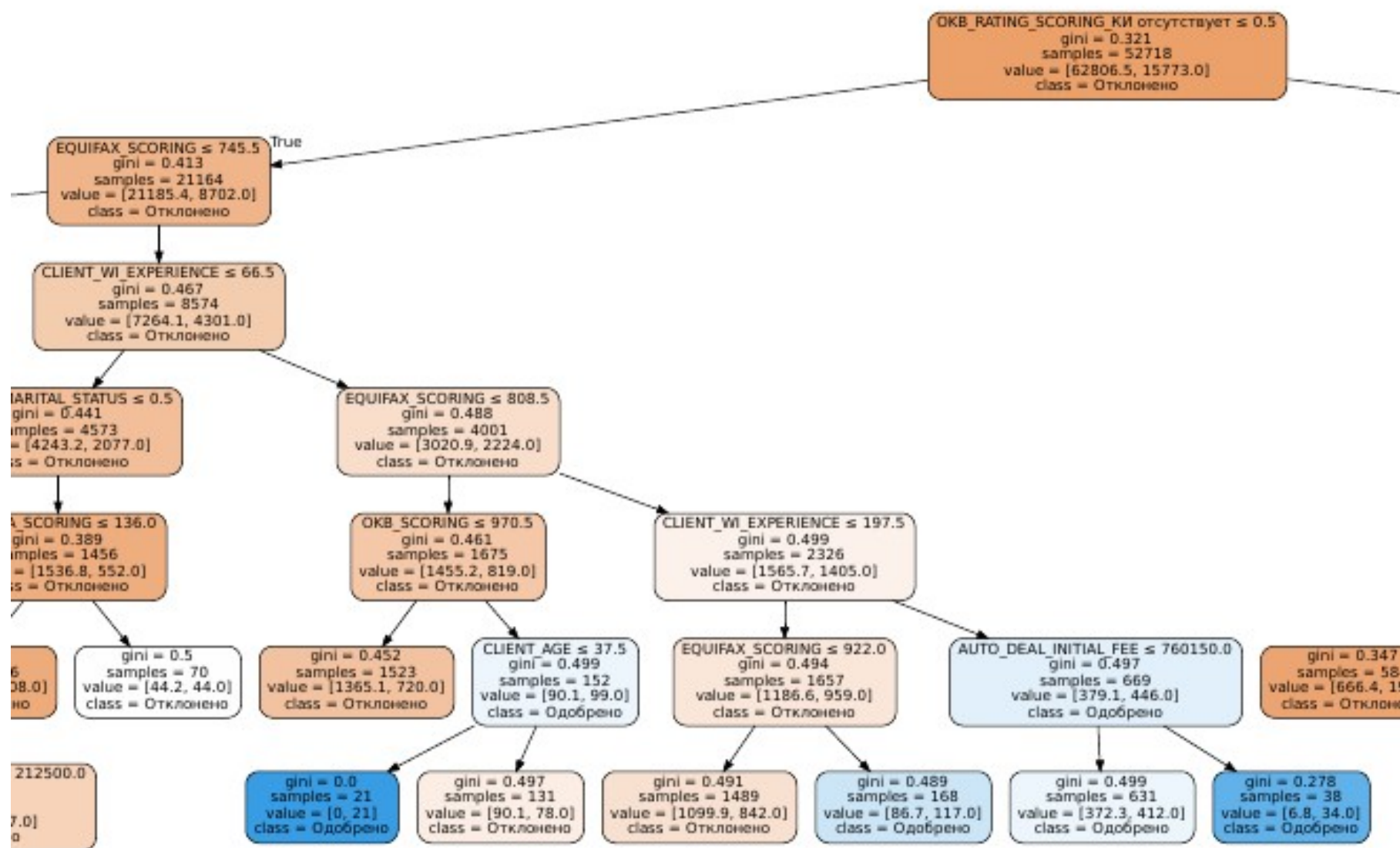
- Сбалансированность модели определяется отношением количества отклоненных к одобренным.
- Это отношение позволяет настроить модель с разными метриками качества.
- Точность — верные/все заявки
- Точно одобренных — точно одобренных / одобренных моделью
- Низкая доля одобренных и высокая точности позволяет оценить качество данных.
- Вес одобренных заявок — параметр модели.



Дерево Решений — ход построения классификатора

- 1) Было построено два дерева с примерно одинаковой точностью, но с разными параметрами.
- 2) Выбраны листья с одобренными заявками.
- 3) Путь от корневого узла к листу выражен в виде кейса с диапазонами признаков.

Дерево Решений



Диапазоны Деревя решений — результат

- Обучение (кросс валидация):

Одобрённых: 15282 (0.37)

Отклонённых: 25102

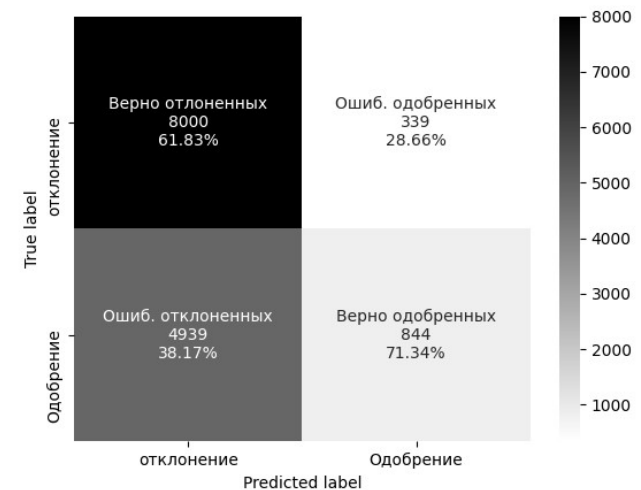
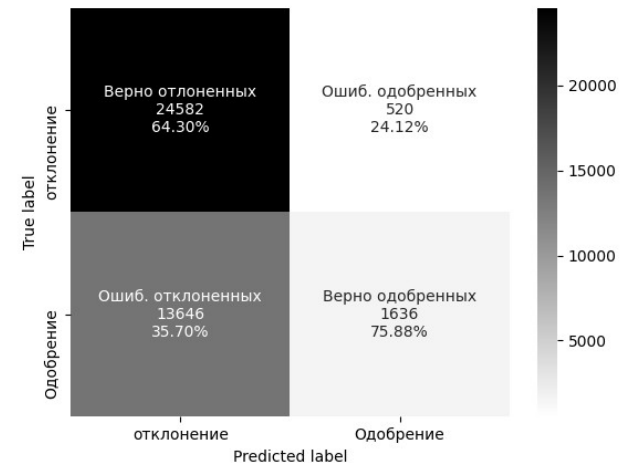
Вероятность точного одобрения 75%, доля одобрённых 5% от общего числа заявок.

- Проверка:

Одобрённых: 5783 (0.4)

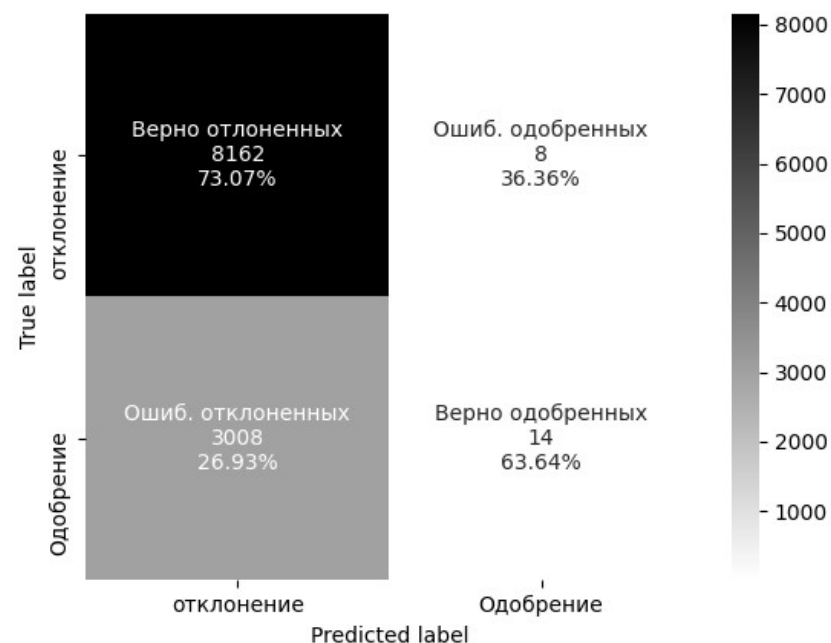
Отклонённых: 8339

Вероятность точного одобрения 71%, доля одобрённых 8% от общего числа заявок.



Дерево решений Результат 1

Атрибут, кейс 1	Значение	Вклад по отдельности
OKB_RATING_SCORING_КИ отсутствует	0	0.21
EQUIFAX скоринг	> 745.5	От 0 до 0.4
Стаж в организации (мес)	< 66.5	От -0.2 до 0
Семейное положение	Не в браке	0.05
Анкетный скоринг	> 90.5	0.05
Первонач. взнос	<= 40500	0.2



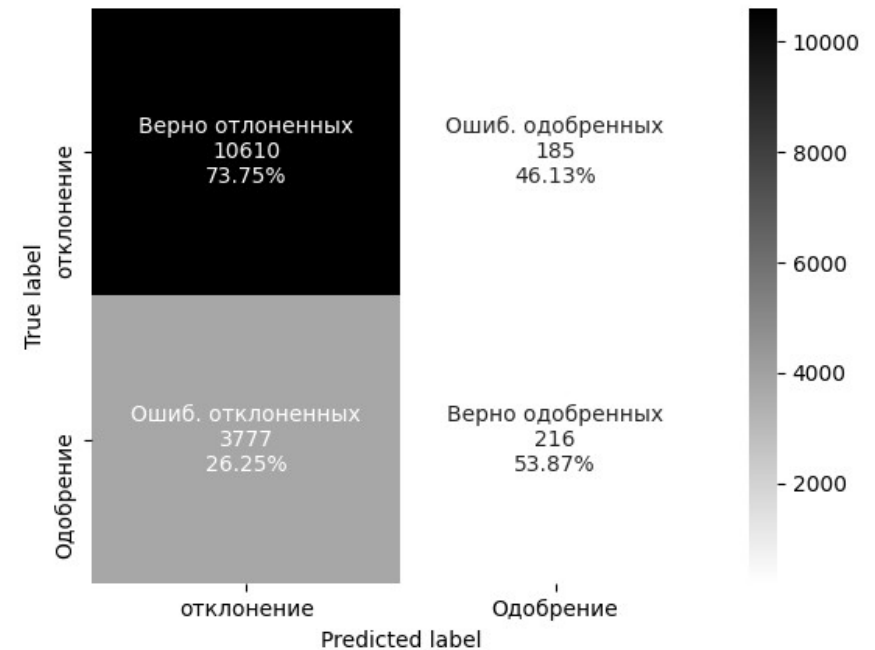
Проверка на тестовой выборке, Кейс 1.

Стаж в организации (мес) — не влияет на результат.

OKB_RATING_SCORING_КИ отсутствует = 0 — это любое значение кроме «КИ отсутствует»

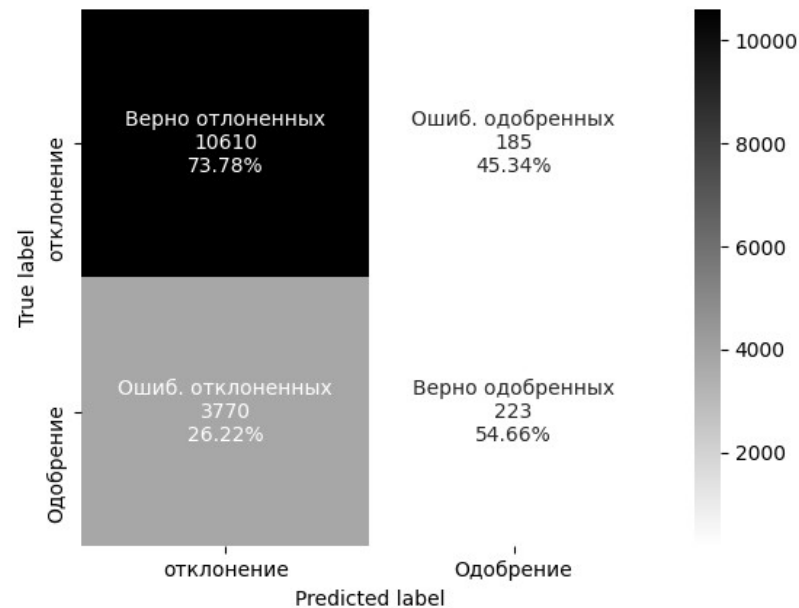
Дерево решений Результат 2

Атрибут, кейс 2	Значение
OKB_RATING_SCORING_КИ отсутствует	0
EQUIFAX скоринг	> 808.5
Стаж в организации (мес)	> 197.5



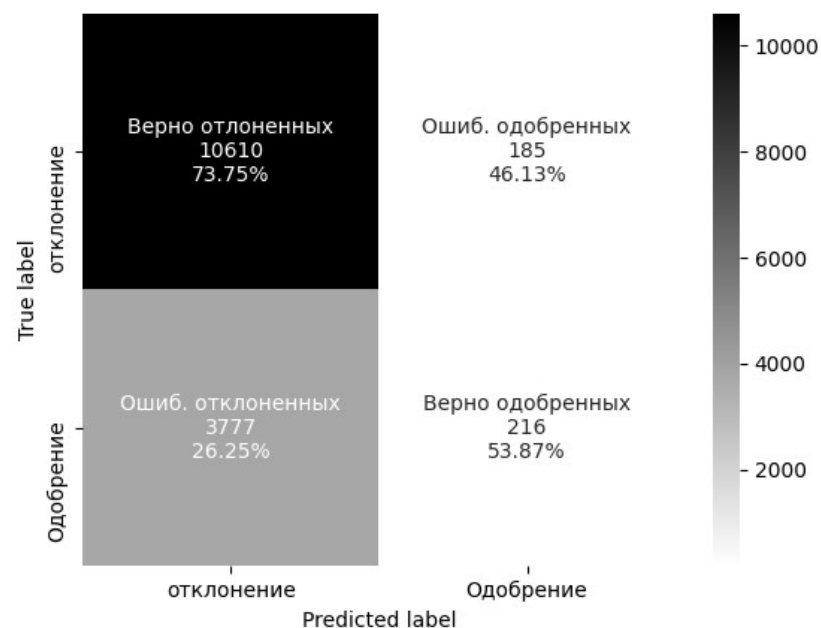
Дерево решений Результат 3

Атрибут, кейс 3	Значение
ОКВ_RATING_SCORING_KI отсутствует	0
Стаж в организации (мес)	≤ 197.5
Эквивалентный скоринг	> 922



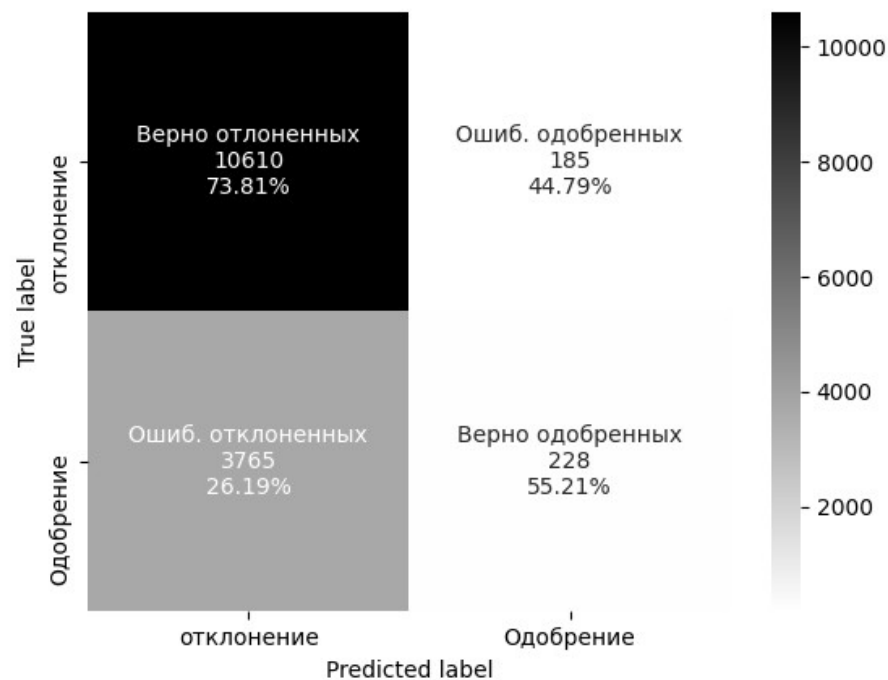
Дерево решений Результат 4

Атрибут, кейс 4	Значение
ОКВ_RATING_SCORING_КИ отсутствует	0
Эквифакс скоринг	> 808.5
Стаж в организации (мес)	> 197.5
Первонач. взнос	> 760150



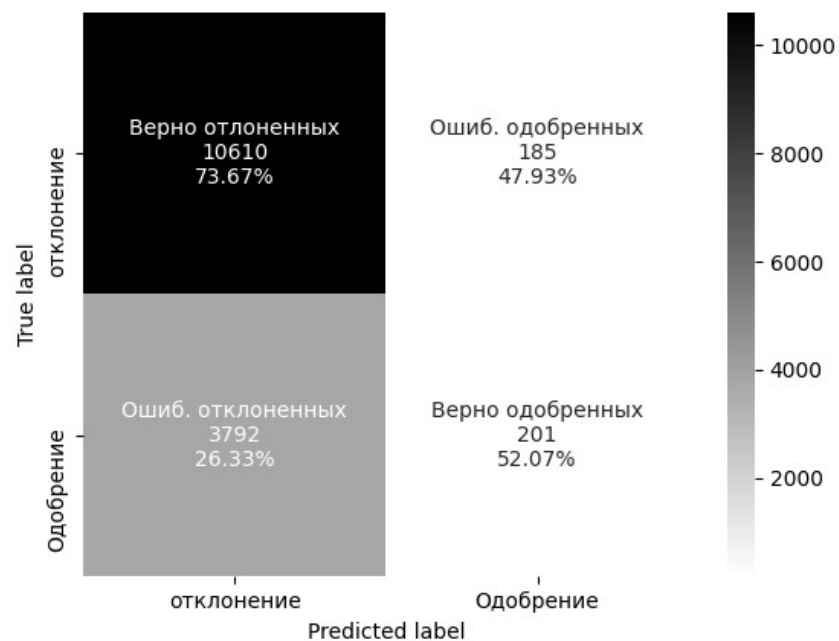
Дерево решений Результат 5

Атрибут, кейс 5	Значение
ОКВ_RATING_SCORING_КИ отсутствует	0
Эквивалент скоринг	≤ 745.5
Стаж в организации (мес)	От 500 до 22500



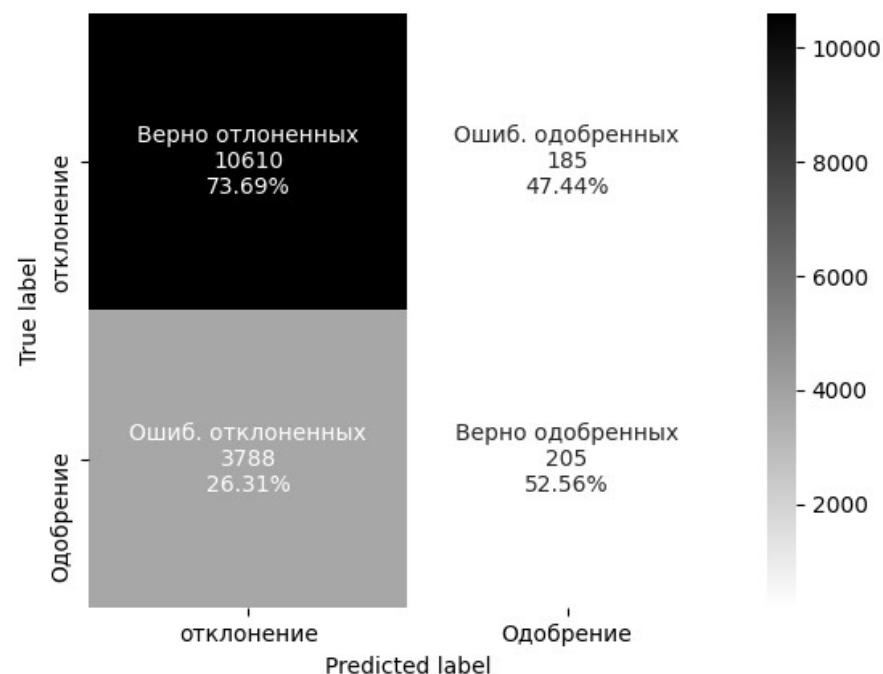
Дерево решений Результат 6

Атрибут, кейс 6	Значение
• ОКВ_RATING_SCORING_KИ отсутствует	• 0
• Эквифакс скоринг	От 745 до 846
• Стаж в организации (мес)	< 48500



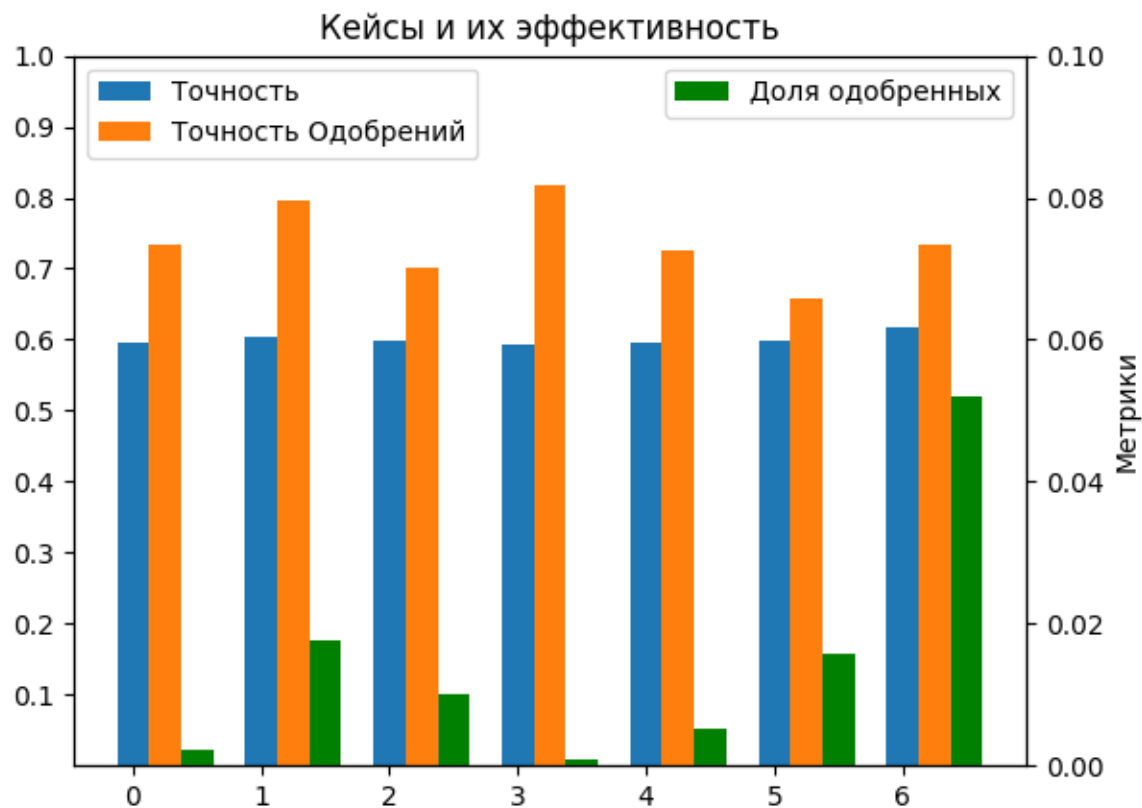
Дерево решений Результат 7

Атрибут, кейс 6	Значение
• OKB_RATING_SCORING_KИ отсутствует	• 0
• Эквифакс скоринг	>846.5
Семейное положение	в браке
Оценка кредитной истории ОКБ	<= 810



Оценка кредитной истории ОКБ ≤ 810 — улучшает точность на кросс-валидации,
На тестовой выборке никак не влияет.

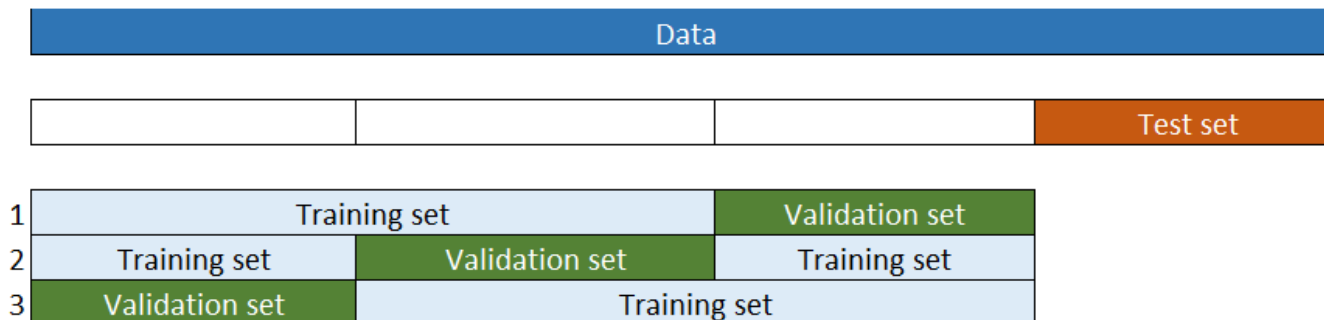
Дерево решений — Кейсы (устарел)



Проверка на тестовой выборке

Проверка кросс-валидация

- Кросс-валидация (скользящий контроль) — стандартная практика тестирования, позволяет эффективнее использовать датасет и более объективно оценивать модели.



Проверка Hold-out

- Hold-out практика - разбиение данных на два множества обучения и валидации/тестирования.
- Недостатки:
 1. переобучение, если использовать для подбора параметров
 2. возможность оптимистической оценки модели человеком (переобучение).
 3. неэффективное использование датасета.

Проверка — выбранный подход

Для подбора параметров использован стратифицированный K-Fold, с количеством сплитов равных 2, 3 на более узком диапазоне параметров. Последнюю проверку модели будем производить на сохраненном Hold-out 20 процентах последних заявок, с даты 2021-07-10.

В Hold-out:

- Одобренных: 3991
- Отклоненных: 9189

Стратификация — это сохранение пропорции одобренных заявок к отклоненным при разбиении.

Важность признаков

Важность признаков до суммирования измеряется в процентах от общего вклада, нам важно только их отношение.

Проблемы вычисления важности или вклада признаков:

- вычисляются для конкретной модели
- модели основанные на деревьях очень нестабильны (вариации)

Для решения этих проблем важность была вычислена для двух лучших моделей из двух библиотек и их большого числа вариаций.

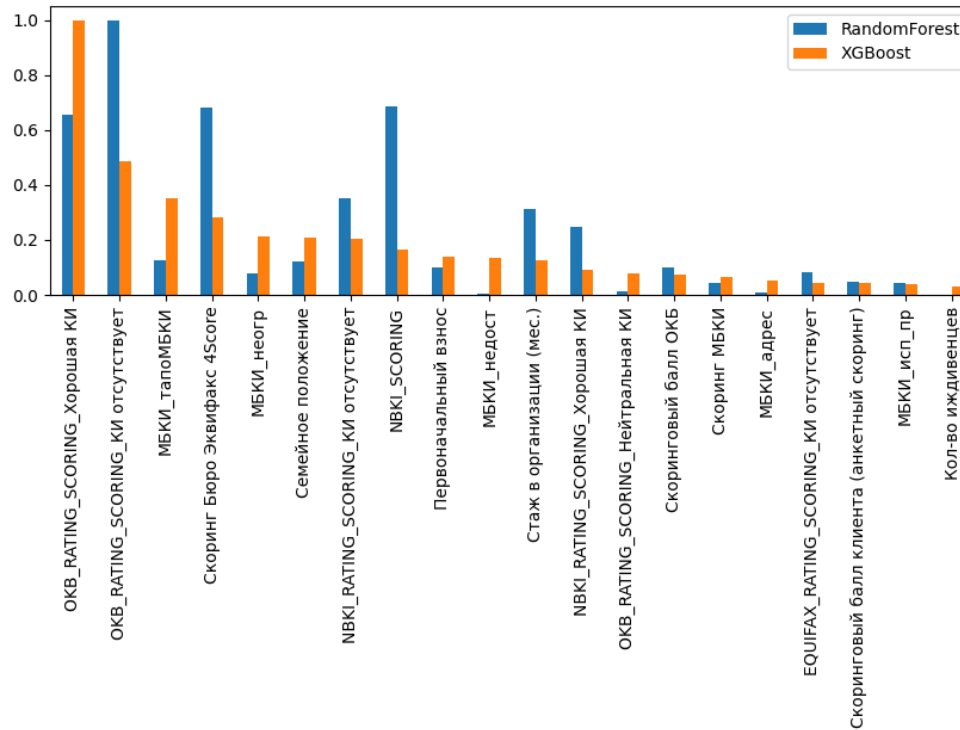
Поэтому, были использованы простые быстрые методы:

- метод основанный на среднем падении загрязнения (MDI) для Random Forest
- „Gain“ метод для XGBoost

Недостатки этих методов были максимально нивелированы на этапе подготовки данных.

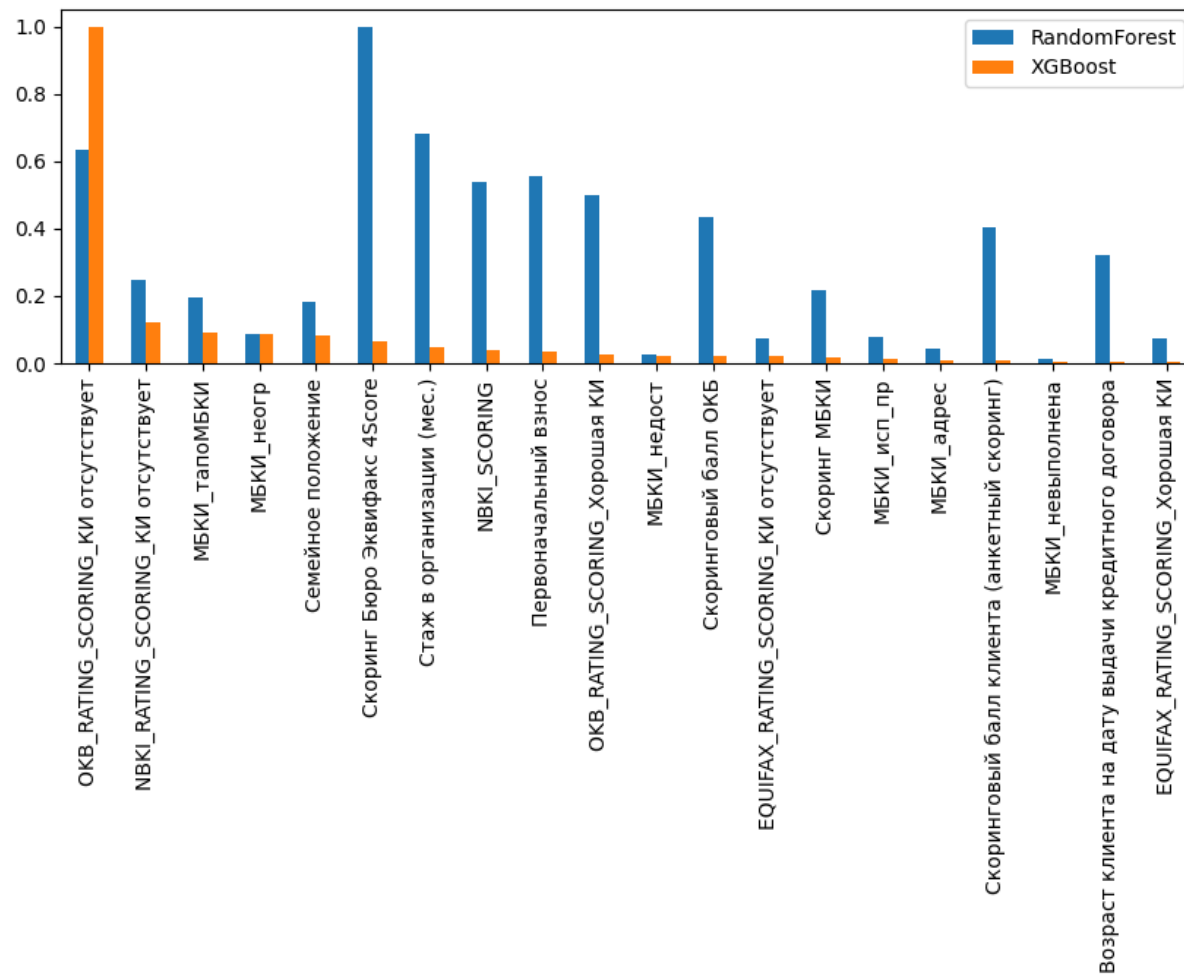
Модели с высоким Precision и низким Recall

Важность признаков для XGBoost и RandomForest



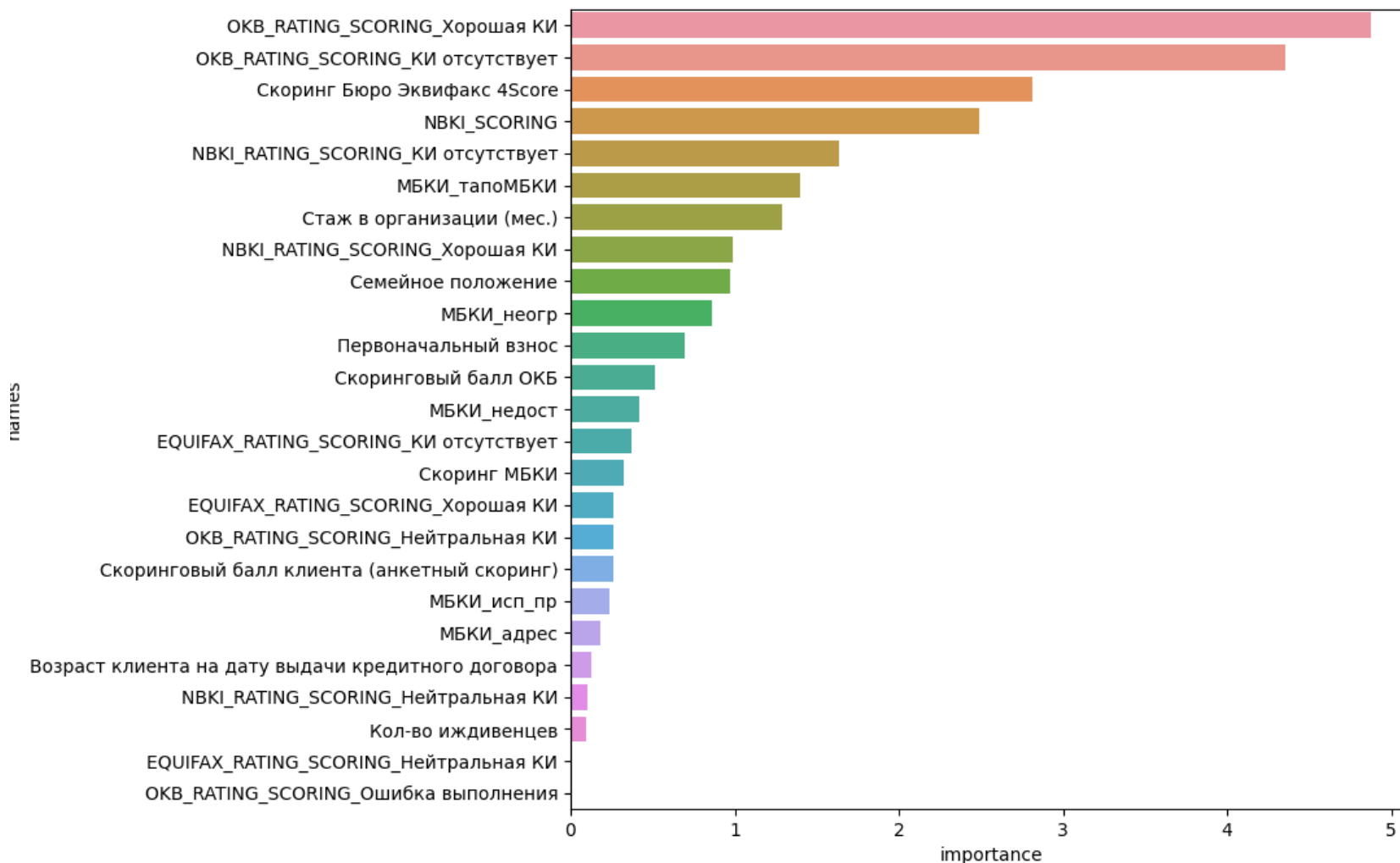
Модель с низкой Precision и Высоким Recall

Важность признаков для XGBoost и RandomForest



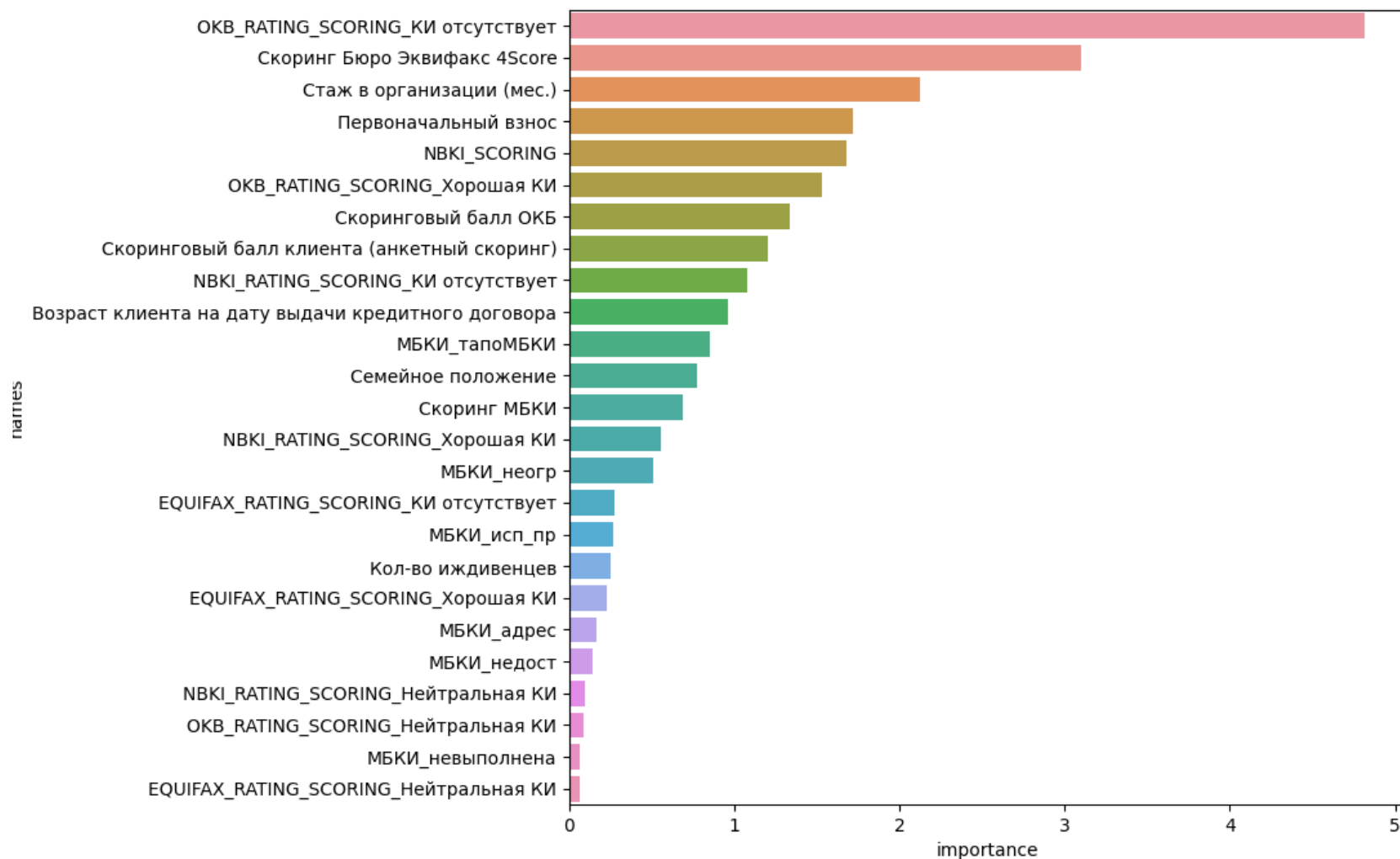
Высокая Точность, низкий охват

Важность признаков сумма XGBoost и RandomForest

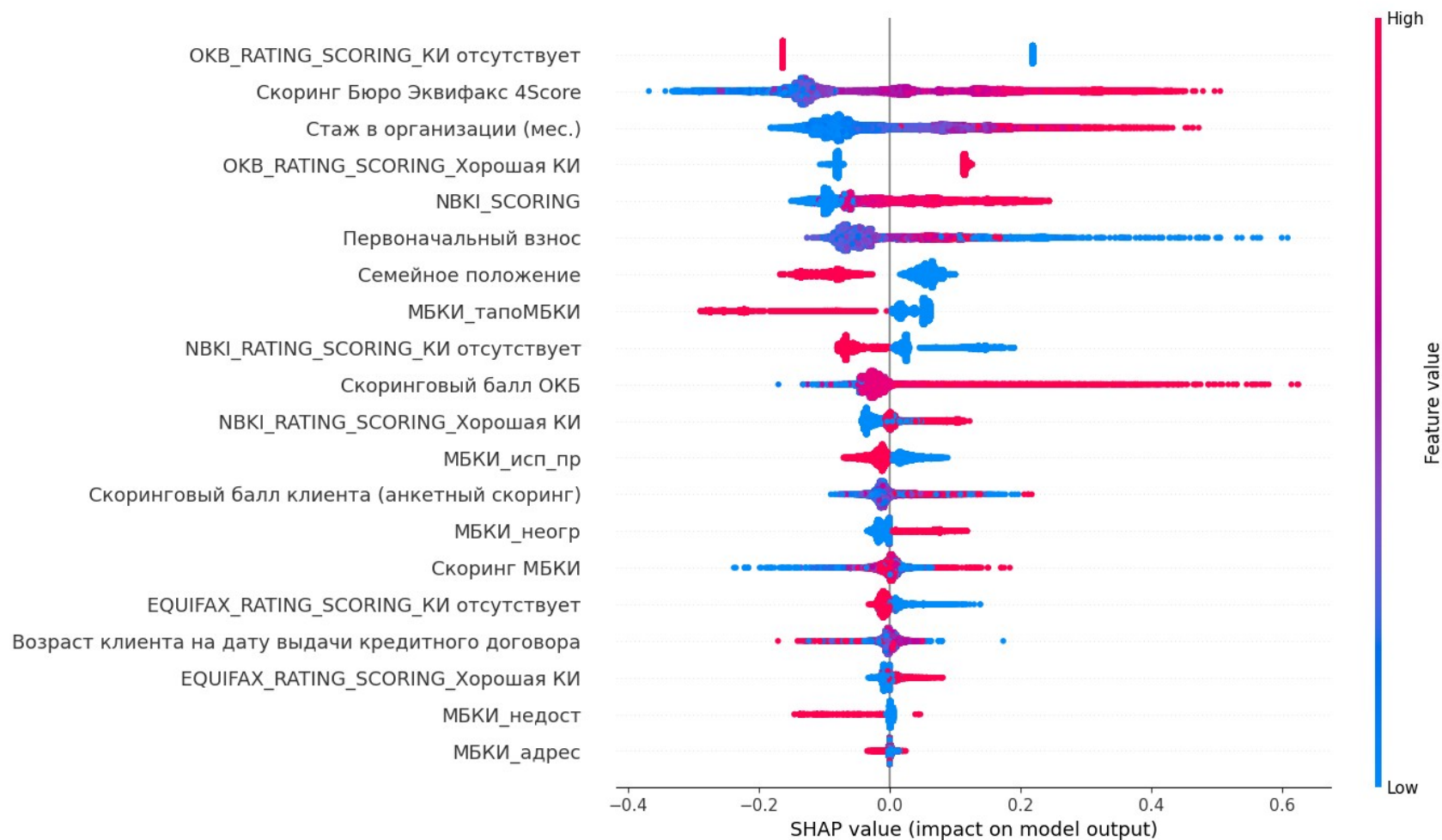


Низкая Точность, высокий охват

Важность признаков сумма XGBoost и RandomForest

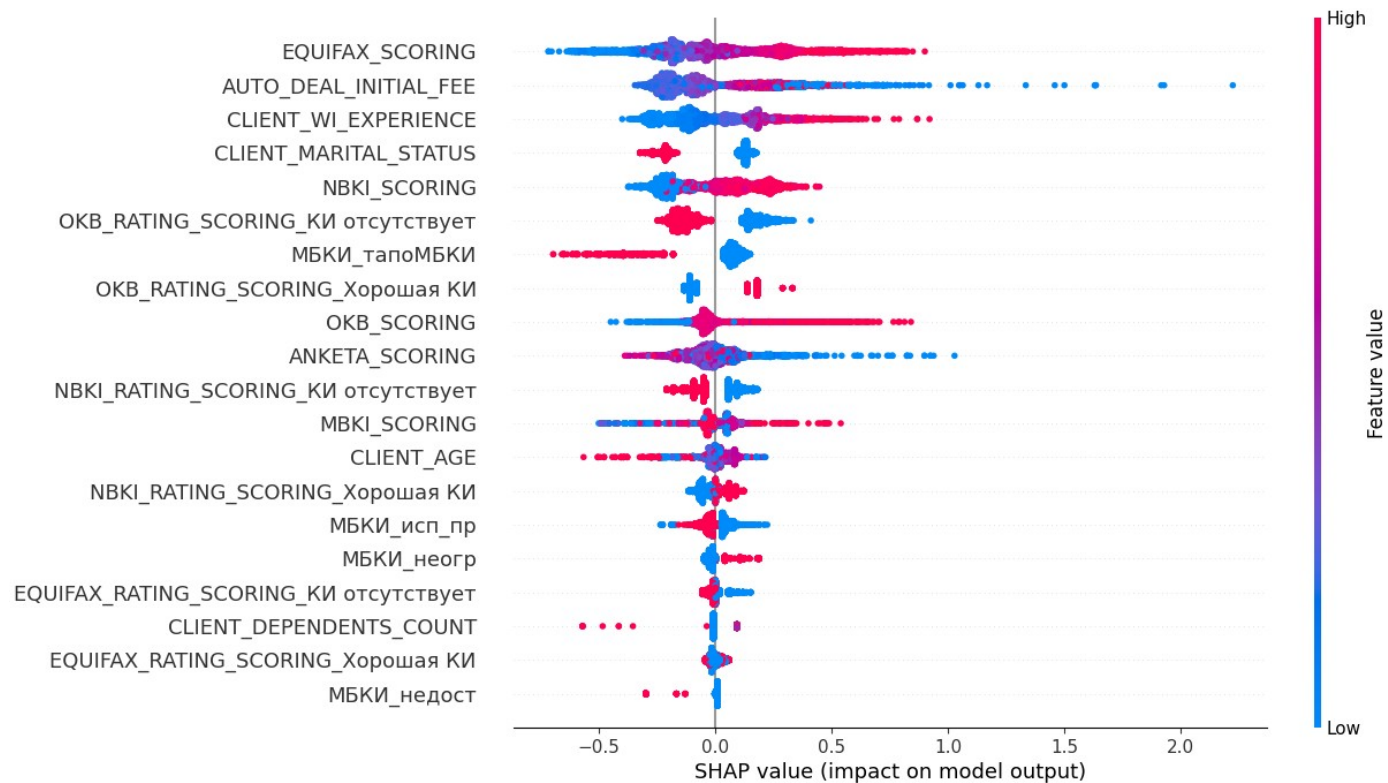


XGBoost — влияние признаков — низкий Precision



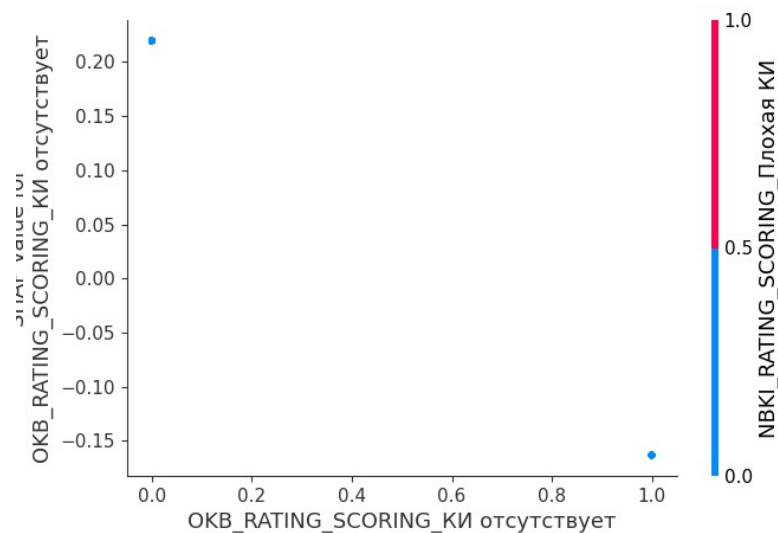
SHAP value — вклад — в принятие решения моделью (положительный — одобрение)

XGBoost — влияние признаков — высокий Precision

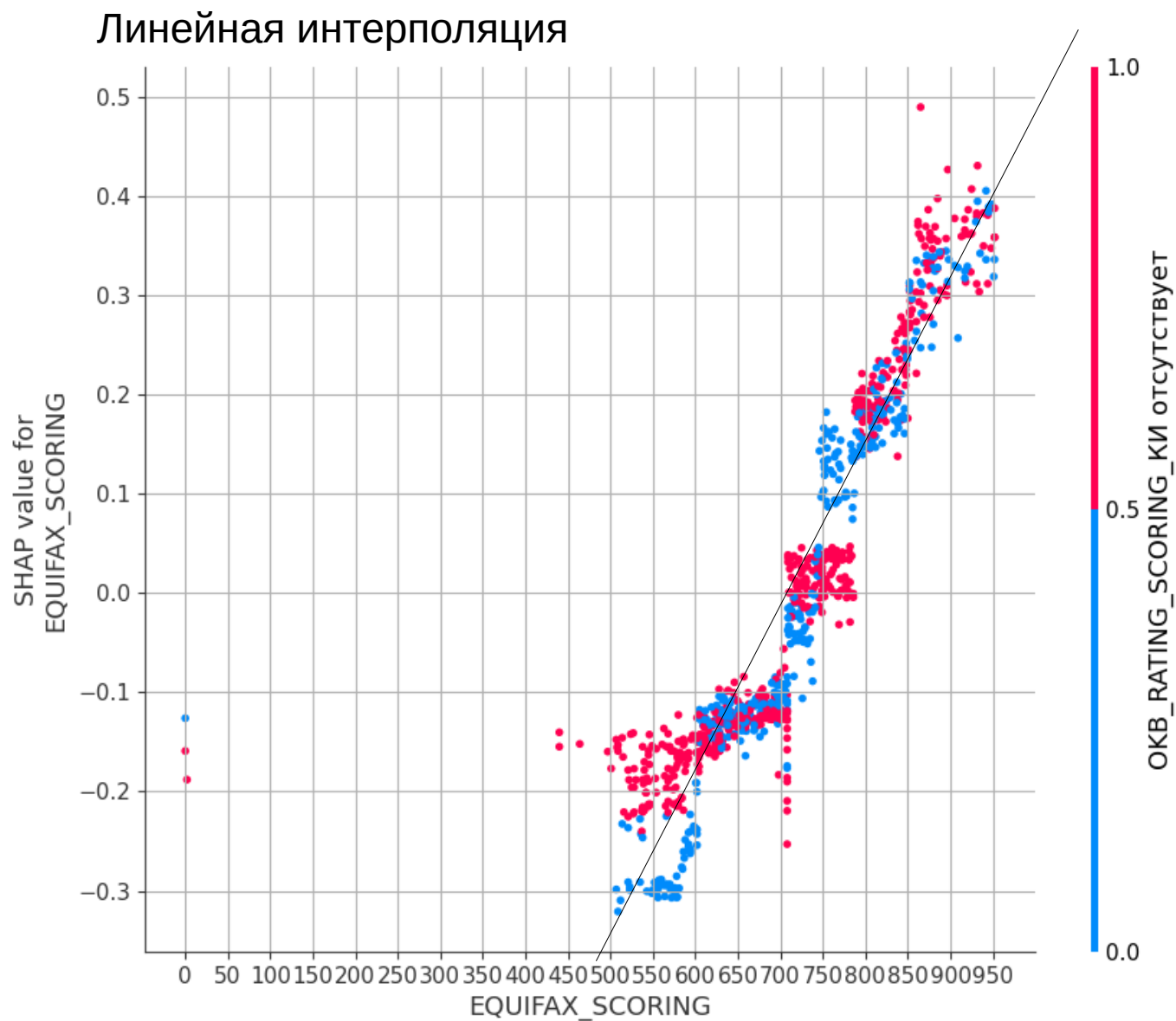


Классификатор — синтез 1

- Построим модель, пройдемся по лестнице от самого важного признака, учитывая существенные черты модели.
- +/- вес — одобрение или отклонение
- OKB_RATING_SCORING_KI отсутствует = 1 — отклонение с весом -0.15
- OKB_RATING_SCORING_KI отсутствует = 0 — одобрение с весом 0.22



Классификатор — синтез 2



Классификатор — синтез 2

X - «Эквивалент скоринг», Y — влияние отрицательное и положительное с центров — 0.

При 750 «Эквивалент скоринг» переходит в положительное влияние на одобрение.

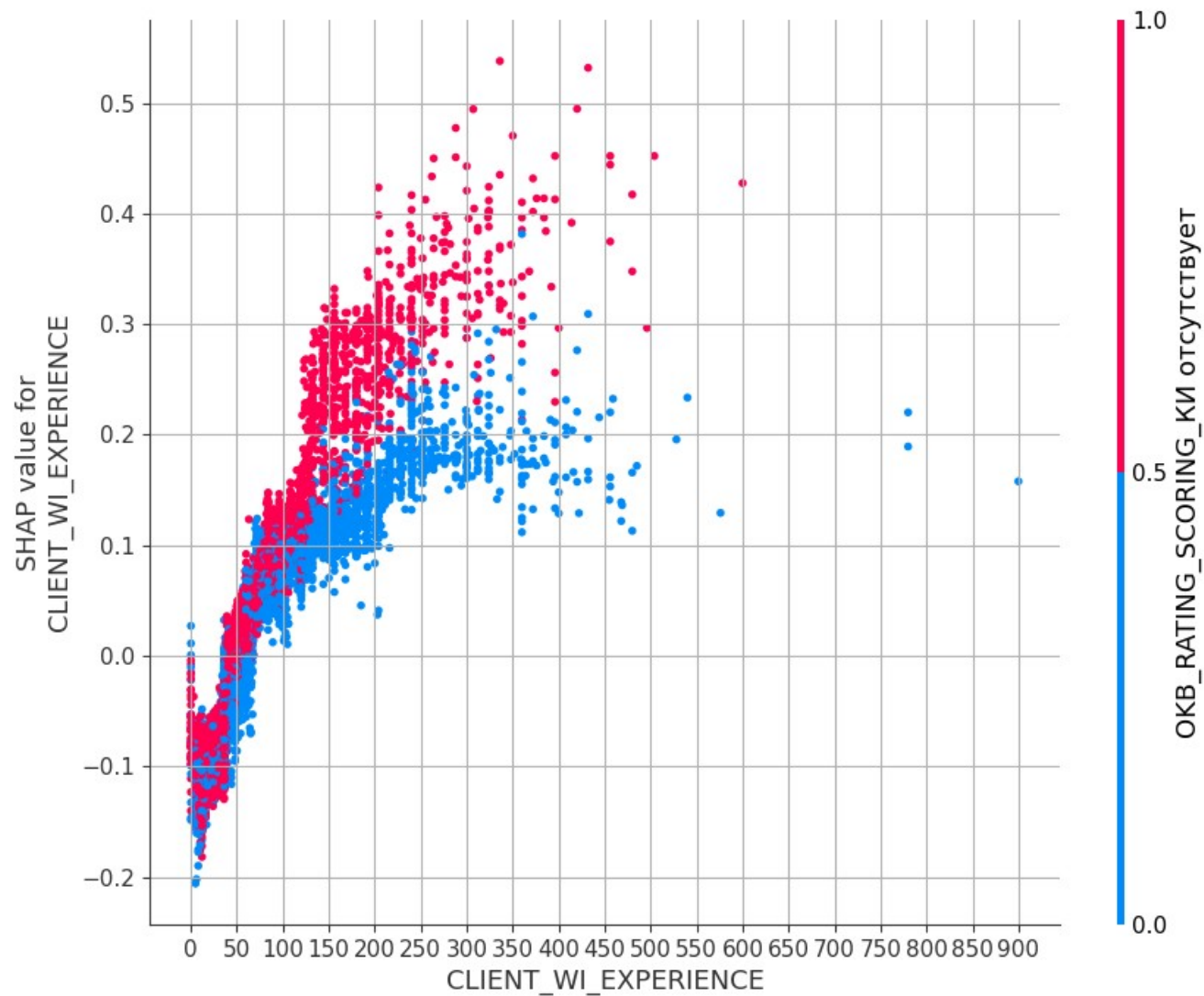
Так как все заявки лежат на одной прямой, OKB_RATING_SCORING_KI отсутствует не влияет на «Эквивалент скоринг».

По формуле прямой проходящей через две точки вес «Эквивалент скоринг» можно задать формулой $(\text{«Эквивалент скоринг»} * 0.7) / 245 - 2.1$

Классификатор построенный по двум этим признакам имеет:

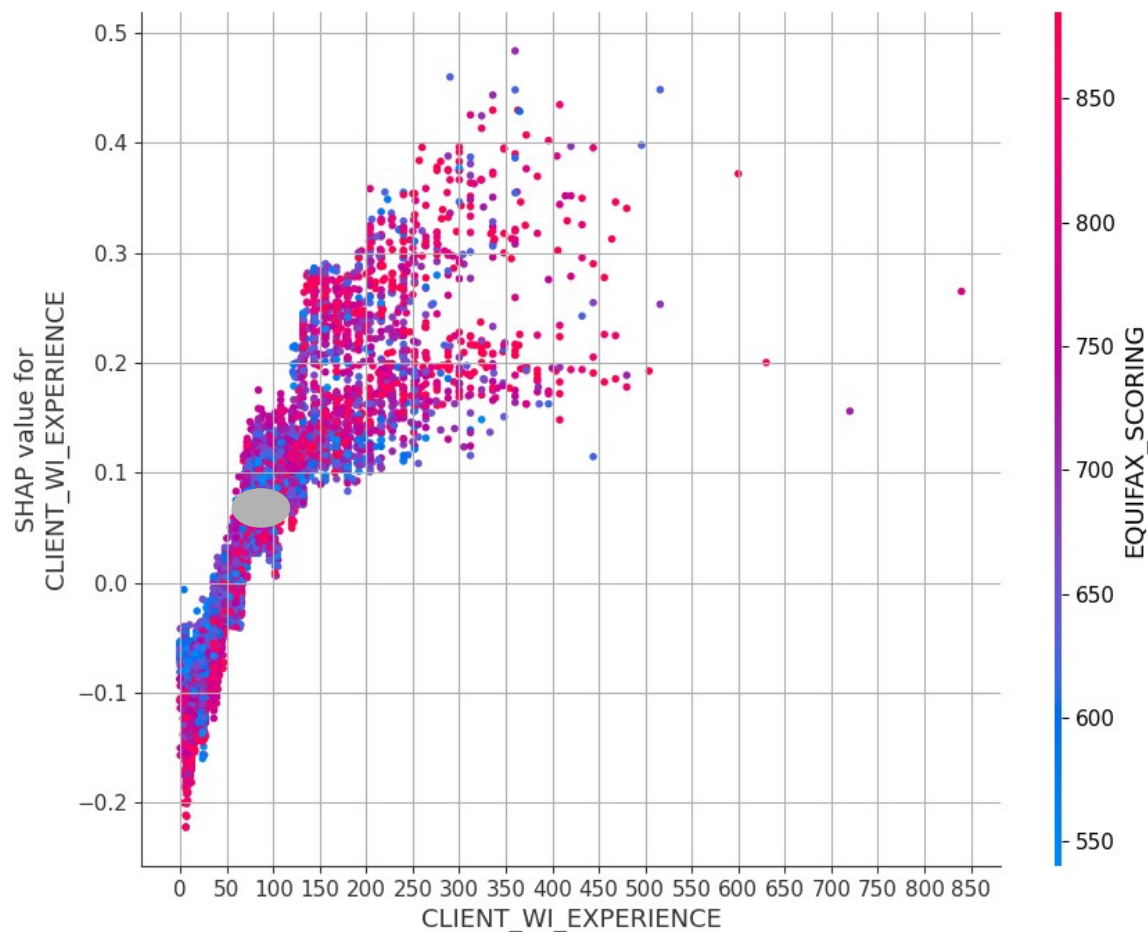
- Accuracy: 0.623748
- Precision: 0.415058
- Recall: 0.592583

Классификатор — синтез 3



Классификатор — синтез 3

Брем точку чуть выше 0 и считаем ее переломной



Классификатор — синтез 3

- Сравним третий признак «Стаж работы» с вышестоящими.
- Мы видим, что от 1 признака зависимости нет, а от 2 есть.
- При значении меньше 80 «Стаж работы» влияет отрицательно на одобрение, положим -0.15, при значениях больше 80 — с разной степенью на одобрение, положим +0.1
- Получена модель по 3 признакам:

Accuracy: 0.662822

- Precision: 0.447582
- Recall: 0.484590

Классификатор — синтез 4

OKB_RATING_SCORING_Хорошая КИ не имеет явных зависимостей, кроме OKB_RATING_SCORING_КИ отсутствует

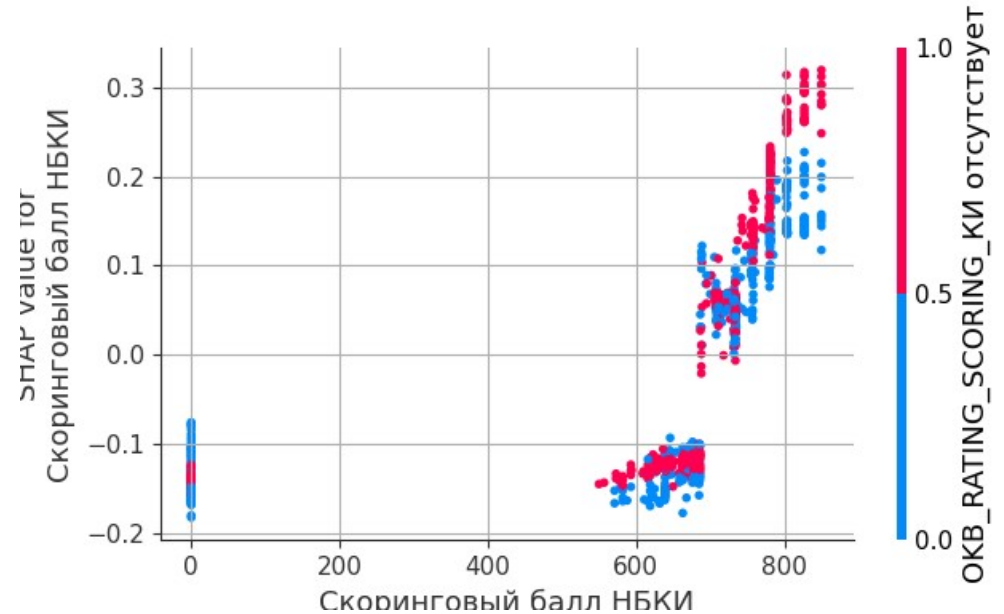
- Применим вес если OKB_RATING_SCORING равно Хорошая КИ, то вес будет 0.2, иначе -0.2
- Результат после применения:

Accuracy: 0.663885

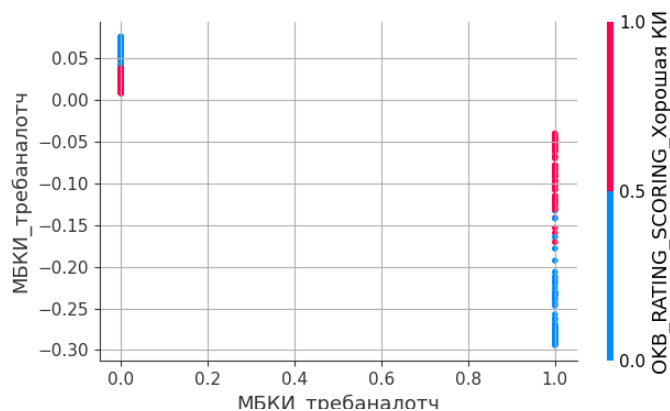
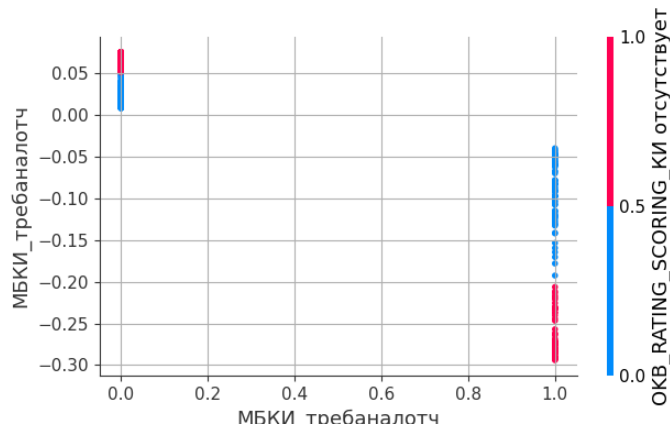
- Precision: 0.451599
- Recall: 0.513155

Классификатор — синтез «NBKI»

- Имеет небольшую зависимость от ОКБ КИ Отсутствует
- При менее 700 - имеет одно значение -0.12, более — линейная зависимость
- $X1=700$
- $X2=900$
- $Y1=0.05$
- $Y2=0.3$
- $(x-700)/200 = (y-0.05)/0.25 \Rightarrow (x-700)*0.25 = (y-0.05)*200 \Rightarrow x*0.25 - 175 = y*200 - 10 \Rightarrow y = x*0.25/200 - 165/200 \Rightarrow y = x*0.00125 - 0.825$



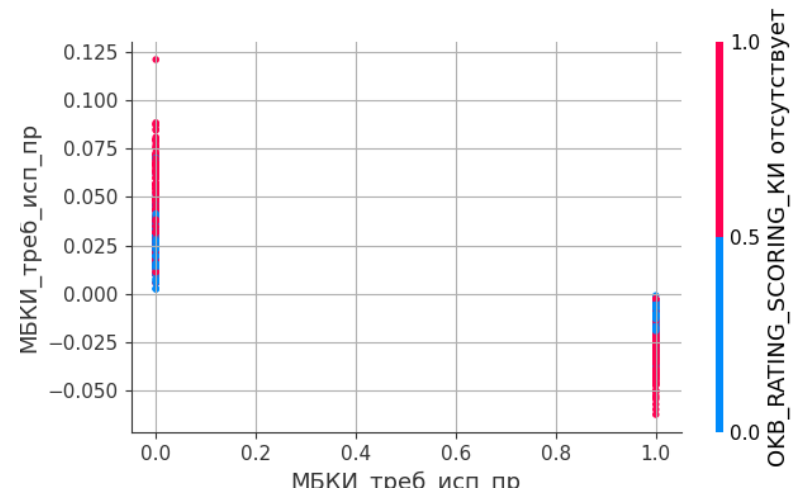
Классификатор — синтез «Скоринг МБКИ»



- Требуется анализ полного отчета МБКИ:
- - имеет четкое влияние
- - зависит от КИ отсутствует и Хорошая КИ
- - Наличие КИ Отсутствует усиливает важность.
- Для 0 = 0.03, для 1 = -0.12 И умножаем на 2 если КИ Отсутствует = 1

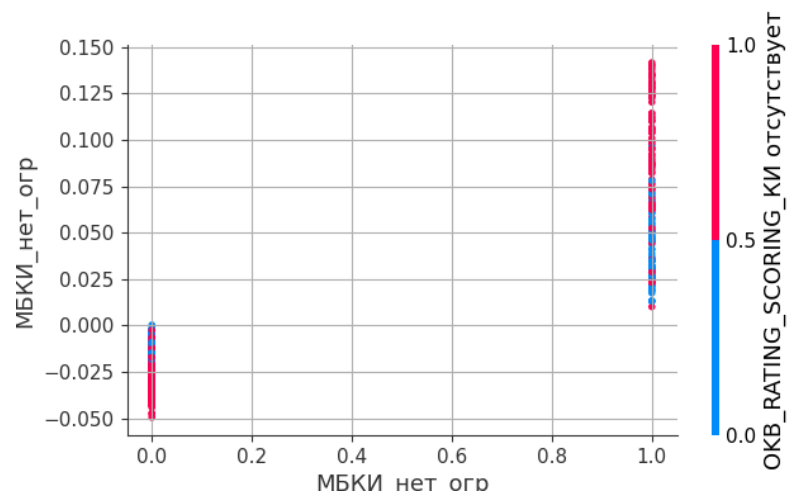
Классификатор — синтез «Скоринг МБКИ»

- Требуется проверка данных о наличии исполнительных производств — влияет однозначно, имеет сильную зависимость от Скоринга Эквивафакс
- Для 0 = 0.25
- Для 1 = -0.25



Классификатор — синтез «Скоринг МБКИ»

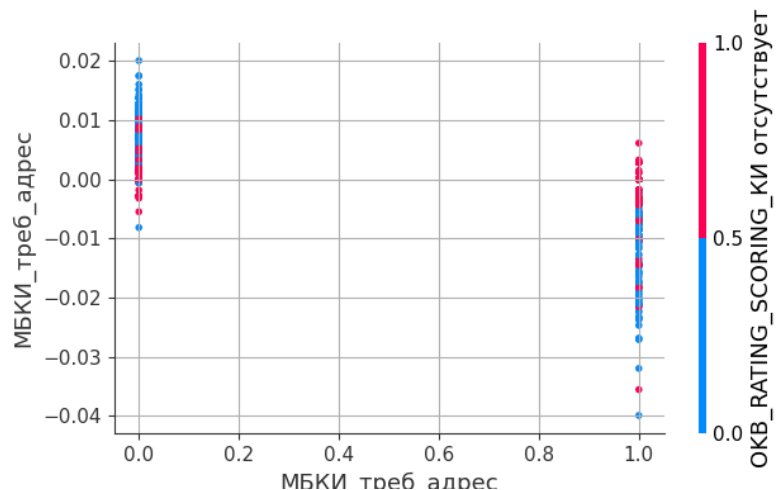
- Нет ограничений
- КИ отсутствует усиливает
- Возмем при 1 = 0.05
- При 0 = -0.025



Классификатор — синтез «Скоринг МБКИ»

Влияет незначительно,
поэтому добавлять в
модель не будем.

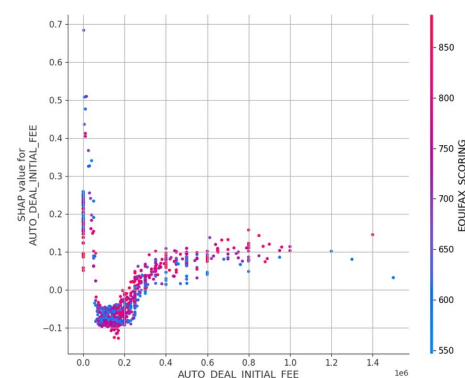
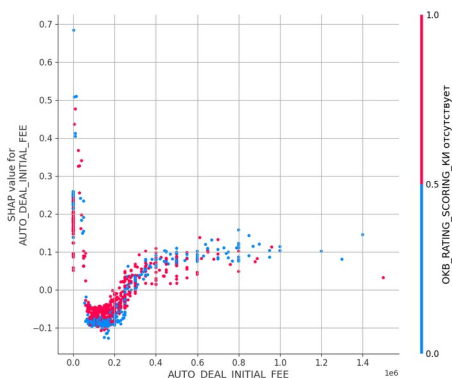
- Как мы видим, все МБКИ показатели усиливаются значением КИ отсутствует, поэтому будем использовать общий мультипликатор 2 при наличии КИ отсутствует.



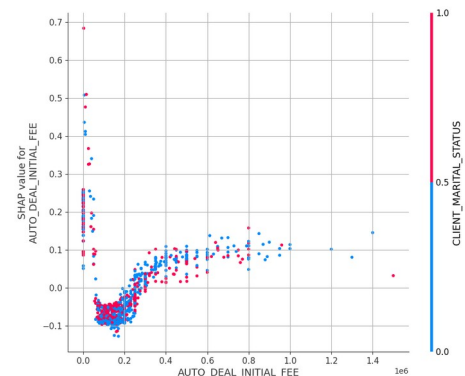
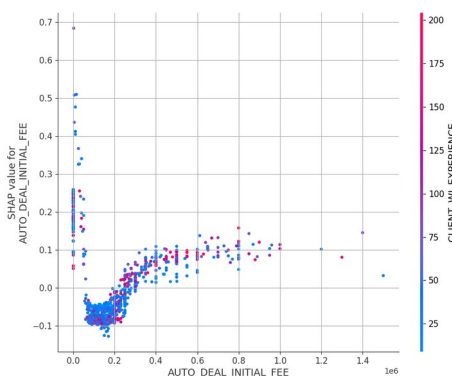
Классификатор — синтез

«Первоначальный взнос» - нет вклада

- Зависимостей нет



- Выделяются диапазоны
- $\geq 0 < 50000$
- $\geq 5000 < 300000$
- > 300000
- Добавление признака не дало результат



Вероятностный классификатор

— Результат - модель

Сложим следующие веса и если сумма будет положительная, то одобряем заявку, а если отрицательная — то отклоняем.

Четыре столбца:

1) Если «Оценка кредитной истории ОКБ» равно «КИ отсутствует», то вес ставим -0.32, равно Хорошая КИ, то +0.175, иначе 0.023

2) По Формуле вычисляем вес «Эквивалент скоринг» * 0.00286 - 2.46

3) Если «Стаж работы» < 61, то вес -0.02, иначе вес +0.23

4) Если «Семейное положение» = 'не в браке', то вес -0.26, если = 'в браке', то вес 0.03

5) Если $0 \leq \text{«Первоначальный взнос»} \leq 50000$, то 0.21 если $50000 < \text{ПВ} \leq 300000$, то -0.2, если $\text{ПВ} > 300000$, то 0.36

6) Если «НБКИ Скоринговый Бал» ≤ 700 , то -0.07, иначе «НБКИ Скоринговый Бал» * 0.001 — 0.84

7) «МБКИ», Если ОКБ КИ отсутствует, то мультипликатор = 0.9

- МБКИ_требаналотч присутствует, то -0.18*мультип

- МБКИ_треб_исп_пр присутствует, то -0.07*мультип

- - МБКИ_нет_огр присутствует, то 0.1*мультип

- Погрешность порядка 0.02

Плюсы данного подхода по сравнению с диапазонами:

- - Дает степень уверенности и вероятности одобрения/отклонения.
- - Точность Модели настраивается через подбор автоматический подбор параметров.

* Параметры классификатора (подбираются автоматически)

Вероятностный классификатор

— Результат

- Обучение (кросс валидация)

Одобрённых: 15282 (0.38)

Отклонённых: 25102

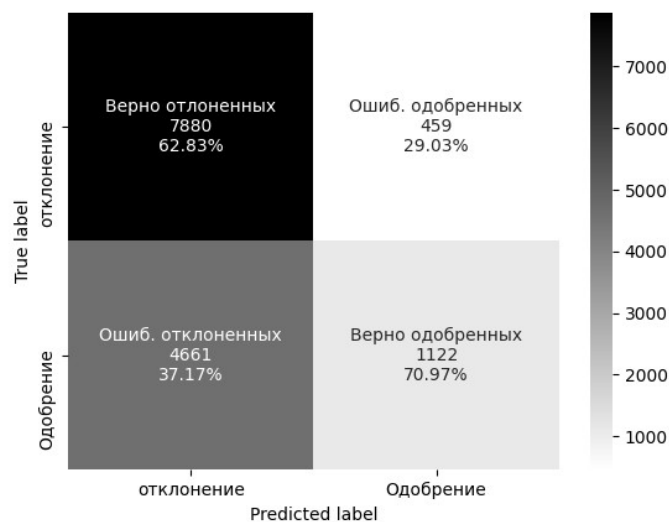
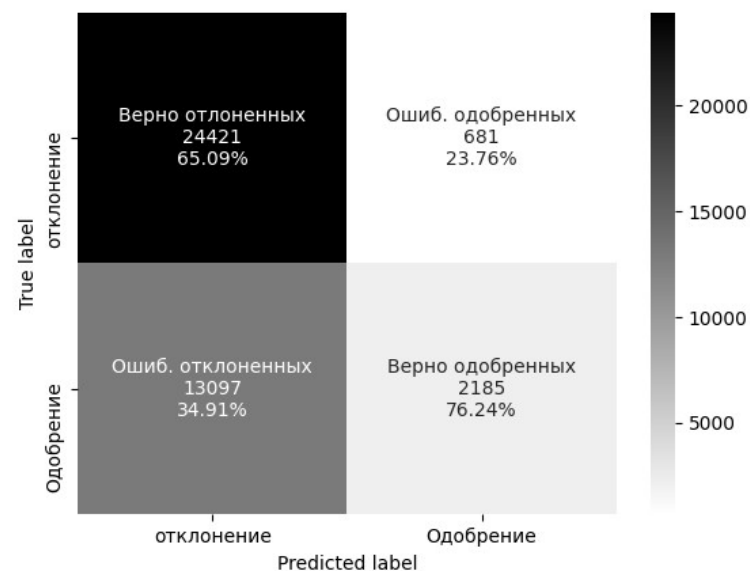
Вероятность точного одобрения 77%,
доля одобрённых 7% от общего числа
заявок (в которых 38% одобрённых).

- Проверка с 2021-07-25 по 2021-10-10
на ~20%

Одобрённых: 5783 (0.4)

Отклонённых: 8339

Вероятность точного одобрения 71%,
доля одобрённых 11% от общего
числа заявок



Вероятностный классификатор

— Результат — метрики (old)

Проверка на ~20% Hold-out заявок с даты 2021-07-10

- Одобрённых в Hold-out: 3991
- Отклонённых в Hold-out: 9189

Вероятность точного одобрения 62%, доля одобренных 2% от общего числа заявок (в которых 30% одобренных должно быть).

Одобрённых 286 из которых 178 одобрены верно.

Ошибочно одобрены 108, ошибочно отклонены 3813.

- Одобрённых: 0.021700 — количество одобренных / всего заявок
- Accuracy: 0.702504 — доля корректно идентифицированных заявок
- Precision: 0.622378 — корректность одобренных моделью заявок
- Recall: 0.044600 — верно одобренных из всех одобренных 4%

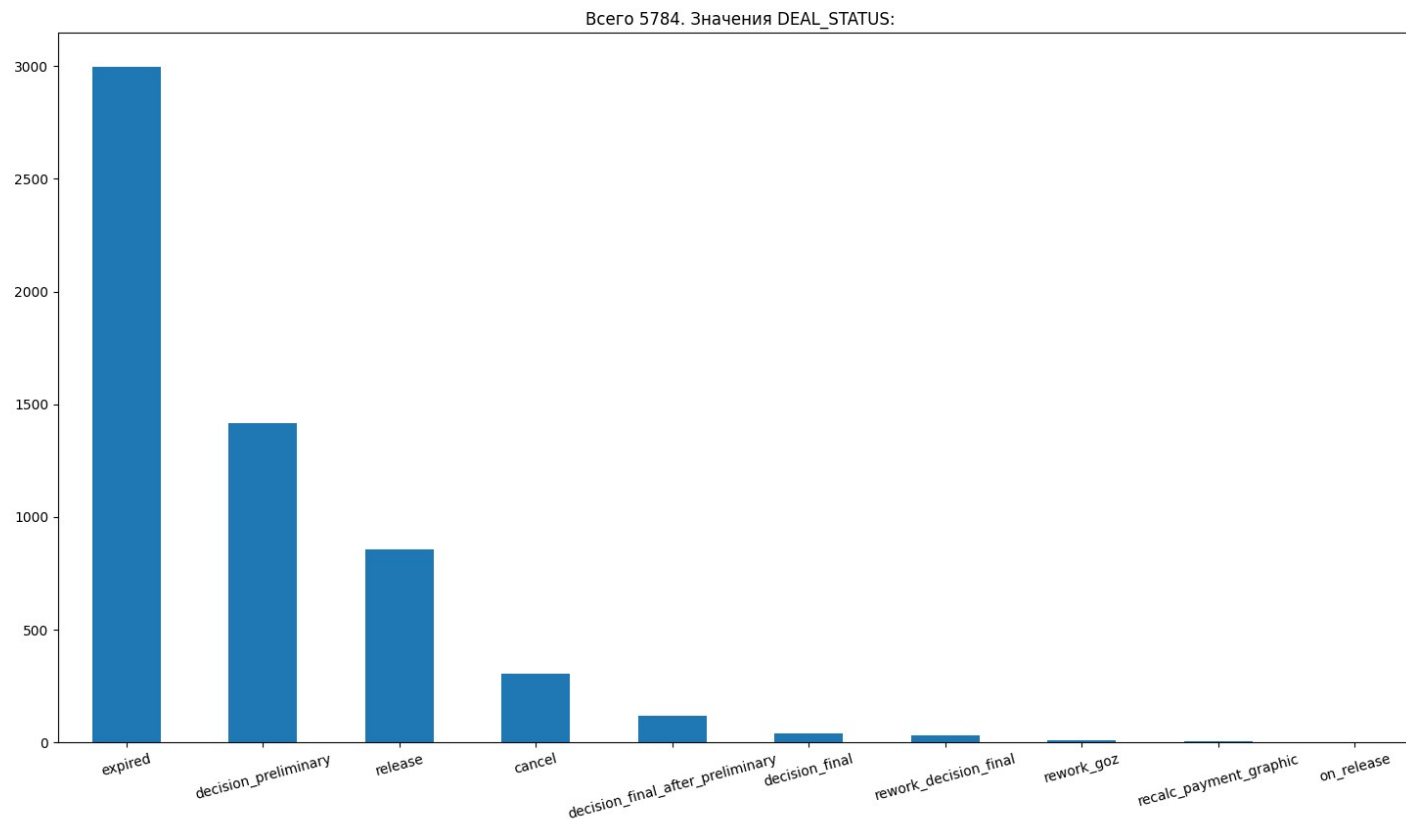
Вероятностный классификатор — шкалирование вероятностей

1) Определим пределы для классов одобрения и отклонения:

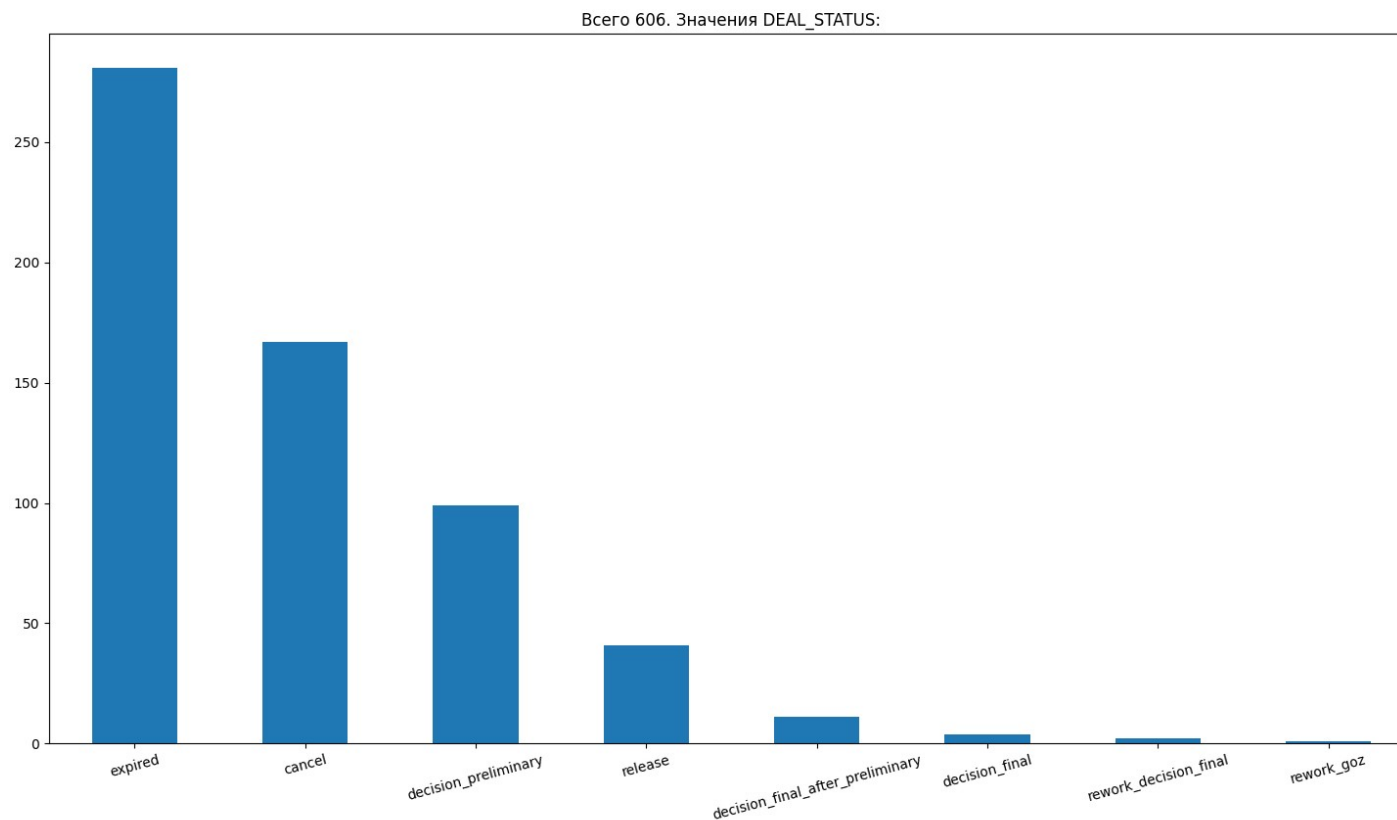
- Одобрения: $0.1 + (950 * 0.6) / 255 - 2.2 + 0.1 + 0.02 = 0.255$
- Отклонения: $-0.45 - (0 * 0.6) / 255 - 2.2 - 0.3 - 0.15 = -3.1$

2) Степень уверенности будет = вес/0.255 и вес/-3.1

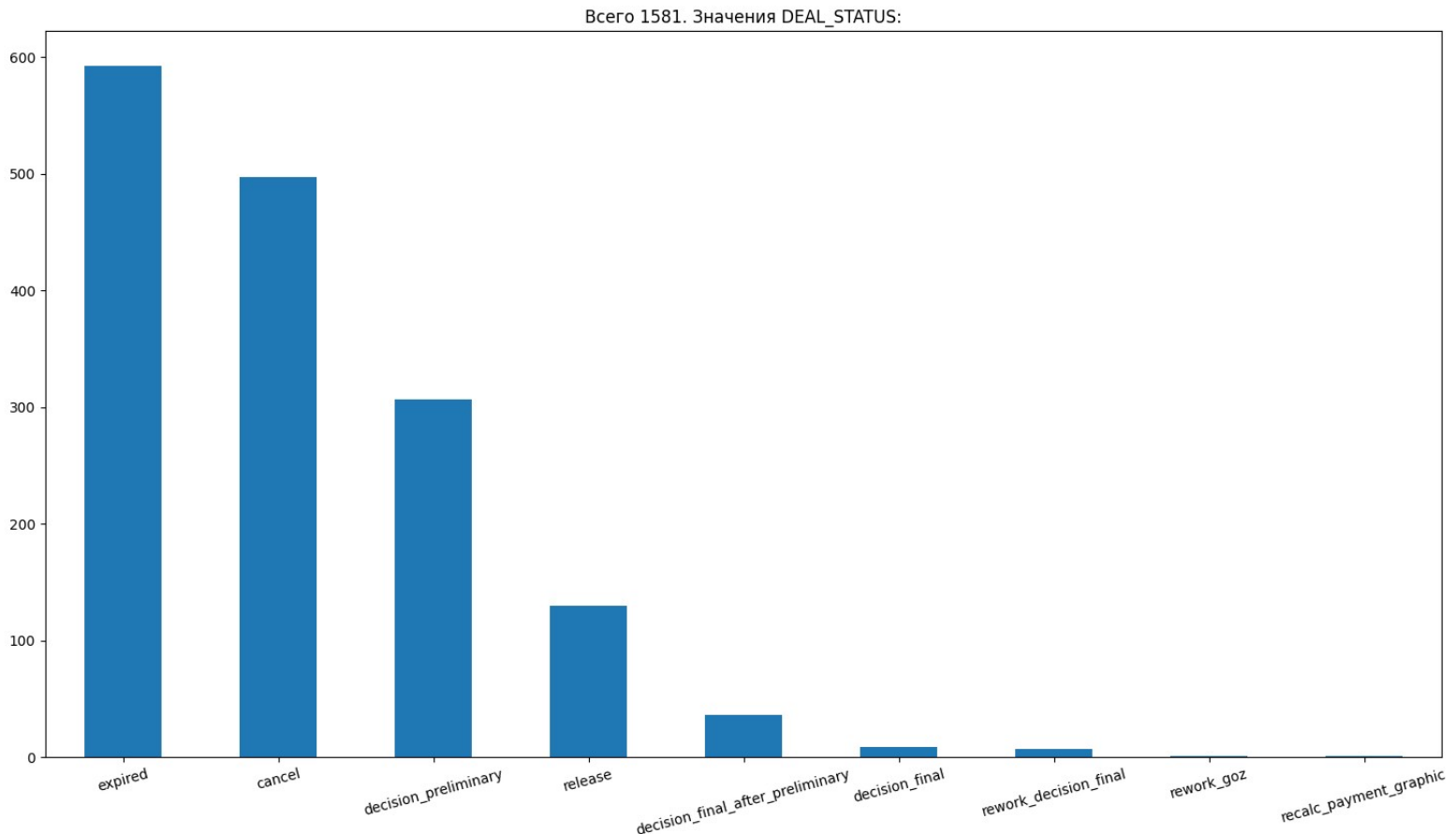
Статус договора на выборке валидации у одобренных



Статус договора на одобренных лучшей моделью



Статус договора на одобренных вероятностным классификатором



Большее количество cancel по сравнению с одобренными изначально, это все ошибочно одобренные.



Грубый перебор - описание

Признаки разбитые на 10 диапазонов:

Бинарные признаки:


Грубый перебор - недостатки

- Недостатки:
- - слабый и сильный признак делят данные одинаково
- - границы не четкие
- - игнорируются разряженные скопления

Грубый перебор — Анализ 1

Метрики для 6 ячеек:

- Одобренных: 0.000228
- Accuracy: 0.697420
- Precision: 1.000000
- Recall: 0.000752



Кластеризация и грубый перебор, минусы

- Перебор — это ячейки с частым попаданием одобренных
 - выч. сложность
 - жесткие границы ячеек
 - нетривиальность
- Кластеризация — одобренные сгруппированные по близким ячейкам
 - субъективность

Кластеризация - подготовка

Взяты 10000 случайных заявок

Столбцы стандартизированы и нормализованы.

Удалены столбцы коррелирующие с «Оценка кредитной истории Эквифакс»:

- OKB_RATING_SCORING_Хорошая КИ
- OKB_RATING_SCORING_КИ отсутствует
- NBKI_RATING_SCORING_КИ отсутствует
- NBKI_RATING_SCORING_Хорошая КИ

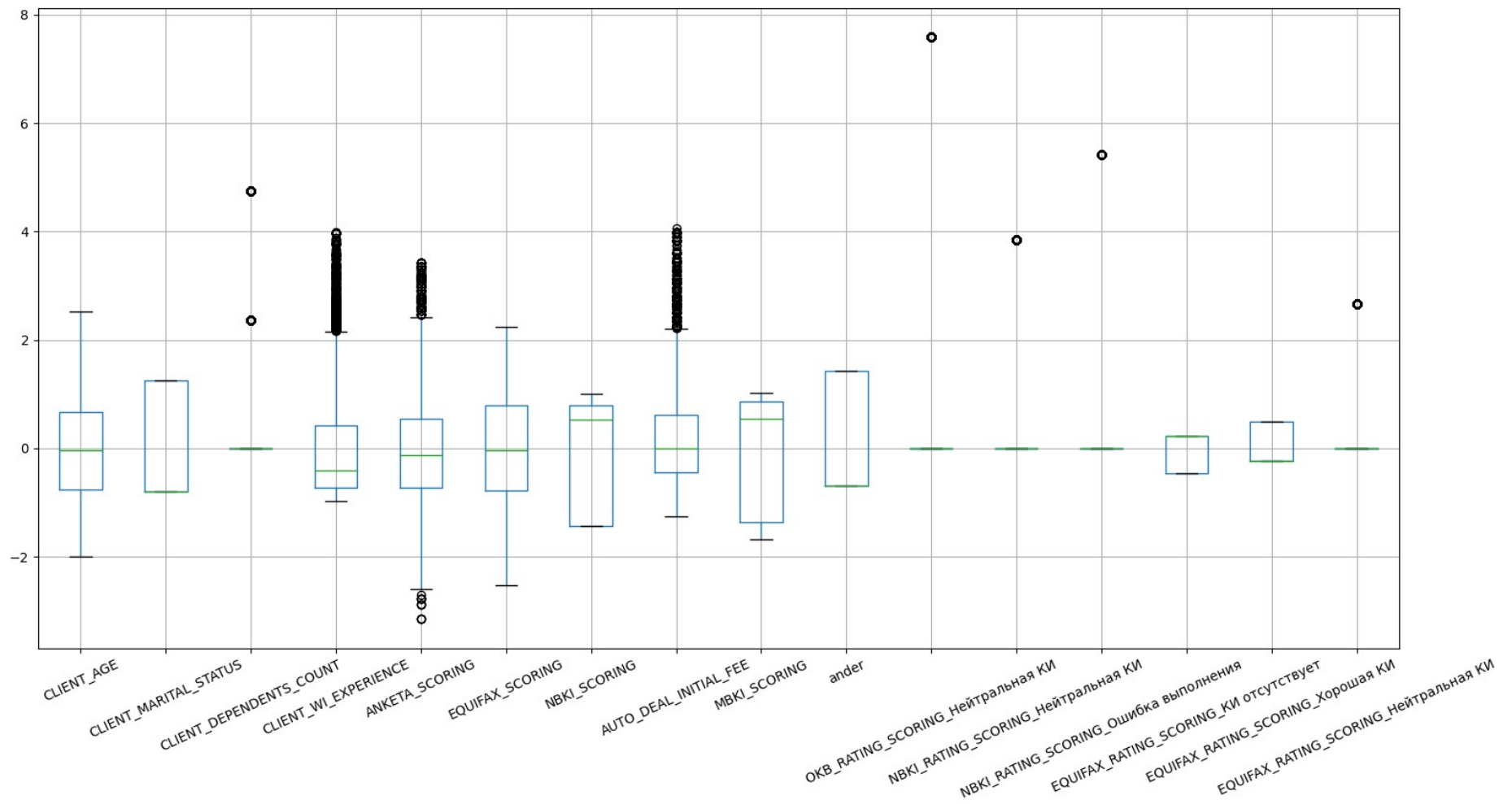
Удалены строки:

- Первоначальный взнос > 700000 - 2519
- Стаж в организации (мес.) > 300 - 1081
- Скоринговый балл клиента (анкетный скоринг) > 150 — 679
- Кол-во иждивенцев >= 3 — 888
- Скоринг Бюро Эквифакс 4Score < 100 — 304

Удалены столбцы с малым количеством значений:

- 'OKB_RATING_SCORING_Плохая КИ'
- OKB_RATING_SCORING_Ошибка выполнения
- EQUIFAX_RATING_SCORING_Ошибка выполнения

Кластеризация - подготовка



Кластеризация - подготовка

Алгоритмы кластеризации, как правило, используют меру расстояния $|x_1 - x_2|$ требующие, чтобы стандартное отклонение столбцов было равно единицы, чтобы обеспечить равный вклад.

PCA алгоритм требует центрирование данных, так чтобы центр находился в области интересующих значений. Медиана для смещенных данных, это которые не соответствуют нормальному распределению, дает среднее ближе к частым значениям, а среднее арифметическое дальше.

Сильно смещены:

- CLIENT_WI_EXPERIENCE — арифметическое: 58, медиана: 36
- AUTO_DEAL_INITIAL_FEE — арифметическое: 180076, медиана: 158000

Из анализа лучшей модели, мы знаем, что для CLIENT_WI_EXPERIENCE переломной точкой является значение 58, поэтому будем использовать среднее арифметическое. А для AUTO_DEAL_INITIAL_FEE, значения от 50000, до 300000, среднее 125000, что ближе к медиане, поэтому будем использовать медиану.

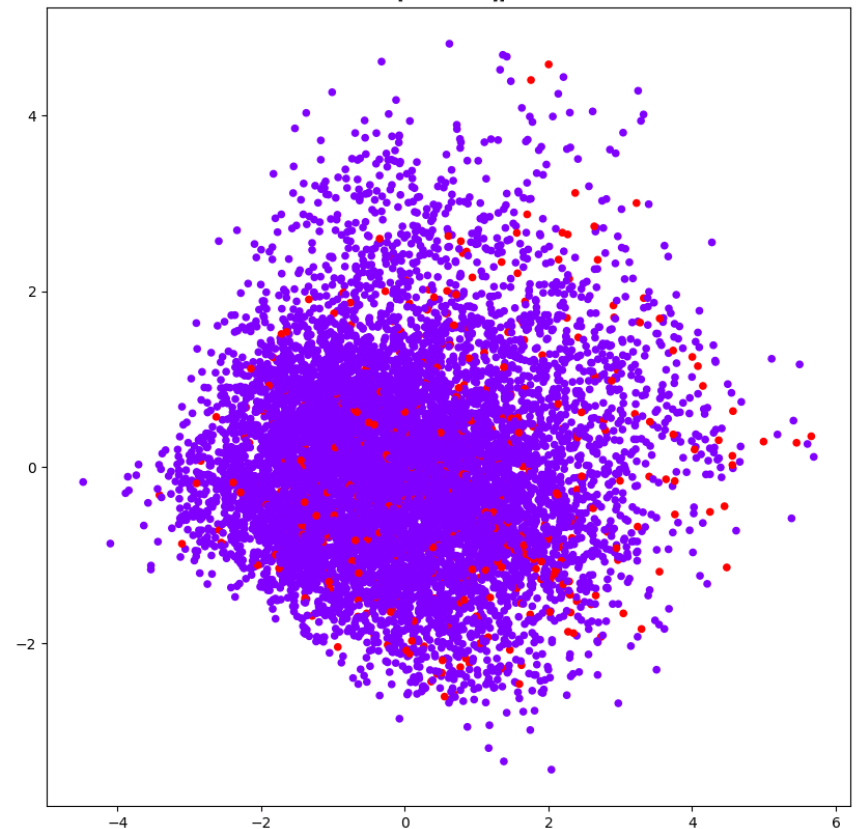
Для бинарных-смещенных столбцов будем использовать медиану, потому что она соответствует самому частому значению, что исключает ненужное смещение для PCA.

При One-Hot кодировании уменьшим вес полученных столбцов на количество новых столбцов.

Кластеризация

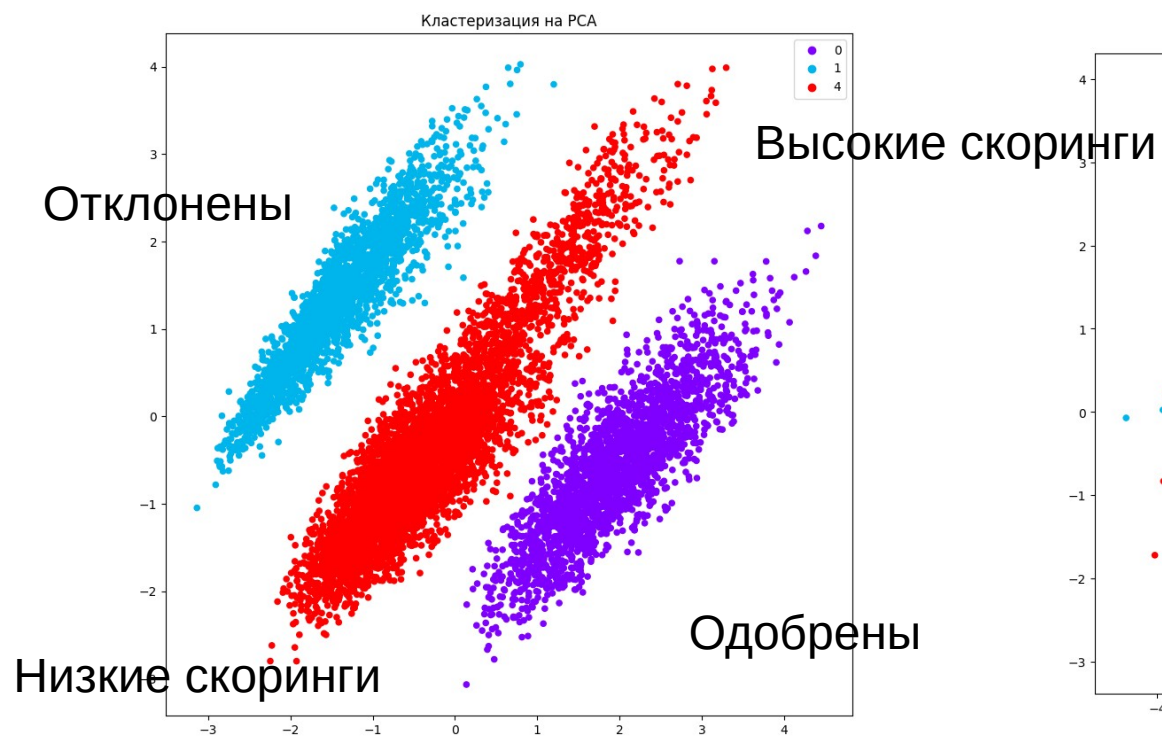
- Кластеры — это сгустки похожих заявок.
- Изменим вес «Решения Андерайтера» от 0.001 до 3 и проанализируем разбиение на 2 кластера.
- При увеличении веса признака «Решение андерайтера» кластеры из своей оригинальной формы разделяются на два облака.
- Как можно видеть, кластеры похожи на вытянутые скопления

Условная форма кластеров построенная по уменьшенной размерности 0.001.png
calinski_harabasz_score 45.37523172763923
contingency_matrix [[6450 288]
[3033 229]]

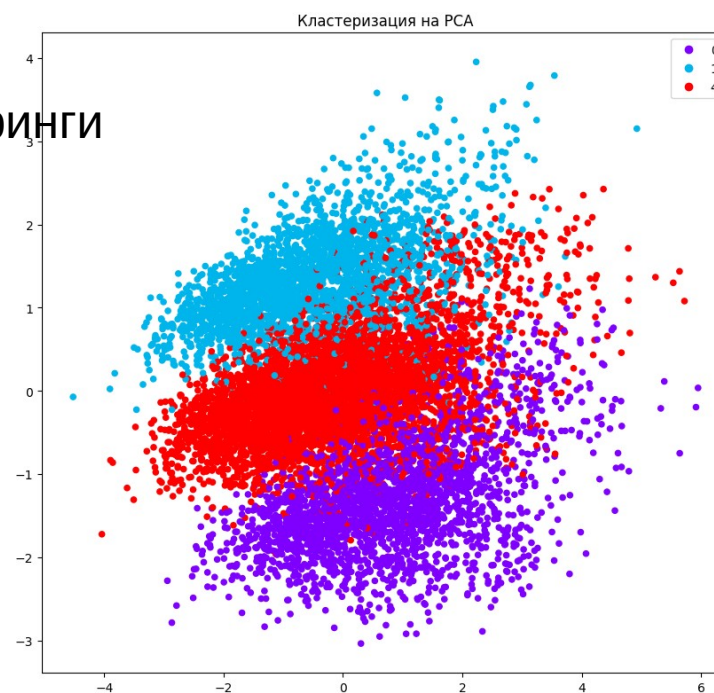


3 кластера

При весе=2.8 «Решение
Андерайтера», с
раставленными акцентами



При весе=2.2 «Решение
Андерайтера», без акцентов



Вытянуты вдоль числовых скорингов

3 кластера

- 1 Отклонено/Одобрено - 2450/0
- 2 Отклонено/Одобрено - 4286/764
- 3 Отклонено/Одобрено - 2/2498

В 3 кластере:

- Кредитная история «Хорошая» или «КИ отсутствует»
- доля клиентов в браке примерно 70%
- «НБКИ скоринговый бал» высокий
- Часто без первоначального взноса
- Эквифакс скоринговый бал смещен в область высоких значений

В 1 кластере:

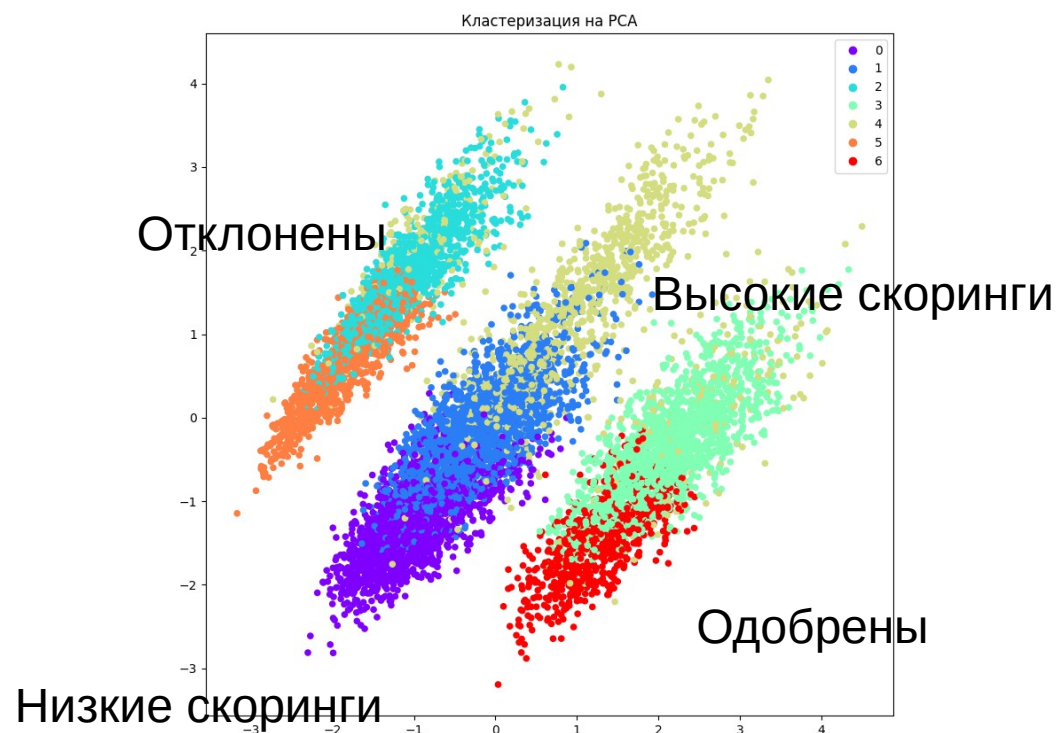
- Кредитная История почти всегда имеет значение «КИ отсутствует»
- «НБКИ скоринговый бал» всегда около нуля или отсутствует

Кластер 2 выделяет наличие НБКИ скорингового бала отличного от нуля.

Как можно видеть, заметную роль играет признак «КИ отсутствует», «Эквифакс скоринговый бал» и коррелирующий с ним «скоринговый бал ОКБ» заметно изменяют свою форму, что подтверждает их важность.

3 кластера - детально

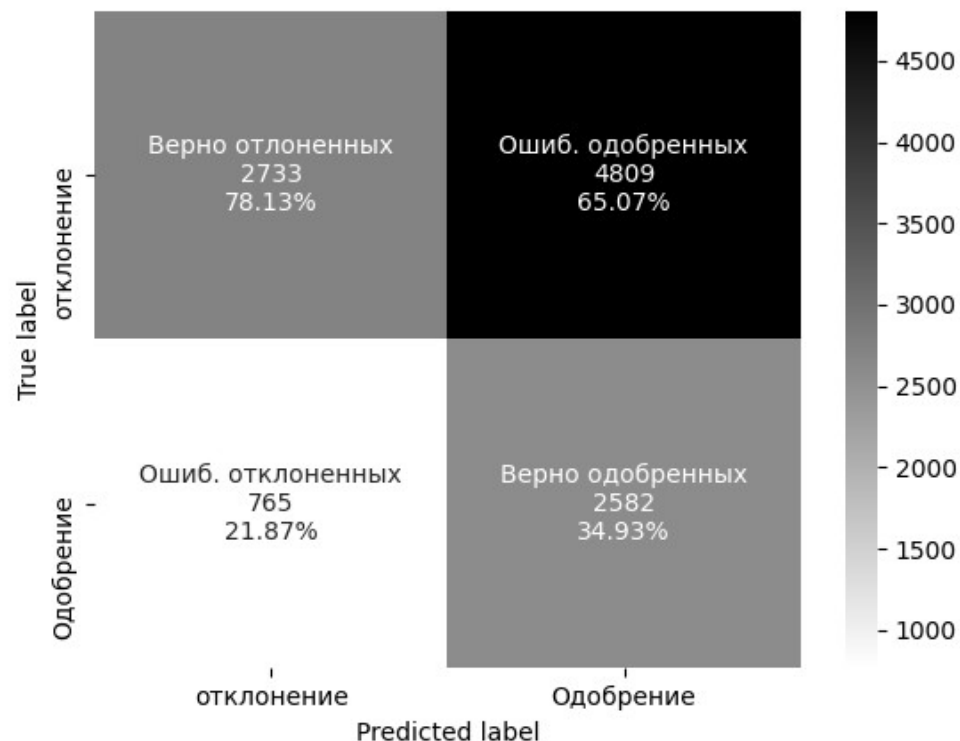
- 4 кластер, это НБКИ скоринг отсутствовал
- В 3 кластере более высокие скоринги, чем в 6
- В 2 кластере более высокие скоринги, чем в 5



3 кластера — модель — неудача

Обучим классификатор находить 3-й кластер или (3-й и 6-й), в котором все одобрены и проверим, какая его эффективность определять решения андерайтера.

- Построенная лучшая модель по выделению кластера имеет:
- Вероятность точного одобрения 35%, доля одобренных 68% от общего числа заявок.
- Такая низкая эффективность объясняется тем, что при построении кластера мы знали Решение Андерайтера, а классификатор был обучен на отобранных одобренных, что помешало ему сделать лучшее обобщение. Или наше предположение неверно.
- Это значит, что рассмотрение одной этой области одобренных заявок не позволяет предсказать решение андерайтера.



Кластеризация одобренных

Релаксация RBF преобразованием

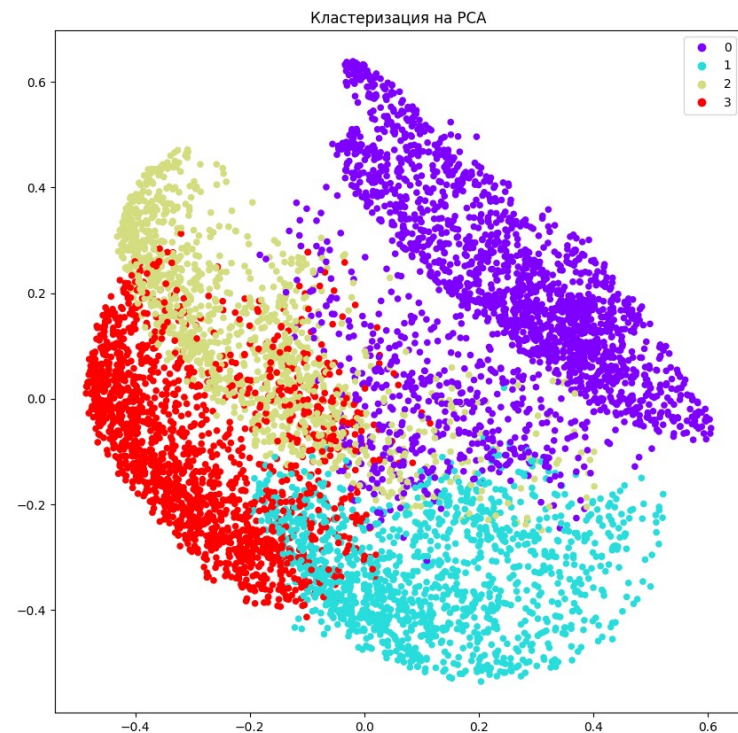
Кластер — количество — характер

- 0 — 2359 — с опытом работы
- 1 — 1275 — в браке с опытом
- 2 — 1180 — не в браке
- 3 — 1693 — с малым опытом

Высокие скоринги в 1 кластере.

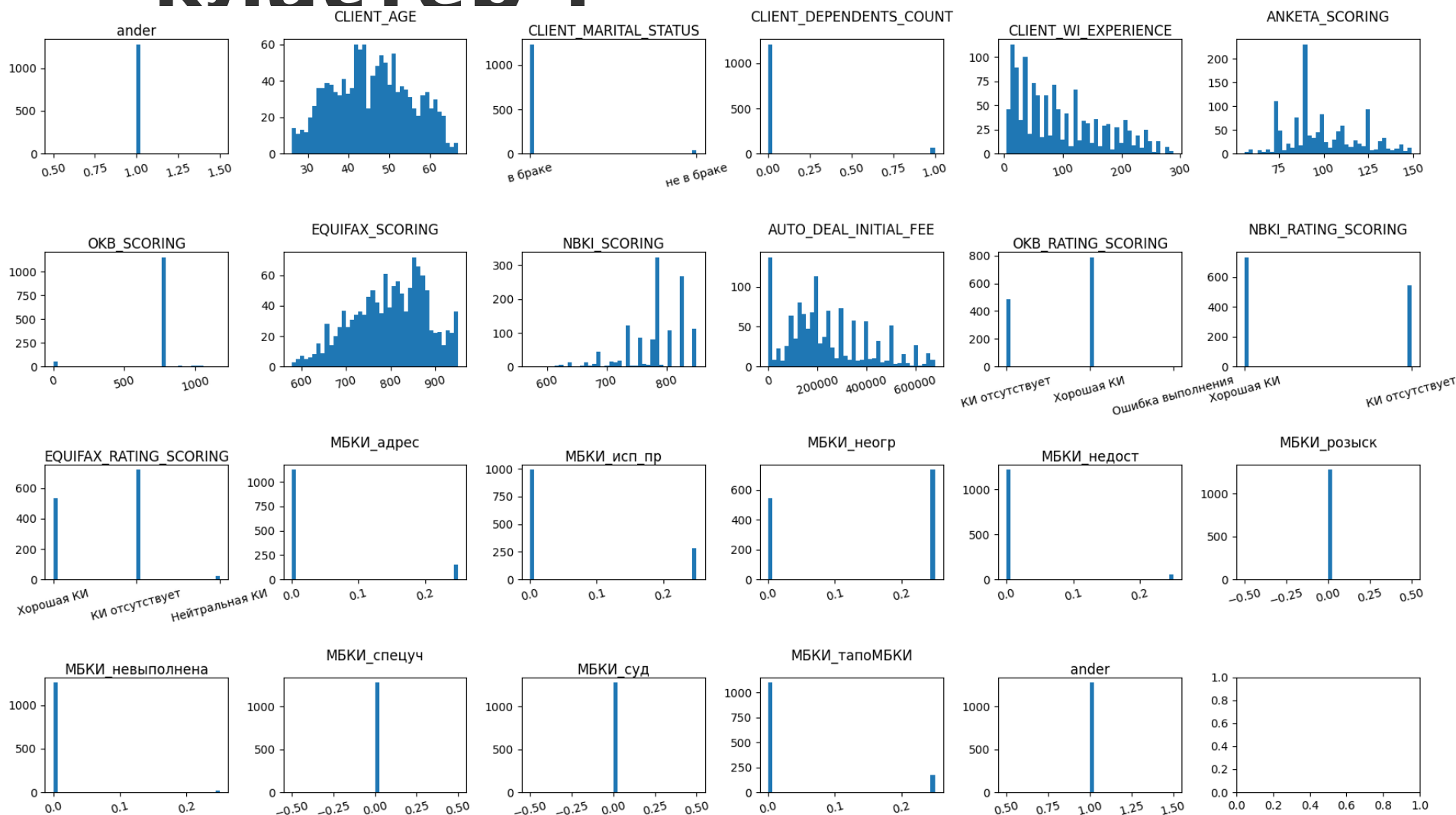
Характеризуется:

- возрастом выше среднего
- Высоким Первоначальным взносом
- Хорошая КИ
- в браке



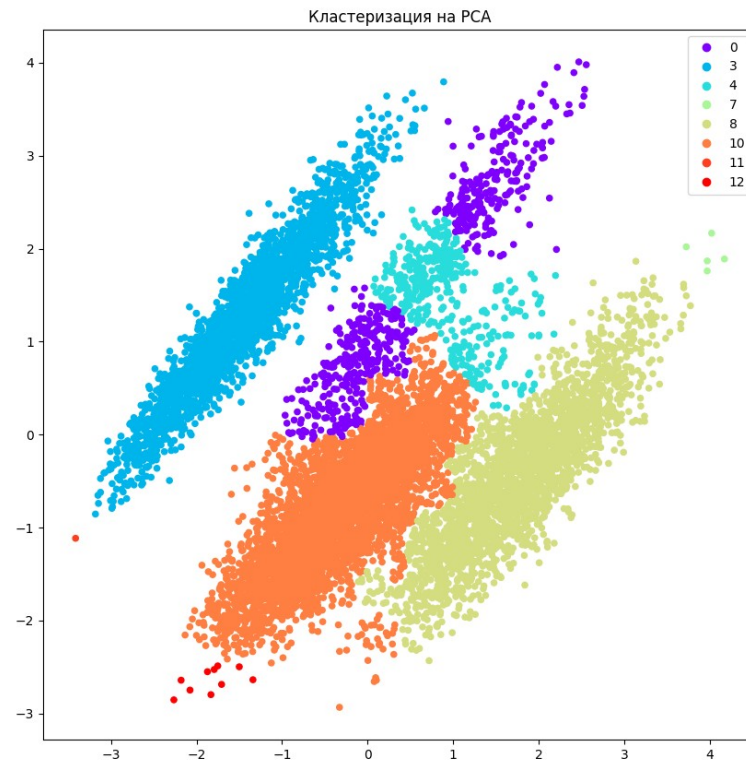
Кластеризация одобренных

— кластер 1



4 кластера

- При весе=2.2-2.5 «Решения Андерайтера», можно обнаружить 4 кластера.



4 кластера (1)

В динамической картине, мы видим поворот, потому, что происходит отделение одобренных заявок похожих на отклоненные и отклоненных похожих на одобренные.

Крайние Одобрённые и отклонённые остаются на местах, потому, что они больше всего соответствуют себе, это кластеры 1 и 3 из пункта «3 кластера»

4 кластера (2)

- 3 Отклонено/Одобрено — 2446/0
- 0 Отклонено/Одобрено — 72/477
- 10 Отклонено/Одобрено — 4057/144
- 8 Отклонено/Одобрено — 25/2371

Кластер 10 и 8 объединяет наличие положительного скоригового бала НБКИ

Кластер 0 и 3 объединяет преимущественно нулевое значение скоригового бала НБКИ


Крайние кластеры 3 и 8 совпадают с кластерами 1 и 3 из пункта «3 Кластера»

Итоги

- Была найдена важность признаков
- Были найдены зависимости решения андерайтера от признаков
- Было создано 2 синтезированные модели, которые объясними

Далее:

- - рассмотреть редкие случаи отклонения
- - улучшать модель путем: очистки и обогащением данных
- - добавить в кластеризацию сомнительные признаки (полагаем, что большинство правдивые)



Кластеризация — дополнительные признаки (в процессе)

- Для кластеризации добавим поля, которые не являются достоверными, но в большинстве характеризуют клиента:
 - ы