Обзор решения задачи прогнозирования вероятности блокировки счета

Модель предсказывает вероятность блокировки РКО после его открытия в течение 6 месяцев

Перечень данных, имеющихся для решения задачи:

1. Список всех поступивших заявок в период с 2018-06-18 по 2019-12-14. (Заявки.xlsx)
2. Список всех открытых счетов. (Открытые рс.xlsx)
3. Список компаний, по которым счет был заблокирован. (Блокировки.xlsx)
4. Сведения из СПАРК по всем компаниям из заявок на открытие РКО. (СПАРК.csv)
5. Список ИНН, у которых открыт кредитный продукт в Банке. (Наличие_кредитного_продкута.xlsx)

Информация из данных для формирования выборки:

- Размер данных: 324 тысячи наблюдений
- Количество уникальных ИНН: 243 тысячи

- Размер данных: 246 тысяч наблюдений
- Количество уникальных ИНН: 218 тысяч

- Размер данных: 31 тысяча наблюдений
- Количество уникальных ИНН: 30 тысяч

- Размер данных: 124 тысячи наблюдений
- Количество уникальных ИНН: 100 тысяч

- Количество уникальных ИНН: 9 тысяч

Формирование выборки для обучения

Определения целевого события и разметка данных на основе таблиц об открытии и блокировке РКО

Логика: Если один из счетов заблокирован в течение 6 месяцев после открытия, то блокируется и ИНН.

Рассматриваются только те клиенты, дата блокировки которых позже, либо равна дате открытия

-11 % наблюдений

Формирование признаков из доступных данных на момент подачи и обработки заявки открытия РКО

Логика: Если подано несколько заявок, рассматривается только последняя обработанная сотрудником заявка по ИНН

Рассматриваются только те клиенты, дата загрузки заявки которых позже, либо равна дате подачи заявки

-68 % наблюдений

Рассматриваются только те клиенты, дата загрузки заявки которых раньше либо равна дате открытия счета

Итого 65 тысяч уникальных ИНН
11% целевого события

Создание и отбор признакового пространства

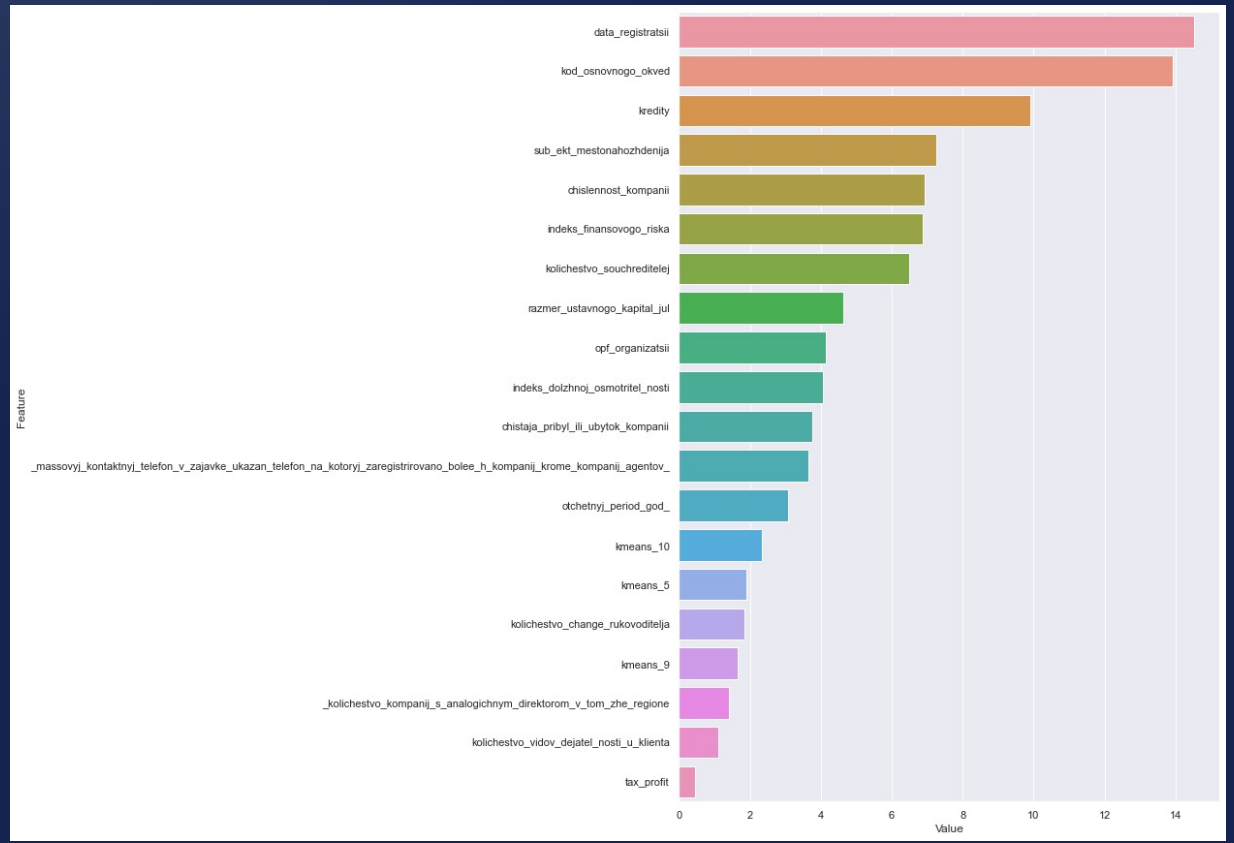
Создание признаков на основе имеющейся информации:

1. Создание признаков на основе финансовой отчетности: ROA, ROE, Сумма налогов/ чистая прибыль.
2. Создание признаков на основе количества заявок, открытий, закрытий счетов, а также смен руководителей.
3. Создание признаков на основе кластеризации алгоритмом Kmeans от 2 до 10 классов.
4. Редактирование и исправление ошибок в отчетности по выручке, уставному капиталу.

Методы отбора признаков:

1. Удалены признаки с количеством пропусков > 90%.
2. Удалены признаки с корреляцией > 90%.
3. Удалены признаки с gain = 0, при обучении алгоритма CatBoostClassifier.
4. Удалены признаки с PermutationImportance < 0, при обучении алгоритма от библиотеки Eli5.
5. Удалены признаки с помощью алгоритма Backward selection.

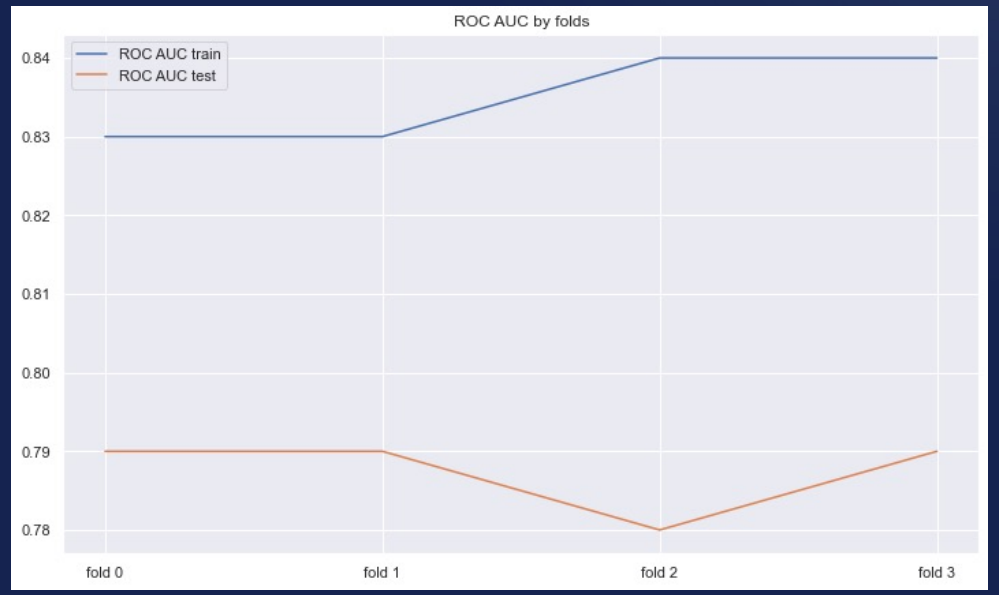
Значимость признаков в финальной модели:



Результаты итоговой модели

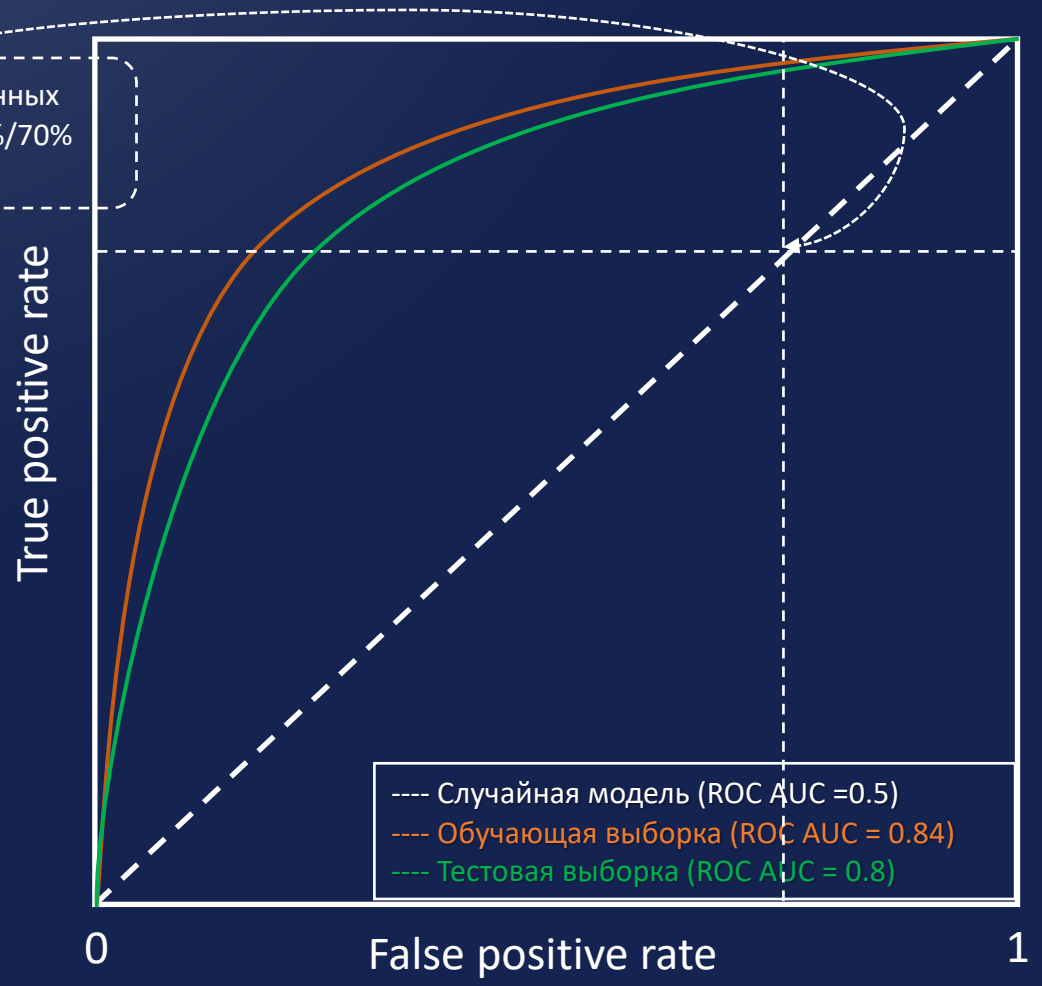
Алгоритм, используемый в обучении – градиентный бустинг от библиотеки CatBoost. *

Обучение проводилось при разбиении на Обучение и Тест с помощью StratiffiedGroupFold на 4-х фолдах в отношении 80/20 от генеральной выборки.



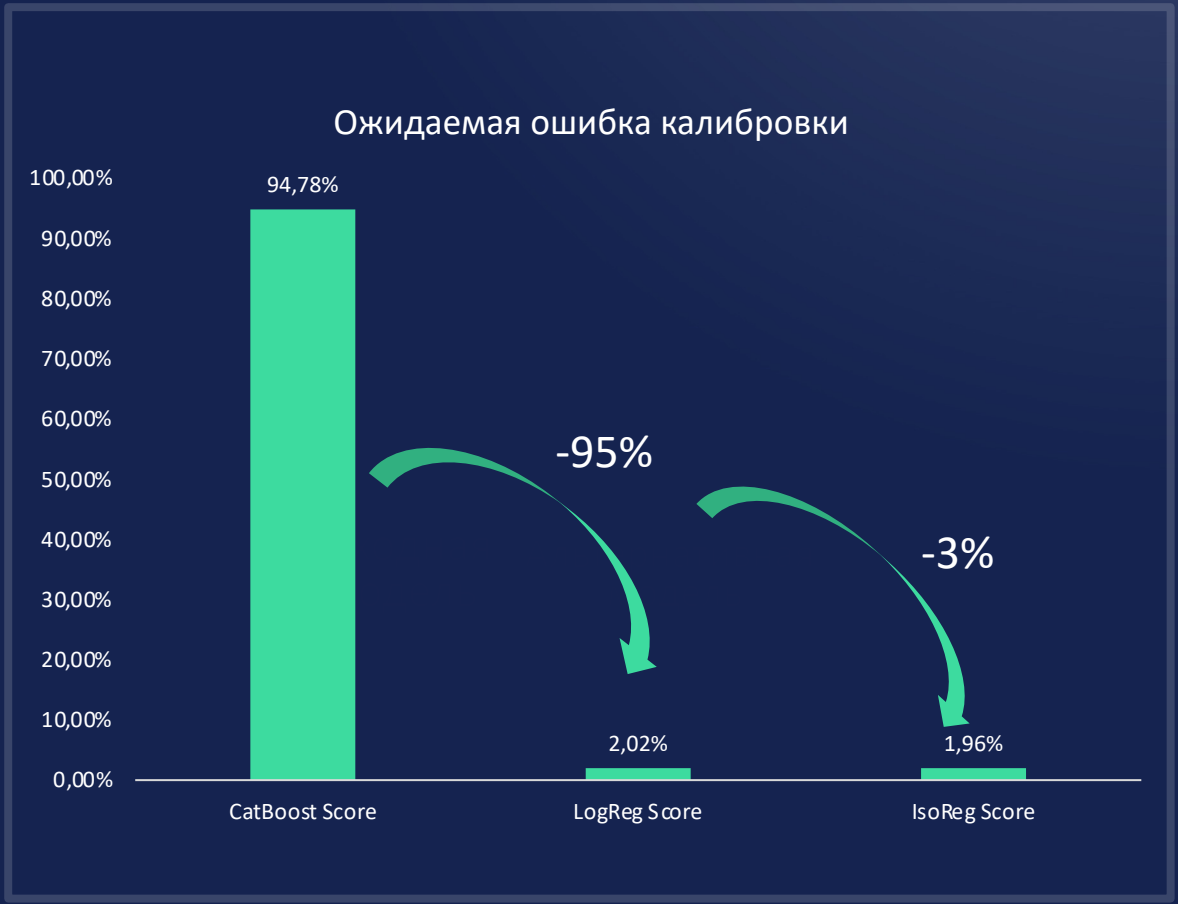
Модель является стабильной и не склонна к переобучению

На 30% от предсказанных клиентов Recall = 77%/70% (Обучение/Тест)



Калибровка, ожидаемый финансовый результат и дальнейшие пути развития модели

Результаты калибровки скорра, полученного градиентным бустингом:



Предположения о расчете финансового результата от применяемой модели:

Ошибка 2-го является более значимой, относительно ошибки 1-го рода, так как мошенничество или дефолт организации принесет высокие издержки Банку. Следовательно бизнес метрикой для расчета финансового эффекта является Recall. То есть количество верно предсказанных клиентов * на издержки для Банка (сумма на момент дефолта или сумма денежных средств, попавших под мошенничество)

Зона для улучшения и рекомендации:

1. Рассмотреть количество поданных заявок и открытий на момент обработки сотрудником / моделью.
2. Рассмотреть расширенный список факторов

Спасибо за внимание

Приложение

Интерпретация количественных признаков модели