

Классификация: предсказание оттока клиентов телеком- компании



Проблема

В телекоммуникационной индустрии удержание клиентов является одной из ключевых задач для повышения прибыльности и снижения расходов.

Зачастую многие клиенты могут отказаться от услуг компании, что негативно сказывается на стабильности доходов.

В этой связи возникает необходимость точно прогнозировать отток клиентов, чтобы своевременно принимать меры по их удержанию.

Необходим анализ данных о клиентах, таких как демографические показатели, услуги, платежи и историю использования услуг. Разработка модели, которая сможет предсказать вероятность оттока на основе имеющихся признаков, что позволит компании:

- сфокусировать маркетинговые усилия на группах риска;
- повысить эффективность программ лояльности;
- снизить уровень оттока и увеличить пожизненную ценность клиента (Customer Lifetime Value)

Критерии успеха

Доказательство качества модели

- ROC-AUC не менее 0.75 на валидационной выборке, что свидетельствует о хорошей способности модели отличать клиентов, склонных к оттоку, от тех, кто останется.

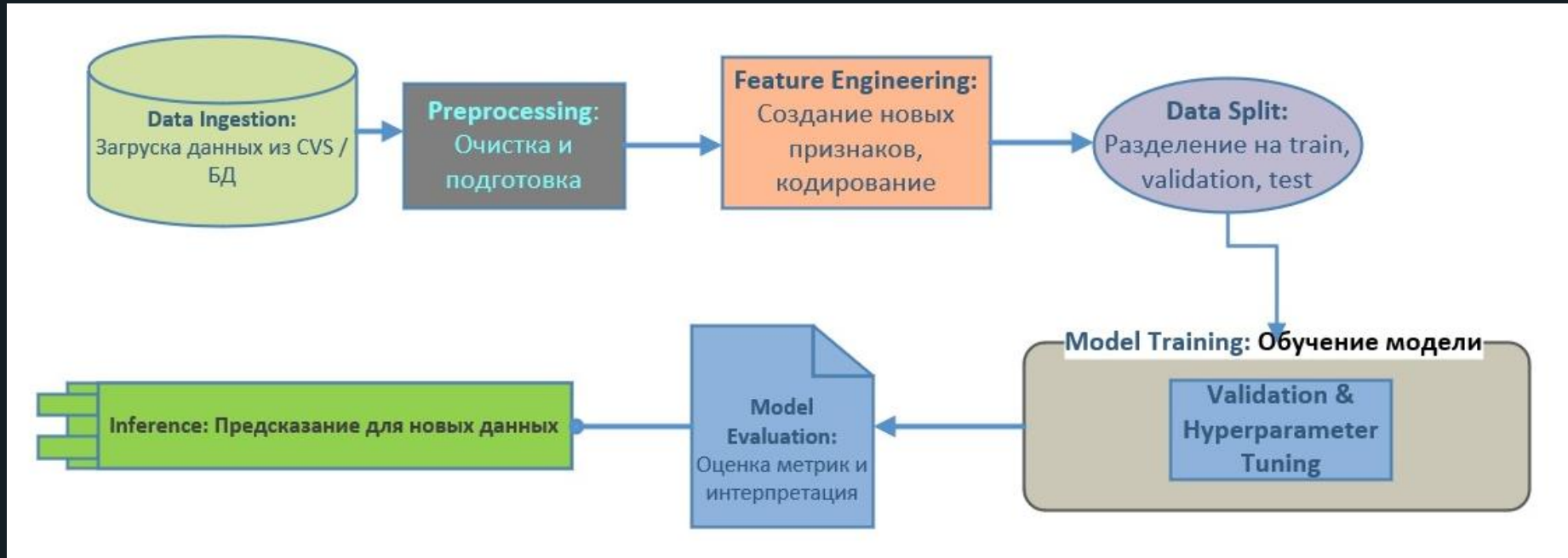
Интерпретируемость и анализ факторов

- В модели нужно выявить ключевые особенности, влияющие на решение о покидании услуги (например, использование определённых услуг, платежные показатели, демографические признаки).
 - Итоговые отчёты должны включать список наиболее значимых факторов риска с бизнес-обоснованиями.

Документированность и воспроизводимость

- Вся разработка сопровождается полной документацией (внутри проекта и в репозитории).
- Проведены тесты на воспроизводимость и стабильность модели.

Архитектура и ключевые решения



Структура репозитория

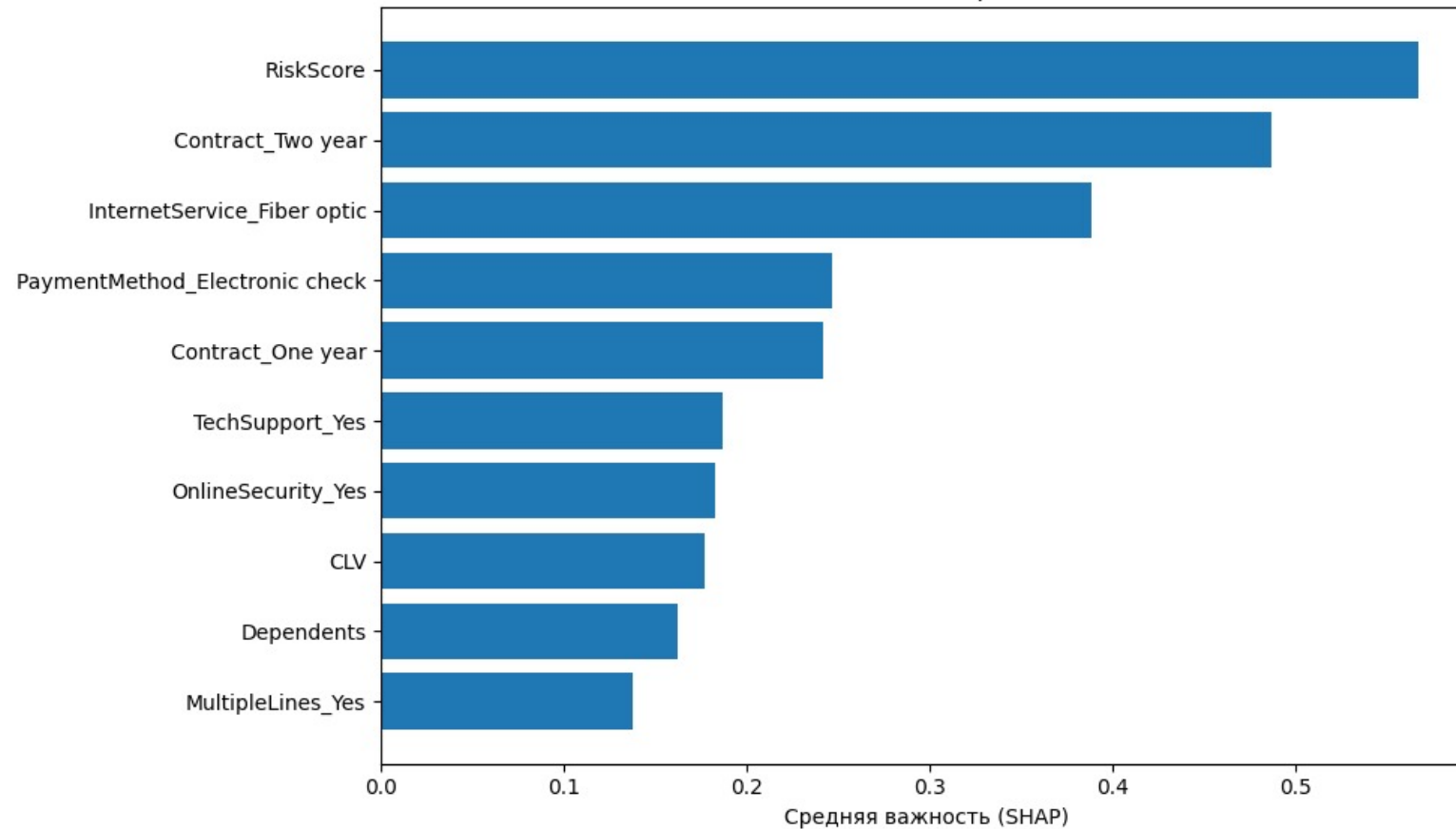
```
/project
├─ data/
│  └─ processed/ # Обработанные данные
│  └─ raw_data/  # Исходные данные
├─ docs/         # Документация
│  └─ images/    # Изображения для документации
├─ models /      # Модели
├─ notebooks/    # Jupyter-ноутбуки и .py-файлы с кодом
│  └─ encoders/  # Содержит словарь расшифровки данных
│  └─ pckgs/     # Самописные функции
├─ plots /       # графики
├─ tests /       # Тесты
├─ requirements.txt
└─ README.md
```

Ключевые архитектурные решения

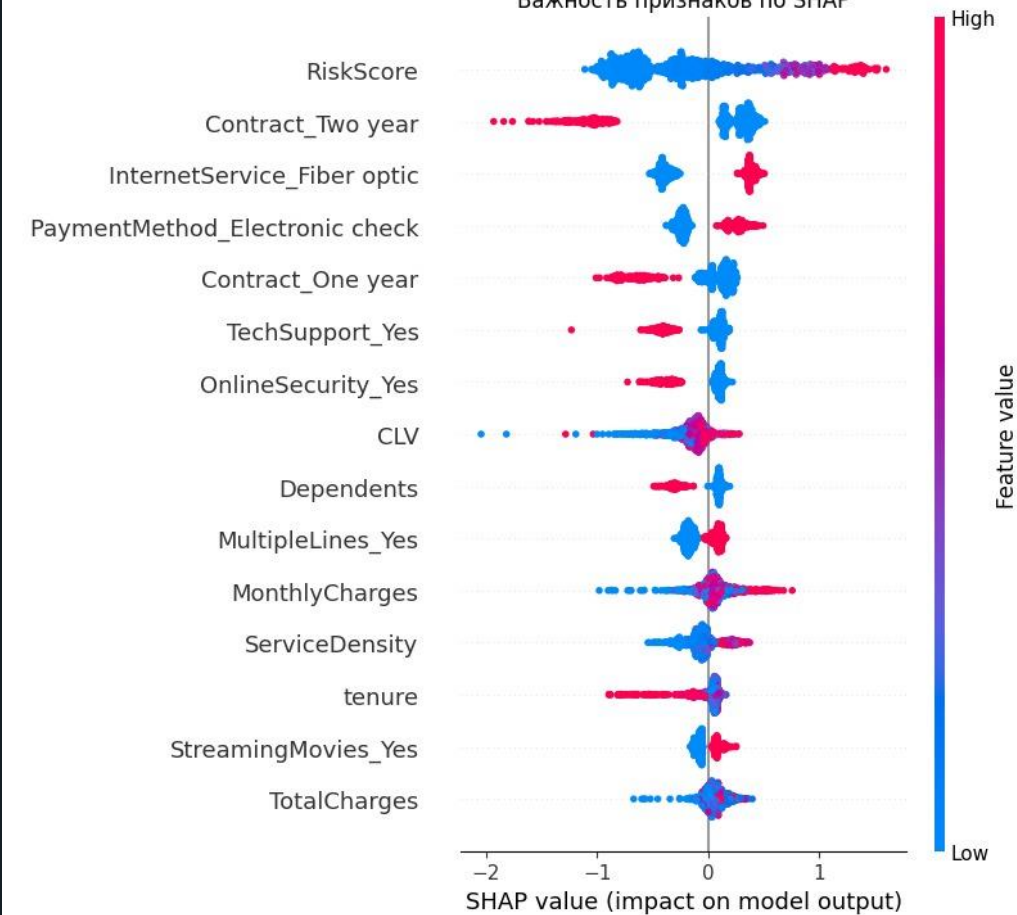
- **Выбор алгоритмов:** LightGBM и CatBoost 😬 за быструю работу с категориальными признаками и хорошую точность.
- **Обработка категориальных признаков:** встроенные механизмы CatBoost или One-Hot кодирование.
- **Валидация:** стратифицированная кросс-валидация для сохранения пропорций классов.
- **Воспроизводимость:** зафиксированные сиды (random_state=42), версии пакетов.
- **Обработка дисбаланса:** использование метрик с учетом дисбаланса, а также возможной взвешенной выборки.

Метрики качества

Топ-10 важных признаков



Важность признаков по SHAP



- Модель опирается на логичные с бизнес-точки зрения факторы (риск-скор, тип и длительность контракта, набор услуг, способ оплаты и стаж клиента), что подтверждает её адекватность и интерпретируемость, а значит — высокое практическое качество прогнозов.

Риски

Таблица TRADE-OFF 'чувствительность-точность'

| Порог | Precision | Recall | F1-Score | Клиентов с риском | Доля рисковых |
|-------|-----------|--------|----------|-------------------|---------------|
| 0.1 | 0.388 | 0.947 | 0.551 | 912 | 64.8% |
| 0.2 | 0.449 | 0.890 | 0.597 | 742 | 52.7% |
| 0.3 | 0.482 | 0.799 | 0.602 | 620 | 44.1% |
| 0.4 | 0.545 | 0.741 | 0.628 | 508 | 36.1% |
| 0.5 | 0.583 | 0.636 | 0.609 | 408 | 29.0% |
| 0.6 | 0.631 | 0.527 | 0.574 | 312 | 22.2% |
| 0.7 | 0.670 | 0.374 | 0.480 | 209 | 14.9% |
| 0.8 | 0.765 | 0.209 | 0.328 | 102 | 7.2% |
| 0.9 | 0.933 | 0.075 | 0.139 | 30 | 2.1% |

1. Максимальный F1-Score:

Порог: 0.4, F1: 0.628

Precision: 0.545, Recall: 0.741

Клиентов для обзвона: 508

2. Найти >80% ушедших:

Порог: 0.1, Recall: 0.947

Клиентов для обзвона: 912

3. Высокая точность (>70%):

Порог: 0.9, Precision: 0.933

Клиентов для обзвона: 30

- **Компромисс между точностью и полнотой.**

Оптимальный по F1 порог 0.4 (Precision 0.541, Recall 0.741)

– из 508 клиентов для обзвона ~47% не уйдут, а ~25% реально уходящих мы не найдем.

- Риск: часть бюджета уйдёт на «ложные срабатывания», при этом мы всё равно теряем четверть уходящих.

- **Сценарий «поймать максимум уходящих» (порог 0.1, Recall 0.947).** • Обзваниваем 991 клиента, «рисковыми» помечено 70% базы.

- Риск: перегруз контакт-центра и рост затрат на удержание при заметной доле ошибочных обращений.

- **Сценарий «очень высокая точность» (порог 0.9, Precision 0.933).** • Всего 38 клиентов для обзвона, из базы попадает только 2.7%.

- Риск: модель почти не используется — подавляющее большинство уходящих клиентов не будет обнаружено.

РИСКИ (стоимость ошибки)

Были взяты следующие стоимости ошибок:

False Positive (ложное срабатывание) = 100 руб

Стоимость звонка менеджера: 50 руб

Предоставляемая скидка клиенту: 50 руб

False Negative (пропуск оттока) = 500 руб

Потеря среднемесячного дохода с клиента

На основе этих стоимостей был рассчитан бизнес-оптимальный порог.

Очень низкий бизнес-порог (0.20) — это значит, что бизнесу дешевле звонить многим клиентам, чем пропускать уходящих!

Порог 0.20 означает:

Бизнес говорит: "Лучше позвонить 10 лишним, чем пропустить 1 уходящего"

Причина: Потеря клиента (500 руб) дороже ложного звонка (100 руб)

Соотношение: $500/100 = 5 \rightarrow$ готовы к 5 ложным звонкам, чтобы не пропустить 1 уходящего

Сегментация клиентов по риску оттока

| Сегмент | Клиентов | Доля | Ср. вероятность | Ожидаемый отток |
|------------------|----------|-------|-----------------|-----------------|
| Критический риск | 102 | 7.2% | 86.8% | 89 чел |
| Высокий риск | 210 | 14.9% | 70.0% | 147 чел |
| Средний риск | 196 | 13.9% | 50.0% | 98 чел |
| Низкий риск | 899 | 63.9% | 12.3% | 110 чел |

Бизнес-порог: 0.20

Минимальная стоимость: 61,400 руб

РИСКИ (РЕКОМЕНДАЦИИ ПО УДЕРЖАНИЮ ДЛЯ КАЖДОГО СЕГМЕНТА)

| Сегмент | Приоритет | Вероятно сть ухода | Меры удержания | Бюджет на клиента | Цель |
|---------------------|---|-----------------------|---|------------------------------------|---|
| Критический риск | ВЫСШИЙ (действо вать немедленно) | >80% | <ul style="list-style-type: none">• Личный звонок топ-менеджера в течение 24 часов• Персональное предложение: скидка 20-30% на 6 месяцев• Бесплатный апгрейд тарифа на 3 месяца• Назначить персонального менеджера | Высокий (500-1000 руб/клиент) | Снизить отток на 60-70% |
| Высокий риск | ВЫСОКИЙ (действ овать на этой неделе) | 60-80% | <ul style="list-style-type: none">• Звонок менеджера по удержанию в течение 3 дней• Предложение: скидка 15% на 3 месяца• Бесплатная дополнительная услуга на 1 месяц• Опрос о причинах недовольства | Средний (200-500 руб/клиент) | Снизить отток на 40-50% |
| Средний риск | СРЕДНИЙ (проакти вная работа) | 40-60% | <ul style="list-style-type: none">• Автоматическое email-письмо с опросом• Предложение: скидка 10% при продлении• Напоминание о преимуществах тарифа• Приглашение на вебинар о новых функциях | Низкий (50-100 руб/клиент) | Снизить отток на 20-30% |
| Низкий риск | НИЗКИЙ (поддерж ание лояльности) | <40% | <ul style="list-style-type: none">• Регулярные информационные рассылки• Программа лояльности: бонусы за длительность• Спасибо-письмо за длительное сотрудничество• Приглашение в реферальную программу | Минимальны й (10-30 руб/клиент) | Поддержани е лояльности, перекрестн ые продажи |

Выводы

В ходе работы была разработана модель бинарной классификации для предсказания ухода клиентов, учитывая дисбаланс классов и важность калибровки вероятностей для бизнес-решений.

Реализация включала **подготовку данных** (кодирование категорий, нормализация), анализ дисбаланса (учитывали при обучении), создание базовых моделей (Majority class, логистическая регрессия), а также более сложных алгоритмов (деревья, ансамбли — RandomForest и Gradient Boosting), с использованием методов интерпретации (SHAP, feature importance).

Анализ trade-off-рисков показал, что оптимальный порог по F1 достигается на уровне 0.4, при котором из 527 клиентов с высокой вероятностью ухода можно целенаправленно обрабатывать большинство случаев риска.

Выводы

Ключевые факторы риска — скоринговый RiskScore, тип интернета (Fiber optic), краткосрочные контракты и низкая стажировка. Эти признаки позволяют выделять сегменты клиентов и целенаправленно работать с ними.

Рекомендации по удержанию — строятся по сегментам риска, от личных звонков и скидок для критических клиентов до кампаний по лояльности для тех, кто в группе низкого риска. Такой подход помогает повысить retention и снизить бизнес-затраты.

Общие выводы — достигнута высокая эффективность модели в сравнении с базовыми подходами, реализованы инструменты интерпретации и сегментации для улучшения бизнес-процессов. В результате — более точные предсказания оттока и возможность целенаправленных мер по удержанию.

Планы развития

Улучшение модели и её расширение

Расширение источников данных

дополнительные данные: данные о звонках и взаимодействиях и т.п.

Автоматизация и интеграция в бизнес-процессы системы автоматизированной работы с клиентами для оперативных мер по удержанию.

Внедрение системы рекомендации мер по удержанию блоков рекомендаций, автоматизированных сценариев и чек-листов для менеджеров по удержанию клиентов для каждого сегмента.

