МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
[Полное название вуза]

Факультет информационных технологий
Кафедра прикладной информатики

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

на тему: «Применение методов машинного обучения для прогнозирования оттока клиентов телекоммуникационной компании»

Направление подготовки 09.04.03 «Прикладная информатика»

Магистерская программа: [название программы]

Выполнил: магистрант гр. ____

_____ / И. О. Фамилия

Научный руководитель:

канд. техн. наук, доцент

_____ / И. О. Фамилия

[Город] 2026

Пример работы, выполненной в Универ AI · univerai.ru

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....3
1. Теоретические основы прогнозирования оттока клиентов.....9
1.1 Отток клиентов как экономическая категория и объект управления.....9
1.2 Постановка задачи прогнозирования оттока как задачи бинарной классификации.....13
1.3 Обзор исследований и методов машинного обучения для прогнозирования оттока.....18
2. Данные и методика исследования.....23
2.1 Характеристика набора данных IBM Telco Customer Churn.....23
2.2 Разведочный анализ данных и факторы оттока.....25
2.3 Подготовка данных, схема эксперимента и критерии оценки.....29
3. Построение и оценка моделей прогнозирования оттока.....34
3.1 Обучение моделей и подбор гиперпараметров.....34
3.2 Сравнение моделей и интерпретация факторов оттока.....37
3.3 Проект встраивания модели в информационную систему оператора..46
3.4 Экономическая оценка применения модели в кампании удержания...51
ЗАКЛЮЧЕНИЕ.....56
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ.....59
Приложение А. Программный код подготовки данных, обучения и оценки моделей.....68

ВВЕДЕНИЕ

Абонент, который расторг договор с оператором связи, уносит с собой не одну ежемесячную плату. Чтобы заменить его новым клиентом, компания снова тратится на рекламу, дилерские комиссии, подключение и скидки первого месяца, а окупаются эти расходы далеко не сразу. Ф. Райхельд и У. Сассер ещё в 1990 г. на примерах компаний сферы услуг показали, что прибыль растёт по мере сокращения доли уходящих клиентов, и предложили считать ориентиром качества сервиса «нулевой отток» [57]. Позднее С. Гупта, Д. Леманн и Дж. Стюарт оценили клиентские базы пяти публичных компаний и получили, что улучшение удержания на 1 % увеличивает стоимость фирмы примерно на 5 %, а такое же улучшение маржи — лишь на 1 % [33]. Для телекоммуникационного бизнеса, где выручка почти целиком складывается из абонентской платы, эти соотношения особенно наглядны.

Российскому абоненту сменить оператора несложно. Закон «О связи» позволяет расторгнуть договор об оказании услуг связи в одностороннем порядке, а с 1 декабря 2013 г. действует право сохранить мобильный номер при переходе к другому оператору [1, 3]. Главный барьер переключения исчез, и конкуренция за абонента сместилась от цены подключения к качеству сервиса и к умению вовремя заметить недовольного клиента. Заметить его можно по данным, которые у оператора уже есть: тип и срок договора, набор услуг, способ оплаты, история начислений.

Прогнозирование оттока по таким данным стало классической прикладной задачей машинного обучения. Её решали на данных американских, европейских и азиатских операторов логистической регрессией, деревьями решений, нейронными сетями, методом опорных векторов и ансамблями [34, 51, 65]. Однако в прикладных работах до сих пор встречаются три слабых места. Модели сравнивают по доле верных ответов, хотя при дисбалансе классов она мало что говорит о качестве. Разницу в третьем знаке ROC-AUC объявляют преимуществом алгоритма, не проверяя её устойчивость. Наконец, качество модели почти никогда не переводят в

деньги, хотя именно деньги решают, будет ли кампания удержания запущена [13, 66]. Поэтому тема работы сохраняет практический смысл и для исследователя, и для ИТ-подразделения оператора.

Степень научной разработанности проблемы

Экономическое значение удержания и пожизненной ценности клиента исследовали Ф. Райхельд, С. Гупта, Д. Леманн, В. Кумар, В. Райнарц, П. Фейдер и Б. Харди [27, 33, 47, 57]. Модели оттока для операторов связи строили и сравнивали С. Неслин с соавторами, А. Леммен и К. Кру, В. Вербеке и Б. Баезенс, К. Куссман и Д. Ван ден Поль, Й. Бюрез, А. Де Кеньи, М. Оскарсдоттир, А. Ахмад [11, 18, 21, 24, 48, 51, 53, 64, 65]. Алгоритмическую основу составили работы Л. Бреймана о деревьях и случайном лесе, Дж. Фридмана о градиентном бустинге и их современные реализации — XGBoost, LightGBM и разработанная в «Яндексе» библиотека CatBoost [16, 17, 20, 30, 43, 55]. Русскоязычному читателю методы машинного обучения систематически изложены в курсе К. В. Воронцова и в переводах книг П. Флаха, С. Рашки, А. Мюллера и С. Гвидо [6, 8, 9, 10]. Вопросы оценки моделей при дисбалансе классов и связи метрик с прибылью рассмотрены в работах Т. Фосетта, Д. Хэнда, Т. Сайто, Т. Вербракена [28, 35, 59, 66]. Меньше изучено, как совместить статистическую проверку различий между моделями с экономическим расчётом кампании удержания и с требованиями к встраиванию модели в информационную систему оператора. Этот пробел и определил направление исследования.

Цель и задачи исследования

Цель работы — разработать и обосновать методику прогнозирования оттока клиентов телекоммуникационной компании на основе методов машинного обучения, включающую выбор модели по статистическим и экономическим критериям и схему её встраивания в информационную систему оператора.

Для достижения цели поставлены следующие задачи:

1) раскрыть экономическое содержание оттока клиентов и место его прогнозирования в управлении взаимоотношениями с абонентами;

2) формализовать прогнозирование оттока как задачу бинарной классификации и обосновать набор метрик качества с учётом дисбаланса классов;

3) проанализировать исследования по прогнозированию оттока в телекоммуникациях и систематизировать применяемые методы;

4) провести разведочный анализ открытого набора данных IBM Telco Customer Churn и выявить факторы, связанные с уходом клиентов;

5) построить и сравнить логистическую регрессию, дерево решений, случайный лес и градиентный бустинг, проверив устойчивость различий бутстрэпом;

6) разработать схему внедрения модели в информационную систему оператора и оценить экономический эффект кампании удержания при разных допущениях.

Объект исследования — процесс удержания клиентов телекоммуникационной компании. **Предмет исследования** — методы и модели машинного обучения для прогнозирования оттока абонентов и их применение в информационной системе оператора.

Гипотеза исследования: на табличных данных о договоре и услугах абонента ансамбли деревьев решений не дают статистически значимого выигрыша в ROC-AUC по сравнению с регуляризованной логистической регрессией, а экономический результат кампании удержания сильнее зависит от выбранного охвата и от допущений о реакции клиентов, чем от выбора алгоритма.

Научная новизна

1) Предложена методика сравнения моделей оттока, объединяющая стратифицированную кросс-проверку, бутстрэп-интервалы для разности ROC-AUC и кривую прибыли кампании удержания. В отличие от

распространённого сравнения по одной метрике она отделяет устойчивое превосходство модели от случайного колебания выборки.

2) Получены количественные оценки чувствительности оптимального охвата кампании удержания к доле удержанных абонентов и к сроку их последующего обслуживания; показано, что в базовом сценарии выгодно обращаться примерно к трети клиентов с наибольшим риском, а в зависимости от допущений оптимальный охват меняется от 12 до 53 % базы.

3) Уточнена архитектура встраивания модели оттока в информационную систему оператора: пакетный скоринг по витрине признаков, передача списков в CRM, контрольная группа для измерения эффекта и мониторинг дрейфа данных по индексу стабильности популяции.

Положения, выносимые на защиту

1) При доле уходящих клиентов около 26 % доля верных ответов неинформативна: константный прогноз «клиент останется» даёт 73,5 % верных ответов. Модель оттока следует оценивать совокупностью ROC-AUC, средней точности, точности и полноты при выбранном пороге и ожидаемой прибыли кампании.

2) На данных IBM Telco градиентный бустинг, случайный лес и логистическая регрессия достигают ROC-AUC 0,862–0,865 на отложенной выборке, а 95-процентный бутстрэп-интервал разности ROC-AUC бустинга и логистической регрессии включает ноль. Для внедрения предпочтительна логистическая регрессия как более прозрачная модель при сопоставимом качестве.

3) Наибольший вклад в прогноз вносят срок обслуживания, тип договора и вид интернет-подключения: случайная перестановка каждого из этих признаков снижает ROC-AUC модели на 0,05–0,08, тогда как пол, наличие партнёра и подключение защиты устройства практически не влияют на прогноз.

4) Экономический эффект кампании удержания определяется прежде всего охватом: при условных допущениях работы прибыль достигает

максимума при обращении к верхней части списка риска и снижается при попытке охватить всю базу, поэтому порог отсечения нужно выбирать по кривой прибыли, а не по F1-мере.

Теоретическая и практическая значимость

Теоретическая значимость работы состоит в систематизации методов прогнозирования оттока по типу модели, способу учёта дисбаланса классов и критерию оценки, а также в обосновании связки статистических и экономических критериев выбора модели. Практическая значимость определяется тем, что программный код подготовки данных, обучения и оценки моделей можно перенести в рабочий контур оператора, а методику расчёта эффекта — использовать при планировании бюджета удержания. Материалы работы пригодны для учебных дисциплин «Интеллектуальный анализ данных» и «Проектирование информационных систем».

Методы и информационная база исследования

В работе использованы методы описательной статистики и разведочного анализа данных, критерий хи-квадрат, логистическая регрессия, деревья решений, случайный лес и градиентный бустинг, стратифицированная кросс-проверка, бутстрэп, перестановочная оценка важности признаков и экономико-математическое моделирование. Информационную базу составили открытый набор данных IBM Telco Customer Churn о 7 043 клиентах условной телекоммуникационной компании [67, 68], федеральные законы и стандарты в сфере связи, персональных данных и искусственного интеллекта, научные статьи и монографии. Расчёты выполнены на Python 3.13 с библиотеками NumPy, pandas и Matplotlib [37, 41, 50]; алгоритмы обучения реализованы автором на NumPy, что позволило контролировать каждый шаг и сверить результат с описанием методов в литературе.

Апробация результатов. Основные положения работы докладывались на [название научной конференции], [год]; по теме исследования опубликовано [количество] работ: [выходные данные публикаций].

Структура работы. Диссертация состоит из введения, трёх глав, заключения, списка использованных источников и приложения. В первой главе рассмотрены экономика оттока, постановка задачи классификации и обзор исследований. Во второй главе описаны данные, результаты разведочного анализа и схема эксперимента. Третья глава содержит обучение и сравнение моделей, интерпретацию факторов, проект встраивания модели в информационную систему и расчёт экономического эффекта. В приложении приведён программный код.

1. Теоретические основы прогнозирования оттока клиентов

1.1 Отток клиентов как экономическая категория и объект управления

Под оттоком (англ. churn, attrition) понимают прекращение отношений клиента с компанией в течение заданного периода. Для оператора связи момент ухода обычно виден точно: абонент расторгает договор, переносит номер к конкуренту или перестаёт платить и отключается. Это отличает телекоммуникации от розничной торговли, где покупатель может просто перестать приходить, и о его уходе компания узнаёт лишь по косвенным признакам. П. Фейдер и Б. Харди называют такую ситуацию контрактной и строят для неё отдельные модели прогноза удержания [27].

Простейший измеритель — коэффициент оттока за период:

$$CR = \frac{N_{ушл}}{N_{нач}} \cdot 100\% \quad (1)$$

где N_ушл — число клиентов, расторгнувших договор за период; N_нач — число клиентов на начало периода.

Месячный и годовой коэффициенты связаны нелинейно. Если каждый месяц уходит 2 % базы, то за год сохраняется $0{,}98^{12} \approx 0{,}785$ клиентов, то есть годовой отток составляет около 21,5 %, а не 24 %. Эту разницу полезно помнить при сравнении отчётов разных компаний: одни публикуют среднемесячный показатель, другие — годовой.

Отток неоднороден. Дж. Хадден с соавторами различают добровольный и вынужденный уход; вынужденным считают отключение по инициативе компании, например за долги, а добровольный делят на намеренный, когда клиент уходит к конкуренту, и случайный, вызванный переездом или сменой жизненных обстоятельств [34]. Для управления удержанием интересен прежде всего намеренный добровольный уход: его причины лежат в цене, качестве и сервисе, а значит, на них можно повлиять. Отдельно говорят о скрытом оттоке — абонент формально остаётся, но

переносит основной трафик к другому оператору. Виды оттока сведены в таблицу 1.

Таблица 1 – Виды оттока клиентов оператора связи

Вид оттока	**Пример**	**Чем управляет оператор**	**Пригодность для прогноза**
Добровольный намеренный	Переход к конкуренту с переносом номера	Цена, качество связи, сервис, удерживающие предложения	Высокая: есть поведенческие признаки
Добровольный случайный	Переезд в регион без покрытия, смерть абонента	Почти ничем	Низкая
Вынужденный	Отключение за неуплату	Кредитная политика, напоминания о платеже	Средняя: решается моделью платёжного риска
Скрытый (частичный)	Вторая SIM-карта другого оператора, снижение трафика	Тарифы, пакеты услуг	Средняя: нужны данные о потреблении

Источник: составлено автором по [34, 47].

Экономическую цену оттока удобно выражать через пожизненную ценность клиента (customer lifetime value, CLV) — приведённую сумму маржи, которую клиент принесёт за всё время обслуживания [33, 47]:

$$CLV = \sum_{t=0}^{T} \frac{m_t r^t}{(1+d)^t} - AC \qquad (2)$$

где m_t — маржа от клиента в периоде t; r — вероятность сохранения клиента в течение периода; d — ставка дисконтирования за период; T — горизонт расчёта; AC — затраты на привлечение клиента.

При постоянной марже и бесконечном горизонте ряд сворачивается в выражение m(1 + d)/(1 + d − r). Для иллюстрации возьмём среднюю ежемесячную плату клиента из анализируемого далее набора данных — 64,76 у. е. — и условно примем маржу 40 %, то есть 26 у. е. в месяц, и ставку дисконтирования 1 % в месяц. При месячном удержании 97 % пожизненная ценность составит 26 · 1,01 / 0,04 ≈ 657 у. е., а при удержании 98 % — 26 · 1,01 / 0,03 ≈ 875 у. е. Рост удержания всего на один процентный пункт

увеличивает ценность клиента на треть. Расчёт условный, но он объясняет, почему операторы готовы тратиться на удержание и почему точность выбора адресатов так важна.

Удержание бывает реактивным и проактивным. В реактивной схеме оператор реагирует на уже заявленное намерение уйти: абонент звонит в контактный центр, и специалист отдела удержания предлагает ему скидку или другой тариф. Схема простая, но срабатывает поздно, когда клиент уже сравнил предложения и принял решение. Проактивная схема опирается на прогноз: компания заранее выделяет абонентов с высоким риском ухода и обращается к ним до того, как они начнут искать альтернативу. Именно для неё нужна модель оттока. Цикл проактивного удержания показан на рисунке 1.



Рисунок 1 – Цикл проактивного удержания клиентов

Проактивная кампания имеет свою цену. Каждый контакт стоит денег, предложение со скидкой получают и те клиенты, которые никуда не собирались уходить, а часть уходящих не откликнется ни на какое предложение. Поэтому модель нужна не для того, чтобы угадать всех уходящих, а для того, чтобы ранжировать базу и направить ограниченный бюджет туда, где отдача максимальна. Э. Аскарса на данных полевых экспериментов показала, что абоненты с наибольшим риском ухода не

обязательно сильнее всего реагируют на удерживающее воздействие [13]. Этот вывод не отменяет моделей оттока, но предостерегает от механического правила «чем выше риск, тем нужнее контакт».

У задачи есть правовая сторона. Данные о договоре, услугах и платежах абонента являются персональными данными, и их обработка подчиняется Федеральному закону № 152-ФЗ «О персональных данных» [2]. Статья 16 этого закона запрещает принимать решения, порождающие юридические последствия для субъекта или иным образом затрагивающие его права, на основании исключительно автоматизированной обработки персональных данных, если нет письменного согласия субъекта или прямого указания федерального закона. Персональное предложение скидки обычно таким решением не считается, однако отказ в услуге или изменение условий договора по сигналу модели уже требует участия человека. Закон «О связи», в свою очередь, относит сведения об абонентах и оказываемых им услугах к информации ограниченного доступа [1]. Отсюда практическое требование к проекту: обучающие выборки обезличиваются, а доступ к оценкам риска получает узкий круг сотрудников.

На государственном уровне развитие технологий искусственного интеллекта поддерживается Национальной стратегией, утверждённой Указом Президента Российской Федерации от 10 октября 2019 г. № 490 [4]. Способы обеспечения доверия к системам искусственного интеллекта описывает ГОСТ Р 59276-2020: среди характеристик, недостаток которых снижает качество системы, он называет понятность, объяснимость и предсказуемость, а доверие рассматривает на всех этапах жизненного цикла, включая эксплуатацию [5]. Для модели оттока это означает, что оператор должен уметь объяснить, на каких признаках строится прогноз, и регулярно проверять его качество на новых данных.

1.2 Постановка задачи прогнозирования оттока как задачи бинарной классификации

Формально задача выглядит так. Имеется выборка из n клиентов; каждый описан вектором признаков x, известных на дату среза, и меткой y, равной единице, если клиент ушёл в течение следующего периода наблюдения, и нулю в противном случае. Требуется построить функцию, которая по новому вектору признаков оценивает вероятность ухода $\hat{p} = P(y = 1 \mid x)$. Решение о контакте принимается сравнением оценки с порогом t: клиент попадает в кампанию, если $\hat{p} \geq t$.

За этой короткой постановкой стоят три методических решения. Первое — горизонт прогноза. Модель, предсказывающая уход в ближайшую неделю, бесполезна для кампании, которую готовят месяц; модель с горизонтом в год слишком грубая. На практике выбирают один–три месяца. Второе — разрыв между срезом признаков и окном целевой переменной: признаки должны быть известны до начала окна, иначе в модель просачивается информация из будущего. Классический пример утечки — признак «количество обращений в отдел расторжения» за тот же месяц, в котором клиент ушёл. Третье — единица наблюдения: договор, лицевой счёт или абонент, у которого несколько договоров.

Логистическая регрессия

Базовая модель для бинарной классификации — логистическая регрессия, предложенная Д. Коксом для анализа двоичных последовательностей [22, 39]. Вероятность ухода выражается через линейную комбинацию признаков:

$$\hat{p}(x) = \sigma(\beta_0 + \beta^T x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}} \quad (3)$$

где β_0 — свободный член; β — вектор коэффициентов; σ — логистическая функция.

Коэффициенты находят минимизацией логистической функции потерь с L2-регуляризацией, которая не даёт весам неограниченно расти при коррелированных признаках:

$$L(\beta) = -\sum_{i=1}^{n} w_i \left[y_i \ln \hat{p}_i + (1 - y_i) \ln(1 - \hat{p}_i) \right] + \frac{\lambda}{2} \|\beta\|^2 \quad (4)$$

где w_i — вес наблюдения (позволяет усилить редкий класс); λ — коэффициент регуляризации.

Главное достоинство модели — интерпретируемость. Экспонента коэффициента при стандартизованном признаке показывает, во сколько раз меняются шансы ухода при росте признака на одно стандартное отклонение. Ограничение тоже очевидно: модель линейна в логарифме шансов и сама не улавливает взаимодействий, например того, что помесячный договор опасен прежде всего в сочетании с оптоволоконным подключением.

Деревья решений и ансамбли

Дерево решений последовательно делит выборку по значениям признаков так, чтобы в получившихся группах клиенты были как можно однороднее по метке [17]. Однородность узла в алгоритме CART измеряют индексом Джини:

$$G = 1 - \sum_{k=0}^{1} p_k^2 = 2p(1 - p) \quad (5)$$

где p — доля ушедших клиентов в узле.

Из всех допустимых разбиений выбирают то, которое сильнее всего снижает взвешенную сумму индексов в дочерних узлах. Дерево наглядно и само находит взаимодействия, но неустойчиво: небольшое изменение выборки может перестроить верхние разбиения. Л. Брейман предложил усреднять много деревьев, обученных на бутстрэп-выборках со случайным подмножеством признаков в каждом узле, — так появился случайный лес

[16]. Усреднение снижает дисперсию прогноза, почти не увеличивая смещения.

Градиентный бустинг строит деревья не независимо, а последовательно: каждое новое дерево аппроксимирует антиградиент функции потерь по текущему прогнозу [30]:

$$F_m(x) = F_{m-1}(x) + \nu\, h_m(x), \quad h_m \approx -\frac{\partial L(y, F)}{\partial F}\Big|_{F = F_{m-1}} \qquad (6)$$

где F_m — прогноз в логарифме шансов после m итераций; h_m — дерево, обученное на m-й итерации; ν — темп обучения ($0 < \nu \leq 1$).

Реализация XGBoost дополнила схему регуляризацией листьев и вторым порядком разложения потерь [20], LightGBM — гистограммным поиском разбиений и выборочным использованием наблюдений с малым градиентом [43], CatBoost — упорядоченным бустингом и кодированием категориальных признаков по целевой статистике без утечки [55]. На табличных данных умеренного размера ансамбли деревьев обычно не уступают глубоким нейронным сетям, а нередко и превосходят их; это показали сравнительные исследования Л. Гриншцайна с соавторами и Р. Шварц-Зива и А. Армона [32, 61].

Метрики качества

Результат классификации при фиксированном пороге описывают матрицей ошибок (таблица 2).

Таблица 2 – Матрица ошибок классификатора оттока

	Модель: уйдёт	**Модель: останется**
Клиент ушёл	TP — верно найденный уходящий	FN — пропущенный уходящий
Клиент остался	FP — ложная тревога	TN — верно найденный лояльный

Из неё выводят точность (precision), полноту (recall) и их гармоническое среднее — F1-меру:

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN} \tag{7}$$

где TP, FP, FN — элементы матрицы ошибок (таблица 2).

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{8}$$

Для задачи оттока точность отвечает на вопрос, какая доля бюджета кампании уйдёт на действительно уходящих клиентов, а полнота — какую долю уходящих кампания вообще заметит. Доля верных ответов (accuracy) здесь вводит в заблуждение. В наборе данных, который анализируется во второй главе, ушли 1 869 клиентов из 7 043, или 26,5 %. Модель, которая всем отвечает «останется», получит 73,5 % верных ответов и не найдёт ни одного уходящего.

Метрики при фиксированном пороге зависят от выбора порога. Чтобы оценить модель как ранжирующее правило, используют ROC-кривую — зависимость доли верно найденных уходящих от доли ложных тревог при всех возможных порогах [28]. Площадь под ней, ROC-AUC, равна вероятности того, что случайно выбранный ушедший клиент получит от модели более высокую оценку риска, чем случайно выбранный оставшийся [36]. Её вычисляют через ранги, как статистику Манна — Уитни:

$$AUC = \frac{\sum_{i:\, y_i = 1} R_i - n_1(n_1 + 1)/2}{n_1 n_0} \tag{9}$$

где R_i — ранг оценки i-го ушедшего клиента в общей упорядоченной выборке; n_1 и n_0 — число ушедших и оставшихся клиентов.

Значение 0,5 соответствует случайному ранжированию, 1,0 — идеальному. При сильном дисбалансе классов ROC-кривая выглядит оптимистично, поэтому её дополняют кривой «точность — полнота» и площадью под ней (средней точностью, AP) [23, 59]. Д. Хэнд указал и на более глубокий недостаток ROC-AUC: метрика неявно усредняет качество по разным соотношениям цен ошибок, причём это распределение своё для

каждого классификатора [35]. Отсюда интерес к метрикам, напрямую связанным с прибылью.

Для оценки качества вероятностей, а не только ранжирования, используют оценку Брайера — средний квадрат разности между прогнозом и фактической меткой. Она важна, когда вероятность ухода подставляют в экономический расчёт: модель может хорошо упорядочивать клиентов, но систематически завышать или занижать риск. А. Никулеску-Мизил и Р. Каруана показали, что бустинг деревьев склонен сдвигать вероятности к краям, и предложили калибровать их после обучения [52].

Дисбаланс классов и экономические критерии

Уходящих клиентов заметно меньше, чем остающихся, и стандартные алгоритмы, минимизирующие среднюю ошибку, смещаются в сторону большинства. Применяют три группы приёмов: изменение весов наблюдений в функции потерь, случайное прореживание большинства и синтез новых наблюдений меньшинства, например методом SMOTE [19]. Й. Бюрез и Д. Ван ден Поль сравнили эти приёмы на нескольких наборах данных об оттоке и пришли к выводу, что простое прореживание даёт выигрыш, а более сложные схемы выборки не всегда его увеличивают [18]. Если же модель оценивают по ROC-AUC, перевзвешивание классов на ранжирование влияет слабо: оно сдвигает вероятности, а порядок клиентов почти не меняет.

Экономический критерий связывает классификацию с кампанией удержания. С. Неслин с соавторами записали прибыль кампании так [51]:

$$\Pi = N\alpha[\beta\gamma(CLV - c - \delta) + \beta(1-\gamma)(-c) + (1-\beta)(-c-\delta)] - A \quad (10)$$

где N — размер клиентской базы; α — доля базы, охваченная кампанией; β — доля уходящих среди охваченных (точность модели при выбранном пороге); γ — доля уходящих, принявших предложение и оставшихся; c — затраты на контакт; δ — стоимость удерживающего предложения; A — постоянные затраты кампании.

Формула показывает, что модель влияет на прибыль через β при каждом значении α, то есть через всю кривую «охват — точность». В. Вербеке с соавторами предложили выбирать модель по максимальной прибыли, которую она позволяет получить при оптимальном охвате [65], а Т. Вербракен, В. Вербеке и Б. Баезенс обобщили эту идею до ожидаемой максимальной прибыли с учётом неопределённости параметра γ [66]. В третьей главе формула (10) применяется к данным с явно оговорёнными условными параметрами.

Близкую логику предлагают Ф. Провост и Т. Фосетт в рамке ожидаемой ценности: каждой клетке матрицы ошибок приписывается выигрыш или потеря в деньгах, и модель оценивается не по одной точке, а по кривой прибыли — зависимости ожидаемого результата от доли охваченных клиентов [56]. Такая кривая сразу показывает и лучший охват, и цену ошибки в выборе порога. Если бюджет кампании фиксирован, вопрос о пороге отпадает: оператор обращается к стольким клиентам из верхней части списка, на скольких хватает денег, и модель отвечает лишь за качество этого списка. Если бюджет гибкий, порог выбирают по максимуму кривой прибыли. В обоих случаях модель нужна прежде всего как ранжирующее правило, и поэтому метрики ранжирования, а не доля верных ответов, служат основными критериями её качества.

1.3 Обзор исследований и методов машинного обучения для прогнозирования оттока

Литературу о прогнозировании оттока в телекоммуникациях удобно читать по трём линиям: какие алгоритмы сравнивают, какие данные используют и как оценивают результат. Обзор Дж. Хаддена с соавторами 2007 г. зафиксировал состояние области на тот момент: регрессионные модели, деревья решений и нейронные сети, данные биллинга и обращений в службу поддержки, оценка по доле верных ответов и лифту [34].

Важной вехой стал турнир моделей оттока, проведённый Центром управления взаимоотношениями с клиентами Университета Дьюка на данных крупного американского оператора мобильной связи. С. Неслин с соавторами проанализировали присланные участниками решения и показали, что разница между методами существенна и измерима в деньгах: лучшие модели приносили кампании удержания заметно больше прибыли, чем средние [51]. А. Леммен и К. Кру на данных американского оператора беспроводной связи применили бэггинг и бустинг деревьев классификации и существенно улучшили точность прогноза; они же предложили обучать модель на сбалансированной выборке с последующей поправкой возникающего смещения [48].

Следующие работы расширили набор методов. К. Куссман и Д. Ван ден Поль применили метод опорных векторов к оттоку клиентов подписного сервиса и выяснили, что логистическую регрессию он опережает только при оптимальной процедуре подбора параметров, а случайному лесу уступает в обоих вариантах [21]. В. Вербеке с соавторами в 2011 г. построили понятные модели на основе индукции правил, сочетающие точность с интерпретируемостью, а в 2012 г. сравнили широкий набор алгоритмов на нескольких наборах данных операторов связи и предложили критерий максимальной прибыли [64, 65]. Их главный вывод для практики: многие алгоритмы дают близкое качество, и выбор между ними стоит делать с учётом понятности и экономического результата.

Отдельная линия посвящена признакам. Б. Хуан, М. Кечади и Б. Бакли сформировали расширенный набор признаков для абонентов фиксированной связи и сравнили на нём несколько методов классификации [40]. А. Идрис, М. Ризван и А. Хан сочетали случайный лес с балансировкой данных на основе роя частиц и разными стратегиями отбора признаков [42]. М. Оскарсдоттир с соавторами построили признаки по графу звонков и показали, что сведения о социальном окружении абонента улучшают прогноз [53]. А. Ахмад, А. Джафар и К. Алджумаа обработали данные сирийского

оператора SyriaTel на платформе больших данных Spark; данные охватывали девять месяцев, а лучший результат, ROC-AUC 93,3 %, дал алгоритм XGBoost [11].

Методические работы последних лет смещают акцент с точности прогноза на пользу для решения. А. Де Кеньи, К. Куссман и К. Де Бок предложили гибридную модель, которая сегментирует клиентов деревом и строит в каждом листе логистическую регрессию, сохраняя понятность при качестве ансамблей [24]. А. Амин с соавторами предложили учитывать степень определённости данных (data certainty) и оценивать прогноз раздельно для групп клиентов, в отношении которых классификатор более и менее уверен [12]. Ф. Девриндт, Й. Берревутс и В. Вербеке сравнили модели оттока с моделями прироста отклика (uplift) и пришли к выводу, что для отбора адресатов кампании важнее предсказывать эффект воздействия, чем сам уход [25]. Тот же мотив звучит в работе Э. Аскарсы [13].

Сравнительные исследования на открытых наборах подтверждают, что разрыв между методами на табличных данных невелик. Т. Вафеядис с соавторами сравнили многослойный персептрон, деревья решений, метод опорных векторов, наивный байесовский классификатор и логистическую регрессию с бустингом и без него и нашли лучший вариант в методе опорных векторов с бустингом, однако различия между первыми строками рейтинга были небольшими [63]. А. Керамати с соавторами получили на данных иранского оператора высокие точность и полноту, комбинируя несколько классификаторов [44]. Основные работы сведены в таблицу 3.

Таблица 3 – Исследования по прогнозированию оттока клиентов операторов связи

Авторы, год	Данные	Методы	Основной вывод
Неслин и др., 2006 [51]	Крупный оператор мобильной связи США, турнир моделей	Решения участников: регрессия, деревья, нейросети и др.	Выбор метода заметно влияет на прибыль кампании
Леммен, Кру, 2006 [48]	Оператор беспроводной связи США	Бэггинг и бустинг деревьев	Ансамбли заметно повышают точность прогноза
Бюрез, Ван ден Поль, 2009 [18]	Несколько наборов данных об оттоке	Взвешенный случайный лес, прореживание, бустинг	Прореживание большинства улучшает результат
Вербеке и др., 2012 [65]	Несколько наборов данных операторов	Широкий набор алгоритмов классификации	Качество многих методов близко; нужен критерий прибыли
Оскарсдоттир и др., 2017 [53]	Графы звонков операторов	Реляционные классификаторы, признаки сети	Социальные признаки улучшают прогноз
Де Кеньи и др., 2018 [24]	Наборы данных об оттоке разных отраслей	Дерево + логистическая регрессия в листьях	Понятность без потери качества
Ахмад и др., 2019 [11]	Оператор SyriaTel	Деревья, случайный лес, GBM, XGBoost на Spark	XGBoost лучший, ROC-AUC 93,3 %
Девриндт и др., 2021 [25]	Данные оператора и других компаний	Модели оттока и uplift-модели	Для кампании важнее прогноз эффекта

Источник: составлено автором.

Из обзора можно сделать три вывода, которые определили схему эксперимента в этой работе. Во-первых, логистическая регрессия остаётся сильной базовой моделью, и любое усложнение нужно сравнивать с ней с проверкой значимости разницы. Во-вторых, ансамбли деревьев стабильно входят в число лучших методов, причём градиентный бустинг чаще других. В-третьих, итоговый критерий выбора — не метрика ранжирования, а результат кампании, который зависит от охвата и параметров предложения.

Отдельно стоит вопрос объяснения модели. К. Рудин настаивает, что для решений с серьёзными последствиями лучше сразу строить интерпретируемые модели, тем более что на структурированных данных они

часто не уступают «чёрным ящикам» [58]. Для сложных моделей применяют методы объяснения после обучения: значения Шепли в реализации SHAP [49] или перестановочную важность признаков, которую А. Фишер, К. Рудин и Ф. Доминичи формализовали как меру зависимости модели от признака [29]. Встроенная важность случайного леса по снижению неоднородности смещена в пользу непрерывных признаков и признаков с большим числом категорий; на это обратили внимание К. Штробль с соавторами [62]. Поэтому в работе важность признаков оценивается перестановкой на отложенной выборке.

Выводы по первой главе

Отток клиентов оператора связи — измеримое явление с прямой денежной ценой, которую удобно выражать через пожизненную ценность клиента. Проактивное удержание требует модели, ранжирующей абонентов по риску ухода, и экономического правила, определяющего, скольким из них стоит сделать предложение. Задача формализуется как бинарная классификация с дисбалансом классов; доля верных ответов для неё непригодна, а оценку нужно строить на ROC-AUC, средней точности, точности и полноте при выбранном пороге и прибыли кампании. Литература показывает, что на табличных данных различия между хорошими алгоритмами невелики, поэтому сравнение моделей должно сопровождаться проверкой устойчивости различий.

2. Данные и методика исследования

2.1 Характеристика набора данных IBM Telco Customer Churn

Эмпирическая часть работы выполнена на открытом наборе данных Telco Customer Churn, который компания IBM распространяет как учебный пример для своих аналитических продуктов. По описанию IBM, набор содержит сведения о 7 043 клиентах вымышленной телекоммуникационной компании из Калифорнии, оказывающей услуги домашней телефонии и доступа в интернет, по состоянию на третий квартал [67]. Расширенная версия для IBM Cognos Analytics 11.1.3 состоит из пяти таблиц: демография, местоположение, население по почтовым индексам, услуги и статус клиента с оценкой удовлетворённости и причиной ухода. В исследованиях и учебных курсах чаще используют компактную версию из одной таблицы с 21 полем; она опубликована IBM в открытом репозитории на GitHub [68]. В работе взята именно она: 7 043 строки, по одной на клиента, идентификаторы customerID уникальны.

Выбор открытого набора объясняется двумя причинами. Данные реальных операторов закрыты коммерческой тайной и тайной связи, а результаты на них нельзя воспроизвести. Набор IBM, напротив, доступен любому читателю, поэтому все приведённые далее цифры можно проверить, повторив расчёт. Обратная сторона — данные условные: компания вымышлена, история обслуживания представлена одним срезом, а денежные величины даны без указания валюты. Поскольку компания описана как калифорнийская, суммы естественно читать как доллары США; в работе они обозначаются как условные единицы (у. е.). Состав полей приведён в таблице 4.

Таблица 4 – Поля набора данных IBM Telco Customer Churn

Поле	Тип	Значения	Содержание
customerID	Строка	7 043 уникальных	Идентификатор клиента
gender	Категория	Female, Male	Пол

Поле	Тип	Значения	Содержание
SeniorCitizen	Бинарный	0, 1	Клиент пенсионного возраста
Partner, Dependents	Бинарные	Yes, No	Наличие партнёра, иждивенцев
tenure	Целое	0–72	Срок обслуживания, мес.
PhoneService, MultipleLines	Категории	Yes, No, No phone service	Телефония, несколько линий
InternetService	Категория	DSL, Fiber optic, No	Тип интернет-подключения
OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport	Категории	Yes, No, No internet service	Дополнительные интернет-услуги
StreamingTV, StreamingMovies	Категории	Yes, No, No internet service	Потоковое ТВ и кино
Contract	Категория	Month-to-month, One year, Two year	Тип договора
PaperlessBilling	Бинарный	Yes, No	Электронный счёт
PaymentMethod	Категория	4 способа	Способ оплаты
MonthlyCharges	Число	18,25–118,75	Ежемесячное начисление, у. е.
TotalCharges	Число	18,8–8 684,8; 11 пропусков	Начислено за всё время, у. е.
Churn	Бинарный	Yes, No	Клиент ушёл в отчётном квартале

Источник: составлено автором по данным [68].

Целевая переменная Churn показывает, ушёл ли клиент из компании в отчётном квартале [67]. Ушли 1 869 клиентов, остались 5 174; доля положительного класса — 26,5 %. Для одного квартала это очень высокая доля: при таком темпе компания за год лишилась бы более двух третей базы. Она объясняется учебным характером набора и облегчает обучение моделей, но переносить её на реальную компанию нельзя.

Качество данных проверено по пропускам и согласованности полей. Пропуски обнаружены только в поле TotalCharges: 11 записей содержат пробел вместо числа. У всех этих клиентов срок обслуживания равен нулю, то есть они только что подключились и ещё не получили ни одного счёта. Пропуск в данном случае содержательный, поэтому он заменён нулём, а не средним значением. Других пропусков и дубликатов нет.

Поле TotalCharges почти полностью выводится из двух других: начислено за всё время ≈ срок обслуживания × ежемесячное начисление. Коэффициент корреляции Пирсона между tenure и TotalCharges равен 0,826, между MonthlyCharges и TotalCharges — 0,651, а между tenure и MonthlyCharges — всего 0,248. Высокая связь не мешает деревьям, но в логистической регрессии делает коэффициенты при этих признаках неустойчивыми; в третьей главе это проявится в «неожиданном» знаке коэффициента при TotalCharges.

Следует назвать и ограничения набора, важные для переноса выводов. В нём нет данных о потреблении — минутах, трафике, роуминге, — которые в реальных моделях оттока часто оказываются сильными предикторами [40]. Нет обращений в контактный центр, жалоб, качества связи на базовых станциях, сведений о конкурентных предложениях. Наконец, срез один, поэтому нельзя построить признаки динамики, например падение трафика за последние три месяца. Модель, обученная на таком наборе, показывает нижнюю границу того, что можно получить на данных договора и тарифа.

2.2 Разведочный анализ данных и факторы оттока

Разведочный анализ начат с сопоставления доли ушедших клиентов в группах по каждому признаку. Самая сильная связь видна у типа договора: на помесячном договоре ушли 42,7 % клиентов, на годовом — 11,3 %, на двухлетнем — всего 2,8%. При этом помесячный договор заключили 3 875 клиентов из 7 043, больше половины базы.

Второй по силе фактор — срок обслуживания. Риск ухода максимален в первые месяцы и затем быстро снижается (рисунок 2). Из 624 клиентов со сроком обслуживания не более одного месяца ушли 60,9 %, а из 362 клиентов, обслуживающихся максимальные 72 месяца, — 1,7 %. Медианный срок обслуживания ушедших клиентов — 10 месяцев, оставшихся — 38 месяцев.

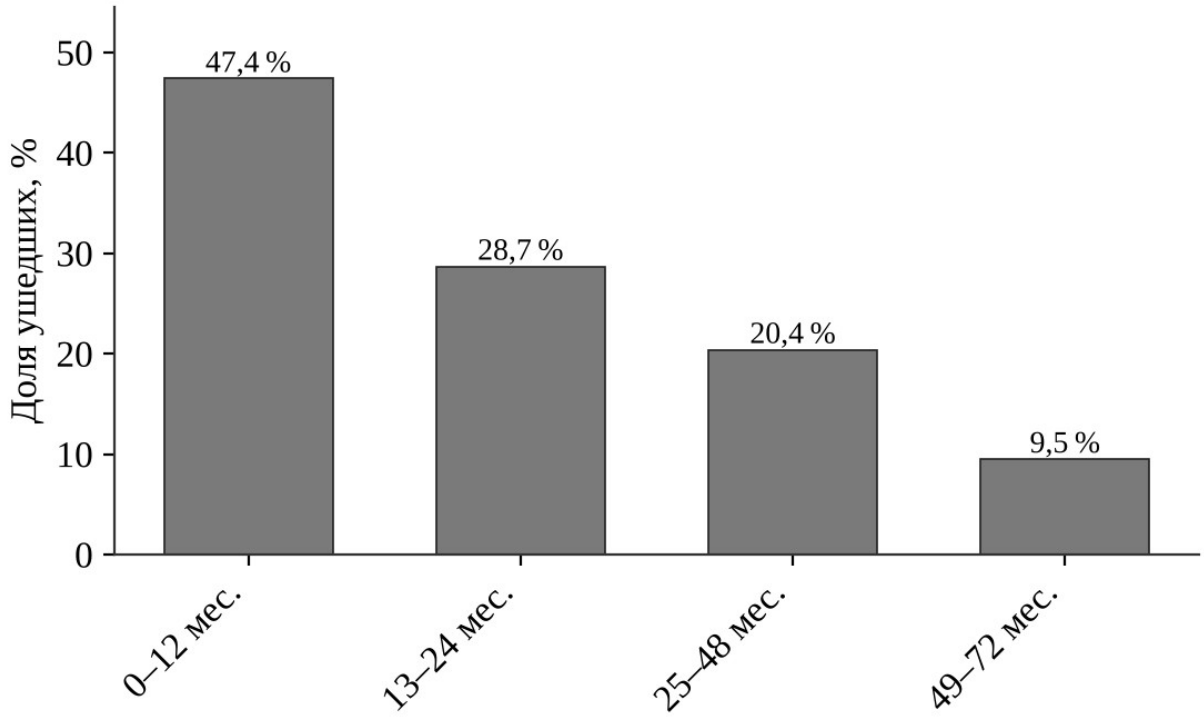


Рисунок 2 – Доля ушедших клиентов по сроку обслуживания

Источник: рассчитано автором по данным [68]; в группах 2 186, 1 024, 1 594 и 2 239 клиентов.

Картина типична для подписочного бизнеса: клиент, переживший первый год, редко уходит, а основной отток формируют новички, которые ещё сравнивают операторов. Отсюда практический вывод для кампании удержания: первые месяцы обслуживания заслуживают особого внимания, и предложение о переходе на долгосрочный договор логично делать именно в это время.

Третий фактор — тип интернет-подключения. Среди клиентов с оптоволоконным доступом ушли 41,9 %, с DSL — 19,0 %, без интернета — 7,4 %. Результат на первый взгляд парадоксален: более современная технология сопровождается вдвое большим оттоком. Объяснение надо искать в цене и ожиданиях. Ежемесячное начисление клиентов с оптоволокном заметно выше, а клиенты, платящие больше, острее реагируют на любые сбои и чаще сравнивают предложения конкурентов. Средняя ежемесячная плата ушедших клиентов — 74,44 у. е., оставшихся — 61,27 у. е.

Тип договора и тип подключения действуют совместно (таблица 5). Наибольший отток, 54,6 %, наблюдается у клиентов с помесячным договором и оптоволокном; таких в базе 2 128 человек. Двухлетний договор снижает отток даже среди пользователей оптоволокна до 7,2 %. Это взаимодействие факторов линейная модель без специальных признаков воспроизвести не может, а дерево решений находит его автоматически.

Таблица 5 – Доля ушедших клиентов по типу договора и интернет-подключения, %

Тип договора	DSL	Оптоволокно	Без интернета	Всего
Помесячный	32,2	54,6	18,9	42,7
На один год	9,3	19,3	2,5	11,3
На два года	1,9	7,2	0,8	2,8
Всего	19,0	41,9	7,4	26,5

Источник: рассчитано автором по данным [68].

Способ оплаты тоже связан с уходом. Клиенты, оплачивающие счета электронным чеком, уходят в 45,3 % случаев, тогда как при автоматическом списании с банковского счёта или карты — в 16,7 % и 15,2 % соответственно, при оплате почтовым чеком — в 19,1 %. Автоплатёж, видимо, говорит о более спокойном и долгом намерении пользоваться услугой, а ручная оплата каждый месяц напоминает клиенту о расходах и создаёт точку выбора.

Дополнительные интернет-услуги ведут себя по-разному. Подключённые защита в сети (OnlineSecurity) и техническая поддержка (TechSupport) сопровождаются оттоком 14,6 % и 15,2 % против 41,8 % и 41,6 % у клиентов с интернетом без этих услуг. Потоковое ТВ и кино такой разницы почти не дают: 30,1 % против 33,5 %. Защитные и сервисные услуги привязывают клиента, развлекательные — нет. Общее число подключённых услуг влияет немонотонно: у клиентов с одной услугой отток 10,9 %, с тремя — 44,9 %, с девятью — 5,3 %. Одна услуга — это чаще всего телефон без интернета, три — интернет с минимумом дополнений, а полный пакет означает глубокую вовлечённость.

Демографические признаки слабее. Клиенты пенсионного возраста уходят в 41,7 % случаев против 23,6 % у остальных, клиенты без партнёра — в 33,0 % против 19,7 %, без иждивенцев — в 31,3 % против 15,5 %. Пол на отток не влияет: 26,9 % у женщин и 26,2 % у мужчин.

Чтобы сравнить силу связи категориальных признаков с оттоком на одной шкале, для каждого признака рассчитаны статистика хи-квадрат по таблице сопряжённости и коэффициент Крамера:

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min(k, 2) - 1)}} \tag{11}$$

где χ^2 — статистика хи-квадрат по таблице сопряжённости признака и метки оттока; n — число наблюдений; k — число категорий признака. Для бинарной метки знаменатель равен n.

Результаты приведены в таблице 6. Критические значения χ^2 при уровне значимости 0,05 составляют 3,84 для одной степени свободы, 5,99 для двух и 7,81 для трёх.

Таблица 6 – Связь категориальных признаков с оттоком

Признак	**χ^2**	**Степени свободы**	**V Крамера**	**Связь при α = 0,05**
Contract	1 184,6	2	0,410	Значима
OnlineSecurity	850,0	2	0,347	Значима
TechSupport	828,2	2	0,343	Значима
InternetService	732,3	2	0,322	Значима
PaymentMethod	648,1	3	0,303	Значима
OnlineBackup	601,8	2	0,292	Значима
DeviceProtection	558,4	2	0,282	Значима
StreamingTV	374,2	2	0,231	Значима
StreamingMovies	375,7	2	0,231	Значима
PaperlessBilling	259,2	1	0,192	Значима
Dependents	189,9	1	0,164	Значима
SeniorCitizen	160,4	1	0,151	Значима
Partner	159,4	1	0,150	Значима
MultipleLines	11,3	2	0,040	Значима, очень слабая
PhoneService	1,0	1	0,012	Незначима
gender	0,5	1	0,009	Незначима

Источник: рассчитано автором по данным [68].

Высокие значения V у дополнительных интернет-услуг частично объясняются категорией «No internet service»: она совпадает для всех шести услуг и выделяет одну и ту же группу клиентов с низким оттоком. Поэтому сумма связей не равна их независимому вкладу, и окончательный ответ о значимости признаков даст многомерная модель. Разведочный анализ задаёт ожидания: в модели должны выделиться тип договора, срок обслуживания, тип подключения и способ оплаты, а пол и наличие телефонной линии должны оказаться практически бесполезными.

Отдельно оценена денежная сторона оттока в наборе. Суммарное ежемесячное начисление всех 7 043 клиентов составляет 456 116,60 у. е., а ушедших клиентов — 139 130,85 у. е., то есть 30,5 % месячной выручки. Доля выручки выше доли клиентов (26,5 %), потому что уходят в среднем клиенты с более дорогими тарифами. Для экономического расчёта в третьей главе это существенно: каждый правильно найденный уходящий клиент в среднем стоит дороже, чем средний клиент базы.

2.3 Подготовка данных, схема эксперимента и критерии оценки

Схема эксперимента показана на рисунке 3. Она построена так, чтобы отложенная выборка ни разу не использовалась для выбора модели, порога или гиперпараметров и давала несмещённую оценку качества.

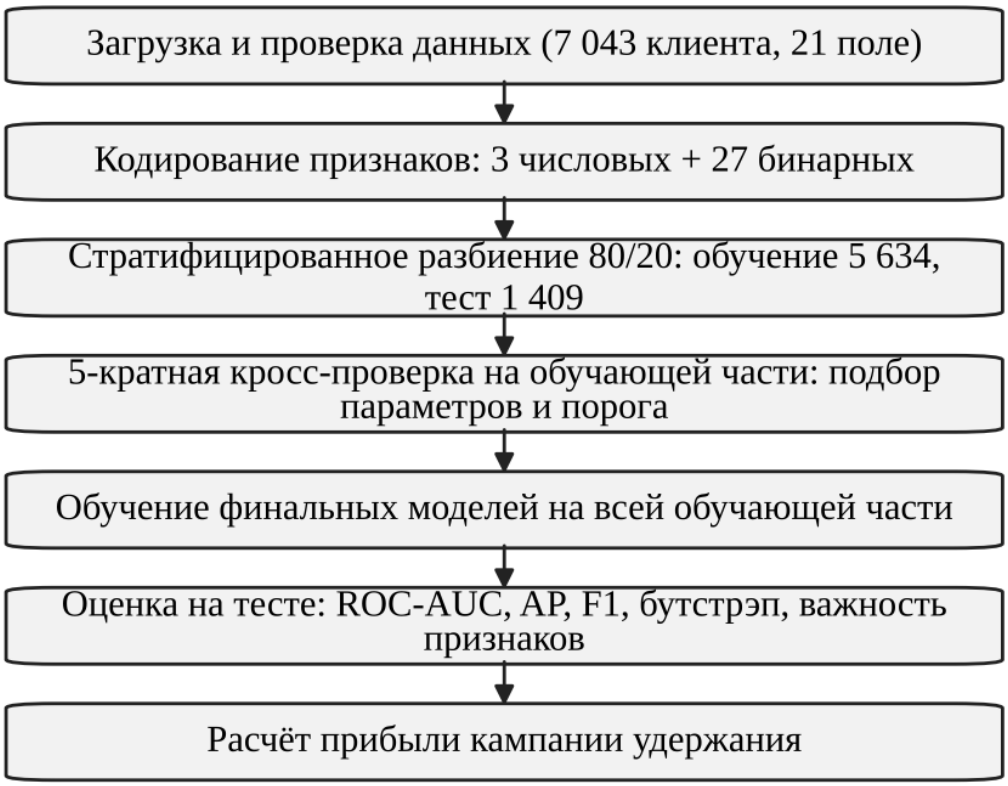


Рисунок 3 – Схема вычислительного эксперимента

Кодирование признаков. Три числовых признака — tenure, MonthlyCharges и TotalCharges — оставлены без изменений. Шестнадцать категориальных признаков преобразованы в бинарные индикаторы с исключением одной категории каждого признака, чтобы избежать полной линейной зависимости столбцов. Получилось 27 индикаторов и всего 30 признаков. Индикаторы «No internet service» для шести интернет-услуг совпадают между собой и с индикатором InternetService = No; эта избыточность оставлена сознательно, поскольку L2-регуляризация в логистической регрессии делит вес между дублирующими столбцами, а деревьям она не мешает.

Масштабирование. Для логистической регрессии каждый признак стандартизован по среднему и стандартному отклонению обучающей части:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{12}$$

где x̄_j и s_j — среднее и стандартное отклонение признака j, рассчитанные только по обучающим наблюдениям.

Параметры стандартизации вычисляются внутри каждого блока кросс-проверки заново, иначе сведения о проверочных наблюдениях просочились бы в обучение. Для деревьев числовые признаки разбиты на 32 интервала по квантилям обучающей выборки; поиск разбиения по гистограмме вместо полного перебора значений применяется в LightGBM и почти не снижает качества [43].

Разбиение выборки. Данные разделены на обучающую (80 %, 5 634 клиента, из них 1 495 ушедших) и тестовую (20 %, 1 409 клиентов, из них 374 ушедших) части со стратификацией по метке: доля ушедших в обеих частях составляет 26,5 %. Генератор случайных чисел зафиксирован (seed = 42), поэтому разбиение воспроизводимо.

Кросс-проверка. Внутри обучающей части выполняется стратифицированная 5-кратная кросс-проверка. Р. Кохави на большом числе наборов данных показал, что стратифицированная кросс-проверка с 5–10 блоками даёт лучшее сочетание смещения и дисперсии оценки, чем бутстрэп и простое отложенное подмножество [45]. По результатам кросс-проверки подбираются гиперпараметры градиентного бустинга и порог классификации; для выбора порога используются вне-блочные прогнозы, то есть прогнозы для каждого наблюдения, полученные моделью, которая его не видела.

Такая двухуровневая схема — отложенная тестовая часть плюс повторная выборка внутри обучающей — рекомендуется в руководствах по прикладному прогнозному моделированию [7, 46]. Она защищает от главной ошибки прикладных работ: многократного подбора параметров по тестовой выборке, после которого оценка на ней перестаёт быть независимой и становится оптимистичной.

Учёт дисбаланса. Основной вариант логистической регрессии обучен без весов, дополнительный — с весом 2,77 для ушедших клиентов, что выравнивает вклад классов (5 634 − 1 495 = 4 139 оставшихся против 1 495 ушедших). Сравнение двух вариантов покажет, влияет ли перевзвешивание на ранжирование. Синтетические наблюдения методом SMOTE не создавались: при доле меньшинства 26,5 % дисбаланс умеренный, а синтез наблюдений в пространстве с 27 бинарными индикаторами порождает невозможные сочетания признаков.

Состав моделей. Нейронные сети в сравнение не включены сознательно. Обучающая часть содержит 5 634 наблюдения и 30 признаков, из которых 27 бинарные; на таблицах такого размера ансамбли деревьев, как правило, не уступают глубоким сетям, а настройка сети требует существенно больше вычислений и подбора параметров [32, 61]. Метод опорных векторов тоже исключён: он не даёт вероятностей без дополнительной калибровки, а в исследованиях оттока уступал случайному лесу [21]. Оставшиеся четыре модели покрывают основные семейства — линейные модели, одиночные деревья, бэггинг и бустинг.

Критерии оценки. Основная метрика — ROC-AUC на тестовой выборке (формула 9). Дополнительно рассчитываются средняя точность AP, оценка Брайера, а также точность, полнота и F1-мера (формулы 7 и 8) при двух порогах: стандартном 0,5 и подобранном по максимуму F1 на внеблочных прогнозах. Для оценки неопределённости построены 95-процентные бутстрэп-интервалы ROC-AUC по 1 000 повторных выборок тестовой части с возвращением; для разности ROC-AUC двух моделей интервал строится по тем же повторным выборкам, что учитывает связанность оценок [26].

Важность признаков. Вклад признака оценивается перестановкой: значения признака в тестовой выборке случайно перемешиваются, и измеряется снижение ROC-AUC [29]:

$$I_j = AUC(f; X, y) - \frac{1}{K}\sum_{k=1}^{K} AUC(f; X^{\pi_k(j)}, y) \quad (13)$$

где f — обученная модель; X^π(j) — тестовая матрица признаков, в которой столбцы исходного признака j переставлены случайной перестановкой π; K = 10 — число повторов.

Все бинарные индикаторы одного исходного признака переставляются вместе, так что важность считается для 19 исходных признаков, а не для 30 закодированных столбцов. Такой подход не смещён в пользу признаков с большим числом значений, в отличие от встроенной важности по снижению неоднородности [62].

Подготовка данных выполнена следующим фрагментом программы (полный листинг — в приложении А):

```
d = pd.read_csv('Telco-Customer-Churn.csv')
d['TotalCharges'] = pd.to_numeric(d.TotalCharges,
                                  errors='coerce').fillna(0.0)
y = (d.Churn == 'Yes').astype(int).values
num = ['tenure', 'MonthlyCharges', 'TotalCharges']
cat = [c for c in d.columns
       if c not in num + ['customerID', 'Churn']]
X_df = pd.get_dummies(d[cat].astype(str),
                      drop_first=True).astype(float)
X_df = pd.concat([d[num].astype(float), X_df], axis=1)
X = X_df.values            # 7043 x 30
```

Выводы по второй главе

Набор IBM Telco Customer Churn содержит 7 043 клиента и 19 содержательных признаков; доля ушедших — 26,5 %, на них приходится 30,5 % месячной выручки. Разведочный анализ выявил сильную связь оттока с типом договора, сроком обслуживания, оптоволоконным подключением, отсутствием защитных услуг и оплатой электронным чеком; пол и наличие телефонной линии с уходом не связаны. Обнаружены взаимодействие договора и типа подключения и избыточность признака TotalCharges. Разработанная схема эксперимента разделяет данные на обучающую и тестовую части, выбирает параметры только по кросс-проверке и оценивает модели совокупностью метрик с бутстрэп-интервалами.

3. Построение и оценка моделей прогнозирования оттока

3.1 Обучение моделей и подбор гиперпараметров

Для сравнения выбраны четыре модели, представляющие основные семейства методов из обзора первой главы: логистическая регрессия как интерпретируемая базовая модель, одиночное дерево решений, случайный лес как пример бэггинга и градиентный бустинг деревьев. Все алгоритмы реализованы автором на NumPy. Библиотека scikit-learn [54], которой обычно пользуются для таких задач, в вычислительной среде отсутствовала, и собственная реализация оказалась полезной: она заставила явно задать каждый шаг — от квантования признаков до правила остановки — и сверить его с первоисточниками [15, 38]. Основные фрагменты кода приведены в тексте, полный листинг — в приложении А.

Логистическая регрессия

Коэффициенты логистической регрессии найдены методом Ньютона, который для логистической функции потерь совпадает с итеративно перевзвешенным методом наименьших квадратов (IRLS) [38]. На каждой итерации вектор коэффициентов обновляется по правилу:

$$\beta^{(k+1)} = \beta^{(k)} - (Z^T W Z + \lambda I)^{-1}\left(Z^T(\hat{p} - y) + \lambda\beta^{(k)}\right) \qquad (14)$$

где Z — матрица стандартизованных признаков с единичным столбцом; W — диагональная матрица с элементами $\hat{p}_i(1 - \hat{p}_i)$; $\lambda = 1$ — коэффициент L2-регуляризации (свободный член не штрафуется).

Итерации прекращаются, когда максимальное изменение коэффициента становится меньше 10^{-8}; на полной обучающей выборке для этого хватает семи–восьми шагов, а обучение занимает сотые доли секунды. Вариант с перевзвешиванием классов отличается только тем, что в матрицу W и градиент входят веса наблюдений: 2,77 для ушедших клиентов и 1 для остальных.

```
class LogReg:
    def fit(s, X, y):
```

```
s.mu, s.sd = X.mean(0), X.std(0) + 1e-9
Z = np.c_[np.ones(len(X)), (X - s.mu) / s.sd]
b = np.zeros(Z.shape[1]); sw = np.where(y == 1, s.w1, 1.0)
R = np.eye(Z.shape[1]) * s.lam; R[0, 0] = 0
for _ in range(50):
    p = 1 / (1 + np.exp(-Z @ b))
    g = Z.T @ (sw * (p - y)) + R @ b
    H = Z.T @ (Z * (sw * p * (1 - p))[:, None]) + R
    step = np.linalg.solve(H, g); b -= step
    if np.abs(step).max() < 1e-8: break
s.b = b; return s
```

Дерево решений и случайный лес

Одиночное дерево строится по индексу Джини (формула 5) с ограничением глубины 5 и минимальным числом наблюдений в листе 30. Ограничения подобраны так, чтобы дерево оставалось читаемым — не более 32 листьев — и не запоминало отдельных клиентов. Прогноз дерева — доля ушедших в листе.

Случайный лес состоит из 200 деревьев глубиной до 10 уровней с минимумом 5 наблюдений в листе. По схеме Л. Бреймана [16] каждое дерево обучается на бутстрэп-выборке того же размера, что и обучающая часть, а в каждом узле выбирает лучшее разбиение среди 6 случайно отобранных признаков из 30. Число признаков в узле выбрано по распространённому правилу — около корня из общего числа признаков ($\sqrt{30} \approx 5{,}5$). Прогноз леса — среднее прогнозов деревьев.

Градиентный бустинг

Градиентный бустинг реализован по схеме Ньютона, которая используется в XGBoost [20]. На каждой итерации для каждого клиента вычисляются градиент $g_i = \hat{p}_i - y_i$ и вторая производная $h_i = \hat{p}_i(1 - \hat{p}_i)$ логистической функции потерь, а затем строится дерево регрессии. Значение в листе и прирост качества при разбиении узла на левую и правую части рассчитываются так:

$$w_{лист} = -\frac{\sum_{i \in лист} g_i}{\sum_{i \in лист} h_i + \lambda} \quad (15)$$

где λ = 1 — регуляризация значения листа.

$$Gain = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \quad (16)$$

где G_L, H_L и G_R, H_R — суммы градиентов и вторых производных в левом и правом дочерних узлах.

Поиск разбиения выполняется по гистограммам: каждый числовой признак заранее разбит на 32 интервала, поэтому суммы градиентов по интервалам считаются одной операцией подсчёта, а перебор порогов сводится к накопленным суммам. На каждой итерации дерево обучается на случайных 80 % клиентов (стохастический бустинг), что снижает переобучение.

Гиперпараметры бустинга — темп обучения и глубина деревьев — подобраны полным перебором по сетке из девяти сочетаний при фиксированном числе деревьев 300; критерием служил средний ROC-AUC пятикратной кросс-проверки на обучающей части. Для небольшой сетки полный перебор оправдан; при большем числе параметров эффективнее случайный поиск [14]. Результаты приведены в таблице 7.

Таблица 7 – Подбор гиперпараметров градиентного бустинга (ROC-AUC кросс-проверки)

Темп обучения ν	**Глубина 2**	**Глубина 3**	**Глубина 4**
0,03	0,8461	0,8458	0,8441
0,05	0,8455	0,8427	0,8381
0,10	0,8420	0,8362	0,8267

Источник: рассчитано автором; 300 деревьев, минимум 20 наблюдений в листе, подвыборка 80 %.

Лучшим оказалось сочетание с самым медленным темпом обучения и самыми мелкими деревьями — ν = 0,03 и глубина 2. Чем глубже деревья и быстрее обучение, тем ниже качество на проверке: модель начинает

подстраиваться под шум обучающих данных. Деревья глубины 2 описывают лишь попарные взаимодействия признаков, и то, что они выигрывают, само по себе говорит о простой структуре зависимости. Для этого набора данных сложные взаимодействия высокого порядка, по-видимому, не нужны, и это первый довод в пользу гипотезы работы.

Время обучения на одном процессорном ядре составило: логистическая регрессия — 0,1 с на все шесть обучений (пять блоков кросс-проверки и финальная модель), дерево решений — 0,5 с, градиентный бустинг — 14 с, случайный лес — 100 с. Для оператора с миллионами абонентов разница существенна: ежемесячный пересчёт логистической регрессии почти ничего не стоит, а ансамбль из сотен деревьев потребует распараллеливания.

3.2 Сравнение моделей и интерпретация факторов оттока

Качество пяти моделей — четырёх основных и логистической регрессии с перевзвешиванием классов — представлено в таблице 8. Средний ROC-AUC кросс-проверки и значение на отложенной выборке согласуются: на тесте все модели получают чуть более высокие оценки, но порядок моделей сохраняется, а расхождение укладывается в пределы разброса между блоками кросс-проверки.

Таблица 8 – Качество ранжирования моделей оттока

Модель	**ROC-AUC, кросс-проверка (среднее ± ст. откл.)**	**ROC-AUC, тест**	**95 % бутстрэп-интервал**	**AP, тест**	**Оценка Брайера**
Логистическая регрессия	0,841 ± 0,011	0,862	0,840–0,883	0,706	0,127
Логистическая регрессия с весами	0,841 ± 0,011	0,862	—	0,705	0,157
Дерево решений	0,823 ± 0,010	0,840	0,817–0,863	0,637	0,137
Случайный лес	0,842 ± 0,011	0,862	0,840–0,882	0,701	0,128
Градиентный бустинг	0,846 ± 0,010	0,865	0,843–0,885	0,710	0,126

Источник: рассчитано автором; тестовая выборка — 1 409 клиентов, 374 ушедших; бутстрэп — 1 000 повторных выборок.

Градиентный бустинг формально лучший по всем трём метрикам, однако разрыв с логистической регрессией и случайным лесом минимален: 0,865 против 0,862 по ROC-AUC и 0,710 против 0,706 по средней точности. Одиночное дерево заметно отстаёт, и это ожидаемо: 32 листа дают лишь 32 различных значения вероятности, и ранжирование внутри листа невозможно. Перевзвешивание классов не изменило ROC-AUC логистической регрессии, но ухудшило оценку Брайера с 0,127 до 0,157. Так и должно быть: веса сдвигают все вероятности вверх, порядок клиентов остаётся почти прежним, а вероятности перестают соответствовать фактической частоте ухода.

Можно ли считать бустинг лучше логистической регрессии? Бутстрэп-интервалы ROC-AUC двух моделей почти совпадают, но для вывода нужна не их взаимная оценка, а распределение разности, полученное на одних и тех же повторных выборках. Средняя разность ROC-AUC бустинга и логистической регрессии по 1 000 повторным выборкам составила 0,0026, а её 95-процентный интервал — от −0,0027 до 0,0084. Интервал включает ноль, следовательно, на данном объёме тестовой выборки превосходство бустинга статистически не подтверждается. Гипотеза работы в части качества ранжирования подтвердилась.

ROC-кривые трёх лучших моделей показаны на рисунке 4. Они практически сливаются; небольшое преимущество бустинга заметно лишь в зоне низкой доли ложных тревог, то есть среди клиентов с самым высоким риском.

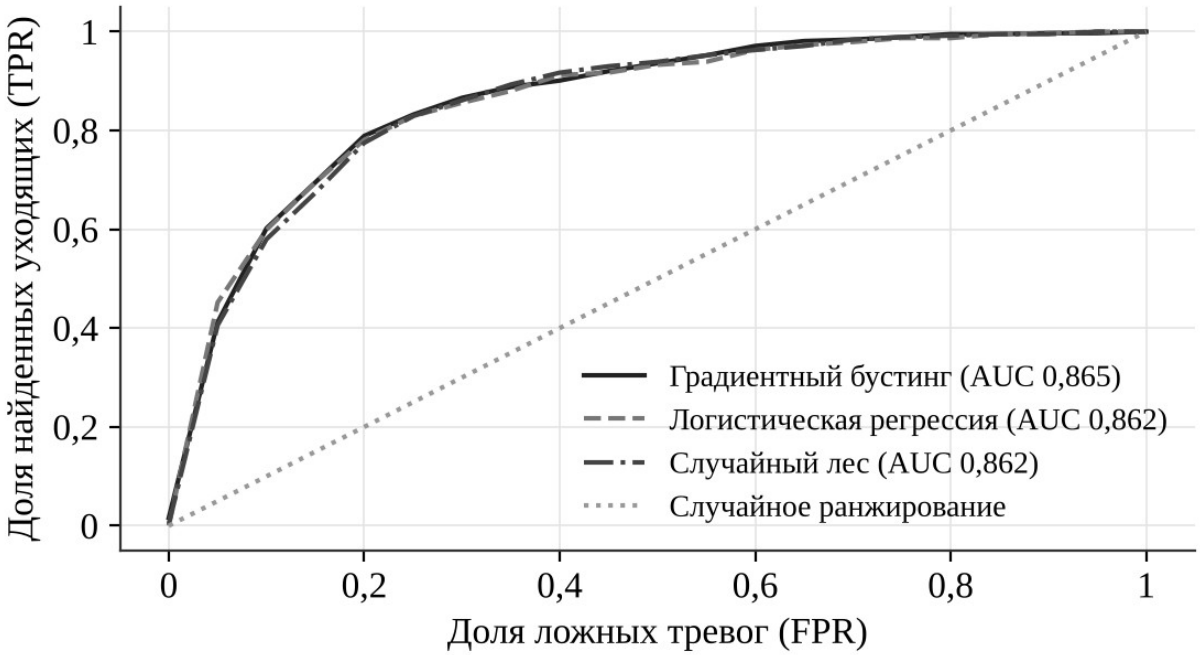


Рисунок 4 – ROC-кривые моделей на тестовой выборке

Источник: рассчитано автором; кривые интерполированы по 21 точке.

Метрики при фиксированном пороге приведены в таблице 9. При стандартном пороге 0,5 все модели, кроме логистической регрессии с весами, осторожны: точность около 0,70, но находят они лишь 53–59 % уходящих. Порог, подобранный по максимуму F1 на вне-блочных прогнозах кросс-проверки, оказался заметно ниже — 0,31–0,36 — и сдвигает баланс к полноте: модели находят 68–79 % уходящих при точности 0,58–0,61.

Таблица 9 – Точность, полнота и F1-мера на тестовой выборке

Модель	Порог	Precision	Recall	F1	Accuracy
Логистическая регрессия	0,50	0,695	0,586	0,636	0,822
	0,31*	0,578	0,794	0,669	0,791
Логистическая регрессия с весами	0,50	0,539	0,821	0,650	0,766
	0,55*	0,571	0,802	0,667	0,788
Дерево решений	0,50	0,700	0,388	0,499	0,793
	0,36*	0,605	0,676	0,639	0,797
Случайный лес	0,50	0,703	0,532	0,606	0,816
	0,34*	0,595	0,730	0,655	0,796
Градиентный бустинг	0,50	0,694	0,535	0,604	0,814
	0,34*	0,604	0,743	0,667	0,803

* Порог подобран по максимуму F1 на вне-блочных прогнозах кросс-проверки обучающей части. Источник: рассчитано автором.

По F1-мере при подобранном пороге логистическая регрессия (0,669) и градиентный бустинг (0,667) снова неразличимы. Accuracy в таблице приведена для полноты картины и хорошо иллюстрирует довод первой главы: при пороге 0,5 она выше, чем при подобранном, хотя модель при этом пропускает вдвое больше уходящих клиентов: у логистической регрессии 155 против 77. Матрица ошибок градиентного бустинга при пороге 0,34 содержит 278 верно найденных уходящих, 96 пропущенных, 182 ложные тревоги и 853 верно определённых лояльных клиента.

Для кампании удержания удобнее всего показатель лифта: во сколько раз доля уходящих в верхней части списка выше средней по базе. Распределение ушедших по децилям прогноза градиентного бустинга показано на рисунке 5. В верхние 10 % списка (141 клиент) попали 116 ушедших, то есть 82,3 % — в 3,1 раза больше средней доли 26,5 %. Верхние три дециля содержат 264 из 374 ушедших, или 70,6 %, а в нижних трёх децилях ушедших всего 11.

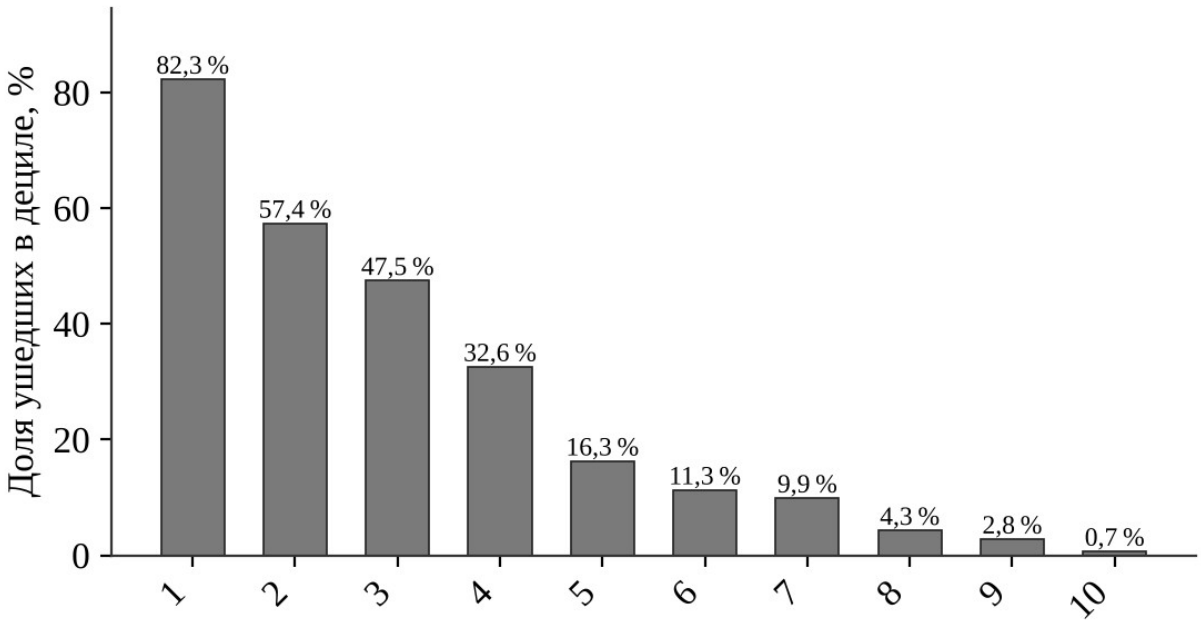


Рисунок 5 – Доля ушедших клиентов по децилям прогноза градиентного бустинга

Дециль 1 — 10 % клиентов с наибольшей оценкой риска. Источник: рассчитано автором на тестовой выборке.

Калибровка вероятностей и анализ ошибок

Для экономического расчёта важно не только упорядочить клиентов, но и правильно оценить саму вероятность ухода. Проверка калибровки выполнена так: клиенты тестовой выборки упорядочены по прогнозу и разбиты на десять равных групп, и в каждой группе средний прогноз сравнивается с фактической долей ушедших (таблица 10). У хорошо откалиброванной модели эти числа близки.

Таблица 10 – Калибровка прогнозов на тестовой выборке: средний прогноз и фактическая доля ушедших по децилям, %

Дециль прогноза	**ЛР: прогноз**	**ЛР: факт**	**ГБ: прогноз**	**ГБ: факт**	**ЛР с весами: прогноз**	**ЛР с весами: факт**
1 (мин.)	0,5	1,4	1,1	0,7	1,3	1,4
2	1,7	2,8	2,2	2,8	4,7	2,8
3	4,2	5,7	4,6	4,3	11,0	5,0
4	8,2	9,2	8,6	9,9	20,1	9,2
5	14,8	10,6	14,6	11,3	32,6	11,3
6	22,3	18,4	22,5	17,0	44,3	18,4
7	33,2	31,2	31,6	32,6	57,9	31,2
8	46,5	41,8	44,4	47,5	70,4	43,3
9	59,8	59,6	57,5	56,7	79,9	57,4
10 (макс.)	71,4	85,0	74,0	82,9	86,8	85,7

ЛР — логистическая регрессия, ГБ — градиентный бустинг. Источник: рассчитано автором.

Логистическая регрессия и бустинг без весов откалиброваны удовлетворительно: не менее чем в восьми децилях из десяти расхождение не превышает пяти процентных пунктов. Обе модели недооценивают риск в верхнем дециле — прогнозируют 71–74 %, а уходят 83–85 % клиентов. Модель с весами, как и ожидалось, систематически завышает вероятность: в пятом дециле она обещает 32,6 % ухода при фактических 11,3 %. Её прогноз годится для ранжирования, но подставлять его в формулу прибыли без перекалибровки нельзя [52]. В расчёте параграфа 3.4 используются прогнозы моделей без весов, причём выбор охвата по кривой прибыли зависит только от порядка клиентов, а не от абсолютных значений вероятностей.

Полезно посмотреть и на то, кого модель пропускает. При пороге 0,31 логистическая регрессия не заметила 77 из 374 ушедших клиентов. Это совсем другие люди, чем найденные: медианный срок обслуживания пропущенных — 48 месяцев против 7 у найденных, помесячный договор у 49 % против 98 %, без интернета — 26 % против 0,3 %. Иначе говоря, модель не видит ухода долгосрочных и спокойных по профилю клиентов. Скорее всего, их уход вызван причинами, которых нет в данных: переездом, сменой домохозяйства, конкретным конфликтом с сервисом. Это не недостаток

алгоритма, а предел информации, содержащейся в признаках договора и тарифа.

Ложные тревоги, напротив, похожи на найденных уходящих как две капли воды: медианный срок обслуживания — 12 месяцев, помесячный договор — 98 %, оптоволокно — 66 %, оплата электронным чеком — 60 %. Это клиенты из той же группы риска, которые в данном квартале остались. Для кампании удержания они не бесполезны: велика вероятность, что часть из них уйдёт в следующих кварталах, и предложение долгосрочного договора для них тоже имеет смысл.

Наконец, модели сходятся в оценках. Коэффициент корреляции прогнозов логистической регрессии и градиентного бустинга на тестовой выборке равен 0,976, а из 510 клиентов с наибольшим риском по каждой модели 476 (93,3 %) совпадают. Две модели разной природы, обученные независимо, выделяют практически одну и ту же группу клиентов; это ещё один довод в пользу того, что выбор алгоритма для данного набора признаков вторичен.

Важность признаков

Перестановочная важность признаков для градиентного бустинга представлена на рисунке 6. Признаки чётко делятся на две группы. Три фактора — срок обслуживания, тип договора и тип интернет-подключения — снижают ROC-AUC при перестановке на 0,078, 0,056 и 0,054 соответственно. Вклад всех остальных признаков не превышает 0,005, а у признаков OnlineBackup, Partner, DeviceProtection и gender он равен нулю в пределах погрешности.

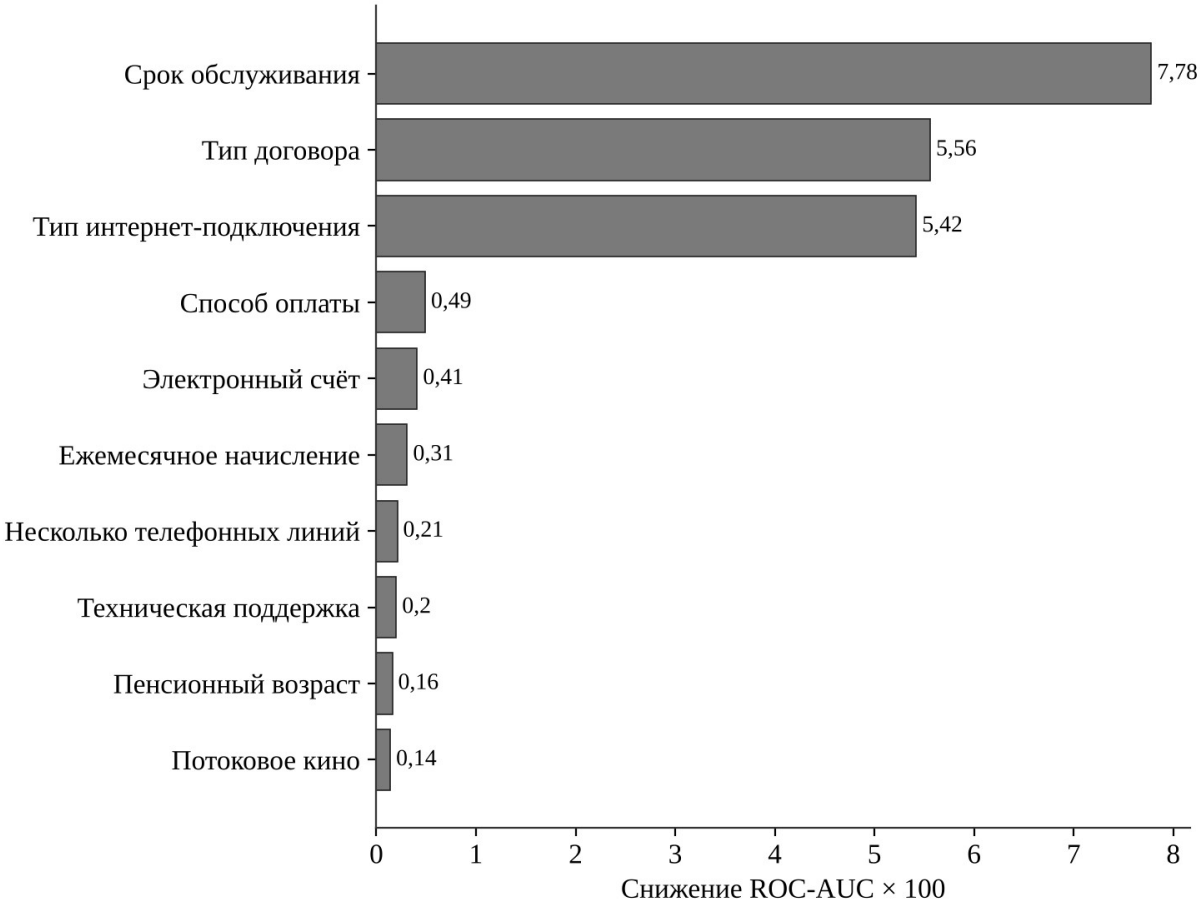


Рисунок 6 – Перестановочная важность признаков градиентного бустинга

Среднее по 10 перестановкам на тестовой выборке; показаны 10 признаков из 19.
Источник: рассчитано автором.

Результат совпадает с выводами разведочного анализа, но уточняет их. По V Крамера защитная услуга OnlineSecurity и техподдержка TechSupport шли сразу за типом договора, а в модели их вклад мал. Объяснение — в пересечении информации: клиенты без этих услуг чаще имеют помесячный договор и оптоволокно, и после учёта главных факторов отдельный вклад дополнительных услуг почти исчезает. Ежемесячное начисление и сумма начислений за всё время тоже малозначимы: первое во многом определяется набором услуг, второе — сроком обслуживания. Модель извлекает нужную информацию из первичных признаков.

Встроенная важность бустинга по суммарному приросту (формула 16) даёт тот же порядок первых трёх признаков, но распределяет вклад иначе: тип договора — 30,3 %, срок обслуживания — 26,5 %, тип подключения — 21,7 %, способ оплаты — 7,3 %. Перестановочная оценка надёжнее, потому

что она измеряется на отложенной выборке и в единицах метрики качества [29, 62].

Логистическая регрессия позволяет посмотреть на направление влияния (таблица 11). Коэффициенты приведены для стандартизованных признаков, поэтому их можно сравнивать между собой; экспонента коэффициента — отношение шансов при росте признака на одно стандартное отклонение [39].

Таблица 11 – Наибольшие по модулю коэффициенты логистической регрессии

Признак	**Коэффициент β**	**Отношение шансов exp(β)**	**Интерпретация**
tenure	−1,494	0,225	Рост срока на 24,6 мес. снижает шансы ухода в 4,5 раза
TotalCharges	+0,784	2,191	Артефакт коллинеарности с tenure
Contract = Two year	−0,630	0,533	Двухлетний договор удерживает
InternetService = Fiber optic	+0,555	1,742	Оптоволокно повышает риск
MonthlyCharges	−0,501	0,606	Артефакт коллинеарности с набором услуг
Contract = One year	−0,284	0,753	Годовой договор удерживает
StreamingTV = Yes	+0,175	1,192	Слабое повышение риска
StreamingMovies = Yes	+0,171	1,187	Слабое повышение риска
OnlineSecurity = Yes	−0,157	0,855	Защитная услуга удерживает
PaymentMethod = Electronic check	+0,150	1,162	Ручная оплата повышает риск
PaperlessBilling = Yes	+0,135	1,144	Слабое повышение риска
MultipleLines = Yes	+0,128	1,136	Слабое повышение риска

Источник: рассчитано автором; признаки стандартизованы по обучающей выборке, $\lambda = 1$.

Большинство знаков согласуется с разведочным анализом. Исключения поучительны. Положительный коэффициент при TotalCharges противоречит

тому, что ушедшие клиенты в среднем заплатили меньше. Противоречие мнимое: при фиксированном сроке обслуживания большая сумма начислений означает более дорогой тариф, а модель уже учла срок отдельным признаком с большим отрицательным весом. Отрицательный коэффициент при MonthlyCharges объясняется тем же образом — ежемесячная плата почти линейно складывается из подключённых услуг, и её отдельный вклад после учёта оптоволокна и потоковых сервисов меняет знак. Для отчёта перед бизнесом такие коэффициенты нужно либо пояснять, либо исключить один из коллинеарных признаков и переобучить модель.

Обе модели рисуют одинаковый портрет уходящего клиента: обслуживается меньше года, договор помесячный, интернет оптоволоконный, защитной услуги и техподдержки нет, счёт оплачивает электронным чеком. Этот портрет полезен отделу маркетинга не меньше, чем сами вероятности: он подсказывает, какое предложение делать. Для клиента на помесячном договоре естественное удерживающее действие — скидка при переходе на годовой договор; для клиента с оптоволокном без техподдержки — бесплатная техподдержка на несколько месяцев.

3.3 Проект встраивания модели в информационную систему оператора

Модель, которая существует только в тетради аналитика, денег не приносит. Д. Скалли с соавторами показали, что в промышленной системе машинного обучения код самой модели занимает малую часть, а основная сложность — в сборе данных, извлечении признаков, проверке, развёртывании и мониторинге [60]. Поэтому в работе предложена архитектура встраивания модели оттока в информационную систему оператора (рисунок 7).

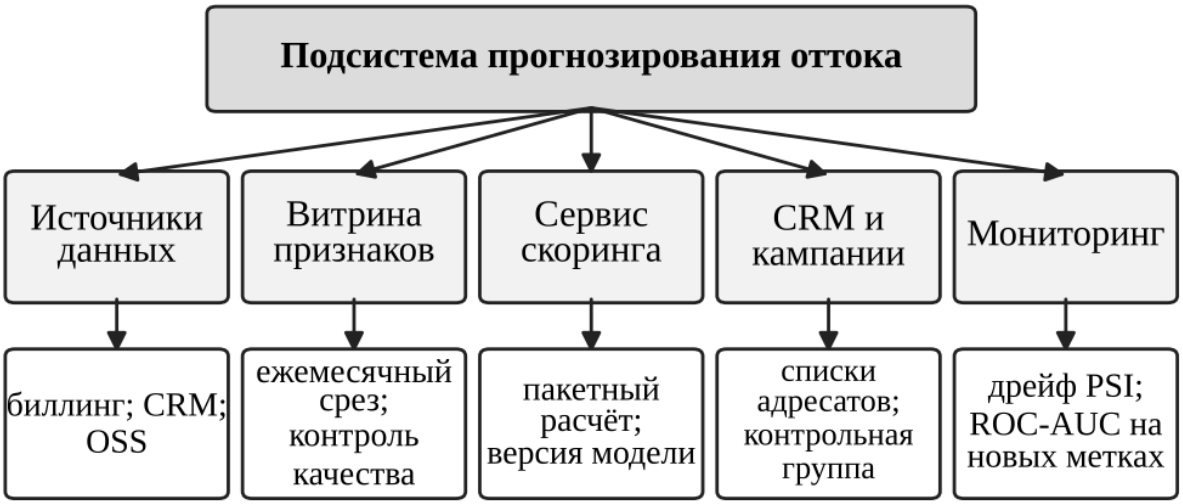


Рисунок 7 – Архитектура подсистемы прогнозирования оттока

Требования к подсистеме сформулированы исходя из того, как модель будет использоваться. Функциональные требования: ежемесячный расчёт оценки риска для всех действующих абонентов; формирование списка адресатов по порогу, утверждённому отделом удержания; выделение контрольной группы случайным образом; сохранение прогноза, версии модели и признаков на дату среза; отчёт о качестве модели и эффекте кампании. Нефункциональные требования: полный пересчёт базы не дольше одной ночи; воспроизводимость прогноза по сохранённой версии модели; доступ к оценкам риска только у сотрудников с соответствующей ролью; журналирование обращений к данным абонентов.

Ежемесячный цикл работы подсистемы выглядит так. В первый рабочий день месяца витрина формирует срез на последний день предыдущего месяца. Сервис скоринга рассчитывает оценки и передаёт их в CRM, где отдел удержания видит распределение риска, утверждает порог и состав предложения. Список, за вычетом контрольной группы, уходит в каналы контакта — контактный центр, SMS, личный кабинет. Через месяц подсистема получает фактические метки ухода, пересчитывает качество модели и сравнивает отток в целевой и контрольной группах. Результаты сравнения становятся исходными данными для настройки следующей кампании.

Источники данных. Признаки, использованные в работе, распределены по трём системам оператора: договор и тариф хранятся в биллинге, данные о клиенте и обращениях — в CRM, сведения об оборудовании и качестве связи — в системах эксплуатации (OSS). В реальной модели к ним добавляются агрегаты трафика и обращений, которых нет в наборе IBM.

Витрина признаков. Ежемесячно, на последний день месяца, формируется срез по всем действующим абонентам с фиксированным набором признаков. Тот же код используется и для обучения, и для расчёта прогноза — это исключает расхождение между обучающими и рабочими признаками, частую причину деградации моделей. При формировании среза выполняются проверки: доля пропусков по каждому полю, допустимые диапазоны значений, совпадение числа абонентов с данными биллинга.

Сервис скоринга. Поскольку кампании удержания планируются раз в месяц, онлайн-прогноз не нужен; достаточно пакетного расчёта по витрине. Для логистической регрессии он сводится к умножению матрицы признаков на вектор из 31 коэффициента и выполняется за секунды даже на миллионах абонентов. Каждый прогноз сохраняется вместе с версией модели и датой среза, чтобы позднее можно было восстановить, на каком основании клиент попал в кампанию.

CRM и кампании. В модуль кампаний передаётся не весь список, а отобранная часть: клиенты с оценкой риска выше порога, рассчитанного по кривой прибыли (параграф 3.4). Случайно выбранные 10 % из них выделяются в контрольную группу, которой предложение не делается. Только сравнение с контрольной группой показывает реальный эффект кампании, ведь часть клиентов с высоким риском осталась бы и без контакта [13, 25]. Накопленные данные о реакции на предложение позже позволят перейти от модели оттока к модели прироста отклика.

Мониторинг. Модель стареет: меняются тарифы, конкурентные предложения и состав базы, и зависимость, найденная на старых данных,

перестаёт выполняться. Это явление называют дрейфом концепции [31]. Простейший сигнал о дрейфе входных данных — индекс стабильности популяции (PSI), который сравнивает распределение признака или прогноза в текущем месяце с распределением на момент обучения:

$$PSI = \sum_{b=1}^{B} (a_b - e_b) \ln \frac{a_b}{e_b} \tag{17}$$

где e_b и a_b — доли наблюдений в интервале b на момент обучения и в текущем срезе; B — число интервалов (обычно 10).

На практике значения PSI ниже 0,1 обычно считают признаком стабильности, а выше 0,25 — сигналом к пересмотру модели. Это эмпирический ориентир, а не статистический критерий, поэтому в проекте оно дополняется прямой проверкой: через месяц после скоринга становится известно, кто из абонентов ушёл, и ROC-AUC пересчитывается на новых метках. Если он опускается ниже 0,80, модель переобучается на свежих данных. Компоненты подсистемы и их назначение сведены в таблицу 12.

Таблица 12 – Компоненты подсистемы прогнозирования оттока

Компонент	Функция	Периодичность	Ответственный
Витрина признаков	Ежемесячный срез признаков по действующим абонентам, контроль качества данных	Ежемесячно	Отдел данных
Сервис скоринга	Расчёт вероятности ухода, хранение прогнозов с версией модели	Ежемесячно	Отдел данных
Модуль кампаний CRM	Отбор адресатов по порогу, выделение контрольной группы, учёт откликов	Ежемесячно	Отдел удержания
Мониторинг	PSI признаков и прогноза, ROC-AUC на новых метках, отчёт об эффекте	Ежемесячно	Отдел данных
Переобучение	Обучение на свежих данных, сравнение с действующей моделью, утверждение	Раз в квартал или по сигналу	Отдел данных, отдел удержания

Источник: разработано автором.

Правовые и этические требования. Обработка данных абонентов для моделирования должна иметь правовое основание по закону «О персональных данных», а обучающие выборки обезличиваются [2]. Решение по прогнозу ограничивается предложением, выгодным для клиента; ухудшать условия договора или отказывать в услуге по сигналу модели нельзя без участия сотрудника — этого требует статья 16 того же закона. Документация модели — состав признаков, качество, дата обучения, ограничения — ведётся в соответствии с принципами ГОСТ Р 59276-2020 [5]. Выбор интерпретируемой логистической регрессии упрощает выполнение этих требований: сотрудник отдела удержания видит, какие признаки повысили риск конкретного клиента, и может проверить разумность прогноза [58].

3.4 Экономическая оценка применения модели в кампании удержания

Экономический эффект оценивается по формуле С. Неслина (формула 10), применённой к каждому клиенту тестовой выборки. Пожизненная ценность удержанного клиента заменена его ежемесячным начислением, умноженным на срок последующего обслуживания; маржа не выделяется, так как набор данных не содержит сведений о затратах. Кампания обращается к k клиентам с наибольшей оценкой риска, и прибыль рассчитывается так:

$$\Pi(k) = \sum_{i \in S_k} [y_i \gamma (M m_i - \delta m_i D) - (1 - y_i)\, \delta m_i D - c] \tag{18}$$

где S_k — k клиентов с наибольшей оценкой риска; y_i — фактический уход клиента; m_i — ежемесячное начисление; γ — доля уходящих, которых удалось удержать; M — срок последующего обслуживания удержанного клиента, мес.; δ — размер скидки; D — срок действия скидки, мес.; c — затраты на контакт.

Формула следует допущению Неслина о том, что скидку получают все охваченные клиенты, которые не собирались уходить, и те уходящие, которые приняли предложение. Это консервативный вариант: на практике часть лояльных клиентов не воспользуется предложением. Значения параметров — условные данные для примера; они приведены в таблице 13 и не взяты из отчётности какой-либо компании.

Таблица 13 – Параметры расчёта экономического эффекта (условные данные для примера)

Параметр	Обозначение	Базовое значение	Диапазон в анализе чувствительности
Затраты на контакт (звонок или сообщение, работа оператора)	c	5 у. е.	—
Скидка на ежемесячное начисление	δ	25 %	—
Срок действия скидки	D	6 мес.	—
Доля удержанных среди уходящих, получивших предложение	γ	30 %	20–40 %
Срок последующего обслуживания удержанного клиента	M	12 мес.	6–24 мес.

Условные данные для примера, выбраны автором для иллюстрации методики. Ежемесячные начисления m_i взяты из набора данных IBM [68].

Кривые прибыли для логистической регрессии, градиентного бустинга и случайного отбора клиентов показаны на рисунке 8. Кривые моделей сначала растут: в верхней части списка большинство клиентов действительно уходит, и удержание части из них окупает скидки для остальных. Затем рост замедляется, достигает максимума и сменяется падением, когда в кампанию начинают попадать в основном лояльные клиенты, которые получают скидку без всякой пользы для оператора. Случайный отбор убыточен при любом охвате.

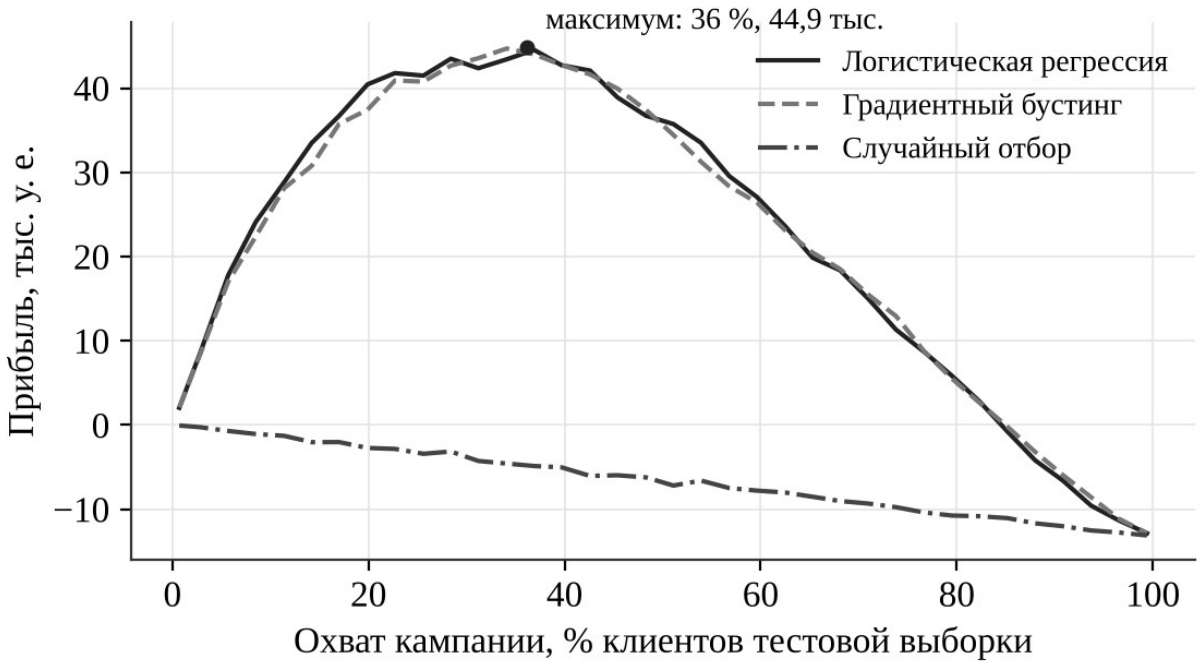


Рисунок 8 – Прибыль кампании удержания в зависимости от охвата (условные данные для примера)

Параметры — таблица 13 (условные данные для примера). Источник: рассчитано автором.

При базовых параметрах максимум прибыли для логистической регрессии достигается при охвате 510 клиентов из 1 409 (36,2 % тестовой выборки) и составляет 44 926 у. е. В этот список попадают 296 из 374 уходящих клиентов; удержание 30 % из них даёт 81 401 у. е. выручки за 12 месяцев. Расходы складываются из скидок удержанным клиентам (10 175 у. е.), скидок 214 лояльным клиентам, попавшим в кампанию (23 750 у. е.), и затрат на контакты (2 550 у. е.). Градиентный бустинг даёт практически тот же результат — 44 741 у. е. при охвате 470 клиентов, случайный лес — 43 405 у. е. при охвате 490. Разница между моделями составляет 0,4–3,4 % прибыли и меньше, чем погрешность любого из параметров таблицы 13.

Для сравнения: если бы предложение получили все 1 409 клиентов, кампания принесла бы убыток 13 154 у. е., а случайный отбор 500 клиентов — убыток около 4 300 у. е. Именно здесь модель создаёт ценность: она превращает убыточное мероприятие в прибыльное за счёт выбора адресатов. Порог вероятности, соответствующий оптимальному охвату, для

логистической регрессии равен 0,31 и почти совпадает с порогом, подобранным по F1-мере. Это совпадение случайно: при других параметрах кампании оптимальный порог по прибыли сместится, а порог по F1-мере останется прежним.

Результаты анализа чувствительности приведены в таблице 14. Оптимальный охват и прибыль сильно зависят от допущений о реакции клиентов.

Таблица 14 – Чувствительность оптимального охвата и прибыли к параметрам кампании (логистическая регрессия)

γ, доля удержанных	**M, мес.**	**Оптимальный охват, клиентов**	**Доля тестовой выборки, %**	**Прибыль, у. е.**
20 %	6	170	12,1	5 553
20 %	12	280	19,9	23 513
20 %	24	530	37,6	75 908
30 %	6	280	19,9	11 420
30 %	12	510	36,2	44 926
30 %	24	590	41,9	128 210
40 %	6	280	19,9	18 676
40 %	12	530	37,6	68 991
40 %	24	750	53,2	182 150

Условные данные для примера: расчёт на тестовой выборке из 1 409 клиентов при c = 5 у. е., δ = 25 %, D = 6 мес.; охват перебирался с шагом 10 клиентов. Источник: рассчитано автором.

При пессимистичных допущениях (удерживается каждый пятый уходящий, и он остаётся лишь на полгода) выгодно обращаться только к 12 % базы, а прибыль падает в восемь раз по сравнению с базовым вариантом. При оптимистичных допущениях охват можно расширить до половины базы. Разброс прибыли между сценариями — более чем тридцатикратный, тогда как разница между моделями не превышает нескольких процентов. Вторая часть гипотезы работы подтвердилась: экономический результат определяется параметрами кампании и выбором охвата, а не алгоритмом.

Если перенести базовый результат на всю клиентскую базу набора — 7 043 клиента, в 5 раз больше тестовой выборки, — ожидаемая прибыль

одной кампании составит около 225 тыс. у. е. Эта величина — оценка по условным параметрам, а не прогноз для реальной компании. Её назначение — показать масштаб и логику расчёта. Для реального оператора параметр γ нужно измерять в пилотной кампании с контрольной группой, а срок М — по истории удержанных клиентов, например с помощью модели удержания П. Фейдера и Б. Харди [27].

Отсюда практическая рекомендация: внедрять логистическую регрессию с порогом, выбранным по кривой прибыли, а первую кампанию проводить как эксперимент на ограниченном охвате с контрольной группой. Полученная оценка γ уточнит оптимальный охват уже для второй кампании. Переход к более сложной модели имеет смысл только тогда, когда появятся новые признаки — данные о трафике, обращениях, качестве связи, — на которых ансамбли деревьев могут показать устойчивое преимущество.

Выводы по третьей главе

Четыре модели оттока обучены и оценены на открытых данных IBM Telco. Градиентный бустинг с подобранными параметрами показал ROC-AUC 0,865, логистическая регрессия и случайный лес — 0,862, одиночное дерево — 0,840. Бутстрэп-интервал разности ROC-AUC бустинга и логистической регрессии (от −0,003 до 0,008) включает ноль, так что преимущество бустинга не доказано. Главные факторы оттока — срок обслуживания, тип договора и оптоволоконное подключение. Предложена архитектура встраивания модели в информационную систему оператора с пакетным скорингом, контрольной группой и мониторингом дрейфа. Экономический расчёт на условных параметрах показал, что модель превращает убыточную сплошную рассылку предложений в прибыльную кампанию, а оптимальный охват зависит главным образом от доли удержанных клиентов и срока их последующего обслуживания.

ЗАКЛЮЧЕНИЕ

В работе решена задача прогнозирования оттока клиентов телекоммуникационной компании методами машинного обучения — от экономической постановки до расчёта эффекта и проекта внедрения. Поставленные во введении задачи выполнены, и по каждой из них получены следующие результаты.

Первое. Отток рассмотрен как экономическая категория. Его цена выражается через пожизненную ценность клиента, и даже условный расчёт показывает, что рост месячного удержания с 97 до 98 % увеличивает ценность клиента примерно на треть. Проактивное удержание требует модели, ранжирующей клиентов по риску, и правила, определяющего охват кампании. Использование модели ограничено законами «О связи» и «О персональных данных»: сведения об абонентах относятся к информации ограниченного доступа, а решения, затрагивающие права клиента, нельзя принимать исключительно автоматически.

Второе. Прогнозирование оттока формализовано как бинарная классификация с дисбалансом классов. Обосновано, что доля верных ответов для неё непригодна, и предложен набор критериев: ROC-AUC и средняя точность для ранжирования, оценка Брайера для качества вероятностей, точность, полнота и F1-мера при выбранном пороге и прибыль кампании по формуле Неслина.

Третье. Обзор исследований показал, что на табличных данных операторов различия между хорошими алгоритмами невелики, ансамбли деревьев стабильно входят в число лучших, а современная литература смещает акцент с точности прогноза на пользу для кампании удержания и на интерпретируемость моделей.

Четвёртое. Разведочный анализ набора IBM Telco Customer Churn (7 043 клиента, доля ушедших 26,5 %, на них приходится 30,5 % месячной выручки) выявил сильную связь оттока с типом договора (от 42,7 % на помесячном договоре до 2,8 % на двухлетнем), сроком обслуживания,

оптоволоконным подключением, отсутствием защитных услуг и оплатой электронным чеком. Пол и наличие телефонной линии с уходом не связаны.

Пятое. Построены и сравнены логистическая регрессия, дерево решений, случайный лес и градиентный бустинг; алгоритмы реализованы на NumPy, гиперпараметры подобраны по кросс-проверке. На отложенной выборке ROC-AUC составил 0,862, 0,840, 0,862 и 0,865 соответственно. Разность ROC-AUC бустинга и логистической регрессии статистически незначима: 95-процентный бутстрэп-интервал от −0,003 до 0,008. Перестановочная оценка важности выделила три ключевых фактора — срок обслуживания, тип договора и тип подключения, — на которые приходится почти весь вклад в качество прогноза.

Шестое. Разработана архитектура подсистемы прогнозирования оттока в информационной системе оператора: витрина признаков, пакетный скоринг, передача списков в CRM с контрольной группой, мониторинг дрейфа по индексу стабильности популяции и ROC-AUC на новых метках. Расчёт на условных параметрах показал, что при базовом сценарии оптимально обращаться примерно к трети клиентов с наибольшим риском, а в зависимости от допущений оптимальный охват меняется от 12 до 53 % базы. Сплошная рассылка предложений в том же сценарии убыточна.

Гипотеза исследования подтвердилась в обеих частях: ансамбли деревьев не дали значимого выигрыша в ROC-AUC по сравнению с регуляризованной логистической регрессией, а экономический результат кампании определяется прежде всего охватом и параметрами реакции клиентов. Для внедрения рекомендована логистическая регрессия: при сопоставимом качестве она обучается за доли секунды, её прогноз объясним сотруднику и клиенту, а документировать её для соответствия требованиям доверия к системам искусственного интеллекта проще.

Ограничения работы связаны с данными. Набор IBM описывает вымышленную компанию одним срезом, в нём нет сведений о потреблении услуг, обращениях и качестве связи, а экономические параметры кампании

условны. Перспективы исследования — проверка методики на данных реального оператора с признаками динамики потребления, переход от модели оттока к модели прироста отклика на основе результатов кампаний с контрольной группой и учёт неопределённости параметров кампании через ожидаемую максимальную прибыль.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Российская Федерация. Законы. О связи : Федеральный закон от 07.07.2003 № 126-ФЗ : [ред. действующая]. – URL: https://www.consultant.ru/document/cons_doc_LAW_43224/ (дата обращения: 22.10.2026). – Текст : электронный.

2. Российская Федерация. Законы. О персональных данных : Федеральный закон от 27.07.2006 № 152-ФЗ : [ред. действующая]. – URL: https://www.consultant.ru/document/cons_doc_LAW_61801/ (дата обращения: 22.10.2026). – Текст : электронный.

3. Российская Федерация. Законы. О внесении изменений в Федеральный закон «О связи» и статьи 333.33 и 333.34 части второй Налогового кодекса Российской Федерации : Федеральный закон от 25.12.2012 № 253-ФЗ. – URL: https://www.consultant.ru/document/cons_doc_LAW_139758/ (дата обращения: 22.10.2026). – Текст : электронный.

4. Российская Федерация. Президент. О развитии искусственного интеллекта в Российской Федерации : Указ Президента Российской Федерации от 10.10.2019 № 490 : (вместе с Национальной стратегией развития искусственного интеллекта на период до 2030 года) : [ред. действующая]. – URL: https://www.consultant.ru/document/cons_doc_LAW_335184/ (дата обращения: 22.10.2026). – Текст : электронный.

5. ГОСТ Р 59276-2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения : национальный стандарт Российской Федерации : издание официальное : утверждён и введён в действие Приказом Федерального агентства по техническому регулированию и метрологии от 23 декабря 2020 г. № 1371-ст : введён впервые : дата введения 2021-03-01. – Москва : Стандартинформ, 2021. – Текст : непосредственный.

6. Воронцов, К. В. Машинное обучение (курс лекций) / К. В. Воронцов. – Текст : электронный // MachineLearning.ru : профессиональный информационно-аналитический ресурс. – URL: http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_(курс_лекций,_К.В.Воронцов) (дата обращения: 22.10.2026).

7. Джеймс, Г. Введение в статистическое обучение с примерами на языке R / Г. Джеймс, Д. Уиттон, Т. Хасти, Р. Тибширани ; перевод с английского С. Э. Мастицкого. – Москва : ДМК Пресс, 2016. – 450 с. – ISBN 978-5-97060-293-5. – Текст : непосредственный.

8. Мюллер, А. Введение в машинное обучение с помощью Python : руководство для специалистов по работе с данными / А. Мюллер, С. Гвидо ; перевод с английского и редакция А. В. Груздева. – Москва [и др.] : Диалектика, 2017. – 472, [1] с. – ISBN 978-5-9908910-8-1. – Текст : непосредственный.

9. Рашка, С. Python и машинное обучение : крайне необходимое издание по новейшей предсказательной аналитике для более глубокого понимания методологии машинного обучения : перевод с английского / С. Рашка. – Москва : ДМК Пресс, 2017. – ISBN 978-5-97060-409-0. – Текст : непосредственный.

10. Флах, П. Машинное обучение : наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах ; перевод с английского А. А. Слинкина. – Москва : ДМК Пресс, 2015. – 399 с. – ISBN 978-5-97060-273-7. – Текст : непосредственный.

11. Ahmad, A. K. Customer churn prediction in telecom using machine learning in big data platform / A. K. Ahmad, A. Jafar, K. Aljoumaa // Journal of Big Data. – 2019. – Vol. 6. – Art. 28. – DOI 10.1186/s40537-019-0191-6.

12. Amin, A. Customer churn prediction in telecommunication industry using data certainty / A. Amin, F. Al-Obeidat, B. Shah [et al.] // Journal of

Business Research. – 2019. – Vol. 94. – P. 290–301. – DOI 10.1016/j.jbusres.2018.03.003.

13. Ascarza, E. Retention futility: targeting high-risk customers might be ineffective / E. Ascarza // Journal of Marketing Research. – 2018. – Vol. 55, № 1. – P. 80–98. – DOI 10.1509/jmr.16.0163.

14. Bergstra, J. Random search for hyper-parameter optimization / J. Bergstra, Y. Bengio // Journal of Machine Learning Research. – 2012. – Vol. 13. – P. 281–305.

15. Bishop, C. M. Pattern Recognition and Machine Learning / C. M. Bishop. – New York : Springer, 2006. – 738 p. – ISBN 978-0-387-31073-2.

16. Breiman, L. Random forests / L. Breiman // Machine Learning. – 2001. – Vol. 45, № 1. – P. 5–32. – DOI 10.1023/A:1010933404324.

17. Breiman, L. Classification and Regression Trees / L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. – Belmont : Wadsworth, 1984. – 358 p.

18. Burez, J. Handling class imbalance in customer churn prediction / J. Burez, D. Van den Poel // Expert Systems with Applications. – 2009. – Vol. 36, № 3. – P. 4626–4636. – DOI 10.1016/j.eswa.2008.05.027.

19. Chawla, N. V. SMOTE: synthetic minority over-sampling technique / N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer // Journal of Artificial Intelligence Research. – 2002. – Vol. 16. – P. 321–357. – DOI 10.1613/jair.953.

20. Chen, T. XGBoost: a scalable tree boosting system / T. Chen, C. Guestrin // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York : ACM, 2016. – P. 785–794. – DOI 10.1145/2939672.2939785.

21. Coussement, K. Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques / K. Coussement, D. Van den Poel // Expert Systems with Applications. – 2008. – Vol. 34, № 1. – P. 313–327. – DOI 10.1016/j.eswa.2006.09.038.

22. Cox, D. R. The regression analysis of binary sequences / D. R. Cox // Journal of the Royal Statistical Society. Series B (Methodological). – 1958. – Vol. 20, № 2. – P. 215–242.

23. Davis, J. The relationship between precision-recall and ROC curves / J. Davis, M. Goadrich // Proceedings of the 23rd International Conference on Machine Learning. – New York : ACM, 2006. – P. 233–240. – DOI 10.1145/1143844.1143874.

24. De Caigny, A. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees / A. De Caigny, K. Coussement, K. W. De Bock // European Journal of Operational Research. – 2018. – Vol. 269, № 2. – P. 760–772. – DOI 10.1016/j.ejor.2018.02.009.

25. Devriendt, F. Why you should stop predicting customer churn and start using uplift models / F. Devriendt, J. Berrevoets, W. Verbeke // Information Sciences. – 2021. – Vol. 548. – P. 497–515. – DOI 10.1016/j.ins.2019.12.075.

26. Efron, B. An Introduction to the Bootstrap / B. Efron, R. J. Tibshirani. – New York : Chapman & Hall, 1993. – 436 p. – ISBN 978-0-412-04231-7.

27. Fader, P. S. How to project customer retention / P. S. Fader, B. G. S. Hardie // Journal of Interactive Marketing. – 2007. – Vol. 21, № 1. – P. 76–90. – DOI 10.1002/dir.20074.

28. Fawcett, T. An introduction to ROC analysis / T. Fawcett // Pattern Recognition Letters. – 2006. – Vol. 27, № 8. – P. 861–874. – DOI 10.1016/j.patrec.2005.10.010.

29. Fisher, A. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously / A. Fisher, C. Rudin, F. Dominici // Journal of Machine Learning Research. – 2019. – Vol. 20, № 177. – P. 1–81.

30. Friedman, J. H. Greedy function approximation: a gradient boosting machine / J. H. Friedman // The Annals of Statistics. – 2001. – Vol. 29, № 5. – P. 1189–1232. – DOI 10.1214/aos/1013203451.

31. Gama, J. A survey on concept drift adaptation / J. Gama, I. Žliobaitė, A. Bifet [et al.] // ACM Computing Surveys. – 2014. – Vol. 46, № 4. – Art. 44. – DOI 10.1145/2523813.

32. Grinsztajn, L. Why do tree-based models still outperform deep learning on typical tabular data? / L. Grinsztajn, E. Oyallon, G. Varoquaux // Advances in Neural Information Processing Systems 35 : Datasets and Benchmarks Track. – 2022..

33. Gupta, S. Valuing customers / S. Gupta, D. R. Lehmann, J. A. Stuart // Journal of Marketing Research. – 2004. – Vol. 41, № 1. – P. 7–18. – DOI 10.1509/jmkr.41.1.7.25084.

34. Hadden, J. Computer assisted customer churn management: state-of-the-art and future trends / J. Hadden, A. Tiwari, R. Roy, D. Ruta // Computers & Operations Research. – 2007. – Vol. 34, № 10. – P. 2902–2917. – DOI 10.1016/j.cor.2005.11.007.

35. Hand, D. J. Measuring classifier performance: a coherent alternative to the area under the ROC curve / D. J. Hand // Machine Learning. – 2009. – Vol. 77, № 1. – P. 103–123. – DOI 10.1007/s10994-009-5119-5.

36. Hanley, J. A. The meaning and use of the area under a receiver operating characteristic (ROC) curve / J. A. Hanley, B. J. McNeil // Radiology. – 1982. – Vol. 143, № 1. – P. 29–36. – DOI 10.1148/radiology.143.1.7063747.

37. Harris, C. R. Array programming with NumPy / C. R. Harris, K. J. Millman, S. J. van der Walt [et al.] // Nature. – 2020. – Vol. 585. – P. 357–362. – DOI 10.1038/s41586-020-2649-2.

38. Hastie, T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction / T. Hastie, R. Tibshirani, J. Friedman. – 2nd ed. – New York : Springer, 2009. – 745 p. – ISBN 978-0-387-84857-0.

39. Hosmer, D. W. Applied Logistic Regression / D. W. Hosmer, S. Lemeshow, R. X. Sturdivant. – 3rd ed. – Hoboken : Wiley, 2013. – 528 p. – ISBN 978-0-470-58247-3.

40. Huang, B. Customer churn prediction in telecommunications / B. Huang, M. T. Kechadi, B. Buckley // Expert Systems with Applications. – 2012. – Vol. 39, № 1. – P. 1414–1425. – DOI 10.1016/j.eswa.2011.08.024.

41. Hunter, J. D. Matplotlib: a 2D graphics environment / J. D. Hunter // Computing in Science & Engineering. – 2007. – Vol. 9, № 3. – P. 90–95. – DOI 10.1109/MCSE.2007.55.

42. Idris, A. Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies / A. Idris, M. Rizwan, A. Khan // Computers & Electrical Engineering. – 2012. – Vol. 38, № 6. – P. 1808–1819. – DOI 10.1016/j.compeleceng.2012.09.001.

43. Ke, G. LightGBM: a highly efficient gradient boosting decision tree / G. Ke, Q. Meng, T. Finley [et al.] // Advances in Neural Information Processing Systems 30. – 2017..

44. Keramati, A. Improved churn prediction in telecommunication industry using data mining techniques / A. Keramati, R. Jafari-Marandi, M. Aliannejadi [et al.] // Applied Soft Computing. – 2014. – Vol. 24. – P. 994–1012. – DOI 10.1016/j.asoc.2014.08.041.

45. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection / R. Kohavi // Proceedings of the 14th International Joint Conference on Artificial Intelligence. – San Francisco : Morgan Kaufmann, 1995. – Vol. 2. – P. 1137–1143.

46. Kuhn, M. Applied Predictive Modeling / M. Kuhn, K. Johnson. – New York : Springer, 2013. – 600 p. – ISBN 978-1-4614-6848-6.

47. Kumar, V. Customer Relationship Management: Concept, Strategy, and Tools / V. Kumar, W. Reinartz. – 3rd ed. – Berlin : Springer, 2018. – 438 p. – ISBN 978-3-662-55380-0.

48. Lemmens, A. Bagging and boosting classification trees to predict churn / A. Lemmens, C. Croux // Journal of Marketing Research. – 2006. – Vol. 43, № 2. – P. 276–286. – DOI 10.1509/jmkr.43.2.276.

49. Lundberg, S. M. A unified approach to interpreting model predictions / S. M. Lundberg, S.-I. Lee // Advances in Neural Information Processing Systems 30. – 2017..

50. McKinney, W. Data structures for statistical computing in Python / W. McKinney // Proceedings of the 9th Python in Science Conference. – Austin, 2010. – P. 56–61. – DOI 10.25080/Majora-92bf1922-00a.

51. Neslin, S. A. Defection detection: measuring and understanding the predictive accuracy of customer churn models / S. A. Neslin, S. Gupta, W. Kamakura, J. Lu, C. H. Mason // Journal of Marketing Research. – 2006. – Vol. 43, № 2. – P. 204–211. – DOI 10.1509/jmkr.43.2.204.

52. Niculescu-Mizil, A. Predicting good probabilities with supervised learning / A. Niculescu-Mizil, R. Caruana // Proceedings of the 22nd International Conference on Machine Learning. – New York : ACM, 2005. – P. 625–632. – DOI 10.1145/1102351.1102430.

53. Óskarsdóttir, M. Social network analytics for churn prediction in telco: model building, evaluation and network architecture / M. Óskarsdóttir, C. Bravo, W. Verbeke [et al.] // Expert Systems with Applications. – 2017. – Vol. 85. – P. 204–220. – DOI 10.1016/j.eswa.2017.05.028.

54. Pedregosa, F. Scikit-learn: machine learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort [et al.] // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830.

55. Prokhorenkova, L. CatBoost: unbiased boosting with categorical features / L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin // Advances in Neural Information Processing Systems 31. – 2018..

56. Provost, F. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking / F. Provost, T. Fawcett. – Sebastopol : O'Reilly Media, 2013. – 414 p. – ISBN 978-1-449-36132-7.

57. Reichheld, F. F. Zero defections: quality comes to services / F. F. Reichheld, W. E. Sasser // Harvard Business Review. – 1990. – Vol. 68, № 5. – P. 105–111.

58. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead / C. Rudin // Nature Machine Intelligence. – 2019. – Vol. 1, № 5. – P. 206–215. – DOI 10.1038/s42256-019-0048-x.

59. Saito, T. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets / T. Saito, M. Rehmsmeier // PLoS ONE. – 2015. – Vol. 10, № 3. – Art. e0118432. – DOI 10.1371/journal.pone.0118432.

60. Sculley, D. Hidden technical debt in machine learning systems / D. Sculley, G. Holt, D. Golovin [et al.] // Advances in Neural Information Processing Systems 28. – 2015..

61. Shwartz-Ziv, R. Tabular data: deep learning is not all you need / R. Shwartz-Ziv, A. Armon // Information Fusion. – 2022. – Vol. 81. – P. 84–90. – DOI 10.1016/j.inffus.2021.11.011.

62. Strobl, C. Bias in random forest variable importance measures: illustrations, sources and a solution / C. Strobl, A.-L. Boulesteix, A. Zeileis, T. Hothorn // BMC Bioinformatics. – 2007. – Vol. 8. – Art. 25. – DOI 10.1186/1471-2105-8-25.

63. Vafeiadis, T. A comparison of machine learning techniques for customer churn prediction / T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, K. Ch. Chatzisavvas // Simulation Modelling Practice and Theory. – 2015. – Vol. 55. – P. 1–9. – DOI 10.1016/j.simpat.2015.03.003.

64. Verbeke, W. Building comprehensible customer churn prediction models with advanced rule induction techniques / W. Verbeke, D. Martens, C. Mues, B. Baesens // Expert Systems with Applications. – 2011. – Vol. 38, № 3. – P. 2354–2364. – DOI 10.1016/j.eswa.2010.08.023.

65. Verbeke, W. New insights into churn prediction in the telecommunication sector: a profit driven data mining approach / W. Verbeke, K. Dejaeger, D. Martens, J. Hur, B. Baesens // European Journal of Operational

Research. – 2012. – Vol. 218, № 1. – P. 211–229. – DOI 10.1016/j.ejor.2011.09.031.

66. Verbraken, T. A novel profit maximizing metric for measuring classification performance of customer churn prediction models / T. Verbraken, W. Verbeke, B. Baesens // IEEE Transactions on Knowledge and Data Engineering. – 2013. – Vol. 25, № 5. – P. 961–973. – DOI 10.1109/TKDE.2012.50.

67. Telco customer churn (11.1.3+) / IBM Samples Team. – Текст : электронный // IBM Community : [сайт]. – 2019. – 11 July. – URL: https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113 (дата обращения: 22.10.2026).

68. Telco Customer Churn on ICP4D : репозиторий кода и данных / IBM. – Текст : электронный // GitHub : [сайт]. – URL: https://github.com/IBM/telco-customer-churn-on-icp4d (дата обращения: 22.10.2026). – Файл data/Telco-Customer-Churn.csv.

ПРИЛОЖЕНИЕ А. ПРОГРАММНЫЙ КОД ПОДГОТОВКИ ДАННЫХ, ОБУЧЕНИЯ И ОЦЕНКИ МОДЕЛЕЙ

Ниже приведён программный код, которым получены результаты второй и третьей глав для логистической регрессии и градиентного бустинга. Код рассчитан на Python 3 с библиотеками NumPy и pandas и читает файл Telco-Customer-Churn.csv из репозитория IBM [68]. Реализация случайного леса и одиночного дерева отличается от приведённой только критерием разбиения (индекс Джини вместо формулы 16) и бутстрэп-выборкой для каждого дерева. Запуск листинга выводит ROC-AUC 0,8622 для логистической регрессии и 0,8646 для градиентного бустинга, а также точность, полноту и F1-меру при пороге 0,5, совпадающие с таблицами 8 и 9.

```
import numpy as np, pandas as pd

# --- 1. Загрузка и кодирование данных ---
d = pd.read_csv('Telco-Customer-Churn.csv')
d['TotalCharges'] = pd.to_numeric(d.TotalCharges,
                                  errors='coerce').fillna(0.0)
y = (d.Churn == 'Yes').astype(int).values
num = ['tenure', 'MonthlyCharges', 'TotalCharges']
cat = [c for c in d.columns
       if c not in num + ['customerID', 'Churn']]
X_df = pd.get_dummies(d[cat].astype(str),
                      drop_first=True).astype(float)
X_df = pd.concat([d[num].astype(float), X_df], axis=1)
X = X_df.values                             # 7043 x 30

# --- 2. Стратифицированное разбиение 80/20 ---
rng = np.random.default_rng(42)
i1 = np.where(y == 1)[0]; i0 = np.where(y == 0)[0]
rng.shuffle(i1); rng.shuffle(i0)
n1, n0 = round(len(i1) * 0.2), round(len(i0) * 0.2)
te = np.sort(np.r_[i1[:n1], i0[:n0]])
tr = np.sort(np.r_[i1[n1:], i0[n0:]])

# --- 3. Метрики ---
def auc(y, p):                              # формула (9)
    r = pd.Series(p).rank().values
    n1 = y.sum(); n0 = len(y) - n1
    return (r[y == 1].sum() - n1 * (n1 + 1) / 2) / (n1 * n0)

def prf(y, p, t):                           # формулы (7), (8)
    yh = (p >= t).astype(int)
    tp = ((yh == 1) & (y == 1)).sum()
    fp = ((yh == 1) & (y == 0)).sum()
    fn = ((yh == 0) & (y == 1)).sum()
    pr = tp / max(tp + fp, 1); rc = tp / max(tp + fn, 1)
    return pr, rc, 2 * pr * rc / max(pr + rc, 1e-9)
```

```
# --- 4. Логистическая регрессия: метод Ньютона (IRLS) + L2 ---
class LogReg:
    def __init__(s, lam=1.0, w1=1.0):
        s.lam, s.w1 = lam, w1
    def fit(s, X, y):
        s.mu, s.sd = X.mean(0), X.std(0) + 1e-9
        Z = np.c_[np.ones(len(X)), (X - s.mu) / s.sd]
        b = np.zeros(Z.shape[1]); sw = np.where(y == 1, s.w1, 1.0)
        R = np.eye(Z.shape[1]) * s.lam; R[0, 0] = 0
        for _ in range(50):
            p = 1 / (1 + np.exp(-Z @ b))
            g = Z.T @ (sw * (p - y)) + R @ b
            H = Z.T @ (Z * (sw * p * (1 - p))[:, None]) + R
            step = np.linalg.solve(H, g); b -= step
            if np.abs(step).max() < 1e-8: break
        s.b = b; return s
    def predict(s, X):
        Z = np.c_[np.ones(len(X)), (X - s.mu) / s.sd]
        return 1 / (1 + np.exp(-Z @ s.b))

# --- 5. Квантование признаков для деревьев (32 интервала) ---
def make_bins(X, nb=32):
    edges = []
    for j in range(X.shape[1]):
        u = np.unique(X[:, j])
        if len(u) <= nb: e = (u[:-1] + u[1:]) / 2
        else: e = np.unique(np.quantile(
            X[:, j], np.linspace(0, 1, nb + 1)[1:-1]))
        edges.append(e)
    return edges

def binize(X, edges):
    return np.stack([np.searchsorted(e, X[:, j], side='right')
                     for j, e in enumerate(edges)], 1)

# --- 6. Дерево регрессии на градиентах (лист по формуле 15) ---
class Tree:
    def __init__(s, depth, min_leaf, nbins, lam=1.0):
        s.depth, s.min_leaf, s.nb, s.lam = depth, min_leaf, nbins, lam
    def fit(s, B, g, h):
        s.nodes = []; s.gain = np.zeros(B.shape[1])
        s._grow(B, g, h, np.arange(len(g)), 0); return s
    def _score(s, G, H):
        return G ** 2 / (H + s.lam)
    def _grow(s, B, g, h, ix, dep):
        nid = len(s.nodes); s.nodes.append(None)
        G, H = g[ix].sum(), h[ix].sum()
        if dep >= s.depth or len(ix) < 2 * s.min_leaf:
            s.nodes[nid] = ('leaf', -G / (H + s.lam)); return nid
        parent = s._score(G, H); best = (0, None, None)
        for j in range(B.shape[1]):
            bj = B[ix, j]; nb = s.nb[j]
            Gc = np.cumsum(np.bincount(bj, g[ix], nb))[:-1]
            Hc = np.cumsum(np.bincount(bj, h[ix], nb))[:-1]
            Nc = np.cumsum(np.bincount(bj, minlength=nb))[:-1]
            ok = (Nc >= s.min_leaf) & (len(ix) - Nc >= s.min_leaf)
            if not ok.any(): continue
```

```
            gain = (s._score(Gc, Hc) + s._score(G - Gc, H - Hc)
                    - parent)
            gain[~ok] = -np.inf; k = int(np.argmax(gain))
            if gain[k] > best[0]: best = (gain[k], j, k)
        if best[1] is None:
            s.nodes[nid] = ('leaf', -G / (H + s.lam)); return nid
        gn, j, k = best; s.gain[j] += gn; m = B[ix, j] <= k
        l = s._grow(B, g, h, ix[m], dep + 1)
        r = s._grow(B, g, h, ix[~m], dep + 1)
        s.nodes[nid] = ('split', j, k, l, r); return nid
    def predict(s, B):
        out = np.empty(len(B)); stack = [(0, np.arange(len(B)))]
        while stack:
            nid, ix = stack.pop(); nd = s.nodes[nid]
            if nd[0] == 'leaf': out[ix] = nd[1]; continue
            _, j, k, l, r = nd; m = B[ix, j] <= k
            stack += [(l, ix[m]), (r, ix[~m])]
        return out

# --- 7. Градиентный бустинг с логистической функцией потерь ---
class GBM:
    def __init__(s, n=300, lr=0.03, depth=2, min_leaf=20, sub=0.8):
        s.n, s.lr, s.depth, s.min_leaf, s.sub = n, lr, depth, \
            min_leaf, sub
    def fit(s, X, y):
        s.edges = make_bins(X); B = binize(X, s.edges)
        nb = [len(e) + 1 for e in s.edges]
        r = np.random.default_rng(11); p0 = y.mean()
        s.f0 = np.log(p0 / (1 - p0)); F = np.full(len(y), s.f0)
        s.trees = []
        for _ in range(s.n):
            p = 1 / (1 + np.exp(-F)); g = p - y; h = p * (1 - p)
            bi = np.where(r.random(len(y)) < s.sub)[0]
            t = Tree(s.depth, s.min_leaf, nb).fit(B[bi], g[bi], h[bi])
            s.trees.append(t); F += s.lr * t.predict(B)
        return s
    def predict(s, X):
        B = binize(X, s.edges); F = np.full(len(X), s.f0)
        for t in s.trees: F += s.lr * t.predict(B)
        return 1 / (1 + np.exp(-F))

# --- 8. Обучение и оценка на отложенной выборке ---
for name, model in [('LR', LogReg(1.0)), ('GB', GBM())]:
    m = model.fit(X[tr], y[tr]); p = m.predict(X[te])
    print(name, 'ROC-AUC = %.4f' % auc(y[te], p),
          'P/R/F1 при 0,5 = %.3f / %.3f / %.3f' % prf(y[te], p, 0.5))
```