ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО»

Институт промышленного менеджмента, экономики и торговли

Высшая инженерно-экономическая школа

Отчет о прохождении учебной практики по научно-исследовательской работе

Ларченко Дмитрия Игоервича

1 курса магистратуры, гр. 3740105/50101

01.04.05 - Статистика

Место прохождения практики: ВИЭШ, ИПМЭиТ

Сроки практики: осенний семестр 2025/2026

Руководитель практики от ФГАОУ ВО «СПбПУ»:
Схведиани Анги Ерастиевич, доцент ВИЭШ, к.э.н.

Руководитель практики от профильной организации:
Не предусмотрен

Оценка:

Руководитель практики
от ФГАОУ ВО «СПбПУ»: Схведиани А.Е.

Руководитель практики
от профильной организации: Не предусмотрен

Обучающийся: Ларченко Д.И.

Дата:

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО»
Высшая инженерно-экономическая школа

ИНДИВИДУАЛЬНЫЙ ПЛАН (ЗАДАНИЕ И ГРАФИК) ПРОВЕДЕНИЯ ПРАКТИКИ

Ф.И.О. обучающегося Ларченко Дмитрия Игоревича

Направление подготовки (код/наименование): 01.04.05 - Статистика

Профиль (код/наименование): 01.04.05_01 «Моделирование и анализ больших данных в экономике»

Вид практики: Учебная

Тип практики: Практика по научно-исследовательской работе

Место прохождения практики: ВИЭШ ИПМЭиТ

Руководитель практики от ФГАОУ ВО «СПбПУ»:
Схведиани Анги Ерастиевич, доцент ВИЭШ, к.э.н.
(Ф.И.О., уч. степень, должность)

Руководитель практики от профильной организации:
Не предусмотрен
(Ф.И.О., должность)

Сроки практики: осенний семестр 2025/2026

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Для современных сервисов своевременное предсказание и предотвращение оттока пользователей является одной из ключевых задач, непосредственно влияющих на получаемую компанией прибыль. Даже небольшое увеличение удержания пользователей влечет за собой значительный рост прибыли как на коротком, так и на длительном промежутке времени. К примеру, в одном из исследований оттока пользователей в банковской сфере, было обнаружено, что всего 5% сокращение оттока влечет за собой увеличение прибыли на 35% и 95% в ИТ компании и рекламном агентстве соответственно [1].

Исследования удержания пользователей проводились в различных областях бизнеса. Подобные работы по анализу оттока пытались идентифицировать или предугадать заранее, что пользователь решит перестать пользоваться услугами сервиса, основываясь на различных показателях. Коэффициент оттока является типичным индикатором, отображающим удержание пользователей. Он показывает отношение пользователей, отменивших подписку за определенный период, к общему числу платных подписчиков. Большинство сервисов с подпиской используют коэффициент оттока для расчета своих показателей удержания и оттока клиентов в самых различных областях благодаря его значимости и интуитивности, адаптируя данный показатель под свою предметную область. В настоящий момент исследования удержания пользователей сильно разнятся для различных типов продукта, что затрудняет составление общей методологической базы и подбор конкретных способов анализа и прогноза данного показателя. Подобная разрозненность вызывает проблемы при анализе эффективности различных методов и способов обработки данных, так как многие исследования не пересекаются друг с другом, хотя и могут исследовать похожую природу явления удержания. В дополнение, так как исследования оттока и удержания прежде всего затрагивают области бизнес-администрирования и инженерии, для ученых из разных сфер научных интересов бывает сложно совместить две раздельных специализированных области в одной работе, либо понять их. [1]

Таким образом, анализ причин оттока пользователей и увеличение точности его прогнозирования являются ключевыми факторами для улучшения удержания клиентов на современном рынке и одной из актуальных задач для научного исследования, а недостаток работ и исследований, приложенным к сервисам музыкального стриминга делает настоящее исследование научно значимы.

1. ОПРЕДЕЛЕНИЕ ОТТОКА

Отток определяется в разных индустриях по-разному. Мы будем использовать определения, которые вывели авторы в работе [1], суммируя типичные статьи с различными критериями идентификации оттока. Общее определение оттока звучит как длительный период неактивности пользователя. Однако «длительность» или «неактивность» в каждой исследовательской сфере определяются по-своему. Подобное непостоянство обуславливается тем, что большая часть современных продуктов теряет клиентов из-за низких инвестиционных затрат клиента. Большинство сервисов в сфере интернет-услуг переходят на свободные условия подписки, благодаря чему стоимость перехода на другую услугу для покупателя становится очень низкой, из-за чего решение об уходе принимаются на основе личных пользовательских предпочтений и успеха персональных маркетинговых кампаний [2]. В то же время в других областях бывает, что клиенты заключают договор об оказании услуг, а потому уход для них сопровождается гораздо большим количеством усилий и денежных затрат из-за расторжения контракта [3]. Таким образом, в общем случае можно разделить отток пользователей на контрактный и неконтрактный. Описание каждого вида будет дано далее.

Первым рассмотрим контрактный уход пользователя из сервиса. Подобное происходит, когда клиент не продлевает договор, даже если наступил срок его продления. Такой уход означает, что пользователь целиком теряет интерес в подобного рода услугах и меняет свою позицию до состояния, где возвращение к прежнему состоянию маловероятно, а зачастую даже невозможно. Обычно такой отток характерен для клиентов банков или теле- и интернет-провайдеров, когда пользователь закрывает свой аккаунт или переходит к другому оператору, а также для сервисов с платной подпиской, таких как сервисы видео и музыкального стриминга.

Вторым вариантом является неконтрактный отток. В обычной неконтрактной ситуации, пользователь может покинуть продукт без привязки ко времени, поэтому, чтобы характеризовать потерянного пользователя, критерий, по которому будут судить о статусе клиента, выбирают заранее, после чего клиент, соответствующий подобному критерию, классифицируется как ушедший. Самым распространенным способом является метод «временного окна», когда выбирается определенный промежуток времени, длина которого зависит от специфики продукта и может варьироваться от нескольких дней, до нескольких месяцев, в течение которого

пользователь должен бездействовать, чтобы его посчитали потерянным. Метод «временного окна» часто используется для анализа логов активности в неконтрактных ситуациях. Когда пользователь не использует сервис определенное количество времени, этот метод определяет его как ушедшего [1].

Анализ оттока пользователей прежде всего используется для улучшения бизнес-результатов. Таким образом в большинстве задач его прогнозирования период оттока определятся как промежуток времени, за который еще можно вернуть доверие пользователя. Если в качестве «временного окна» выбирать период, за который пользователь уже решил уйти из продукта, то полезность такой статистики будет для бизнеса очень низка, так как считается, что клиента, который уже решил уйти, нельзя переубедить. Поэтому на текущий момент большинство задач предсказания оттока используют вероятностный метод для определения намерения пользователя уйти из сервиса, чтобы до момента принятия им окончательного решения, попытаться его вернуть.

1.1. Измерение оттока в различных областях бизнеса.

Большинство ранних исследований удержания клиентов производились с точки зрения менеджмента, в особенности управления взаимоотношения с клиентами (Customer relationship management – CRM). CRM охватывает все проблемы оттока, которые могли бы возникнуть в процессе идентификации, привлечения, удержания и развития клиентов. Современные же методы прогнозирования оттока в большинстве своем используют данные активности пользователей в приложении, после чего определяют, подходит ли пользователь под профиль потенциально уходящего.

1.1.1. Телеком

Рассмотрим подробнее каким образом исследуется отток в самых распространенных отраслях. Значительное количество работ посвящено удержанию пользователей в секторе телекоммуникаций. Это обусловлено тем, что, для телекома отток пользователей — ключевая метрика, отображающая их доход, так подавляющее количество покупателей пользуются их услугами в тарифном плане, продляемого ежемесячно, при этом подключение нового клиента обходится в несколько раз дороже удержания старого, а отток бьет напрямую по выручке и окупаемости. Ежегодно телеком-компания теряет от 20% до 40% своей клиентской базы, а финансовые потери для крупных операторов доходят до $20 млрд в год [4],

что объясняет, почему компании инвестируют в прогнозирование удержание, так как даже небольшое положительное изменение в динамике потерь клиентов окупает подобные вложения.

Телеком состоит из нескольких сегментов, каждый из которых имеет свою динамику оттока. Самая большая клиентская база у мобильной связи. В данной категории присутствуют оба типа оттока: контрактный и неконтрактный, — так как для разных тарифов по-разному происходит оплата услуг связи. В качестве примера контрактного оттока можно взять вариант, когда человек покупает тариф на 24 месяца, но отменяет его через какое-то время с выплатой штрафа; неконтрактный, в свою очередь, представляет собой длительное бездействие и отсутствие пополнения баланса. Также динамика удержания отличается для различных типов предоставляемых услуг из-за непрерывного развития технологий и эволюции рынка. Очень большой отток наблюдается у клиентов, пользующихся классическим телефоном или кабельным ТВ, так как клиенты переходят к стриминговым сервисам в случае ТВ, и предпочитают интернет-телефонию классическому телефону. От конкуренции на местном рынке сильно зависит динамика оттока у кабельного интернета и оптоволокна, так как зачастую для каждого потенциального покупателя существует множество вариантов подключения и клиент выбирает лучшее соответствие качества цене. Для провайдера выгодно продавать клиенту наборы из нескольких услуг — телефон, интернет и ТВ вместе, — таким образом клиенту дороже поменять провайдера, если он пользуется всеми услугами.

Таким образом, явное расторжение контракта или непродление тарифа считается контрактным оттоком. Клиент либо лично отказывается от услуги, либо перестает оплачивать, а через 90 дней система блокирует счет. Такой промежуток обусловлен тем, что через 90 дней компания считает клиента ушедшим, но до этого времени клиента еще можно восстановить [4], [5]. Неконтрактный отток считается при отсутствии пополнения 60-90 дней и/или отсутствия использования услуг (звонков, SMS, интернета). Подобным образом считать потерянных пользователей труднее, потому что даже после выделенного промежутка времени клиент может вернуться, поэтому для каждой компании очень важно правильно выбрать «временное окно», чтобы с высокой вероятностью определять потенциально уходящих клиентов, но еще сомневающихся, чтобы можно было их вернуть. [1]

1.1.1.1. Используемые данные и признаки

Основным источником данных в телекоме являются CDR (Call data reports) — журнал данных о звонках, SMS и интернет-трафике. В своей работе авторы [4] показали, что паттерны использования имеют высокое корреляцию с будущим оттоком. Если у клиента внезапно меняется в худшую сторону то, как он пользуется продуктом, то это может быть сильным сигналом предполагаемого ухода из сервиса. Еще одним немаловажным признаком является история платежей. Клиент, который платит регулярно, менее вероятно захочет уйти из сервиса при прочих равных, при этом клиент с 2-мя и более просроченным платежами имеет шансы уйти из сервиса в 5 раз выше [4]. Третьим по счету является качество предоставляемых услуг. Если клиент не удовлетворен получаемым качеством, то с большой вероятностью он захочет поменять сервис даже если платит регулярно и стабильно пользуется услугами сервиса. Немаловажным является взаимодействие пользователя с поддержкой и удовлетворенность этим. Резкий скачок количества обращений в поддержку может быть тревожным сигналом.

1.1.2. Банкинг и финансы

Банки имеют гораздо ниже ставку оттока, чем телеком, по нескольким причинам. Прежде всего для клиента сложно и дорого поменять банк по своему желанию. Необходимо переносить платежи, перезаключать договоры, закрывать все счета и выводить средства, а также выполнять все регуляторные требования. Из-за подобного количества сложностей в данной сфере достаточно низкая инертность, потому что во многих случаях клиентам легче смириться с неудобством, чем проходить все процедуры для перехода к пользованию услугами другого банка. У различных типов счетов варьируется отток: для текущих счетов показатель составляет 8-15% в год, от кредитных карт отказываются 15-25% человек в год за счет высокой конкуренции, похожий показатель оттока показывают клиенты сберегательных счетов, потому что в данной категории очень много неактивных пользователей и низкая стоимость перехода в другой банк. Также отличается лояльность у клиентов в зависимости от длительности пользования конкретным банком. Например, банки Великобритании показали, что только 28% были лояльны 1-5лет, а 67% - 6 и более лет, что показывает, что очень сильно влияет фактор доверия, которое строится за длительное время между банком и его клиентом. [5].

Банкинг можно разделить на несколько сегментов, в каждом из которых выделить свои закономерности оттока. Как уже было упомянуто, в банкинге для физических лиц выделяют текущие счета с достаточно низким оттоком, сберегательные счета и кредитные карты с высоким процентом потерь клиентов из-за высокой конкуренции и простоты смены банка (в случае со сберегательным счетом). Очень активно развивается отрасль интернет-банкинга. Мобильный банкинг и предоставление услуг через интернет очень сильно влияют на удержание и повышают качество обслуживания. Также высокая стоимость смены банка в B2B направлении, если фирма целиком пользуется услугами какого-то банка и управляет своими финансами через него, хотя для малых компаний показатели оттока могут быть выше из-за высокой конкуренции и малых затратах на смену на начальных этапах.

Для банка существует несколько типов оттока. Первым учитывают полный отказ клиента от услуг банка, то есть закрытие счетов клиентом лично или прекращение обслуживания со стороны банка из-за бездействия, — оба этих случая включаются в полный отток. Большое количество банков в настоящий момент предоставляют своим клиентам возможность улучшить качество получаемых услуг, либо добавить какие-то новые при помощи платной подписки. Если клиент уже пользуется преимуществами платной подписки, то отказ от нее хоть и не является фактом расторжения договора, но может показывать о планах об уходе и являться ранним сигналом полного ухода [6]. Частичный уход представляет из себя закрытие клиентом одного или нескольких своих счетов, сокращение использования услуг банка или перевод денег в другой банк. Также часто выделяют непродление контракта для продуктов с явным периодом (кредит, инвестиции), что относится к контрактному оттоку.

1.1.2.1. Используемые данные и признаки

Банки оценивают каждого клиента со стороны его надежности и уверенности в наличии у него денежных средств, поэтому зачастую демографические показатели всегда учитываются при составлении портрета пользователя. Например, 20-летний студент с достаточно малой вероятностью будет располагать большими денежными средствами и будет пользоваться базовой версией аккаунта, когда как 45-летний предприниматель, в свою очередь, скорее всего будет активно пользоваться услугами банка и премиум функциями. Соответственно, у молодежи будет гораздо более высокий процент оттока, чем у старших поколений из-за высокой

мобильности и отсутствия больших сбережений. Вторым признаком можно считать транзакционное поведение и состояние банковского аккаунта. Частые переводы, высокий баланс и стабильные платежи могут сказать о том, что клиент активно пользуется банком и не испытывает с этим критических проблем, в то время как низкая частота переводов, резкое опустошение баланса и отсутствие регулярных пополнений свидетельствует о том, что скорее всего клиент может легко уйти. Ключевым фактором в портрете пользователя является платежное поведение. Значительная часть банковских клиентов берет в этом банке кредит или кредитную карту, поэтому характер платежей может многое сказать о том, готов ли клиент и дальше пользоваться подобными услугами. Платежное поведение является одним из самых важных признаков в предсказании оттока. Просроченный на 30 дней платеж в 10 раз увеличивает риск ухода, а клиент с 2-мя и более просроченными платежами обычно уходит в следующие 3 месяца [4]. Также хорошим показателем лояльности клиента может быть количество продуктов, которые имеет клиент. Люди, которые пользуются только текущим счетом более мобильны из-за низкой стоимости смены сервиса. При этом клиенты с несколькими продуктами, как правило, остаются со своим банком на длительное время, так как им гораздо сложнее оформлять все заново в новом банке. Немаловажным критерием является качества сервиса и пользовательский опыт. Выделяют несколько показателей, описывающих удовлетворение пользователя опытом пользования банком: качество сервиса, качество функционала, соответствие цена-качество, доверие, интуитивность интерфейса и осознаваемый риск потери денег [7].

1.1.3. Потоковые сервисы и медиа

Потоковые сервисы имеют очень высокие показатели оттока по сравнению с другими областями. Прежде всего это связано с тем, что у большей части из них отсутствуют контракты и существует только месячная подписка, а в некоторых есть еще бесплатная версия, монетизируемая за счет рекламы. Подписка легко отменяется одной кнопкой, а у каждого пользователя есть множество альтернатив, поэтому при относительно малых изменениях клиенты готовы сразу поменять один сервис на другой. Структуру индустрии можно поделить на два больших сектора: музыкальные платформы и видео-платформы. Большинство из низ работает исключительно по подписке, но в некоторых есть бесплатный план, включающий в себя базовый функционал (такая модель с платной подпиской называется freemium). В стриминговых сервисах можно выделить 3 типа оттока: полный отказ,

понижение (применимо к freemium, — переход из платной подписки на бесплатный план), неактивность. Полный отказ и понижение определяются однозначно, в случае отказа пользователь сам отменяет подписку, либо она отменяется из-за проблем с оплатой, понижение происходит так же, но вместо того, чтобы прекратить пользоваться сервисом окончательно пользователь остается на бесплатном уровне. Неактивным пользователь считается, когда он не заходит в приложение определенное количество времени, но подписка все еще активна — клиент продолжает оплачивать подписку, но уже не пользуется приложением и может прекратить платить в любой момент, —что создает уникальную проблему — клиент платит, но не пользуется.

1.1.3.1. Freemium отток в стриминговых сервисах

Особая ситуация возникает в freemium-сервисах. В данной категории можно добавить отток еще из бесплатной версии. Хоть пользователи и не приносят доход напрямую, не платя за подписку, каждый из бесплатных пользователей может перейти на премиум режим, поэтому отток из базового плана все еще уменьшает потенциальное количество пользователей платной подписки и, соответственно, возможную прибыль. Из-за чего возникает парадокс: как много бесплатного контента может предоставить сервис, чтобы не потерять бесплатного пользователя, но мотивировать его улучшить текущий план? Есть два варианта решений данной ситуации в различных компаниях. Первый — сильно ограничить бесплатный функционал, чтобы пользователь мог познакомиться с сервисом, но не смог в дальнейшем полноценно им пользователя (пробный период, либо функциональные ограничения). Подобный подход сильно мотивирует бесплатных пользователей перейти на премиум, но в то же время у большой части клиентов вызывает негативные эмоции, желание прекратить пользоваться сервисом и сделать выбор в пользу конкурента. Второе решение — предоставить практически полноценный доступ к продукту в бесплатной версии, а в премиум подписку включить улучшения качества предоставляемых услуг, разнообразить контент, отключить рекламу и т.д. Данный способ очень сильно выигрывает в ситуации с потерей бесплатных пользователей у первого, но у клиентов меньше мотивации платить за улучшение плана. Подобные два подхода уже сравнивались в работе [8], но в отношении freemium мобильных игр, а не стриминговых сервисов, так что подробное изучение влияния бесплатного опыта пользователя на отток и конверсию требует дальнейшего изучения.

1.1.3.2. Используемые данные и признаки

Для видео- и музыкальных платформ разнятся конкретные характеристики пользователей, но все равно можно выделить общие черты, на которые нужно обращать внимание при категоризации клиентов. Прежде всего необходимо оценивать частоту и регулярность пользования сервисом: сколько времени проведено, прослушивая музыку/просматривая видео, как часто открывается приложение и какая средняя продолжительность сессии. Персональные предпочтения и открытие нового контента: если пользователь начинает потреблять один и тот же контент длительное время это может обозначать, что его рекомендательная система может работать некорректно, что тоже влияет на вероятность ухода из сервиса. Паттерны взаимодействия: как часто пропускаются треки/эпизоды, частота отметок положительных реакций, поисковые запросы и использование социальных функций. Важно оценивать, насколько часто и эффективно используются функции платной подписки, чтобы пользователь понимал, за что он платит и был этим доволен.

2. ТЕХНОЛОГИИ И МЕТОДЫ АНАЛИЗА И ПРОГНОЗИРОВАНИЯ УДЕРЖАНИЯ ПОЛЬЗОВАТЕЛЕЙ В СЕРВИСАХ МУЗЫКАЛЬНОГО И ВИДЕО СТРИМИНГА

После анализа существующей литературы по теме предсказания и анализа оттока пользователей было выявлено множество моделей и методов прогнозирования оттока, эффективных в различных ситуациях, отраслях бизнеса, наборах данных, которые можно разделить на несколько групп: глубокое обучение, машинное обучение, сравнительный анализ и гибридные методы.

2.1. Логистическая регрессия

Стоит начать с классических методов машинного обучения. Первой рассмотрим логистическую регрессию — метод классификации, который предсказывает принадлежность примера к определенному классу. Для задачи прогноза оттока пользователей логистическая регрессия служит двум важным целям: используется как базовая модель для сравнения с более сложными методами и позволяет получить легко интерпретируемые результаты, которые легко объяснить представителям бизнеса. Авторы [6] применили этот метод в задаче прогноза оттока в телеком компании. На датасете с историческими данными в результате получили accuracy 62% и F-measure 0.585 и выделили важные признаки для последующего их использования. У логистической регрессии есть ряд преимуществ для включения ее в исследование: она очень проста в реализации, почти не требует вычислительных ресурсов, быстро обучается на небольших датасетов, дает возможность получать вероятностные прогнозы и очень легко интерпретируется. Однако, несмотря на все положительные качества, есть ряд ограничений, сильно усложняющих эффективное использование данного метода. Прежде всего логистическая регрессия предполагает линейную зависимость, чего часто не происходит в реальности. Также метод показывает низкую точность на сложных данных по сравнению с другими, разница доходит до 40%, что критически важно для доверия результатам прогнозов.

Тем не менее логистическая регрессия остается ценным методом для прогноза оттока, подходящая для того, чтобы использовать ее в качестве отправной точки, первичного анализа и выделения ключевых факторов.

2.2. Деревья решений

Один из самых интуитивно понятных и интерпретируемых методов машинного обучения. Метод моделирует процесс принятия решений путем построения иерархической структуры, где каждый узел представляет собой проверку какого-то параметра, ветви — результат проверки, а листья — предсказания класса. Для задачи прогноза оттока пользователей деревья сами по себе оказывают среднюю производительность и результаты: 85-90% accuracy на данных телекома. У деревьев есть ряд плюсов, выгодно выделяющих их среди других методов. Прежде всего деревья очень быстро обучаются и делают предсказания даже на больших датасетах, также деревья не требуют нормализации признаков и работают напрямую с категориальными переменными. В дополнение деревья хорошо определяют важность признаков на основе того, насколько близко они находятся к корню. Однако, деревья очень подвержены риску переобучения и достаточно нестабильны — даже маленькие изменения в данных делают модель неустойчивой. Также деревья смещаются в сторону большинства класса и не могут эффективно захватывать линейные зависимости. Подводя итог, деревья решений можно использовать для того, чтобы определить важные признаки, определяющие пользователей, подверженных риску.

2.3. Случайный лес

Случайный лес один из самых популярных и мощных методов машинного обучения. Он строит ансамбль независимых деревьев решений, каждое из которых обучается на случайной подвыборке обучающих данных. Для задачи прогноза оттока пользователей метод показывает достаточно высокие результаты accuracy и f-measure – 94% и 0.945 соответственно на телекоммуникационных данных [9]. Таким образом, одним из явных преимуществ случайного леса является его высокая точность. Благодаря ансамблированию метод менее подвержен переобучению, чем отдельное дерево. Случайный лес хорошо захватывает нелинейные взаимодействия между признаками, в отличие от логистической регрессии, что уменьшает необходимость ручного поиска подобных зависимостей. Также метод хорошо справляется с выбросами за счет интервалов признаков и экстремальные значения не оказывают большого влияния на модель. Однако, есть и некоторые ограничения в его работе, которые стоит учитывать. Случайный лес требует гораздо больше памяти за счет того, что нужно хранить несколько полных деревьев, где каждое дерево может хранить тысячи узлов. Без правильной предобработки метод плохо

работает с несбалансированными данными. В случае оттока пользователей, где данные скорее всего будут подвержены дисбалансу классов, случайный лес может смещаться в сторону предсказания большинства класса. Также модели может не хватить информации на очень малых датасетах (<1000 строк), и она может переобучиться. Таким образом случайный лес является одним из наиболее практичных и эффективных методов для прогноза оттока пользователей. Случайный лес может быть хорошей первой моделью для прогноза оттока благодаря простоте реализации и интерпретируемости.

2.4. Метод опорных векторов (SVM – support vector machines)

SVM — это линейный алгоритм используемый в задачах классификации и регрессии. Данный алгоритм имеет широкое применение на практике и может решать как линейные, так и нелинейные задачи. Суть работы “Машин” Опорных Векторов проста: алгоритм создает линию или гиперплоскость, которая разделяет данные на классы. Основной задачей алгоритма является найти наиболее правильную линию, или гиперплоскость, разделяющую данные на два класса. SVM это алгоритм, который получает на входе данные, и возвращает такую разделяющую линию.

В задачах классификации пользователей как уходящих, метод показывает следующим образом: для датасета с данными телекома accuracy 92.44%, f-measure 88.8%, при стандартных параметрах и достигает 98% точности и f-measure 97.5% при правильной настройке гиперпараметров [10]. SVM отлично работает с большим количеством признаков и достаточно надежен, чтобы быть в нем уверенным. Тем не менее, метод требует значительного количества времени для обучения на больших датасетах, а таже требует критической нормализации признаков, так как очень чувствителен к их масштабу и выбросам. Еще одним минусом является то, что на датасетах музыкального стриминга он не тестировался в рассматриваемых исследованиях, а потому нельзя с уверенностью говорить об его эффективности в новой области.

2.5. XGBoost (Extreme gradient boosting)

XGBoost является одним из самых популярных и эффективных ансамбльных методов для прогноза оттока пользователей благодаря высокой точности и скорости обучения. Метод отлично справляется с задачами классификации и регрессии, широко используется за счет своей скорости и точности на табличных наборах

данных. Базовой моделью экстремального градиентного бустинга является дерево. Техника строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений. Обучение ансамбля происходит последовательно. На каждой итерации вычисляются отклонения предсказаний уже обученного ансамбля на обучающей выборке. Для задачи прогноза оттока метод показывает результаты сравнимые с SVM и превосходящие случайный лес, а также показывает устойчивость к переобучению.

2.6. LightGBM (Light gradient boosting machine)

Метод представляет собой быструю, распределенную платформу градиентного бустинга для классификации, ранжирования и регрессии. По сравнению с экстремальным бустом, LightGBM быстрее обучается, требует меньше памяти и имеет близкую точность, что ставит их на один уровень во многих ситуациях. Вместо того, чтобы расти слой за слоем, LightGBM выбирает лист, который максимизирует уменьшение потерь, и растит его дальше, благодаря чему деревья точнее приспосабливаются к данным. Как уже говорилось, скорость обучения LightGBM в 10-20 раз быстрее XGBoost на больших датасетах, а также метод использует в 2-3 раза меньше памяти благодаря листовому росту. Метод хорошо работает для больших структурированных датасетов, категориальных данных, на результаты предсказаний слабо влияет дисбаланс классов, что важно для прогноза оттока. На данных телекома со стандартными настройками метод показывает 92.71% accuracy, 89.20% F-measure [10].

2.7. Гибридные модели

2.7.1. LSTM + GRU + LightGBM

Далее рассмотрим две самых эффективных модели, которые были использованы при прогнозировании оттока пользователей. Первую использовали в своем исследовании [11]: LSTM + GRU + LightGBM — это передовая гибридная архитектура, которая комбинирует рекуррентные нейросети для захвата временных зависимостей в поведении пользователей с градиентным бустингом (LightGBM) для финального предсказания на основе комбинированных признаков. Модель показала AUC 0.9258 и F1-Score 0.85 на данных потокового сервиса, что превышает результаты классических методов и отдельных нейросетевых структур.

Модель является трехуровневой архитектурой: LSTM (Long Short-Term Memory) – нейросеть обучается захватывать долгосрочные временные зависимости и предсказывает долгосрочные тренды в поведении. На втором уровне находится GRU (Gated Recurrent Unit) – обучается захватывать краткосрочные паттерны в активности пользователя, быстро реагирует на изменения в поведении и фокусируется на недавних действиях. Комбинирует обе нейросети уже описанный выше LightGBM и дополняет модель исходными табличными признаки, делая финальное предсказание. Благодаря такому подходу модель одновременно обучается на нескольких временных масштабах и лучше захватывает сложные зависимости, эффективно комбинируя нейросети и ансамбли. Нейросетевые составляющие дают иерархичное представление данных и знания о структуре временных рядов, в то время как градиентный бустинг добавляет обработку категориальных признаков, хорошую интерпретируемость и значительный рост эффективности для потоковых данных за счет оптимизации для непрерывно поступающих данных. Благодаря этому модель может обновляться без полного переобучения. К сожалению, подобная модель будет плохо подходить к статическим данным без истории или малым датасетам, при этом требуя значительной предобработки временных рядов.

2.7.2. CCP-Net

Customer churn prediction network была разработана для прогноза оттока клиентов в работе [12]. Метод комбинирует несколько нейросетей для эффективного захвата как глобальных, так и локальных признаков поведения клиентов. Модель была протестирована на датасете данных из телекома и показала Accuracy 91.17%, что в этой работе превышало результаты как отдельных методов машинного обучения, так и других гибридных архитектур.

CCP-Net – это трехмодульная архитектура, которая последовательно обрабатывает данные через три типа нейронных сетей. На первом уровне глобальные зависимости захватывает Multi-Head Self-Attention и позволяет модели выявлять долгосрочные паттерны в поведении. Данная нейросеть избегает проблемы, свойственной для LSTM/GRU — исчезания градиента, а также обрабатывает информацию параллельно. Второй уровень используется нейросетью BiLSTM(Bi-derectional LTSM), которая тоже захватывает долгосрочные зависимости с двух направлений, благодаря чему получает двустороннюю информацию и лучше предсказывает будущие изменения в поведении, обеспечивая

контекст как из прошлого, так и из «будущего». Локальные признаки и краткосрочные признаки захватывает CNN (Convolutional Neural Network), используя различные размеры фильтров для захвата локальных особенностей.

Таким образом, CCP-Net является мощным инструментом для предсказания оттока пользователей, но отсутствие результатов по стриминговым сервисам оставляет необходимость дополнительных исследований ее эффективности в этой отрасли.

Рассмотрев большую часть основных моделей, используемых в ансамблях, можно оценить принципы работы каждого из них, исходя из составляющих частей. Далее приведем сводные таблицы методов, применяемых в литературе для прогнозирования оттока пользователей в различных областях науки и бизнеса.

2.8. Варианты применения для музыкального стриминга

Музыкальные стриминг-платформы существенно отличаются от телекома и банков по характеру данных о клиентах, что требует специфичной адаптации моделей машинного обучения. В отличие от табличных признаков оттока, у телеком индустрии, стриминг-данные образуют временные ряды, требующий захвата последовательных паттернов поведения. В этом случае гибридная модель LSTM+GRU+LightGBM особенно эффективна: LSTM и GRU слои улавливают долгосрочные тренды использования, а LightGBM дополняет это статичными признаками вроде «дней с регистрации», «среднего количества сессий в месяц», «процента лайков» На 12.5 GB датасете, содержащем 22 277 уникальных пользователей и 26 млн событий, эта гибридная архитектура достигает AUC 95.60% и F1-score 90.09%, что значительно превосходит чистый LightGBM или XGBoost на одних лишь табличных признаках [11]. Балансировка класса остаётся критичной, поскольку потенциальные уходящие пользователя составляют лишь ~22% датасета. RF-AdaBoost и XGBoost могут остаться резервными опциями: они обучаются на 2–5 минут на CPU и имеют сравнимую точность, демонстрируя хорошие результаты.

ЗАКЛЮЧЕНИЕ

В ходе работы было сравнено и проанализировано множество различных алгоритмов и архитектур, применённых к задачам предсказания оттока пользователей в телекоммуникационной, банковской, потоковых сервисах и электронной коммерции.

Анализ показал чётко выраженную иерархию методов по производительности. По достигаемому качеству (метрика F1-score) лидируют гибридные архитектуры: CCP-Net (Multi-Head Self-Attention + BiLSTM + CNN) с F1 = 91–92%, комбинация TDA + XGBoost (F1 = 98.5%) и гибридная модель LSTM+GRU+LightGBM (F1 = 85.4%, AUC = 92.58%). Традиционные методы градиентного бустинга (XGBoost, LightGBM, CatBoost) обеспечивают F1 = 85–96% при заметно меньших вычислительных затратах, тогда как базовые классификаторы (логистическая регрессия, KNN, деревья решений) показывают приемлемые результаты только на специально подготовленных датасетах.

Важным выводом является то, что выбор метода должен определяться еще и структурой данных, а не только критериями точности. На табличных данных с фиксированным набором признаков (телеком, банк) методы глубокого обучения (BiLSTM-CNN, LSTM) не превосходят XGBoost/LightGBM по качеству (различие <5% по F1), но требуют порядок больше вычислительных ресурсов и времени обучения. Напротив, на данных с реальными временными рядами и последовательностями событий (потоковые сервисы, мобильные приложения) архитектуры, захватывающие временные зависимости (LSTM, BiLSTM), демонстрируют заметное преимущество по F1-score.

Исследование подтвердило, что качество конструирования признаков и подготовки данных оказывает большое влияние на итоговую производительность. Применение топологического анализа данных (TDA) позволило достичь F1 = 98.5% на телеком-датасете, однако потребовало 1–2 часов дополнительной подготовки данных [10]. Гибридная архитектура LSTM+GRU+LightGBM достигла AUC = 95.60% на 12.5 GB датасете потокового сервиса благодаря комбинированию двух типов моделей: глубокие слои для захвата последовательных паттернов и традиционные методы для работы со статичными признаками.

В общем и целом, исследование демонстрирует, что задача прогнозирования оттока клиентов имеет множество адекватных решений разной сложности и

эффективности. Выбор конкретного метода определяется сочетанием факторов: структура данных, наличие временных компонент, требуемые метрики качества и вычислительные ограничения. Результаты работы подтверждают актуальность развития гибридных архитектур, комбинирующих сильные стороны классических машинных методов обучения и глубоких нейросетей, особенно для сценариев с разнородными и высокомерными данными.

Далее будет приведены две сводные таблицы с количественными и качественными характеристиками применяемых методов в рассматриваемой литературе

Таблица 1 — Сводная таблицы численных характеристик применяемых методов

Источник	Метод	Accuracy	Auc	recall	precision	f-score	Отрасль
Rivaldo, Rahman Taufik, Igit Sabda Ilman, Ossy Dwi Endah Wulansari	XGBoost / LightGBM / CatBoost	98.3	-	-	-	-	Banking
Yixin Li, Bingzhang Hou, Yue Wu, Donglai Zhao, Aoran Xie, Peng Zou	Logistic Regression	62	-	-	-	-	Broadcasting/TV
Soumaya Lamrhari, Hamid El Ghazi, Mourad Oubrich, Abdellatif El Faker	LDA + Sentiment Analysis + Fuzzy-Kano + Random Forest + K-means	98	-	-	-	-	E-commerce
Xinyu Liu, Guoen Xia, Xianquan Zhang, Wenbin Ma, Chunqiang Yu	Multi-Head Self-Attention + BiLSTM + CNN + ADASYN	92.19	-	-	92.19	-	Multi-industry
Chenggang He, Chris H. Q. Ding	Ensemble-Fusion (17 ML algorithms)	95.35	91	96.96	96.01	96.96	SaaS
Marcel Sagming, Reolyn Heymann, Maria Vivien Visaya	TDA Barcode Statistics + XGBoost	98.5	-	98.5	98.5	98.5	Telecom
Mohammed Affan Shaikhsurab, Pramod Magadum	Adaptive Stacking Ensemble (XGBoost, LightGBM, LSTM, MLP, SVM)	99.28	-	-	-	-	Telecom
Marcel Sagming, Reolyn Heymann, Maria Vivien Visaya	TDA Barcode Statistics + XGBoost	98.5	-	98.5	98.5	98.5	Telecom
Kaveh Faraji Googerdchi, Shahrokh Asadi, Seyed Mohammadbagher Jafari	MOEEC-1 & MOEEC-2 (NSGA-II + Clustering)	97.3	93.76	88.66	93.48	91	Telecom

Продолжение таблицы 1

You-wu Liu, Jing Wang, Chibiao Liu	Improved ELM + Autoencoder + Gaussian Kernel	92.7	-	-		83.5	-	Telecom
Fatima Enehezei Usman-Hamza, Abdullateef Oluwagbemiga Balogun, et al.	Decision Forest (LMT, RF, FT) + Weighted Voting/Stacking	90	-	-	-		-	Telecom
Tjeng Wawan Cenggoro, Raditya Ayu Wirastari, Edy Rudiantoa, Mochamad Ilham Mohadi, Dinne Ratj, Bens Pardamean	Deep Learning + Vector Embedding + Weighted Softmax Loss	89.82	-	-	-		81.16	Telecom
Asad Khattak, Zartashia Mehak, Hussain Ahmad, Muhammad Usama Asghar, Muhammad Zubair Asghar, Aurangzeb Khan	BiLSTM + CNN + Hybrid Deep Learning	81	-	-	-		-	Telecom
Yancong Zhou, Wenyue Chen, Xiaochen Sun, Dandan Yang	Random Forest + AdaBoost Dual-Ensemble + BPNN	93	-		93	99	96	Telecom

Таблица 2 — сводная таблица качественных характеристик методов

Метод	**Требования к данным (объем, шумы, баланс, разнообразие)**	**Скорость обучения**	**Обработка дисбаланса**	**Работа с временными рядами**	**Конструирование признаков**
Логистическая регрессия (LR)	Малые	<1 сек	SMOTE	Нет	Минимум
Случайный лес (RF)	Малые	1-5 сек	Балансировка	Слабо	Среднее
Дерево решений (DT)	Малые	<1 сек	Чувствительно	Слабо	Среднее
K-ближайших соседей (KNN)	Средние	1-3 сек	SMOTE	Нет	Низкое
SVM / SVC	Средние	5-20 сек	SMOTE	Нет	Высокое

Продолжение таблицы 2

XGBoost	Средние	5-30 сек	scale_pos_weight	Слабо	Низкое
LightGBM	Средние	2-10 сек	is_unbalanced	Слабо	Низкое
CatBoost	Средние	30-120 сек	auto_class_weights	Слабо	Категории
AdaBoost	Малые	2-5 сек	Встроен	Слабо	Среднее
Gradient Boosting	Средние	5-15 сек	SMOTE	Слабо	Среднее
LSTM	Большие	1-4 часа	SMOTE	Отлично	Высокое
GRU	Большие	1-3 часа	SMOTE	Хорошо	Высокое
RNN	Большие	2-5 часов	SMOTE	Хорошо	Высокое
Многослойный персептрон (MLP)	Большие	1-3 часа	SMOTE	Слабо	Высокое
BiLSTM-CNN	Очень большие	2-4 часа	SMOTE	Отлично	Очень высокое
LSTM+GRU+LightGBM (Гибрид)	Большие	2-3 часа	SMOTE+вес	Отлично	Высокое
Наивный Байес (NB)	Малые	<1 сек	SMOTE	Нет	Минимум
SVM-RBF	Средние	5-20 сек	SMOTE	Нет	Высокое
CCP-Net (Multi-Head)	Очень большие	4-8 часов	ADASYN	Отлично	Очень высокое
TDA + XGBoost + Barcode	Средние	30-120 сек	XGBoost	Слабо	Высокое (TDA)

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Ahn, Jaehyun, Junsik Hwang, Doyoung Kim, Hyukgeun Choi, and Shinjin Kang. 'A Survey on Churn Analysis in Various Business Domains'. *IEEE Access* 8 (2020): 220816–39. https://doi.org/10.1109/ACCESS.2020.3042657.

[2] Lamrhari, Soumaya, Hamid El Ghazi, Mourad Oubrich, and Abdellatif El Faker. 'A Social CRM Analytic Framework for Improving Customer Retention, Acquisition, and Conversion'. *Technological Forecasting and Social Change* 174 (January 2022): 121275. https://doi.org/10.1016/j.techfore.2021.121275.

[3] Prasad, U Devi, and S Madhavi. *Prediction Of Churn Behaviour Of Bank Customers Using Data Mining Tools*. n.d.

[4] Mozer, M.C., R. Wolniewicz, D.B. Grimes, E. Johnson, and H. Kaushansky. 'Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry'. *IEEE Transactions on Neural Networks* 11, no. 3 (2000): 690–96. https://doi.org/10.1109/72.846740.

[5] Usman-Hamza, Fatima Enehezei, Abdullateef Oluwagbemiga Balogun, Luiz Fernando Capretz, et al. 'Intelligent Decision Forest Models for Customer Churn Prediction'. *Applied Sciences* 12, no. 16 (2022): 8270. https://doi.org/10.3390/app12168270.

[6] Li, Yixin, Bingzhang Hou, Yue Wu, Donglai Zhao, Aoran Xie, and Peng Zou. 'Giant Fight: Customer Churn Prediction in Traditional Broadcast Industry'. *Journal of Business Research* 131 (July 2021): 630–39. https://doi.org/10.1016/j.jbusres.2021.01.022.

[7] Mbama, Cajetan & Ezepue, Patrick & Alboul, Lyuba & Beer, Martin. (2018). Digital banking, customer experience and financial performance: UK bank managers' perceptions. Journal of Research in Interactive Marketing. 12. 432-451. 10.1108/JRIM-01-2018-0026.

[8] Ascarza, Eva, Oded Netzer, and Julian Runge. 'The Twofold Effect of Customer Retention in Freemium Settings'. *SSRN Electronic Journal*, ahead of print, 2020. https://doi.org/10.2139/ssrn.3725224.

[9] He, Chenggang, and Chris H. Q. Ding. ‘A Novel Classification Algorithm for Customer Churn Prediction Based on Hybrid Ensemble-Fusion Model’. *Scientific Reports* 14, no. 1 (2024): 20179. https://doi.org/10.1038/s41598-024-71168-x.

[10] Sagming, Marcel, Reolyn Heymann, and Maria Vivien Visaya. ‘Using Topological Data Analysis and Machine Learning to Predict Customer Churn’. *Journal of Big Data* 11, no. 1 (2024): 160. https://doi.org/10.1186/s40537-024-01020-6.

[11] Gani Joy, Usman, Kazi Ekramul Hoque, Mohammed Nazim Uddin, Linkon Chowdhury, and Seung-Bo Park. ‘A Big Data-Driven Hybrid Model for Enhancing Streaming Service Customer Retention Through Churn Prediction Integrated With Explainable AI’. *IEEE Access* 12 (2024): 69130–50. https://doi.org/10.1109/ACCESS.2024.3401247.

[12] Liu, Xinyu, Guoen Xia, Xianquan Zhang, Wenbin Ma, and Chunqiang Yu. ‘Customer Churn Prediction Model Based on Hybrid Neural Networks’. *Scientific Reports* 14, no. 1 (2024): 30707. https://doi.org/10.1038/s41598-024-79603-9.

[13] Cenggoro, Tjeng Wawan, Raditya Ayu Wirastari, Edy Rudianto, Mochamad Ilham Mohadi, Dinne Ratj, and Bens Pardamean. ‘Deep Learning as a Vector Embedding Model for Customer Churn’. *Procedia Computer Science* 179 (2021): 624–31. https://doi.org/10.1016/j.procs.2021.01.048.

[14] Hussain, Mazhar, Asad Javed, Samar Hayat Khan, and Muhammad Yasir. ‘Pillars of Customer Retention in the Services Sector: Understanding the Role of Relationship Marketing, Customer Satisfaction, and Customer Loyalty’. *Journal of the Knowledge Economy* 16, no. 1 (2024): 2047–67. https://doi.org/10.1007/s13132-024-02060-2.

[15] Rivaldo, Rivaldo, Rahman Taufik, Igit Sabda Ilman, and Ossy Dwi Endah Wulansari. ‘A Comparative Study of XGBoost, LightGBM, and CatBoost Models for Customer Churn Prediction in the Banking Industry’. *Jurnal Pepadun* 6, no. 2 (2025): 178–87. https://doi.org/10.23960/pepadun.v6i2.277.

[16] Ross, Nicholas. ‘Customer Retention in Freemium Applications’. *Journal of Marketing Analytics* 6, no. 4 (2018): 127–37. https://doi.org/10.1057/s41270-018-0042-x.

[17] Shaikhsurab, Mohammed Affan, and Pramod Magadum. *Enhancing Customer Churn Prediction in Telecommunications: An Adaptive Ensemble Learning Approach*. n.d.

[18] Thangeda, Rahul, Niraj Kumar, and Ritanjali Majhi. 'A Neural Network-Based Predictive Decision Model for Customer Retention in the Telecommunication Sector'. *Technological Forecasting and Social Change* 202 (May 2024): 123250. https://doi.org/10.1016/j.techfore.2024.123250.