Текст защиты дипломной работы

(Слайд 1: Титульный лист)

Добрый день, уважаемые члены аттестационной комиссии! Меня зовут Доготарь Дмитрий, я представляю вашему вниманию свою дипломную работу по теме:

Прогнозирование оттока клиентов телеком-компании и формирование персонализированных рекомендаций

Исследование проводилось на общедоступном датасете Telco Customer Churn с Kaggle, содержащем 7043 записи о клиентах.

(Слайд 2: Актуальность и цель)

На высококонкурентном рынке телекоммуникаций удержание клиентов — это ключ к успеху. Терять клиентов дорого, а привлекать новых — еще дороже. Поэтому целью моей работы была не просто разработка еще одной модели, которая предсказывает отток, а создание комплексной системы, которая:

* Точно определяет клиентов в группе риска.
* Находит экономически выгодную стратегию их удержания.
* Дает конкретные, персонализированные советы, что именно предложить каждому такому клиенту.

(Слайд 3: План работы)

Для достижения этой цели я прошел через несколько ключевых этапов: от анализа данных и создания новых признаков до обучения моделей, экономического анализа и, наконец, формирования итогового отчета с рекомендациями.

Весь код для анализа данных, обучения моделей и получения результатов, представленных в докладе, является полностью воспроизводимым и доступен в моем GitHub-репозитории по ссылке на этом слайде и в конце презентации.

(Слайд 4–6: Анализ данных — что влияет на отток?)

Я начал с исследовательского анализа данных. Главные выводы:

Был выявлен дисбаланс классов: около 27% клиентов ушли, что я учёл при обучении моделей.

Контракт решает всё: клиенты с помесячным контрактом уходят значительно чаще, чем те, у кого годовой или двухлетний.

Новички в зоне риска: чем меньше клиент с компанией, тем выше вероятность его ухода.

Дополнительные услуги важны: отсутствие техподдержки и онлайн-безопасности - сильные индикаторы будущего оттока.

(Слайд 7: Анализ существующих подходов)

При решении задач прогнозирования оттока в телекоме традиционно используются такие модели, как логистическая регрессия и случайный лес. Однако современные исследования показывают высокую эффективность градиентного бустинга, что, как мы убедимся далее, и подтвердилось в моей работе.

Отличительной особенностью моего подхода является не только применение мощной модели XGBoost, но и последующий экономический анализ для определения оптимальной стратегии удержания, что часто упускается в аналогичных исследованиях.

(Слайд 8: Выбор лучшей модели)

Для корректной оценки качества модели данные были разделены на обучающую и тестовую выборки в соотношении 80 на 20. Применялось стратифицированное разделение для сохранения исходного процентного соотношения ушедших клиентов в обеих выборках.

Я протестировал 8 различных моделей — от классической логистической регрессии до современных нейронных сетей, таких как TabNet.

Сначала, с базовыми настройками, лучшие результаты показала TabNet.

Однако после тщательной и автоматизированной настройки гиперпараметров с помощью фреймворка Optuna, лидером стала модель XGBoost, показав ROC AUC 0.8398. Это высокий результат, говорящий о хорошей предсказательной силе. Именно ее я и использовал в дальнейшей работе.

В качестве ключевой метрики для сравнения моделей я использовал ROC AUC (Площадь под кривой ошибок). Она показывает, насколько хорошо модель способна различать классы — в нашем случае, уходящих и лояльных клиентов.

Также использовал чистую экономическую прибыль (Net Profit): это ключевая бизнес-метрика для оценки эффективности стратегии удержания. Рассчитывается как разница между доходом от успешно удержанных клиентов и затратами на кампанию по удержанию.

Выбранная модель XGBoost, как ансамбль деревьев решений, по своей природе устойчива к выбросам в данных и не требует сложного масштабирования признаков, что подтверждает её робастность и применимость на реальных данных.

(Слайд 9: Главная идея - от точности к прибыли)

И здесь начинается самое интересное. Высокая точность модели — это хорошо, но бизнесу нужна прибыль. Просто предлагать скидки всем, кто «похож» на уходящего, — прямой путь к убыткам. Мы можем потратить на удержание больше, чем заработаем.

(Слайд 10: Поиск точки максимальной прибыли)

Поэтому я провел экономический анализ. Я задал простые бизнес-допущения: сколько мы тратим на удержание, сколько в среднем принесет удержанный клиент, и какова вероятность, что он примет наше предложение.

Затем я рассчитал чистую прибыль для каждого возможного порога вероятности оттока — от 0 до 100%.

На графике вы видите результат. Максимальная прибыль достигается не при максимальном охвате, а в конкретной точке. Оптимальный порог составил 64,2%. Это и есть наши «целевые» клиенты. Такая стратегия на тестовой выборке принесла бы компании условную прибыль в $25,662.

(Слайд 11: Кого нашли. Что им предложить?)

Итак, мы знаем, кого нужно удерживать. На тестовой выборке это 410 клиентов. Но что им предложить? Рассылать всем одинаковую скидку — неэффективно.

Я использовал метод k-ближайших соседей (k-NN). Для каждого "рискового" клиента я нашел в данных самого похожего на него, но лояльного клиента — его "двойника".

(Слайд 12: Почему клиент уходит? Делаем модель понятной)

Чтобы менеджер по удержанию мог вести диалог с клиентом более предметно, важно понимать причины прогноза. Для этого я использовал инструмент LIME.

На слайде — анализ для клиента с высокой вероятностью ухода. LIME наглядно показывает, что главные факторы риска для него — это:

Отсутствие долгосрочных обязательств (самый весомый фактор): клиент пользуется услугами на условиях ежемесячной оплаты. Это классический признак клиента, который не видит для себя долгосрочной ценности в услугах компании и готов уйти в любой момент.

Высокая финансовая нагрузка: ChargeRatio> 0.11 (вклад +0.147) говорит о том, что ежемесячный платеж для клиента является существенным. Возможно, его текущий тариф кажется ему завышенным по сравнению с предложениями конкурентов или его собственными ожиданиями. Чувствительность к цене — его вторая ключевая характеристика.

Недавний клиент: tenure <= 9.00 (вклад +0.082) означает, что клиент пользуется услугами менее 9 месяцев. Он еще не успел стать лояльным и, столкнувшись с высокой ценой, легко рассматривает альтернативы.

Фактор, снижающий риск оттока (Что его еще держит):

Тип интернет-услуги: InternetService\_Fiber optic <= 0.00 (вклад -0.101) — это самый интересный и ключевой фактор в его профиле. Это условие означает, что клиент НЕ пользуется оптоволоконным интернетом. Скорее всего, он использует более старую технологию, например, DSL.

Почему это снижает риск ухода?

Меньшая конкуренция: в его районе может быть ограниченное количество провайдеров, предлагающих альтернативу DSL. Переход на оптоволокно от другой компании может быть невозможен или сопряжен с трудностями.

Более низкая цена: DSL-тарифы обычно дешевле, чем оптоволокно. Несмотря на то, что ChargeRatio для него высок, абсолютная цена может быть не такой большой, и он это понимает.

Меньшая требовательность: Этот клиент, возможно, не является технически продвинутым пользователем, которому нужны сверхвысокие скорости. Его устраивает стабильность и базовая функциональность DSL.

Этот анализ дает менеджеру конкретные "болевые точки" для обсуждения.

(Слайд 13: Пример персонализированной рекомендации)

Логика проста: если "двойник" пользуется какой-то услугой, а наш "рисковый" клиент — нет, то именно эту услугу и стоит ему предложить, возможно, со скидкой.

Например, для клиента с ID 1334, у которого вероятность оттока 73%, система нашла похожего лояльного клиента, который пользуется телефонной связью. Наш клиент — нет. Рекомендация: "Предложить подключить PhoneService". Это просто, логично и персонализировано.

(Слайд 14: Выводы и результаты)

В результате проделанной работы была достигнута главная цель проекта. Я создал систему, которая решает все три поставленные задачи:

* Во-первых, она точно определяет клиентов в группе, используя модель XGBoost.
* Во-вторых, она разрабатывает прибыльную стратегию удержания, находя оптимальный сегмент для удержания.
* И в-третьих, она дает персонализированные и объяснимые советы, используя LIME и k-NN, помогая бизнесу принимать взвешенные решения.

(Слайд 15: Дальнейшие шаги)

Проект можно развивать дальше: провести A/B тестирование рекомендаций, чтобы измерить их реальную эффективность, а также обогатить модель новыми данными, например, из логов службы поддержки, модифицировать k-NN, чтобы он сравнивал не только отсутствующие услуги, но и ключевые факторы оттока, выявленные LIME (например, тип контракта).

(Слайд 16: Спасибо за внимание!)

На этом мой доклад окончен. Спасибо за внимание! Я готов ответить на ваши вопросы.