Итоговый отчет по дипломному проекту

Тема:

Прогнозирование оттока клиентов телеком-компании и формирование персонализированных рекомендаций

Автор: Доготарь Дмитрий

Руководитель: Корбут Даниил

Оглавление

[1. Постановка задачи 3](#_Toc210719771)

[Актуальность 3](#_Toc210719772)

[Цель проекта 3](#_Toc210719773)

[Метрики качества 3](#_Toc210719774)

[2. Анализ 4](#_Toc210719775)

[Исследование аналогов 4](#_Toc210719776)

[Предварительный анализ данных (EDA) 5](#_Toc210719777)

[Обоснование выбора алгоритмов 5](#_Toc210719778)

[3. Методика решения 6](#_Toc210719779)

[Подготовка данных и Feature Engineering 6](#_Toc210719780)

[Разделение выборки 7](#_Toc210719781)

[Выбор и оптимизация модели 8](#_Toc210719782)

[4. Результаты 8](#_Toc210719783)

[Качество итоговой модели 8](#_Toc210719784)

[Определение оптимальной бизнес-стратегии 8](#_Toc210719785)

[Формирование персонализированных рекомендаций (k-NN) 10](#_Toc210719786)

[Интерпретация прогноза для конкретного клиента (LIME) 10](#_Toc210719787)

[5. Выводы и заключение 11](#_Toc210719788)

[Основные выводы 11](#_Toc210719789)

[Пути развития проекта 12](#_Toc210719790)

[6. Список источников 12](#_Toc210719791)

# 1. Постановка задачи

## Актуальность

В условиях высокой конкуренции на телекоммуникационном рынке удержание существующих клиентов является ключевым фактором успеха. Потеря клиентов (отток) приводит к прямым финансовым потерям, связанным с прекращением получения абонентской платы, и к дополнительным затратам на привлечение новых абонентов, которые, по статистике, в 5–7 раз выше затрат на удержание. Таким образом, разработка эффективной системы прогнозирования и предотвращения оттока является критически важной бизнес-задачей.

## Цель проекта

Разработать комплексную систему для прогнозирования оттока, которая не просто выявляет клиентов из группы риска, а определяет экономически обоснованную стратегию по их удержанию и формирует для них персонализированные рекомендации.

Цель делится на следующие подзадачи:

* + - * Создать модель машинного обучения, точно предсказывающую вероятность оттока клиента.
      * На основе экономических показателей определить оптимальный порог вероятности для сегментации клиентов, на которых следует нацелить маркетинговые усилия.
      * Для клиентов из целевого сегмента сгенерировать конкретные, персонализированные предложения по удержанию.
      * Обеспечить интерпретируемость результатов для бизнес-пользователей.

## Метрики качества

Для оценки решения использовались две ключевые метрики:

* + - * ROC AUC (Receiver Operating Characteristic Area Under the Curve): Основная метрика для оценки качества классификационной модели. Она показывает, насколько хорошо модель способна разделять классы (ушедших и лояльных клиентов). Эта метрика особенно важна в условиях дисбаланса классов, который наблюдается в исходных данных (около 27% клиентов ушли). Значение ROC AUC варьируется от 0.5 (случайное угадывание) до 1.0 (идеальная классификация).
      * Чистая экономическая прибыль (Net Profit): Ключевая бизнес-метрика для оценки эффективности стратегии удержания. Рассчитывается как разница между доходом от успешно удержанных клиентов и затратами на кампанию по удержанию.

Эта метрика позволяет найти оптимальный баланс между охватом "рисковых" клиентов и затратами на них, максимизируя итоговую выгоду для компании.

# 2. Анализ

## Исследование аналогов

Прогнозирование оттока клиентов — классическая задача в области Data Science. Большинство существующих решений фокусируются на построении точной модели классификации. В качестве базовых подходов часто используются логистическая регрессия и методы, основанные на деревьях решений (Random Forest, Gradient Boosting). Более продвинутые решения активно применяют градиентный бустинг (XGBoost, LightGBM, CatBoost) и нейронные сети (MLP, TabNet), которые показывают высокую точность на табличных данных.

Однако стандартные подходы часто останавливаются на этапе прогноза, оставляя бизнесу задачу по определению стратегии. Данный проект идет дальше, интегрируя экономический анализ для поиска оптимального порога и используя метод k-ближайших соседей (k-NN) для генерации персонализированных рекомендаций, что является его ключевым отличием и преимуществом.

## Предварительный анализ данных (EDA)

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

Пропуски и аномалии:

В столбце TotalCharges (общие расходы) было обнаружено 11 пропущенных значений, которые были удалены, так как составляли менее 0.2% от всех данных. Тип данных этого столбца был исправлен с object на числовой.

Дисбаланс классов:

Выявлен умеренный дисбаланс: 73.4% клиентов остались лояльны, 26.6% ушли. Это было учтено при обучении моделей с помощью взвешивания классов.

Анализ признаков:

Числовые признаки:

Новые клиенты (tenure) и клиенты с высокими ежемесячными платежами (MonthlyCharges) более склонны к оттоку.

Категориальные признаки:

Наибольшее влияние на отток оказывают тип контракта (помесячный контракт — самый рискованный), отсутствие услуг онлайн-безопасности (OnlineSecurity) и технической поддержки (TechSupport), а также тип интернет-соединения (оптоволокно).

Эти выводы легли в основу создания новых признаков и подтвердили важность выбранных переменных для моделирования.

## Обоснование выбора алгоритмов

Для решения задачи был выбран комплексный подход, включающий несколько типов моделей:

Базовые модели (Logistic Regression, Random Forest):

Для получения начальной оценки качества и сравнения.

Градиентный бустинг (XGBoost, LightGBM, CatBoost):

Являются стандартом де-факто для задач на табличных данных благодаря высокой точности и эффективности. CatBoost был выбран в качестве итоговой модели, так как он хорошо справляется с категориальными признаками и часто показывает лучшие результаты "из коробки".

Нейронные сети (MLP, TabNet):

Для проверки, могут ли современные архитектуры глубокого обучения дать прирост в качестве на данной задаче.

Optuna:

Фреймворк для автоматической оптимизации гиперпараметров, который позволил систематически и эффективно найти лучшие параметры для каждой из моделей, что является более надежным подходом, чем ручной подбор.

k-Nearest Neighbors (k-NN):

Выбран для генерации рекомендаций из-за своей простоты и интуитивной понятности. Метод находит «похожего» лояльного клиента (двойника), чьи услуги можно порекомендовать клиенту из группы риска.

LIME (Local Interpretable Model-agnostic Explanations):

Используется для интерпретации прогнозов «черного ящика» (например, CatBoost), объясняя, какие признаки и как повлияли на конкретное предсказание. Это делает модель понятной для бизнес-пользователей.

# 3. Методика решения

## Подготовка данных и Feature Engineering

Очистка:

Удален столбец customerID как неинформативный. Обработаны пропуски в TotalCharges.

Создание новых признаков (Feature Engineering):

NumAdditionalServices: Количество подключенных дополнительных услуг (например, онлайн-безопасность, стриминг). Позволяет оценить "вовлеченность" клиента в экосистему провайдера.

ChargeRatio: Отношение ежемесячной платы к общей. Может указывать на новых клиентов с высокими платежами, которые еще не накопили большой TotalCharges.

TenureGroup: Категоризация клиентов по "стажу" (New, Established, Loyal). Упрощает для модели выявление нелинейных зависимостей.

Кодирование и масштабирование:

Бинарные категориальные признаки (например, gender) были преобразованы в 0 и 1 с помощью LabelEncoder.

Остальные категориальные признаки преобразованы с помощью One-Hot Encoding (get\_dummies).

Все числовые признаки были масштабированы с помощью StandardScaler для корректной работы моделей, чувствительных к масштабу (например, Logistic Regression, k-NN, нейронные сети).

## Разделение выборки

Данные были разделены в три этапа для обеспечения объективной оценки и качественной настройки моделей:

Исходное разделение: 80% на обучающую выборку (train) и 20% на тестовую (test). Тестовая выборка откладывалась и использовалась только для финальной оценки качества и экономического анализа.

Разделение обучающей выборки: train (80%) была разделена еще раз: 75% для обучения (X\_train\_new) и 25% для валидации (X\_val).

Обоснование: Такая схема (train\_new / val / test) является лучшей практикой при использовании фреймворков, подобных Optuna. Гиперпараметры подбираются на val выборке, а финальная оценка происходит на полностью «невидимой» test выборке, что предотвращает утечку данных и переобучение.

## Выбор и оптимизация модели

Процесс выбора лучшей модели состоял из двух этапов:

Baseline: было обучено 8 моделей с параметрами по умолчанию (Logistic Regression, Random Forest, 3 вида бустинга, MLP, TabNet). Лучший результат на этом этапе показала TabNet (ROC AUC = 0.8134).

Оптимизация с Optuna: для каждой из 8 моделей был проведен автоматический подбор гиперпараметров на 50 итерациях. После оптимизации наилучший результат показала модель XGBoost с ROC AUC = 0.8398.

Итоговая модель: XGBoost с оптимальными параметрами:

{'n\_estimators': 325, 'learning\_rate': 0.019763284969574767, 'max\_depth': 3, 'subsample': 0.6289804108291868, 'colsample\_bytree': 0.8753930344701659}.

Эта модель была обучена на полной обучающей выборке (80% данных) и использовалась для всех последующих шагов.

# 4. Результаты

## Качество итоговой модели

Итоговая модель XGBoost показала ROC AUC 0.8398 на валидационной выборке, что является высоким показателем для данной задачи и говорит о хорошей предсказательной способности. Модель устойчива к выбросам и не требует сложной предобработки категориальных признаков, что подтверждает ее робастность.

## Определение оптимальной бизнес-стратегии

Это ключевой результат проекта. На основе тестовой выборки был проведен экономический анализ для определения оптимального порога вероятности оттока.

Экономические допущения:

Средний жизненный цикл удержанного клиента (LTV): 12 месяцев.

Стоимость удержания: 2 ежемесячных платежа клиента (например, скидка).

Вероятность успешного удержания (Success Rate): 40%.

Результаты анализа:

Максимальная чистая прибыль: $25,662 (на тестовой выборке из 1407 клиентов).

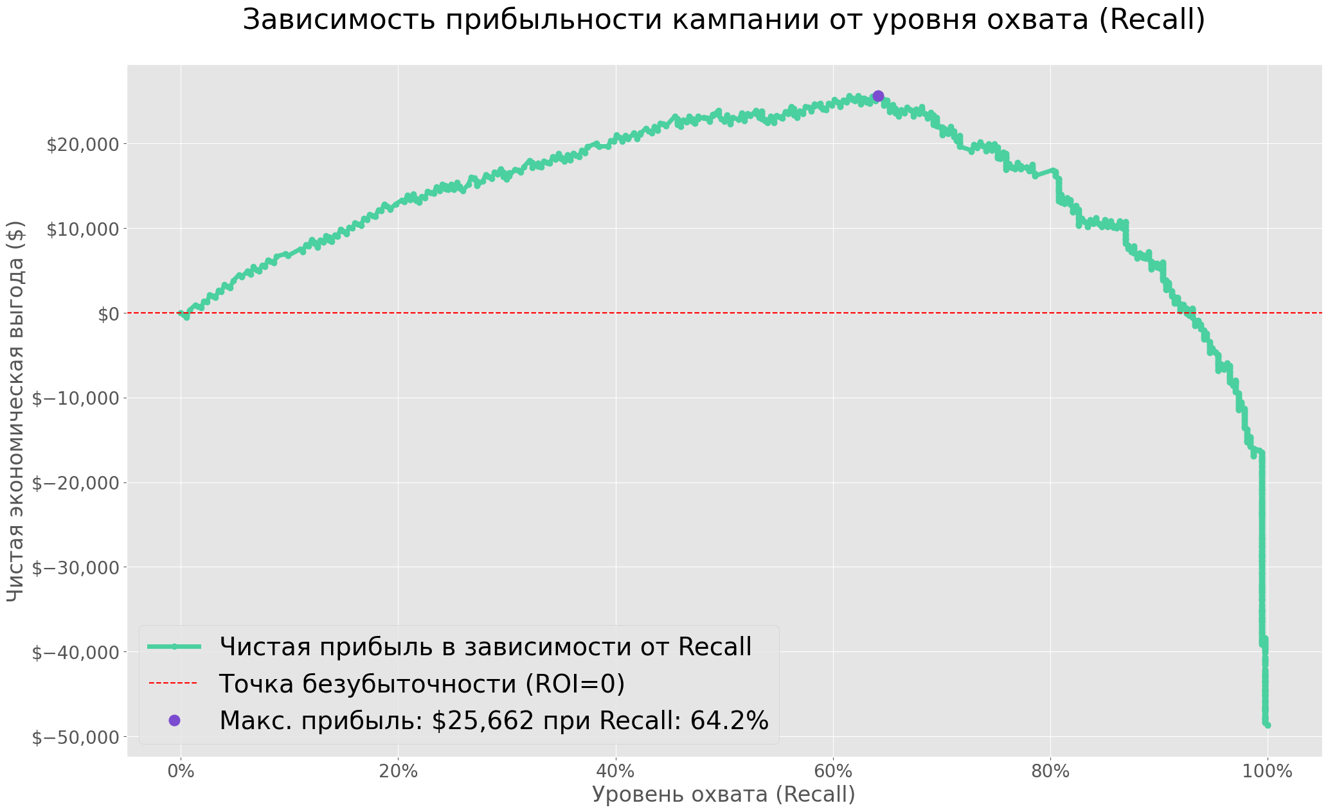
Оптимальный порог вероятности: 0.3624.

Охват (Recall): 64.2% (мы выявляем 64% реально уходящих клиентов).

Точность (Precision): 58.54 % (59% из тех, на кого мы воздействуем, действительно собирались уйти).

Вывод: вместо того чтобы реагировать на всех клиентов с вероятностью> 20% или> 50%, стратегия, нацеленная на клиентов с вероятностью оттока выше 64.2%, приносит максимальную финансовую выгоду.

График зависимости прибыли от охвата (Recall)



## Формирование персонализированных рекомендаций (k-NN)

На основе оптимального порога было идентифицировано 410 клиентов для удержания. Для каждого из них был применен метод k-NN.

Логика:

Для «рискового» клиента находится самый похожий на него лояльный клиент («двойник») из обучающей выборки.

Сравниваются их наборы услуг.

Если у «двойника» есть услуга, которой нет у «рискового» клиента, эта услуга рекомендуется к подключению (например, в виде специального предложения).

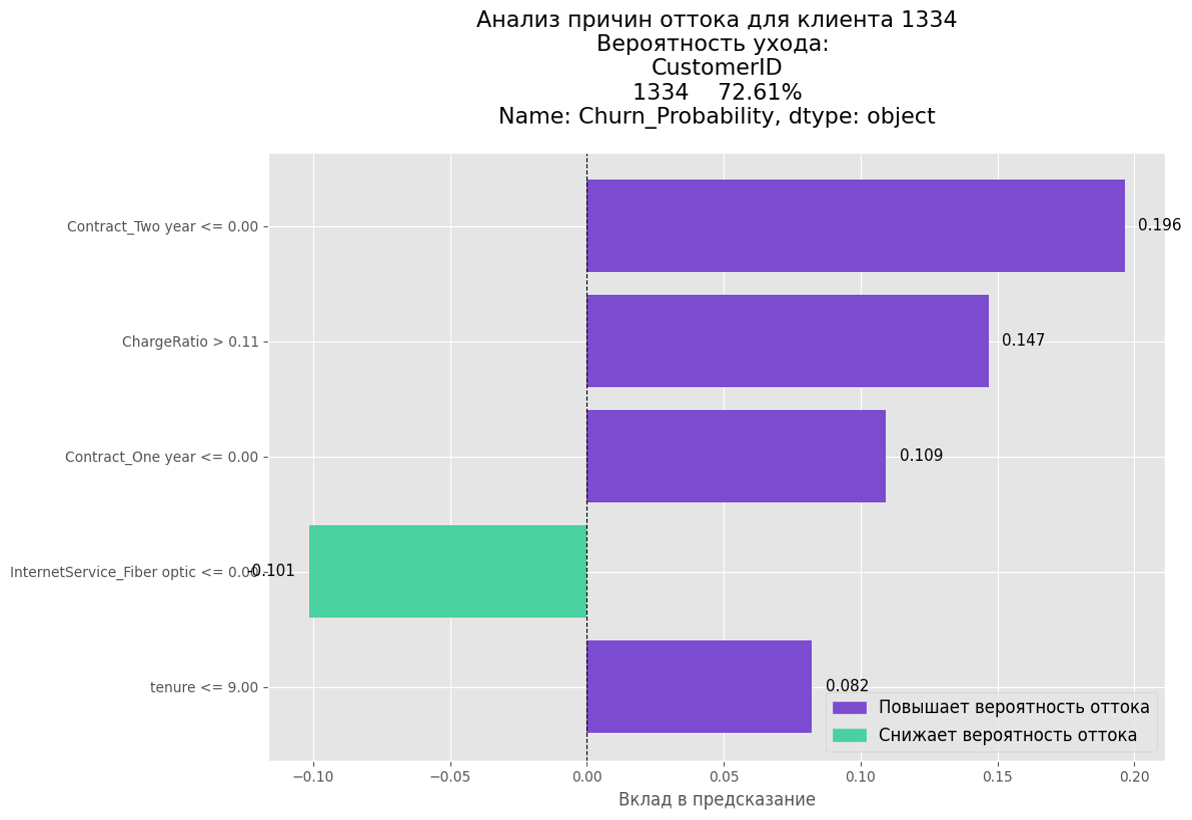
Если наборы услуг совпадают, рекомендуется предложить более выгодный контракт (например, годовой вместо помесячного).

Пример отчета:

|  |  |  |
| --- | --- | --- |
| CustomerID | Churn\_Probability | KNN\_Recommendation |
| 3380 | 82,48% | Предложить более выгодный контракт. |
| 1334 | 72,61% | Предложить подключить: PhoneService |

## Интерпретация прогноза для конкретного клиента (LIME)

Для клиента 1334 с высокой вероятностью оттока (72,61%) был проведен LIME-анализ, который показал ключевые факторы, повлиявшие на прогноз:



Этот анализ дает менеджеру по удержанию четкое понимание «болевых точек» клиента и позволяет сделать ему более релевантное предложение (например, подключить PhoneService).

# 5. Выводы и заключение

## Основные выводы

В ходе проекта была успешно разработана и протестирована комплексная система поддержки принятия решений для задачи удержания клиентов.

Ключевые достижения:

* + - * Высокоточная модель: построена модель XGBoost с ROC AUC = 0.8398, превосходящая 7 других алгоритмов после оптимизации.
      * Экономическая эффективность: доказано, что максимизация бизнес-метрик (прибыли), а не только технических (точности), приводит к созданию рентабельной и осмысленной стратегии удержания.
      * Персонализация и интерпретируемость: система не только определяет, кого удерживать, но и предлагает, как это сделать (k-NN) и почему клиент находится в группе риска (LIME).
      * Готовый инструмент: итоговый отчет представляет собой готовый к использованию список клиентов для отдела маркетинга, что делает решение практически применимым.

## Пути развития проекта

* + - * A/B тестирование: проведение A/B тестов для проверки эффективности предложенных рекомендаций.
      * Расширение данных: включение дополнительных данных (например, история обращений в поддержку, данные об использовании трафика) для улучшения точности модели.
      * Динамический анализ: внедрение модели в production для отслеживания вероятности оттока в режиме реального времени и адаптации предложений.
      * Альтернативные рекомендации: использование более сложных алгоритмов рекомендаций (например, коллаборативной фильтрации) при наличии достаточных данных.

## 6. Список источников

Проект основан на анализе и обработке данных, выполненных в среде Jupyter Notebook с использованием стандартных библиотек Python для анализа данных и машинного обучения (Pandas, Scikit-learn, XGBoost, CatBoost, Optuna, LIME и др.). Весь код, воспроизводящий результаты, представлен в приложенном файле [diplom\_telco\_customer.ipynb](https://github.com/DgtrDN/diplom/blob/main/diplom_telco_customer.ipynb).