НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

ВЫСШАЯ ШКОЛА БИЗНЕСА

ПРОЕКТ “ПРОГНОЗИРОВАНИЕ ОТТОКА КЛИЕНТОВ”

По дисциплине:
"Прикладная наука о данных"

Направление обучения: 38.04.05 Бизнес-информатика
Магистерская программа “Бизнес-аналитика и системы больших данных”

Выполнили:

студенты БАСБ 252

Бобров Валерий, Гришкина Анна, Милицына Полина,

Мороз Екатерина, Соболева Татьяна, Шпилёва Лина

Проверил:

д-р физ.-мат. наук, профессор,

Попов Виктор Юрьевич

МОСКВА 2025

Описание проекта

Актуальность

Модель прогнозирования оттока клиентов позволяет банку заранее выявлять клиентов с высоким риском ухода и предпринимать проактивные меры для их удержания. Это способствует:

- Сокращению расходов на привлечение новых клиентов
- Увеличению LTV (пожизненной ценности клиента)
- Достижению более стабильных пассивов
- Повышению общей прибыльности бизнеса

Основная цель

Построение модели машинного обучения для бинарной классификации (прогнозирования ухода клиента) на основе анализа профиля клиента, истории взаимодействия и финансовых показателей.

Задачи:

1. Провести разведочный анализ данных (EDA)
2. Подготовить данные для дальнейшего обучения модели машинного обучения
3. Сформировать новые признаки для улучшения предсказательной способности моделей
4. Построить несколько моделей машинного обучения
5. Оценить модели по метрикам качества и выбрать итоговую модель для дальнейшего внедрения
6. Сформулировать выводы по итогам проделанной работы
7. Подготовить рекомендации для бизнеса

Данные

Использован датасет Kaggle *Bank Churn Competition by ipii.hs.ex.mts*, включающий данные о демографии, финансовом поведении и факте ухода клиентов. Размер выборки: 15 000 клиентов. Пропусков в данных нет.

Метрики

Для оценки качества моделей: F1-Score, Precision, Recall, ROC-AUC.

Для бизнес-оценки использовались Churn Rate, Retention Rate, ARPU, LTV, ROI.

Команда проекта

Роль	Описание роли	Участник
Project Manager	Общее руководство проектом	Шпилева Лина
Data Analyst	Исследовательский анализ данных	Мороз Екатерина
Business Analyst	Анализ и оценка влияния на бизнес, юнит-экономика	Гришкина Анна
Data Engineer	Подготовка данных и feature engineering для дальнейшего обучения ML-модели	Милицына Полина
Data Scientist	Построение и оценка моделей	Соболева Татьяна
BI Analyst	Создание визуализаций и представление результатов	Бобров Валерий

Исследовательский анализ данных (EDA)

Первичный анализ:

- **Всего записей**: 15000
- Пропущенные значения в данных отсутствуют, заполнение пропусков не требуется

Анализ целевой переменной:

- Целевая переменная Exited не сбалансирована: 80% клиентов остаются, 20% уходят. Требуется балансировка весов в при обучении моделей.

Описание датасета:

Целевая переменная	
Exited	Факт ухода клиента (0/1)
Числовые признаки	
Credit Score	Кредитный рейтинг клиента
Age	Возраст клиента
Tenure	Количество лет обслуживания в банке
Balance	Баланс на счете
NumOfProducts	Количество банковских продуктов

EstimatedSalary	Предполагаемая заработная плата
Категориальные признаки	
Geography	Страна проживания (Франция, Испания, Германия)
Gender	Пол клиента (Мужской/Женский)
HasCrCard	Наличие кредитной карты (0/1)
IsActiveMember	Активность клиента (0/1)
Остальные признаки	
Customer ID	Уникальный идентификатор клиента * Неинформативен для прогноза оттока
Surname	Фамилия клиента * Не использовался для прогноза оттока

Основные зависимости:

Визуализация признаков позволяет сформулировать предположения о закономерностях в данных:

- *Взрослые клиенты (40-45+) чаще уходят из банка.* Возможно это связано с изменением финансовых привычек и поиском лучших условий для формирования пенсионных накоплений. Другая возможная причина – сервис банка, ориентированный на более молодую аудиторию.
- Клиенты в основном используют 1-2 продукта, более 2-х продуктов у одного клиента – редкость. *Высокий отток наблюдается среди тех, кто имеет 1 продукт или больше 2-х продуктов.* Возможно, низкая вовлеченность делает менее лояльными и более склонными к смене банка, тех кто пользуется одним продуктом. Те, кто пользуется более чем двумя продуктами, могут сталкиваться со сложностью и неудобством управления своим портфелем, что также повышает вероятность ухода.
- *Отток среди женщин значительно выше чем среди мужчин*, что может быть связано с гендерными различиями в финансовом поведении и чувствительности к качеству банковских услуг
- *Среди стран наиболее высокий уровень оттока наблюдается в Германии*. Возможно, рынок в Германии более конкурентный, и клиенты легче меняют банк. Или же клиенты в Германии могут быть более требовательными к финансовым условиям и уровню сервиса

Выбросы:

Используя метод детекции выбросов IQR можно заключить, что данные содержат незначительное количество выбросов. Наибольшая доля выбросов присутствует в переменной **Age (**2.76%):

- **CreditScore:** 9 выбросов (0.06%)
- **Age:** 414 выбросов (2.76%)
- **Tenure:** 1 выброс (0.01%)
- **NumOfProducts:** 19 выбросов (0.13%)

Корреляции между переменными:

- Корреляция с целевой переменной наблюдается у **Age** (0.45), **NumOfProducts** (-0.29), **IsActiveMember** (-0.20), **Balance** (0.15)
- Также наблюдается корреляция между парами признаков **Balance** и **NumOfProducts** (-0.41) и **Age** и **NumOfProducts** (-0.16)
- **Вероятно, имеют место нелинейные зависимости** между признаками и между признаками и целевой переменной, что удастся выявить благодаря feature engineering

Методология

Feature Engineering

Для повышения предсказательной способности модели были применены несколько методов feature engineering, направленных на выявление скрытых взаимосвязей между признаками и уменьшение шума в данных.

Во-первых, использовалось комбинирование категориальных признаков. Например, объединение признаков *страна проживания* и *пол клиента*. Такое преобразование позволило учитывать взаимодействие демографических факторов и их влияние на вероятность оттока.

Во-вторых, было реализовано комбинирование числовых признаков. Например произведение *возраста* и *кредитного рейтинга*. Это дало возможность отразить нелинейные эффекты и скорректировать влияние возраста на финансовое поведение клиентов.

Третьим шагом стало логарифмирование непрерывных числовых переменных (в частности, заработной платы), что позволило нормализовать распределения.

Наконец, был применён target encoding – замена категориальных значений на усредненные значения целевой переменной по группам, например, для комбинации *страна* × *возраст*. Такой подход помог отразить устойчивые различия между клиентскими когортами.

Модели и метрики

Для решения задачи классификации (предсказание оттока клиентов) были последовательно построены и протестированы несколько вариантов моделей на базе CatBoost. CatBoost был выбран в качестве основной алгоритмической платформы благодаря

его высокой устойчивости к переобучению, встроенной поддержке категориальных признаков и стабильным результатам на табличных данных (Яндекс.Практикум, n.d.).

В процессе экспериментов были разработаны следующие версии моделей:

1. Базовая модель CatBoost (модель 1) – без дополнительных преобразований признаков, использовалась как отправная точка для оценки качества.
2. CatBoost с продвинутым feature engineering (модель 2) – добавлены новые признаки, отражающие взаимодействие категориальных и числовых переменных.
3. CatBoost с feature engineering и подбором гиперпараметров с помощью Optuna (модель 3) – автоматизированный поиск оптимальных параметров позволил улучшить обобщающую способность модели.
4. CatBoost с feature engineering, Optuna и кросс-валидацией (модель 4) – для более надёжной оценки качества и снижения влияния случайности.
5. Ансамбль моделей (модель 5) – объединение нескольких моделей градиентного бустинга для повышения точности и устойчивости предсказаний.

Для оценки эффективности каждого варианта использовались основные метрики качества классификации:

- F1-score – интегральная метрика, отражающая баланс между точностью (Precision) и полнотой (Recall);
- Precision – доля корректно предсказанных уходящих клиентов среди всех предсказанных как уходящие;
- Recall – доля действительно уходящих клиентов, которых модель смогла правильно выявить;
- ROC-AUC – показатель общей производительности модели и способности различать классы.

Основное внимание уделялось метрике Recall, поскольку для задачи удержания клиентов приоритетом является максимальное выявление пользователей с риском оттока, даже если это снижает точность предсказаний.

Результаты моделирования

Полученные метрики качества в перечисленных выше моделя представлены в таблице:

Модель	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**
CatBoost (модель 1)	0.8570	0.6056	0.8275	0.6994	0.9250
CatBoost с feature engineering (модель 2)	0.8737	0.6462	0.8209	0.7232	0.9334
CatBoost с Optuna и feature engineering (модель 3)	0.8693	0.632	0.8375	0.720	0.936
CatBoost с Optuna и feature engineering и с кросс-валидацией (модель 4)	0.8800	0.6714	0.7894	0.7256	0.9297
Ансамбль моделей (модель 5)	0.8966	0.803	0.644	0.715	0.926

Исходя из значений в таблице лучшей моделью является CatBoost с feature engineering и подбором гиперпараметров (модель 3). У данной модели самый высокий recall (0.8375), а как было сказано ранее, на этой метрике делается особый фокус. Ансамбли градиентных бустингов, в свою очередь, показывают лучший Precision, но теряют в Recall, поэтому их лучше не использовать для решения данной бизнес-задачи. В итоге, было принято решение считать модель 3 лучшей по качеству, так как для бизнес-задач удержания ключевым приоритетом является Recall – важно выявить максимальное число клиентов с риском оттока, а не просто повысить точность прогноза**.** Именно данная модель и является конечной моделью, которая будет предсказывать отток клиентов из банка.

Вклад feature engineering

По результатам анализа важности признаков можно отметить, что восемь из десяти признаков с наибольшим вкладом в качество модели были созданы именно в рамках feature engineering. Это подтверждает ключевую роль производных признаков в улучшении объясняющей способности модели.

Высокие значения важности для взаимодействий (*Кредитный скоринг x Количество продуктов*, *Возраст x Количество продуктов*) и нормировок (*Количество продуктов/ год возраста*) показывают, что модель улавливает не просто влияние отдельных характеристик клиента, а их взаимосвязи, отражающие общее поведение пользователей. Так, произведение кредитного рейтинга и количества продуктов отражает степень вовлеченности надёжных клиентов в экосистему банка: пользователи с высоким скорингом и большим числом активных продуктов демонстрируют минимальную вероятность оттока. Напротив, при низком скоринге рост количества продуктов может сигнализировать о попытке удержания менее надёжных клиентов, что увеличивает риск ухода. Признак *Возраст x Активность* отражает поведенческий компонент: влияние активности клиента на вероятность оттока варьируется между возрастными когортами.

Показатель *Логарифм зарплаты x Кредитный рейтин*г характеризует совокупную платежеспособность клиентов. Различные сочетания этих двух признаков по-разному коррелируют с оттоком: высокие значения обоих факторов связаны с высокой вероятностью удержания, а сочетание высокого дохода и низкого кредитного рейтинга, наоборот, может указывать на склонность к переходу в другой банк[1].

Кроме того, наличие в топе таргет-кодированного признаков (*Target-encoded группа: Страна x Пол x Возрастная группа*) подтверждает устойчивые когортные различия по полу, возрасту и географии, которые модель успешно использует для повышения точности прогнозов.

Следует отметить, что метрики важности в бустинговых моделях отражают вклад признаков в качество прогноза, но не определяют направление эффекта.

[1] Подразумевается банк, который использует данные о кредитном рейтинге другого бюро кредитных историй (БКИ).

Финансовые метрики (юнит-экономика)

Метрика	**Значение**	**Комментарий**
Churn rate	20.11%	Если Churn = 20%, значит, 1 из 5 клиентов перестаёт пользоваться услугами. Мы брали столбец Exited (1 – ушёл, 0 – остался).
Retention rate	79.89%	
Средний баланс	42 636€	
ARPU (средняя выручка)	584.94€	ARPU = доход от % по остатку + услуг банка Так как в исходных данных нет прямого дохода, мы смоделировали его логически: доход от баланса – 1% (предположим, банк зарабатывает на остатках и комиссиях); доход от каждого продукта – 100 условных единиц. То есть: EstimatedRevenue = Balance * 0.01 + NumOfProducts *100 Это позволяет оценить, сколько в среднем приносит один клиент
LTV (ценность клиента)	1177.08€	≈ доход × срок × маржа

CAC (привлечение)	350€	задан вручную
ROI	236.31%	клиенты окупаются в 2+ раза

Анализ показал, что банк имеет устойчивую клиентскую базу с умеренным уровнем оттока:

- Churn rate – 20,1 %, что означает, что около 80 % клиентов продолжают пользоваться услугами.
- Средний баланс клиента – 42 636 €, что указывает на наличие клиентов со значительными остатками средств.
- ARPU (доход на клиента) – 584,94 €
- LTV (жизненная ценность клиента) – 1 177,08 €. То есть, за весь период сотрудничества один клиент приносит банку примерно 1,2 тыс. евро прибыли после учёта маржинальности

При CAC = 350 €:

- ROI = 236 %, то есть на каждый 1 € вложений в привлечение клиентов банк получает 2,36 € прибыли.
- Это свидетельствует о высокой окупаемости инвестиций в маркетинг и эффективной модели привлечения, а также о хорошей окупаемости самого бизнеса.

Визуализация данных

Был построен дашборд, на нём отражены ключевые KPI клиентского оттока (уровень оттока, коэффициент удержания, количество ушедших и общая клиентская база, средний кредитный скор, экспозиция при оттоке), а также диагностические визуализации: динамика оттока по стажу обслуживания, двухфакторная сегментация древовидной картой по географии и возрастным сегментам, зависимость оттока от числа продуктов, воронка формирования пула клиентов оттока (последовательность фильтров Баланс = 0 → Неактивный → 1 продукт → Credit Score < 600), а также распределения по группам баланса и полу.

По данным текущего среза уровень оттока ≈ 20,1% (удержание ≈ 79,9%). Наибольший вклад в отток формируют клиенты с нулевым балансом, неактивные, с одним продуктом и низким кредитным скором; минимальный риск наблюдается у клиентов с двумя продуктами. Профиль по стажу близок к U-образному (повышенный риск у «новых» клиентов и рост на длинном стаже); сегментация по географии и возрасту указывает на концентрацию риска в старших возрастных группах в ряде стран. Воронка количественно подтверждает траекторию (15K клиентов → 10K с балансом 0 → 8K неактивных → 6K с 1 продуктом → 4K со скором <600 → 3K churners). Выводы предполагают приоритизацию реактивации неактивных клиентов с нулевыми балансами, перевод клиентов из 1→2 продуктов и адресные меры в выявленных географическо-возрастных кластерах. Данные выводы подтверждаются результатами моделей, что говорит, что дашборд будет предоставлять полезную информацию пользователям.

Основные выводы и инсайты

1. Кредитный рейтинг и баланс не являются решающими для прогноза оттока. Поведенческие признаки (активность, количество продуктов, взаимодействие) несут большую информативность в предсказании оттока.
2. Риск оттока выше для отдельных сегментов:
    a. Клиенты из Германии демонстрируют заметно больший процент оттока по сравнению с Францией и Испанией. Это связано с региональными особенностями и, возможно, продуктовой политикой банка.
    b. Владельцы только одного банковского продукта гораздо чаще уходят, чем клиенты с несколькими продуктами. Из этого следует, что лояльность выше у клиентов, использующих несколько сервисов одновременно.
    c. Возраст также коррелирует с вероятностью оттока: более высокая вероятность ухода наблюдается среди клиентов среднего возраста, особенно в возрасте 40-50 лет.
3. Ансамбли моделей усиливают Precision, но теряют Recall. Такой подход может быть полезен, если банк ориентирован на снижение издержек от неверных обращений к лояльным клиентам. Однако приоритетнее минимизировать пропуски уходящих (FN), а не минимизировать ложные срабатывания (FP).
4. Модель CatBoost с feature engineering и подбором гиперпараметров с помощью Optuna даёт наилучший результат по ключевым метрикам (Recall, Precision, ROC-AUC). В частности, модель выигрывает по показателю полноты (Recall), что важно, т.к. это позволяет идентифицировать максимальное число фактически уходящих клиентов.
5. Добавление искусственно созданных признаков (feature engineering) улучшает предсказательную способность модели (например, показатель AUC увеличился на +1,1%).

Рекомендации для бизнеса

Приоритет удержания необходимо сосредоточить на клиентах из Германии, пользователях только с одним продуктом и возрастной группе 40-50 лет, поскольку в этих сегментах риск оттока выше при значимом потенциале роста ценности через расширение продуктовой корзины. Для снижения оттока необходимо внедрить персонализированные кросс-продажи (карты, депозиты, пакеты услуг) и акции с бонусами, а также улучшить клиентский опыт: упростить ключевые сценарии в приложении, обеспечить оперативную поддержку и сделать коммуникации проактивными и понятными. Модель оттока необходимо интегрировать с CRM/CDP так, чтобы при вероятности ухода выше 0,5 автоматически запускались сценарии удержания в релевантных каналах с учётом частоты контактов и согласий клиента; модель необходимо переобучать ежеквартально на обновлённых данных с мониторингом дрейфа признаков и тестированием альтернатив. Эффект необходимо оценивать по снижению оттока в целевых сегментах, приросту LTV и ROI кампаний, сопоставляя затраты на удержание с ожидаемыми потерями LTV при уходе и подтверждая причинно-следственный эффект A/B-тестами и когортным анализом.

Заключение

Проект достиг поставленной цели: разработана интерпретируемая модель CatBoost с расширенным feature engineering и подбором гиперпараметров, обеспечившая приоритетно высокий Recall (0,836) при ROC-AUC 0,936, что соответствует задаче максимального выявления “рисковых” клиентов. В ходе проекта были выполнены основные задачи: проведён EDA, подготовлены данные, сконструированы признаки, протестированы несколько вариантов моделей и выбрана итоговая, рассчитана юнит-экономика и, наконец, предложены практические рекомендации по удержанию клиентов. Экономическая оценка подтверждает практическую значимость использования модели: при некоторых допущениях LTV и ROI указывают на существенную окупаемость внедрения. Для операционализации предусмотрены интеграция с CRM/CDP (автозапуск сценариев удержания при порогах вероятности), регулярное переобучение и мониторинг дрейфа с последующей оценкой эффекта по LTV/ROI и A/B-тестам.

Список источников

1. Kaggle. (n.d.). *Bank Churn Competition by ipii.hs.ex.mts* [Dataset]. Kaggle. https://www.kaggle.com/
2. CatBoost. (n.d.). *CatBoost documentation*. Yandex. https://catboost.ai/en/docs/
3. Optuna. (n.d.). *Optuna: A hyperparameter optimization framework*. Preferred Networks. https://optuna.org/
4. Géron, A. (2023). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.
5. Google. (n.d.). *Intro to Financial Metrics & Unit Economics* [Online course]. *Google Data Analytics Certificate*. https://www.coursera.org/
6. Kaggle. (n.d.). *Churn prediction and customer lifetime value notebooks*. Kaggle. https://www.kaggle.com /
7. Яндекс.Практикум. (n.d.). *CatBoost: алгоритм градиентного бустинга. Практикум Яндекс. https://practicum.yandex.ru/blog/algoritm-gradientnogo-bustinga-catboost/*