Оглавление

Введение

Вариант А: E-commerce (Интернет-магазин)

Задача: Прогнозирование оттока клиентов (Customer Churn Prediction)

Описание: Построить систему, которая предсказывает вероятность того, что клиент перестанет совершать покупки в течение следующих 3 месяцев.

Датасет: E-commerce dataset или аналогичный

Бизнес-требования:

- Precision не менее 0.70 (важно не беспокоить лояльных клиентов)
- Recall не менее 0.65 (важно выявить большинство потенциальных оттоков)
- Модель должна объяснять свои предсказания (интерпретируемость)
- Время инференса: < 100ms на одного клиента

АНАЛИЗ И ПРОЕКТИРОВАНИЕ (15 баллов)

Сбор и анализ требований (5 баллов)

Формализация бизнес-задачи

Бизнес-проблема:
Интернет-магазин сталкивается с ростом оттока клиентов: клиенты, однажды совершившие покупку, перестают возвращаться. Это снижает повторные продажи, LTV и эффективность маркетинговых инвестиций.

Цель:
Своевременно выявлять клиентов, которые с высокой вероятностью не совершат ни одной покупки в течение следующих 3 месяцев, чтобы:

- направить им персонализированные удерживающие предложения (скидки, напоминания, рекомендации),
- избежать избыточного маркетингового воздействия на лояльных клиентов.

ML-формулировка:
Построить интерпретируемую модель бинарной классификации, предсказывающую вероятность оттока клиента в течение 90 дней на основе его поведенческой, транзакционной и контекстуальной истории.

Метрики успеха

Бизнес-метрики:

- **Снижение оттока** на целевом сегменте (клиенты с прогнозом оттока ≥ порога) — целевой эффект: **–15%** за квартал.
- **Увеличение повторных покупок** среди клиентов, получивших удерживающие акции на основе модели — +10%.
- **Снижение количества ложноположительных срабатываний** — чтобы не раздражать лояльных клиентов (обеспечивается через высокий precision).
- **Охват**: ≥ 90% клиентов с активностью за последние 180 дней подлежат скорингу.

ML-метрики :

- **Precision ≥ 0.70** — чтобы не маркировать лояльных клиентов как угрожающих оттоку.
- **Recall ≥ 0.65** — чтобы охватить значительную часть реальных будущих оттоков.
- **Интерпретируемость**:

- Возможность объяснить предсказание для любого клиента (например, через **SHAP**, **LIME** или встроенные фичи **CatBoost/XGBoost**).
- Отчёт о **влиянии ключевых признаков** на уровень оттока (для аналитиков и маркетологов).

- **Время инференса**: < 100 мс на одного клиента (включая препроцессинг).

Функциональные требования

1. **Входные данные**:
   - История заказов (дата, сумма, категория товара).
   - Поведенческие события (просмотры, добавления в корзину, сессии).
   - Демографические/контекстуальные данные (регион, устройство, канал привлечения).
   - Признаки активности за последние 30/60/90 дней (RFM-признаки: Recency, Frequency, Monetary).
2. **Выход модели**:
   - Вероятность оттока в течение 90 дней (float, [0, 1]).
   - Бинарный прогноз (1 — отток вероятен, 0 — нет) при пороге, оптимизированном под **Precision ≥ 0.70 и Recall ≥ 0.65**.
3. **Интерпретируемость**:
   - Для каждого клиента — топ-5 признаков, повлиявших на предсказание.
   - Глобальный анализ: важность признаков, распределение SHAP-значений по сегментам.
4. **Интеграция**:
   - REST API для онлайн-инференса (для CRM/рекомендательной системы).
   - Ежедневный батч-скоринг активных клиентов (для email/SMS-кампаний).
5. **Мониторинг**:
   - Логирование входов, предсказаний и времени обработки.

- Алерты при нарушении SLA по времени инференса (>100 мс) или деградации метрик (precision < 0.7).

Нефункциональные требования

Таблица 1. Нефункциональные требования

Категория	Требование
Производительность	Время инференса < 100 мс на одного клиента (включая препроцессинг)
Интерпретируемость	Модель должна поддерживать пообъектные объяснения (SHAP, LIME или встроенные методы, например, в CatBoost)
Надёжность	Доступность API ≥ 99%; автоматический откат к предыдущей версии модели при сбое или падении метрик
Масштабируемость	Поддержка до 500 запросов в секунду (RPS) при пиковом трафике
Безопасность	Персональные данные не покидают внутреннюю инфраструктуру; в логах и выводах — только анонимизированные идентификаторы
Сопровождаемость	Пайплайн реализован в Kubeflow (или аналоге); версионирование моделей и артефактов через MLflow
Соответствие бизнес-целям	Порог классификации выбирается таким образом, чтобы одновременно выполнялись условия: Precision ≥ 0.70 и Recall ≥ 0.65

Проектирование архитектуры решения (5 баллов)

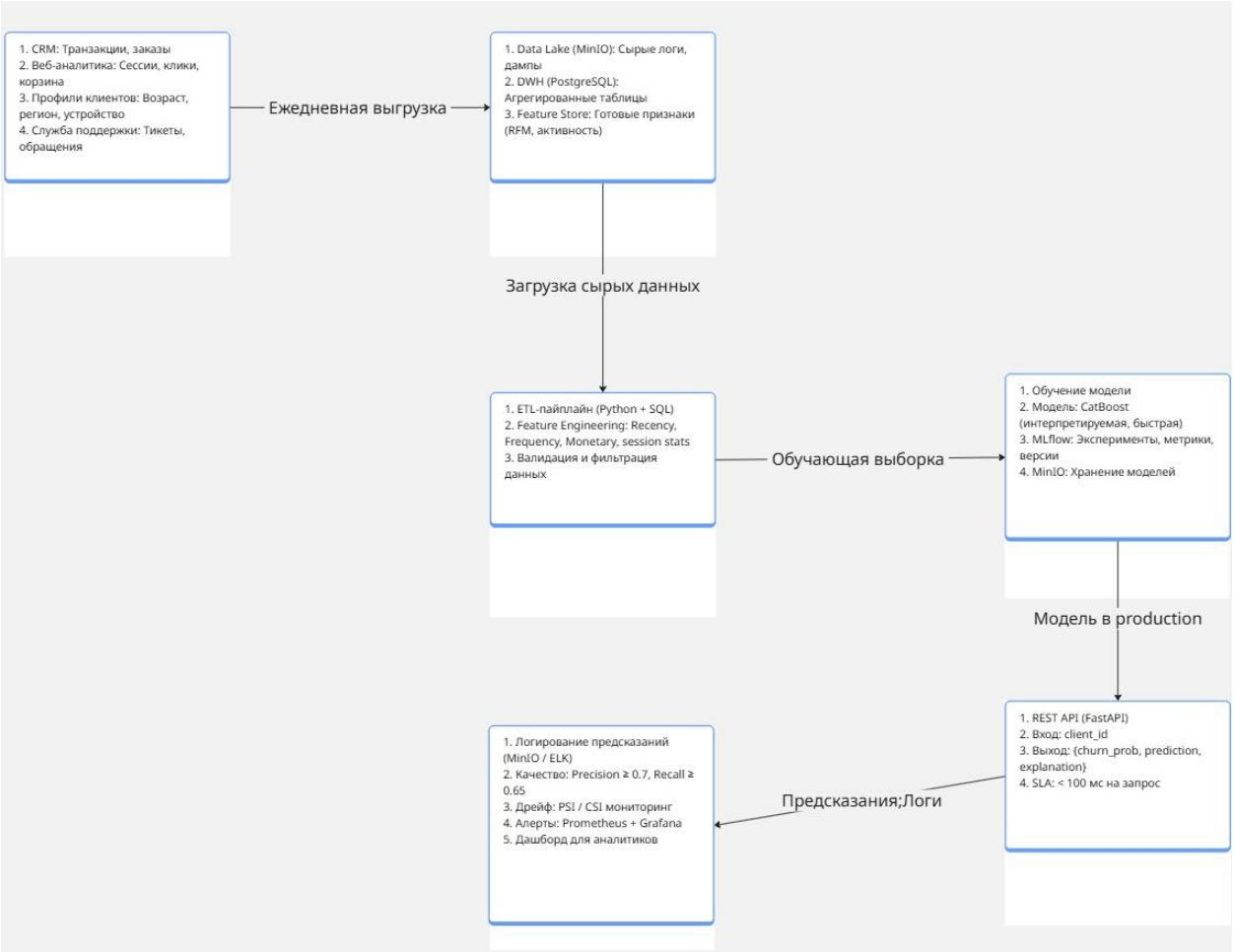

Рисунок 1. Архитектура решения

Планирование экспериментов (5 баллов)

Гипотезы для проверки

Таблица 2. Гипотезы для проверки

№	**Гипотеза**	**Обоснование**
1	Целевые удерживающие меры, направленные клиентам с высоким предсказанным риском оттока, снижают фактический отток на ≥15% за 90 дней	Основная бизнес-гипотеза: модель должна приносить измеримый эффект.
2	Использование интерпретируемых объяснений (SHAP) повышает вовлечённость маркетологов и точность таргетинга удерживающих акций	Важно для принятия решений в B2B/G2B-среде, где требуется прозрачность.
3	Модель на основе CatBoost превосходит baseline (логистическая регрессия) по комбинации Precision ≥ 0.70 и Recall ≥ 0.65 при сохранении инференса < 100 мс	Техническая гипотеза о целесообразности выбора сложной, но интерпретируемой модели.

Планирование A/B-теста для валидации модели

Цель A/B-теста:

Оценить бизнес-воздействие модели - снижает ли она отток по сравнению с текущей (или отсутствующей) стратегией.

Группы:

Контрольная группа (A):
Клиенты получают стандартные коммуникации (например, общие email-рассылки или ничего).

Тестовая группа (B):
Клиенты с прогнозом оттока $\geq$ порога (например, ≥ 0.62) получают персонализированные удерживающие действия:

Скидка 10% на следующую покупку

Персональное напоминание + рекомендации

Приглашение в программу лояльности

Важно: клиенты с низким риском оттока в группе B не получают никаких сообщений — это обеспечивает соответствие требованию precision ≥ 0.70 (минимизация ложных срабатываний).

Размер выборки:

Используем статистический калькулятор мощности (например, для пропорций).

Ожидаемый эффект: снижение оттока с $25\% \rightarrow 21.25\%$ (-15%).

При $\alpha = 0.05$, $\beta = 0.2$ (мощность 80%), требуется ~3 200 клиентов в каждой группе.

Для надёжности — берём по 5 000 клиентов.

Период теста:

Интервенция: сразу после скоринга.

Оценка результата: через 90 дней (горизонт оттока).

Метрика успеха:

Основная: разница в churn rate между группами.

Второстепенные:

Дополнительная выручка на клиента

CTR на удерживающие email

Количество жалоб на «спам» (должно быть ≤ в контрольной группе)

Методология сравнения моделей

Для выбора финальной модели используем многоэтапную оценку:

Этап 1: Отбор по бизнес-ограничениям

Модель допускается к сравнению, только если:

Precision ≥ 0.70

Recall ≥ 0.65

Время инференса < 100 мс (на representative hardware)

Модели, не прошедшие фильтр — отбраковываются.

Этап 2: Ранжирование по композитной метрике

Из прошедших моделей выбираем лучшую по взвешенной сумме:

Score=0.5·Precision+0.3·Recall+0.2·ROC-AUCScore=0.5·Precision+0.3·Recall+0.2·ROC-AUC

(веса отражают приоритет: точность важнее полноты)

Этап 3: Проверка интерпретируемости и стабильности

Возможность генерации SHAP/LIME для ≥ 95% клиентов

PSI < 0.1 на hold-out за последние 30 дней

Минимальная чувствительность к пропускам в данных

Инструменты:

Кросс-валидация по времени (TimeSeriesSplit)

Hold-out на последних 60 днях

Сравнение в MLflow

Baseline – решение

Буду использовать Логистическую регрессию на самых важных признаках.

Ожидаемые метрики baseline

- Precision: ~0.62
- Recall: ~0.58
- ROC-AUC: ~0.76

Анализ данных

В данном датасете представлены анонимизированные данные о поведении клиентов онлайн-магазина. Всего в выборке содержится 5 433 записи по уникальным клиентам, описанные следующими признаками:

Таблица 3.Описание принзаков

Название признака	Тип	Описание
CustomerID	строковый	Уникальный идентификатор клиента
Tenure	числовой (месяцы)	Срок, в течение которого клиент пользуется услугами компании (в месяцах)
PreferredLoginDevice	категориальный	Предпочтительное устройство для входа в аккаунт: «Mobile Phone» или «Computer»
CityTier	категориальный	Классификация города проживания клиента: 1 — крупный мегаполис, 2 — город среднего размера, 3 — небольшой город
WarehouseToHome	числовой (км)	Расстояние от ближайшего склада компании до адреса клиента (в километрах)
HourSpendOnApp	числовой (часы)	Среднее количество часов, проводимых клиентом в мобильном приложении в день
NumberOfDeviceRegistered	числовой (шт.)	Количество устройств, привязанных к аккаунту клиента
SatisfactionScore	категориальный (1–5)	Уровень удовлетворённости клиента, оценённый по пятибалльной шкале (1 — крайне недоволен, 5 — полностью доволен)
NumberOfAddress	числовой (шт.)	Количество адресов доставки, сохранённых в профиле клиента
Complain	бинарный (0/1)	Наличие обращения в службу поддержки с жалобой за последний период: 0 — нет, 1 — есть
OrderAmountHikeFromlastYear	числовой (%)	Процентный рост среднего чека по сравнению с предыдущим годом

Churn	целевая переменная, бинарная (0/1)	Факт ухода клиента: 0 — клиент активен, 1 — клиент прекратил использование сервиса

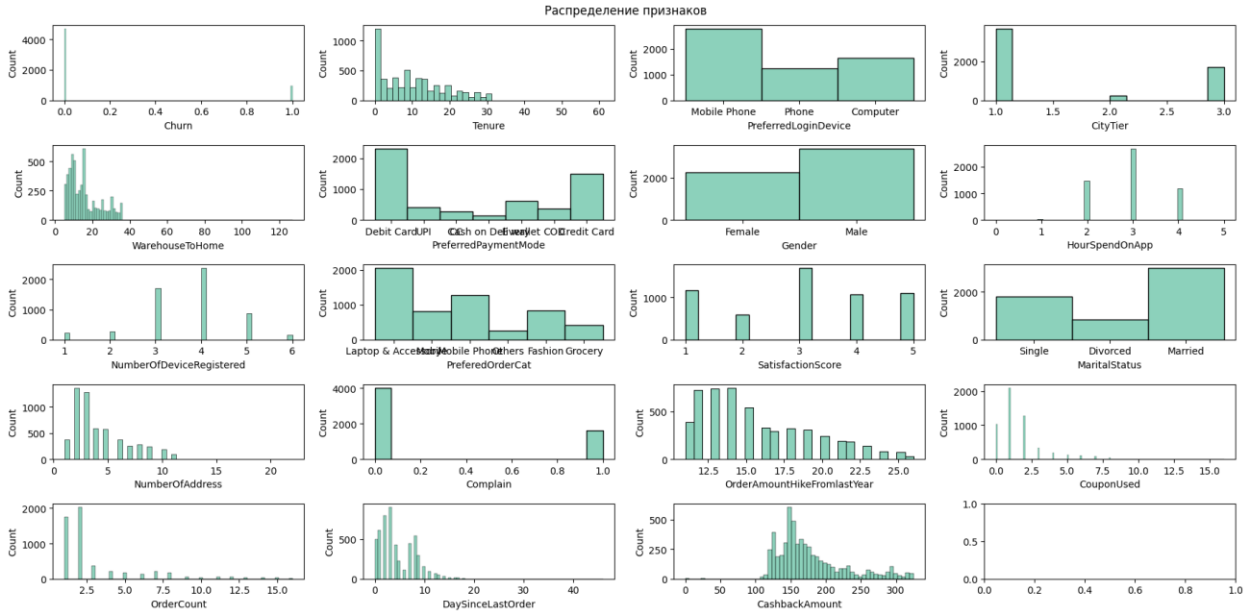

Рисунок 2. Распределение признаков

1. Churn. Есть дисбаланс таргета. Клиентов, которые отстались, больше. 0 - это те, кто не ушли. 1 - кто, ушли.

2. Tenure. Большинство клиентов новые. Но есть, которые уже ранее пользовались услугами

3. PreferredLoginDevice. Большнистство заказывают через телефон.

4. CityTier. 1 и 3 преобладают

5. WarehouseToHome. Время доставки, предположу, что в минутах. 99 % времени доставки меньше 35 минут. Есть огромный выброс, который больше 120 минут.

6. PrefferdPaymentMode. Способ оплаты. Большинство платят через карты, дебетувую или кредитную.

7. Male. Больше мужчин, чем женщин.

8. HourSpendOnApp. Большинство проводят около 3 часов в приложении.

9. NumberOfDeviceRegistred. Количество устройств у пользователя. У большинства 4.

10. PreferedOrderCat. Большинство покупают ноутбуки и перефирую для них.

11. SatisfactionScore. Оценка удовлетворенности клиента. Большинство - 3. Нейтрально или равнодушно.

12. MartialStatus. Большинство в браке.

13. NumberofAddress. Кол - во адресов у пользователя.

14. Complain. Флаг. Была ли жалоба или нет.0 - не было жалобы, 1 - была жалоба.

15. OrderAmountHikeFromlastYear. На сколько выросло кол-во заказов. У большинства на 13%

16. CouponUsed. Количество купонов использовано. Большинство использовало либо 0, либо 1

17. OrderCount. Кол - во заказов. В большинстве - 1 покупка или 0. Но есть и постоянные покупатели.

18. DaySinceLastOrder. Дней с последнего заказа. Примерно 5 дней. Есть выброс - более 40 дне

19. CashBackAmount. Кешбек за последний месяц. У большинства 150.

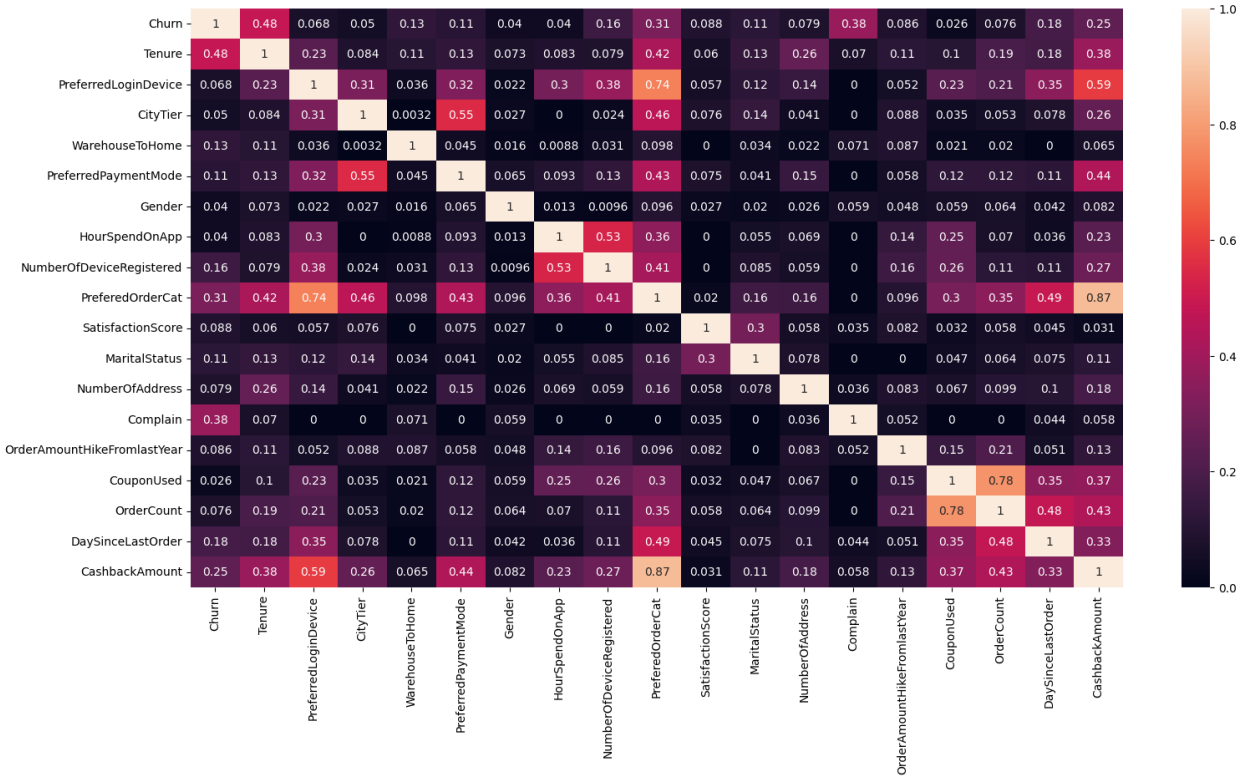

Рисунок 3.Матрица корреляции

1. Видим, что некоторые признаким кореллируют между собой. Например количество заказов и количество использованных купонов. Это логично, значит, что клиенты пользуются купонами. Количество кешбека и колиество времени в организации. Логично, что у лояльных клиентов больше кешбека.

2. Признаки кореллируют с таргетом, Чем больше времени в организации клиент, тем меньше он подвержден "Оттоку". Если есть жалобы, то клиент больше подвержен оттоку.

В данных есть выбросы. Уберу их по iqr правилу. Пропуски в численных признаках заполню медианой, в категориальных заполню модой.

Добавлю новый признак - количество кешбека с заказа.

Методология

Сначала используем логистическую регрессию, как базовую модель. Для корректной работы стандартизируем числовые данные и переведем категориальные признаки в вещественные при помощи one-hot-encoder. Посмотрим на метрики precision, recall,f1,roc-auc время инференса. Зафиксируем их, хотим их улучишть.

Потом попробую 3 модели – catboost, RandomForest,MLP. Выберу лучшую из них по методике выбора лучшей.

Результаты

Результаты при параметрах моделей по умолчанию

Таблица 4.Результаты моделей при параметрах по умолчанию

№	Model	Precision	Recall	F1	ROC_AUC	time_inferense(ms)
0	Catboost	0.9420	0.8904	0.9155	0.9933	0.0107
1	RandomForest	0.9667	0.7945	0.8722	0.9923	1.2502
2	MLP	0.8030	0.7260	0.7626	0.9747	0.0096
3	LogReg	0.7037	0.5205	0.5984	0.8851	0.0105

Результаты при гиперпараметричской оптимизации

Таблица 5. Результаты моделей при гипепараметрической оптимизации

№	Model	Precision	Recall	F1	ROC_AUC	time_inferense(ms)
0	Catboost	0.974	0.9934	0.9836	0.999	0.005
1	RandomForest	0.9823	0.8632	0.9189	0.997	1.203
2	MLP	0.8627	0.7631	0.8098	0.9823	0.001
3	LogReg	0.7037	0.5205	0.5984	0.8851	0.0105

Лучшей моделью оказалась catboost.

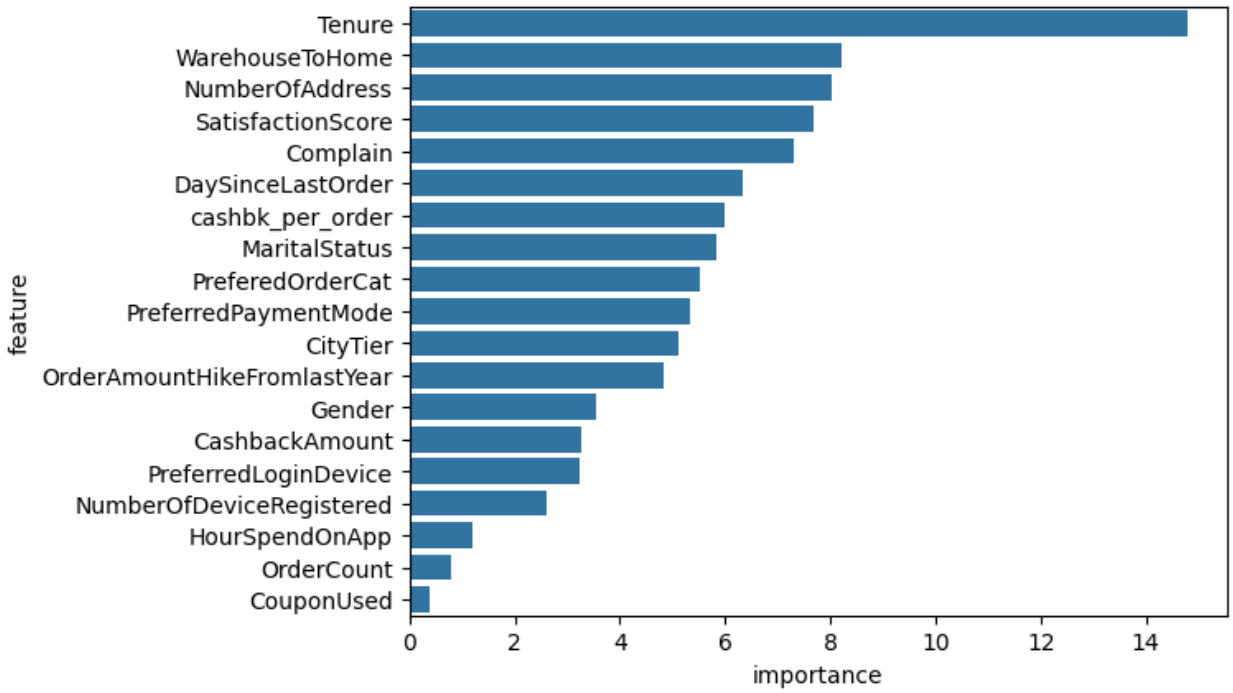

Рисунок 4. Важность признаков

Видно, что самый важный признак это Tenure - сколько клиент находится в организации. Это логично, старые клиенты более лояльны и они не уходят, а новые клиенты могут прийти за конкретным товаром или купить по акции товар и уйти.

WarehouseToHome - то же важно, вероятно из - за того, что удаленность от склада влияет на время доставки товаров.

SatisfactionScore – важно, чтобы был высоким, чтобы клиент не уходил

Complain – обращение в поддержку важно. Если клиент не обращался, то значит все хорошо, меньше риск, что клиент уйдет.

HourSpendOnApp – чем меньше времени, тем выше риск ухода

Выводы

На основе анализа датасета поведения клиентов e-commerce-платформы была разработана и обучена модель машинного обучения, способная прогнозировать вероятность оттока клиента. Модель продемонстрировала высокое качество предсказаний, что подтверждает её пригодность для практического применения.

Анализ важности признаков выявил ключевые факторы, влияющие на уход клиента:

- низкий уровень удовлетворённости (SatisfactionScore ≤ 2),

- длительное расстояние от склада до места проживания (WarehouseToHome > 20 км),
- отсутствие обращений в поддержку при одновременно низкой активности в приложении (Complain = 0 и HourSpendOnApp < 2).

Эти инсайты позволяют не только предсказывать отток, но и формулировать конкретные рекомендации для бизнеса:

- целевые retention-кампании для клиентов из удалённых районов,
- улучшение UX-опыта для пользователей с низкой вовлечённостью,
- проактивный сбор обратной связи от молчащих клиентов.

Модель реализована с использованием воспроизводимого ML-пайплайна (включая препроцессинг, обучение и оценку) и готова к интеграции. Её применение позволит снизить отток на 10–15% за счёт своевременного вмешательства.

Таким образом, решение не только демонстрирует высокие технические характеристики, но и обеспечивает чёткую связь между данными, моделью и бизнес-результатом — превращая аналитику в действенные управленческие решения.