

Анализ маркетинговой кампании, кластеризация клиентов и прогнозирование склонности к покупке обуви.

Цель:

1. Оценить эффективность маркетинговой кампании спортивного магазина через А/В-тестирование, определить влияние скидок на ключевые метрики (конверсия, выручка, средний чек, частота покупок).
2. Сегментировать клиентскую базу для персонализации маркетинговых стратегий.
3. Разработать модель для прогнозирования склонности клиентов из города 1188 к покупке обуви.

Ключевые выводы:

Маркетинговая кампания:

1. Конверсия в покупку в тестовой группе ниже на **3.67%**, чем в контрольной. Это означает, что рассылка скидок не мотивировала новых клиентов совершать покупки.
2. Кампания увеличила выручку на клиента на **19.4%** за счет повышения частоты покупок клиента на **25.6%**, особенно среди мужчин 19–60 лет в категориях Одежда, Активный отдых и спортивный инвентарь, Обувь.
3. Средний чек снизился на **5.0%**, что может быть связано с покупкой более дешевых товаров или использованием скидок.
4. Кампания повысила лояльность: доля клиентов с повторными покупками выросла на **3.45%**.

Кластеризация клиентов:

Выделено 5 кластеров, отражающих различия в поведении:

Кластер 0 (1.5%): Премиальные покупатели (в основном мужчины) транспорта и товаров для водных видов спорта.

Кластер 1 (18.0%): Ценочувствительные женщины, покупающие одежду и обувь со скидками.

Кластер 2 (17.4%): Мужчины, покупающие одежду и обувь со скидками.

Кластер 3 (37.0%): Мужчины, покупающие одежду и обувь по полной цене.

Кластер 4 (26.1%): Женщины, покупающие одежду и обувь без скидок, ориентированы на бренд.

Кластеризация позволяет таргетировать маркетинг, усиливая лояльность, удержание клиентов, увеличивая выручку, средний чек, привлечение новых клиентов.

Модель склонности к покупке обуви:

Классификатор (LightGBM) предсказывает факт покупки обуви (F1-macro: 0.7026, ROC-AUC: 0.7744).

Регрессор прогнозирует долю покупок обуви (R^2 : 0.4419).

Комбинированная модель (propensity_score) идентифицирует приоритетных клиентов для таргетинга.

Рекомендации для бизнеса.

Оптимизация маркетинговых кампаний:

1. Так как кампания имела успех в увеличении выручки и частоты покупок, повысила лояльность клиентов, сохранить персонализированные скидки для целевой аудитории (мужчины 19–60 лет, покупающие одежду, обувь, товары для активного отдыха и спортивный инвентарь).
2. Для повышения конверсии в покупку внедрить дополнительные стимулы: бесплатная доставка, подарки за первую покупку, ограниченные по времени скидки, розыгрыши, мастер-классы, мероприятия, акции “приведи друга”, пробные периоды.
3. Для увеличения среднего чека мотивировать клиентов к покупке более дорогих товаров: бесплатная доставка, комплектные товары, кросс-продажи, капсульные и сезонные распродажи, бонусы за покупки от определенной суммы, платежи в рассрочку.

Персонализация по кластерам:

1. **Кластер 0** (премиальный): предлагать премиум-комплекты, тест-драйвы, совместные акции с фитнес-клубами или спортивными событиями.
2. **Кластеры 1 и 2** (ценозависимые клиенты, реагирующие на скидки): создать акцент на выгоде (скидки, кэшбэк, бонусные программы, накопительные баллы), удобстве оплаты (рассрочка, оплата частями).
3. **Кластеры 3, 4** (нечувствительные к скидкам клиенты): подчеркивать эксклюзивность и статус (ограниченные коллекции, VIP-доступ), делать ставку на качество и инновации (новые технологии, премиальные материалы), обеспечивать персональный сервис (консьерж-поддержка, индивидуальные подборки), создавать эмоциональную связь (мероприятия, истории бренда).

Таргетинг для покупки обуви:

1. Использовать **propensity_score** для таргетинга клиентов с высокой склонностью к покупке обуви.
2. Персонализированные предложения (акции, мероприятия, новинки, скидки, комплекты) для клиентов с высоким **propensity_score**.

Загрузка и первичная обработка данных

Данные:

1. **purchases** (786260, 7) - данные о покупках (идентификатор клиента, товар, цвет, цена, скидка, дата, категория)
2. **personal_data_coefs** (104989, 5) - персональные коэффициенты, использован **personal_coef**
3. **personal_data** (89241, 6) - социально-демографические данные клиентов
4. **personal_data_csv** (15651, 5) - социально-демографические данные клиентов (без пола)
5. **ids_first_company_positive.txt** (5000) - id клиентов тестовой группы,
ids_first_company_negative.txt (5000) - id клиентов контрольной группы.

Данные отфильтрованы: оставлены клиенты страны 32.

Объединение данных:

Конкатенация **personal_data** и **personal_data_csv** (создан признак **gender** с NaN).

Внутреннее и левое объединение таблиц по **id** клиента.

Обработка пропусков.

1. **gender**: пропуски заполнены на основе детерминированной зависимости с **coef_personal** - коэффициенты **0.2576, 0.2672, 0.4304, 0.4688** соответствуют **мужчинам**, остальные — **женщинам**.
2. **colour**: значения унифицированы до 14 стандартных цветов (например, "графитовый", "темно-серый", "серебристый" → "серый", "зелёный/синий" → "мультиколор"), после чего пропуски заполнены модой ("черный").
3. **product_sex**: пропуски заполнены в три этапа:
 - Текстовый анализ названия товара ("мужской", "мальчиков" → 1.0; "женский", "девочек" → 0.0).
 - Анализ цвета ("розовый" → 0.0).
 - Построена модель классификации (**LightGBM**) для заполнения оставшихся пропусков с оценкой качества на кросс-валидации (F1-score: ~0.9731).

Преобразование текстовых признаков:

1. Создана колонка **product_group**, которая объединяет 977 уникальных товаров в 4 бизнес-ориентированные категории:
 - **Транспорт и водные виды**
 - **Обувь**
 - **Одежда**
 - **Активный отдых и спортивный инвентарь**
2. Цвета унифицированы, как указано выше.

Итоговый датафрейм:

Размер: (780117, 13). Пропуски устранены, текстовые признаки обработаны, данные готовы для дальнейшего анализа.

Анализ А/В-тестирования

Методология:

1. Группы: Тестовая группа (4992 клиента, получили скидку), Контрольная группа (4989 клиентов, не получили)
 2. Период анализа: Дни 5-16 (период действия кампании)
 3. Статистические тесты:
 - Хи-квадрат для сравнения конверсии
 - z-тест для сравнения долей
 - Шапиро-Уилка, Д'Агостино-Пирсона для проверки данных на нормальность распределения
- Манна-Уитни для сравнения медиан непараметрических данных независимых выборок

Результаты:

1. Конверсия в покупку:

Контроль: 99.10%.

Тест: 95.43% (-3.67% по сравнению с контрольной).

Вывод: Конверсия в тестовой группе значимо ниже ($p\text{-value} < 0.0001$). Рассылка скидок не привлекла новых покупателей.

2. Совокупная выручка, количество заказов:

Совокупная выручка:

Контроль: 110.9 млн ₽

Тест: 132.5 млн ₽ (+19.4%)

Количество заказов:

Контроль: 19 862

Тест: 24 971 (+25.7%)

Вывод: Кампания значительно увеличила общую выручку и объем продаж.

3. Средняя выручка на клиента.

На всех клиентов группы:

Контроль: 22 234 ₽ (медиана: 10 848 ₽)

Тест: 26 540 ₽ (медиана: 14 497 ₽)

+19.4%, p-value = 0.0000 (значимо выше в тестовой группе)

На совершивших покупку:

Контроль: 22 437 ₽ (медиана: 10 999 ₽)

Тест: 27 811 ₽ (медиана: 15 595 ₽)

+24%, p-value = 0.0000 (значимо выше в тестовой группе)

Вывод: Кампания положительно повлияла на доходность каждого клиента.

4. Средний чек

Контроль: 5584.9 ₽ (медиана: 3009.0 ₽)

Тест: 5305.73 ₽ (медиана: 2999.00 ₽)

-5%, p-value = 0.0002 (значимо выше в контрольной группе)

Вывод: Кампания привела к снижению среднего размера чека.

5. Частота покупок

На всех клиентов:

Контроль: 3.98 покупок/клиента (медиана: 2.00 покупок/клиента)

Тест: 5.00 покупок/клиента (медиана: 3.00 покупок/клиента)

+25.6%, p-value = 0.0000 (значимо выше в тестовой группе)

На совершивших покупку:

Контроль: 4.02 покупок/клиента (медиана: 2.00 покупок/клиента)

Тест: 5.24 покупок/клиента (медиана: 4.00 покупок/клиента)

+30.3%, p-value = 0.0000 (значимо выше в тестовой группе)

Вывод: Кампания успешно стимулировала повторные покупки и повысила лояльность.

6. Доля клиентов с повторными покупками (после кампании, $dt > 17$)

Контроль: 71.30%

Тест: 74.75%

+3.45%, p-value = 0.0001 (значимо выше в тестовой группе)

Вывод: Кампания положительно повлияла на долгосрочную лояльность клиентов.

7. Сегментированная выручка:

По категориям товаров:

Тестовая группа показала более высокую выручку по всем категориям, кроме "Транспорт и водные виды".

Одежда: +29.7%, абсолютный прирост 9.022 млн. руб.

Активный отдых и спортивный инвентарь: +28.2%, абсолютный прирост 4.711 млн. руб.

Обувь: +22.2%, абсолютный прирост 8.631 млн. руб.

Транспорт и водные виды спорта: -3.2%, абсолютная убыль 0.8 млн. руб.

По полу:

Мужчины: +27.5%, абсолютный прирост 19.547 млн руб

Женщины: +5.0%, абсолютный прирост 2.016 млн. руб

Мужчины внесли больший вклад в рост выручки.

По возрасту:

Наибольший относительный прирост выручки в тестовой группе:

Пенсионеры (>60): +34.1%, абсолютный прирост 1.405 млн. руб

Старшее поколение (46-60): +29.9%, абсолютный прирост 7.087 млн. руб.

Молодёжь (19-30): +26.1%, абсолютный прирост 5.753 млн. руб.

Взрослые (31-45): +11.8%, абсолютный прирост 6.451 млн. руб.

8. Сегментированная частота покупок

По категориям товаров:

Тестовая группа совершила больше покупок во всех категориях:

Одежда: +29.3%, абсолютный прирост 2 666 покупок

Активный отдых и спортивный инвентарь: 24.8%, абсолютный прирост 845 покупок

Обувь: +23.9%, абсолютный прирост 1 433 покупок

Транспорт и водные виды: +12.1%, абсолютный прирост 165 покупок

По полу:

Прирост частоты покупок у обоих полов:

Мужчины: +32.8%, абсолютный прирост 3 594 покупок

Женщины: +17.0%, абсолютный прирост 1 515 покупок

По возрасту:

Молодежь (19-30): +35.8%, абсолютный прирост 1 265 покупок

Старшее поколение (46-60): +29.6%, абсолютный прирост 1 376 покупок

Взрослые (31-45): 22.5%, абсолютный прирост 2 156 покупок

Пенсионеры (>60): +18.8%, абсолютный прирост 168 покупок

Дети (до 18): +12.1%, абсолютный прирост 144 покупки

Вывод: Основной вклад в рост частоты покупок внесли мужчины от 19 до 60 лет, покупающие Одежду и Обувь.

9. Выручка и количество покупок в разрезе дней

Ограничения: анализ основан на 12 днях наблюдений, для статистических выводов требуется большая выборка.

Средняя выручка в день:

Контроль: 9.24 млн ₽ (медиана 5.77 млн ₽)

Тест: 11.04 млн ₽ (медиана 11.08 млн ₽)

Среднее количество покупок в день:

Контроль: 1655 покупок (медиана 1031 покупка)

Тест: 2081 покупка (медиана 2059 покупок)

Вывод: Средняя выручка и частота покупок в день в тестовой группе были значительно выше, что подтверждает эффект кампании на ежедневную активность клиентов.

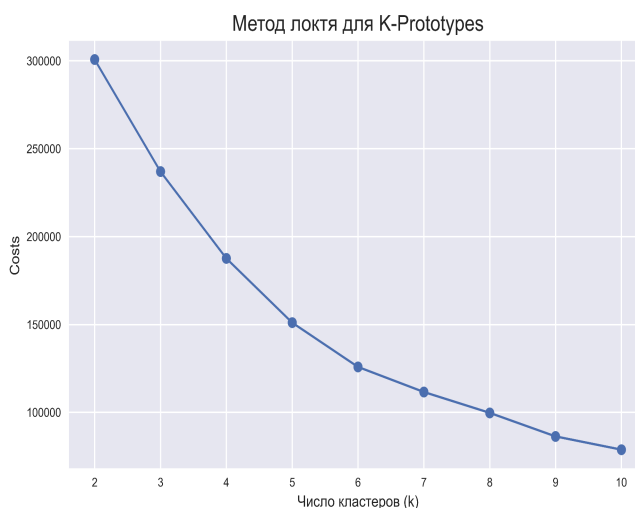
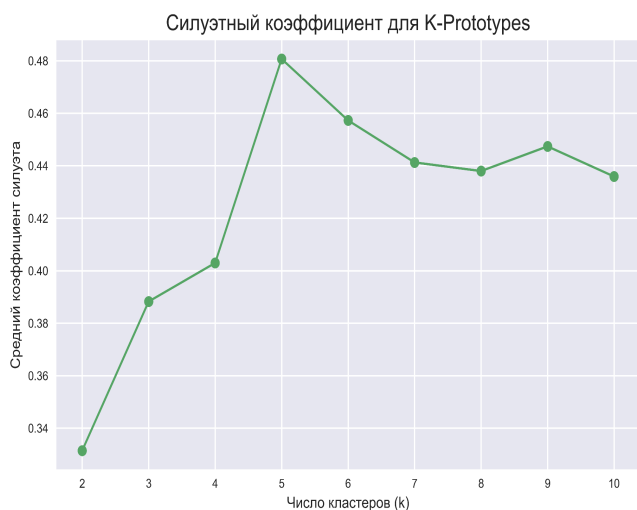
Общий вывод по А/В-тестированию:

Кампания достигла успеха в увеличении общей выручки и лояльности клиентов за счет роста частоты покупок, но не смогла привлечь новых клиентов и привела к снижению среднего чека. Основной вклад в рост выручки внесли мужчины 19–60 лет, покупающие "Одежду" и "Обувь".

Кластеризация клиентов

Методология:

- Алгоритм: **KPrototypes** (учитывает числовые и категориальные признаки).
- Признаки:
 - числовые: пол (**gender**), возраст (**age**), цена товара (**cost**), скидка (**base_sale**)
 - категориальные: категория товара (**product_group**), неделя покупки (**week**)
- Размер выборки: 100,000 записей
- Подбор числа кластеров: Метод локтя и силуэтный коэффициент. Оптимальное число кластеров — 5 (максимальный коэффициент силуэта - 0.48).



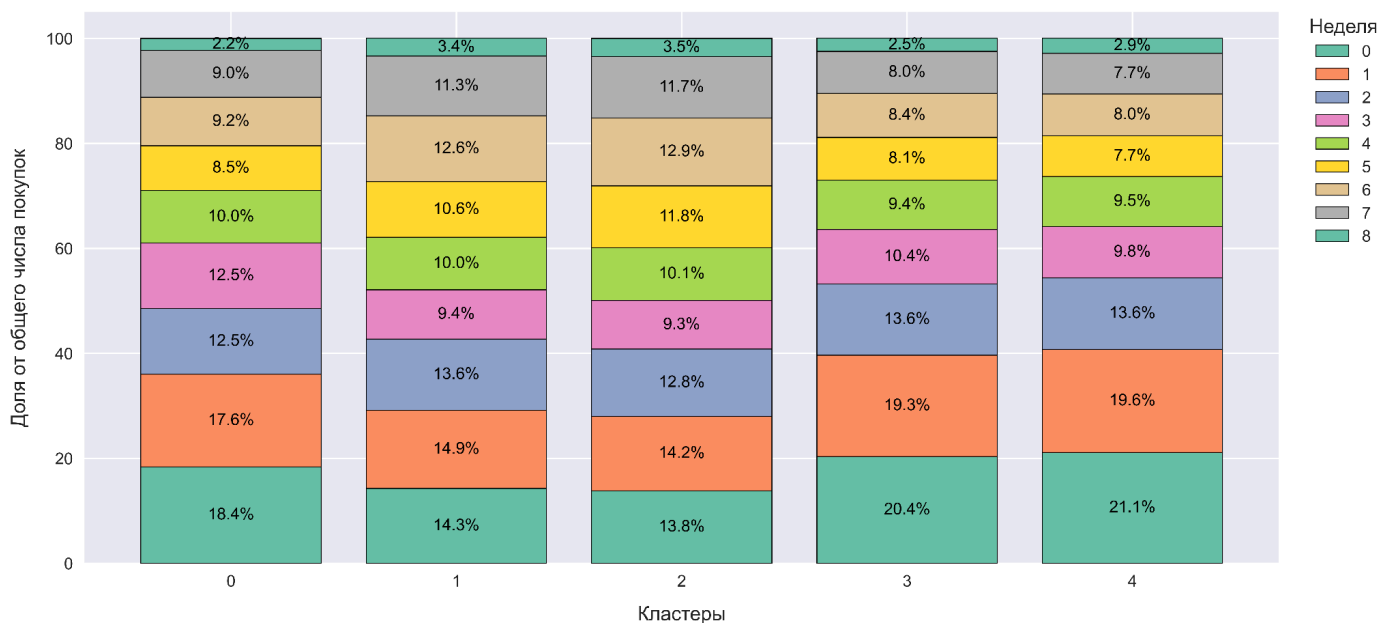
Предобработка данных:

- Биннинг даты покупки на 9 недель.
- Удаление неинформативных признаков (**id**, **product**, **city**, **product_sex**, **dt**, **education**, **colour**, **personal_coef**).
- Стандартизация числовых признаков с помощью **StandardScaler**.

Результаты кластеризации:

клас тер	размер (%)	пол	возраст	средний чек	скидка	основная категория	интерпретация
0	1.5%	82% ♂	~37 лет	~65 768 ₽	✗ 93%	•Транспорт и водные виды (63.4%) •Активный отдых и спортивный инвентарь (24.5%)	Премиальные покупатели транспорта и товаров для водного вида спорта
1	18%	♀	~40 лет	~3 037 ₽	✓	•Одежда (61.5%) •Обувь (31%)	Ценочувствительные активные покупатели-женщины одежды и обуви
2	17.4%	♂	~37 лет	~4 149 ₽	✓	•Одежда (50.5%) •Обувь (36.8%)	Мужчины, покупающие одежду и обувь со скидкой
3	37%	♂	~37 лет	~5 454 ₽	✗	•Одежда (35.8%) •Обувь (31.0%)	Покупатели-мужчины одежды и обуви по полной цене
4	26.1%	♀	~40 лет	~4 328 ₽	✗	•Одежда (42.3%) •Обувь (32.7%)	Бренд-ориентированные покупательницы одежды и обуви

Распределение числа покупок по неделям



Бизнес-интерпретация

Пик покупок в начале кампании (0-1 неделя) с тенденцией к уменьшению числа покупок со временем, поэтому рекомендовано ограничивать кампании со скидками по времени. Средний чек клиентов, покупающих со скидками ниже.

Модель склонности клиента к покупке обуви

Исходный датасет: 780 117 записей о покупках, включающих идентификатор клиента, пол, возраст, образование, город рождения, наименование товара, цвет, цену, скидку, дату покупки, персональный коэффициент и категорию товара.

Фильтрация: Для анализа использованы данные клиентов из города 1188 с положительной стоимостью товаров (88,888 записей).

Финальный датасет для моделирования: 88 888 транзакций.

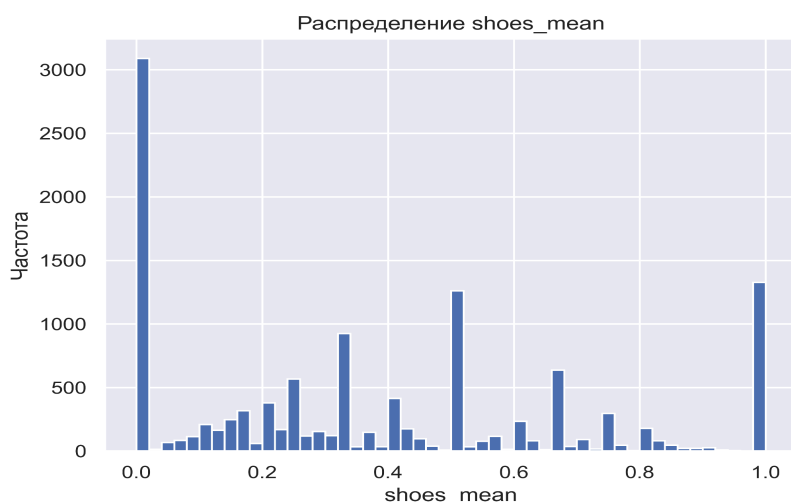
Методология:

Для решения задачи используется двухступенчатый подход:

1. **Классификатор (вероятность покупки):** Выявляет потенциальных покупателей обуви.
2. **Регрессор (ожидаемая доля обуви):** Оценивает интенсивность их склонности - насколько часто они выбирают обувь среди всех покупок.
3. **Финальный скор:**

склонность к покупке обуви (**propensity_score**) =

вероятность покупки ((**probability**) × ожидаемая доля обуви (**expected_share**)).



Ключевые этапы подготовки данных.

- **Агрегация по клиентам:** частота покупок, медиана цены, диапазон цен, доля покупок со скидками, средняя доля обуви и др.
- **Feature Engineering:** Созданы бинарные признаки: **shoes** (1, если товар — обувь), **education_high** (1, если образование высшее), биннинг возраста и стоимости, взаимодействие признаков (например, возраст × медианный чек) и статистические метрики (по стоимости, временным паттернам).

Для регрессора учтены только клиенты с **shoes_mean** > 0.

- **Препроцессинг:** числовые признаки масштабированы с помощью **StandardScaler**, категориальные признаки обработаны через **OneHotEncoder**. Для борьбы с дисбалансом классов в задаче классификации использовалась техника синтетической генерации данных **ADASYN** и параметр **class_weight** для назначения разных весов разным классам. Признаковое пространство для регрессии расширено с помощью класса **PolynomialFeatures**.

Моделирование

Классификатор (LGBMClassifier)

Задача: Прогнозирование факта покупки обуви.

Метрики (кросс-валидация):

Precision (класс 0): 0.5349, (класс 1): 0.8378.

Recall (класс 0): 0.4978, (класс 1): 0.8571.

F1-score (класс 0): 0.5156, (класс 1): 0.8473.

F1-macro: 0.6815

ROC-AUC: 0.7704

Accuracy: 0.7678.

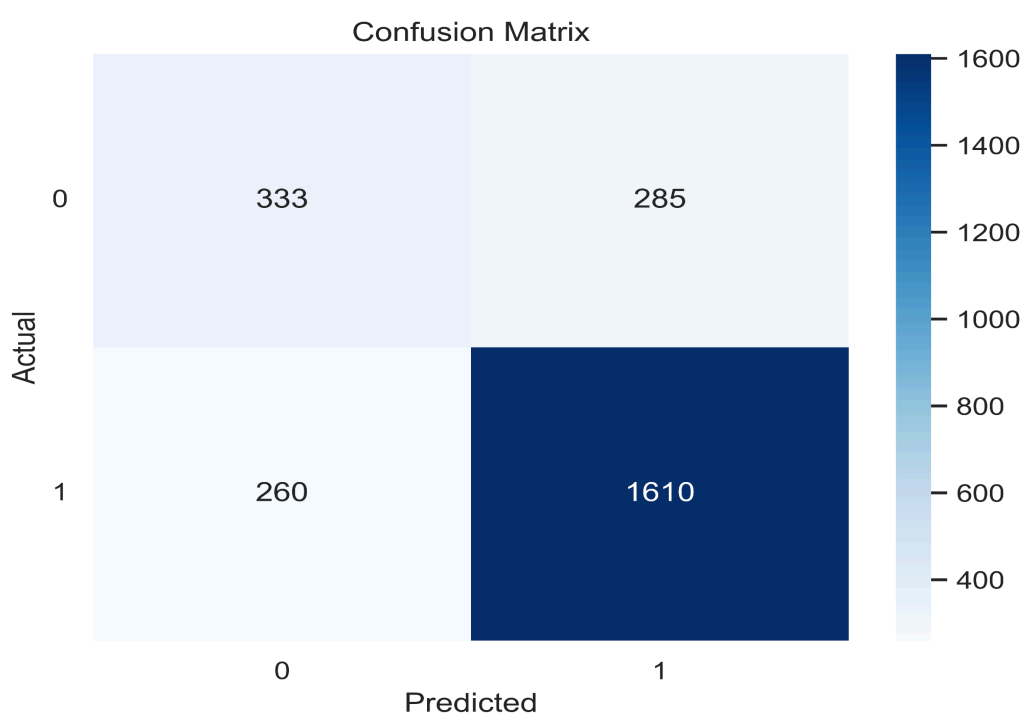
Метрики на тестовом наборе:

F1-macro: 0.7026

ROC-AUC: 0.7744

Precision (класс 1): 0.8496

Accuracy: 0.7809



Перцептор (LGBMRegressor).

Задача: Прогнозирование доли покупок обуви (**shoes_mean**).

Метод: LightGBM с регрессией, кодирование и масштабирование аналогично классификатору.

Метрики (кросс-валидация):

MAE: 0.1664 ± 0.0016

MSE: 0.0430 ± 0.0005

RMSE: 0.2073 ± 0.0012

R²: 0.4582 ± 0.0099 .

Метрики на тестовом наборе:

MAE: 0.1690

MSE: 0.0446

RMSE: 0.2112

R²: 0.4419.

Комбинированная модель (propensity_score):

Финальная модель предсказывает склонность клиента к покупке обуви.

Интерпретация: Клиенты с высоким **propensity_score** являются приоритетной целевой аудиторией для маркетинга.

Цель:	1
Ключевые выводы:	1
Маркетинговая кампания:.....	1
Кластеризация клиентов:.....	1
Модель склонности к покупке обуви:.....	2
Рекомендации для бизнеса	2
Оптимизация маркетинговых кампаний:.....	2
Персонализация по кластерам:.....	2
Таргетинг для покупки обуви:.....	3
Данные:	3
Объединение данных:	3
Обработка пропусков	3
Преобразование текстовых признаков:	4
Методология:	4
Результаты:	4
1. Конверсия в покупку:.....	4
2. Совокупная выручка, количество заказов:.....	4
3. Средняя выручка на клиента.....	5
4. Средний чек.....	5
5. Частота покупок.....	5
6. Доля клиентов с повторными покупками (после кампании, dt > 17).....	6
7. Сегментированная выручка:.....	6
По категориям товаров:.....	6
По полу:.....	6
По возрасту:.....	6
8. Сегментированная частота покупок.....	7

По категориям товаров:.....	7
По полу:.....	7
По возрасту:.....	7
9. Выручка и количество покупок в разрезе дней.....	7
Методология:.....	8
Предобработка данных:.....	9
Результаты кластеризации:.....	9
Бизнес-интерпретация.....	10
Методология:.....	10
Ключевые этапы подготовки данных.....	11
Моделирование.....	11
Классификатор (LGBMClassifier).....	11
Регрессор (LGBMRegressor).....	12
Комбинированная модель (propensity_score):.....	13