Итоговый аналитический отчёт

Автор: Лебедева Евгения

Проект: Анализ маркетинговых кампаний и клиентских данных

Платформа: Skillbox

Год: 2025

1. Введение.

В рамках проекта необходимо было проанализировать данные крупного спортивного магазина, включающие информацию о клиентах, их покупках и социально-демографических признаках.

Основные задачи включали: оценку эффективности маркетинговых кампаний, восстановление утерянной информации о поле клиентов, кластеризацию аудитории и построение модели склонности к покупке.

Результаты анализа призваны помочь в повышении точности маркетинговых коммуникаций и росте продаж за счёт персонализации.

2. Предобработка данных.

Была произведена загрузка и объединение таблиц из базы данных (SQL):

- ✓ personal_data,
- ✓ personal_data_coeffs
- ✓ purchases.

Из данных были удалены дубликаты, строки с критичными пропусками, а также отфильтрованы клиенты из страны с кодом 32, на которых фокусировался анализ.

Наименования товаров были стандартизированы — объединены различные варианты написания в единую категорию. Цвета товаров нормализованы: наборы цветов через слеш (/) разбиты, оставлен первый цвет как основной.

Данные успешно очищены и приведены к анализируемому виду, что позволило обеспечить корректность при обучении моделей и кластеризации.

Размер итогового дата-фрейма составляет: 16 столбцов и 664665 строк.

3. Восстановление пола клиентов.

Был добавлен CSV-файл с утерянной информацией о клиентах.

Так как часть данных о поле клиентов была утеряна, была поставлена задача бинарной классификации: предсказать пол (0/1) на основе других признаков.

В качестве модели была выбрана **Random Forest Classifier** — устойчивая и интерпретируемая модель, хорошо работающая с табличными данными и нечувствительная к масштабированию признаков.

В качестве признаков были использованы: возраст, образование, страна, город, персональный коэффициент.

Модель обучалась на части данных, где пол известен, а затем использовалась для восстановления пропущенных значений.

4. A/B-тестирование.

Для оценки эффективности первой маркетинговой кампании (email-скидка 5000 клиентам с 5 по 16 день) был проведён A/B-тест.

Методы и шаги анализа:

- ✓ **Выбор временного периода кампании:**
  Отобраны данные только с **5-го по 16-й день** (период проведения email-кампании).

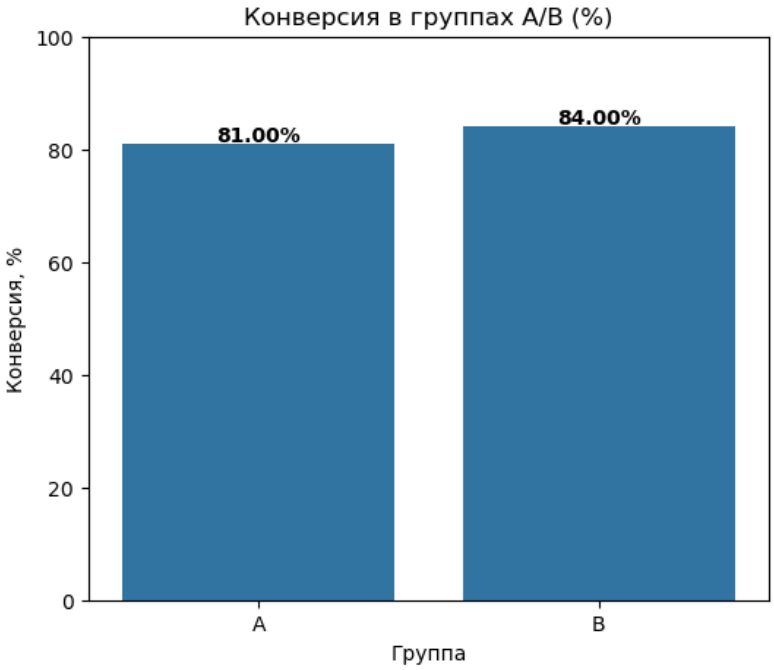
- ✓ **Импорт групп A и B**
  Группа A: ID клиентов, получивших предложение со скидкой
  Группа B: ID клиентов из контрольной группы
  (Данные загружены из ids_first_company_positive.txt и ids_first_company_negative.txt)

- ✓ **Создание целевой переменной**
  Для каждого клиента указано, совершал ли он покупку в указанный период (made_purchase = 1 или 0).

- ✓ **Расчёт конверсий**
  Для каждой группы подсчитаны:
  Кол-во клиентов, кол-во совершивших покупку и **конверсия (conversion_rate)**

- ✓ **Визуализация**
  Построен столбчатый график с конверсиями в группах A и B с подписями значений.

- ✓ **Z-тест на равенство долей**
  successes = [4063, 4199] — число покупок в группах A и B
  nobs = [5023, 5021] — размер групп
  Вычислены Z-статистика и p-value.



Z-статистика: −3.6
P-value: < 0.0001

Мы протестировали гипотезу:

- H_0 (нулевая гипотеза): Конверсии в группах A и B равны;
- H_1 (альтернатива): Конверсии различаются.

  Z-статистика: −3.6 P-value: < 0.0001

- P-value < 0.05 → мы отвергаем нулевую гипотезу;
- Разница между группами статистически значима;
- Email-скидка **не сработала** — контрольная группа показала **лучший результат**, причём не случайно, а статистически достоверно.

Вывод:

1. Предоставление персональной скидки не повысило, а снизило вероятность покупки.
2. С высокой долей вероятности — скидка воспринималась как лишняя, неуместная или подозрительная.
3. Кампания оказалась неэффективной и, возможно, даже вредной для выручки.

Рекомендации:

- Остановить повторение email-скидок в текущем формате.
- Требуется пересмотр содержания кампании, её каналов и целевой аудитории.
- Протестировать другие виды стимулов: Бонусы, подарки, лотереи;
- Информирование, а не навязывание скидки.
- Провести кластеризацию клиентов: выявить группы, которым действительно нужны скидки; предлагать персонализированные стимулы по сегментам.

5. Кластеризация клиентов.

Цель данного этапа - разделить клиентов на поведенческие сегменты для персонализации маркетинга — предложить для каждого сегмента релевантные товары, каналы и скидки, т.е. кластеризировать.

Использован метод KMeans, так как он хорошо работает при известных числах кластеров и позволяет чётко выделять группы на основе сходства по признакам.

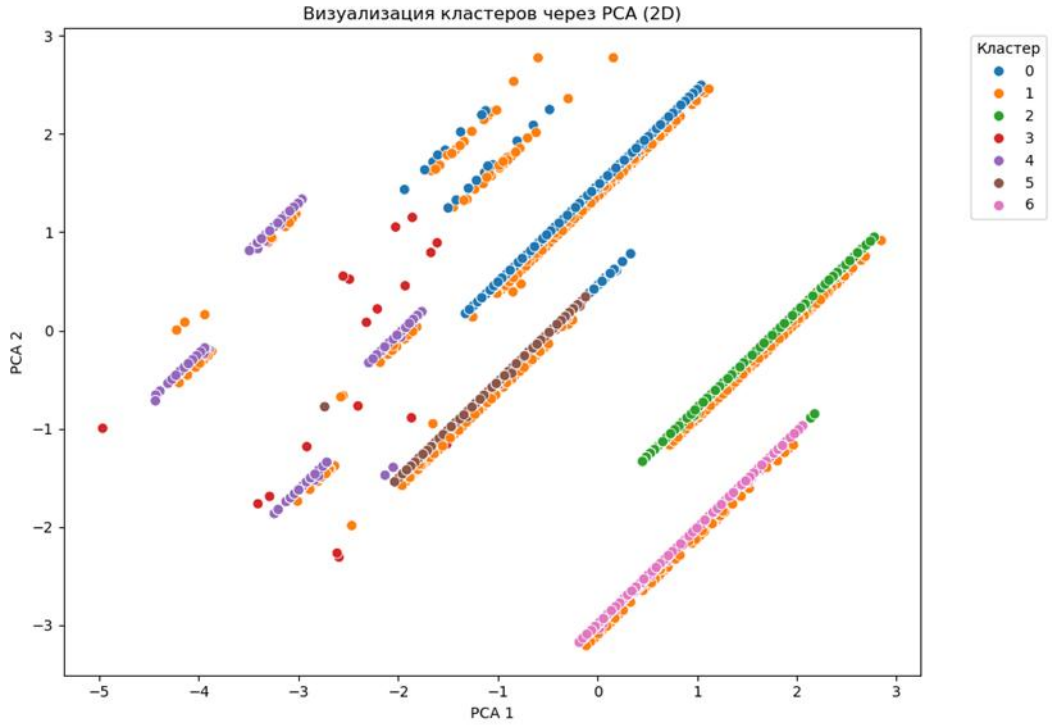
Для выбора оптимального числа кластеров применены методы:

- ✓ **Метод силуэта (silhouette score)** → максимален при **7 кластерах.**
- ✓ **Метод локтя (inertia)** → «локоть» около 4–5.

Выбрано **k = 7 кластеров**, как оптимальное по силуэту, поскольку silhouette_score учитывает не только плотность, но и качество границ между кластерами.

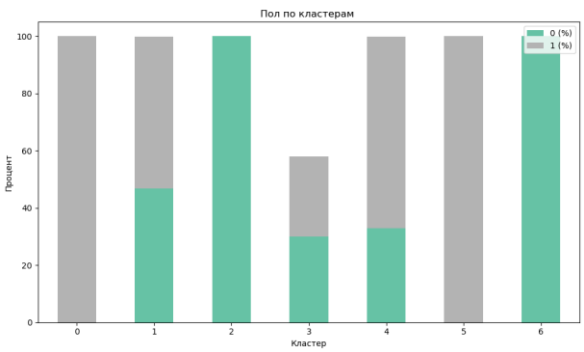
Признаки: возраст, пол, образование, коэффициент, чувствительность к скидке, количество покупок и средний чек.

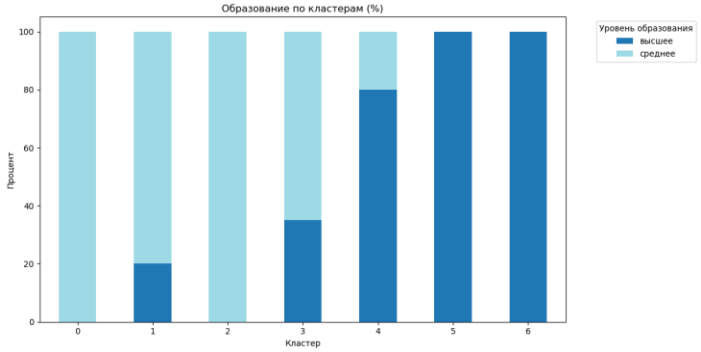
Для визуализации результатов использовано понижение размерности (PCA).

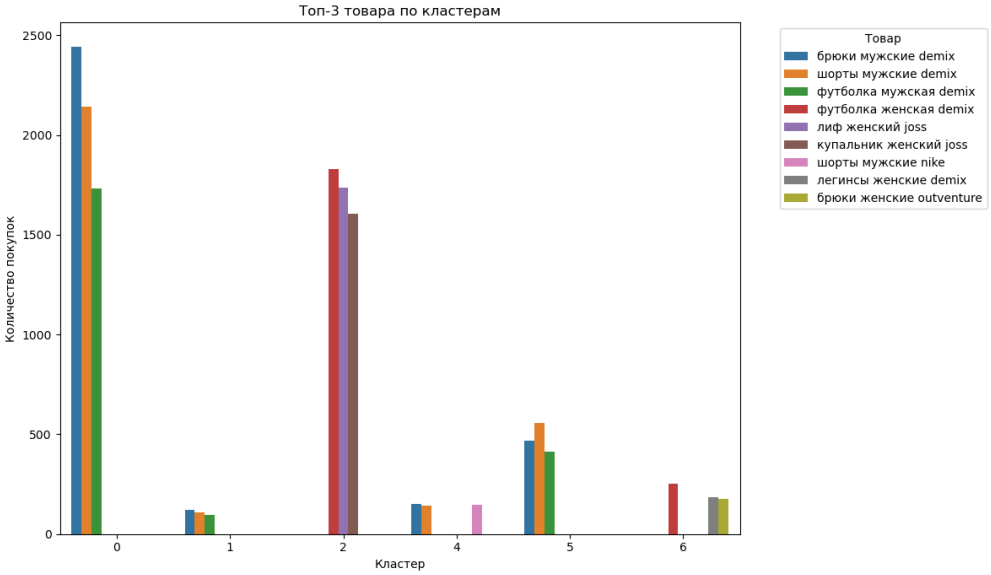


Далее сделан анализ и визуализация результатов сегментации:

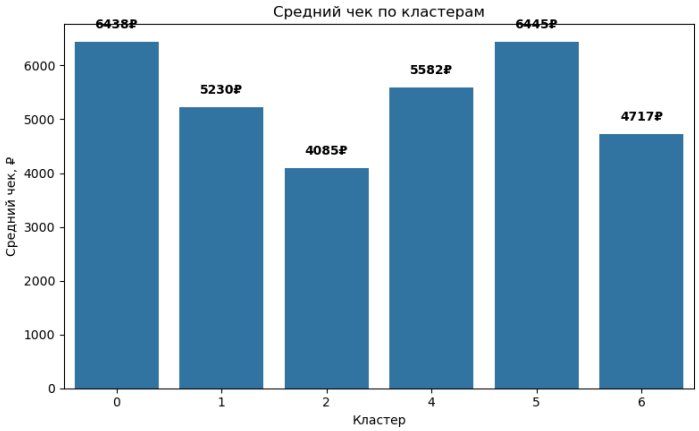
- ✓ Социально-демографические различия: распределение **возраста**, **пола** и **образования** по кластерам.
- ✓ Предпочтения по товарам: выделены ТОП-3 товара для каждого кластера.
- ✓ Чувствительность к скидкам: расчёт доли покупок по скидке (base_sale).
- ✓ Средний чек по кластерам: средняя стоимость покупки рассчитана и отображена графически.





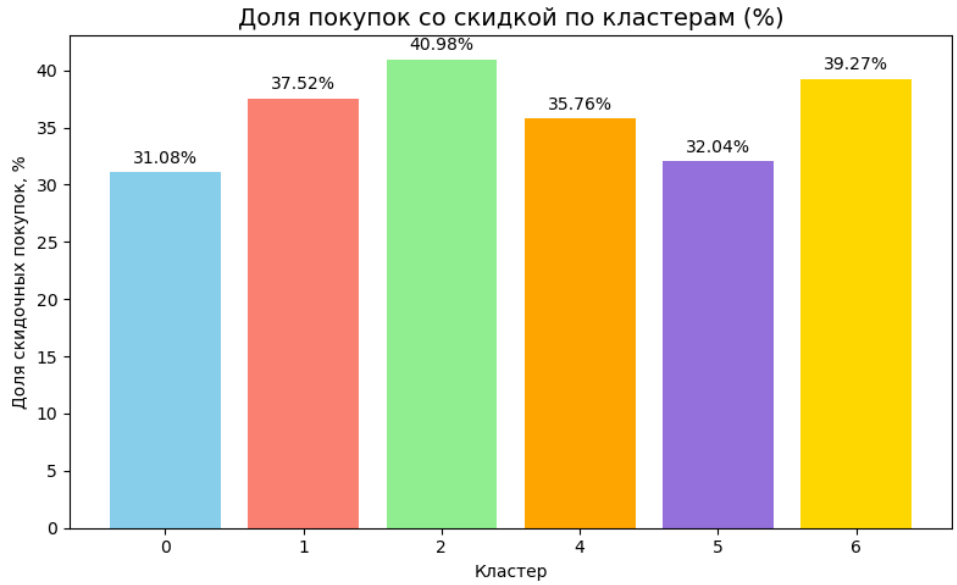


Кластер	Мин. возраст	Макс. возраст	Средний возраст
0	30	42	35.4
1	28	36	31.1
2	35	45	38.7
3	25	50	42.0
4	32	39	33.3
5	38	48	41.5
6	30	40	34.8



Если сравнить набор товаров и распределение полов, то мы видим, что:

- Кластеры 0 и 5 — полностью мужские,
- Кластеры 2 и 6 — полностью женские,
- Кластер 1 — наиболее сбалансированный,
- Кластер 4 — с преобладанием мужчин.



Выводы и рекомендации:

Кластер	Целевая аудитория	Поведение	Топ-3 товара	Скидки	Средний чек	Каналы и рекомендации
0	Мужчины ~35 лет, среднее образование	Рациональны, покупают без ориентации на акции	брюки, шорты, футболки мужские Demix	~31% — наименьшая доля	~6 438 ₽	Ремаркетинг, рекомендации в ЛК, push без скидок, уведомления о наличии и новинках
1	Молодые (28–32), сбалансированный пол	Предпочитают недорогую мужскую одежду, реагируют на скидки	брюки, шорты, футболки мужские Demix	~37.5% — высокая чувствительность	~5 229 ₽	Таргетированная реклама, кэшбек, mobile-first маркетинг
2	Женщины 38+, среднее образование	Часто покупают одежду для плавания и отдыха, реагируют на акции	футболки, лифы, купальники женские (Joss, Demix)	~41% — наибольшая зависимость	~4 084 ₽	Рекламные баннеры, офлайн-акции, SMS и Viber рассылки
3	Малочисленный сегмент, нестабильное поведение	Поведение не выявлено, выбор случайный	-	-	-	Требует дополнительного анализа или исключения из таргетинга
4	Мужчины и женщины ~33, преимущественно высшее образование	Чёткий выбор, брендовые предпочтения	брюки, шорты Nike и Demix	~36% — важны акции на бренды	~5 582 ₽	Ремаркетинг, персональные предложения в кабинете
5	Мужчины ~40+, высшее образование	Покупают качественную одежду, менее чувствительны к скидкам	шорты, брюки, футболки мужские Demix	~32% — скидки важны, но не критичны	~6 445 ₽	YouTube/Telegram, спец предложения по предзаказу
6	Женщины 30–40 лет, высшее образование	Спорт и досуг, средняя чувствительность к скидке	футболка, легинсы женские Demix, брюки Outventure	~39% — чувствительность высокая	~4 716 ₽	Инфлюенсер-маркетинг, подборки для женщин, каналы для спорта/фитнеса

6. Модель склонности к покупке.

Наша цель: создать модель, которая предсказывает, совершит ли клиент покупку после маркетингового взаимодействия (например, email, баннер, push). Это позволит использовать модель в качестве фильтра в будущих коммуникациях: не отправлять сообщения тем, кто с высокой вероятностью не заинтересован.

В качестве **таргета** (y) использовала бинарный признак:

target — 1, если покупка была, 0 — если нет.

Изначально в данных были только положительные примеры (то есть зафиксированные покупки), а **отрицательные (target = 0)** мы **сформировали самостоятельно**.

Использованы признаки:

- ✓ Демографические: возраст, образование, пол (восстановлен моделью),
- ✓ Поведенческие: личный коэффициент (personal_coef),
- ✓ География: страна, город (32 и 1188)
- ✓ Продуктовые характеристики: категория товара, наличие скидки (base_sale).

Использовалась модель "Случайный лес" (Random Forest), поскольку она устойчива к шуму и выбросам, хорошо работает на табличных данных без необходимости масштабирования.

Класс	Precision	Recall	F1-score
0 (не покупка)	0.80	0.78	0.79
1 (покупка)	0.58	0.61	0.59
Accuracy	—	—	**0.72**

Интерпретация:

- Модель достаточно хорошо распознаёт тех, кто не покупает товар (класс 0) — точность 80%.
- Для класса 1 (тех, кто покупает товар) модель работает удовлетворительно, но не идеально:
- F1-score = 0.59 — это компромисс между точностью и полнотой.
- Recall (60%) — модель ловит 60% потенциальных покупателей, что может быть приемлемо для маркетинга.

Общая точность — 72%, что для задачи рекомендации товаров — вполне рабочий результат.

Выводы и рекомендации:

- ✓ Модель можно использовать для приоритизации рассылок и товарных предложений: запускать кампании только на клиентов с высокой вероятностью покупки.
- ✓ Чтобы улучшить качество: учесть повторные покупки, добавить данные о прошлых реакциях на акции (например, по base_sale), использовать признаки категорий товара или бренда.
- ✓ F1-score в районе 0.59 — это хороший старт для рекомендательной модели без поведенческих меток (например, кликов, просмотров).

7. Общие выводы и рекомендации.

На основании комплексного анализа клиентских данных, маркетинговых кампаний, кластеризации и модели склонности к покупке, можно сформулировать следующие ключевые рекомендации для бизнеса:

1. Сегментированный маркетинг:

- Отказаться от массовых коммуникаций по всей базе клиентов.
- Использовать результаты кластеризации для **таргетированных предложений**:
  - Кластер 1 и 2 — хорошо реагируют на скидки и мобильные каналы.
  - Кластер 4 и 5 — ценят персональные предложения и брендовые товары.
  - Кластер 6 — подходит для кампаний через фитнес- и лайфстайл-блогеров.

2. Применение модели склонности к покупке:

- Внедрить модель Random Forest для оценки **вероятности отклика клиента**.
- Использовать модель при планировании email- и push-рассылок, чтобы:
  - снизить нагрузку на каналы,
  - не раздражать «холодных» клиентов,
  - повысить конверсию и ROI коммуникаций.
- Обновлять модель **раз в 3–6 месяцев** с учётом новых покупательских паттернов.

3. Улучшение кампаний:

- **Email-скидки** в первой маркетинговой кампании не показали статистически значимого эффекта. Рекомендуется:
  - Проводить A/B-тесты **с более чёткой сегментацией и персонализацией**,
  - Увеличить длительность отклика (например, 7 дней), отслеживая динамику.

4. Продолжение аналитики:

- Кластер 3 требует дополнительного анализа: нестабильное поведение, возможно — случайные покупки. Возможны сценарии:
  - проведение опроса,
  - запуск экспериментальных компаний с отслеживанием реакции.
- Построить **модель оттока** и **анализ жизненного цикла клиента (CLV)** для дальнейшей персонализации маркетинга.