Отчет по результатам проекта

Выполнила: Айым Темирбаева

Содержание

1. Вступление

Заказчик: крупный магазин спортивных товаров. Задача — с помощью данных о покупках клиентов и их социально-демографических признаках проанализировать эффективность уже проведенных ранее маркетинговых кампаний и выявить факторы, способные повысить продажи.

Представлены данные о покупках клиентов за два месяца. Данные хранятся в базе данных (shop_database.db).

База данных содержит три таблицы personal_data, personal_data_coeffs, purchases. Также есть csv файл (personal_data.csv) с восстановленными данными о клиентах, id клиентов участвовавших в первой маркетинговой кампании (ids_first_company_positive.txt, ids_first_company_negative.txt).

2. Исследовательский анализ данных

2.1 Обработка данных

Есть 3 таблицы в базе данных. Информация о данных в таблицах (таблица 1 и 2):

personal_data			
№	Признак	Количество	Тип данных
1.	id	89241	int64
2.	gender	89241	int64
3.	age	89241	int64
4.	education	89241	object
5.	city	89241	int64
6.	country	89241	int64

personal_data_coeff			
№	Признак	Количество	Тип данных
1.	id	104989	int64
2.	lbt_coef	104989	float64
3.	lbt_coef	104989	float64
4.	sm_coef	104989	float64
5.	personal_coef	104989	float64

Таблица 1.

Данный этап состоит из формирования набора данных, обработки пропущенных значений и нормализации данных для дальнейшего анализа. Для начала были объединены таблицы personal_data и personal_data_lost, что позволило восстановить недостающие данные о клиентах. Затем к этим данным были добавлены персональные коэффициенты personal_coef из таблицы personal_data_coeffs. В результате было получено 104 989 записей с полной информацией о возрасте, образовании, поле, стране, городе проживания и персональном коэффициенте клиентов. В процессе анализа было выявлено значительное количество пропущенных значений в колонках colour и product_sex таблицы purchases (таблица 3).

Информация о данных в файле personal_data_lost

№	Признак	Количество	Тип данных
1.	id	15748	int64
2.	age	15748	int64
3.	education	15748	object
4.	city	15748	int64
5.	country	15748	int64

purchases			
№	Признак	Количество	Тип данных
1.	id	786260	int64
2.	product	786260	object
3.	colour	786260	object
4.	cost	786260	int64
5.	product_sex	786260	float64
6.	base_sale	786260	int64
7.	dt	786260	int64

Таблица 2.

Для обработки colour было принято решение учитывать только первый цвет из списка, если он был записан через /. Также пропущенные значения были заменены на unknown, поскольку определить цвет по другим характеристикам товара невозможно. Пропущенные значения в product_sex были восстановлены с использованием информации из названия продукта. Если название содержало слова, характерные для женской одежды (девоч, женс), поле product_sex заполнялось 0 (женский). Если в названии встречались слова, относящиеся к мужской одежде (мальчик, мужс), использовалось значение 1 (мужской). В случаях, когда определить пол продукта было невозможно, проставлено значение 2 (неопределенный/унисекс). В результате обработки пропущенные значения в ключевых полях были устранены, данные нормализованы и приведены к единому формату. Это позволило подготовить их для дальнейшего этапа, бинарной классификации пола клиентов.

Поле	Количество пропущенных значений
colour	119 524
product_sex	314 712

Таблица 3.

2.2 Бинарная классификация

В ходе работы проведена бинарная классификация для предсказания пола клиентов, у которых это значение отсутствовало в исходных данных. Данный этап был выполнен с использованием модели Random Forest, что позволило достичь высокой точности предсказаний. На начальном этапе было выявлено, что в таблице personal_data содержится 104 989 записей, но информация о поле (gender) отсутствовала у 15 748 клиентов. Для решения этой проблемы использовались имеющиеся социально-демографические признаки, которые были выявлены с помощью

корреляции. На основе корреляции выбраны признаки такие как возраст (age), город (city), образование (education) и персональный коэффициент (personal_coef). До обучения была проведена обработка категориальных данных. Поле education содержало строковые значения, поэтому оно было закодировано с использованием LabelEncoder. После этого данные были разделены на две выборки:

- Обучающая выборка (train_data) – клиенты, у которых известен пол (89 241 записей).
- Тестовая выборка (test_data) – клиенты с пропущенными значениями пола (15 748 записей).

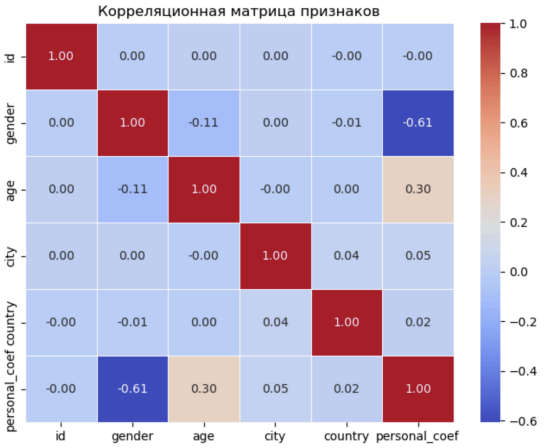

Рисунок 1. Корреляционная матрица

После подготовки данных была обучена модель Random Forest Classifier (100 деревьев), которая позволила предсказать gender на основе четырех признаков. Данные были разделены на обучающую (80%) и валидационную (20%) выборку, а затем модель была обучена и протестирована. F1-score на валидационной выборке составил 0.9999, что свидетельствует о практически идеальном качестве предсказаний. Это указывает на то, что выбранные признаки достаточны для точного предсказания пола клиентов. После успешного тестирования модель была применена к данным, где gender отсутствовал. Полученные предсказания были записаны обратно в personal_data, заполнив все пропущенные значения. Итоговая проверка показала, что теперь все 104 989 записей имеют корректные значения gender, и пропущенные данные полностью устранены. Таким образом, данный этап позволил полностью восстановить информацию о поле клиентов, что критически важно для дальнейшего анализа покупательского поведения и сегментации аудитории.

2.3 A/B тестирование

В ходе A/B-тестирования была проанализирована эффективность первой маркетинговой кампании, направленной на предоставление персональных скидок 5 000 клиентам через email-рассылку. Для сравнения были выделены две группы: тестовая (получившие скидку) и контрольная (не получавшие скидку). Оценка эффективности проводилась на основе четырех ключевых метрик: конверсия (CR), средний чек (AOV), выручка на пользователя (RPU), среднее количество заказов на пользователя (OPU)

Основные результаты:

- Конверсия (CR) в тестовой группе (0.9998) выше, чем в контрольной (0.9997), однако различие не является статистически значимым (p = 0.3543). Это говорит о том, что наличие скидки не оказало значимого влияния на само решение о покупке.

- Средний чек (AOV) в тестовой группе составил 5244.07 ₽, а в контрольной – 5372.25 ₽ ($p = 0.0258$). Клиенты, получившие скидку, совершали покупки на немного меньшую сумму, но разница статистически значима.
- Выручка на пользователя (RPU) в тестовой группе (5242.93 ₽) оказалась ниже, чем в контрольной (5370.53 ₽), различие статистически значимо ($p = 0.0265$). Это означает, что скидка не увеличила общую выручку на клиента, а скорее снизила её.
- Среднее количество заказов на пользователя (OPU) значительно увеличилось в тестовой группе (12.87 заказов против 11.21 заказов в контрольной), различие статистически значимо ($p < 0.0001$). Персональная скидка мотивировала клиентов совершать больше заказов, но с более низким средним чеком.

Скидка оказала сильное влияние на поведение покупателей: клиенты в тестовой группе совершали больше покупок, но их средний чек был ниже. Конверсия практически не изменилась, что означает, что наличие скидки не повлияло на само решение о покупке. Выручка на пользователя снизилась, что говорит о том, что тестируемый уровень скидки мог быть слишком высоким.

2.4 Кластеризация

Перед выполнением кластеризации были проведены этапы предобработки. Данные о покупках клиентов из таблицы purchases были объединены с персональными данными (personal_data) на основе id. Оставлены только клиенты, относящиеся к стране с кодовым значением 32. Это позволило исключить нерелевантные записи и сфокусироваться на нужной аудитории. Введен новый признак category, который определяет категорию товара на основе названия продукта. Для этого были выделены основные категории, такие как:

- Одежда (футболки, куртки, лосины и т.д.)
- Обувь (кроссовки, ботинки, сандалии и т.д.)
- Аксессуары (рюкзаки, шапки, перчатки и т.д.)
- Спортивный инвентарь (мячи, лыжи, велосипеды и т.д.)
- Тренажеры и утяжелители (гантели, эспандеры)
- Туризм (палатки, спальные мешки)
- Защита (шлемы, налокотники)
- Питание и добавки (протеин, витамины)

Для кластеризации была взята случайная выборка из 100 000 строк с учетом равномерного распределения данных, чтобы сократить время выполнения вычислений. Метод кластеризации был выбран K-Means. Для определения оптимального числа кластеров использованы метод локтя (Elbow Method) и силуэтный коэффициент (Silhouette Score = 0.1403). Было выбрано 3 кластера, так как это обеспечивало наилучший баланс между детализацией сегментов и интерпретируемостью (рисунок 2). Анализ кластеров клиентов магазина спортивных товаров показал, что основными группами покупателей являются мужчины среднего возраста, молодые мужчины и женщины, предпочитающие фитнес-одежду.

Первая группа – мужчины 35–50 лет, составляющие крупнейший кластер. Они приобретают в основном одежду и обувь, включая популярные бренды Demix и Outventure. Основной цвет товаров – черный, а средняя стоимость покупки составляет 2 999 рублей. Эти покупатели не так сильно ориентированы на скидки, поскольку их покупки стабильны. Для повышения продаж среди данной группы рекомендуется

продвигать спортивные бренды, запускать капсульные коллекции и вовлекать клиентов в программы лояльности вместо стандартных скидок.

Вторая группа – молодые мужчины 19–25 лет, которые демонстрируют высокую чувствительность к скидкам и акционным предложениям. Они отдают предпочтение бюджетной спортивной одежде и обуви, включая шорты и футболки брендов Demix и FILA. Основной цвет товаров – черный и белый, а наиболее частая цена покупки – 1 999 рублей. Для этого сегмента следует продвигать товары через социальные сети, запускать акции «2+1» и предоставлять скидки на комплекты спортивной одежды.

Третья группа – женщины 35–39 лет, которые ориентируются на спортивную и фитнес-одежду, а также купальники и легкие спортивные комплекты. Они чаще всего покупают товары в черном и белом цветах, а их средний чек аналогичен предыдущим кластерам – около 1 999 рублей. Женщины более активно используют скидки, поэтому рекомендуется запускать специальные фитнес-коллекции, привлекать клиентов через email-рассылки и предлагать программы лояльности с накопительными скидками.

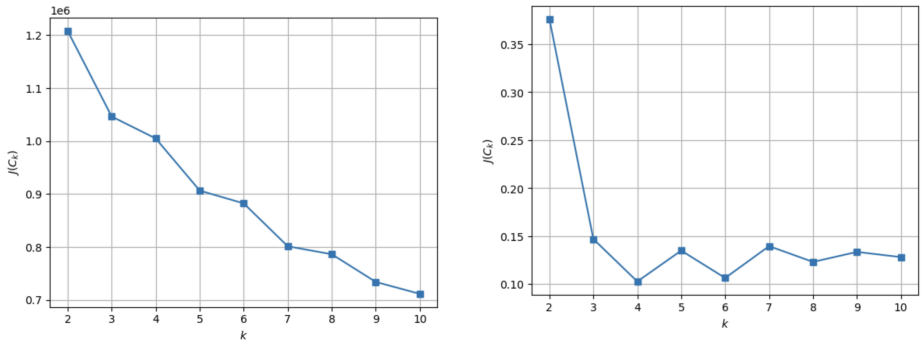

Рисунок 2. График метода локтя и силуэтного коэффициент

Общий вывод по кластеризации

1. Кластер 1 – мужчины 35-50 лет, покупают одежду и обувь, редко используют скидки.
2. Кластер 2 – молодые мужчины 19-25 лет, активно реагируют на скидки, покупают бюджетную одежду и обувь.
3. Кластер 3 – женщины 35-50 лет, покупают одежду для фитнеса и плавания, часто пользуются скидками.

Эти данные помогут магазину разработать персонализированные маркетинговые стратегии и увеличить продажи. Кластеризация позволила лучше понять аудиторию и сформировать маркетинговую стратегию для каждого сегмента, что поможет повысить продажи и оптимизировать маркетинговые затраты.

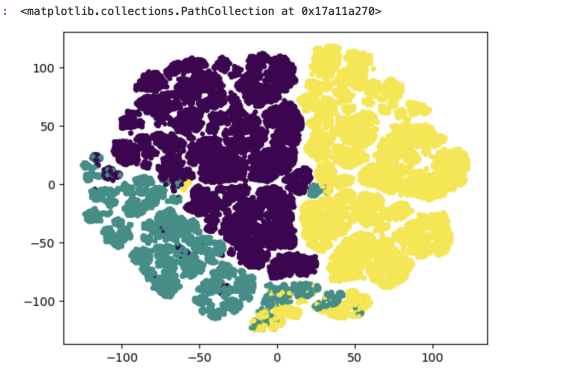

Рисунок 3. Кластеризация Kmeans

2.5 Модель склонности клиента к покупке

Перед построением модели были проведены шаги по объединению данных о клиентах (personal_data) с информацией о покупках (purchases) и добавлены записи пользователей, которые не совершали покупку (из A/B-теста). Обработаны пропуски с данных А/Б теста, добавлен новый признак – категория товара (category).

Далее были сформированы целевые переменные:

- purchase – факт покупки (1 – купил, 0 – не купил).
- category – категория товара, который клиент купил.

Так как большинство клиентов совершают покупки (99.99% положительных меток), модель была обучена модель Random Forest Classifier (100 деревьев) с параметром class_weight='balanced', чтобы учесть редкие случаи отказа от покупки.

Было проведено два обучения

- Обучение модели покупки (purchase) для выявления купит ли клиент товар.
- Обучение модели предсказания категории (category) для выявления товара какой категории купить клиент.

Результаты

- Точность предсказания покупки (purchase):
    - F1-score = 1.00 (почти идеальное предсказание).
- Точность предсказания категории (category):
    - F1-score ≈ 0.97 (хорошая способность различать категории товаров).

Согласно полученным результатам предсказания для города 1188

- Большинство клиентов были предсказаны как совершившие покупку. Это указывает на высокую склонность данной группы к покупкам.
- Категории товаров, предсказанные моделью, распределены неравномерно. Некоторые группы товаров имеют значительно большее количество предсказанных покупок.

Анализ категорий товаров, которые с наибольшей вероятностью будут куплены, показывает следующее распределение:

- Одежда – 40 296 предсказанных покупок.
- Обувь – 19 565 предсказанных покупок.
- Другое (прочие товары, не вошедшие в основные категории) – 24 962 предсказанных покупок.
- Аксессуары – 1 218 предсказанных покупок.
- Спортинвентарь – 2 696 предсказанных покупок.
- Тренажеры и утяжелители – всего 45 предсказанных покупок.
- Туризм – 92 предсказанных покупки.
- Велосипедные аксессуары, защита и питание – предсказаны в минимальном количестве.

Это говорит о том, что ключевой интерес клиентов сосредоточен на одежде, обуви и спортивных аксессуарах. А категории «тренажеры и утяжелители», «туризм», «велосипедные аксессуары» имеют небольшие показатели предсказанных покупок.

3. Рекомендации

1. Выявление факторов, способных повысить продажи

На основе проведенного анализа можно выделить несколько ключевых факторов, которые влияют на продажи:

- **Социально-демографические характеристики клиентов**
  Кластеризация показала, что мужчины 35–50 лет (кластер 0) и женщины 35–50 лет (кластер 2) составляют основную покупательскую аудиторию. Они совершают покупки чаще, чем молодые покупатели 19–25 лет (кластер 2), но менее чувствительны к скидкам.
- **Категория покупаемых товаров**
  Одежда и обувь являются основными драйверами продаж, особенно среди мужчин и женщин среднего возраста. Спортивные аксессуары и инвентарь также востребованы, но в меньшей степени.
- **Влияние скидок**
  A/B тестирование показало, что скидки приводят к увеличению количества заказов, но при этом снижается средний чек и выручка на пользователя. Это говорит о том, что важно не просто предлагать скидки, а грамотно управлять их размером.
- **Персонализированный подход**
  Кластеризация клиентов позволяет точнее нацеливать маркетинговые кампании. Например, молодым клиентам 19–25 лет (кластер 2) лучше предлагать скидки и акции, а мужчинам среднего возраста (кластер 1) – программы лояльности вместо стандартных скидок.
- **Ценовая политика**
  Анализ показал, что наиболее популярные товары имеют стоимость в диапазоне 999–5 999 рублей. Продвижение товаров в этом ценовом сегменте может увеличить продажи.

2. Оценка запуска маркетинговой кампании в городе 1188

Модель склонности к покупке показала, что клиенты из города 1188 имеют высокую вероятность совершения покупок. Основные предсказанные категории товаров:

- Одежда (40 296 покупок)
- Обувь (19 565 покупок)
- Другое (24 962 покупок) – включает прочие спортивные товары
- Аксессуары (1 218 покупок)
- Спортивный инвентарь (2 696 покупок)
- Туризм (92 покупки)
- Тренажеры и утяжелители (45 покупок)

Это говорит о том, что запуск маркетинговой кампании в этом городе оправдан, но с акцентом на одежду и обувь, так как они имеют наибольший спрос. Категории

«тренажеры и утяжелители», «туризм», «велосипедные аксессуары» не пользуются высокой популярностью, поэтому их продвижение может не принести ожидаемого эффекта.

3. Рекомендации для повышения продаж

На основе анализа можно предложить следующие рекомендации:

- **Сегментация аудитории и персонализированные предложения**
  Использовать данные кластеризации для таргетированной рекламы.
  - Для мужчин 35–50 лет (кластер 1) – предлагать программы лояльности, коллекции брендовой одежды и обуви.
  - Для молодых мужчин 19–25 лет (кластер 2) – акции «2+1», скидки на комплекты.
  - Для женщин 35–50 лет (кластер 3) – запускать специальные фитнес-коллекции и email-рассылки с персональными предложениями.
- **Оптимизация скидочных предложений**
  A/B тестирование показало, что скидки стимулируют покупки, но снижают средний чек. Рекомендуется тестировать разные уровни скидок (например, 5%, 10%, 15%) и анализировать их влияние на выручку.
- **Развитие программ лояльности**
  Предложение бонусных баллов вместо скидок для постоянных клиентов поможет удерживать аудиторию без сильного снижения среднего чека.
- **Продвижение популярных товаров**
  Основные категории покупок – одежда, обувь и аксессуары. Продвижение этих категорий в рекламе (например, таргетированная реклама в соцсетях) может увеличить конверсию.
- **Локальные маркетинговые кампании**
  Для города 1188 рекомендуется запуск персонализированной кампании с акцентом на одежду и обувь. Использование онлайн-рекламы, email-рассылок и социальных сетей поможет повысить охват.

4. Заключение

Проведенный анализ позволил выявить ключевые факторы, влияющие на продажи в магазине спортивных товаров. Было установлено, что мужчины и женщины среднего возраста являются основными покупателями, а молодые люди 19–25 лет более чувствительны к скидкам. A/B тестирование показало, что скидки увеличивают количество заказов, но снижают средний чек. Кластеризация позволяет сегментировать аудиторию и предложить персонализированные маркетинговые стратегии. Модель склонности к покупке подтвердила, что город 1188 является перспективным регионом для запуска маркетинговой кампании, особенно в сегментах одежды и обуви. Использование предложенных стратегий поможет магазину повысить продажи и оптимизировать маркетинговые затраты.