**1 Аналитический отчёт**

**Шарипов Ильнур - Итоговый проект курса «Data Analyst. Junior»**

****Проект: Анализ эффективности маркетинговых кампаний и сегментация клиентов для магазина спортивных товаров****

****1. Введение****

* ****Цель проекта:**** Проанализировать эффективность проведённых маркетинговых кампаний, сегментировать клиентскую базу и построить модель для прогнозирования склонности к покупке.
* ****Исходные данные:**** Предоставлены данные о покупках клиентов за два месяца в виде БД shop\_database.db и дополняющего CSV-файла personal\_data.csv.gz.
* ****Задачи:****
  1. Предобработка и объединение данных.
  2. Восстановление пропущенных значений пола клиента с помощью модели бинарной классификации.
  3. Оценка эффективности первой маркетинговой кампании методом A/B-тестирования.
  4. Кластеризация клиентов для выявления сегментов аудитории.
  5. Построение модели склонности к покупке для жителей города 1188.

****2. Предобработка данных****

* ****Объединение данных:**** Таблицы personal\_data, personal\_data\_coeffs и purchases объединены по id. Данные из personal\_data.csv.gz были добавлены для заполнения пропусков.
* ****Фильтрация:**** Оставлены только клиенты из страны с кодом 32, как указано в ТЗ.
* ****Обработка пропусков:****
  + Пропуски в признаке gender заполнены с помощью модели (задача 2).
  + Пропуски в признаке colour заполнены константой «другой».
  + Все строки в которых были пропуски по столбцам city, education, age, gender и частично в product\_sex, colour - удалил, так как много пропущенных элементов в одной строке, таких строк насчитывалось 6143.
  + Пропуски в признаке product\_sex заполнил на основе названия продукта, там присутствуют слова мужской, женский, для мальчиков, для девочек, детский и т.д. Остальные пропуски заполню цифрой 3, так как есть товары без гендерной принадлежности. Создал функцию для определения слов, далее применил к строкам где product\_sex is NaN.
* ****Обработка текстовых полей:****
  + Поле colour обработал таким образом, сгруппировал цвета, отсортировал по убыванию и оставил первые 50 строк, к остальным цветам применил константу «другой».
  + В поле product уменьшил число уникальных названий товара до 7 тысяч, было 23 тысячи. Создал функцию которая в названии оставляет только первые три слова.
* ****Итог:**** Сформирован очищенный датафрейм, готовый для дальнейшего анализа.

****3. Бинарная классификация (Определение пола клиента)****

* ****Цель:**** Восстановить пропущенные значения в столбце gender (0/1).
* ****Методология:****
  + ****Признаки:**** Использованы следующие признаки 'colour', 'cost', 'product\_sex', 'base\_sale', 'age', 'education', 'city', 'personal\_coef'. У каждого пола свой коэффициент, и признак personal\_coef очень хорошо поможет для прогноза пола клиента, возможно для этого добавили эту фичу.
  + ****Модель:**** Случайный лес (RandomForestClassifier)
  + ****Валидация:**** Применил кросс-валидацию и поделил на три фолда.
* ****Результат:****
  + На тестовой выборке достигнута ****F-мера = 1.0**** (без признака personal\_coef ****F-мера = 0.77****).
  + Модель показала высокую точность предсказания. Восстановленные значения были использованы для заполнения пропусков в итоговом наборе данных.

****4. A/B-тестирование первой маркетинговой кампании****

* ****Исходная гипотеза:**** Предоставление персональной скидки (тестовая группа, ids\_first\_company\_positive.txt) приводит к ****увеличению частоты покупок**** и ****росту общей выручки****  по сравнению с контрольной группой (группа B, ids\_first\_company\_negative.txt).
* ****Анализ:****  
  Для оценки эффективности кампании были рассчитаны и проанализированы следующие метрики:
  + ****Среднее количество покупок на пользователя**** (ключевая метрика)
  + ****Средняя выручка на пользователя**** (ключевая метрика)
  + ****Средний чек**** (вторичная метрика для глубины анализа)
  + ****Общая выручка по группе****

Для проверки статистической значимости различий по всем метрикам был проведён тест Манна-Уитни (распределения метрик не являются нормальными, выборки не зависимы).

* ****Ключевые результаты:****
  + ****Частота покупок:**** Наблюдается ****статистически значимое увеличение на 15%**** в тестовой группе (p-value < 0.01). Это означает, что кампания успешно стимулировала клиентов совершать покупки чаще.
  + ****Выручка на пользователя:**** Наблюдается ****статистически значимое увеличение на 11.6%**** в тестовой группе (p-value < 0.01). Это прямое следствие увеличения частоты покупок и ключевой показатель роста доходов.
  + ****Средний чек:**** Наблюдается ****статистически значимое, но небольшое снижение на 3%**** (p-value ≈ 0.022). Это указывает на сдвиг в поведении: клиенты стали покупать более часто, но менее дорогие товары.
  + ****Общая выручка:**** ****Рост на 35.6 млн рублей**** в тестовой группе является финальным доказательством финансовой эффективности кампании.
* ****Итоговый вывод:****  
  ****Отвергается нулевая гипотеза и применяется альтернативная гипотеза.**** Маркетинговая кампания является статистически и экономически эффективной. Она достигла своей цели, значимо увеличив ключевые для бизнеса метрики: ****частоту покупок на 15%**** и ****выручку на пользователя на 11.6%****.
* ****Бизнес-рекомендации:****
  + ****Признать кампанию успешной и масштабировать данную стратегию.**** Основной фокус при её масштабировании должен быть на ****увеличении частоты покупок****.
  + ****Оптимизировать коммуникацию**** для смещения поведения клиентов в сторону более крупных покупок:
    - Внедрить механики ****перекрёстных продаж**** («клиенты, купившие этот товар, также покупают...»).
    - Использовать ****персональные предложения**** на товары более высокого ценового сегмента для клиентов, которые начали покупать чаще.
    - Рассмотреть возможность установления ****минимального порога корзины**** для действия скидки в будущем.
  + ****Проанализировать, какие именно товары стимулировали рост частоты покупок.**** Это поможет точечно предлагать скидки на аналогичные товары другим сегментам клиентов и ещё больше увеличить эффективность будущих кампаний.

****5. Кластеризация клиентов****

* ****Цель:**** Выявить сегменты клиентов для разработки персональных маркетинговых стратегий.
* ****Методология:****
  + ****Признаки:**** RFM-метрика (Recency - дней с последней покупки, Frequency - количество покупок, Monetary - общая сумма покупок), социально-демографические данные (возраст, город), доля покупок со скидкой.
  + ****Алгоритм:**** K-Means (на основе метода локтя было выбрано оптимальное количество кластеров - 7).
  + ****Размер данных:**** Для ускорения работы алгоритма использована случайная подвыборка (100 000 строк).
* ****Результаты и рекомендации:****

****Кластер 0 (213516 покупок-самые большие покупки):****

****- преобладают мужчины (65%), с возрастом 38 лет и средним образованием (82%).****

****- с высоким средним чеком 7088, низкая чувствительность к базовой скидке (19%).****

****- предпочитают товары без гендерной принадлежности (100%) и чаще велосипед горный stern.****

*Использовать программы лояльности вместо скидок, предлагать премиум товары без гендерной принадлежности.*

****Кластер 1 (113663 покупок):****

****- преобладают мужчины (72%), с возрастом 39 лет и средним образованием (81%).****

****- со средним чеком 4083, полная чувствительность к базовой скидке (100%).****

****- предпочитают только мужские товары (100%) и чаще кроссовки мужские puma.****

*Рекламу настраивать на мужской пол выше среднего возраста и средним образованием, товары знаменитых брендов, использовать программы скидки и кэшбэки.*

****Кластер 2 (103104 покупок):****

****- преобладают женщины (81%), с возрастом 41 года и средним образованием (87%).****

****- с низким средним чеком 3205, полная чувствительность к базовой скидке (100%).****

****- предпочитают исключительно женские товары (100%) и чаще кроссовки для девочек.****

*Рекламу настраивать на женский пол ниже среднего возраста и средним образованием, использовать программы скидкок и кэшбэка.*

****Кластер 3 (106321 покупок):****

****- преобладают женщины (81%), с возрастом 42 года и средним образованием (90%).****

****- со средним чеком 4916, отсутствует чувствительность к базовой скидке (0%).****

****- предпочитают исключительно женские товары (100%) и чаще кроссовки женские demix.****

*Рекламу настраивать на женский пол ниже среднего возраста и средним образованием, использовать программы лояльности вместо скидок.*

****Кластер 4 (179704 покупок):****

****- преобладают мужчины (74%), со средним возрастом 40 лет и средним образованием (82%).****

****- со средним чеком 5947, отсутствует чувствительность к базовой скидке (0%).****

****- предпочитают мужские товары (99.9%) и чаще кроссовки мужские nike.****

*Рекламу настраивать на мужской пол выше среднего возраста и средним образованием, использовать программы лояльности вместо скидок.*

****Кластер 5 (18314 покупок - самые маленькие покупки):****

****- преобладают женщины (65%), со средним возрастом 38 лет и средним образованием (85%).****

****- с низким средним чеком 3357, умеренная чувствительность к базовой скидке (33%).****

****- предпочитают детские товары (100%) и чаще сабо детские crocs.****

*Рекламу настраивать на женский пол ниже среднего возраста и средним образованием, предлагать бюджетный сегмент детских товаров, использовать умеренные скидки и кэшбэки.*

****Кластер 6 (45491 покупок):****

****- преобладают мужчины (69%), со средним возрастом 16 лет и высшем образованием (81%).****

****- со средним чеком 5741, умеренная чувствительность к базовой скидке (33%).****

****- предпочитают мужские(41%), женские(31%), неопределённые(28%) товары и чаще кроссовки мужские nike.****

*Предлагать детские товары детям до 18 лет, использовать умеренные скидки и кэшбэки.*

****6. Модель склонности к покупке****

* ****Цель:**** Построить модель склонности клиента к покупке определённого товара при коммуникации
* ****Методология:****
  + ****Целевая переменная:**** Таргет выбирал таким образом, если товар покупался больше медианного значения то таргет будет 1, в противном случае 0.
  + ****Признаки:**** Данные о профилях клиентов, данные товаров и данные о прошлых маркетинговых кампаниях.
  + ****Модель:**** Была обучена модель **Дерево решений** для предсказания склонности к покупке.
* ****Результат:**** Построена модель, которая предсказывает склонность к покупке определённый товар **для клиентов из города 1188**. Для запуска новой кампании рекомендуется выбрать топ-10 товаров с наибольшей предсказанной вероятностью, что позволит оптимизировать маркетинговый бюджет.

| **Количество покупок предсказанных** | **Имя товара** |
| --- | --- |
| 905 | велосипедки мужские asics |
| 894 | велосипедки мужские craft |
| 816 | велосипедки женские freddy |
| 747 | велосипед детский унисекс |
| 634 | велосипед подростковый scott |
| 605 | велосипедки женские odlo |
| 524 | велосипед детский трехколесный |
| 502 | велосипед шоссейный polygon |
| 469 | ветровка женская shu |
| 464 | велосипед подростковый trek |

**Анализ покупок города 1188:**

**- немного больше преобладают мужчины (57%), со средним возрастом 39 лет и средним образованием (81%).**

**- со средним чеком 5231, умеренная чувствительность к базовой скидке (35%).**

**- предпочитают разнообразные товары(мужские-41%, женские - 28%, неопределённые - 29%) и чаще кроссовки мужские puma.**

****7. Анализ второй маркетинговой компании****

**- немного больше преобладают мужчины (56%), со средним возрастом 39 лет и средним образованием (84%).**

**- со средним чеком 5800, умеренная чувствительность к базовой скидке (35%).**

**- предпочитают разнообразные товары, кроме детских (мужские-40%, женские - 29%, неопределённые - 30%, детские-2%) и чаще кроссовки мужские nike.**

**Вторая маркетинговая кампания принесла 9896 покупки, покупки увеличились на 1.4 процента.**

Вторая маркетинговая кампания принесла 57.9 млн выручки, выручка увеличилась на 1.51 процента.

****Вторая маркетинговая компания была эффективна, на 1.5% увеличилась выручка и на 1.4% увеличились покупки.****

****8. Заключение****  
Проведённый анализ позволил дать конкретные бизнес-рекомендации по оптимизации маркетинговых активностей, сегментировать аудиторию и создать инструмент для таргетирования будущих кампаний. Ключевой вывод: все маркетинговые кампании эффективны.