Собиржонов Акбар

Итоговый проект по «Data Analyst Junior»

Введение:

Цель проекта:

Цель проекта — на основе клиентских данных и истории покупок провести анализ эффективности маркетинговых кампаний, выявить закономерности в поведении покупателей и построить модель склонности к покупке.

Анализ охватывает пять ключевых этапов:

- предобработка данных,
- бинарная классификация,
- A/B-тестирование,
- кластеризация,
- модель склонности к покупке.

Описание данных:

Исходные данные включали три таблицы:

- `personal_data`: ID клиентов, их пол, возраст, образование, страна и город.
- `personal_data_coeffs`: персональные коэффициенты клиентов (`personal_coef` и др.).
- `purchases`: информация о покупках (товар, цвет, стоимость, скидка, дата).

Дополнительно были предоставлены:

- Утерянные данные о клиентах в архиве (`personal_data.csv.gz`).
- Списки клиентов, участвовавших в первой маркетинговой кампании (`ids_first_company_positive.txt` и `ids_first_company_negative.txt`).

Для анализа отобраны только клиенты из страны с кодом 32, как было указано в задании.

1. Предобработка данных

Исходные данные были получены из базы данных `shop_database.db`, а также файлов `personal_data.csv.gz`, `ids_first_company_positive.txt` и `ids_first_company_negative.txt`.
Выполнены следующие действия:

- Оставлены клиенты из страны с кодом `32`;
- Объединены данные из таблиц и файла с пропущенной информацией;
- Приведены к единому формату наименования товаров и цветов (разделённые через `/` цвета были нормализованы);
- Удалены явные дубликаты и устранены пропуски (например, пропуски в возрасте заменялись медианой);
- Для удобства анализа добавлены признаки: наличие скидки, дата покупки в числовом формате и др.

2. Бинарная классификация

Для восстановления пола клиентов в утерянной части данных была обучена модель бинарной классификации. Использовались полные данные, где пол был известен.
Основные этапы:

- Объединение признаков из таблицы `personal_data_coeffs` и других источников;
- Кодирование категориальных признаков (например, образование и город);
- Разделение на тренировочную и тестовую выборки;
- Использовалась модель `RandomForestClassifier`, которая показала высокое качество на тестовых данных (AUC ROC > 0.85).
  Модель применялась для предсказания пола в данных, где он отсутствовал.

3. A/B-тест первой маркетинговой кампании

Для оценки эффективности персональной email-кампании:

- `Группа A` — 5000 клиентов, получивших персональную скидку (из файла `ids_first_company_positive.txt`);
- `Группа B` — аналогичные клиенты, но без скидки (из `ids_first_company_negative.txt`).
  Проведены расчёты:
- Конверсия: отношение количества покупателей к общему числу пользователей;
- Тест на пропорции (Z-тест): показал статистически значимое превосходство конверсии в группе A;
- Значение p-value < 0.05, эффект подтверждён.
  Вывод: Кампания эффективна. Рекомендуется использовать подобную персонализацию в будущем.

4. Кластеризация клиентов

Цель — определить группы клиентов с похожими характеристиками.
Методы:

- Признаки: возраст, коэффициент `personal_coef`, активность, предпочтения по товарам и цветам;
- Использовалась стандартизация и метод `KMeans` с определением оптимального числа кластеров через `elbow method`;
- Выделено **4 кластера**.

Анализ кластеров:

- **Кластер 0**: Молодые, активные, часто покупают со скидками, склонны к импульсивным покупкам.
- **Кластер 1**: Старшая возрастная группа, чаще покупают товары для мужчин, предпочитают товары без скидок.
- **Кластер 2**: Активные пользователи с высоким `personal_coef`, склонны к дорогим покупкам.
- **Кластер 3**: Малозаметная активность, редкие покупки, чаще — с большой скидкой.

Рекомендации:

- Кластеру 0 эффективна персонализация и быстрые акции.
- Кластер 2 — целевая аудитория для премиум-товаров.
- Кластер 3 — имеет смысл активировать через дополнительные стимулы или лояльность.

5. Модель склонности клиента к покупке (для города 1188, страна 32)

Задача — предсказать вероятность покупки товара клиентом из города 1188 в стране 32.
Использовались:

- Данные профиля клиентов,
- История покупок,
- Характеристики товаров (включая цвет, скидку, гендерную ориентацию товара).
  Построена модель `XGBoostClassifier`, показавшая:
- AUC ROC > 0.80 на валидационной выборке;
- Основные влияющие признаки: наличие скидки, пол, возраст, коэффициент `personal_coef`, и тип товара.

Вывод: Модель пригодна для использования в таргетированных маркетинговых рассылках.

Общие выводы и рекомендации:

1. Персональные email-кампании доказали свою эффективность.
2. Разделение клиентов на кластеры позволяет точнее настраивать коммуникацию.
3. Модель склонности даёт возможность экономить бюджет, отправляя сообщения только заинтересованным клиентам.
4. Город 1188 — перспективное направление для запуска следующей кампании.

Рекомендуемые шаги:

- Использовать модель склонности в связке с email-рассылкой.
- Тестировать разный контент для разных кластеров.
- Повторно провести A/B-тест на новых городах.