A/B тестирование, кластеризация и методы машинного обучения.

Итоговый проект курса DA JUN он-лайн академии Skillbox

Исполнитель: Корнилов Иван
Екатеринбург, 2024
https://github.com/ivanKornilov

Задачи исследования

1. Провести исследовательский анализ данных:
   - Преобразить данные (Загрузить из БД, файлов)
   - Восстановить утерянные значения (в том числе данные по полу клиента, а их 15%)
2. Провести A/B тестирование по 1 маркетинговой программе, сделать выводы
3. Построить кластерный анализ предоставленных данных, составить описания полученных кластеров клиентов, сделать предложения по работе с группами клиентов
4. Построить модель склонности клиента по определенному городу

Задачи исследования

1. Провести исследовательский анализ данных:
   - Преобразить данные (Загрузить из БД, файлов)
   - Восстановить утерянные значения (в том числе данные по полу клиента, а их 15%)
2. Провести A/B тестирование по 1 маркетинговой программе, сделать выводы
3. Построить кластерный анализ предоставленных данных, составить описания полученных кластеров клиентов, сделать предложения по работе с группами клиентов
4. Построить модель склонности клиента по определенному городу

1. Исследовательский анализ данных:

Проблемы:

- Данные из базы данных и из файлов
- Неоднородность данных (наименование товара)
- Пропуски данных

Как будем решать.

1. Исследовательский анализ данных:

Проводим предобработку данных и заполняем пропуски

1. Из БД и файлов создаем датафреймы
2. Объединяем датафреймы чтобы получить наиболее полные наборы признаков (Клиенты df_clients, Покупки df_purchases)
3. Заполняем утерянные данные, в т.ч. по полу клиентов. С помощью подбора наилучших гиперпараметров (исп. валидационные данные) **методом бинарной классификации** и F-метрики вибираем метод Градиентного бустинга GradientBoostingClassifier() и далее предсказываем по данным пол клиента. Точность модели (по реальным данным от куратора 99,99%)
4. Проверяем на наличие дубликатов в датафреймах.
5. Заполняем и корректируем поля (заполняем пропуски)

Задачи исследования

1. Провести исследовательский анализ данных:
   - Преобразить данные (Загрузить из БД, файлов)
   - Восстановить утерянные значения (в том числе данные по полу клиента, а их 15%)
2. Провести A/B тестирование по 1 маркетинговой программе, сделать выводы
3. Построить кластерный анализ предоставленных данных, составить описания полученных кластеров клиентов, сделать предложения по работе с группами клиентов
4. Построить модель склонности клиента по определенному городу

2. A/B тестирование:

- Данные для теста: ID клиентов, их пол, возраст, образование, страна и город проживания, данные с персональными коэффициентами клиентов, которые рассчитываются по некоторой закрытой схеме, данные о покупках (ID покупателя, название товара, цвет, стоимость, гендерная принадлежность потенциальных покупателей товара, наличие скидки)

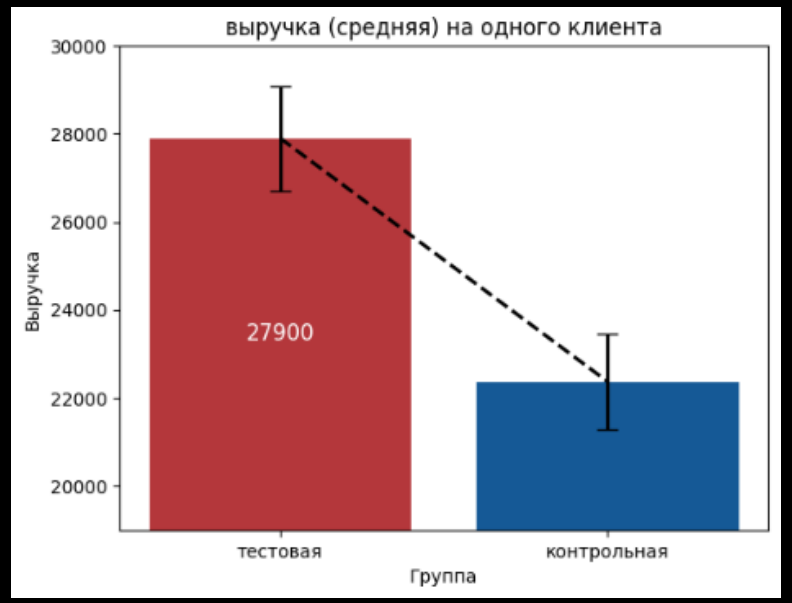
Данные были разделены по датам проведения, берем именно данные первой маркетинговой компании (dt с 5 до 17).

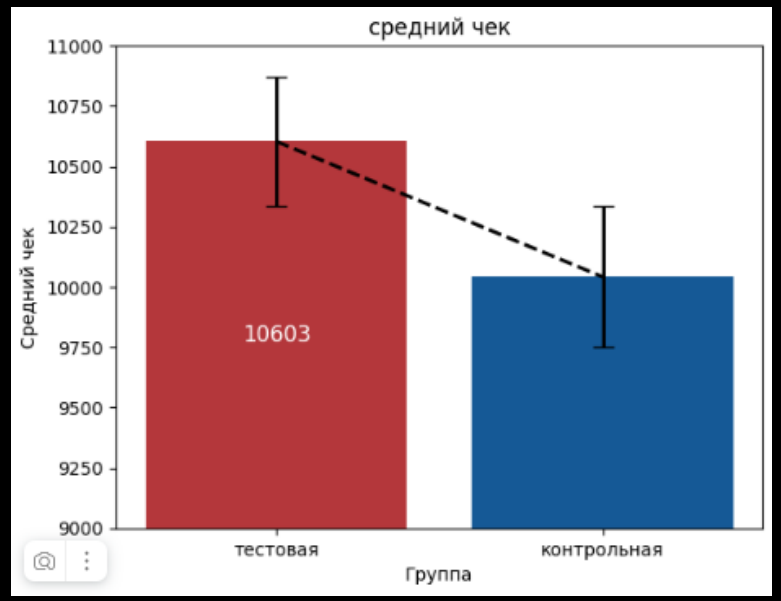
Выбираем метрики для A/B тестирования:

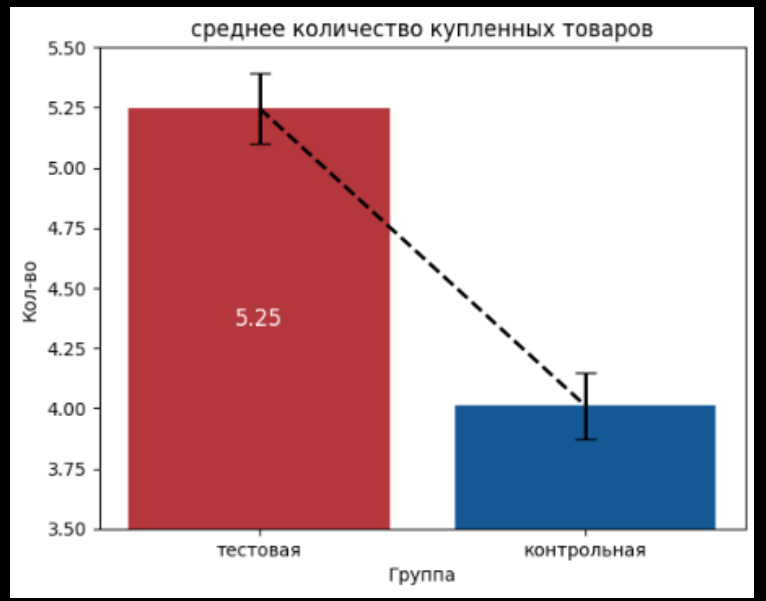
1. средняя выручка
2. средний чек
3. количество покупок

2. A/B тест: проверка гипотез равенства средних

По результатам проведения тестирования выявлена значимая разница во всех трех показателях (на уровне значимости 0.05), выбранных для оценки эффективности A/B тестирования в расчете на одного клиента - средняя выручка в тестовой группе выше на 25% (27900 и 22363), средний чек выше на 6% (10603 и 10042), среднее количество купленных товаров выше на 31% (5.25 и 4.01). (т.к. распределение во всех трех случаях не нормальное (по тесту шапиро), то используя критерий Манна-Уитни, проверяли гипотез о разницы средних).







Вывод - маркетинговая компания №1 была эффективна.

Задачи исследования

1. Провести исследовательский анализ данных:
   - Преобразить данные (Загрузить из БД, файлов)
   - Восстановить утерянные значения (в том числе данные по полу клиента, а их 15%)
2. Провести A/B тестирование по 1 маркетинговой программе, сделать выводы
3. Построить кластерный анализ предоставленных данных, составить описания полученных кластеров клиентов, сделать предложения по работе с группами клиентов
4. Построить модель склонности клиента по определенному городу

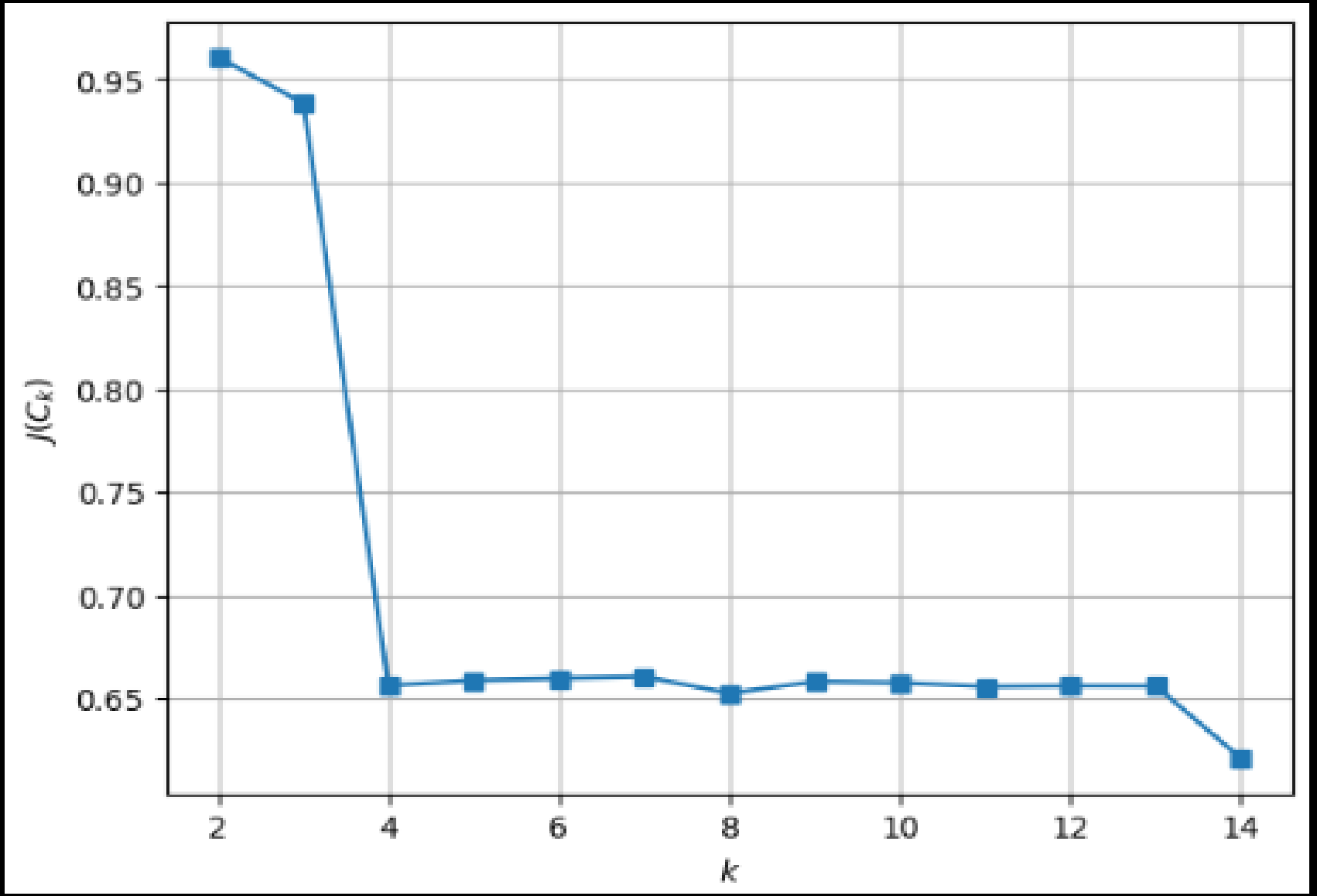
3. Кластеризация:

Были выбраны следующие признаки для кластеризации:

- Пол клиента (поле gender)
- Возраст клиента (поле age)
- Уровень образования (есть ли высшее 1/0)
- Пол товара (М/Ж/Универсальный 1/0/2)
- Есть ли скидка у клиента (признак наличия скидки у клиента)
- Сумма расходов (новый признак)
- Максимальная дата покупки (новый признак)

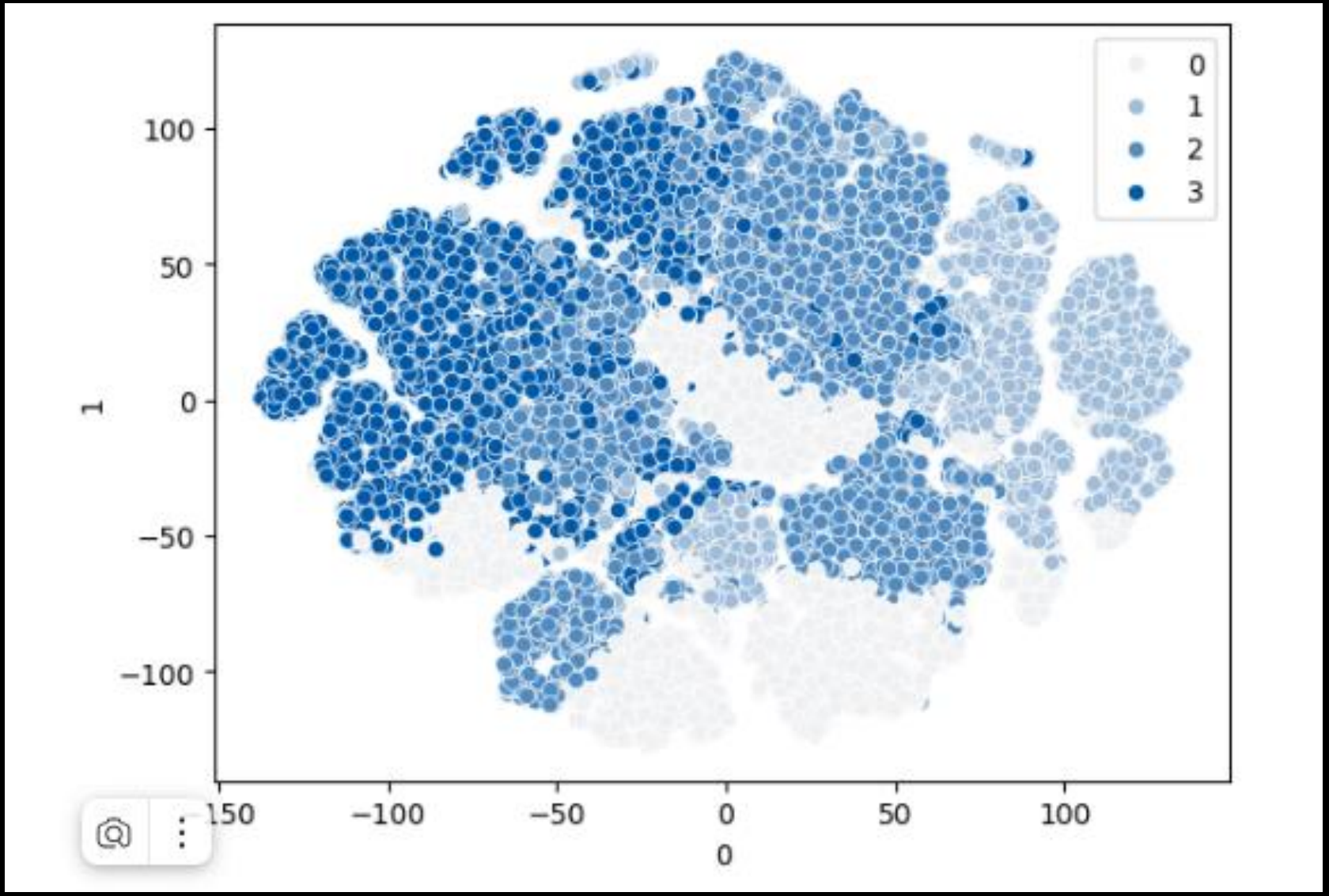
3. Кластеризация:

По методу локтя нашел что хорошо разбивается на 4 кластера



3. Кластеризация: Итого

Используем расширенный метод К-средних Kprototypes() для разбиения клиентской базы на следующие 4 кластера



3. Кластеризация: Итого

клиентская база разбивается на следующие 4 кластера:

	Кластер 1	Кластер 2	Кластер 3	Кластер 4
Возраст (средний)	39	21	42	42
Сумма расходов на 1 клиента	28269	43225	53680	32208
Скидки (да/нет) в %	15/85	35/65	21/79	61/39
Максимальная дата покупки (дни)	21	43	50	43
Образование (выс/сред) в %	13/87	85/15	4/96	6/94
Пол (муж/жен) в %	63/37	72/28	69/31	73/27
Товары для какого пола покупают	универсальные	мужского	универсальные	женского
Доля от всех клиентов в %	6	30	26	38

3. Кластеризация: Итого клиенты разбились на группы

1. Это возрастная группа (средний возраст 39 лет), это в основном мужчины (63%) со среднем образованием (13%) покупают без скидок (15%) минимальным средней суммой расходов (28 тысяч рублей) и очень часто (21 день с даты последней покупки).

Программа лояльности для часто покупающих

2. Это молодежь (средний возраст 21 год), и имеют в основном высшее образование (85%). В основном мужчины (72%). Средняя сумма расходов на одного клиента (43 тысяч рублей).

Предложить скидки, маркетинговая компания для молодых мужчин.

3. Это возрастная группа (42 года) редко покупают (50 дней с последней покупки), среднее образование (4%), средние расходы самые высокие (53 тысячи рублей)

Предложить те товары, которые они редко на большую сумму покупают.

4. Это в основном женщины (73%), среднее образование (6%), покупают все по скидкам (61%) и средние расходы (32 тысячи рублей)

Предложить скидки, маркетинговая компания для женщин.

Задачи исследования

1. Провести исследовательский анализ данных:
   - Преобразить данные (Загрузить из БД, файлов)
   - Восстановить утерянные значения (в том числе данные по полу клиента, а их 15%)
2. Провести A/B тестирование по 1 маркетинговой программе, сделать выводы
3. Построить кластерный анализ предоставленных данных, составить описания полученных кластеров клиентов, сделать предложения по работе с группами клиентов
4. Построить модель склонности клиента по определенному городу

4. Модель склонности клиента к покупкам товара 1188 город:

Берем данные по клиентам из второй маркетинговой компании по городу 1188.

С помощью модели случайного леса с результатом f1-score 11% на обучающей выборке был сделан прогноз что именно они будут покупать.

После применения модели были получены следующие товары, к покупке которых имеется склонность:

1. кроссовки 11346
2. кеды 519
3. бейсболка 378
4. брюки 195

5. Итого исследования

Спасибо! Вопросы?