1. **Подготовка данных для выявления данных кто участвовал в рекламной компании.**

Исходные файлы в текстовом формате. Загружаем файлы в пайтон и удаляем все буквы в файле , т. к. нам нужны только айди клиентов. Разделяем данные по следующим значениям : запятая, точка с запятой, новая строка и пробел. Удаление строк, которые не являются числами. Получаем датафрейм размерностью – 5022 строк кто участвовал в акции и 5019 строк кто не участвовал в акции.

2. **Загружаем данные.**

3. **Удаление пропусков и приведение название продукта к одному формату**

Пропуски в product_sex заполняем 2 – тип товара унисекс.

Цвет товара разбиваем по разделителю / и создаем новую колонку как число цветов товара. Колонку с цветом удаляем.

Наименование товара разбиваем по пробелу . В новый столбец product1

Добавляем первый элемент из названия товара. Старую колонку с наименованием товара удаляем.

Рекомендации по составлению наименования товара: не использовать обратный слеш при разделении наименования товара и первым словом писать группу товара.

4. **Модель для прогноза пола клиентов.**

При передаче данных выяснилось, что часть информации о клиентах из таблицы personal_data была утеряна.

Необходимо построить модель классификации на полных данных, чтобы, соответственно, восстановить утерянные.

Так как для предсказания используются малоинформативные данные , такие как возраст, образование и город проживания , то предсказания не могут быть точными. Использовалось три модели для предсказания пола. В итоге получились следующие результаты :

Accuracy логистической регрессии на трейне = 0.5734936287379138
на тесте = 0.5681470137825421

Accuracy случайного леса на трейне = 0.5753505794967023
на тесте = 0.5673999925297875

Accuracy многослойного персептрона трейне = 0.5773195876288659
на тесте = 0.5676614499682516

Все три модели имеют схожий результат , я взял для предсказания логистическую регрессию.

5. A/B-тестирование.

Первая кампания проводилась в период с 5-го по 16-й день.. Эта кампания включала в себя предоставление персональной скидки 5 000 клиентов через email-рассылку.

- создаем датафреймы для тех кто участвовал в акции и не участвовал.

- результат:

средний чек на одного покупателя, участвующих в акции = 16271.838038
общая сумма покупок покупателей, участвующих в акции = 48424990
количество покупателей, участвующих в акции = 2976

средний чек на одного покупателя, не участвующих в акции = 15473.838481
общая сумма покупок покупателей, не участвующих в акции = 43589803
количество покупателей, не участвующих в акции = 2817

Проводим тест Шапиро-Уилка для проверки распределения на нормальность. Результаты теста, что данные не являются нормально распределенные.

Провели тест Манна-Уитни для не нормально распределенных выборок.

Отвергаем нулевую гипотезу распределение выборок статистически значимо различаются.

Вывод: акцию можно признать удавшейся , т.к. есть изменения по всем выше пере численным показателям.

Вторая кампания проводилась на жителях города 1 134 и представляла собой баннерную рекламу на билбордах: скидка всем каждое 15-е число месяца (15-й и 45-й день в нашем случае).

- создаем датафреймы для для жителях города 1 134 и формируем их кто купил и кто не купил.

- результат:

средний чек на одного покупателя, участвующих в акции = 9497.335714
общая сумма покупок покупателей, участвующих в акции = 18614778
количество покупателей, участвующих в акции = 1960

средний чек на одного покупателя, не участвующих в акции = 10220.941
общая сумма покупок покупателей, не участвующих в акции = 10220941
количество покупателей, не участвующих в акции = 1000

Провели тест Манна-Уитни для не нормально распределенных выборок.

Отвергаем нулевую гипотезу: распределение выборок статистически значимо различаются.

Вывод: акцию можно признать удавшейся , т.к. средний чек не сильно отличается , а два других показателя имеют лучшие показатели у тех кто участвовал в акции .

6. **Кластеризация**.

Для кластерного анализа рассматриваем критерий personal_coef.
Используем метод KMeans. С помощью методов ‘локтя и ‘силуета’ принимаем решение задать 4 кластера, но на графике распределения значений personal_coef по кластерам видно, что два кластера схожи и можно выделить 3 кластера.

Осуществляем разбиение на 3 кластера , используя метод KMeans.
С использованием графика выделяем значения трех кластеров , а именно :

personal_coef > 0,5
personal_coef < 0,5 и > 0, 45
personal_coef < 0,45

Создаем 3 датафрейма по кластерам и находим наиболее популярные товары в этих кластерах.

1 кластер (personal_coef > 0,5)

Кроссовки 8568
Рюкзак 4662
Сандалии 4544
Футболка 3562
Бейсболка 3405
Носки 3141
Сумка 3121
Брюки 2994
Велосипед 2981
Спортивный 2910
Шорты 2872
Куртка 2724
Кеды 2412
Набор 2301
Палатка 2224
Шлепанцы 2068
Худи 2000
Сабо 1856
Мяч 1714
Полуботинки 1508

2 кластер (personal_coef < 0,5 и > 0, 45)

Велосипед	1927
Мяч	1213
Рюкзак	1199
Носки	1115
Сумка	1084
Бейсболка	1074
Кроссовки	861
Палатка	774
Набор	740
Сабо	563
Солнцезащитные	527
Перчатки	511
Спальный	400
Коврик	369
Футболка	367
Шлепанцы	352
Кеды	301
Шорты	297
Панама	292
Худи	278

3 кластер (personal_coef < 0,45)

Велосипед	7509
Рюкзак	5310
Бейсболка	4726
Носки	4223
Кроссовки	4174
Сумка	4071
Палатка	3999
Набор	3760
Мяч	3000
Солнцезащитные	2387
Сабо	2249
Коврик	2066
Спальный	2040
Перчатки	1666
Сандалии	1653
Футболка	1629
Шлепанцы	1543
Самокат	1327
Шорты	1263
Панама	1207

Провели анализ по кластерам насколько на покупку влияет наличие скидки.

вывод: дисконт не повлиял ни на общую сумму продаж ни на число покупателей, во всех кластерах покупателей.

7. Модель склонности клиента к покупке.

- Создаем единый датафрейм, где есть вся информация о покупателе и продаже товаров.

- Создаем датафрейм для покупателей из страны 32 города 1 188 и устанавливаем условие , что в dt != 0, т.е. только когда были покупки.

-Создаем столбец purchase, который и будет целевой переменной.

- Значение в этом столбце равно 1, если была одна покупка у покупателя и 2, если было покупок больше 1 раза.

- Данные в purchase оказались не сбалансированные , т.е. когда значение 2 = 181 55, а 1 = 2919.

Кодируем все наименования товара с помощью LabelEncoder и получаем датафре йм со следующей информацией:

personal_coef – персональный коэффициент

education1 – образование

base_sale – персональная скидка

product_sex продукт для конкретного пола (0- для женьщин, 1- для мужчин, 2- унисекс)

number_of_colors – количество цветов у продукта

age – возраст покупателя

product2 – наименование продукта

purchase - количество покупок одним покупателем , 1 – одна покупка , 2 – более 1 раза покупал один клиент

Так как у нас не сбалансированные данные , то разделяем датафрейм на два по столбцу purchase (где значения 1 и 2) и делим полученные датафрейма на трейн и тест в соотношении 70:30. Потом разделенные датафреймы обьединяем , что бы в трейне и тесте не было дисбаланса по значениям 1 и 2. Получили 4 датафрейма X_train7, X_test7, y_train7, y_test7. Их используем для предсказания , преварительно сделав стандартизацию с помощью MinMaxScaler.

Результаты моделей:

Accuracy логистической регрессии на трейне: 0.8615009151921904
Accuracy логистической регрессии на тесте : 0.8614581685908588

Accuracy случайного леса на трейне : 0.863399091587011
Accuracy случайного леса на трейне: 0.8611418630396963

Так как модели предсказывают на данных которые не видели так же как на данных для обучения, то можно сделать вывод что модели не переобучились и предсказы вают корректно.