Ситуация

В рамках улучшения продукта SoS по анализу поисковых запросов требуется анализ аномальных значений активностей отдельных респондентов по отдельным брендам и решение по работе с ними.

Описание данных

Поле	Пример	Тип данных	Описание
SubjectID	1585271561336880000	bigint	Индивидуальный ID респондента
QueryText	13 с redmi	string	Поисковый запрос
BrandID	216181	string	Индивидуальный ID бренда, который был определен в запросе
Category1ID	18	string	Идентификатор категории бренда
Category2ID	1801	string	Идентификатор подкатегории бренда
Category3ID	180199	string	Идентификатор подкатегории бренда второго уровня
Brand	redmi	string	Название бренда, который был определен в запросе
Category1	Гаджеты	string	Название категории бренда
Category2	Смартфоны	string	Название подкатегории бренда
Category3	Другое	string	Название подкатегории бренда второго уровня
CategoryDelivery	Смартфоны	string	Категория, идущая в поставку SoS
ResourceName	Aliexpress	string	Название ресурса
ResourceType	Универсальные маркетплейсы	string	Название типа ресурса Универсальные маркетплейсы, Другие ecom ресурсы, Поисковики
UseType	Mobile App	string	Тип использования Mobile App, Mobile Web, Desktop

Platform	Mobile	string	Источник данных Mobile, Desktop
Пол	мужчины	string	Пол респондента
Возраст	45-54	string	Возраст респондента
Регион	москва	string	Регион респондента
Федеральный_округ	центральный	string	Федеральный округ респондента
Количество_детей	нет	string	Количество детей у респондента
Занятость	работает	string	Занятость респондента
Доход	хватает на еду и одежду, но не на дорогие вещи	string	Доход респондента
Weight	2259.726	decimal(12,3)	Вес респондента (сколько людей в тысячах представляет респондент в конкретный день)
week_weight	2560.570	decimal(12,3)	Вес респондента недельный
month_weight	1993.047	decimal(12,3)	Вес респондента месячный
Start	2024-12-31 09:29:25	timestamp	Время отправки запроса
researchdate	2024-12-31	date	Исследовательские сутки
week	2024-12-30	date	Неделя Понедельник каждой недели
Month	2024-12-01	date	Месяц Первое число каждого месяца
inDelivery	1	int	Поставка SoS, 1 - входит в поставку SoS, 0 - не входит в поставку SoS

Задача

Нужно разработать алгоритм поиска аномальных респондентов по их активностям – взвешенному количеству запросов (OTS).

- I - множество всех респондентов в рассматриваемый период. Если респондент не попал в отчетную выборку какого-либо дня, полагаем его поисковую активность равной нулю.

- J - множество брендов некоторой категории, которые соответствуют условиям значимости бренда.
- Взвешенное число запросов респондента $i \in I$ бренда $j \in J$ в день k равно произведению дневного веса респондента на количество его обращений к бренду. Например, если респондент i с весом 5 трижды посещал бренд j в один день разными способами, то его число запросов по бренду равно 15.
- Для определения бренда как значимого необходимо, чтобы за рассматриваемый период взвешенное количество запросов (OTS) было больше или равно 20 000 хотя бы в одном месяце ***И*** размер выборки был больше или равен 6 респондентам хотя бы в одном месяце за рассматриваемый период. Для брендов в условиях отличия от 0 их OTS значений за месяц используются среднемесячные веса респондентов.

Пояснения

Вы должны разработать алгоритм, который будет находить дневные аномальные активности респондентов в значимых брендах. Для выявленного аномального респондента впоследствии применяется операция удаления его из выборки для получения более корректного анализа популярности брендов. При этом данный респондент удаляется не полностью из выборки, а только в тот день, когда он был признан аномальным. Но также важно, что в этот день удаляются все его активности, даже по тем брендам, контакт с которыми был признан нормальным. То есть при удалении аномалии в одном бренде уменьшаются запросы и в других брендах, и даже в других категориях.

Важно, что на поставку заказчику идут категории из столбца CategoryDelivery, именно по ним нужно проводить исследование.

А теперь объясняю «на пальцах»:

- У вас есть датасет с номерами респондентов, их характеристиками, типами и способами просмотра разных ресурсов, брендами и категориями просмотренных респондентами товаров.
- Вы находите ежедневные взвешенные количества запросов респондентов в брендах, а затем применяете свой алгоритм поиска аномалий среди них.

Метрики качества

Требования к решению со стороны заказчика:

- Оптимальный процент аномальных респондентов от общего количества респондентов в выборке равен такому числу x_0, при котором достигает своего минимального значения следующая функция:

$$a = 4 - 15 \cdot \tanh(1.2 \cdot (x - 4) - 1) + (x - 3)^2, x \in [0; 100]$$
$$x_0 \approx 5.8199, a_0 \approx -0.4782 \quad (1)$$

- Оптимальный средний процент аномальных респондентов в каждой категории от общего числа респондентов в этой категории равен такому числу x_0, при котором достигает своего минимального значения следующая функция:

$$b = \tanh(-0.5 \cdot x) + 0.1 \cdot x, x \in [0; 100]$$
$$x_0 \approx 2.8873, b_0 \approx -0.6057 \quad (2)$$

- Оптимальный процент суммарного OTS по всем брендам всех категорий после удаления всех аномалий от общего суммарного OTS до удаления равен такому числу x_0, при котором достигает своего минимального значения следующая функция:

$$c = 0.05 \cdot (x - 93)^2 + \tanh(93 - x), x \in [0; 100]$$
$$x_0 \approx 94.5753, c_0 \approx -0.7938 \quad (3)$$

- Оптимальный средний процент суммарного OTS по каждой категории после удаления всех аномалий от суммарного OTS до удаления в данной категории равен такому числу x_0, при котором достигает своего минимального значения следующая функция:

$$d = 0.05 \cdot |x - 93|^{1.5} + \tanh(95 - x), x \in [0; 100]$$
$$x_0 \approx 96.6283, d_0 \approx -0.5803 \quad (4)$$

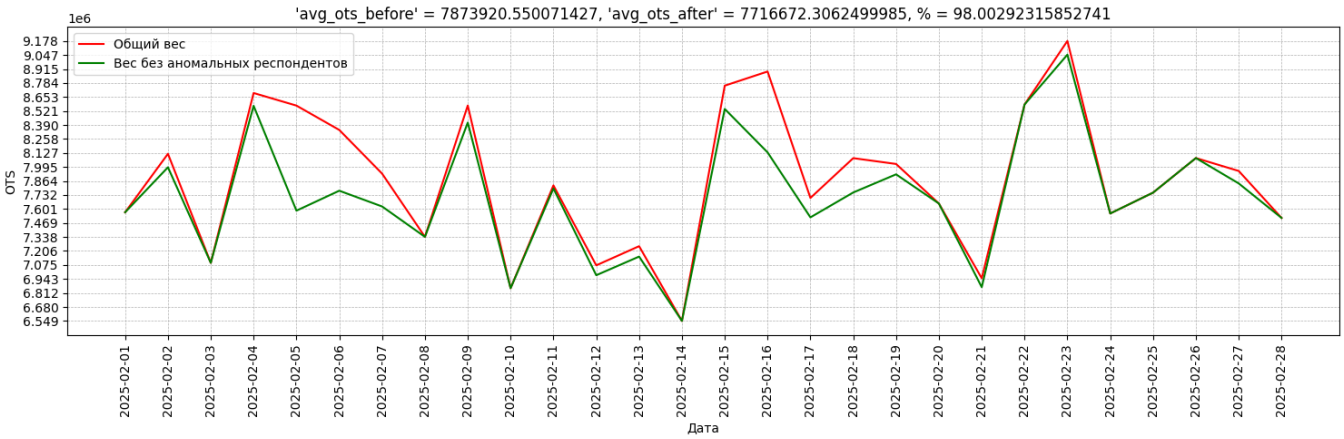
- Количество аномальных респондентов в день не превышает 8 человек
- Интерпретируемость алгоритма (он не должен брать числа и параметры «из воздуха»)
- Возможность смотреть аналитику по характеристикам респондентов и ресурсов
- Время работы программы не более 7 минут
- Аномальные активности характеризуются большим OTS, малый OTS не аномален

Результаты

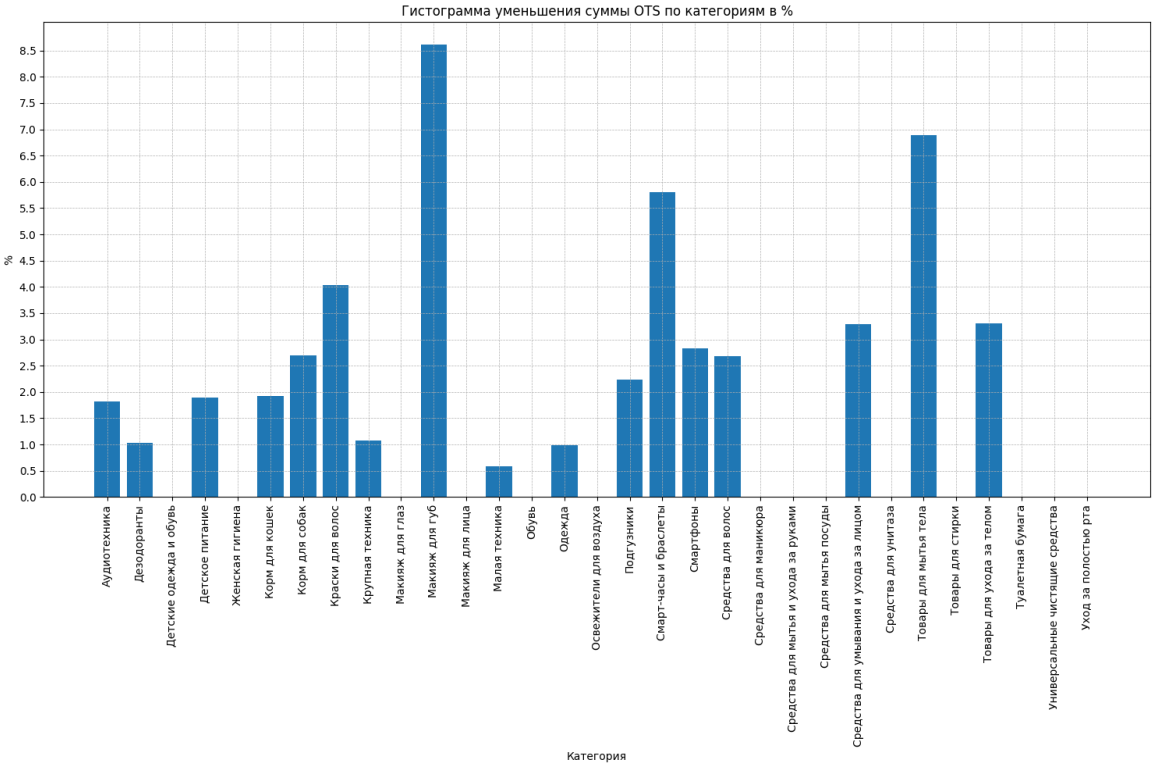
Результатом работы программы являются:

- Файл .py или .ipynb с решением
- Текстовое подробное объяснение алгоритма

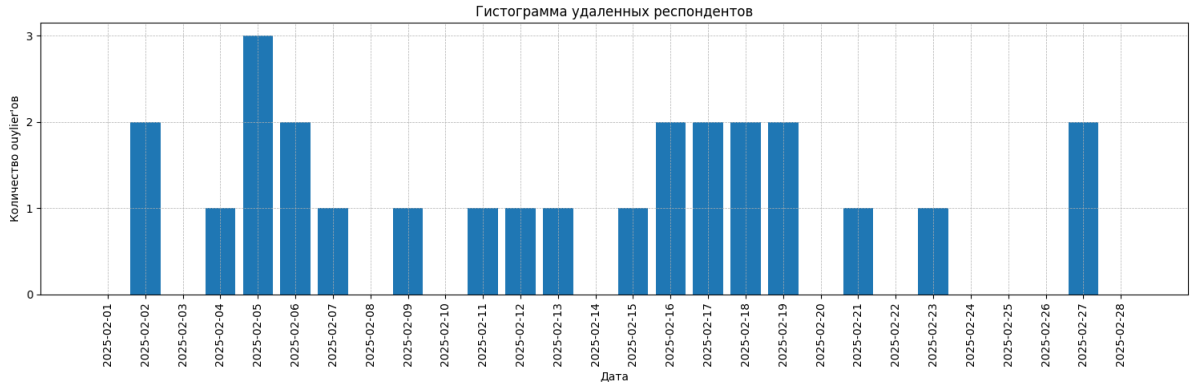
- Файл csv с аномальными респондентами (ID и день аномальности)
- График, на котором отражено изменение общего OTS (см. пример 1)



- Гистограмма, на которой показано, на сколько % изменилась OTS в каждой категории (см. пример 2)



- Гистограмма, на которой показано ежедневное количество аномальных респондентов (см. пример 3)



Критерии оценивания

Необходимые (но не достаточные) условия получения положительной оценки:

1. Понятное описание алгоритма, его интерпретируемость (а не просто «нейросеть выдала результат, почему такой – сам не знаю»)

2. Работа алгоритма на уровне брендов (то есть вы не можете строить анализ на уровне категорий, нужно погружаться глубже и анализировать временной ряд OTS для брендов)
3. Работающий код, запускающийся одним нажатием кнопки (в идеале файл .py) + файл csv с аномалиями и три картинки, создающиеся в процессе работы программы (например, в специальной папке в директории с основной программой)
4. Время работы алгоритма не более 7 минут

На основании ваших ответов будут посчитаны:

- % аномальных респондентов от общего количества респондентов
- Средний % аномальных респондентов в категории от общего количества респондентов категории
- % суммарного OTS по всем брендам всех категорий после удаления всех аномалий от общего суммарного OTS до удаления
- Средний % суммарного OTS по каждой категории после удаления всех аномалий от суммарного OTS до удаления в данной категории.

Полученные значения будут подставлены в формулы 1-4, каждая в свою соответствующую. В итоге будут получены четыре числа: *a, b, c, d*.

Последнее число *f* формируется так: за каждый день, в который количество аномальных респондентов в день превышает 8 человек, вы получаете 0,05 балл, а за каждую реализованную возможность аналитики из суммы будет вычитаться 0,1 балла. Всего предусмотрено 5 аналитических возможностей, поэтому минимальное число:

$$f_0 = \sum_{1}^{5} -0.1 + \sum_{1}^{0} 0.05 = -0.5 \qquad (5)$$

Под аналитическими возможностями подразумевается:

1. Возможность построения графиков до и после очистки от аномалий для разных характеристик респондентов («было\стало» в разбивке по полу, возрасту, округу и т.д.)
2. Возможность построения графиков до и после очистки от аномалий для разных характеристик ресурсов («было\стало» в разбивке по имени, платформе, типу и т.д.)
3. Возможность построения графиков до и после очистки от аномалий для разных уровней описания категорий («было\стало» в разбивке по Category1, Category2, Category3)
4. Возможность просмотра текста поисковых запросов аномальных респондентов
5. Возможность просмотра изменения OTS по дням для отдельного бренда (см. пример 1, но для конкретного бренда)

С полученными числами *a, b, c, d, f* будут проведены некоторые операции (какие конкретно – не скажу, потому что тогда начнете подгонять ответ или отдавать предпочтение одному критерию в ущерб другим). По итогам этих операций будет составлен ваш рейтинг и, соответственно, присвоены итоговые баллы. Для вас важно одно – чем меньше каждое из этих чисел, тем лучше. Идеальные значения данных чисел указаны в формулах выше.

Сроки выполнения задания

До утра 9:00 8 июня (воскресенье)

Формат ответа

Присылать файл .py (.ipynb) и текстовое описание на почту kuzovchikov98@mail.ru. Можно присылать несколько вариантов, но прошу особо не увлекаться и не спамить) В любом случае я буду проверять последний присланный вариант. В теме письма указать «Полусеместровый контроль ML – Группа – ФИО»

Дополнение

В качестве небольшой помощи высылаю 3 статьи с классификацией различных методов выявления аномалий (всего в статьях представлено более сотни алгоритмов).

Желаю творческих успехов!