По условию одна из основных целей этого сервиса, чтоб как можно больше операций клиентов проходили через офферы. Все действия для достижения этой цели можно разделить на две категории — изменение нашего предложения (списка офферов) и изменение поисковой системы офферов (интерфейса приложения). В первом варианте мы работаем со случаем, когда клиенты не нашли подходящий оффер ввиду его отсутствия. Во втором — со случаем, когда клиенты не нашли оффер в списке дотупных, не активировал его и операция прошла по стандартной схеме или покупка была совершена у конкурента нашего партнера (тоже без нашего оффера). Выбрав с каким сценарием работать, необходимо сформулировать гипотезу, спланировать эксперимент, собрать данные и обработать результаты.

**1. Подготовка к тесту**

Поиск слабого места

В нашем случае — это наши плохие сценарии (мало операций проходят по офферам и нужно увеличивать это число).

*Формирование гипотезы*. Зависит от выбора действий.

В случае изменения поиска нужно тестировать просто разный дизайн или разную разбивку по категориям (и соответственно новый дизайн), тогда альтернативной гипотезой будет то, что новый интерфейс или новые категории более интуитивно понятны и приводят к большему числу операций по офферам. Разрабатывать новый интерфейс нужно по каким-то соображениям — изменять только дизайн или саму логику представления наших данных.

Например, можно опросить коллег и узнать, что бы они изменили как пользователи в дизайне, и составить дизайн, учитывая эти рекомендации, но тогда естественно нужно еще учитывать возможность ухудшения и проводить two-side тест. Также можно действовать из каких-то других предположении, проанализировав какая именно часть интерфейса дает наш плохой сценарий (какие офферы пропускают чаще, где они находятся на экране, как эти блоки переставить и т.д.).

При поиске лучшей подачи данных (категории, поиск внутри категорий и т.д.), это тоже можно делать исходя из действий пользователей (заходил ли пользователь в нужную категорию при поиске, скроллил ли на нужной вкладке/экране и т.д.)

Если же идем в сторону улучшения предложения, то можно попытаться улучшить алгоритм персонализации. Альтернативной гипотезой будет то, что новый алгоритм персонализации лучше старого. Здесь нужно разрабатывать новый алгоритм. А еще необходимо учитывать, что для оценки работы алгоритма нужно n-ное количество времени, так как для корректных персональных рекомендаций необходимы данные для анализа. Соответственно, тут значительно возрастает время эксперимента. Пример как можно изменить алгоритм: добавлять бОльшую значимость категориям, где в теории будут чаще совершаться покупки (пользователи чаще покупают продукты, чем спортивные товары и т.д.).

Выбор метрики

Кажется логичным использовать процент покупок по офферам из всех покупок — это для каждого пользователя. Для обеих выборок эти данные надо будет как-то обобщить — средний процент покупок по офферам в группе, медиана и т. д. Сложно делать какие-то выводы только по среднему значению без визуализации и других оценок, то есть мы можем выяснить как друг от друга отклонились результаты средних в первой и во второй группе, но считать ли увеличение среднего значения улучшением — непонятно, поэтому я бы смотрела, у какого процента пользователей увеличился процент покупок по офферам в каждой группе. Такая метрика хорошо отражает идею альтернативной гипотезы, которую мы надеемся подтвердить — наши изменения приближают нас к цели, заданной в условии.

Оценка сложности внесения изменений и сложности поиска подходящих выборок

Для каждой из этих гипотез (про алгоритм и про новый интерфейс) могут понадобиться разные группы пользователей. Пользователи должны репрезентативно отражать всех клиентов, но при этом эти группы еще должны быть между собой однородны для каждой гипотезы. У обеих групп не должно отличаться поведение, поэтому должны соотноситься их проценты покупок по офферам, их поведение внутри приложения. Для тестирования алгоритма еще важно, первый ли это месяц использования или нет. Частота совершения покупок в принципе тоже играет роль. Возможно стоит определить способ определения похожести пользователей, какую-то сегментацию по признакам, важным для нашей ситуации. В тинькофф часто тестируют интерфейс и скорее всего уже есть методы отбора пользователей для таких тестов.

Кроме того, после сформулированной гипотезы, выбранной метрики, оценки сложностей, подбора пользователей и длительности неплохо было бы принять решение о том, что делать при подтверждении/отклонении гипотезы и заранее понять, какими способами будут проверяться гипотезы, так как есть много методов (и классических математических, и эмпирических). Для этого нужно понимания того, с каким типом данных работаем. Например, конверсия близка к нормальному распределению.

**2. Запуск и ожидание**

Здесь уже больше технических моментов. При полностью спланированном тесте необходимо будет лишь внести изменения и получить результаты, если сложность внесения изменений и необходимые для этого действия оценены в предыдущем этапе.

В нашем конкретном примере этот этап будет состоять из замены интерфейса или из замены самого предложения у группы, на которой будет проводиться тестирование. Ни первое, ни второе не должны вызвать особых сложностей, так как в Тинькофф тестируются различные дизайны, а изменение предложения на новое предполагает обновление данных.

**3. Подведение итогов**

Если решение о дальнейших действиях при всех возможных вариантах итога теста принято заранее, то никаких проблем с подведением итогов быть не должно. Однако, нельзя забывать об оценке практической значимости и целесообразности действий. В общем, нужно знать что делать с результатами. В некоторых случаях можно предложить какие-то действия по улучшению, если их нет.

Действиями после результатов будут внесения изменений у всех пользователей (изменение интерфейса, изменение алгоритма для формирования предложения) либо же наоборот оставление старых вариантов.