Как проводить АБ-тесты (часть 1). Задание 1.

Кейс

Тинькофф.Таргет - это агрегатор кэшбэков от партнеров (супермаркеты, магазины электроники, онлайн-магазины, заправки и т.п.).

Под каждого партнера собирается целевая аудитория и запускается таргет на ограниченный период, поэтому у каждого клиента свой набор офферов и они периодически меняются.

Механика: когда клиент совершает покупку по нашей карте в магазине X и у него есть **оффер** от этого партнера, он получает часть денег (какой-то фиксированный процент) за эту операцию обратно на карту.

Одна из основных целей этого сервиса - чтобы как можно больше операций клиентов проходили через **офферы**.

Например, если клиент хочет купить велосипед, то он сначала идет в наш сервис, смотрит есть ли у него кэшбэк в таком магазине и покупает велосипед по **офферу**.

Плохой сценарий: если у клиента нет подходящего **оффера** или клиент его не находит среди других (у каждого клиента в моменте 200-400 **офферов**)

Придумайте 2 кейса-гипотезы для достижения основной цели и опишите процесс проведения исследования и АВ теста для проверки этой гипотезы.

Комментарий:

При решении данного кейса я пользовалась следующей статьей: https://academy.yandex.ru/posts/kak-provesti-a-b-testirovanie-6-prostykh-shagov

Основная цель: чтобы как можно больше операций клиентов проходили через офферы.

Предположу, что численно эту метрику можно выразить как отношение количества операций клиента, совершенных через офферы, к общему количеству операций клиента.

На мой взгляд, есть несколько вариантов негативных сценариев:
- Операция есть, а оффера не было.
- Операция есть, оффер есть, но он не активирован.
- Оффер есть, операций нет.

Предположим, что мы посчитали в целом по всем клиентам следующие показатели:

- Количество операций, совершенных без активации оффера (при наличии оффера), деленное на общее количество операций. Оказалось, что этот показатель равен 10%.
- Количество операций, совершенных без активации оффера (при наличии оффера), деленное на количество операций, совершенных по офферам. Оказалось, что этот показатель равен 40%.

Мы считаем, что это слишком высокие показатели, которые сигнализируют нам о том, что люди пропускают офферы слишком часто.

У нас возникло 2 идеи:

1. Что если дать клиенту самому возможность выбирать офферы с повышенным кэшбэком? Например, можно дать клиенту каждую неделю выбирать 3 «любимых» магазина-партнера из предложенного списка.
2. Что если уменьшить количество офферов? Возможно их слишком много, и клиент не может найти подходящий, или взаимодействие с сервисом ему кажется неудобным.

Вариант 1:

Шаг 1. Определим цели.

В процессе исследования имеющих данных мы заметили довольно высокие показатели «анти-конверсии» офферов. Мы считаем, если дать клиенту возможность самому выбирать для себя офферы, то это должно повысить конверсию.

Шаг 2. Определим метрику.

В качестве метрики будем смотреть на 2 показателя:

1. Отношение количества операций клиента, совершенных через офферы, к общему количеству операций клиента. Назовем это OC (offer conversion).
2. Отток клиентов. Churn.

Будем надеяться, первый показатель вырастет, а второй хотя бы не изменится.

Шаг 3. Гипотеза.

Наша гипотеза заключается в следующем: Если мы дадим клиенту возможность самому выбирать интересные ему офферы, то, во-первых, клиент привыкнет пользоваться сервисом, научится находить то, что ему нужно, а, во-вторых, будет чаще совершать покупки по офферам (в том числе и тем, которые он не выбирал).

Однако, нельзя исключать, что клиент не пользуется офферами по каким-то другим причинам (например, потому что ему все равно, есть они или нет, и он не хочет тратить время на их поиски по сервису). А наши навязчивые предложения – выбрать «любимые» офферы, будут его раздражать, и он решит отказаться от услуг Тинькофф.

Получается, что нам нужно проверять 2 пары гипотез.

1. Нулевая гипотеза H0: OC останется без изменений.
   Альтернативная гипотеза H1: OC изменится после внедрения нашей идеи.
   Будем рассматривать двухстороннюю альтернативу. Мы не можем исключать, что наша идея плохая и конверсия станет хуже.

2. Нулевая гипотеза H0: Churn останется без изменений.
   Альтернативная гипотеза H1: Churn увеличится после внедрения нашей идеи.
   Будем рассматривать одностороннюю альтернативу. Наша идея не направлена на уменьшение Churn.

Шаг 4. Готовим эксперимент.

- **Версия A – это то, как есть сейчас.**
- **Создаем новую версию B:**

- Добавляем возможность выбрать 3 оффера в неделю из предложенного списка.

- Создаем объявление, которое будет отправлено пуш-уведомлением, о том, что у клиента появилась замечательная возможность самому решать, что покупать и за что получать кэшбэк. Добавляем 2 возможности: во-первых, если клиент не выберет, то система сама подберет ему лучшие предложения на основании анализа его прошлых предпочтений; во-вторых, возможность отключить напоминания о выборе «любимых» офферов (заодно потом можно проанализировать и этот показатель).

- Создаем систему пуш-уведомлений, которые 1 раз в неделю (по воскресеньям днем) будут напоминать, что клиент может выбрать 3 «любимых» оффера на следующую неделю.

- p-уровень значимости возьмем 0.05. А мощность 0.95.

- Выбираем контрольные группы:

Поскольку мы можем получить нежелательные результаты, стоит ограничиться каким-то определенным регионом. Если эксперимент окажется удачным, то можно расширить его на другие регионы. Например, возьмем Екатеринбург. Население – 1.5 млн человек. Предположим, что клиентами Тинькофф, являются 100 000 человек.

Воспользуемся этим калькулятором: https://mindbox.ru/ab-test-calculator/

Предположим, что ОС был 10%, мы хотим увидеть 11%. Тогда размер каждой выборки должен быть по 24 000 человек.

Предположим, что Churn был 5%, мы «не хотим видеть» 5.5%. Тогда размер каждой выборки должен быть по 50 000 человек.

Отлично, значит мы разделим клиентов Екатеринбурга на 2 выборки случайным образом. И только для одной из них поменяем интерфейс.

- Будем проводить тест в течение 2-х недель.

Шаг 5. Проводим эксперимент.

- Разработаем вариант Б.
- Посмотрим, как выглядит (проверим на нескольких сотрудниках).
- Согласуем тексты с маркетологами.
- Когда все будет готово, начнем проводить тест и через 2 недели посчитаем наши метрики.

Шаг 6. Анализируем результаты.

- Теперь воспользуемся этим калькулятором: https://abtestguide.com/calc/
- Вспомним, что наши выборки были по 50 000 тыс. человек.
- Предположим, что для группы А показатели остались такими же, как и были (ОС 10%, Churn 5%). А для группы В ОС составила 12%, Churn – 5.2%.

Для ОС мы получили статистически значимые различия (даже с учетом двухсторонней альтернативы). Доверительный интервал группы А не включает среднее группы В.

Для Churn мы получили p-value 0.0753, что не позволяет отвергнуть нулевую гипотезу (на что мы и надеялись). Однако, если бы Churn для группы В составил 5.24%, то p-value составил бы 0.0426, и мы могли бы предполагать, что наша «инициатива» все таки увеличивает Churn, что очень и очень плохо!

Прежде, чем внедрять эту инициативу повсеместно, я бы рекомендовала:

1. Продолжить эксперимент в Екатеринбурге еще на 2 недели.
2. Повторить эксперимент еще в 2-3 городах. Включить Санкт-Петербург или Москву.
3. Посмотреть на показатель отключения пуш-уведомлений на выбор «любимых» офферов. Если часть испытуемых сразу отключила эти уведомления, то результаты эксперимента стоит пересчитать (уменьшить размер выборки В на количество человек, которые сразу отключили эти уведомления). Если на второй неделе еще довольно значительное количество испытуемых отключили уведомления (например, 10%), то, возможно, стоит подумать над тем, как изменить систему уведомлений.
4. Я бы посмотрела и на другие показатели – например, увеличились ли показатели прибыли Тинькофф в этом регионе, увеличилось ли количество транзакций, какие «любимые» категории выбирают клиенты – совпадает ли это с прогнозами системы (которая делает рекомендации по офферам), угадывают ли клиенты свои самые крупные покупки (может быть они думают, что кэшбэк пятерочки будет выше, потому что они ходят каждый день туда, но на самом деле кэшбэк какой-то единичной покупки был бы выше).

Вариант 2:

Шаг 1. Определим цели.

Мы полагаем, что уменьшение количества офферов приведет к увеличению конверсии этих офферов в покупки. Мы основываем наше предположение на гипотезе, что 200-400 офферов – это слишком много, неудобно искать и, следовательно, какие-то офферы просто не находятся, или клиенты перестают пользоваться сервисом. Давайте сократим количество офферов до 20. Пусть их будет доступно одновременно только 20 и будем всем новым клиентам показывать только новую версию.

Настроим рекомендательную систему таким образом, чтобы она показывала клиенту только 10 офферов, которые, по ее мнению, будут наиболее востребованы клиентом, и 10 офферов рандомно.

Шаг 2. Определим метрику.

Будем проверять ту же метрику ОС - отношение количества операций клиента, совершенных через офферы, к общему количеству операций клиента. Ожидаем, что она вырастет.

Шаг 3. Гипотеза.

H0: Метрика останется такой же.

H1: Метрика изменится. Проверяем двухстороннюю гипотезу, потому что мы не уверены, что это «ухудшение» может позитивно сказаться на метрике.

Шаг 4. Готовим эксперимент.

- **Версия А** – так, как есть сейчас.
- **Версия В**:
  - Настраиваем рекомендательную систему.
  - Создаем объявление для пуш-уведомления, что вам каждый день доступно 20 предложений по кэшбэкам, спасибо, что выбрали Тинькофф.
  - p-value возьмем 0.05. А мощность 0.95.
  - после завершения эксперимента и в случае, если мы примем решение от него отказаться, стоит предусмотреть для клиента возможность выбора – оставить 20 предложений или увеличить до стандартных.
- **Выбираем контрольные группы.**
  - Поскольку эксперимент проводится только на новых клиентах, можно не ограничиваться каким-то одним регионом, а провести сразу по всей России (если проводить одновременно с первым экспериментом, то надо исключить Екатеринбург).
  - Воспользуемся этим калькулятором: https://mindbox.ru/ab-test-calculator/

Предположим, что ОС был 10%, мы хотим увидеть 11%. Тогда размер каждой выборки должен быть по 24 000 человек.

Предположим, что приток новых клиентов в Тинькофф составляет 10000 человек в неделю. Тогда для обеспечения требуемого размера выборки В, нам нужно 3 недели + хотя бы 2 недели на наблюдение за теми, кто пришел. Довольно много. 5 недель.

- Получается, что фактически, выборка А - это все клиенты Тинькофф на момент начала теста. Это не очень хорошо, слишком существенная разница в степенях свободы. Попробуем рандомно выбрать для выборки А 24000 человек, за которыми и будем наблюдать (мы считаем, что новые клиенты из выборки В и уже существующие клиенты – это выборки из одной генеральной совокупности, а по ЦПТ их выборочные средние и дисперсии должны быть близки со значениями генеральной совокупности при таком объеме выборки).

- Будем проводить тест в течение 5 недель.

Шаг 5. Проводим эксперимент.

- Разработаем вариант Б.
- Посмотрим, как выглядит (проверим на нескольких сотрудниках).

- Согласуем текст с маркетологами.
- Когда все будет готово, начнем проводить тест и через 5 недель посчитаем наши метрики.

Шаг 6. Анализируем результаты.

- Теперь воспользуемся этим калькулятором: https://abtestguide.com/calc/
- Вспомним, что наши выборки были по 24 000 тыс. человек.
- Предположим, что для группы А показатели остались такими же, как и были (ОС 10%). А для группы В ОС составила 9,5%.

Для ОС мы не получили статистически значимых различий. Доверительный интервал группы А включает среднее группы В. Калькулятор показывает p-value 0.0324 и мощность 74% (но там что-то не то. Если поменять выборки местами, то калькулятор фиксирует статистически значимые различия с тем же p-value и мощностью 83%).

Какие выводы можно сделать:

- Тест показал результат конверсии хуже, чем в базовом варианте. Не смотря на то, что статистически, это не значимые различия, внедрять данное изменение не стоит.
- То, что тест не показал значимых различий, но и не показал существенного снижения конверсии, может свидетельствовать о том, что рекомендательная система подбирает офферы не очень эффективно. Действительно, сокращение вариантов в 20-40 раз не привело к существенному сокращению конверсии. Стоит проанализировать более внимательно, как клиенты выбирают офферы:
    - Какой % офферов не активирован. Это происходит у одних и тех же клиентов или характерно для одних и тех же офферов.
    - Какие офферы выбирают реже всего. Это характерно в целом для категории, или для каких-то отдельных офферов.
    - Если ли связь между размером кэшбэка и частотой его активации и т.д.