A/B для продактов и аналитиков

В этом разделе инструкция для менеджеров по дизайну теста, постановке задачи на разработку и вопросам, связанным с расчетом метрик и подведением итогов.

ПОМНИТЕ: Основным владельцем A/B-теста, как и фичи, является продакт менеджер. Поэтому только он ответственен за то, что все будет задизайнено, запущено и посчитано вовремя.

Дизайн теста

Дизайн теста делается в первую очередь продакт менеджером, **аналитик выступает как консультант**, который помогает с определением конкретных метрик и техническими деталями заведения событий.
Идея заключается в том, что продакт менеджер является владельцем своего куска продукта и должен сам понимать, какие метрики являются ключевыми в его зоне ответственности. А значит, задачей аналитика является вносить уточнения в логику расчета ***уже после этапа первичного выбора метрик.***
*Далее помечены места, где точно следует просить помощи аналитика. В остальных пунктах нужно стараться принимать решения самостоятельно: * помощь аналитика*

Перед проведение теста жизненно необходимо ответить на следующие вопросы

(и, в идеале, описать их в документе в Confluence)

1. Какую гипотезу тестируем?

2. Что видят пользователи каждой из тестовых групп?

1. готов дизайн;
2. проведено UX-исследование (если не было проведено на этапе тестирования продукта);

3. Что мы ожидаем от проведения теста?

Тут может быть один из нескольких вариантов:

1. Мы хотим подтвердить гипотезу путем **значимого улучшения метрик.** В случае серого или красного эксперимента, мы не выкатываем этот вариант, а возвращаемся на этап генерации гипотез. Решение принимается по росту целевых метрик и **о тстутствию падения** ключевых бизнес метрик. Подробно см. пункт 4

2. Мы в любом случае должны выкатить какое-либо **продуктовое** изменение. Тогда нам нужно убедиться, что эксперимент не портит ключевые метрики. **Выкатываются серые или зеленые эксперименты.**

   Важно:
   Серый тест - это тест в котором ключевые метрики и бизнес метрики не показали значимого изменения.
   Что это в действительности означает?
   Это означает, что в текущей конфигурации чувствительности метрик и тестов, а так же объема трафика мы не смогли увидеть значимое изменение в рамках MDE (минимального эффекта, который можно задетектить). **Но это не значит, что эффекта нет!**
   Поэтому, **КАЖДЫЙ** серый тест - это риск, на который мы идем только в том случае, когда **НЕТ** варианта не выкатить какое либо изменение (например у нас есть обязательство перед инвесторами на внедрение нового способа авторизации).

   Отсюда - если мы провели эксперимент с новой фичей и она не показала значимого улучшения ключевых метрик, то мы считаем этот эксперимент **неудачным** и возвращаемся к проработке решения или вообще отбрасываем гипотезу. *Важно помнить, что только 10-20% наших гипотез срабатывают (опыт крупных компаний, например Microsoft), а это значит, что большинство гипотез скорее всего вредны, так как усложняют продукт.*
3. *Технический эксперимент* - по сути аналогичен варианту выше, но тестируется техническое изменение, которое не отражается на пользовательском опыте.
4. Ухудшающий или обратный эксперимент. Этот тип эксперимента подразумевает отключение уже внедренной фичи с целью проверки / актуализации оценок эффекта.

4. Какие метрики смотрим в тесте?

Все метрики в эксперименте делятся на три группы:

1. ***Базовые*** *метрики бизнеса.*
   a. Сейчас сюда входят:
      i. GMV per User
      ii. Average Check
      iii. Conversion
      iv. Gross Profit
      v. Average orders per User
      vi. Косты
   b. Эти метрики должны быть посчитаны в **каждом эксперименте** (и считаются авторасчетом для всех экспериментов). При формулировании гипотез желательно ориентироваться на их улучшение или как минимум отсутствие падения
2. ***Целевые*** *метрики для принятия решения.*
   a. Целевые метрики выбираются менеджером и аналитиком в процессе дизайна эксперимента и их набор определяется в первую очередь тем, на какие метрики и как влияет тестируемый функционал.
   b. Целевых метрик не должно быть больше 3. Эти метрики учитываются при принятии решения о цвете эксперимента.
3. ***Информационные*** *метрики*
   a. Все дополнительные метрики, которые аналитик и менеджер считают нужным смотреть в эксперименте.
   b. **Не следует ориентироваться** на них при принятии решений, но используем в качестве дополнительного источника гипотез или проверки на то, что все проходит нормально.

5. Какие срезы сравниваем в эксперименте? * помощь аналитика

В авторасчете для каждого эксперимента автоматически считаются результаты в разрезах новый/старый пользователь и по платформам (iOS / android / web). Остальные срезы сейчас необходимо считать вручную
Следует помнить: Чем больше сравнений в эксперименте - тем больше вероятность увидеть прокрас метрики, там где его нет. Количество сравнений - это # срезов * # метрик. Это происходит потому, что статистический тест по умолчанию может показать значимое изменение в 5% случаев, соответственно, при большом количестве сравнений вероятность увидеть хотя бы один прокрас стремится к единице(!)

6. Какой срок длительности у A/B теста? * помощь аналитика

Минимальный срок длительности эксперимента - **2 недели.** Такой срок выбран как минимальный по двум причинам:

1. При более коротких экспериментах мы рискуем не прокрасить метрики
2. При тестах, длительность которых не кратна 7 дням, мы неправильно учитываем внутринедельную сезонность

Возможны случаи, когда требуется проводить более длинный эксперимент (3/4 недели):

1. Ожидаемые эффекты малы и за две недели не получится получить требуемый объем трафика
2. В эксперименте хочется пронаблюдать длинную метрику (например, retention)

В некоторых случаях можно рассмотреть опцию проведения более короткого эксперимента. Если у вас имеется такая необходимость –– пишите в ~proj-exp-platform

ВАЖНО:

При оценке длительности эксперимента нужно по мере возможности планировать такой объем трафика, чтобы **минимальный детектируемый эффект (MDE) был >= ожидаемому эффекту.**

Имея оценки ожидаемого эффекта по целевым метрикам, оценки MDE для выбранных платформ и доли трафика можно получить из:

1. (Пока работает нестабильно) Таблица sandbox.mde_metrics
2. [OUTDATED] MDE базовых метрик
3. Для кейсов, которые не покрываются пунктами (1) и (2): [OUTDATED] Как оценить MDE/traffic size/мощность любой метрики

7. Как фильтровать трафик? * помощь аналитика

При желании можно пофильтровать трафик, чтобы убрать из эксперимента пользователей, на которых эксперимент не повлиял. Делается это с помощью фильтрации пользователей по факту наличия некоторого события за период эксперимента ('фильтрующего события'). Это можно сделать как с использованием авторасчета (указав фильтрующее событие при заведении теста в А/Б платформе, так и при ручной оценке эксперимента.

Замечание: *по нашим оценкам использование фильтрующих событий в большинстве случаев не оказывает существенного влияния на чувствительность экспериментов.*

8. Как будет приниматься решение?

Необходимо описать, какие решения будут приняты при всех возможных изменениях метрик. В целом, рекомендуемые типы решений такие:

1. Эксперимент зеленый или серый, никаких сомнений нет - катим
2. Эксперимент зеленый или серый, но есть сомнения - проводим тщательный пост-анализ, если все еще непонятно –– перезапускаем (например в другом месяце или в другом варианте реализации)
3. Эксперимент красный, гипотезу подтвердить не удалось
    a. Сначала следует искать потенциальные ошибки в реализации
    b. И только после этого думаем, почему гипотеза не валидна и ищем альтернативные гипотезы

9. Что делать в случае пересечения тестов? * помощь аналитика

При запуске тестов в параллель есть риск получить смещенную оценку эффекта – оценку не только нашей фичи, но и ее взаимовлияния с фичами, которые тестировались в это время на пересекающемся наборе пользователей.

Почему тесты могут быть подвержены эффекту от пересечений?

1. Неаддитивность фичей
    a. Один тест перекрашивает страницу в синий, другой – кнопку на этой странице в синий. Обе фичи положительно влияют на метрику, но если они пересекаются, мы получаем смещенный эффект – синюю кнопку на синей странице нельзя увидеть. Получившийся эффект не равен сумме эффектов первой и второй фичи.

Если два теста затрагивают одну и ту же функциональность и есть подозрение в неаддитивности:

- Рекомендуемый безопасный вариант (1) – разводить такие тесты во времени и запускать их друг после друга
- Рекомендуемый безопасный вариант (2) – разводить такие тесты по трафику и запускать одновременно на непересекающихся фрагментах трафика. При использовании не всего трафика особое внимание рекомендуется уделить оценке необходимого трафика для теста (см п.6)

  Как разводить тесты по трафику?

  - Тесты должны быть синхронизированы по соли. Что это значит: алгоритм сплитования будет использовать одну и ту же соль при разделении пользователей, и, тем самым, один и тот же пользователь в этих тестах будет попадать в одни и те же бакеты. К примеру, если тесты X и Y имеют одну соль, пользователь с определенным идентификатором N, попадающий в бакет 399 в тесте X, попадет в бакет 399 также и в тесте Y.
  - Тесты должны быть разнесены по разным бакетам. Например, если тест X занимает бакеты 0-500, то тест Y должен занимать **другие** бакеты, не пересекаясь с тестом X (например, 500-1000, 800-900 и другие вариации)

- Возможный, но не рекомендуемый вариант – проводить тесты одновременно и не разводить по трафику. После окончания теста оценивать, исказило ли пересечение результаты ([DRAFT] Пересекающиеся тесты). В зависимости от размера искажения оба теста будут корректными / оба неккоректными, но их можно будет полечить без перезапуска / оба нужна будет перезапускать.

Разработка и запуск теста

Дальнейшие шаги, которые должны быть предприняты для запуска теста:

1. **Менеджер создает задачу на разработку**, в которой должно быть описано:
    a. Поведение сервиса во всех вариантах тестирования; количество и названия групп.
    b. Соответствие параметра (параметр задается в админке A/B-платформы) или названия группы и желаемого поведения. В новом аппе на текущий момент используется параметр *feature_state* со значениями (0,1,2…)
    c. Сроки проведения теста *(длительность согласована с аналитиком)*
    d. Название теста
    e. Метка теста - это тот параметр, на который завязывается разработка и который будет использоваться как идентификатор теста в админке. *Отличие метки от id теста в том, что ее можно использовать в новых тестах, что позволяет перезапускать тесты, не дожидаясь новых клиентских релизов*
    f. Событие которое нужно разработать для фильтрации трафика, если такового в текущий момент нет *(согласовано с аналитиком)*
2. Менеджер **заводит эксперимент в админке.**
    a. Админка A/B платформы есть на проде и на стейдже. Тестовую админку **используют разработчики при создании теста** (просите разработчиков делать это самим). Ссылки на админки:
        i. Продакшн админка - https://bs-ui-ab-test-platform.k-prod.sbermarket.tech/ab
        ii. Тестовая админка - https://bs-ui-ab-test-platform.k-stage.sbermarket.tech/ab
        Basic Auth
        User: *fmcg-user*
        Password: *fmcg-pass*
        Авторизация в админку
        User: fmcg-user@instamart.ru
        Password: rxDQrjlY9uQ4JGSKyQj3
    b. В поле **"метка"** необходимо написать название метки, на которую будет ориентироваться разработка. Пример метки **"my_new_feature".** Это поле является аналогом **exp_id**, но позволяет не заязываться на него при разработке теста, а значит, позволяет его перезапускать.
    c. В эксперименте пока стоит указывать дату проведения больше, чем планируется, чтобы была возможность его продлить. На текущий момент это необходимо, так как нет функциональности продления A/B-теста.
    d. Для лучшего контроля за пересечениями тестов следует заполнить поле 'место в продукте' –– этап пользовательского флоу, на котором ваша фича оказывает воздействие.
    e. Необходимо прикрепить ссылку на тикет доски TESTS. Это нужно, чтобы по вашему эксперименту был автоматически посчитан экономический эффект. Подробнее про TESTS: Доски PROJ и TESTS в Jira
    f. Одна из групп должна быть **помечена как контрольная** - с ней будут сравниваться все остальные тестовые группы. В интерфейсе для этого есть специальная галочка
    g. **Названия групп. Крайне рекомендуется называть группы смысловыми названиями. Примеры:**
        - *A, B, C* - совсем неинформативно
        - *control, test* - все еще неинформативно но уже лучше
        - *control / without_bages, with_bages* - **идеально**
        - *without_bages, with_bages_option1, with_bages_option2* - **идеально для A/B/C-тестов**
    h. При необходимости, эксперимент можно синхронизировать с другим, то есть сделать так, чтобы тестовые и контрольные группы совпадали в нескольких экспериментах. Для этого существует параметр "соль".
    i. **Метрики**. В этом разделе необходимо добавить целевые и информационные метрики, чтобы они были посчитаны авторасчетом. Желательно указывать ожидаемые эффекты по целевым метрикам.
3. В случае, если тест будет считаться не через авторасчет, аналитик должен подготовить дашборд для мониторинга теста. Ссылку на дашборд следует прикрепить в документ с дизайном эксперимента.
4. **Как определить подходящую дату для запуска? * помощь аналитика**
    a. По календарю в админке A/Б платформы следует завалидировать, что в желаемые даты проведения нет потенциально опасных пересечений. Если пересечения есть / не получается определить по календарю, следует синхронизироваться с аналитиками соответствующего домена о возможности разнесения тестов.
5. **Запуск теста**
    a. Когда все предыдущие шаги сделаны, разработка готова и релиз с функциональностью выехал на прод, менеджер нажимает кнопку запуска теста в админке.

ВАЖНО: Автоматическая оценка результатов теста (авторасчет)

Для тестов по **anonymousId** доступен автоматический расчет результатов теста ([OUTDATED] Автоматический расчет метрик). **Для того, чтобы ваш тест считался авторасчетом, необходимы события AB Test Used On Client.** Подробнее о том, что нужно сделать, чтобы они присутствовали, описано в A/B для разработчиков. Рекомендуем перед запуском теста уточнить у разработчиков, что они читали эту доку и для общения с A/B платформой используют методы и ручки, обеспечивающие отправку событий AB Test Used on Client.

В случае, если необходимые события есть, **результаты (промежуточные и итоговые) вы сможете наблюдать в дашборде по эксперимента**. Ссылка на дашборд также появится на странице вашего теста в админке (кнопка 'Результаты').

Подведение итогов

1. Посчитать эффекты на метрики и их статистическую значимость

Целевым считается автоматическая оценка результатов эксперимента с помощью авторасчета. Если по каким-то причинам авторасчет не доступен, можно воспользоваться функциями и запросами из репозитория платформы. Следует самостоятельно убедиться, что запросы актуальны.

1. Мы считаем прокрашенными только те тесты, у которых значимо изменились базовые или целевые метрики. Это помогает уменьшить вероятность ложного прокраса и манипуляций с цифрами за счет того, что гипотеза четко определена перед началом теста.
2. Статистически значимым эффект считает при p-value < 0.05.

3. Подвести итоги

В документе с дизайном или новом документе подводятся результаты теста. Для подведения результатов рекомендуется использовать Шаблон результатов по-юзерного теста

3. Выкатка новой функциональности

1. Если в результате эксперимента было принято решение о том, чтобы включать какую-либо новую функциональность на всех, то есть возможность сделать это, не дожидаясь следующего релиза.
   Сделать это крайне рекомендуется в случае, если метрики зеленые и эксперимент показал хорошие результаты. В том случае, если победила контрольная группа, так же **очень рекомендуется** включить ее на 100% через админку (это поможет избежать дополнительных ошибок, которые могли произойти на стороне разработки).
2. Следующий этап - удалить A/B-тест в одном из следующих релизов. Задачку на удаление A/B желательно создать сразу, чтобы разработчики про нее не забыли.

Куда прийти с вопросом по A/B платформе?

- Канал ~proj-exp-platform в Mattermost