

Название эксперимента

Описание бизнес-проблемы, текущей ситуации, бейслайна

Какова текущая ситуация? Зачем вообще что-то менять?

Например: “Сейчас при первой покупке в skyeng доступны пакеты уроков: 8, 16, 32, 64, конверсия в покупку по каждому пакету такая-то. Пакеты стоят от X тысяч рублей, что для пользователя может быть значительной суммой и в случае неуверенности, он откажется от покупки. Хотим добавить пакет 4 урока, чтобы повысить конверсию в первую покупку, запустив в воронку “сомневающимся”. Ожидаем, что прирост конверсии перекроет снижение среднего чека первой покупки и позволит больше зарабатывать”

Описание user story

“Как (роль), Я (действие), Чтобы (ценность)”

Пример: “Как новый клиент, никогда не занимавшийся в Skyeng, я хочу взять сначала маленький пакет уроков, чтобы решить, подходит мне это или нет.”

Гипотеза

“Если сделать <то-то> для <такого-то пользователя>, то <метрика Z> повысится на <10(%)>, потому что <то-то>”.

Пример: “Если для новых клиентов добавить на страницу оплаты пакет 4 урока, то суммарная конверсия из заявки в оплату повысится с 10% до 12%, потому что согласно исследованию такому-то у нас более холодный трафик по сравнению с прошлым годом”

Подсчет прибыли от решения проблемы

Сколько денег компания заработает на решении этой проблемы?

*Не забудьте учесть, на какую долю пользователей мы сможем оказать влияние.
Пример: “Прирост конверсии в первую покупку на 2% при сохранении среднего чека позволит в течение год заработать на 5 млн рублей больше. Однако эксперимент направлен только на пользователей мобильного приложения, которые составляют 20% от всех клиентов, то есть ожидаемый профит - $5 \cdot 0,2 = 1$ млн рублей”*

Варианты решения / эксперимента

Какой самый дешевый вариант по разработке теста?

Тут считаем ROI и принимаем решение, запускать вообще эксперимент или нет

Описание эксперимента

Какая выборка будет в тесте, однородная ли она? Как именно их будем отбирать?

Пример 1: в эксперименте участвуют 100% новых пользователей, группы определяются случайно в момент первого захода на страницу сайта. Однородность разделения протестирована на АА-тесте системы сплитования.

Пример 2: группы эксперимента формируются вручную таким образом, чтобы группы были однородны по влияющим на метрику сегментам и разница метрики не была статистически значимой на исторических данных. Также доказано, что с течением времени пользователь сохраняет свои свойства, то есть его метрика в течение времени отличается не более, чем на X%, что ниже заложенного эффекта (в противном случае разделение на исторических данных не гарантировало бы однородность групп в течение теста, так как пользователи могли бы менять свое поведение в течение времени вне зависимости от введенных изменений)

Какие именно будут группы/варианты? Что будут видеть пользователи в каждой из групп?

Помните, что тестировать нужно только одно изменение на раз, иначе вы не сможете отличить, что именно повлияло на поведение пользователей и не сможете сделать выводы по итогам теста.

Что будет происходить на бэкенде в каждой из групп?

Чем будет отличаться бизнес-процесс?

Пример: пользователям из контрольной группы будут

Необходим ли контроль за исполнением БП? Если да, то как он будет организован?

Пример: необходимо отслеживать, что работники отдела sales проводят продажи для пользователей тестовой группы по тестовому скрипту, а для пользователей контрольной группы - по текущему скрипту. Ежедневно административными ассистентами будут прослушиваться 50 случайных записей разговоров и размечаться по чек-листу (ссылка на чек-лист).

Ссылка на описание задачи разработчикам

Ссылка на макеты дизайна

Ссылка на чат с обсуждением эксперимента

Метрики, по которым будем оценивать эксперимент

Основные метрики (что хотим улучшить)

Пример: C1 - конверсия из регистрации в первую оплату

Добавочные метрики (что хотим не уронить)

Пример: средний чек первой оплаты, C2 - конверсия в повторную покупку

Планируемый прирост и эффект (S/M/L эксперимент)

При каких развитиях событий и получения каких результатов **какие решения я принимаю**

Какой должен быть размер выборки?

Когда запускаем эксперимент?

Когда останавливаем набор в группы?

Сколько дней от запуска эксперимента до остановки набора пользователей (*ожидаемая продолжительность*)?

Не забудьте учесть, на какую долю пользователей мы сможем оказать влияние.

Пример: размер выборки - по 2500 наблюдений в каждую группу. Группы делятся 50/50, дневная аудитория сайта - 2000 пользователей, доля мобильного трафика, для которого производится изменение - 20%. Таким образом, понадобится $2 \cdot 2500 / (0,20 \cdot 2000) \sim 13$ дней.

Нежелательно проводить тест “несимметричное неделе” число дней.

Пример: не стоит проводить тест с понедельника 1.01 по субботу 13.01, несмотря на достижение необходимого числа наблюдений. В таком случае в тест попадет по 2 дня с понедельника по субботу и только одно воскресенье, что при наличии “внутринедельной сезонности” может исказить результаты. К примеру, у вас есть какой-то сегмент пользователей, которые заходят на сайт в воскресенье и их доля в эксперименте будет в 2 раза ниже, чем в реальности.

Сколько ждем после остановки набора до подсчета целевой метрики?

Пример: на исторических данных мы определили, что 95% оплат приходят в течение 7 дней после того, как пользователь положил товар в корзину. Таким образом, после остановки эксперимента стоит дождаться закрытия 7-дневного окна конверсии для подсчета результатов.

Не поменяется ли аудитория/условия эксперимента, пока будет идти тест?

Пример: в период проведения теста попадают майские праздники. В это время мы усиливаем/замораживаем маркетинговые активности, пользователи отличаются от тех, которые приходят в “непраздничные” периоды, и это отличие может оказать эффект на метрику. Поэтому наблюдения, собранные в этот временной интервал, мы при подведении итогов теста не учитываем.

План работ и ответственные

Кто когда что делает?

Кто в целом отвечает за эксперимент?

Ход эксперимента

Дата фактического запуска:

Критические проблемы, искажающие результат:

Комментарии:

Дата фактического завершения:

Фактическая продолжительность:

Итоговые расчеты

Статистически значима ли разница между группами?

Группы действительно однородны по основным факторам, влияющим на целевую метрику? (например, распределение по маркетинговым каналам трафика или географии)

Были ли нарушения бизнес-процесса, могли ли они повлиять на результат?

Были ли пересечения с другими экспериментами, могли ли они повлиять на результат?

Ссылка на итоговый расчеты

Результаты эксперимента и выводы

Какие выводы сделали:

Какое решение приняли в результате и почему:

- Неуспешный, не катим на всех
- Неуспешный, НО катим на сегмент1, сегмент2, сегмент3
- Успешный и катим на всех
- Успешный, НО не катим на всех